# INFORME SOBRE SISTEMAS CONVERSACIONALES MULTIMODALES MULTILINGÜES

Tecnologías y Arquitecturas para el Desarrollo de Asistentes Virtuales, Sistemas de Diálogo y Otros Interfaces Conversacionales

**Autores:**
Quesada Moreno, Jose Francisco (Universidad de Sevilla)
Callejas Carrión, Zoraida (Universidad de Granada)
Griol Barres, David (Universidad de Granada)

**Plan de Impulso de las Tecnologías del Lenguaje**

Noviembre 2019

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA
SECRETARÍA DE ESTADO PARA EL AVANCE DIGITAL
ontsi observatorio nacional de las telecomunicaciones y de la SI
red.es
Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

**Colaboradores:**

Pro Martín, José Luis (Universidad de Sevilla)

Sierra Márquez, Pablo (Universidad de Sevilla)

**Contribuciones:**

Aguirre Bengoa, Eneko (Universidad del País Vasco)

Alías Pujol, Francesc (La Salle)

Dahl, Deborah (Conversational Technologies)

Hillmann, Stefan (Technical University Berlin)

Möller, Sebastian (Technical University Berlin)

Mariani, Joseph (LIMSI-CNRS)

Martínez Hinarejos, Carlos David (Universitat Politècnica de València)

McTear, Michael F. (University of Ulster)

Riccardi, Giuseppe (Universidad de Trento)

Ureña López, Luis Alfonso (Universidad de Jaén)

Wacker, Philippe (LT-Innovate)

Wanner, Leo (Universitat Pompeu Fabra)

Weiss, Benjamin (Technical University Berlin)

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital y Red.es, que no comparten necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

# RESUMEN EJECUTIVO

Establecer una conversación natural, ágil y fluida con una máquina utilizando el lenguaje natural como herramienta básica de interacción ha sido uno de los retos de investigación en los campos del Procesamiento del Lenguaje Natural, Lingüística Computacional y Tecnologías del Lenguaje [1].

Este reto ha captado constantemente el interés en el ámbito académico, comercial e industrial, teniendo en cuenta especialmente el amplio rango de aplicaciones de este tipo de sistemas. De hecho, diferentes informes recientes resaltan cómo el uso del lenguaje natural (y principalmente de la voz) está cambiando la forma en la que nos relacionamos con la tecnología, con unas perspectivas de crecimiento en tecnologías, sectores y dispositivos que indican que nos encontramos en la "era de los interfaces conversacionales":

- Gartner estima que en 2020 el 75 % de los hogares norteamericanos contará con un dispositivo de voz;
- Según Comscore, un 20 % de las búsquedas en Android son vía voz y se espera que para 2020 sean un 50 %;
- Edison Research indica que existen 43 millones de personas que poseen un asistente de voz en Estados Unidos a finales de 2018;
- Data Bridge Market Research describe que se espera que el mercado mundial de la voz, conversación y tecnologías asociadas alcance los 6.770 millones de dólares para 2025, con un crecimiento anual del 25.7 % en el período de 2018 a 2025;
- Capgemini señala que más de la mitad de los usuarios de dispositivos móviles (51 %) son ya usuarios de asistentes de voz.

Los asistentes personales, conocidos también como asistentes personales virtuales, asistentes personales inteligentes, asistentes personales digitales, asistentes móviles o asistentes conversacionales, se han convertido en una de las herramientas más innovadoras a la hora de simplificar y hacer más natural la interacción entre humanos y máquinas. Ejemplos bien conocidos de estos interfaces son Siri de Apple, Google Now, Microsoft Cortana, Amazon Alexa, Samsung S Voice, M de Facebook y Nuance Dragon.

Como ejemplo gráfico de esta evolución en el desarrollo y uso de interfaces conversacionales, la Figura 1 muestra la tendencia en las búsquedas de la palabra *chatbot* en el buscador Google desde el año 2004. De forma similar, una búsqueda de "asistentes personales" en Google Play hacia finales de marzo de 2019 genera como resultado más de 240 aplicaciones. Muchos de estos asistentes ayudan a los usuarios a realizar una variedad de tareas en sus teléfonos inteligentes, como obtener información mediante la búsqueda por voz, encontrar restaurantes, obtener indicaciones sobre rutas, configurar alarmas, realizar conversiones entre monedas, actualizar el calendario y participar en conversaciones generales. Otros asistentes proporcionan funciones más especializadas, como el control de la condición física, la preparación personalizada de bebidas o la planificación de recetas.

Aunque los interfaces conversacionales se han generalizado con el auge y evolución de los teléfonos inteligentes, actualmente también se integran en otros dispositivos, como relojes inteligentes, robots sociales y altavoces inteligentes como Amazon Echo o Google Home. En el futuro, podemos prever que los interfaces conversacionales serán un componente integral del llamado Internet de las Cosas (IoT), una red masiva de objetos conectados, sensores y dispositivos que se "comunican" entre sí y, en muchos casos, también se comunican con los seres humanos.

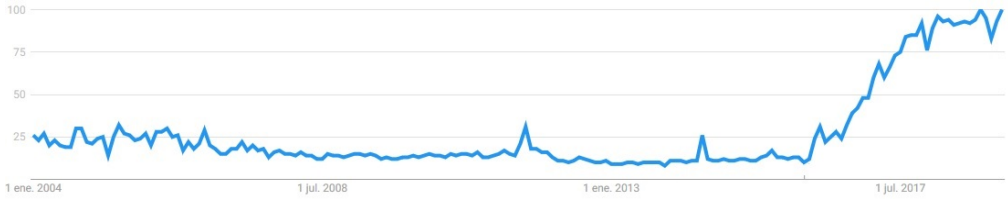


*Figura 1: Tendencia desde 2004 en las búsquedas en Google de la palabra Chatbot*

El interés tanto académico como industrial en el desarrollo de estos asistentes ha fomentado múltiples proyectos de investigación financiados por distintas agencias tanto a nivel internacional (a nivel europeo fundamentalmente a través de los distintos programas marco de la Unión Europea entre los que se incluye H2020), a nivel nacional, autonómico, etc. De forma paralela, han surgido distintas empresas centradas en el diseño de sistemas o interfaces hombre-máquina dirigidos a la implementación de sistemas conversacionales. Cabe además resaltar que el ritmo de creación de empresas en este sector ha aumentado considerablemente en los últimos años [2].

El interés académico e industrial, además de fomentar los proyectos y líneas de investigación o la creación de un elevado número de empresas tanto creadoras de tecnología como interesadas en su uso, ha provocado una cierta confusión terminológica en este campo, con términos tales

como *Sistemas de Diálogo Hablado*, *Asistentes Virtuales*, *Chatbots* o *Sistemas conversacionales* entre otros, cada uno de ellos con mayor o menor éxito en distintos escenarios.

Desde el punto de vista de los objetivos de este informe, utilizaremos el término *sistema conversacional* para hacer referencia a **las aplicaciones o sistemas informáticos con los que es posible comunicarse sosteniendo una conversación en lenguaje natural, bien escrito o hablado.**

Este enfoque enfatiza dos ideas clave. En primer lugar, la utilización de lenguaje natural (escrito o hablado) como soporte a la comunicación. En este sentido, los sistemas conversacionales se incluyen dentro del ámbito más general de las Tecnologías del Lenguaje. En segundo lugar, el soporte a una interacción conversacional o dialogada. Es decir, la interacción se realiza a través de una serie de turnos en los que usuario y sistema van intercambiando mensajes en lenguaje natural.

Establecido el objeto de estudio de este informe, cabe resaltar que el enfoque seguido para su elaboración se inserta dentro de las directrices clave del **Plan de Impulso de las Tecnologías del Lenguaje**[1], una iniciativa actualmente dependiente de la *Secretaría de Estado para el Avance Digital (Ministerio de Economía y Empresa)*. Entre los objetivos clave de este Plan merece la pena destacar los siguientes:

- Desarrollo de infraestructuras lingüísticas;
- Promoción de la Industria de las Tecnologías del Lenguaje Humano;
- Impulso de la Administración Pública como promotor de la Industria de las Tecnologías del Lenguaje.

A pesar de la difusión y uso actual de las tecnologías conversacionales, nuestra experiencia en el sector nos demuestra constantemente un cierto desconocimiento de los enfoques, tecnologías y modelos clave relacionados con estas tecnologías. Es por ello que, teniendo en cuenta los objetivos del Plan de Impulso de las Tecnologías del Lenguaje, el presente informe pretende presentar a todo el tejido industrial los conceptos clave relativos a los sistemas conversacionales.

El informe cubre un doble reto: por un lado, servir como documento genérico capaz de introducir los conceptos básicos, enfoques, técnicas, metodologías, etc. de los sistemas conversacionales a

[1] https://www.plantl.gob.es/Paginas/index.aspx

un público genérico, perteneciente a distintos sectores, pero especialmente a los sectores industrial y a la Administración Pública. En segundo lugar, presentar de forma sistemática y científicamente motivada dichos conceptos, enfoques, técnicas y metodologías. Es decir, creemos que, aunque existe una gran cantidad de documentación básica (en algunos casos creada por las propias empresas del sector), ésta no aborda una descripción amplia y detallada de los principios básicos de esta disciplina.

## ESTRUCTURA DEL INFORME

Siguiendo los principios descritos en la sección anterior, la estructura del informe se ha organizado en varios bloques temáticos.

### *Introducción a los Sistemas Conversacionales*

Los primeros tres capítulos abordan una panorámica general, así como de los fundamentos de los sistemas conversacionales, en el contexto de las tecnologías del lenguaje y el estudio del ecosistema tecnológico que enmarca el diseño e implementación de este tipo de sistemas:

- Capítulo 1: *Sistemas conversacionales y tecnologías del lenguaje: panorámica general*
- Capítulo 2: *Sistemas conversacionales: fundamentos*
- Capítulo 3: *Ecosistema tecnológico*

En primer lugar, se presenta una introducción de la noción de sistema conversacionales y se describe la arquitectura básica de un sistema conversacional, cuyos módulos se asocian con las principales tecnologías del lenguaje y niveles de descripción lingüística que utilizan estos sistemas: reconocimiento y síntesis del habla, comprensión y generación del lenguaje natural y gestión del diálogo. Estas tecnologías se describen en profundidad en el Capítulo 3, centrado en el estudio del ecosistema tecnológico que rodea la construcción de un sistema conversacional.

En el Capítulo 2 se describe cómo se han ido desarrollado los sistemas conversacionales a lo largo del tiempo, terminando con el contexto actual y haciendo especial énfasis en su relación con el ámbito de la ciencia y tecnología de la conversación, especialmente en el estudio de la estructura del discurso orientada a la información.

***Características básicas de un Sistema Conversacional***

La siguiente parte del informe incluye los Capítulos 4 a 7, a través de los cuales se abordan aspectos y funciones clave de un sistema conversacional.

- Capítulo 4: *Gestión del diálogo: control y modelado del diálogo*
- Capítulo 5: *Multimodalidad, multilingüismo, emociones y sistemas conversacionales afectivos*
- Capítulo 6: *Protocolos y estándares*
- Capítulo 7: *Adaptación y modelado del usuario y el contexto*

Este parte comienza con el Capítulo 4 profundizando en la pieza que constituye el núcleo de un sistema conversacional: el control y modelado del propio diálogo.

Interactuar con un sistema conversacional exige utilizar lenguaje natural a través de al menos un canal. No obstante, en muchas ocasiones se deben diseñar sistemas que soporten más de un lenguaje o incluso más de una fuente o canal de información además de la señal acústica (en lenguaje hablado) o la cadena de texto (en lenguaje escrito). Los retos que aparecen tras estos objetivos se explican en el Capítulo 5. Por otro lado, en el Capítulo 6 se describe que establecer la propia comunicación de entrada y salida, así como la gestión del diálogo y otra información, tal como las emociones que forman parte de la comunicación exige la utilización de distintos estándares para la representación de esta información y protocolos para su transmisión.

El usuario como agente relevante en la comunicación juega un rol especial en el diseño, implementación y la evaluación de un sistema conversacional. Por ello, el Capítulo 7 aborda el papel del usuario, así como el contexto comunicativo, desde los puntos de vista de la adaptación y el modelado en un sistema conversacional.

***Sistemas conversacionales: Diseño, implementación y evaluación***

Diseñar un sistema conversacional se puede concebir en última instancia como el desarrollo de una aplicación informática. Por tanto, es importante conocer qué tipo de arquitecturas, herramientas, entornos y técnicas se han desarrollado para tal fin.

La incorporación del lenguaje natural añade además al proceso de desarrollo y evaluación unos condicionantes específicos que se deben tener en cuenta, lo que justifica la importancia de los corpus de datos las metodologías estadísticas en el desarrollo y evaluación de sistemas conversacionales.

La tercera parte del informe aborda las cuestiones relativas a esta dimensión de los sistemas conversacionales, desde su diseño, hasta la implementación y evaluación funcional de los mismos:

- Capítulo 8: *Plataformas, arquitecturas y herramientas*
- Capítulo 9: *Corpus de datos y evaluación*

El Capítulo 8 describe las principales arquitecturas, plataformas y herramientas propuestas en el ámbito académico e industrial para la implementación de sistemas conversacionales. El capítulo 9 se centra en la presentación de las metodologías y corpus de datos disponibles tanto para tareas de entrenamiento y modelado de sistemas de diálogo como para tareas y retos (challenges) de evaluación y control de calidad.

***Sistemas conversacionales: Investigación y Retos***

Por último, la cuarta parte del informe analiza los retos actuales en el desarrollo de sistemas conversacionales teniendo en cuenta la situación actual tanto a nivel de investigación como de aplicaciones industriales y comerciales:

- Capítulo 10: *Panorama actual, tendencias y oportunidades*

## CONTRIBUCIONES

El informe incluye así mismo las siguientes contribuciones de investigadores y expertos en el campo de los sistemas conversacionales:

- **Aguirre Bengoa, Eneko** (Universidad del País Vasco - Coordinador del proyecto LIHLITH): *LIHLITH: Mejorando los sistemas de aprendizaje a través del aprendizaje continuo*
- **Alías Pujol, Francesc** (La Salle, Universitat Ramon Llull - Coordinador del Grupo de Investigación en Tecnologías Media): *El proyecto GENIOVOX*, *El proyecto UNISON*

- **Dahl, Deborah** (Conversational Technologies, Estados Unidos): *Retos para la adopción de estándares en sistemas conversacionales*
- **Mariani, Joseph** (LIMSI-CNRS, Université Paris-Saclay, Orsay, France): *Sistemas conversacionales: ¿el reto definitivo?*
- **Martínez Hinarejos, Carlos** (Universidad de Valencia - Coordinador de la RTTH) *Presentación de la Red Temática de Tecnologías del Habla*
- **McTear, Michael F.** (University of Ulster, Reino Unido) *Interfaces conversacionales: pasado y futuro*
- **Riccardi, Giuseppe** (Universidad de Trento, Italia - Coordinador del proyecto SENSEI): *SENSEI: dar sentido a los datos de conversaciones persona-persona*
- **Ureña López, Luis Alfonso** (Universidad de Jaén, Presidente de la SEPLN): *Presentación de la SEPLN - Sociedad Española para el Procesamiento del Lenguaje Natural*
- **Wacker, Philippe** (LT-Innovate): *El sonido de la inteligencia lingüística*
- **Wanner, Leo** (Universitat Pompeu Fabra - Coordinador del proyecto KRISTINA): *KRISTINA: Un agente virtual socialmente competente en el ámbito de la salud*
- **Weiss, Benjamin; Hillmann, Stefan; Möller, Sebastian** (Technical University Berlin, Alemania): *Calidad del sistema en sistemas de diálogo conversacionales y hablado*

Plan TL

# Índice

Plan TL

**Parte I**

# INTRODUCCIÓN A LOS SISTEMAS CONVERSACIONALES

# 1. SISTEMAS CONVERSACIONALES Y TECNOLOGÍAS DEL LENGUAJE: PANORÁMICA GENERAL

Este capítulo ofrece una visión de conjunto de los conceptos e ideas clave de los sistemas conversacionales.

## 1.1. SISTEMAS CONVERSACIONALES

La primera sección del capítulo describe las principales tecnologías de Procesamiento de Lenguaje Natural que se utilizan para el desarrollo de interfaces conversacionales, su definición básica, diferentes terminología empleada para referenciarlos y principales aplicaciones.

### 1.1.1. *Tecnologías del lenguaje natural orientadas a la interacción dialogada entre usuario y sistema*

Construir un sistema automático que sea capaz de mantener una conversación con una persona ha sido uno de los retos de la investigación y desarrollo en tecnologías del lenguaje desde sus mismos orígenes. Durante las últimas décadas, han aparecido múltiples enfoques, técnicas y sistemas tanto a nivel académico como industrial centradas en este objetivo, lo que a su vez ha provocado una proliferación de términos para nombrar esta línea de trabajo, por ejemplo, *sistemas de diálogo hablado*, *chatbots*, *asistentes virtuales* o *sistemas conversacionales*.

Desde el punto de vista de los objetivos de este informe, utilizaremos el término *sistema conversacional* para hacer referencia a **las aplicaciones o sistemas informáticos con los que es posible comunicarse sosteniendo una conversación en lenguaje natural, bien sea escrito o hablado.**

Tal y como se ha destacado en el resumen ejecutivo, este enfoque enfatiza dos ideas clave. En primer lugar, la utilización de lenguaje natural como soporte a la comunicación. En este sentido, los sistemas conversacionales se engloban dentro del ámbito más general de las Tecnologías del Lenguaje. En segundo lugar, el soporte a una interacción conversacional o dialogada. Este tipo de aplicaciones requieren una secuencia de interacciones (turnos de diálogo) entre la persona y la máquina para conseguir que el usuario consiga su propósito. Así, el objetivo del usuario se alcanza gradualmente a través de una serie de turnos en los que usuario y sistema van intercambiando mensajes en lenguaje natural.

### 1.1.2. *Las plataformas conversacionales: uno de los principales retos tecnológicos*

En un reciente informe sobre las 10 tendencias tecnológicas estratégicas de 2018 [2] la consultora Gartner menciona las *plataformas conversacionales* ( *Conversational Platforms*) como una de las líneas clave. La noción de plataforma conversacional es similar a la de sistema conversacional utilizado en este informe, y compartimos la idea de que uno de los principales retos que estos sistemas deben abordar es la limitación que muchos entornos comerciales actuales imponen al usuario al obligarle a utilizar un modelo de comunicación fijo y muy estructurado. Por el contrario, los sistemas o plataformas conversacionales deben resolver el reto que supone una mayor robustez en los modelos conversacionales, así como la implementación de mejoras en las APIs y los modelos de eventos usados para acceder, invocar y orquestar los servicios de terceros para conseguir resultados complejos.

### 1.1.3. *Definición básica de sistema conversacional*

Resumiendo, un sistema conversacional puede entenderse como un sistema automático capaz de emular a un ser humano en un diálogo con otra persona, recibiendo como entrada y generando salidas en lenguaje natural.

Este hecho obliga a dotar al sistema de la funcionalidad necesaria para que pueda:

- Hacer referencia durante el diálogo a la información que haya aparecido anteriormente;
- Tomar la iniciativa para reconducir el diálogo dentro del o los dominio/s en los que se ha definido;
- Solicitar información necesaria para cumplir el objetivo solicitado;
- Requerir aclaraciones cuando existan dudas sobre la información aportada por el usuario, etc.

Desarrollar una aplicación informática que pueda mantener una conversación con una persona de manera natural sigue siendo hoy en día un reto, debido a la gran cantidad de fuentes de conocimiento de distinta naturaleza que hay que tener en cuenta y las limitaciones de las tecnologías

[2] *https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018/*

utilizadas para obtener información del usuario. No obstante, los constantes avances de la investigación en Procesamiento de Lenguaje Natural, Tecnologías del Habla, Inteligencia Artificial y Dispositivos Móviles han permitido que sea factible actualmente la comunicación persona-máquina mediante la voz con un alto grado de flexibilidad. Si bien dicha comunicación se suele restringir a dominios concretos, también existen otros sistemas de propósito general.

### 1.1.4. *Aplicaciones de los sistemas conversacionales*

El número de entornos y tareas en los que pueden aplicarse los sistemas conversacionales es enorme, por ejemplo:

- Sistemas que proporcionen información sobre transportes públicos;
- Sistemas de atención en el ámbito médico;
- Servicios de banca electrónica;
- Turismo;
- Entornos industriales;
- Aplicaciones accesibles desde los vehículos;
- Sistemas que faciliten el acceso a la información a personas con discapacidades,
- Aplicaciones de tele-educación;
- Apps y asistentes para dispositivos móviles;
- Acceso a servicios y control de máquinas vía telefónica;
- Interacción en el hogar y control domótico;
- Interacción con robots y dispositivos wearables.

### 1.1.5. *Ambigüedad terminológica*

Como se ha mencionado previamente, existe una ámplia terminología relacionada con los sistemas de diálogo, en la que sobresalen los términos:

- Chatbots y sistemas de diálogo hablado;
- Interfaces orales de usuario (VUI) y sistemas de respuesta oral interactiva (IVR);
- Sistemas conversacionales.

**Chatbots y sistemas de diálogo**

El término **chatbot** se suele utilizar para describir sistemas con los que usualmente se interactúa en modo texto (a través de un chat), con los que se charla sobre cualquier temática, habitualmente para entretenerse. Este tipo de agentes son los que se emplean en competiciones internacionales relacionadas con el test de Turing, como el Loebner Prize[3], donde la finalidad es obtener un sistema con una capacidad de conversación tan sofisticada que pueda hacer creer a sus interlocutores que están hablando con un ser humano. También se ha vinculado recientemente el término *chatbot* con aplicaciones web que permiten resolver consultas comunes o frecuentes de los usuarios utilizando un enfoque tipo *Pregunta-Respuesta* (*Q&A: Question & Answer*) con una capacidad de gestión del diálogo limitada a la detección de intenciones (*intent detection*) y reconocimiento de entidades (*named entity recognition)* que permitan activar acciones concretas.

En contraposición con los chatbots, en los **sistemas de diálogo** se entiende que el usuario tiene un objetivo o tarea más compleja que desea cumplir por medio de la interacción con el sistema. Por tanto, la conversación usuario-máquina persigue un fin o tarea específica y el sistema debe ser capaz de averiguar cuál es y cómo satisfacerla lo mejor posible.

Esta característica no impide que los sistemas de diálogo alcancen grandes cotas de sofisticación y que incluso un mismo sistema deba ser capaz de gestionar diálogos que versan sobre distintas temáticas (*sistemas de diálogo multidominio*). Por ejemplo, un sistema gestor de viajes podría ser capaz de conversar sobre reservas de hotel, de medios de transporte e incluso de actividades en el destino.

La distinción anterior entre chatbot y sistema de diálogo es común en el ámbito académico, aunque en ocasiones en este contexto se engloban ambos bajo el término “sistema de diálogo’’ y se indica explícitamente cuándo éste persigue resolver tareas específicas (sistema “orientado a la tarea’’, *task-oriented dialogue system*) o por el contrario sólo pretende mantener conversaciones intrascendentes (*non-task oriented dialogue*), como es el caso de los sistemas orientados a la charla (“chat-oriented dialogue systems’’ o sistemas que interactúan con los usuarios a modo de

[3] *https://www.aisb.org.uk/events/loebner-prize*

compañeros virtuales (“companion dialogue systems’’).

Sin embargo, en los últimos tiempos, con la proliferación de sistemas basados en tecnologías del lenguaje, se está popularizando el término chatbot como cualquier sistema con el que es posible comunicarse mediante lenguaje natural, también para sistemas que cumplen tareas específicas.

Por ejemplo, un vistazo rápido a la revista *chatbotsmagazine*[4] da una idea de que actualmente en el contexto empresarial esta es la acepción más común. Así, los asistentes virtuales como Siri, Alexa, Cortana o Google Assistant se consideran chatbots al igual que las aplicaciones que pueden prestar servicios ininterrumpidos en Facebook, Telegram o Slack.

**Interfaces orales de usuario (VUI) y sistemas de respuesta oral interactiva (IVR)**

Respecto a la distinción entre **interfaz oral de usuario** (Voice User Interface, VUI) y sistema de diálogo, se puede entender por interfaz oral a cualquier sistema que sea capaz de procesar el habla, mientras que el sistema de diálogo debe poder establecer una conversación. En este sentido, muchos de los sistemas actuales son capaces únicamente de responder a preguntas o comandos aislados (por ejemplo, podemos preguntar “¿Cuánto mide el Teide?’’ y obtener una respuesta), pero no pueden sostener una conversación de varios turnos (por ejemplo, realizar una transacción que requiera aportar varios datos o para aclarar entradas anteriores).

Un término relacionado es el de **sistema de respuesta oral interactiva** (Interactive Voice Response, IVR), que en este caso suele estar relacionado con sistemas de procesamiento de llamadas telefónicas tipo Call-Center.

**Sistemas conversacionales**

A los sistemas de diálogo también se les puede denominar **sistemas conversacionales**. En este informe consideramos que ambos términos son sinónimos. Si bien, algunos autores interpretan el adjetivo “conversacional’’ como conversación intrascendente sin un objetivo particular.

Otra distinción sutil es expresada mediante el uso del término sistema conversacional frente a **agente conversacional**. Usualmente, se emplea la palabra “agente’’ cuando el usuario puede figurarse al sistema como un interlocutor identificable. Esto suele ocurrir con los **agentes conversacionales personificados** (*embodied conversational agents, ECAs*) que tienen una apariencia física mostrada a través de avatares u otras representaciones gráficas; o los robots, para los cuales también existe el término **robot conversacional** (*conversational robot*). No obstante, incluso cuando

[4] *chatbotsmagazine.com*

el sistema sólo se comunica oralmente, se le puede describir como un agente cuando se le dota de una personalidad específica a través de su voz o comportamiento.

## 1.2. COMPONENTES Y ARQUITECTURA BÁSICA DE UN SISTEMA CONVERSACIONAL

La segunda sección del capítulo describe las funciones básicas de un sistema conversacional y su traducción en una arquitectura modular básica.

### 1.2.1. *Funciones básicas de un sistema conversacional*

La Figura 2 resume las acciones básicas que debe realizar un sistema de diálogo para cumplir la finalidad global para la que fue diseñado.

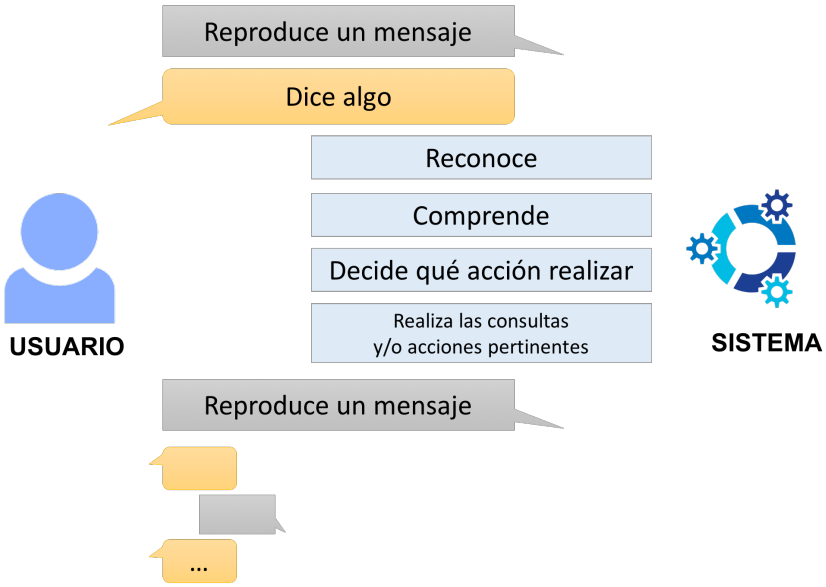


*Figura 2: Diagrama de acciones de un sistema de diálogo*

Tal y como se observa en esta figura, el sistema genera un mensaje inicial, normalmente para dar la bienvenida o informar al usuario sobre las características y funcionalidades del sistema. Tras cada intervención del usuario, el sistema repite cíclicamente un conjunto de acciones básicas como respuesta a cada acción del usuario:

- Reconocer el mensaje emitido por el usuario.
- Extraer el significado de dicho mensaje, es decir, comprender qué información es útil en el dominio del sistema.

- Realizar operaciones de acceso a base de datos u otros recursos del sistema (servicios web, etc.), en los que se almacena la información que solicita el usuario o se registran las operaciones que desea conocer.
- Decidir qué acción o acciones deben realizarse a continuación de cada solicitud del usuario, es decir, qué respuesta debe suministrar el sistema.
- Planificar la construcción de la respuesta que se debe enviar al usuario, generando el texto correspondiente.
- Reproducir un mensaje hablado con el texto previamente generado, o incluso añadir información multicanal adicional (gráficos, listados de opciones, etc.) en caso de usar una interfaz que permita tanto la comunicación hablada como escrita (como, por ejemplo, si se está usando una aplicación en un dispositivo móvil).

### 1.2.2. *Arquitectura modular básica de un Sistema Conversacional*

Dado el gran número de operaciones que deben realizarse, es habitual desarrollar los sistemas conversacionales de forma modular, empleando los componentes que se muestran en la Figura 3:

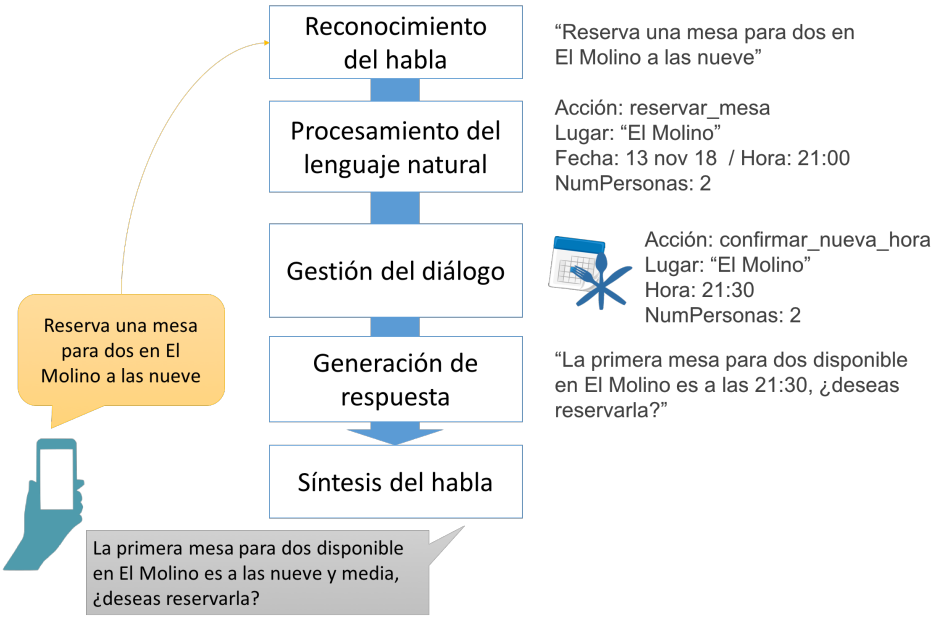


*Figura 3: Arquitectura modular de un interfaz conversacional hablado*

- **Módulo de Reconocimiento Automático del Habla**: reconoce la señal vocal pronunciada

por el usuario y proporciona la secuencia de palabras reconocida más probable (o las *k* más probables).

- **Módulo de Comprensión del Habla**: a partir de la(s) secuencia(s) de palabra(s) reconocida(s), el sistema obtiene una representación semántica de su significado.
- **Gestor de Diálogo**: considera la interpretación semántica de la petición del usuario, la historia del proceso de diálogo, la información de la aplicación disponible en ese punto y el estado del sistema, y determina la siguiente acción que debe tomar el sistema siguiendo la estrategia del diálogo. Este módulo accede también a los **repositorios de datos** asociados con la aplicación para realizar consultas y procesar sus resultados.
- **Módulo de Generación de Respuestas**: recibe la respuesta del sistema como una representación formal y tiene como función la generación de un mensaje en lenguaje natural, gramaticalmente correcto, que transmita el mensaje generado por el gestor de diálogo. La respuesta del sistema proporcionada por el generador de respuestas puede incorporar otras modalidades de información (vídeo, imágenes, tablas con datos, gestos a reproducir por un avatar, etc.).
- **Sintetizador de Texto a Voz**: recibe la respuesta del sistema como texto y genera la correspondiente señal de audio que se transmite al usuario.

## 1.3. ARQUITECTURAS DESTACADAS

El esquema estructural y la organización modular presentada en la sección anterior describe lo que se puede considerar como arquitectura clásica de un sistema de diálogo hablado o sistema conversacional.

La modularización facilita la implementación de este tipo de sistemas al aislar funcionalmente cada una de las fases por las que pasa el procesamiento de un turno de diálogo. No obstante, en ocasiones, una separación estricta o rígida entre cada uno de los componentes supone una simplificación o limitación excesiva en el diseño de los sistemas conversacionales, como se verá en la siguiente sección al describir el estado del arte y retos de investigación y desarrollo, donde entre otras cuestiones se comparará la integración a bajo nivel de los módulos clásicos con el desarrollo de modelos de diálogo incrementales.

En cualquier caso, la arquitectura modular anterior permite un análisis más detallado y exhaus-

tivo de las conexiones y arquitecturas empleadas a lo largo de los años para orquestar todos los componentes descritos.

### 1.3.1. *Galaxy*

Los investigadores del Spoken Language Systems Group (SLS) del MIT desarrollaron el trabajo pionero GALAXY, que planteaba una arquitectura para la implementación de sistemas de diálogo en el marco del programa DARPA Communicator [3, 4]. GALAXY se basa en el modelo cliente-servidor, en el que cada uno de los módulos del sistema actúa como servidor de una determinada tarea (reconocimiento, comprensión, etc.).

En la arquitectura se define un módulo (*Hub*) encargado de centralizar la comunicación entre los diversos módulos. El *Hub* recibe los envíos de los diferentes módulos y los redirecciona a su destinatario. La Figura 4 muestra los módulos que conforman la arquitectura, definiéndose módulos dependientes del dominio del sistema (por ejemplo, el gestor de diálogo y el módulo de comprensión del lenguaje) e independientes del mismo (como el servidor de audio). En esta arquitectura, el flujo de información viene controlado por la programación que se realice en el hub central.

Basándose en esta arquitectura, se desarrollaron sistemas de diálogo referentes en el área para:

- facilitar información en tareas de información meteorológica,
- planificación de viajes aéreos,
- información de carreteras.

Ejemplos de estos sistemas son:

- VOYAGER [5], sistema de información para viajeros,
- PEGASUS [6], sistema de acceso a un sistema de reservas, on-line
- o JUPITER [7], sistema de información meteorológica.

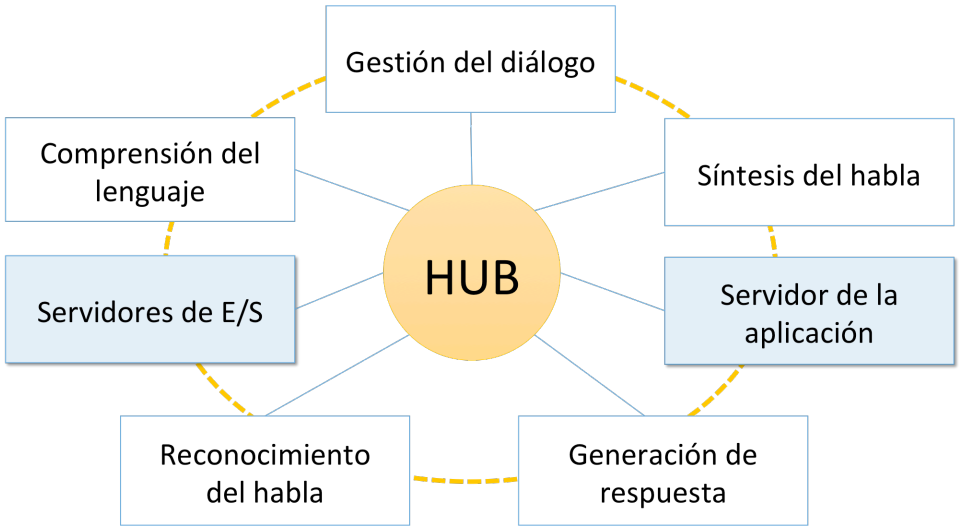


*Figura 4: Arquitectura definida en Galaxy Communicator*

### 1.3.2. *Olympus/Ravenclaw*

La arquitectura Olympus/RavenClaw [8] continuaba con la idea del hub, aunque proponía una secuencia más definida. La Figura 5 muestra el conjunto de módulos que componen la arquitectura. En la configuración por defecto, Olympus utiliza el reconocedor automático del habla Sphinx; el sintetizador de texto a voz Festival, Theta o Swift; el módulo de comprensión Phoenix; el módulo para tratamiento de medidas de confianza Helios; el generador de respuestas en lenguaje natural Rosetta (basado en plantillas de respuestas) y el gestor del diálogo Ravenclaw. Cualquiera de estos módulos puede reemplazarse por otro, posibilitándose asimismo la incorporación de nuevos módulos. Esta arquitectura se empleó en el desarrollo de múltiples sistemas, por ejemplo:

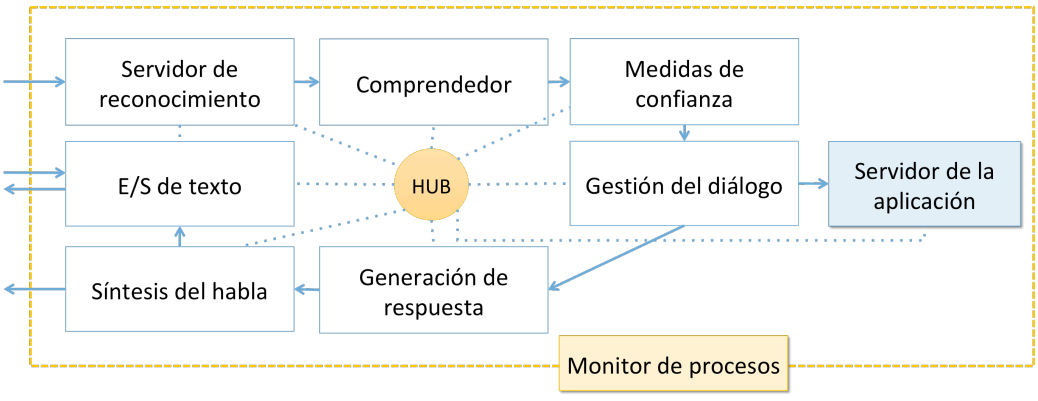


*Figura 5: Arquitectura Olympus para el desarrollo de sistemas de diálogo*

- *Conquest* [9] fue un sistema de diálogo hablado que proporcionaba información sobre con-

ferencias y congresos,

- el sistema de diálogo hablado *Let's Go!* [10], que proporciona información sobre la red de autobuses metropolitana de Pittsburgh (E.E.U.U) y que se ha utilizado como referencia para la aplicación de diferentes técnicas y modelos en el desarrollo de sistemas de diálogo, así como en la evaluación de los mismos al liberarse las grabaciones realizadas durante varios años.

### 1.3.3. *USI y Speech Graffiti*

El proyecto USI [11], también conocido como *Speech Graffiti*, tuvo como principal objetivo el diseño de un interfaz universal (independiente de las aplicaciones) para la comunicación hablada de forma que pueda utilizarse para facilitar el desarrollo de sistemas de diálogo. Para cumplir este objetivo, se definió un protocolo de interacción con el sistema que proporcionase una alternativa intermedia entre el lenguaje natural y los diálogos en los que la iniciativa recae en el sistema. Este protocolo se basaba en estructuras de comunicación independientes del dominio del sistema y palabras clave que posibilitan la interacción del usuario con un servicio determinado (por ejemplo, "starting over" para borrar la historia previa del diálogo).

La idea fundamental es que un usuario previamente instruido en las construcciones básicas del protocolo de comunicación pudiese interactuar fácilmente con cualquier aplicación compatible con Speech Graffiti. En [12] se realiza un estudio comparativo entre un mismo sistema con dos interfaces de comunicación: utilizando lenguaje natural o Speech Graffiti. Los resultados mostraron que un 74 % de los usuarios prefirieron Speech Graffiti, aún siendo los porcentajes de éxito de los diálogos muy similares en ambos interfaces.

Utilizando el paradigma USI se desarrollaron los sistemas que se citan a continuación [13]:

- *MovieLine* (proporcionaba información actualizada semanalmente sobre la cartelera de cines en Pittsburgh),
- *ApartmentLine* (acceso a información sobre alojamiento en Pittsburgh),
- *FlightLine* ( información sobre salidas, llegadas y terminales de vuelos)
- *Gadget* (control remoto de dispositivos: sistemas Hifi, aparatos domésticos, etc.).

### 1.3.4. *ATLAS*

Por otra parte, ATLAS es una plataforma creada para facilitar el desarrollo de aplicaciones multilingües y multimodales, y en especial sistemas de diálogo. El diseño de la plataforma se basa en un modelado en capas del sistema, siendo ATLAS una capa intermedia entre la capa dependiente de la aplicación y la capa que contiene los diferentes módulos del sistema de diálogo (Figura 6). Su objetivo es implementar la mayoría de las funcionalidades del sistema que sean independientes de la tarea, así como definir funciones para dar soporte a los módulos dependientes de la misma.

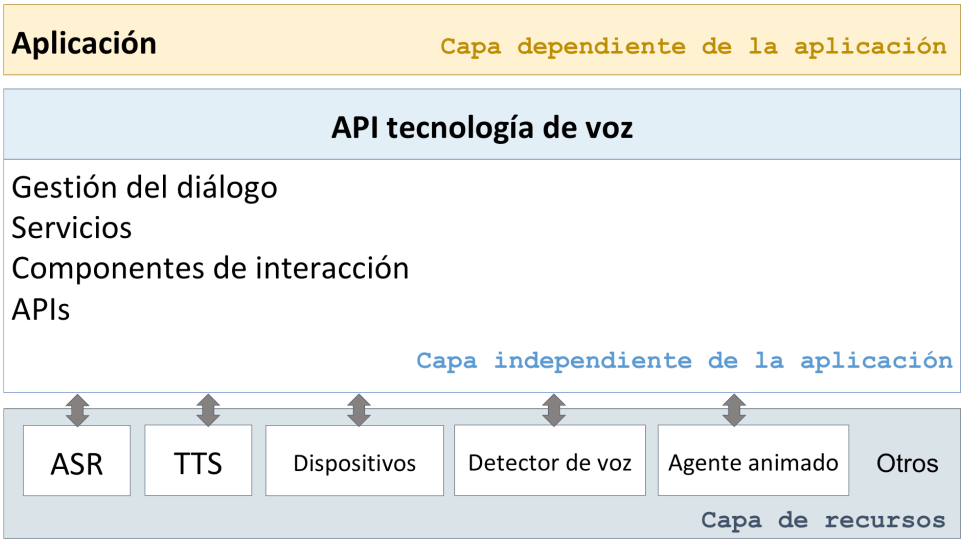


*Figura 6: Arquitectura definida en la plataforma ATLAS*

### 1.3.5. *JASPIS*

El grupo de investigación **Speech-based and Pervasive Interaction** de la Universidad de Tampere desarrolló la plataforma JASPIS [14] para implementar sistemas de diálogo multilingües, distribuidos, adaptados al usuario y al entorno de la interacción. La arquitectura se basa en la utilización de tres elementos diferenciados: agentes, gestores y evaluadores.

El nivel superior de esta arquitectura está formado por los gestores, conectados a un gestor central mediante una topología en estrella basada en el modelo cliente-servidor. Los agentes son componentes que implementan las acciones de interacción con el usuario: reproducir mensajes del sistema, tomar decisiones durante el diálogo, etc. Los evaluadores se utilizan para verificar diferentes aspectos de los agentes, con la finalidad de determinar si son apropiados para tareas específicas del sistema. La información del sistema se almacena en bases de datos compartidas

que pueden ser consultadas por todos los componentes de la arquitectura mediante el gestor de la información. La Figura 7 muestra la arquitectura de la plataforma JASPIS.

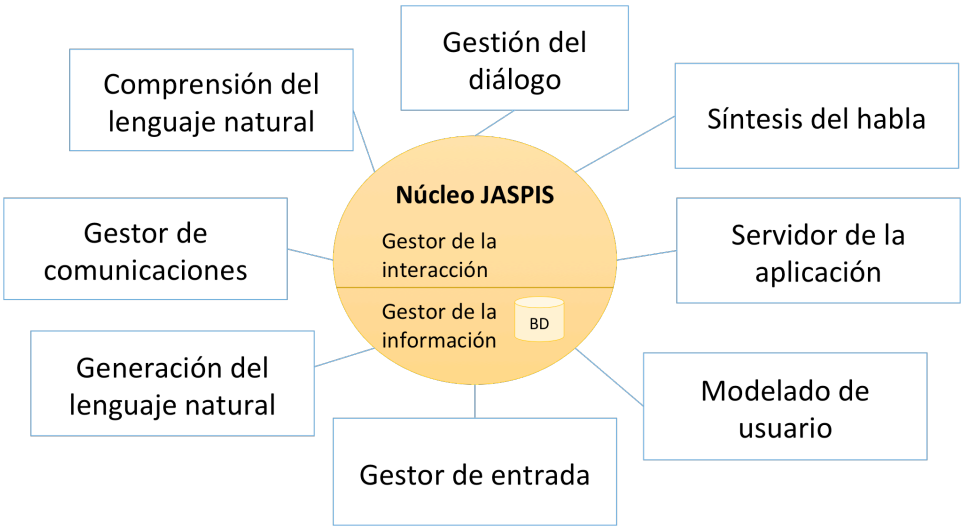


*Figura 7: Arquitectura definida para la plataforma JASPIS*

Mediante la arquitectura JASPIS se desarrollaron los siguientes sistemas:

- *Mailman / AthosMail* [15] (cliente de correo multilingüe, inglés y finlandés, diseñado para facilitar la lectura de los correos electrónicos mediante el uso del teléfono),
- *Busman* (información sobre horarios de autobuses en finlandés),
- *Doorman* [16] (lectura de correos electrónicos, realización de operaciones como la apertura de puertas y suministro de información sobre la situación de despachos y personal del departamento).

### 1.3.6. *TRIPS y KPML*

La arquitectura definida en el proyecto TRIPS [17] utiliza, al igual que DARPA Communicator, una serie de módulos que se comunican entre sí mediante un *hub* central. Para realizar esta comunicación se definió un lenguaje propio denominado KPML (*Knowledge Query and Manipulation Language*). El sistema consta de tres componentes fundamentales: el gestor de interpretaciones (*Interpretation Manager*), el agente de comportamiento (*Behavioral Agent*) y el gestor de generación (*Generation Manager*).

La función del gestor de interpretaciones es procesar la entrada del usuario a partir de la inter-

pretación semántica (secuencia de actos de diálogo) de la respuesta proporcionada por el reconocedor. El agente del comportamiento planifica que comportamiento va a seguir el sistema tras el análisis de la acción del usuario. El gestor de generación proporciona la respuesta final del sistema. La Figura 8 muestra la arquitectura desarrollada para el sistema TRIPS.

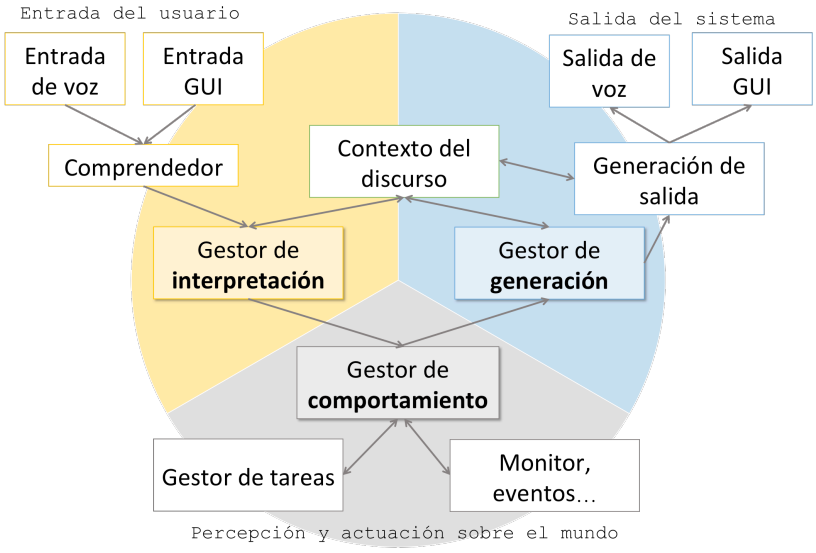


*Figura 8: Visión sintética de la arquitectura definida en el proyecto TRIPS*

### 1.3.7. *SMARTKOM*

SMARTKOM [18] es un sistema de diálogo multimodal que combinaba voz y gestos como modalidades de entrada y de salida. SMARTKOM definió una arquitectura de referencia para el desarrollo de sistemas de diálogo multimodales, que se resume en la Figura 9 y está conformada por:

- Módulos interfaz: A la entrada existe un módulo de audio y a la salida, un gestor de visualización.
- Reconocedores y sintetizadores: A la entrada del sistema existen módulos para el reconocimiento de voz, gestos y prosodia. A la salida, se integra un sintetizador de voz y el gestor de visualización.
- Módulos de procesamiento semántico: Este grupo de módulos realiza la comprensión y transformación de las representaciones semánticas (análisis de gestos y voz, reconocimiento de la intención del usuario, modelado del discurso y del dominio, planificador de acciones, transformación de conceptos a voz).

- Servicios externos: bases de datos y módulos encargados de extraer información de la web.

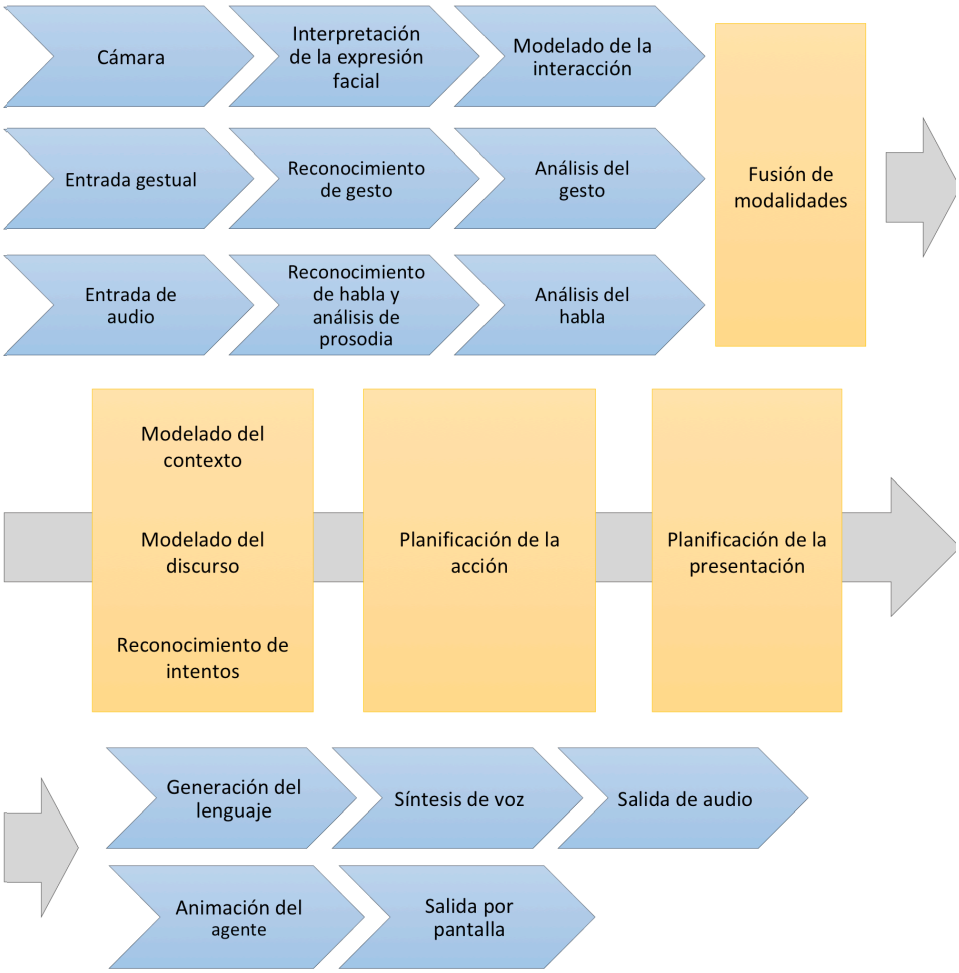


*Figura 9: Ilustración de los componentes principales de la arquitectura Smartkom*

### 1.3.8. *HOPS*

En [19] se presenta un sistema de diálogo multimodal desarrollado para el proyecto europeo HOPS. La arquitectura del sistema está conformada por los siguientes módulos:

- Módulo de voz: Controla la interacción a través del teléfono. Sus componentes son un reconocedor, un intérprete VoiceXML y un sintetizador.
- Módulo de texto: Controla la interacción a través de la web. Consiste en un servidor de texto y un analizador sintáctico-semántico.

- Gestor de diálogo: Se trata de un gestor de propósito general, independiente del servicio concreto. Utiliza la información semántica generada por los módulos de texto y voz.
- Gestor de ontologías: Envía al gestor de diálogo una representación del servicio basada en precondiciones, que son evaluadas por el gestor de diálogo para seleccionar el estado activado que posea una precondición más estricta.
- Gestor de acciones y consultas: Accede a las ontologías y a las bases de datos de las aplicaciones de forma transparente para el gestor del diálogo.
- Gestor de aplicaciones: Controla los recursos que requiere cada servicio implementado en el sistema.

### 1.3.9. *Sistemas de diálogo end-to-end*

Los sistemas de diálogo end-to-end [20, 21, 22, 23] tienen como principal objetivo el desarrollo completo del sistema utilizando únicamente tres módulos: el reconocedor automático del habla, el sintetizador de texto a voz y el módulo de diálogo end-to-end. Este último módulo se basa en un único modelo que toma como entrada la frase(s) proporcionada(s) por el reconocedor automático del habla para generar en un único paso la entrada del sintetizador de texto a voz. Las principales ventajas de este paradigma son la posibilidad de desarrollar fácilmente sistemas de diálogo multidominio y el hecho de no requerir datos supervisados para desarrollar los módulos descritos en la arquitectura clásica de estos sistemas.

Para ello, se apoyan en el aprendizaje de modelos de diálogo basados en aprendizaje profundo (deep learning) y aprendidos a partir de grandes corpus de diálogos que no requieren ser etiquetados. No obstante, la disponibilidad de estos grandes volúmenes de diálogos es todavía muy limitada. Además, existen muchos turnos del diálogo en los que no es posible extraer el contexto de la interacción para predecir en un único paso la siguiente respuesta del sistema a partir de la frase mencionada por el usuario en el turno actual.

## 1.4. LOS SISTEMAS CONVERSACIONALES EN EL CONTEXTO DE LAS TECNOLOGÍAS DEL LENGUAJE Y LA LINGÜÍSTICA

En su artículo *Language Technology: A first overview*[5], Uszkoreit señala:

[5] *http://www.dfki.de/~hansu/LT.pdf, 1998*

> *Las tecnologías del lenguaje son tecnologías de la información especializadas en el tratamiento del más complejo medio de información: el lenguaje humano. Por tanto, estas tecnologías se incluyen habitualmente bajo el término Tecnologías del Lenguaje Humano. El lenguaje humano puede aparecer tanto en forma hablada como escrita. Mientras que el habla es el modo más natural y antiguo de la comunicación con lenguaje, la información más compleja, así como la mayor parte del conocimiento humano, se mantiene y transmite mediante textos escritos. Las tecnologías de habla y texto procesan o producen lenguaje en estas dos modalidades.*
>
> *Pero el lenguaje incluye otros aspectos adicionales compartidos por el habla y el texto tales como diccionarios, la mayor parte de los modelos gramaticales y el significado de las oraciones. Por tanto, estos importantes componentes de la tecnología del lenguaje no están incluidos en lo que podemos denominar como tecnologías del habla y del texto. Entre otras se incluyen tecnologías que enlazan el lenguaje con el conocimiento. No conocemos cómo se representan en el cerebro humano el lenguaje, el conocimiento y el pensamiento. No obstante, la tecnología del lenguaje debe crear sistemas basados en representaciones formales que enlacen el lenguaje con conceptos y tareas del mundo real. Esto establece el interfaz con el área de las tecnologías del conocimiento, actualmente en rápido crecimiento*[6].

Como se desprende de esta cita, las tecnologías del lenguaje se enfrentan al reto de enlazar lenguaje con conocimiento y pensamiento. La primera consecuencia es que será necesario llevar a cabo una integración entre múltiples áreas de especialización, desde la lingüística, en tanto que estudio del lenguaje humano, hasta las ciencias de la computación (metodologías y tecnologías para el desarrollo de sistemas computacionales), la inteligencia artificial (incluyendo cuestiones tales como representación del conocimiento, planificación, aprendizaje automático, razonamiento, etc.) y muchas otras áreas de interconexión tales como psicolingüística, sociolingüística, dialectología, teoría de lenguajes formales, teoría de la señal, teoría de la comunicación, etc.

Como se ha indicado, un sistema conversacional se puede describir como un sistema computacional capaz de conducir una conversación o diálogo con un usuario a través de un conjunto de interacciones (turnos de diálogo) con el objetivo de completar una tarea. Para ello, el medio de comunicación básico será el lenguaje humano (bien hablado o escrito) y posiblemente enriqueci-

[6] *Traducción de los autores.*

do con otras modalidades comunicativas o fuentes de información paralingüística, tales como la detección de emociones, gestos y movimientos, utilización de dispositivos táctiles, etc.

La utilización del lenguaje como mecanismo básico, así como la propia naturaleza de la interacción conversacional, hacen que el corpus de conocimiento lingüístico juegue un papel relevante en todo el ciclo de desarrollo de estos sistemas.

### 1.4.1. *Reconocimiento del habla y síntesis de la voz: fonética y fonología*

Ambas áreas están relacionadas con el nivel de producción y reconocimiento del lenguaje hablado. La **fonética** se centra en el estudio de las propiedades físicas de los sonidos utilizados en el habla humana, desde su producción por parte de los sistemas fonador y articulario, su transmisión como señal acústica y su recepción por parte de los órganos auditivos (fonética articularia, acústica y perceptiva).

Por otro lado, la **fonología** aborda el estudio de los sonidos acústicamente discriminativos (*fonemas*) en general y en cada lengua en particular. Ambas ramas son especialmente relevantes en los sistemas conversacionales hablados para modelar tanto la fase de **reconocimiento del habla** como **síntesis de la voz**, que se abordarán con más detalle en el Capítulo 3.

### 1.4.2. *Comprensión y generación de lenguaje natural: morfología, sintaxis y semántica*

Los sistemas de reconocimiento del habla no utilizan exclusivamente la señal acústica para llevar a cabo la función de reconocimiento, ya que se requiere información lingüística adicional, tales como modelos de lenguaje. Podemos considerar que el resultado de este módulo es la obtención de una o más hipótesis de reconocimiento (entendidas como secuencias de palabras, posiblemente enriquecidas con otra información acústica suprasegmental tal como prosodia, pausas, etc.).

Con el objetivo de poder utilizar esta información, un sistema conversacional debe *entender* el significado asociado al mensaje emitido por el usuario. Para llevar a cabo esta tarea se requiere la incorporación de modelos en niveles adicionales de la estructura lingüística.

El **análisis morfológico** permitirá detectar la estructura de las propias palabras asignando la información relevante tal como categoría gramatical (*Part Of Speech - POS*), así como otros rasgos léxicos y morfológicos. A partir de esta información y del propio contenido léxico de las palabras

detectadas, el **análisis sintáctico** permitirá descubrir el análisis de constituyentes de las expresiones que intervienen en cada una de las frases, asignando funciones gramaticales y roles temáticos.

Por último, el nivel de **análisis semántico** se encargará de construir una representación semántica abstracta que contenga el significado asociado a la expresión original emitida por el hablante.

Estas tres etapas se agrupan en lo que tradicionalmente se denomina **Comprensión del Lenguaje Natural** (*NLU: Natural Language Understanding*). Por otro lado, el proceso simétrico que parte de la representación semántica del mensaje que se debe comunicar al usuario y termina construyendo la realización del texto correspondiente se conoce como **Generación del Lenguaje Natural** (*NLG: Natural Language Generation*). Durante la fase de generación también pueden intervenir los niveles semántico, sintáctico y léxico-morfológico, pero en este caso en sentido descendente, para a partir de la representación semántica, construir el nivel sintáctico de constituyentes y finalmente la secuencia de palabras final. Veremos estas fases con más detalle en la Sección 3.4.

### 1.4.3. *Medidas de confianza*

Con la finalidad principal de corregir los posibles errores generados por los módulos de reconocimiento y comprensión, la salida generada por el reconocedor puede no limitarse a una única frase, sino, como se ha comentado previamente, a múltiples hipótesis en forma de grafo de palabras [24] o de las $k$ mejores frases (las más probables para el reconocedor).

Un complemento a estos métodos consiste en la utilización de medidas de confianza (típicamente un valor real entre 0 y 1) que representen la fiabilidad de cada una de las palabras. El objetivo de la estimación de medidas de confianza es evaluar la calidad de las palabras reconocidas y de los conceptos semánticos extraídos en la fase de compresión de lenguaje natural, facilitando en lo posible la detección y corrección de errores por parte del gestor de diálogo. Las medidas de confianza obtenidas en estos dos módulos tienen como objetivo evaluar su comportamiento de forma que el gestor de diálogo pueda medir la calidad de la información recibida y en consecuencia, elegir la acción concreta a realizar: rechazar la frase, preguntar otra vez, o pedir confirmación de alguno de los datos obtenidos. De este modo, se posibilita que los módulos siguientes del sistema puedan operar con varias alternativas y tengan en cuenta en su funcionamiento la fiabilidad de las palabras.

Las medidas de confianza pueden clasificarse en tres niveles diferenciados:

- Nivel de palabra: en este caso el objetivo es detectar palabras mal reconocidas. Para ello, se utilizan parámetros obtenidos del módulo de reconocimiento de voz, procedentes de la decodificación y del modelo de lenguaje.

- Nivel de concepto: en este caso se pretende detectar conceptos erróneos dentro de una frase determinada. Estas medidas de confianza son muy importantes para la gestión de diálogo, puesto que la información semántica se utiliza para realizar la gestión y decidir cuáles van a ser las acciones del sistema en su interacción con el usuario. En este caso suelen utilizarse parámetros obtenidos del reconocedor de voz y del módulo de comprensión.

- Nivel de frase: en este nivel, el objetivo es detectar, por un lado, frases fuera del dominio de la aplicación, y por otro, frases del dominio con problemas en el reconocimiento que no tienen ninguna información semántica o concepto correcto. Se pretende detectar frases que no van a ser correctamente reconocidas y comprendidas por el sistema desarrollado, evitando realizar interpretaciones erróneas.

Aunque las medidas de confianza asociadas a las palabras en el proceso del reconocimiento son la medida más frecuentemente utilizada para denotar la fiabilidad de la información que se suministra al gestor de diálogo [25] [26] [27] [28] [29] [30], existen diferentes trabajos en los que se muestra además la utilidad de las medidas de confianza asociadas a conceptos y atributos durante el proceso de comprensión [31] [32] [33] [34].

### 1.4.4. *Gestión del diálogo: Pragmática*

En la arquitectura general del sistema conversacional el módulo de Gestión de Diálogo se ha asociado con las tareas correspondientes a coordinación y dirección de la propia conversación. Este módulo debe ser capaz de integrar dinámicamente la información que va recibiendo del usuario (una vez analizada y representada semánticamente), utilizando para ello un sofisticado modelo de memoria, decidir cuándo debe conectarse con módulos externos para recabar información adicional (por ejemplo, la disponibilidad y precios de una reserva de hotel), y decidir qué información se le debe enviar al usuario para continuar con la conversación (facilitarle toda la información solicitada, pedir información adicional relevante para continuar con el diálogo, desambiguar información que no está totalmente clara, detectar y resolver errores de reconocimiento o comprensión, etc.).

Estas tareas pueden requerir el uso de estrategias tradicionalmente asociadas con el campo de la **pragmática**. Esta disciplina se centra en el estudio de la influencia de los factores extralingüísticos en el análisis del lenguaje natural, desde la situación comunicativa, hasta las expectativas de los interlocutores y todo el contexto comunicativo en general.

## 1.5. LOS SISTEMAS CONVERSACIONALES EN LA ADMINISTRACIÓN PÚBLICA

En el informe de la OCDE sobre estrategias de gobierno para transformar los servicios públicos en áreas de bienestar social, se estudian como ejemplo las políticas de bienestar más innovadoras en el contexto de dos países que están haciendo grandes avances: Dinamarca y Suecia [35].

En concreto, la estrategia actual de Dinamarca (**The Common Public Strategy for Digital Welfare** 2013-2020) tiene siete pilares y uno de ellos es el uso de nuevos enfoques digitales para la gestión de casos, donde se hace especial hincapié en la liberación de recursos a través de reconocimiento del habla.

Ciertamente el reconocimiento y procesamiento del habla ha tenido tradicionalmente más peso en las administraciones que los sistemas conversacionales, destacando las aplicaciones de dictado, transcripción y traducción automática. No es de extrañar que las grandes empresas que generan estas soluciones, usualmente tengan a las administraciones públicas en su portfolio[7].

No obstante, los sistemas de diálogo pueden aportar grandes beneficios al sector público:

- Disponibilidad. Los ciudadanos pueden acceder a los servicios en cualquier momento y por tanto es posible atenderlos fuera del horario de oficina.
- Facilidad de uso. Puesto que se pueden utilizar sosteniendo una conversación en lenguaje natural, son más intuitivos que otro tipo de interfaces como la web.
- Universalidad. Estos sistemas son accesibles desde cualquier tipo de dispositivo, incluso mediante llamada telefónica desde un teléfono fijo, con lo que no sería necesaria ni siquiera conexión a Internet.
- Inmediatez. Se reducen los tiempos de espera con las preguntas frecuentes que el sistema sea capaz de responder sin redirigir a un operador humano.

[7] *Grundig: https://www.grundig-gbs.com/en/industry-sectors/public-administration/general-overview/*

Algunos ejemplos de uso de esta tecnología en el sector público son:

**Servicios de atención general al ciudadano**. En España el teléfono 060 ofrece información general al ciudadano y le permite realizar trámites.

El equivalente en Estados Unidos es el número 311 para llamadas que no sean de emergencia en los ayuntamientos para realizar quejas (p.ej. baches en la calle, graffitis, basuras, etc.). En las ciudades donde estos servicios se han automatizado se están obteniendo muy buenos resultados. Por ejemplo, en Nueva York[8], la automatización únicamente del sistema de enrutamiento de llamadas (preguntar al ciudadano qué desea y en función de lo que diga re-dirigir la llamada al número adecuado), ha permitido incrementar la eficacia del 311 en más de un 20 %. En ciudades grandes con un volumen alto de llamadas, este incremento supone una mejora considerable.

**Servicios de emergencias**. Los sistemas conversacionales pueden ser de gran ayuda en la fase de triaje en sistemas de emergencia. Durante el triaje, se clasifican las situaciones de emergencia para evaluar las prioridades de atención. Uno de los componentes principales de las emergencias extrahospitalarias es el triaje telefónico, donde el centro regulador de emergencias categoriza las llamadas en base a su grado de urgencia.

La optimización del triaje telefónico tiene implicaciones muy importantes para la calidad del servicio. Por ejemplo, en [36] se muestra cómo la definición de un buen protocolo de triaje telefónico ayudó al disminuir las consultas no justificadas de pediatría asegurando un mejor tratamiento a los pacientes graves.

El uso de sistemas de diálogo en este contexto permitiría contar con ayuda para clasificar las llamadas [37] y realizar traducción automática de llamadas de emergencia cuando quien llama habla en un idioma minoritario [38]. En el caso en el que la persona que recibe la asistencia es conocida, la clasificación automática puede personalizarse. Este es el caso de sistemas asistenciales para personas mayores, donde se han obtenido buenos resultados en el triaje automático y la selección de la mejor respuesta [39].

**Servicios sanitarios**. Los sistemas de diálogo tienen una larga trayectoria en el ámbito sanitario [40]. Por una parte, pueden emplearse para la gestión de citas médicas. De esta forma, los pacientes pueden pedir una cita sosteniendo una conversación en lenguaje natural con un agente. Por otra parte, están apareciendo nuevas aplicaciones ligadas al ámbito de la medicina preven-

[8] *https://www.nuance.com/omni-channel-customer-engagement/case-studies/nyc-311.html contiene la historia de éxito del uso de servicios IVR de Nuance.*

tiva. Así, se están desarrollando sistemas que asesoran para adquirir hábitos de vida saludable como realizar ejercicio regular, llevar una dieta sana o usar protector solar [41]. También se están creando oportunidades para el autoseguimiento de enfermedades crónicas o la recuperación en casa tras altas hospitalarias. En este caso, los sistemas pueden aconsejar cómo realizar ejercicios o actividades encaminadas a la recuperación y también realizar preguntas que permitan evaluar la situación del paciente y en su caso re-dirigirlo a su médico [42, 43].

**Otros servicios asistenciales**. En el ámbito del cuidado de las personas mayores, los sistemas de diálogo permiten la detección y tratamiento de desórdenes cognitivos [44], la asistencia en el hogar y luchar contra el aislamiento mediante sistemas que acompañan y dan conversación [45]. Si bien estos últimos, los sistemas que acompañan, son controvertidos y han generado debate respecto al rol de la tecnología en el cuidado de las personas mayores [46].

## 1.6. SISTEMAS CONVERSACIONALES, ¿EL OBJETIVO FINAL?

Sistemas conversacionales

Texto original en inglés:

**Joseph MARIANI, LIMSI-CNRS, Université Paris-Saclay, Orsay, France**

Traducción al castellano: Pablo Sierra y José Luis Pro

El procesamiento del lenguaje hablado ha conseguido un gran progreso desde las primeras investigaciones, hace 50 años, y ahora disponemos de sistemas de activación por voz que se han hecho indispensables en nuestro día a día, en coches o en nuestros hogares, como Google Home, Amazon Echo, MS Homepad o Alibaba Tmall Genie, por mencionar unos cuantos, debido a que han alcanzado una calidad suficiente para un despliegue a gran escala.

Esto ha sido posible gracias a la investigación científica basadas en aprendizaje automático, desde técnicas de búsqueda de patrones en los 80 hasta los modelados estadísticos en los 90 y las redes neuronales más recientes, que han alcanzado mayor profundidad. Para ello fue necesaria la disponibilidad de grandes cantidades de datos para poder reflejar los distintos idiomas, dialectos, acentos, timbres, ambientes acústicos y aplicaciones posibles, gracias al progreso en el almacenamiento en ordenadores y velocidad de procesamiento. La introducción del paradigma de evaluación por el US Darpa a mediados de los 80 también fue decisivo en este marco como un instrumento fundamental para evaluar las distintas perspectivas, así como seleccionar las más eficientes.

Muchas de las aplicaciones que fueron previstas a comienzos de los 80 se han logrado a día de hoy, aunque este exitoso proceso de caja negra no nos ha hecho comprender mejor el lenguaje humano. Funciona, pero ¿cómo y por qué? La esperanza es que las tecnologías disponibles serán útiles como herramientas en las manos de investigadores en el área de lingüística.

A pesar de los logros, sigue habiendo algunos temas científicos por resolver, y los sistemas conversacionales aparecen como el desafío de mayor complejidad. Me siento perplejo al ver que el tópico de mi tesis doctoral en ingeniería que defendí en 1977, llamada diálogo piloto-avión, sigue sin estar resuelto. ¡Parece ser que ahora es el problema más complejo de resolver! Cuando miro atrás, los progresos en investigación buscaban dividir este problema en pequeños pasos, que han sido resueltos poco a poco.

En la tesis doctoral que defendí en 1982, propuse una lista de los diferentes niveles de decodificación del habla: los llamados “bajos” niveles, acústico, fonético, léxico y los “altos” niveles, sintáctico, semántico, pragmático, diálogo, y los mismos para la codificación del habla: generación y síntesis de habla. La mayoría de estos temas han sido exitosamente tratados hoy en día en términos de procesamiento automático, a través de modelos acústicos, fonéticos y lingüísticos y complementación semántica de posiciones, aunque los modelos pragmáticos y de diálogo que son obligatorios para lograr sistemas conversacionales no están disponibles en condiciones operativas. Esto incluye el manejo de anáfora, elipsis, objetivos, enfoques, creencias, sentimientos, metáforas y actos de habla indirectos. En mi tesis incluí un ejemplo de Barbara Grosz: como partir de una afirmación simple: “La caja de herramientas está cerrada”. hacia una “respuesta”: “La llave está en la estantería”, en 12 pasos que un humano normal es capaz de realizar de manera fácil y rápida, mientras que para la máquina más poderosa será imposible. De hecho, uno puede pensar que el objetivo final será logrado cuando una máquina sea capaz de interpretar un silencio.

El paradigma de evaluación objetiva y cuantitativa introducido por US DARPA a mediados de los 80 ha tenido un impacto decisivo permitiendo el progreso en nuestro campo de investigación, a través de la organización de campañas de evaluación anuales. Su primer resultado consistió en la migración de acercamientos basados en el conocimiento a búsqueda de patrones y aprendizaje automático. Esto dirigió el progreso normal mediante la identificación de subproblemas de dificultad gradual y comprobando cómo los mejores sistemas alrededor del mundo podían resolverlos. Comenzó con habla leída (con un vocabulario limitado y una complejidad del lenguaje simple), a dictado de voz (con un vocabulario ilimitado, pero afrontando la dificultad práctica del usuario para preparar frases correctas), al uso de micrófonos variados, para transcripciones de la retrans-

misión de noticias, finalmente a transcripción de habla de diálogos (tableros de conmutadores), incluyendo habla telefónica e idiomas distintos al inglés, y finalmente alcanzado la transcripción. La síntesis de texto a voz consiguió progresos en paralelo que también fueron medidos con métodos cualitativos y cuantitativos, así como la comprensión del habla para tareas específicas, tales como vuelos o reservas de restaurantes.

Los sistemas disponibles hoy en día necesitan una frase directa seguida de una petición simple, para poder manejar un diálogo limitado. Pero aun así son incapaces de manejar una conversación completa. Mientras que la investigación se beneficia de la gran cantidad de datos acumulada cada día por los distribuidores de tecnologías que despliegan sus aparatos para hogares, uno de los obstáculos es la dificultad para juzgar la calidad de los sistemas de diálogo comparados con la evaluación del reconocimiento automático de voz, y que el ciclo“ensayo-prueba-mejora” no puede ser implementado tan fácilmente como en un ASR. La razón es que un diálogo es dinámico y que puede haber un número infinito de turnos por diálogo correcto, al igual que la Traducción Automática afronta el mismo problema con el posible número de traducciones correctas. La adecuación relativa de una respuesta debe ser considerada junto a su fluidez para conseguir un diálogo eficiente y cómodo. Ha habido varios intentos para medir la calidad de los chatbots y una necesidad de introducir el concepto de manera más genérica en la robótica, en un esquema multi modal de comunicación y un ambiente complejo cada vez más poblado por robots, avatares y cuestiones relativas a la red.

Los sistemas conversacionales pueden parecer el Santo Grial, específicamente si consideramos que las conversaciones deberían ser construidas en cualquier idioma, y que por lo tanto implican solucionar la traducción hablada o alternativamente la comprensión universal.

## 2. SISTEMAS CONVERSACIONALES: FUNDAMENTOS

En esta sección se presenta una breve reseña histórica de los sistemas conversacionales, su contexto actual y auge, y los modelos y teorías surgidos durante las últimas décadas para el análisis semántico de la conversación.

## 2.1. BREVE RESEÑA HISTÓRICA

Los seres humanos han imaginado siempre la posibilidad de comunicarse oralmente con seres artificiales. Existen muchos ejemplos en el cine y literatura, algunos de los más antiguos se pueden encontrar en la mitología griega y romana en las que los héroes conversaban con estatuas de divinidades o guerreros de bronce. Los primeros intentos serios de construir sistemas parlantes se remontan a los siglos XVIII y XIX en los que se construyeron los primeros autómatas que imitaban la conducta humana, en su mayoría estudios e ingenios para la síntesis de sonidos similares a la voz humana (como los desarrollados por Leonhard Euler, C. G. Kratzenstein, Charles Wheatstone, el barón Von Kempelen, Josef Faber), o las primeras máquinas eléctricas desarrolladas por J.Q. Sterwart, R.H. Dudley, R. Reiz y S. Watkins durante la década de los años 20 y 30 del siglo XX.

Durante los años 40, se desarrollaron las primeras computadoras y algunos científicos prominentes como Alan Turing precisaron su potencia para el desarrollo de sistemas “inteligentes” y para medir la capacidad de inteligencia de las máquinas propuso el denominado Test de Turing [47], una prueba de la habilidad de una máquina para exhibir un comportamiento inteligente similar al de un ser humano de tal manera que, interactuando con ella en una conversación en lenguaje natural, una persona pueda determinar si su interlocutor es una máquina o una persona.

Éste fue el punto de partida que fomentó las iniciativas de investigación que en los años 60 originaron los primeros agentes conversacionales, los cuales aún no interpretaban semánticamente las frases proporcionadas por los usuarios. Por ejemplo, ELIZA de Weizenbaum [48], basado en la localización de palabras clave y el uso de plantillas predefinidas. Las plantillas transformaban la entrada del usuario en respuestas del sistema. Así, cuando el usuario escribía una frase como“Estoy X”, ELIZA contestaba“¿Cuánto hace que estás X” independientemente del significado de ’X’. El sistema PARRY [49] imitaba a una persona con esquizofrenia paranoide. Fue desarrollado en 1972 por el psiquiatra Kenneth Colby y se le considera un programa más avanzado que ELIZA, al utilizar algoritmos que modelizaban las teorías del Dr. Colby sobre la paranoia. A principios de los años 70, un grupo de 33 psicólogos utilizaron una variante del Test de Turing para evaluar el sistema, logrando PARRY engañar a sus examinadores humanos el 52 % del tiempo y, considerándose por ello en algunos ámbitos, como la primera máquina en superar el Test de Turing. Son célebres también las “conversaciones” que mantuvieron ELIZA y PARRY durante la década de los 70.

Durante los años 70 surgió la investigación en lingüística computacional partiendo de los trabajos teóricos desarrollados desde los años 50 por Chomsky, Montague y Wood. Al mismo tiempo, apa-

recen los primeros sintetizadores del habla basados en reglas. En los años 70 también se desarrollan los primeros reconocedores continuos del habla basados en décadas de investigación sobre el habla discreta en la que los estímulos verbales se alternaban con pausas largas.

En este contexto, los sistemas de diálogo en las décadas de 1960 y 1970 estaban basados en texto y motivados por los esfuerzos para aplicar técnicas de la lingüística a aplicaciones que involucraban el diálogo, por ejemplo, en sistemas como BASEBALL [50], SHRDLU [51] y GUS [52].

La investigación en las tecnologías de los sistemas de diálogo, tal y como los conocemos hoy en día, se remonta a finales de la década de los ochenta como resultado de dos grandes proyectos con financiación gubernamental, que posicionaron los interfaces conversacionales como área clave de investigación dentro de los campos de las tecnologías del habla y el procesamiento del lenguaje natural: el programa sistemas de lenguaje hablado DARPA en los Estados Unidos y el programa Esprit SUNDIAL en Europa.

El dominio del programa DARPA fue la consulta de información relativa a vuelos (Air Travel Information Services, ATIS). Su principal objetivo fue el estudio y desarrollo de las tecnologías relativas al reconocimiento del habla y comprensión del lenguaje, bajo el dominio de la reserva de vuelos utilizando el canal telefónico [53] [54]. Dado que todos los participantes del proyecto utilizaron la misma base de datos, fue posible comparar el funcionamiento de los diferentes prototipos desarrollados, llevándose a cabo un gran esfuerzo para realizar evaluaciones periódicas de los diferentes sistemas. El corpus de diálogos ATIS sigue estando disponible para desarrolladores y evaluadores de sistemas de diálogo.

La tarea ATIS sirvió de marco para el desarrollo de los primeros proyectos, como los llevados a cabo en AT&T, concretamente el proyecto AMICA [55], donde se aplicaron diferentes métodos estadísticos para el desarrollo de un sistema de diálogo con iniciativa mixta. ATIS también fue el punto de partida de las investigaciones del MIT y de la CMU, dentro de este proyecto se desarrollaron los sistemas CMU ATIS [56] y MIT ATIS [57] [58].

El proyecto SUNDIAL (Speech Understanding and Dialogue) [59], financiado por la Comunidad Europea, tuvo como dominio la consulta de horarios de trenes y aviones en inglés, francés, alemán e italiano. El objetivo del proyecto fue construir sistemas de diálogo en tiempo real capaces de mantener una conversación con el usuario siguiendo una estrategia cooperativa. Además de tratar aspectos de reconocimiento y comprensión del habla, un tema de estudio en el que se centró la investigación fue el modelado del diálogo hablado, desarrollándose diferentes aproximaciones

para realizar la gestión del diálogo.

La investigación llevada a cabo en SUNDIAL condujo a un gran número de proyectos con financiación europea centrados en el modelado del diálogo, como VerbMobil [60], DISC [61] y ARISE [62]. También cabe mencionar el sistema Philips [63], sistema comercial de consultas sobre horarios de trenes desarrollado a partir de la investigación llevada a cabo en SUNDIAL. Dentro del proyecto europeo ARISE se desarrollaron seis sistemas en paralelo: dos prototipos italianos basados en la tecnología desarrollada por el CSELT [64] [65], un prototipo francés desarrollado por el LIMSI [66] [62] y dos prototipos en holandés y uno en francés basados en la tecnología Philips.

Durante la década de 1990, entre los programas de investigación de mayor importancia, cabe destacar DARPA Communicator. En este programa, con financiación gubernamental y centrado en el desarrollo de tecnologías del habla, participaron centros de investigación en Estados Unidos y Europa. El objetivo establecido fue el desarrollo de una nueva generación de sistemas de diálogo, utilizando como entrada del sistema no únicamente la voz sino existiendo también la posibilidad de incorporar otros tipos de información multimodal. Los sistemas desarrollados dentro de este programa soportan interacciones complejas con el usuario, desde el punto de vista que tanto el sistema como el usuario pueden iniciar la conversación, cambiar de tema o interrumpir al otro participante del diálogo. Los dominios de aplicación (en los que se incluye la planificación de viajes) requieren usualmente el acceso a múltiples fuentes de información, representando un avance con respecto a los sistemas desarrollados en los programas ATIS o SUNDIAL.

Los investigadores de la CMU desarrollaron el sistema Carnegie Mellon Communicator [67], que permite obtener información de itinerarios complejos que incluyen reservas múltiples de vuelos, hoteles y alquiler de coches. La arquitectura del sistema se basa en la definición de módulos agentes de dominio que se encargan de la gestión de la información más específica. Paralelamente, investigadores de la Universidad de Colorado desarrollaron el sistema CU Communicator [68], que aborda la misma tarea.

De los diversos proyectos que siguieron, sobresalen los sistemas implementados por el grupo del LIMSI: el sistema PARIS-SITI, dedicado a facilitar información turística sobre la capital francesa, el sistema MASK (Multimodal Multimedia Service Kiosk) [69] y el sistema sobre información de trenes RAILTEL [70].

En paralelo con las numerosas iniciativas de investigación surgidas en los 90, se comenzaron a implementar sistemas comerciales con el objetivo principal de utilizar la voz para automatizar

las tareas de autoservicio, como el enrutamiento de llamadas, consultas de información y otras transacciones simples. Uno de los primeros fue el sistema de procesamiento de llamadas mediante reconocimiento de voz (VRCP) de AT&T [71] que recibió a las personas que llamaban con el mensaje “Please say collect, calling card, or third party”. Tras reconocer la opción seleccionada, el sistema transfería la llamada para proporcionar el servicio solicitado. El sistema VRCP tuvo un gran éxito comercial y se utilizó para atender millones de llamadas por año.

Hasta mediados de la década de 1990, el autoservicio automatizado generalmente se gestionaba mediante sistemas IVR, que proporcionaban indicaciones habladas pre-grabadas (por ejemplo, “pulse 1 para ver el saldo, pulse 2 para pagar una factura”). Este modo de interacción con el usuario se denomina DTMF (Dual-tone multi-frequency signaling, marcación por tonos). Una de las aplicaciones más conocidas de los IVR son los sistemas de correo de voz, pero esta tecnología se ha utilizado en multitud de aplicaciones para automatizar el enrutamiento de llamadas y el autoservicio [72].

El sistema HMIHY (**How may I help You?**), desarrollado por AT&T [73], fue uno de los primeros sistemas de voz interactivos implementados comercialmente para abordar la tarea de enrutamiento de llamadas a mayor escala. La tarea de HMIHY es identificar el motivo de la llamada del cliente utilizando entradas en lenguaje natural en lugar de pulsaciones en el teclado correspondientes a un menú, para luego encaminar la llamada al destino apropiado. Las personas que llamaban al sistema eran recibidas con un mensaje de bienvenida abierto (”¿En qué puedo ayudarle?”), que fomentaba respuestas del usuario con menos restricciones y más elaboradas. Seguidamente, el sistema decidía si era necesario confirmar la entrada proporcionada, aceptar las correcciones, o solicitar información adicional. A finales de 2001, HMIHY atendía más de 2 millones de llamadas por mes y mostraba mejoras significativas en la satisfacción del cliente con respecto a sistemas anteriores.

Si bien los sistemas de diálogo hablado y las VUI utilizan tecnologías de lenguaje hablado similares para el desarrollo de aplicaciones interactivas de voz, los desarrolladores de VUI se centran sobre todo en las necesidades empresariales, como el retorno de la inversión, problemas relacionados con factores humanos, la facilidad de uso y la satisfacción del usuario.

Los Agentes conversacionales personificados (ECAs) utilizan personajes animados generados por ordenador para combinar expresiones faciales, posturas corporales, gestos con las manos y el habla para proporcionar una interacción más natural y similar a la de los seres humanos. Este tipo

de sistemas se emplean cada vez más en aplicaciones comerciales, por ejemplo, para leer noticias y responder preguntas sobre transacciones y productos en páginas web de comercio electrónico. Debido a que se consideran sistemas más confiables, creíbles y entretenidos, se utilizan además en entornos de atención médica y también en juegos de roles, simulaciones y entornos virtuales inmersivos.

Los compañeros virtuales y los robots sociales pueden tomar la forma de objetos físicos, como mascotas digitales o robots, o pueden existir virtualmente como aplicaciones software. Este tipo de sistemas pueden servir de apoyo en las actividades de la vida diaria y facilitar la vida independiente en el hogar para personas mayores y para personas con discapacidades. También se utilizan para desempeñar un papel educativo para los niños.

### 2.1.1. *Chatbots*

Los chatbots, también conocidos como chatterbots, facilitan usualmente conversaciones simuladas en las que el usuario puede proporcionar únicamente entradas textuales. Una de las motivaciones para los desarrolladores de chatbots es tratar de engañar al usuario para que piense que está conversando con otro humano (Imitation Game) [47]. Los Premios Loebner, iniciados en 1991 por el Dr. Hugh Loebner, tiene como objetivo encontrar un chatbot que pueda superar el Juego de Imitación.

Hasta la fecha, la mayoría de las conversaciones con chatbots se han basado en texto, aunque algunos chatbots más recientes utilizan el habla como entrada y salida y, en algunos casos, también incluyen avatares o cabezas parlantes para dotar al chatbot de una apariencia más humana. En general, la interacción con los chatbots toma la forma de conversaciones más simples que los diálogos en tareas relacionadas con los SDS y VUIs. Otra diferencia es que los chatbots generalmente responden a las entradas del usuario sin tomar la iniciativa en la conversación o realizar preguntas al usuario.

Los chatbots se utilizan cada vez más en áreas como tele-educación, recuperación de información, negocios y comercio electrónico, por ejemplo, como asistentes automatizados en línea para complementar o incluso reemplazar el servicio proporcionado por un ser humano en un centro de llamadas. Por ejemplo, IKEA tiene un asistente automatizado en línea llamado Anna, desarrollado por Artificial Solutions. Anna responde preguntas sobre los productos y horarios de apertura de IKEA, así como también muestra emociones si no puede encontrar la información solicitada.

## 2.2. CONTEXTO ACTUAL Y AUGE DE LOS SISTEMAS CONVERSACIONALES

Las interfaces orales están experimentando un gran auge en los últimos tiempos y algunas fuentes apuntan a que el mercado de las interfaces orales llegará hasta los 18k millones de dólares en 2023[9]. A continuación, presentamos algunas de las razones que han causado este incremento en su interés comercial.

### 2.2.1. *Inteligencia Artificial: de los Sistemas Expertos al Aprendizaje Automático*

En primer lugar, cabe destacar el renacimiento de la Inteligencia Artificial. Desde mediados de la década de 1950, los investigadores en Inteligencia Artificial (IA) han competido contra el desafío de crear aplicaciones capaces de comportarse de manera inteligente. Al principio, se creía que el comportamiento inteligente podía reproducirse utilizando modelos de razonamiento simbólico basados en reglas de lógica formal (aproximaciones basadas en conocimiento). Este enfoque hacía hincapié en problemas que son difíciles de resolver para los seres humanos pero fáciles para las computadoras, por ejemplo, la toma de decisiones en el juego de ajedrez. Los sistemas basados en el conocimiento (también conocidos como sistemas expertos) se desarrollaron en los años 70 y 80 como ayuda en la toma de decisiones en problemas complejos como el diagnóstico médico.

Sin embargo, bien pronto se hizo evidente que varios aspectos del comportamiento inteligente que son fáciles para los seres humanos pero difíciles para los ordenadores, como el reconocimiento de voz y el reconocimiento de imágenes, no podían resolverse utilizando estos enfoques simbólicos, y requerían procesos como la extracción de patrones de datos y el aprendizaje de la experiencia. Como resultado, los enfoques subsimbólicos que utilizan redes neuronales y métodos de aprendizaje estadístico han llegado a dominar el campo.

Los avances en Inteligencia Artificial y Aprendizaje Profundo han posibilitado obtener mejores tasas de acierto y adaptarse a nuevos usuarios y contextos de interacción. El desarrollo de la propia interfaz oral puede beneficiarse de tratar con grandes cantidades de datos que permitan aprender parte de las reglas de interacción. Varios factores han contribuido al éxito reciente de los enfoques subsimbólicos:

- los avances en las unidades de procesamiento de gráficos (GPU) que han permitido los cálculos masivos paralelos necesarios para ejecutar redes neuronales;

[9] *https://www.marketsandmarkets.com/PressReleases/speech-voice-recognition.asp*

- la disponibilidad de cantidades ingentes de datos (conocido como Big Data) que permiten a los sistemas de inteligencia artificial aprender y ser cada vez más inteligentes;
- el desarrollo de nuevos algoritmos (conocidos como Deep Learning o Aprendizaje Profundo) que se ejecutan en GPUs y son capaces de procesar estas enormes cantidades de datos. Como consecuencia de estos avances en Inteligencia Artificial, muchas compañías importantes como Google, Microsoft, Amazon, Facebook o Baidu, han reclutado a los principales expertos mundiales en aprendizaje profundo en áreas tales como búsqueda, aprendizaje, comprensión del lenguaje natural y desarrollo de asistentes personales.

El reconocimiento automático del habla ha mejorado notablemente desde 2012 con la utilización de técnicas de aprendizaje profundo. También ha habido avances importantes en la comprensión del lenguaje hablado. Las aproximaciones basadas en el aprendizaje automático de modelos de gestión de diálogo a partir de los datos han mejorado su rendimiento en comparación con los enfoques tradicionales basados en la definición manual de reglas.

### 2.2.2. *De la Web Semántica al Question-Answering en los buscadores web*

En segundo lugar, la visión de la Web Semántica es que todo el contenido en la Web debe ser estructurado y legible por una máquina, de modo que las aproximaciones tradicionales de búsqueda de palabras se han sustituido por la búsqueda semántica basada en el significado de la entrada. El etiquetado semántico de las páginas con codificaciones como el Resource Description Framework in Attributes (RDFa) y las grandes bases de conocimiento estructuradas, como el Knowledge Graph (gráfico de conocimiento) de Google, han permitido a los motores de búsqueda interpretar mejor la semántica de la intención del usuario, devolver respuestas estructuradas a las consultas y diseñar modos de interacción avanzados de pregunta-respuesta para asistentes personales virtuales como Google Now.

### 2.2.3. *Dispositivos portables inteligentes*

Por otro lado, los teléfonos inteligentes (smartphones) y otros dispositivos inteligentes portables ofrecen actualmente un potencial y funcionalidades mucho mayores que los ordenadores personales de hace pocos años. Además de disponer de un micrófono y altavoz, estos dispositivos tienen acceso a una amplia gama de información contextual, como la ubicación del usuario, el

tiempo y la fecha, los contactos y el calendario. La integración de esta información contextual en los interfaces conversacionales permite personalizar los sistemas.

### 2.2.4. *Nuevos ecosistemas en tecnología y telecomunicaciones*

Las velocidades inalámbricas más rápidas, la disponibilidad casi omnipresente de conexiones WiFi, los procesadores más potentes en dispositivos móviles y el advenimiento de la computación en la nube han posibilitado que tareas como el reconocimiento de voz se puedan realizar en la nube mediante potentes ordenadores.

En cuanto a estos nuevos escenarios de uso, el desarrollo en los últimos años en las tecnologías de reconocimiento de habla de largo alcance (*far field speech recognition*, FFSR) permite que se pueda reconocer la voz con éxito en distintos tipos de dispositivos. En los inicios del área, los sistemas de diálogo estaban restringidos a sistemas telefónicos. En la actualidad, son accesibles desde multitud de dispositivos, incluyendo electrodomésticos, coches, etc. La tecnología que permite el FFSR, por ejemplos los *arrays* de micrófonos, ha sido empleada con éxito no sólo para mejorar el reconocimiento del hablar en dispositivos generales como los teléfonos inteligentes, sino que incluso ha permitido crear nuevos tipos de hardware ideados específicamente para dar soporte a la interacción oral. Es el caso de los dispositivos de interacción oral en el hogar Amazon Echo y Google Home.

La existencia estos dispositivos no sólo genera nuevos escenarios de uso, sino que también hacen que los usuarios se familiaricen con este tipo de interfaces y las asimilen a su vida diaria

### 2.2.5. *El interés general por los sistemas conversacionales*

Por último, mientras que hace pocos años el interés en las interfaces conversacionales se limitaba a empresas relativamente pequeñas y a entusiastas del sueño de la IA, ahora muchas de las empresas más grandes del mundo compiten para crear sus propios dispositivos conversacionales, por ejemplo, Siri de Apple y Google Now de Google, Alexa de Amazon, Cortana de Microsoft, M de Facebook y Duer de Baidu.

Estos dispositivos avanzados permiten a las empresas crear perfiles más precisos de sus usuarios y personalizar sus servicios de comercio electrónico, e incluso la capacidad de disponer de interfaces orales para el control domótico de dispositivos en la nube (Internet de las Cosas, IoT, *Internet*

*of Things*). Los servicios web y el Internet de las Cosas proveen de numerosas aplicaciones listas para usarse. Por ejemplo, la bombilla inteligente *Philips Hue* provee de una API que permite activarla, desactivarla, cambiarla de color... todas estas órdenes están accesibles desde una aplicación para móviles, pero igualmente se pueden conectar de forma sencilla con una interfaz oral que permita traducir comandos orales en cambios en el estado de la bombilla.

Han proliferado además las herramientas que permiten desarrollar sistemas basados en tecnologías del habla: Pandorabots[10], dialogflow [11], chatfuel[12], Amazon Lex[13], IBM Watson [14], Lekta [15].

La diferencia de esta nueva situación con los inicios del área es que los desarrolladores se pueden centrar en la interfaz en sí, en generar diálogos usables y personalizados, sin tener que implementar a su vez la conexión con los dispositivos que controlan.

De forma parecida ocurre con los servicios web: la gran oferta de servicios accesibles a través de diferentes APIs permite a los desarrolladores centrarse en cómo dar acceso a los mismos a través de un sistema de diálogo sin tener que lidiar de forma tan detallada como antes con los aspectos relacionados con el servicio que se presta.

A pesar de estos avances, todavía queda mucho camino por recorrer antes de que las interfaces conversacionales alcancen un rendimiento similar al de los seres humanos.

## 2.3. ESTRUCTURA DEL DISCURSO ORIENTADA A LA INFORMACIÓN

Presentamos a continuación varios modelos y teorías que han surgido durante las últimas décadas dirigidas a obtener un análisis semántico de la información en un discurso o conversación.

### 2.3.1. *DRT: Teoría de Representación del Discurso*

La Teoría de Representación del Discurso (DRT) [74] representa el significado del discurso (Estructura de Representación del Discurso, DRS) mediante un lenguaje lógico similar a la lógica de

[10] *https://www.pandorabots.com/*
[11] *https://dialogflow.com/*
[12] *https://chatfuel.com/*
[13] *https://aws.amazon.com/es/lex/*
[14] *https://www.ibm.com/watson/*
[15] *https://lekta.ai/*

predicados de primer orden, extendida para representar el contexto considerando las frases precedentes.

### 2.3.2. *Modelo de Discurso Lingüístico*

En el Modelo de Discurso Lingüístico (LDM) [75], la representación semántica se basa en Árboles Sintácticos de Discurso (DPT) que consideran las relaciones estructurales y las categorías contextuales para capturar el contenido proposicional y el contexto del discurso. Cada nodo en este árbol (Unidades Constitutivas del Discurso, DCU) representa una unidad semántica descrita mediante operadores del discurso que denotan las relaciones entre las DCU. Suelen utilizarse gramáticas del discurso para construir el DPT a partir de DCU.

### 2.3.3. *Teoría de la Representación del Discurso Segmentado*

La Teoría de representación del discurso segmentado (SDRT) [76] amplía el modelo DRT estableciendo una relación entre la frase actual y la anterior y determinando cómo combinar estas dos oraciones con el análisis semántico de las oraciones anteriores. La principal diferencia entre este modelo y el LDM es que las relaciones en SDRT afectan no solo a la estructura del discurso, sino también a su representación semántica.

### 2.3.4. *RST: Teoría de la Estructura Retórica*

La Teoría de la Estructura Retórica (RST) [77, 78] describe la estructura de un discurso por medio de las relaciones de coherencia entre sus partes. Estas relaciones se clasifican en relaciones temáticas (informativas) y relaciones de presentación (intencionales). Estos modelos se han utilizado principalmente para la generación automática de textos y de resúmenes.

DRT, LDM y SDRT difieren en cómo la estructura del discurso afecta a su representación semántica y no modelan explícitamente las intenciones de los usuarios. Aunque RST considera tanto las perspectivas informativas como las intencionales, la teoría también se centra más en la perspectiva informacional y no considera las relaciones entre oraciones para crear la representación semántica del discurso.

### 2.3.5. *La teoría de los Actos de Habla*

Las teorías orientadas a la intención modelan la estructura del diálogo a través de las relaciones entre actos del habla o actos de diálogo. La teoría de los Actos de Habla [79, 80] analiza la estructura del discurso teniendo en cuenta la intención del locutor y los efectos sobre el oyente. Los actos de diálogo básicos se clasifican en ilocucionarios, comisivos, expresivos y declaraciones. Muchos investigadores han modificado esta taxonomía básica del habla agregando más actos específicos de dominio. Además, los actos de diálogo suelen ser un componente clave en otras teorías que modelan la estructura del discurso, como las gramáticas de diálogo, los modelos basados en planes y la teoría de los actos de conversación.

### 2.3.6. *El esquema de anotación DAMSL*

El esquema de anotación DAMSL (*Dialog Act Markup in Several Layers*) [81, 82] amplía la teoría del acto de habla para anotar conversaciones orientadas a tareas mediante el uso de varias etiquetas definidas en tres capas ortogonales (Funciones comunicativas hacia adelante, Funciones comunicativas hacia atrás y Expresión Caracteristica). Estas capas se usan respectivamente para definir una taxonomía similar a la teoría del acto de habla, indican las relaciones entre el enunciado actual y los anteriores, y describen el contenido y la forma de un enunciado. Este esquema de anotación se puede extender con capas y actos específicos de dominio adicionales. El principal inconveniente del modelo es que la única relación que se captura es la que existe entre el enunciado actual y el anterior.

### 2.3.7. *Gramática de diálogo*

Las gramáticas de diálogo se basan en la detección de los patrones regulares que están presentes en las conversaciones colaborativas que contienen patrones regulares, como los pares pregunta / respuesta [83]. Estos patrones regulares jerárquicos se pueden expresar mediante gramáticas en las que las reglas especifican cómo se puede segmentar el diálogo en unidades más pequeñas (por ejemplo, objetivos, subobjetivos y actos de diálogo). Aunque las gramáticas de diálogo se pueden utilizar para predecir el siguiente elemento de una conversación, es muy difícil generar el conjunto completo de reglas que cubre todas las posibles variaciones de conversaciones en un dominio complejo.

### 2.3.8. *Sistemas de diálogo basados en planes*

En el modelo [84] basado en el plan, los diálogos se perciben como un plan que los participantes siguen para proporcionar respuestas adecuadas y alcanzar objetivos específicos. El plan describe cómo el discurso actúa y las intenciones del hablante modelo se relacionan con el objetivo de la conversación. El enfoque de diálogo basado en planes para la gestión de diálogos se ha aplicado en muchos sistemas de diálogo complejos [80]. Sin embargo, estos modelos generalmente hacen suposiciones fuertes sobre el plan y el entorno en el que se ejecutará (por ejemplo, las creencias de los participantes del diálogo no cambian), que no son prácticas en situaciones reales. Además, los modelos basados en planes fallan para los diálogos que no siguen de cerca la estructura de la tarea (por ejemplo, cambios de tema, subdiálogos de aclaración o corrección, etc.). Se han propuesto varios modelos aumentados basados en planes para abordar estos problemas.

### 2.3.9. *Teoría de la Conversación*

La teoría de la Conversación [85] modela el diálogo como un conjunto de acciones de hablante en lugar de acciones de agente único para hacer que las acciones de toma de tierra sean más explícitas. Se definen cuatro niveles de acciones para mantener el contenido y la coherencia del diálogo: actos de toma de turnos, actos de toma de tierra, actos centrales de habla y actos de argumentación. Estos niveles capturan información clara de diálogo y se pueden emplear de forma independiente el uno del otro en un sistema de diálogo.

### 2.3.10. *Teoría de la Estructura Discursiva*

La Teoría de la Estructura Discursiva (GST) [86] de Grosz y Sidner modela la estructura del diálogo utilizando los conceptos de unidad discursiva y coherencia discursiva. Las unidades de discurso, que se definen en función de la intención de los participantes del diálogo, son un conjunto de enunciados que cumplen una función específica en el objetivo general del diálogo. La estructura intencional se utiliza para capturar la coherencia del discurso. Aunque GST describe un modelo abstracto de estructuras discursivas, la teoría no detalla cómo resolver la segmentación del discurso, el reconocimiento del segmento del discurso y la detección de las relaciones entre ellos.

#### 2.3.11. *La teoría de la actualización del estado de la información*

La *Information State Theory* [87] usa el concepto de estado de la información para representar la información que se puede usar para diferenciar un diálogo de los demás. En lugar de definir una representación específica para la estructura de diálogo, la teoría proporciona un marco general para el modelado de diálogo que puede implementarse en el contexto de cualquier teoría de estructura de diálogo.

# 3. SISTEMAS CONVERSACIONALES: ECOSISTEMA TECNOLÓGICO

Este capítulo introduce las principales tecnologías que se requieren para crear un sistema conversacional, junto con los procesos y alternativas utilizadas habitualmente para su resolución.

## 3.1. RECONOCIMIENTO DEL HABLA

El reconocimiento automático del habla ha sido un área de investigación activa durante más de 50 años. A lo largo de las últimas décadas, se ha progresado desde el reconocimiento de palabras aisladas dentro de conjuntos reducidos de vocabulario hasta el reconocimiento de habla continua y con conjuntos de vocabulario cada vez mayores. Estos avances han supuesto que la comunicación con los sistemas conversacionales pueda efectuarse cada vez de una forma más natural. No obstante, ha sido recientemente, motivado con el auge de los asistentes con voz en teléfonos inteligentes y la utilización de redes neuronales de aprendizaje profundo, cuando la precisión en el reconocimiento ha mejorado considerablemente. Diferentes aspectos que suelen tenerse en cuenta a la hora de establecer clasificaciones de los reconocedores son el tipo de usuarios permitidos (sistemas independientes o dependientes del usuario), el estilo de habla soportado (reconocedores de palabras aisladas, palabras conectadas o de habla continua) o el tamaño del vocabulario (pequeño, medio o gran vocabulario).

La complejidad de la tarea de reconocimiento radica en diversos problemas:

- La variabilidad acústica: cada persona pronuncia los sonidos de manera diferente cuando habla. Originalmente, los usuarios tenían que leer horas de texto a sistemas de dictado de voz para crear un modelo dependiente del hablante. Este pre-entrenamiento del reconocedor no es posible en sistemas que ofrecen información y servicios a usuarios desconocidos

y/u ocasionales (por ejemplo, en los call centers), por lo que se ha realizado un esfuerzo considerable para que el reconocedor sea independiente del hablante. Se presentan todavía problemas cuando los hablantes tienen patrones de habla atípicos, como acentos fuertes o discapacidades del habla.

- La confusión acústica: muchas palabras suenan de forma similar, lo que dificulta su distinción.
- El problema de la coarticulación: las características de los sonidos pronunciados pueden variar en función de los sonidos vecinos.
- El tamaño del vocabulario: los primeros reconocedores eran capaces de reconocer solo una pequeña cantidad de palabras, como variantes de “sí” o “no” o cadenas de dígitos. Gradualmente, los vocabularios se han expandido, primero a decenas de miles de palabras y en los sistemas actuales a vocabularios de millones de palabras.
- Los fenómenos propios del habla espontánea y continua: interjecciones, pausas, dudas, falsos comienzos, repeticiones de palabras, autocorrecciones, etc. El habla continua es difícil de procesar porque no hay una marca clara de los límites entre las palabras, y la pronunciación de un sonido individual puede verse afectada por la coarticulación. Los primeros sistemas de reconocimiento automático del habla utilizaban el reconocimiento de palabras aislado en el que el hablante tenía que hacer una breve pausa entre cada palabra, pero hoy en día, los sistemas de reconocimiento son capaces de gestionar habla continua y producida espontáneamente. El reconocimiento del habla en conversaciones en las que existan varias personas que puedan incluso hablar simultáneamente representa todavía un reto.
- Las condiciones del entorno (ruido, distorsiones del canal, limitaciones de ancho de banda, etc.). El reconocimiento tiende a degradarse en condiciones ruidosas, donde el ruido puede ser transitorio (por ejemplo, tráfico) o constante (por ejemplo, música de fondo o voces adicionales). Las técnicas para mejorar la robustez del reconocimiento incluyen la separación de señales, la mejora de características, la compensación de modelos y la adaptación de modelos. El uso de modelos acústicos basados en redes neuronales de aprendizaje profundo también ha permitido mejorar la robustez de estos sistemas.
- Calidad de los micrófonos. El rendimiento del reconocimiento es dependiente de la calidad de los micrófonos. Además, cuando los usuarios hablan a través de un teléfono inteligente, a veces sostienen el dispositivo cerca de su boca y otras veces lejos de ellos para ver la

pantalla, de modo que el micrófono no está siempre a la misma distancia y debe abordarse el reconocimiento desde cualquier dirección y a cierta distancia de los micrófonos.

Por estos motivos, cuando se utilice un sistema de reconocimiento automático del habla, se debe analizar la calidad de las palabras reconocidas o de los conceptos comprendidos por el sistema con el fin de detectar posibles errores o zonas de gran ambigüedad. Esta necesidad es aún más importante en los sistemas de diálogo, donde una mala interpretación de la frase pronunciada puede llevar al sistema a realizar un comportamiento erróneo, ya que la salida del reconocedor es el punto de partida del resto de módulos del sistema. Entre las diferentes técnicas utilizadas para el desarrollo de reconocedores, sin duda, la aproximación estadística es actualmente la más utilizada. En esta aproximación, el problema del reconocimiento puede entenderse como encontrar la secuencia de palabras $W$ pronunciadas dada una secuencia de datos acústicos $A$. Esta secuencia puede determinarse siguiendo la expresión:

$$W = \underset{W}{\text{máx}}\, P(W|A)$$

Utilizando la regla de Bayes, la expresión anterior puede reescribirse de la siguiente forma:

$$P(W|A) = \frac{P(A|W)P(W)}{P(A)}$$

donde $P(A|W)$ se denomina modelo acústico (probabilidad de obtener la secuencia acústica $A$ cuando se ha pronunciado la secuencia de palabras $W$) y $P(W)$ es proporcionada por el modelo de lenguaje (probabilidad de pronunciar la secuencia de palabras). Dado que la probabilidad de la secuencia acústica es independiente de la secuencia de palabras, la expresión puede escribirse del siguiente modo:

$$\text{W} = \underset{W}{\text{máx}} P(A|W)P(W)$$

Para llevar a la práctica esta aproximación, la solución más utilizada desde mediados de los 70 hasta recientemente consiste en el modelado de las unidades acústicas mediante modelos ocultos de Markov (HMM), como es el caso de reconocedores como HTK (*Hidden Markov Model Toolkit*) [88] o Sphinx [89]. El éxito de los HMM se basa principalmente en la existencia de algoritmos de aprendizaje automático de los parámetros del modelo [90], así como en su capacidad para representar el habla como un fenómeno secuencial en el tiempo. Se han estudiado múltiples aproximaciones, como son los modelos discretos, semicontinuos o continuos, así como diversas topologías de los modelos.

Uno de los elementos imprescindibles para disponer de un reconocedor de habla continua es el modelo de lenguaje. Dado que la percepción de la acústica es a menudo insuficiente, incluso para

las personas, para reconocer la secuencia de fonemas o palabras pronunciadas, es necesario tener un modelo de concatenación de palabras. Los modelos de lenguaje más utilizados son los basados en N-gramas [91] [92] y los basados en gramáticas regulares [93] o independientes del contexto [94]. Las gramáticas suelen ser adecuadas para tareas reducidas, ya que permiten más precisión en el tipo de restricciones que imponen, pero son incapaces de representar la gran variabilidad del habla natural.

Los N-gramas, sin embargo, recogen de una forma más sencilla las concatenaciones entre palabras, pero son muy adecuados cuando se dispone de un número suficiente de muestras de entrenamiento. En ambos casos la existencia de técnicas de aprendizaje automático (Inferencia Gramatical) para las Gramáticas [95], técnicas de estimación de parámetros para los N-gramas y técnicas de suavizado permite la obtención de modelos adecuados para el reconocimiento del habla. En el caso de HMM como representación acústico-fonética y de N-gramas como modelos de lenguaje, se construye una red de estados en que las palabras están representadas por la concatenación de modelos de los fonemas que las componen. Los avances también son significativos en cuanto a la correcta identificación de los fonemas en entornos ruidosos. Actualmente, los métodos estadísticos proporcionan las mejores tasas de reconocimiento en estos entornos de acústica deficiente, como es el caso de la locución telefónica.

## 3.2. COMPRENSIÓN DEL LENGUAJE NATURAL

El proceso de comprensión puede entenderse como un cambio en el lenguaje de representación, de lenguaje natural a un lenguaje semántico, de forma que se mantenga el significado del mensaje. Al igual que en el reconocedor de voz, el módulo de comprensión puede trabajar con varias hipótesis (tanto de reconocimiento como de comprensión) y con medidas de confianza.

Para afrontar el problema de la comprensión existen actualmente dos grandes aproximaciones: la comprensión basada en reglas y la comprensión basada en modelos estadísticos estimados a partir de datos.

Las alternativas basadas en reglas extraen la información semántica a partir del análisis sintáctico-semántico de las frases, utilizando gramáticas definidas para la tarea, o a partir de la detección de palabras (o secuencias de palabras) clave, con significado semántico. Algunos analizadores, con el objetivo de mejorar la robustez del análisis, combinan los aspectos sintácticos y semánticos de la tarea. Otras técnicas se basan en aplicar un análisis a dos niveles, en el cual se utilizan gramáticas

para llevar a cabo un análisis detallado de la frase y extraer la información semántica relevante. Además, existen sistemas que utilizan analizadores basados en reglas aprendidas de forma automática a partir de un corpus de entrenamiento utilizando técnicas de procesamiento del lenguaje natural.

En el caso de los métodos estadísticos, el proceso se basa en la definición de unidades lingüísticas con contenido semántico y en la obtención de modelos a partir de muestras etiquetadas. Este tipo de análisis [96] [97] emplea un modelo probabilístico para identificar los conceptos, marcadores y valores de los casos, para representar las relaciones entre los marcadores de los casos y sus valores y para decodificar semánticamente las pronunciaciones del usuario.

El modelo se genera durante una fase de entrenamiento (aprendizaje), donde sus parámetros capturan las correspondencias entre las entradas de texto y su representación semántica. Una vez el modelo de entrenamiento se ha aprendido, se emplea a modo de decodificador para generar la mejor representación semántica de la entrada. De este modo, el proceso de comprensión se realiza de forma similar al reconocimiento del habla. Mediante el algoritmo de Viterbi puede interpretarse como un proceso de traducción de una frase de entrada (secuencia de palabras) en una frase de salida (secuencia de unidades semánticas).

La definición del lenguaje semántico se basa en un gran número de casos en la utilización del concepto de *frame*. En esta aproximación, la representación generada por el módulo de compresión contiene *conceptos* (los diferentes tipos de consultas que puede realizar el usuario) y *atributos* (información que debe aportar el usuario para completar o modificar la consulta requerida al sistema). De este modo, todo mensaje enviado por el módulo de comprensión al gestor del diálogo tras cada intervención del usuario es un frame.

En comparación con el reconocimiento automático del habla, la tarea de comprensión del lenguaje es mucho más diversa, ya que abarca una variedad mayor de tecnologías y enfoques diferentes. La elección de un enfoque particular depende especialmente de las tareas que realice el comprendedor; por ejemplo, puede ser importante realizar tareas de bajo nivel, como la normalización de la entrada antes de pasar a tareas de nivel superior. Extraer el significado puede en algunos casos requerir solo la identificación de palabras clave, mientras que en otros casos puede requerirse una comprensión más profunda. A continuación, revisamos varias tecnologías y enfoques diferentes. Estas tecnologías y enfoques no son necesariamente mutuamente excluyentes y es frecuente que se apliquen varias tecnologías en las diferentes etapas del proceso de interpretación semántica.

**Reconocimiento de actos de diálogo**: El reconocimiento de actos del diálogo, también conocido como determinación de la intención del usuario o la clasificación de elocuciones, implica determinar la función de una elocución en un diálogo, por ejemplo, si se trata de una pregunta, sugerencia, orden, etc. En la década de los 80 y principios de los 90, el reconocimiento de actos de diálogo utilizaba modelos de inferencia y razonamiento. A finales de la década de 1990, se desarrollaron aproximaciones alternativas basadas en métodos estadísticos, que modelan el reconocimiento de actos de diálogo como una tarea de clasificación supervisada.

Stolcke et al. [98] utilizaron una combinación de características léxicas, colocacionales y prosódicas, así como secuencias de actos de diálogo. Modelaron los diálogos como un Modelo Oculto de Markov. En otros trabajos, se han utilizado Redes Bayesianas para la clasificación [99] [100]. Para conocer una descripción general de los enfoques para el reconocimiento de actos de diálogo, se recomienda consultar [101] [102].

El reconocimiento estadístico de actos de diálogo implica entrenar clasificadores con un corpus de turnos de diálogo donde cada expresión se anota en términos de actos de diálogo. El corpus Switchboard [103] se ha utilizado en muchos estudios, y los conjuntos de etiquetas para anotar el corpus incluyen la propuesta *Dialog Act Markup in Several Layers* (DAMSL) [81], que proporciona un conjunto de tipos de actos de diálogo independientes del dominio de interacción. Otros sistemas, como Verbmobil, utilizan actos de diálogo definidos específicamente para el dominio del sistema. Entre las características que se utilizan a menudo para la clasificación se encuentran las siguientes: palabras y expresiones en la elocución (generalmente etiquetados utilizando n-gramas), prosodia, información sintáctica y semántica obtenida de un procesamiento lingüístico (en forma de n-gramas de actos de diálogo).

## 3.3. GESTIÓN DEL DIÁLOGO

Uno de los aspectos centrales en el desarrollo de interfaces conversacionales es diseñar el modelo de gestión de diálogo. Este modelo define el comportamiento del sistema como respuesta a las intervenciones de los usuarios, el contexto de la interacción y la consulta a repositorios de información. El diseño de este modelo generalmente se lleva a cabo en los sistemas comerciales mediante la elaboración manual de estrategias de diálogo estrechamente vinculadas al dominio de la aplicación para optimizar el comportamiento del sistema en dicho dominio. Más recientemente, está ganando popularidad el diseño de estrategias de diálogo mediante el uso de modelos

estadísticos entrenados con conversaciones reales.

La gestión de diálogo se puede entender como una de las tareas más importantes de un interfaz conversacional dado que este componente encapsula la lógica de la aplicación e influye notablemente en la satisfacción del usuario. La complejidad del gestor de diálogo depende especialmente de la tarea, la flexibilidad e iniciativa de los diálogos. El gestor de diálogo procesa diferentes fuentes de información y la selección de la siguiente acción del sistema depende de múltiples factores (los resultados del módulo de comprensión y del reconocedor automático del habla, los resultados de las consultas a las bases de datos, el conocimiento del dominio de la aplicación y de sus restricciones, el conocimiento sobre los usuarios y la historia del diálogo, el número de confirmaciones previas, el estado de los dispositivos que controla el sistema, etc.).

Dado que los resultados de los módulos de reconocimiento de voz y comprensión del lenguaje pueden contener errores, uno de los principales objetivos del gestor de diálogo es realizar la detección y corrección de dichos errores. Una forma común de realizar esta función es emplear un nivel de confianza para el resultado de los módulos de Reconocimiento automático del habla y Comprensión del Lenguaje Natural, y utilizarlos para decidir cuándo aceptar las hipótesis de estos módulos, solicitar una confirmación al usuario, o rechazar las hipótesis y solicitar de nuevo la información del usuario. Es importante reducir al mínimo el número de confirmaciones y rechazos, al tiempo que se preserva un nivel de precisión razonable.

La estrategia de interacción de una interfaz conversacional determina quién toma la iniciativa en el diálogo: el sistema, el usuario o ambos. En la literatura a menudo se distinguen tres tipos de estrategias de interacción: dirigida por el usuario, dirigida por el sistema e iniciativa mixta. Cuando se utiliza una iniciativa dirigida por el usuario, el usuario siempre tiene la iniciativa en el diálogo, y el sistema se limita a responder a las consultas y los comandos del usuario. El principal problema con esta estrategia es que el usuario puede pensar que es libre de decir lo que quiera, lo que tiende a causar errores de Reconocimiento automático del habla y Comprensión del Lenguaje Natural.

Cuando se utiliza una iniciativa dirigida por el sistema, éste tiene la iniciativa en el diálogo, y el usuario simplemente responde sus consultas. La ventaja de esta estrategia es que ayuda a restringir las respuestas del usuario, lo que lleva a diálogos más eficientes. La desventaja es la falta de flexibilidad, ya que el usuario está restringido a comportarse de acuerdo con las expectativas del sistema, proporcionando los datos necesarios para realizar alguna acción en el orden especificado

por el sistema.

Cuando se utiliza la iniciativa mixta, tanto el usuario como el sistema pueden tomar la iniciativa en el diálogo. La ventaja es que el sistema puede guiar al usuario en las tareas que se llevarán a cabo, mientras que el usuario puede tomar la iniciativa, hacer preguntas, presentar nuevos temas y proporcionar respuestas que aporten más información que la requerida por el sistema. El mayor inconveniente de la iniciativa mixta es que el usuario puede decir cualquier cosa e introducir un tema diferente lo que puede hacer que el sistema pierda de vista su agenda. Por lo tanto, los diálogos de iniciativa mixta requieren capacidades avanzadas de Reconocimiento Automático del Habla y Comprensión del Lenguaje Natural, así como la capacidad de mantener y controlar el historial de diálogo y la agenda del sistema.

A menudo se emplean dos tipos de estrategias de confirmación: confirmación explícita y confirmación implícita. Con confirmación explícita, el sistema confirma los datos aportados por el usuario tal y como muestra el siguiente ejemplo:

**Usuario:** *Quiero saber los horarios desde Madrid.*

**Sistema:** *¿Quieres salir de Madrid?*

**Usuario:** *Sí.*

La desventaja de las confirmaciones explícitas es que el diálogo tiende a alargarse debido a estos turnos de confirmación adicionales, y esto hace que la interacción sea menos eficiente e incluso excesivamente repetitiva si todos los elementos de datos proporcionados por el usuario deben ser confirmados.

Cuando se utiliza la estrategia de confirmación implícita, el sistema incluye parte de la información anterior del usuario en su próxima pregunta. Si el usuario responde la pregunta directamente, por ejemplo, en este caso indicando una hora de salida, se supone que la información anterior sobre el destino se confirma implícitamente y no se requieren turnos adicionales.

Estas estrategias de confirmación son útiles para evitar malentendidos, por ejemplo, cuando el sistema ha entendido algo de lo que no está seguro con precisión. Una situación relacionada, pero diferente es la no comprensión, que ocurre cuando el sistema no ha podido recopilar ningún dato tras la intervención del usuario. En este caso, dos estrategias típicas para manejar el error son pedir al usuario que repita la entrada, o solicitarle que la reformule.

Se han desarrollado varias aproximaciones diferentes de gestión de diálogo dentro de la comunidad investigadora y en la industria [104, 105]. Estos enfoques se pueden clasificar en dos categorías principales: aproximaciones basadas en reglas y aproximaciones estadísticas o basados en datos y metodologías de aprendizaje automático. Las aproximaciones híbridas combinan estos dos enfoques principales.

## 3.4. GENERACIÓN DEL LENGUAJE NATURAL

La generación del lenguaje natural es el proceso de obtención de textos en lenguaje natural a partir de una representación no lingüística. Es importante obtener mensajes legibles, optimizando el texto empleando expresiones referenciales y nexos y adaptando el vocabulario y la complejidad de las estructuras sintácticas a la destreza lingüística del usuario.

El enfoque más sencillo consiste en emplear mensajes de texto predefinidos (como por ejemplo mensajes de error o avisos). Aunque es intuitivo, este enfoque carece de flexibilidad. El siguiente nivel de sofisticación es la generación basada en plantillas, en las que una misma estructura de mensaje se utiliza incluyendo ligeras alteraciones. El enfoque de plantilla se emplea principalmente para la generación de frases en aplicaciones con texto de estructura muy regular como informes de negocios.

Los sistemas basados en frases emplean lo que puede considerarse como plantillas generalizadas, a nivel de oración (las frases se asemejan a reglas gramaticales), o a nivel del discurso (en este caso a menudo se denominan planes de texto). En estos sistemas se selecciona en primer lugar un patrón para emparejar el nivel superior de la entrada y seguidamente cada parte del patrón se amplía en uno más específico que se empareja con una determinada porción de la misma. El proceso en cascada se detiene cuando cada patrón ha sido substituido por una o más palabras.

Finalmente, los sistemas basados en características representan, en cierto modo, el nivel máximo de generalización y flexibilidad. En estos sistemas, cada alternativa mínima posible de expresión se representa por una sola característica; por ejemplo, si la frase es positiva o negativa, si es una pregunta o un imperativo o una declaración, o su tiempo verbal. Para ordenar las características es necesario emplear conocimiento lingüístico, alternativamente se puede generar el lenguaje natural basándose en corpus [106], construyendo las elocuciones del sistema de forma estadística.

### 3.5. SÍNTESIS DEL HABLA

Los sintetizadores de texto a voz transforman un texto en una señal acústica. Un sistema de síntesis oral se compone de dos partes: un “front-end” y un “back-end”. El front-end realiza dos tareas fundamentales. En primer lugar, convierte el texto plano, (que contiene símbolos tales como números y abreviaturas) en sus palabras asociadas. Este proceso se denomina usualmente normalización, proceso previo, o tokenización del texto. En segundo lugar, asigna transcripciones fonéticas a cada palabra, y divide y marca el texto en unidades prosódicas, es decir frases, cláusulas, y oraciones.

El proceso de asignar transcripciones fonéticas a las palabras se llama conversión texto-a-fonema o conversión de grafema-a-fonema. La salida del back-end es la representación simbólica constituida por las transcripciones fonéticas y la información prosódica. El back-end (denominado a menudo sintetizador) convierte la representación lingüística simbólica en sonido. Por una parte, la síntesis del habla se puede basar en la producción de voz humana, este es el caso de la síntesis paramétrica (que simula los parámetros fisiológicos del tracto vocal) y de la síntesis basada en armónicos (que modela la vibración de las cuerdas vocales). En esta última técnica, parámetros tales como la frecuencia fundamental, la expresión, y los niveles de ruido se varían en el tiempo para crear la onda del discurso artificial. Otra aproximación basada en modelos fisiológicos es la síntesis articulatoria, que comprende técnicas de cálculo para modelar el tracto vocal humano y los procesos de articulación.

Parte II

# CARACTERÍSTICAS BÁSICAS DE UN SISTEMA CONVERSACIONAL

# 4. GESTIÓN DEL DIÁLOGO: CONTROL Y MODELADO DEL DIÁLOGO

Como se ha indicado en el Capítulo 3 al estudiar el ecosistema tecnológico que compone los principales módulos o componentes de un sistema conversacional, el método de gestión de diálogo se encarga de la coordinación de todas las tareas implicadas en una conversación, desde la captura de la entrada del usuario, integración con sistemas funcionales externos tales como bases de datos, servicios web, etc., representación de la historia del diálogo, perfil del usuario, ayuda de la interacción, proponiendo la salida que el sistema comunicará al usuario en la próxima interacción, o encargándose incluso de operaciones que enlazan el sistema conversacional con analítica de eventos y datos, *business intelligence*, generación de informes operativos, etc. A continuación, este capítulo describe los principales enfoques para el control y modelado del diálogo, haciendo un especial énfasis en las aproximaciones estadísticas más recientes.

## 4.1. DIÁLOGOS COMO GRAFOS DE TRANSICIONES ENTRE ESTADOS

En los sistemas de estados finitos el flujo del diálogo puede determinarse con anterioridad y representarse por medio de una red o gramática. Utilizando la representación del diálogo mediante una red de transiciones entre estados, los nodos representan las preguntas del sistema y las transiciones determinan el conjunto de caminos que pueden establecerse en la red, de modo que la interacción está completamente estructurada: el gestor de diálogo se desplaza a través de la red, en la que se indica qué información debe ser intercambiada en cada estado del diálogo, y obtiene la información del usuario necesaria para llevar a cabo una determinada tarea.

La principal ventaja de esta aproximación es su simplicidad, facilitando el desarrollo de gestores cuando las tareas sean muy sencillas, estas tareas estén claramente estructuradas y en dichas tareas exista un número pequeño de tipos de respuestas. Los principales inconvenientes residen en el hecho de no ser adecuados para gestionar diálogos complejos y su falta de flexibilidad, dado que los usuarios no pueden desviarse de los caminos establecidos para cada estado. El diseño del gestor de diálogo requiere una labor intensiva, específica para cada dominio, y basada en la depuración de errores conforme se van detectando, ya que el control del flujo del diálogo se determina de forma manual. Para su implementación suelen utilizarse reglas gramaticales y diversos tipos de máquinas de estados. Diferentes trabajos relativos al uso de gramáticas para representar el diálogo son el sistema SUNDIAL [107] y SUNSTAR [108]. El gestor de diálogo del sistema TOOT utiliza una máquina de estados finitos para controlar la interacción, basada en el estado actual

del sistema y los resultados suministrados por el reconocedor. El gestor consta de 168 estados, cada uno de los cuales está asociado con una de las 12 gramáticas definidas, que especifican el modelo de lenguaje del reconocedor en ese punto del diálogo.

## 4.2. CONTROL DEL DIÁLOGO BASADO EN FRAMES

El objetivo de esta aproximación es solucionar la falta de flexibilidad de los modelos de estados finitos. Se asemejan a esta aproximación en que ambas son capaces de gestionar tareas basadas en completar un formulario de datos solicitando información al usuario para posteriormente realizar una consulta a una fuente de conocimiento externa (form-filling tasks). La diferencia estriba en la no necesidad de seguir un orden preestablecido para completar los diferentes campos requeridos, de forma que se le dote al sistema de iniciativa mixta. Para permitir este grado de flexibilidad mayor, es necesario dotar al sistema de tres componentes:

- Un frame que haga referencia a los conceptos y atributos definidos para la tarea.
- Una gramática o modelo de lenguaje para el reconocedor más extensa.
- Un algoritmo de control del diálogo que determine las próximas acciones del sistema basándose en los contenidos del frame.

En la definición del frame pueden considerarse valores adicionales a la simple anotación de si el campo (slot) posee o no un valor, por ejemplo, utilizar las medidas de confianza para hacer referencia a la fiabilidad del dato almacenado. En [109] se presenta un gestor de diálogo en el dominio de anuncios de compra-venta de coches usados en el que se extiende esta idea. La variante de frame definida, denominada E-form (electronic form), incluye además información sobre las preferencias del usuario en forma de prioridades. Estos formularios se utilizan en la gestión de diálogo desarrollada en el sistema Bell Labs Communicator. Ejemplos de utilización de la aproximación basada en frames son los sistemas JUPITER, ARISE, WITAS, COMIC, etc.

Una alternativa para la elaboración de estrategias de gestión basadas en la definición de reglas, es el lenguaje VoiceXML [110]. VoiceXML surge del trabajo conjunto de varias importantes compañías (AT&T, IBM, Lucent, Motorola, etc) que formaron el denominado VoiceXML Forum. VoiceXML permite crear diálogos con audio, posibilitando la síntesis, el reconocimiento y la grabación del habla, así como el desarrollo de conversaciones con iniciativa mixta. A todas estas funcionalidades, VoiceXML aporta las ventajas del desarrollo basado en las aplicaciones Web.

## 4.3. ENFOQUES BASADOS EN PLANES

Los sistemas basados en planes consideran que los seres humanos se comunican para conseguir objetivos. Las intervenciones del usuario suelen representarse como actos de diálogo [**?**] y se utilizan para alcanzar estas metas. Las teorías de modelado del diálogo basadas en planes [111, 112, 113] argumentan que los actos de diálogo del usuario forman parte de un plan u objetivo global, que el sistema tiene que identificar y para a continuación responder de la forma más adecuada. Las principales críticas que se realizan a esta aproximación se basan en la dificultad de realizar el reconocimiento del plan del usuario.

La Teoría de los Juegos Conversacionales [114] utiliza técnicas mixtas correspondientes a las gramáticas de discurso y a la aproximación de sistemas basados en planes. Esta teoría se ha utilizado para la representación de modelos de diálogo orientados a la tarea [115], en los que el diálogo consiste en una o más transacciones definidas como subtareas. Otra variante, denominada aproximación colaborativa, se basa en la visión del diálogo como un proceso en el que el usuario y la máquina deben trabajar conjuntamente para conseguir la comprensión mutua del diálogo. Este tipo de aproximación trata de capturar el conjunto de motivaciones que se producen durante el diálogo, en lugar de centrarse en la estructura de la tarea. Un ejemplo de utilización de esta aproximación es el gestor de diálogo del sistema TRAINS-93.

## 4.4. ENFOQUES BASADOS EN AGENTES

En los sistemas basados en agentes el flujo del diálogo se determina de forma dinámica a través de un proceso en el que el gestor de diálogo lleva a cabo un cierto razonamiento para determinar las próximas acciones. Estas aproximaciones utilizan técnicas de Inteligencia Artificial y orientan el modelado del diálogo como una colaboración entre agentes inteligentes para solucionar un determinado problema o tarea. Son adecuadas para la gestión del diálogo en tareas complejas, como las negociaciones y la resolución de problemas.

La comunicación en este tipo de aproximación puede verse como la interacción entre dos agentes, cada uno de los cuales dispone de su propio conjunto de acciones. Estas aproximaciones incluyen normalmente mecanismos para la detección y corrección de errores, definen estructuras para tener en cuenta la historia del diálogo y se basan usualmente en iniciativas mixtas para el control del diálogo.

Ejemplos de estas aproximaciones son el gestor Agenda desarrollado por la CMU y la arquitectura Ravenclaw, Queen's Communicator, SesaMe y la arquitectura JASPIS. En estas aproximaciones se utilizan aproximaciones orientadas a objetos y se intenta separar la gestión genérica del diálogo de las acciones específicas del dominio.

## 4.5. LA TEORÍA DE LOS ESTADOS DE LA INFORMACIÓN

El proyecto Trindi [116] propuso una arquitectura para el desarrollo de gestores de diálogo basada en el concepto de estado de la información (*information state*) [117]. La aproximación de estados de información intenta combinar los puntos fuertes de los modelos de diálogo basados en estados y los basados en planes, utilizando aspectos del modelado del diálogo mediante estados e incluyendo representación semántica detallada y nociones de obligaciones, compromisos, creencias y planes. La teoría del modelado del diálogo basada en el concepto de estado de la información se caracteriza por los siguientes conceptos:

- Una descripción de los componentes de la información (*informational components*), que incluyen el contexto y factores de motivación internos (intenciones, creencias, compromisos, etc.).
- Representaciones formales de los componentes anteriores (como listas, conjuntos, vectores de características, etc.).
- Un conjunto de acciones o cambios en el diálogo que provocarán la actualización del estado de información.
- Un conjunto de reglas de actualización que gobiernan las transiciones entre estados de información cuando se cumplan condiciones preestablecidas en el estado de información actual y se hayan llevado a cabo acciones previas durante el diálogo.
- Una estrategia de actualización para decidir qué regla (o reglas) se selecciona del conjunto de reglas aplicables en un determinado instante del diálogo.

El estado de la información hace referencia a toda la información necesaria para distinguir un diálogo de otros diálogos, incluyendo una representación de la historia de las acciones previas llevadas a cabo durante el diálogo, utilizada para decidir la nueva acción que realizará el sistema.

Se trata de un concepto muy abstracto, instanciado muchas veces como modelo mental, contexto del diálogo, medida conversacional, etc.

Cabe realizar una distinción entre las aproximaciones basadas en el estado de la información y las aproximaciones basadas en el estado del diálogo. En éstas últimas, el diálogo se comporta de acuerdo a una gramática en la que los estados representan los resultados de las acciones llevadas a cabo en el estado anterior y cada estado dispone de un conjunto de transiciones o acciones permitidas. La información es implícita únicamente del estado y de la relación que pueda tener con los otros estados.

TrindiKit [118] fue la primera versión completa desarrollada para esta aproximación del estado de la información. Escrita en Prolog, modelaba el sistema de diálogo definiendo estados de la información, reglas de actualización y selección, y algoritmos de control que gobiernan las reglas a aplicar en el estado de información. Los estados de información se definen usualmente como una estructura recursiva de la forma *Nombre: Tipo*, donde *Nombre* es un identificador y *Tipo* es un tipo específico de datos.

Ejemplos de gestores de diálogo desarrollados utilizando este paradigma son los implementados para los proyectos WITAS y COMIC. El gestor de diálogo desarrollado para WITAS se basa en la utilización de una estructura en forma de árbol de estados de diálogo, donde los nodos representan los cambios en el diálogo y las ramas denotan las acciones llevadas a cabo. Utilizando TrindiKit se han desarrollado sistemas de diálogo como Go- Dis [119] o EDIS [120].

## 4.6. ENFOQUES ESTADÍSTICOS

Los enfoques de aprendizaje automático para la gestión de diálogo tienen como principal objetivo reducir el esfuerzo y el tiempo necesarios para elaborar modelos de diálogo y, al mismo tiempo, facilitar el desarrollo de nuevos gestores de diálogo, su adaptación a distintos perfiles de usuario y a cambios en el dominio de aplicación. La idea principal es aprender estrategias óptimas utilizando métodos automatizados en lugar de confiar en los principios de diseño empírico [121].

### 4.6.1. *Aprendizaje reforzado (*Reinforcement Learning)

Las técnicas de aprendizaje reforzado (*reinforcement learning*) para la gestión de diálogo se fundamentan en explorar el conjunto de acciones que puede llevar a cabo el sistema, y determinar

la mejor selección de acciones, o estrategia del gestor de diálogo (*policy*), que optimizará el funcionamiento del sistema, representado mediante una función de utilidad (como, por ejemplo, la evaluación del sistema llevada a cabo por los usuarios). Para realizar esta optimización puede utilizarse un corpus de diálogos adquirido mediante la interacción de usuarios reales con el sistema o simular el comportamiento del usuario para obtener los datos de entrenamiento requeridos.

Para explorar las diferentes elecciones de acciones que puede realizar el sistema para cada estado del diálogo, es necesario definir una representación explícita de todos los estados del diálogo y de las alternativas que puede llevar a cabo el sistema en cada uno de ellos. Cada estado incorpora información que es relevante para realizar la elección de acciones, por ejemplo, las medidas de confianza de la información y la historia del diálogo. Dada la naturaleza de los algoritmos de aprendizaje que se utilizan, la información debe resumirse en función de un conjunto pequeño de características.

### 4.6.2. *Noción de recompensa*

El componente final del modelo es la recompensa (refuerzo, *reward*) asociada a cada estado. La recompensa hace referencia a las consecuencias inmediatas de ejecutar una acción en un determinado estado, determinándose la recompensa final acumulada en el diálogo teniendo en cuenta parámetros como el número de correcciones realizadas, el número de accesos a las bases de datos, los errores que hayan podido cometerse en el reconocimiento de la voz, la duración del diálogo, etc. La utilidad de tomar una decisión $a$ en el estado $s$ es la recompensa asignada a dicha acción más la suma de las recompensas acumuladas hasta el instante actual del diálogo, asumiendo que se ha utilizado la mejor estrategia de diálogo.

Puede consultarse numerosa bibliografía referente a la utilización de técnicas de aprendizaje reforzado para obtener la mejor estrategia de gestión de diálogo. Uno de los trabajos pioneros fue el de los investigadores de AT&T [122], en el que exponen que la gestión del diálogo puede analizarse como un problema de optimización de una función objetivo, representativa de las componentes relevantes de la tarea de diálogo considerada.

Bajo este planteamiento, el sistema de diálogo se implementa mediante un modelo estocástico, el MDP (Proceso de Decisión de Markov), que posibilita el aprendizaje automático de las estrategias de diálogo mediante la interacción con un usuario simulado.

Esta técnica se aplicó para obtener automáticamente una estrategia óptima para la tarea ATIS. El aprendizaje de los modelos se realiza mediante una combinación de aprendizaje reforzado (el sistema aprenderá la estrategia óptima en su interacción con los usuarios) y aprendizaje supervisado (modelos de simulación de usuarios). El diálogo puede formalizarse como un proceso de decisión secuencial por medio de un conjunto de acciones, un espacio de estados y una estrategia.

#### 4.6.3. *Proceso de decisión de Markov*

Los MDP sirven como representación formal del diálogo persona-máquina y sirven como base para formular los problemas del aprendizaje de una estrategia para gestionar el diálogo [123, 124, 125, 122]. Un MDP se describe formalmente mediante un espacio de estados finitos $S$, un conjunto finito de acciones $A$, un conjunto de probabilidades de transición $T$ y una función de recompensa $R$. En un instante de tiempo $t$, el gestor del diálogo se encuentra en un determinado estado $s_t \in S$, ejecuta una acción discreta $a_t \in A$, transita a un nuevo estado $s_{t+1}$ de acuerdo con la probabilidad $p(s_{t+1}|s_t, a_t)$ y recibe una recompensa $r_{t+1}$. La propiedad de Markov asegura que el estado y la recompensa en el instante $t+1$ dependen únicamente del estado y acción del instante $t$.

Modelando el diálogo utilizando MDPs, el gestor del diálogo puede entenderse como un agente que se desplaza a través de una red de estados de diálogos interconectados. Comenzando en un determinado estado inicial, el gestor va transitando entre estados tomando decisiones y recibiendo recompensas tras cada una de ellas. Dado que la respuesta del usuario a una determinada acción del sistema se desconoce, las transiciones son no-deterministas, es decir, la elección de una acción $a$ en un estado $s$ en el instante $t$ no conduce siempre al mismo estado y recompensa en el instante $t+1$.

Mediante los MDP, la estrategia del diálogo puede entenderse como el establecimiento de una correspondencia entre estados y acciones: para cada estado $s$, la estrategia selecciona la próxima acción del sistema $a$. De esta forma, la gestión del diálogo puede formalizarse como un problema matemático de optimización. La estrategia óptima es aquella que maximiza la recompensa acumulada en el tiempo, es decir, aquella en la que durante la evolución del diálogo se seleccionan las acciones con mayor recompensa asociada. Aunque puede conseguirse un comportamiento estocástico del modelo utilizando para ello una correspondencia probabilística para las acciones de cada uno de los estados, se suelen utilizar estrategias deterministas.

**Espacio de estados y conjunto de acciones.** La definición correcta del espacio de estados del diálogo $S$ y del conjunto de acciones del sistema $A$ es fundamental para el funcionamiento del modelo MDP. En tareas en las que el usuario debe aportar información para que el sistema responda a sus consultas (tareas *slot-filling*) la aproximación más utilizada consiste en definir un número de variables de estado en función del número de campos (atributos o *slots*) que debe completar el usuario para realizar las consultas permitidas.

En este caso, el número posible de estados viene determinado por la suma del número de estados en los que pueden encontrarse cada uno de los $slots$. Para que el número de estados no se haga inabordable, es habitual utilizar categorías para definir el estado del slot (como "desconocido", "conocido" y "confirmado"). Además, debe limitarse el conjunto de acciones del sistema. Una práctica común suele consistir en representar dichas acciones como tuplas consistentes en un acto de diálogo, el nombre del slot y un valor asociado al mismo (por ejemplo: $<$*confirmación implícita, ciudad destino, Londres*$>$).

**Determinación de la estrategia óptima de diálogo.** Las ecuaciones que definen el comportamiento del MDP son:

- La función de transición.

$$T(s', a, s) = P(s_{t+1} = s' \,|a_t = a, s_t = s) \tag{1}$$

- La matriz de la estrategia (*policy*) del diálogo.

$$\pi(s, a) = P(a_t = a \,|s_t = s) \tag{2}$$

- La recompensa esperada.

$$R(s', a, s) = \varepsilon(r_{t+1} \,|s_{t+1} = s', a_t = a, s_t = s) \tag{3}$$

La recompensa hace referencia al cumplimiento del objetivo del diálogo. Representando el diálogo como el recorrido por una secuencia de estados $s_0, s_1, ..., s_T$, la recompensa global del diálogo puede calcularse mediante la expresión:

$$R = \sum_{t=1}^{T} R(s_{t+1}, a_t, s_t) \tag{4}$$

y el objetivo del diálogo es obtener una función que maximice dicha expresión.

Sea $V$ una función que representa la recompensa esperada por transitar del estado $s$ al estado terminal $s_T$ dada la estrategia $\pi$, el cálculo de la estrategia óptima del diálogo puede realizarse de forma recursiva mediante la expresión

$$V^{\pi}(s) = \sum_{a} \pi(s,a) \sum_{s'} T(s',a,s) \left[ R(s',a,s) + V^{\pi}(s') \right] = \sum_{a} \pi(s,a) Q^{\pi}(s,a) \tag{5}$$

donde $Q(s,a)$ proporciona la recompensa asociada a seleccionar la acción $a$ en el estado $s$.

La estrategia del diálogo óptima es aquella que maximiza la función $V$ para todos los estados del espacio $S$.

$$V^{*}(s) = \max_{\pi} V^{\pi}(s') \quad \forall s \in S \tag{6}$$

La función V óptima puede obtenerse mediante:

$$V^{*}(s) = \max_{a} \sum_{s'} T(s',a,s) \left[ R(s',a,s) + V^{*}(s') \right] \tag{7}$$

De forma similar, la función $Q$ óptima puede hallarse resolviendo:

$$Q^{*}(s,a) = \sum_{s'} T(s',a,s) \left[ R(s',a,s) + \max_{a'} Q^{*}(s',a') \right] \tag{8}$$

y, finalmente, la estrategia óptima del diálogo viene dada por:

$$\pi^{*}(s) = \arg\max_{a} Q^{*}(s,a) \tag{9}$$

Si las probabilidades de transición y la función de recompensa se conocen a priori, la estrategia óptima ($\pi^{*}$) puede obtenerse mediante técnicas de programación dinámica (Iteración de Valores o Iteración de Estrategias).

**El algoritmo Q-Learning.** El algoritmo Q-Learning [126] es una de las soluciones más sencillas cuando estas funciones se desconocen. Se basa en utilizar valores de calidad (*Q-values*) para cada par $(s,a)$. Estos valores estiman la respuesta esperada tras seleccionar la acción $a$ en el estado $s$ siguiendo la estrategia $\pi$.

El proceso comienza inicializando arbitrariamente los valores de calidad para todos los pares estado-acción, que suelen disponerse en forma de matriz. Durante la interacción del sistema con el usuario (simulado), los valores de la matriz van actualizándose iterativamente para obtener una

mejor estimación de los mismos. Después de llevar a cabo una acción $a$ en el estado $s$, la respuesta del usuario origina una transición a un estado $s'$ y una recompensa $r$, recalculándose el valor de calidad utilizando:

$$Q(s,a) := (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a')) \tag{10}$$

donde $\alpha$ representa una tasa de aprendizaje entre 0 y 1; y $\gamma$ es un factor de descuento entre 0 y 1 que pondera las diferentes recompensas. Una vez se ha estimado el valor de calidad óptimo del conjunto de pares estado-acción (teóricamente tras visitar un número infinito de veces todas las posibles combinaciones de pares), la estrategia óptima viene dada por la Ecuación 9.

Una forma típica de actuar para resolver este problema consiste en utilizar una estrategia $\epsilon$-*greedy*: en cada iteración se selecciona un número aleatorio $0 < \beta < 1$, si $\beta < \epsilon$ la siguiente acción se selecciona aleatoriamente; en caso contrario, se selecciona la mejor acción. El valor de $\epsilon$ va disminuyéndose conforme aumentan el número de ciclos.

En [127] se aplica el MDP y el aprendizaje reforzado a un dominio de planificación de viajes. El principal objetivo del trabajo es describir una técnica para reducir el espacio de estados-acciones, con la finalidad de realizar un aprendizaje más rápido y reducir la carga computacional. La metodología propuesta se aplica para el aprendizaje de estrategias de diálogo con múltiples objetivos que optimicen las confirmaciones de uno o más datos.

**El algoritmo sapReduction.** Este algoritmo genera espacios reducidos con restricciones utilizando para ello tres niveles de reducción:

- Reducción de estados: El espacio de estados completo $S$ está conformado por todas las posibles combinaciones de slots $Q$ y variables de estado $V$, en las que se incluyen combinaciones válidas y no permitidas, que son las que este paso del algoritmo trata de eliminar. Para llevar a cabo esta reducción se tiene en cuenta que un slot terminal (donde típicamente se lleva a cabo la transacción) requiere que los slots no terminales hayan sido confirmados. Además, se tiene en cuenta qué slots no son necesarios para el objetivo del diálogo.
- Reducción de acciones: El espacio de acciones completo $A$ lo componen todas las combinaciones de slots $Q$ y de acciones simples $A^s$. Este paso del algoritmo tiene como objetivo eliminar las acciones que no son válidas para cada estado. Para ello se han definido cate-

gorías que representan las acciones válidas del sistema, agrupando las combinaciones de slots $Q$ y acciones $A^s$ permitidas.

- Reducción del espacio de búsqueda para cada partición $SA_i$: Para ello, se tratan de fijar fronteras entre los diversos objetivos del diálogo. En este paso se propone la generación de múltiples espacios de búsqueda de acuerdo con las particiones especificadas en el diseño del diálogo, equivalentes a los objetivos definidos. La finalidad es agrupar múltiples espacios de estados en un único espacio.

Los resultados presentados en el trabajo muestran que la reducción de espacios posibilita menores requerimientos de memoria (94 % de reducción), un aprendizaje más rápido (la convergencia se alcanza un 93 % más rápidamente) y mejor funcionamiento (8,4 % de reducción y un 7,7 % de mayor recompensa final en el modelo) frente a la utilización del espacio de estados sin aplicar el algoritmo de reducción. En la experimentación se han utilizado técnicas de simulación de usuarios.

#### 4.6.4. *Modelos Ocultos de Markov Parcialmente Observables: POMDP*

En [128] [129], investigadores de la Universidad de Cambridge consideran también la aproximación markoviana, el MDP aplicado al sistema de diálogo, y el uso del aprendizaje reforzado para que el gestor de diálogo aprenda la estrategia más adecuada. En particular, la cuestión objeto de su estudio es cómo crear el sistema inicial (*bootstrapping*). Posteriormente, en [130] [131] [132] [133], los mismos investigadores mejoran sus resultados mediante el uso de un MDP parcialmente observable (POMDP). En los resultados experimentales, el modelo POMDP supera a los MDP tradicionales. Además, el método descrito permite mejorar automáticamente el funcionamiento de gestores de diálogo sencillos (diseñados a mano). Formalmente un POMDP puede definirse como una tupla $\{S, A, T, R, O, Z, \lambda, b_0\}$ donde:

- $S$ es el conjunto de estados que definen el comportamiento del agente.
- $A$ es el conjunto de acciones que puede tomar el agente.
- $T$ define una probabilidad de transición $P(s'|s, a)$.
- $R$ define la recompensa esperada $r(s, a)$.
- $O$ es el conjunto de observaciones que el agente puede recibir sobre el mundo.

- Z define la probabilidad de observación $P(o'|s', a)$.
- $\lambda$ es un factor de descuento geométrico $0 \leq \lambda \leq 1$.
- $b_0$ es el estado inicial $b_0(s)$.

Un POMDP funciona del siguiente modo. En cada instante, el sistema se encuentra en un estado no observado $s$. El sistema selecciona una acción $a_m$, recibe una recompensa $r$ y transita a un estado (no observado) $s'$, donde $s'$ depende únicamente de $s$ y $a_m$. El sistema recibe una observación $o'$ que depende de $s$ y $a_m$. Aunque la observación permite que el sistema disponga de alguna evidencia sobre el estado $s$ en el que se encuentra, $s$ no se conoce con exactitud, definiéndose una probabilidad $b(s)$ (*belief state*) que indica la probabilidad de que el sistema se encuentre en un determinado estado s. En cada instante, esta probabilidad se actualiza basándose en $o'$ y $a_m$:

$$b'(s') = P(s' \left| o', a_m, b\right) = k \cdot P(o' \left| s', a_m\right) \sum_{s \in S} P(s' \left| a_m, s\right) b(s) \tag{11}$$

siendo $k = P(o'|a_m, b)$ un factor de normalización. En cada instante $t$ el sistema recibe una recompensa $r(b_t, a_{m,t})$ dependiente del estado $b_t$ y la acción seleccionada $a_{m,t}$. La recompensa acumulada durante el diálogo (*return*) puede calcularse mediante:

$$R = \sum_{t=0}^{\infty} \lambda^t R(b_t, a_{m,t}) = \sum_{t=0}^{\infty} \lambda^t \sum_{s \in S} b_t(s) r(s, a_{m,t}) \tag{12}$$

Cada acción $a_{m,t}$ viene determinada por la estrategia $\pi(b_t)$ y la construcción del modelo POMDP implica encontrar la estrategia $\pi^*$ que maximiza el *return*. Una resolución exacta de este problema es intratable [134] [135], aunque pueden obtenerse soluciones aproximadas que proporcionan estrategias de diálogo útiles. La solución más simple consiste en discretizar el espacio de estados y utilizar los métodos de optimización descritos para el MDP [136].

**Modelado de la gestión del diálogo basado en POMDP.** En [137] se describe como modelar la gestión del diálogo mediante la utilización de POMDP. A esta aproximación se le conoce como SDS-POMDP. En ella, la variable de estado $s \in S$ se descompone en tres términos:

$$s = (s_u, a_u, s_d) \tag{13}$$

- el objetivo del usuario, $s_u \in S_u$;
- la acción del usuario, $a_u \in A_u$;
- la historia del diálogo, $s_d \in S_d$.

Para factorizar el modelo, la función de transición del modelo se descompone de la siguiente forma:

$$P(s' \,|s, a_m) = P(s'_u \,|s_u, s_d, a_u, a_m)\; P(a'_u \,|s'_u, s_u, s_d, a_u, a_m)\; P(s'_d \,|a'_u, s'_u, s_u, s_d, a_u, a_m) \tag{14}$$

**Modelo de objetivo del usuario.** Al primer término se le denomina modelo de objetivo del usuario (*user goal model*) e indica cómo varía el objetivo del usuario en cada instante del diálogo. Se asume que el objetivo del usuario en un determinado instante depende únicamente del objetivo actual y de las acciones llevadas a cabo por el sistema hasta dicho instante:

$$P(s'_u \,|s_u, s_d, a_u, a_m) = P(s'_u \,|s_u, a_m) \tag{15}$$

**Modelo de acción del usuario.** El segundo término, denominado modelo de acción del usuario (*user action model*) indica qué acciones del usuario son más probables para cada instante. Se realiza la misma suposición que para el primer término:

$$P(a'_u \,|s'_u, s_u, s_d, a_u, a_m) = P(a'_u \,|s'_u, a_m) \tag{16}$$

**Modelo de la historia del diálogo.** El tercer término, modelo de historia del diálogo (*dialogue history model*) indica cómo han afectado las acciones del usuario y el sistema a la historia del diálogo:

$$P(s'_d \,|a'_u, s'_u, s_u, s_d, a_u, a_m) = P(s'_d \,|a'_u, s_d, a_m) \tag{17}$$

Por tanto, la función de transición puede reescribirse de la siguiente forma

$$P(s' \,|s, a_m) = P(s'_u \,|s_u, a_m)\; P(a'_u \,|s'_u, a_m)\; P(s'_d \,|a'_u, s_d, a_m) \tag{18}$$

Con esta representación se reduce el número de parámetros requeridos para la función de transición y permite la estimación de los diferentes parámetros por separado.

**Estrategia de descomposición de la observación.** La observación $o$ del POMDP se puede descomponer en dos términos: las hipótesis del reconocedor $\tilde{a}_u \in A_u$ y las medidas de confianza $c \in R$. La función de observación viene dada por:

$$P(o' \,|s', a_m) = P(\tilde{a}_u, c' \,|s'_u, s'_d, a'_u, a_m) \tag{19}$$

Asumiendo que la observación depende únicamente de la acción llevada a cabo por el usuario y por la gramática $g$ seleccionada por el gestor de diálogo:

$$P(\tilde{a}_u, c' \,|s'_u, s'_d, a'_u, a_m) = P(\tilde{a}_u, c' \,|a'_u, g) \tag{20}$$

que representa la distribución de probabilidad de "observar" las hipótesis del reconocedor $\tilde{a}_u$ con medida de confianza $c$ cuando el usuario realmente seleccionó la acción $a_u$ y se activó la gramática $g$. Por último, utilizando las expresiones anteriores, la probabilidad $b(s)$ puede escribirse de la forma siguiente:

$$\begin{aligned} b'(s'_u, s'_d, a'_u) = k \cdot P(\tilde{a}'_u, c'|a'_u, g) \cdot P(a'_u \,|s'_u, a_m) \cdot \sum_{S_u} P(s'_u \,|s_u, a_m) \cdot \\ \cdot \sum_{S_d} P(s'_d \,|a'_u, s'_u, s_d, a_m) \cdot \sum_{a_u} b(s_u, s_d, a_u) \end{aligned} \tag{21}$$

#### 4.6.5. *Optimización para POMDP*

En [138, 139, 137] se resume una solución al problema del gran número de estados que pueden presentarse en los sistemas de diálogo reales cuando se modelan mediante los POMDPs, que hace que la estimación de la estrategia óptima sea intratable.

La metodología consiste en centrarse durante el proceso de estimación de la estrategia únicamente en los estados más probables para la acción actual del sistema. De este modo, se definen dos espacios de estados acoplados: el espacio total de estados (*master state space*) y un espacio de estados resumido (*summary state space*) mucho más simple. El espacio de estados resumido contiene los N mejores estados de objetivos ($s_u$) del espacio total (N es normalmente 1 ó 2) y una codificación simplificada de la acción del usuario $a_u$ y de la historia del diálogo $s_d$.

La optimización de la estrategia se lleva a cabo utilizando estos dos espacios de estados acoplados, utilizando para ello técnicas basadas en el modelo (por ejemplo, PBVI, Point-based Value Iteration [140]) o aproximaciones como *Q-learning* conjuntamente con simuladores de usuarios.

#### 4.6.6. *HIS: Hidden Information State*

En [131] [141] [142] se introduce el modelo de diálogo HIS (*Hidden Information State*), una simplificación específica del modelado del diálogo mediante POMDPs para facilitar su uso en aplicaciones reales, dado que el tamaño del espacio de estados necesario para representar los sistemas de diálogo del mundo real hace que la implementación del modelo SDS-POMDP sea intratable.

La idea principal del modelo HIS consiste en particionar el espacio de estados asumiendo que en un instante $t$, el espacio de todos los objetivos del usuario $S_u$ puede agruparse en clases de equivalencia (particiones) donde todos los miembros de la clase están enlazados y no son distinguibles. De este modo, las probabilidades se calculan para cada una de las particiones y no para cada uno de los estados que las componen.

La forma de actuar es la siguiente. Inicialmente todos los estados $s_u \in S_u$ se sitúan en una única partición $p_0$. A medida que los diálogos progresan, esta partición raíz se va dividiendo repetidamente en particiones más pequeñas de modo binario. Cada vez que se genera una nueva partición

$$p \rightarrow \{p', p - p'\} \tag{22}$$

la probabilidad del modelo puede recalcularse de la siguiente forma:

$$b(p') = P(p'\,|p)\; b(p) \quad y \quad b(p - p') = \big(1 - P(p'\,|p)\big)\; b(p) \tag{23}$$

Considerando que la información proporcionada por el usuario viene representada implícitamente por las particiones realizadas, puede asumirse:

$$P(a'_u\,|s'_u, a_m\,) = P(a'_u\,|p', a_m) \tag{24}$$

$$P(s'_d\,|s'_u, a'_u, s_d, a_m) = P(s'_d\,|p', a'_u, s_d, a_m) \tag{25}$$

y, de este modo, la probabilidad final del modelo HIS se expresa mediante la siguiente ecuación:

$$b'(p', a'_u, s'_d) = k \cdot P(o' \,|a'_u) P(a'_u \,|p', a_m) \sum_{s_d} P(s'_d \,|p', a'_u, s_d, a_m) P(p' \,|p) b(p, s_d) \quad (26)$$

La ecuación consta de cuatro distribuciones de probabilidad:

- Modelo de observación: Se aproxima por la probabilidad de la mejor opción del módulo de comprensión:

$$P(o' \,|a'_u) \approx k' \cdot P(a'_u \,|o) \quad (27)$$

- Modelo de acto de diálogo de usuario. Se compone de dos partes, la probabilidad del bigrama del tipo de acto de diálogo de usuario actual dado el tipo de acto de diálogo de sistema que le ha precedido, y una probabilidad que denota el grado de consistencia del acto de diálogo de usuario con la partición ($p'$).

$$P(a'_u \,|p', a_m) \approx P(\tau(a'_u) \,|\tau(a'_m)) P(\mathcal{M}(a'_u) \,|p') \quad (28)$$

  donde $\tau(a)$ hace referencia al tipo de acto de diálogo y $\mathcal{M}(a)$ denota la correspondencia o no del acto de diálogo $a$ con la partición actual $p'$.

- Modelo de diálogo: Es completamente heurístico.

$$\begin{aligned} P(s'_d \,|p', a'_u, s_d, a_m) &= 1 \quad si\ s'_d\ es\ consistente\ con\ p', a'_u, s_d, a_m \\ &= 0 \quad resto\ de\ los\ casos \end{aligned} \quad (29)$$

- Refinamiento del espacio: Depende de las reglas de ontología que definen el dominio de la aplicación. Los objetivos del usuario se construyen utilizando reglas con probabilidades establecidas a priori. La probabilidad correspondiente a utilizar la secuencia de reglas $r_1, r_2, \ldots, r_k$ para dividir la partición $p$ en la sub-partición $p'$ se define de la siguiente forma:

$$P(p' \,|p) = \prod_{i=1}^{k} P(r_i) \quad (30)$$

Para implementar en la práctica el modelo HIS, el espacio de los objetivos del usuario se representa mediante una estructura de árbol jerárquico, donde un conjunto de reglas de ontología describe la estructura jerárquica de la información y los valores específicos que pueden asignarse a los nodos terminales. Cada uno de los nodos no terminales dispone de probabilidades asociadas

a establecer particiones en cada uno de ellos $P(p'|p)$. Las particiones en el espacio de objetivos de usuario se representan mediante un bosque de árboles, representando cada uno de ellos una única partición. Al comienzo del diálogo existe una única partición con un único nodo con probabilidad uno asociada.

Cada uno de los actos de diálogo de usuario entrante se intenta emparejar con alguna de las particiones disponible en el turno actual. En caso de no encontrarse el emparejamiento, se consultan las reglas de la ontología y el sistema intenta buscar el emparejamiento expandiendo el árbol actual.

La técnica de optimización para realizar la búsqueda de la estrategia óptima suele basarse en utilizar Q-learning conjuntamente con un simulador de usuario externo [141]. Esta metodología se ha aplicado para la optimización de la estrategia del diálogo en una tarea de información turística. El aprendizaje se lleva a cabo iterando mediante el algoritmo Monte Carlo, manteniendo constante la estrategia actual durante 5000 diálogos, actualizándose de forma continua los valores de confianza asociados a cada uno de los espacios de estados resumidos.

### 4.6.7. *HAM: Hierarchical Abstract Machines*

En [143] se describe una aproximación, conocida como *Hierarchical Abstract Machines* (HAMs), para la obtención de la estrategia óptima del diálogo basada en la combinación de máquinas de estados finitos y MDPs. Las ventajas mencionadas para esta aproximación son:

- Las estrategias del diálogo se especifican de forma que el desarrollador del sistema decide qué aspectos dejar a mano y qué optimizar.
- Posibilita un aprendizaje más rápido, dado que se incorpora conocimiento del dominio para reducir el número de estados del espacio.
- Transferencia de conocimiento, dado que las HAMs pueden reutilizarse.

Una HAM puede entenderse como un MDP en el que se restringe las acciones que pueden tomarse en cada estado durante el aprendizaje reforzado. Son similares a las máquinas de estados finitos no deterministas: se parte de un estado inicial y se recorren los estados hasta alcanzar un estado de parada, en el que se devuelve el control al usuario.

Formalmente, una HAM es una colección de tres tuplas $H_i = (\mu, I, \delta)$, donde $\mu$ es un conjunto finito de máquinas de estados, $I$ es el estado inicial y $\delta$ es la función de transición que determina el siguiente estado mediante transiciones deterministas o estocásticas.

#### 4.6.8. *Modelado del diálogo mediante un proceso de clasificación*

En [144] se describe una aproximación estadística para la gestión de diálogo basada en representar la historia previa del diálogo mediante un registro de diálogo, cuyos contenidos puede ir completando el usuario a través de los actos de diálogo que proporcione como respuestas a los turnos del sistema. Los contenidos de este registro se codifican en cada uno de sus campos en términos de tres valores, 0; 1; 2, de acuerdo con el siguiente criterio:

- 0: El usuario no ha realizado una consulta sobre el concepto o no ha proporcionado el valor del atributo correspondiente.
- 1: El concepto o atributo está presente con una medida de confianza superior a un umbral prefijado (un valor entre 0 y 1). Las medidas de confianza se generan durante los procesos de reconocimiento y comprensión.
- 2: El concepto o atributo está presente, pero con una medida de confianza inferior al umbral.

A la hora de implementar un gestor de diálogo que lleve a la práctica el modelo descrito, se realiza una división del modelo estadístico en dos procesos diferenciados. Un primer proceso se encarga de realizar la actualización del registro del diálogo a partir de la información suministrada por el módulo de comprensión tras la intervención del usuario.

Este proceso se realiza de forma automática, incorporándose al registro toda aquella información relevante para la tarea que proporcione comprensión y variando las confianzas almacenadas junto a los valores de los conceptos y atributos una vez el usuario los confirme positivamente. Este último caso implica un cambio en el registro de diálogo distinto a completar su contenido. Por ello, se separan de la representación semántica del turno de usuario los actos de diálogo independientes de la tarea (por ejemplo, Afirmación, Negación y No-Entendido).

Un segundo proceso se encarga de obtener las probabilidades asignadas a cada una de las posibles acciones definidas para el gestor de diálogo considerando el estado actual del registro del diálogo

(una vez se ha actualizado con la información suministrada por el usuario en el turno anterior), los conceptos independientes de la tarea y la última respuesta generada por el sistema. Para ello, se utiliza un proceso de clasificación que toma como entrada la variable definida para codificar el estado actual del diálogo y determina así la clase (respuesta del sistema) más probable para la situación actual del diálogo, independientemente de que se trate de una situación vista o no vista.

El gestor de diálogo opera de acuerdo con el siguiente algoritmo:

- Recibe la representación semántica de la entrada del usuario generada por el módulo de comprensión.
- Actualiza el registro del diálogo de acuerdo con la información anterior.
- Realiza la codificación de la información suministrada por comprensión y del registro de diálogo, siguiendo el formato y orden definido para cada una de las estructuras.
- Genera la secuencia resultante de unir la codificación de la última respuesta del sistema, el contenido del registro de diálogo de acuerdo con la codificación definida, y la información independiente de la tarea aportada por el usuario en su última intervención.
- Clasifica esta secuencia para seleccionar como salida del proceso de clasificación una de las posibles acciones definidas para el gestor de diálogo.

La función de clasificación puede definirse de acuerdo con los requisitos del dominio de aplicación del interfaz conversacional. En [144] se propone el uso de redes neuronales artificiales. En [145] se describe una extensión del modelo que permite modelar estadísticamente el proceso de gestión de forma completa en dominios que conlleven una gran variedad de respuestas del sistema tras realizar una consulta a los repositorios de datos del sistema.

Más recientemente, en [146], se propone el uso de clasificadores evolutivos para la definición de la función de clasificación. Una de las aportaciones más importantes de esta aproximación para la gestión estadística del diálogo, es la posibilidad de ampliar fácilmente el registro del diálogo para tener en cuenta características adicionales a los actos de diálogo del usuario y del sistema (por ejemplo, para considerar información del contexto de la interacción capturada automáticamente por sensores [147], modelos de usuario [148] o información relativa al estado emocional de los usuarios [149]).

#### 4.6.9. *Modelado del diálogo mediante redes bayesianas*

Otra aproximación estadística alternativa al modelado del diálogo mediante Procesos de Decisión de Markov y aprendizaje reforzado, consiste en la utilización de redes bayesianas (*Belief Networks*, BN) para representar las interacciones del diálogo. En [150] se citan las principales ventajas de la utilización de las BN para el modelado de la iniciativa mixta del diálogo:

- Las probabilidades de la BN pueden estimarse automáticamente a partir de un conjunto de datos disponible, lo que facilita la portabilidad y escalabilidad a otros dominios. El objetivo del usuario puede identificarse mediante inferencia probabilística.
- La topología de la BN puede aprenderse también a partir de un corpus de datos de entrenamiento. La topología puede capturar las dependencias entre los nodos de la BN, representando cada uno de estos nodos un concepto semántico del dominio del sistema
- La probabilidad de propagación en una red BN se corresponde con el cómputo de probabilidades de los eventos que pueden producirse. Mediante este procedimiento pueden detectarse qué conceptos restan por solicitar al usuario o cuáles de ellos deben confirmarse previamente a la consulta.
- La topología y las probabilidades de la BN puede aprenderse de forma automática, definirse a mano o utilizar ambas aproximaciones.

**Aplicación de las redes bayesianas a la comprensión del lenguaje natural.** Las redes bayesianas se han utilizado tradicionalmente en el marco de la comprensión del lenguaje natural. En este contexto, existe un conjunto finito de conceptos semánticos (M) y de objetivos de usuario (N). Los objetivos $G_i$ y los conceptos $C_j$ son binarios, siendo ciertos si aparecen en el turno correspondiente. Por tanto, el problema de la comprensión puede verse como llevar a cabo $N$ decisiones binarias con $N$ BNs (cada una para uno de los objetivos del usuario). La BN para el objetivo $G_i$ tiene como entrada un conjunto de conceptos semánticos $C$ extraídos de la intervención del usuario. La red proporciona la probabilidad a posteriori $P(G|C)$, a partir de la que se toma una decisión binaria mediante su comparación con un umbral. La expresión de esta probabilidad coincide con la formulación de Naive Bayes:

$$P(G_i = 1|\vec{C}) = \frac{P(\vec{C}|G_i = 1)P(G_i = 1)}{P(\vec{C})} \quad (31)$$

Se asume que el objetivo $G_i$ está presente si la probabilidad $P(G_i|C)$ es mayor que un umbral $\theta$. Utilizando esta formulación, las consultas del usuario que no superen el umbral de ninguna de las redes se suponen fuera del dominio del sistema.

**Modelado de diálogo mediante backward inference.** La idea principal es utilizar las redes bayesianas para detectar conceptos automáticamente de acuerdo a las restricciones del dominio capturadas por sus probabilidades. La detección automática de conceptos se lleva a cabo mediante la técnica de *backward inference*. Una vez inferido el objetivo del diálogo ($G_i$) en una consulta determinada del usuario, el nodo objetivo de la BN correspondiente se instancia (a 0 ó 1) para examinar la confianza de la red para cada uno de los conceptos de entrada. Si la topología de la red BN asume la independencia condicional entre los conceptos, su probabilidad actualizada es simplemente $P(C_j|G)$. Si existe dependencia entre los diferentes conceptos, la probabilidad actualizada del objetivo ($P^*(C_i)$) se calcula mediante:

$$P^*(\vec{C}|G_i) = P(\vec{C}|G_i)P^*(G_i) = P(\vec{C}, G_i)\frac{P^*(G_i)}{P(G_i)} \tag{32}$$

donde $P^*(G_i)$ se actualiza instanciando el nodo objetivo, $P(\vec{C}, G_i)$ se obtiene a partir del corpus de entrenamiento.

Basándose en el valor de $P^*(C_j)$, se decide mediante la utilización del umbral si $C_j$ está presente o no. Esta decisión se compara con la ocurrencia real de $C_j$ en la intervención del usuario. Si la decisión indica que $C_j$ no debería estar presente, pero está realmente en dicha intervención, el concepto se marca como espurio y el modelo de diálogo solicitará su confirmación. Si la decisión binaria indica que $C_j$ debería estar presente pero no es así en la consulta del usuario, el concepto se etiqueta como requerido y el modelo de diálogo lo solicitará al usuario.

- En [150] se aplican las redes bayesianas para el desarrollo de un modelo de diálogo para el sistema CU FOREX [151]. El dominio de este sistema es proporcionar información referente a dos tipos de consultas bancarias: cambios de moneda y tipos de intereses. Para este dominio se han definido cinco conceptos. Cada una de las dos BN definidas (una para cada consulta del sistema), recibe como entrada estos cinco conceptos. La topología de las redes se entrenó automáticamente a partir de un corpus. La evaluación del modelo se llevó a cabo a partir de 550 diálogos, obteniéndose porcentajes de éxito de la tarea del 96 %.

  En este trabajo se realiza también la aplicación de las BN para el dominio ATIS. A partir de

un corpus de datos extraído de ATIS-3, se implementaron 11 BN (una para cada objetivo detectado) y se definieron un total de 60 etiquetas semánticas (conceptos y atributos necesarios para realizar las consultas en la base de datos). Para solucionar la redundancia del modelo de diálogo (debida a la no detección de las equivalencias entre atributos de la tarea) se propuso la definición de dos umbrales y el refinamiento a mano de las probabilidades.

- Quartet es una plataforma desarrollada por Microsoft Research para implementar sistemas de diálogo multimodales. La incertidumbre del dialogo se representa mediante redes bayesianas. Se han definido cuatro niveles de representación que soportan inferencia y toma de decisiones (canal, señal, intención y conversación).

# 5. MULTIMODALIDAD, MULTILINGÜISMO, EMOCIONES Y SISTEMAS CONVERSACIONALES AFECTIVOS

La interacción conversacional dialogada entre un usuario y un sistema informático comprende dos componentes básicos: por un lado, el propio lenguaje natural o humano utilizado para la interacción, y en segundo lugar la modalidad de interacción, que tal y como se ha indicado puede ser la expresión hablada o escrita, tanto en la fase de recepción (reconocimiento de voz y comprensión del lenguaje) como en la de emisión (generación de lenguaje y síntesis de voz). Ahora bien, dependiendo de la naturaleza de la interacción persona-máquina que pretendamos modelar y mediante un sistema conversacional, será adecuado o necesario utilizar más de un lenguaje y más de una modalidad de interacción, tanto en la fase de entrada como en la de salida del sistema.

Esto nos lleva a abordar dos características relevantes en los sistemas conversacionales como son la multimodalidad y el multilingüismo.

## 5.1. MULTIMODALIDAD

Aunque se han propuesto distintas definiciones para la noción de multimodalidad, un punto de partida simple y a la vez genérico asocia la multimodalidad con el uso de dos o más sentidos para el intercambio de información.

De esta forma, un sistema conversacional multimodal permitiría al usuario interactuar con más de una modalidad, por ejemplo, el usuario podría introducir información en el sistema escribiendo

mensajes con el teclado, hablándole al sistema o utilizando una pantalla táctil. O visto desde el punto de vista del sistema conversacional, este sistema podría incorporar información capturando la señal acústica que le llega del usuario, mensajes escritos, reconocimiento facial y gestual, eventos capturados en dispositivos táctiles, etc.

De forma simétrica, en el lado de la salida del sistema conversacional, un entorno multimodal podrá utilizar simultáneamente un sintetizador de voz para emitir un mensaje, mostrar ese mismo mensaje o versiones enriquecidas del mismo mediante gráficos, tablas, etc., mediante distintos tipos de pantallas, usar distintos tipos de avatares con capacidad de modelado facial y gestual, etc.

La proliferación de múltiples tipos de dispositivos que permiten la interacción hombre-máquina ha motivado el interés por el desarrollo de sistemas conversacionales con capacidad multimodal. Pensemos por ejemplo en una aplicación móvil. A través de la misma podemos hablar y recibir mensajes acústicos, escribir y leer textos, el sistema puede mostrar imágenes, mapas, rutas sobre los mismos, generar gráficos y animaciones. Podemos así mismo utilizar la capacidad táctil de la pantalla para apuntar objetos o regiones sobre el dispositivo. El mismo dispositivo puede vibrar, etc. En definitiva, los dispositivos ponen a disposición de los entornos conversacionales una amplia diversidad y riqueza de modalidades.

El siguiente reto que se debe abordar es la manipulación, captura y representación de estos eventos multimodales, y en particular su sincronización. Este problema se ha ilustrado habitualmente con expresiones del tipo *Coloca esto aquí* en el que un usuario a la vez que va pronunciando esta frase va tocando distintas zonas de la pantalla o monitor.

Pero así mismo, podríamos entender la multimodalidad como un rasgo que permite enriquecer y mejorar el propio diseño y comportamiento de un sistema conversacional. Si en lugar de restringir la entrada del sistema conversacional al texto (para sistemas conversacionales escritos) o la señal acústica (para sistemas conversacionales hablados utilizando un módulo de Reconocimiento de Voz), el sistema puede utilizar información paralingüística que incluya por ejemplo la detección de emociones, conseguiremos sistemas conversacionales que se adaptarán de una forma más efectiva a las necesidades del usuario y las características del contexto conversacional.

## 5.2. MULTILINGÜISMO

Normalmente un sistema conversacional establecerá la comunicación con un usuario utilizando un único lenguaje. Ahora bien, cada vez es más habitual encontrar proyectos que requieren el uso de sistemas conversacionales para hablantes en distintos idiomas. De esta forma se plantea un requisito funcional especialmente relevante: ¿cómo construir sistemas conversacionales multilingües? Ante este objetivo, una solución que podemos calificar como ingenua se plantearía el diseño de distintos sistemas, cada uno de ellos para uno de los lenguajes que se deben soportar. Esta solución, además de ser poco operativa desde el punto de vista del diseño y reusabilidad de componentes, aumenta considerablemente el coste, los recursos y el tiempo necesario para llevar a cabo la implementación.

Por tanto, surge la necesidad de abordar estrategias que soporten una capacidad de interacción multilingüe: es decir, un único sistema conversacional que pueda operar con distintos lenguajes en las fases de Comprensión y Generación.

### 5.2.1. *Idiomas con pocos recursos*

Como se introdujo en la sección 3.1, los últimos avances en el uso de aprendizaje profundo para el reconocimiento del habla han permitido una mejora en el rendimiento de estos sistemas que ha favorecido la adopción de las aplicaciones con interfaz oral.

Sin embargo, estas ventajas se derivan del entrenamiento de redes sobre ingentes cantidades de datos. Esto implica que sólo los idiomas para los que existe un gran volumen de recursos pueden beneficiarse de las altas tasas de acierto que se obtienen con el uso de redes neuronales profundas.

Precisamente la compilación y preparación de los materiales necesarios para el entrenamiento de los modelos acústicos (grabaciones de audio en el idioma objetivo) y lingüísticos (textos en el idioma objetivo) es uno de los procesos más caros en cuanto al tiempo y esfuerzo necesarios. En el caso de idiomas con pocos hablantes, ese esfuerzo no se ve suficientemente recompensado. Por esto, han surgido diversas iniciativas que tratan de paliar este problema centrándose en el desarrollo rápido de tecnologías del habla para idiomas con pocos recursos.

Uno de los primeros enfoques planteados fue crear un registro fonético universal con corpus de

todos los idiomas de los que se disponga. Después, para cada idioma, se hace un mapeo entre su registro fonético y el registro universal.

Se trata de una idea parecida al alfabeto fonético universal, donde cada sonido se representa con una grafía y las palabras se representan usando las grafías correspondientes a los sonidos con los que se pronuncian[16]. Este enfoque se explica con detalle en el libro de referencia [152].

Actualmente, los enfoques más extendidos son aquellos que reutilizan elementos a través de diversos idiomas (cross-language) [153]. Se trata de aprovechar los recursos disponibles para un idioma con más recursos con la finalidad de reconocer otro idioma con menos recursos. En los últimos años se han desarrollado técnicas cada vez más sofisticadas para reutilizar recursos. Por ejemplo en [154] se explica un sistema para desarrollar de forma rápida y casi automática un reconocedor para cuatro lenguas eslavas basado en los recursos para el checo.

### 5.2.2. *Traducción automática y uso multilingüe*

Actualmente el enfoque más usado en traducción automática está basado en el uso de los denominados "word embeddings". Usualmente, para entrenar un traductor automático se necesitaban grandes cantidades de textos "paralelos", es decir, el mismo texto en el idioma inicial y en el idioma objetivo. Los *embeddings* no consideran a las palabras de forma aislada sino como vectores con gran cantidad de información, entre ella semántica y de relación con otros términos. Estas relaciones generan espacios parecidos entre idiomas, de forma que los *embeddings* correspondientes a distintos idiomas se pueden alinear y construir espacios multilingües sin la necesidad de textos paralelos.

### 5.2.3. *Sistemas conversacionales afectivos*

La emoción juega un papel clave en la interacción entre los seres humanos. Por esta razón, la comunicación con un sistema conversacional debería ser más efectiva si el sistema puede procesar y comprender las emociones de sus usuarios, así como mostrar sus propias emociones. Piccard [155] acuñó el término computación afectiva en un momento en que la emoción no se consideraba un aspecto relevante del diseño de sistemas artificiales.

[16] *Se pueden escuchar los sonidos del alfabeto fonético universal en: http://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/*

Actualmente existe una comunidad científica muy activa en el campo de la computación afectiva, con varias conferencias y revistas internacionales dedicadas a esta temática. Sus trabajos han demostrado los amplios beneficios que la emoción puede aportar a los sistemas artificiales, incluido el aumento de la facilidad de uso, la eficiencia, confianza, eficacia y mejora en la comunicación. La computación afectiva también se utiliza en aplicaciones comerciales como el análisis de sentimientos, la minería de opinión para marketing y la creación de marca.

La computación afectiva es interdisciplinar, involucrando investigación en los campos de las ciencias de la computación e ingeniería, ciencias cognitivas, neurociencia y psicología. Cada uno de estos campos tiene su propia terminología y, a veces, dentro del mismo campo hay una falta de consenso sobre el significado de conceptos fundamentales y cómo se pueden representar y formalizar. Por ejemplo, se utilizan términos como emoción, sentimiento, afecto, estados afectivos, estado de ánimo y posturas sociales e interpersonales. Para una revisión de diferentes tradiciones de investigación y teorías de las emociones, se recomienda [156, 1].

Los procesos involucrados en la construcción de un reconocedor de emociones engloban las siguientes fases fundamentales: recopilación y anotación de datos, aprendizaje y optimización. Durante la fase de recogida de datos, se graban diferentes señales del usuario y se procesan previamente para eliminar el ruido y otros fenómenos que pueden degradarlas. Las decisiones fundamentales en esta fase son qué características del usuario utilizar para evaluar su estado emocional y cómo adquirirlas. El reconocimiento de emociones se puede realizar usando cualquiera de las modalidades de entrada de la interfaz conversacional (por ejemplo, detectando emociones en la voz o expresión facial del usuario) o utilizando una combinación de ellas. También puede considerarse el efecto que la interacción con el interfaz puede tener en sus respuestas emocionales.

Las características que suelen considerarse no son solo valores en bruto calculados a partir de la señal de entrada (por ejemplo, la frecuencia cardíaca o el volumen de voz), sino también medidas estadísticas (por ejemplo, la variación de la frecuencia cardíaca y el volumen promedio). Se suele utilizar también un proceso de clasificación para obtener características significativas para ingresar al reconocedor de emociones. Shuller y Batliner [157] presentan una discusión de las ventajas y desventajas de diferentes tipos de características.

Una vez que se ha obtenido una base de datos con todas las unidades de reconocimiento representadas como vectores de características, la base de datos debe anotarse para asignar una emoción a cada una de las unidades. El procedimiento de anotación depende de cómo se recopi-

laron los datos (emociones actuadas, emociones provocadas o emociones espontáneas).

Las emociones actuadas por profesionales pueden ser apropiadas para algunas modalidades, aunque se pierden algunas de las sutilezas de las respuestas emocionales que no pueden producirse conscientemente. Para otros tipos de señales, como los datos fisiológicos, las bases de datos de emociones actuadas no son adecuadas. Con respecto a las emociones provocadas, es importante evitar inducir emociones diferentes de la emoción objetivo y eliminar las posibilidades de inducir varias emociones [158]. Las emociones espontáneas son las más naturales, pero exigen un complejo proceso de anotación de emociones para obtener una base de datos confiable. Además, pueden tener el inconveniente de que no todas las emociones son frecuentes en todos los dominios de aplicación y, por lo general, las bases de datos de emociones espontáneas están desbalanceadas [159].

Una vez anotada, la base de datos emocional se usa para entrenar un algoritmo de reconocimiento de patrones que a partir de un vector de características genere una hipótesis de clasificación para una determinada emoción. Se han comparado diferentes algoritmos para verificar su idoneidad para esta tarea teniendo en cuenta las señales utilizadas para reconocer el estado emocional de los usuarios: señales fisiológicas (sistema cardiovascular, actividad electrodérmica, sistema respiratorio, sistema muscular, actividad cerebral, etc.), señal de voz (características paralingüísticas e información afectiva extraída del texto resultante de la transcripción de voz a texto), expresiones faciales y gestos, etc.

La síntesis de emociones se basa en gran medida en las mismas características descritas para el reconocimiento de emociones. Existe un extenso número de trabajos que muestran que los seres humanos asignamos características humanas a los interlocutores artificiales. Especialmente relevantes son los resultados experimentales logrados por Nass, que muestra que a menudo se asigna contenido emocional y social cuando el interlocutor es un sistema conversacional [160, 161].

Los agentes conversacionales personificados (*embodied conversational agent*, ECA) pueden mostrar expresiones faciales, gestos y voces con contenidos emocionales. Los comportamientos expresivos para estos agentes implican no solo elegir las características apropiadas, sino también decidir cómo se van a aplicar como respuesta a las acciones de los usuarios.

La personalidad puede definirse como las características de una persona que influyen de manera única en sus cogniciones, motivaciones y comportamientos en diferentes situaciones. Por lo

tanto, la personalidad es un aspecto importante que considerar para entender el comportamiento de los usuarios. Además, varios estudios han demostrado que los usuarios a menudo asignan también personalidad a los sistemas sintéticos [160]. Por lo tanto, hay un gran interés en encontrar modelos apropiados para brindar personalidades consistentes que permitan interacciones creíbles con los interfaces conversacionales [162].

# 6. PROTOCOLOS Y ESTÁNDARES

Challenges to Adoption of Standards in Conversational Systems

**Contribución de: Deborah Dahl (Conversational Technologies, Estados Unidos)**

## 6.1. SISTEMAS CONVERSACIONALES

Los últimos avances en aprendizaje automático, reconocimiento del habla y tecnologías de la comprensión del lenguaje natural han estimulado el desarrollo de una amplia variedad de sistemas conversacionales sofisticados, incluyendo tanto sistemas hablados como basados en texto. Algunas de las plataformas mas conocidas incluyen Amazon Alexa, Microsoft LUIS, Google dialogue flow, y IBM Watson, entre muchos otros. De acuerdo con la página web *Chatbots Journal* hay al menos veinticinco plataformas de asistentes virtuales disponibles en el momento.[17]

Estos sistemas conversacionales tienen muchas aplicaciones útiles como servicio al cliente, entretenimiento, compra, servicios gubernamentales y simplemente asistencia a personas en tareas del día a día. Como resultado, hay varios miles de aplicaciones basadas en estos sistemas que existen a día de hoy. Por ejemplo, se han estimado al menos cincuenta mil aplicaciones que usan la plataforma de Amazon Alexa exclusivamente. Similarmente, la plataforma de chatbot Chatfuel afirma ser la fuente de más de trescientas mil aplicaciones para chatbot.

## 6.2. MODELOS, FORMATOS Y APIS DE TERCEROS

Sin embargo, todas las plataformas comerciales de asistentes conversacionales están basadas en modelos de terceros, formatos de resultado y APIs. En algunos casos hay múltiples formatos y APIs correspondientes a las diferentes tecnologías que conforman un sistema conversacional, re-

[17] *25 platforms: https://chatbotsjournal.com/25-chatbot-platforms-a-comparative-table-aeefc932eaff*

conocimiento de habla, comprensión del lenguaje natural, gestión del diálogo, y salida de texto a voz, consecuentemente, en un sistema de terceros, cada componente podría potencialmente tener su propio modelo de terceros, formato y APIs. Como resultado, una aplicación desarrollada para una plataforma nunca operará en otra plataforma, resultando en una compleja y ardua composición de distintas fuentes en un mismo sistema.

Sin embargo, la mayoría de las funcionalidades que estas plataformas ofrecen son en realidad bastante similares, debido simplemente a la naturaleza de los sistemas conversacionales. Como requisitos para conversar con personas, todas ellas necesitan reconocer la voz, comprender el lenguaje natural, entablar un diálogo y presentar información al usuario a través del uso de la voz, texto o gráficos. Cada una de estas capacidades serán similares en cualquier sistema conversacional, resultando en una gran oportunidad para la estandarización.

## 6.3. EL VALOR DE LOS ESTÁNDARES

Los beneficios de los estándares han sido reconocidos desde hace mucho en áreas técnicas y no técnicas. La adopción de los estándares para sistemas conversacionales tendrá, de maneras similar, muchos beneficios. Aquí enumeramos algunos ejemplos.

### 6.3.1. *Interoperabilidad*

Las aplicaciones que han sido desarrolladas usando formatos estandarizados pueden ser usadas en más de una plataforma. Sus componentes pueden ser mezclados y emparejados de manera que el sistema tenga la opción de un mejor uso de los componentes de clases.

### 6.3.2. *Inspirando un ecosistema de herramienta/entrenamiento*

La existencia de estándares fomenta un ecosistema de herramientas y entrenamiento debido a que los desarrolladores de herramientas no tendrán que desarrollar versiones distintas de su software dependiendo del formato usado por una empresa específica. Similarmente, les resultará a los profesores mucho más sencillo crear materiales de entrenamiento si sólo necesitan proveer de materiales para un conjunto estándar de formatos y protocolos de comunicación.

#### 6.3.3. *Pequeñas organizaciones, equipos de investigación o particulares pueden contribuir*

Las pequeñas organizaciones que no tengan los medios para desarrollar una completa pila tecnológica para sistemas conversacionales complejos serán capaces de desarrollar componentes para sistemas conversacionales reusables. De manera similar, serán capaces de hacer uso de componentes tales como reconocedores de voz creados por otros desarrolladores. Será mucho más sencillo para equipos de investigación con recursos limitados progresar en áreas de investigación avanzadas debido a que se podrán enfocar en innovar en vez de en desarrollar componentes que necesiten pero que no se encuentren dentro de su enfoque principal.

#### 6.3.4. *Costes y licencias*

Los costes y licencias de los componentes de software normalmente serán reducidos si se basan en estándares libres.

#### 6.3.5. *Estabilidad respecto a los cambios en el mercado*

En un área técnica en constante cambio como la de los sistemas conversacionales las empresas vienen y van. Incluso aunque la empresa no deje el negocio completamente, puede ser adquiridas o podrían decidir descontinuar un producto. Por ejemplo, las plataformas conversacionales wir.ai y api.ai han sido recientemente adquiridas por Facebook y Google, respectivamente. Mientras que estas adquisiciones no han resultado en grandes cambios en esas plataformas, podría suceder fácilmente.

Si una plataforma desaparece o cambia substancialmente, puede ser catastrófico para terceras partes que se hallen encerradas dentro del uso de esa plataforma. Esto resulta especialmente problemático si el sistema estaba encargándose de aplicaciones de funcionalidad crítica en áreas como la seguridad pública o defensa. Cambios significativos en el coste también pueden ocurrir en cualquier momento, los cuales pueden tener consecuencias similares. Por otra parte, la pérdida de uno o dos proveedores del ecosistema tecnológico no tiene por qué ser catastrófico si las tecnologías están basadas en estándares, ya que los productos de otros proveedores seguirán estando disponibles y pueden incluirse fácilmente en el sistema como substitutivo de las plataforma descontinuadas.

#### 6.3.6. *Formatos y APIs estándar son más completos que sistemas de terceros*

Protocolos de terceros desarrollados dentro de una organización en un área especializada como la de sistemas conversacionales tienen muchas posibilidades de no incluir funcionalidades esenciales. Esto es debido a que ninguna organización puede tener por si sola la misma capacidad en competencias que una comunidad internacional de investigadores que participan en el desarrollo de estándares. En una organización, es probable que algunas funcionalidades fundamentales sean ignoradas. Un buen ejemplo son los resultados de procesamiento alternativo (llamados “nbest list”). En la mayoría de los casos, los resultados alternativos no están disponibles en plataformas de terceros, a pesar de que se ha demostrado que son esenciales en aplicaciones reales.

### 6.4. ESTÁNDARES EXISTENTES PARA SISTEMAS CONVERSACIONALES

Existen un número de estándares actuales que pueden ser usados en sistemas conversacionales.

#### 6.4.1. *Estándares en uso en sistemas comerciales*

El speech synthesis markup language (SSML) [163] usado para crear texto para sintetizadores de voz. Esta herramienta permite a los desarrolladores controlar la ratio de voz, tono, énfasis y pausas. Se encuentra disponible actualmente en el asistente conversacional de Microsoft, Actions en Google, y Alexa de amazon.

Estándares de protocolos de transporte populares como HTTP [164] y Websockets [165] así como formatos estándares en uso como JSON son muy usados en sistemas conversacionales, aunque tengan muchas más aplicaciones.

#### 6.4.2. *Estándares relevantes no usados en sistemas comerciales*

Existen diversos estándares que podrían ser usados en sistemas conversacionales, pero que no son usados a día de hoy. Algunos son:

- State-chart markup language (SCXML)[166]: define estados y eventos en el procesamiento de diálogos.

- El W3C Multimodal architecture [167]: una API que describe la comunicación entre los componentes de un sistema conversacional.
- Extensible Multimodal Annotation (Emma) [168]: un formato usado para representar el resultado del procesamiento del lenguaje.
- ISO-TimeML [169]: un formato para anotar las características temporales de las sentencias.
- DiAML (ISO 24617-2) [170]: anotación para actos de diálogo.

### 6.4.3. *Estándares en uso en programas de investigación*

Mientras que el uso de estándares en sistemas conversacionales es limitado, los estándares son muy usados en programas de investigación. Esto puede ser causado por el hecho de que las aplicaciones desarrolladas en investigación suelen tener una funcionalidad más compleja que los sistemas comerciales, son desarrolladas por un equipo de desarrolladores más expertos y no están subyugadas al hecho de ser rentables económicamente. Los sistemas descritos en [171] [172] [173] [174] [175] [176] [177] [178] [179] son sólo unos pocos ejemplos de sistemas conversacionales de investigación que usan estándares.

## 6.5. RETOS PARA LA ADOPCIÓN DE ESTÁNDARES EN SISTEMAS CONVERSACIONALES DE TERCEROS

A pesar de los beneficios obvios de los sistemas basados en estándares, podemos ver que sólo SSML es usado actualmente en sistemas conversacionales comerciales. ¿Cuáles son los motivos para el fallo de los usos de estándares? En esta sección examinamos cuatro tipos de motivos.

### 6.5.1. *Falta de conocimiento*

Los desarrolles de plataformas pueden no estar al tanto de la existencia de estándares que son aplicables a sistemas conversacionales. Muchos implementadores nuevos en el campo desconocen los sistemas conversacionales y por esa razón no conocen los sistemas anteriores y estándares. Las nuevas plataformas conversacionales son presentadas en el mercado como nuevas tecnologías, por este motivo puede que los desarrolles asuman la no existencia de estándares relevantes.

### 6.5.2. *Falta de estándares y características*

Algunas plataformas pueden requerir de características que no existen en los estándares actuales, y como consecuencia rechazan los estándares por completo por la simple falta de algunas de esas características. Este hecho ignora que un estándar ampliado con características de terceros y un formato más cercano a un estándar es mejor que un formato que sea completamente ad hoc.

### 6.5.3. *Percepción de que los estándares son irrelevantes*

Los desarrolladores de plataformas pueden estar al tanto de los estándares, pero elegir no aplicarlos debido a la creencia de que los estándares son irrelevantes o no tienen valor para sus plataformas. Puede que también crean que adoptar un estándar no es necesario porque su producto es único y auto contenido, y por lo tanto no hay necesidad de interoperabilidad.

### 6.5.4. *Decisión de ignorar estándares*

Incluso aunque los sistemas comerciales estén al tanto de los estándares aplicables a sus plataformas, pueden decidir no usarlos por uno o varios de los siguientes motivos. Incluso aunque entiendan el valor de los estándares puede que no vean suficiente valor en ellos en compensación con los problemas que acarrea su uso.

**Complejidad percibida por los desarrolladores de plataformas.** Algunos de los estándares mencionados anteriormente son muy ricos en características y pueden resultar en una implementación compleja de todas esas características del estándar. Por ejemplo, EMMA contempla la característica específica de una anotación de la confianza de arcos individuales en un retículo de palabras. Sin embargo, las características complejas son típicamente opcionales y una plataforma no tiene por qué implementar completamente todas las características de un estándar para ser una implementación útil.

Si un estándar requiere de una familiaridad básica con una tecnología subyacente, por ejemplo XML, puede ser difícil para algunas compañías adquirir ese conocimiento. Como resultado, puede que prefieran usar un acercamiento más ad hoc que requiera habilidades de desarrollador más accesibles.

**Complejidad percibida para desarrolladores de aplicaciones.** Algunas compañías pueden llegar

a creer que el uso de estándares resultará en un desarrollo que parezca muy complejo para su equipo de desarrolladores. Como consecuencia, temen que basar sus plataformas en estándares produzca un rechazo por parte de los desarrolladores en el uso de sus plataformas. Debido a que los sistemas conversacionales son inherentemente complejos, algunas compañías pueden ser reacias a añadir la percepción de complejidad en su proceso de desarrollo.

**Percepción de un incremento en el tiempo de implementación.** La comprensión e implementación de estándares conlleva un sobre esfuerzo temporal y puede ralentizar la salida al mercado, lo cual es algo significativo en este mercado. La ralentización del proceso de implementación puede ser percibida como una interferencia en la idea de un producto viable mínimo (MVP), donde el objetivo es obtener feedback de clientes lo más rápido posible.

**Interoperatibilidad percibida como un inconveniente.** Algunas empresas pueden oponerse a la interoperabilidad con el objetivo de dificultar el cambio a otra empresa por parte de clientes y desarrolladores, así como el intercambio de componentes.

**Percepción como una limitación a la innovación.** Puede que también exista una percepción de que los estándares reprimen la innovación por medio de una limitación a la hora de elegir plataforma. Sin embargo, la gran mayoría de plataformas conversacionales de investigación están basadas en una extensa fundamentación en otras investigaciones y no requieren de una nueva innovación en cada una de sus partes. Este caso es muy particular en el área de formatos de datos y APIs que son las más comunes en los tipos de estándares.

**Costes de licencias.** Algunos estándares, como los publicados por ISO, requieren de un coste de licencia; sin embargo, los estándares de la web (publicados por la W3C) y estándares de internet (publicados por la Internet Engineering Task Force) no tienen dicho coste.

## 6.6. SUGERENCIAS PARA LA SUPERACIÓN DE RETOS

### 6.6.1. *Análisis para la identificación de la falta de características*

Si los estándares no están en uso debido a la percepción de la falta o incluso por la falta real de características, la participación de miembros de los grupos de investigación, de las plataformas o de las comunidades de desarrolladores. Los participantes pueden proveer puntos de vista sobre qué características son necesarias y cómo podrían ser incorporadas en futuras versiones de los

estándares. De manera similar, si lo estándares no están en uso porque usan viejos formatos como XML, las versiones de los estándares que usen formatos más populares como JSON, pero con la semántica original, podrían ser desarrollados.

### 6.6.2. *Middleware y herramientas de desarrollo*

Un middleware que soporte la interoperabilidad de los componentes usando formatos distintos puede ser extremadamente útil. Por ejemplo, en "Standard portals for intelligence services" (Dahl, 2017) discutí la posibilidad de un servidor de middleware que mediara entre componentes basados en la nube que usen protocolos distintos, de manera que pareciese de cara al desarrollador que todas las interacciones están basadas en estándares. Un primer paso hacia esta idea se podría conseguir a través de un estudio comparativo de diferentes formatos de terceros y un listado de sus similitudes y diferencias. Las diferencias en los formatos pueden ser reconciliadas mediante middleware.

De hecho, algunos middlewares relevantes están empezando a ser accesibles en el proceso de desarrollo. Como ejemplo tenemos la herramienta de desarrollo cross-platform Jovo[18]. Esta plataforma soporta dos plataformas de terceros muy populares como Google DialogFlow y Amazon Alexa. Los exportadores e importadores de diferentes formatos también pueden ser útiles. Por ejemplo, Google DialogFlow ahora contiene exportadores para Alexa y Microsoft Cortana[19] y un importador para formatos de Alexa.

### 6.6.3. *Demos e implementaciones de código abierto de referencia*

La disponibilidad de implementaciones de referencia con código abierto pueden ser un estímulo de mucho valor para desarrolladores a los que les gustaría empezar con ventaja en el código basado en estándares. La disponibilidad de demos que ilustren el valor de los estándares también puede resultar en un buen estímulo.

[18] *https://www.jovo.tech/*

[19] *https://dialogflow.com/docs/integrations/alexa-exporter*

### 6.6.4. *Acciones por parte del gobierno*

El uso de estándares podría ser requerido por parte de las agencias de financiación en proyectos de investigación para sistemas conversacionales; sin embargo, como mencionamos antes, los programas de investigación usualmente ya apoyan más los estándares que los sistemas comerciales.

Junto a la función que un gobierno puede tener en la promoción de los estándares en los programas de investigación, la exigencia del uso de estándares debería ser considerada en las contrataciones gubernamentales que incluyan sistemas conversacionales. Desde el punto de vista del gobierno, este requisito será muy valioso por al menos dos motivos. En primer lugar, aumentará la competencia y en segundo lugar, reducirá la posibilidad de una pérdida crítica de funcionalidad debido a que las empresas quiebren, sean adquiridas, o simplemente finalicen el apoyo a un cierta plataforma de un sistema conversacional de terceros.

## 6.7. CONCLUSIONES

El panorama actual de sistemas conversacionales se encuentra fracturado entre distintas empresas y existe muy poca estandarización de APIs, modelos, evaluación y formatos definidos para los resultados, a pesar de que existen un número relevante de estándares que podrían ser usados. Este hecho resulta en muchas ineficiencias e introduce un riesgo cuando los sistemas conversacionales de terceros son usados para aplicaciones de funcionalidades críticas. Algunas de las causas de esta fracturación incluyen la falta de conocimiento de los estándares, la percepción de que los estándares existentes no son aplicables a los sistemas que están siendo construidos y a las decisiones conscientes de no usar estándares.

Dichas decisiones pueden estar basadas en la percepción de que los estándares son muy complejos, innecesarios, entorpecen el ritmo de desarrollo, o a la ausencia de ciertas características. Además, los desarrolladores de plataformas puede que simplemente quieran hacer que a sus clientes les resulte más difícil cambiar de plataforma. Algunas sugerencias sobre cómo abordar estos retos incluyen la creación de talleres que exploren los estándares actuales e identifiquen la falta de características, estimulando el desarrollo de middleware y de implementaciones de código abierto de referencia, y la exigencia del uso de estándares en proyectos gubernamentales.

# 7. ADAPTACIÓN Y MODELADO DEL USUARIO Y EL CONTEXTO

En años recientes, una cuestión de gran interés ha sido el desarrollo de herramientas y técnicas que faciliten la evaluación de los sistemas de diálogo basándose en la generación automática de conversaciones entre el sistema de diálogo y un módulo adicional llamado simulador del usuario, que representa la interacción de los usuarios con el sistema de diálogo y cuyo comportamiento puede aprenderse a partir de un corpus de diálogos usuario-máquina. La finalidad de realizar modelos de usuario no reside en modelar las características precisas de cada uno de los usuarios, sino servir como herramienta en la fase de desarrollo de un nuevo sistema de diálogo.

## 7.1. MODELADO Y SIMULACIÓN DE USUARIO

La investigación de técnicas de modelado del usuario tiene una larga historia dentro de los campos del procesamiento del lenguaje natural y los sistemas de diálogo hablado, y en particular en las áreas relativas al diseño de sistemas adaptados al usuario.

Tradicionalmente, la mayoría de los trabajos referentes al modelado de los usuarios se ha caracterizado por la utilización de aproximaciones no estadísticas. En estas aproximaciones, el objetivo es construir un modelo representativo del usuario que describa su estado durante la interacción con el sistema. Para ello, es necesario incorporar al modelo información referente al objetivo del diálogo y la historia del diálogo (intercambio de información que se ha producido entre el usuario y el sistema hasta el instante actual del diálogo). Además, dependiendo de la tarea, puede ser necesario incluir en el modelo información referente a las preferencias del usuario, grado de conocimiento sobre la tarea o el propio sistema de diálogo, conocimientos o capacidades, nivel de satisfacción con respecto al comportamiento del sistema, etc.

En [180] se realiza una revisión de la aplicación de técnicas de modelado del usuario en los campos más representativos de los sistemas de diálogo, como la generación de lenguaje natural, compresión del lenguaje natural y gestión del diálogo. En el caso de la gestión del diálogo, las primeras aplicaciones del modelado de usuario se basan en modelos complejos determinados por el conocimiento de la tarea. Actualmente, los modelos de usuario se han ido simplificando, de modo que el modelado de usuario para la gestión del diálogo se realiza muy frecuentemente mediante unas reglas sencillas o mediante pares atributos-valor.

Como alternativa a las reglas surgió el modelado estadístico de usuarios reales, que se emplea

como solución a la falta de los datos necesarios para realizar el entrenamiento y el test en el aprendizaje automático de la gestión del diálogo.

Estas técnicas de aprendizaje automático tienen como objetivo aprender estrategias de gestión del diálogo óptimas a partir de un corpus de diálogos utilizando métodos de “prueba y error” en lugar de basarse en principios de diseño empíricos (diseño a mano de la estrategia basándose en una serie de reglas). No obstante, el tamaño de los corpus etiquetados suele ser demasiado pequeño para explorar suficientemente el espacio global de posibles estados y estrategias del diálogo.

Además, no existe una garantía de que la estrategia óptima pueda estar presente en el corpus de diálogos disponible, con lo que puede argumentarse que no puede aprenderse una estrategia óptima a partir de un corpus prefijado, independientemente de su tamaño. Una solución interesante para este problema consiste en entrenar un modelo de usuario probabilístico para simular las intervenciones del usuario y utilizarlas en el aprendizaje mediante la interacción entre el gestor del diálogo y el usuario simulado. En esta línea pueden referenciarse los trabajos [181] [182] [183] [184] [185] [186] [187].

El usuario simulado también permite explorar estrategias de diálogo no presentes en el corpus. De esta forma, el gestor del diálogo puede llegar a desviarse de las estrategias iniciales y aprender nuevas metodologías potencialmente mejores. La simulación de usuarios estadística se realiza en dos fases. En la primera, el modelo de usuario se entrena a partir de un corpus de diálogos para aprender qué respuestas proporcionaría un usuario real en una determinada situación del diálogo. En esta fase suelen utilizarse técnicas de aprendizaje supervisado. En la segunda fase, el modelo de usuario entrenado se utiliza para predecir las respuestas a las acciones del usuario.

El sistema aprendido interactúa con el usuario y optimiza la estrategia del diálogo basándose en la realimentación ofrecida por el usuario simulado. De esta forma, pueden adquirirse tantos diálogos de aprendizaje como se desee y además permite estrategias del diálogo que no están presentes en el corpus inicial de diálogos persona-máquina. Por tanto, posibilita que el gestor del diálogo se desvíe de las estrategias conocidas y aprenda una estrategia de gestión mejor.

A la hora de estimar el modelo de usuario, deben tenerse en cuenta tanto factores observables (por ejemplo, la historia del diálogo hasta el momento actual) como no observables (objetivo del usuario, memoria, preferencias, etc.). Existe un gran número de referencias sobre diferentes aproximaciones y metodologías:

- **Modelos de n-gramas**

  La utilización de modelos estadísticos para predecir la próxima acción del usuario se sugirió por primera vez en [181] [188]. En estos trabajos se introduce un modelo de n-gramas para predecir la acción más probable del usuario en el instante $t$ dado el historial anterior de acciones del sistema y del usuario. En la práctica, la casuística existente y la falta de las suficientes muestras de datos imposibilitan la utilización de toda la historia del diálogo. Eckert, Levin y Pieraccini aproximan la historia completa mediante un modelo de bigramas.

  El modelo propuesto tiene la ventaja de ser puramente probabilístico y completamente independiente de la tarea. El punto débil del modelo radica en la no definición de restricciones en el comportamiento del usuario simulado. De este modo, cualquier acción del usuario puede ser válida tras una acción del sistema, a pesar de toda la historia previa del diálogo. Con la utilización de bigramas, la respuesta generada por el modelo puede corresponderse correctamente con la última respuesta del sistema, pero puede carecer de sentido teniendo en cuenta toda la historia previa del diálogo.

  En [122] se describe cómo el modelo puro de bigramas puede modificarse para representar de forma más realista la historia del diálogo (Modelo de Levin). En lugar de permitir que cualquier respuesta del usuario pueda seguir a una determinada acción del sistema, sólo se estiman las probabilidades de algunos pares {respuesta del usuario - acción del sistema}, mientras que todas las demás probabilidades se consideran nulas. El conjunto de parámetros del modelo probabilístico del usuario caracteriza el nivel de cooperación y el grado de iniciativa del usuario simulado.

  Al igual que en el modelo de bigramas, el modelo de Levin no asegura la consistencia entre diferentes acciones del usuario a lo largo del transcurso del diálogo debido a la suposición de que la respuesta del usuario depende únicamente del último turno del sistema. Como en el caso del modelo de bigramas, las acciones del usuario pueden no cumplir restricciones lógicas del diálogo, haciendo que los diálogos continúen indefinidamente debido a que el usuario cambia continuamente de objetivo o repite información previamente suministrada.

- **Modelos basados en grafos**

  Scheffler y Young proponen en [189], [190] [191] [182] un modelo basado en grafos como solución a la inconsistencia en los objetivos del usuario que presenta el modelo de Levin. Para ello, se combinan reglas deterministas para las acciones dependientes del objetivo y un modelo probabilístico para cubrir el resto de las acciones del usuario durante el diálogo.

En el modelo de Scheffler y Young, todos los posibles "caminos" que puede tomar un usuario durante el diálogo deben representarse en forma de red. Los arcos de la red simbolizan acciones y los nodos representan "puntos de elección". Se establece una división de los nodos en nodos de elección probabilísticos (el simulador de usuario toma una decisión aleatoria en base a las probabilidades estimadas a partir de un corpus de datos etiquetados) y nodos de elección determinista, donde la ruta a tomar al llegar a ellos depende del objetivo del usuario. Este objetivo permanece fijo durante todo el diálogo.

El principal inconveniente de este modelo es su alta dependencia del conocimiento que se disponga de la tarea, dado que un aspecto crítico del modelo es la especificación de todos los posibles caminos en el diálogo, que supone un gran esfuerzo manual. Esta tarea puede automatizarse, en parte, si existe un prototipo inicial del sistema del diálogo y el rango de posibles acciones del usuario se definen y acotan correctamente.

- **Redes Bayesianas**

  Pietquin, Beaufort y Dutoit combinan características del modelo de Scheffler y Young y del modelo de Levin, con el principal objetivo de reducir el esfuerzo manual de la construcción de las redes con los puntos de elección [183] [184]. La principal idea de su trabajo se basa en condicionar el conjunto de probabilidades indicadas en el modelo de Levin teniéndose en cuenta el objetivo y la memoria del diálogo.

  De forma similar al modelo de Scheffler y Young, el objetivo del diálogo es una simple tabla que contiene pares $< nombre\ del\ atributo,\ valor\ del\ atributo >$ con variables de estado asociadas. Estas variables se utilizan para estimar la probabilidad del usuario para cada uno de los valores de cada campo y para detectar la frecuencia en la que el usuario ha mencionado una determinada información durante el transcurso del diálogo.

  En este trabajo, Pietquin selecciona a mano todos los parámetros del modelo utilizando principios empíricos y el sentido común. Los valores de las probabilidades se seleccionan también a mano. Se sugiere la utilización de una red bayesiana para implementar y visualizar el modelo propuesto. Las variables de entrada a la red son los tipos de acciones del sistema (por ejemplo: bienvenida, confirmación, respuesta, etc.) y los nombres de los atributos mencionados por el usuario. Las variables de salida son los nombres de los atributos y sus valores proporcionados por el usuario y una variable booleana que indica si el usuario finaliza la interacción con el sistema o no. El objetivo del usuario y la función que memoriza la información que ha mencionado se tratan como variables internas de la red.

- **Técnicas de Aprendizaje Automático (Machine-Learning)**

  Georgila, Henderson y Lemon proponen la utilización de modelos de Markov, realizando una descripción más detallada de los estados del modelo, historias del diálogo más amplias y empleando técnicas de aprendizaje automático [192].

  El diálogo se describe como una secuencia de Estados de Información [193], cada uno de los cuales viene representado por un vector de características que describe el estado actual del diálogo, la historia previa del mismo y cualquier posible respuesta del usuario válida para dicho estado. Aunque la riqueza de información del modelo ayuda a compensar la suposición de Markov, se requiere una gran cantidad de datos de entrenamiento para estimar de manera fiable los parámetros del modelo.

  En su trabajo, se presentan dos métodos diferentes para predecir la próxima acción del usuario dada una historia de estados de información. El primer método reutiliza el modelo de n-gramas propuesto en [181], pero utilizándose valores de $n$ de 2 a 5 para cubrir una historia más amplia del diálogo. Se menciona que los mejores resultados se obtienen con 4-gramas, es decir, utilizándose una historia de 4 estados de información para predecir la próxima acción del usuario. El espacio de estados utilizado en las experimentaciones es del orden de $10^{87}$ estados, con lo que existe un gran número de secuencias de estados no disponibles en los datos de entrenamiento.

  El segundo método se basa en la utilización de una combinación lineal de características para realizar la correspondencia entre un estado $s$ y un vector de características con valores reales $f(s)$. La mayoría de estos valores son binarios, indicando la presencia o ausencia de una determinada información (por ejemplo: destino, fecha de salida...) y el resto de las variables toman valores continuos (por ejemplo, el WER estimado). En total, se utilizan 290 variables [192]. Se utilizan técnicas de aprendizaje supervisado para estimar el conjunto de pesos $wa$ para cada acción $a$ que describen lo apropiado de utilizar cada vector de $f(s)$ para predecir $a$. Una vez se estiman los pesos, se calcula una función $P(a|s)$ para cada una de las acciones. Para ello se aplica una función exponencial normalizada al producto de $f(s)$ y $wa$.

  Dado que cada Estado de Información incluye no sólo el estado actual del diálogo, sino también información sobre la historia previa del diálogo, pueden modelarse los aspectos que pueden contribuir en el comportamiento del usuario para predecir su siguiente acción.

- **Modelado del usuario como un proceso de decisión de Markov (MDP)**.

  Destacan los trabajos [181] [123] [194]. Tal y como se ha comentado en la Sección 2.2.2, un MDP puede describirse formalmente como un espacio de estados finito $S$, un conjunto finito de acciones $A$, un conjunto de probabilidades de transición $T$ y una función de recompensa $R$. En cada instante $t$, el gestor del diálogo se encuentra en un estado $s_t \in S$, ejecuta una acción discreta $a_t \in A$, transita a un nuevo estado de acuerdo a la probabilidad $p(s_{t+1}|s_t, a_t)$ y recibe una recompensa $r_{t+1}$. De este modo, el gestor del diálogo interactúa con el simulador de usuarios, realizando transiciones entre los diferentes estados del modelo a medida que realiza acciones como respuesta a las intervenciones del usuario. El gestor del diálogo recibe una recompensa por cada acción que lleva a cabo, siendo la estrategia de diálogo óptima la que maximice las recompensas recibidas a lo largo del tiempo.

- **Modelos Ocultos de Markov (HMMs)**

  Cuayahuitl, Renals, Lemon y Shimodaira presentan un método para la simulación de diálogos basada en HMMs en el que se generan tanto las acciones del usuario como las del sistema [195]. El principal objetivo es ampliar un corpus pequeño de diálogos persona-máquina con nuevos diálogos simulados.

  Se proponen diferentes variaciones de HMMs. La más avanzada de ellas es un HMM de entrada-salida (IOHMM). El modelo se caracteriza por un conjunto de estados visibles $S = S_1, S_2, \ldots, S_n$ y un conjunto de observaciones $V = V_1, V_2, \ldots, V_m$. Los estados $S$ representan turnos de sistema y las observaciones $V$ se corresponden con el conjunto de acciones del sistema. El estado en el instante $t$ se denota mediante $q_t$, siendo $a_{s,t}$ la acción del sistema en el instante $t$. Las respuestas del usuario se representan mediante un conjunto de intenciones $H = H_1, H_2, ..., H_l$ y la acción del usuario en el instante $t$ se denota mediante $a_{u,t}$. El comportamiento del modelo viene regido por un conjunto de probabilidades de transición $P(q_{t+1}|q_t, a_{s,t})$ y un conjunto de probabilidades de salida $P(a_{s,t}|q_t, a_{u,t-1})$. Las respuestas del usuario se predicen utilizando un conjunto de probabilidades que modelan al usuario $P(a_{u,t}|q_t, a_{s,t})$

  En lugar de entrenar un único modelo IOHMM genérico para simular cualquier tipo de diálogo, los diálogos del corpus inicial se agrupan según objetivos, entrenándose un submodelo para cada uno de los objetivos y utilizándose un modelo de bigramas para estimar la secuencia de objetivos.

Dentro del proyecto BASURDE se elaboró un simulador de usuarios [187] [196] consistente en dos

módulos: un gestor de diálogo de usuario (UDM) y un generador de respuestas de usuario (URG). El UDM es un módulo simétrico al gestor del diálogo estocástico definido para BASURDE. Este módulo recibe frames de sistema, realiza operaciones de lectura y escritura en su propio registro histórico (UHR), lee y realiza transiciones en el mismo modelo utilizado por el gestor del diálogo, aplica un conjunto de reglas para seleccionar las transiciones adecuadas, y genera los frames de usuario. El URG es un módulo simétrico al generador de respuestas del sistema. Recibe los frames de usuario y genera las correspondientes frases en lenguaje natural.

Al principio de cada diálogo, el simulador de usuario lee los parámetros y objetivos del escenario simulado y almacena esta información en el UHR. Seguidamente, lee la versión estática del modelo (sDM), realiza la búsqueda del estado siguiendo los objetivos marcados, y genera el frame correspondiente. En cada turno de diálogo, el simulador de usuario realiza la lectura de los frames de sistema, compara estos frames con los posibles actos de diálogo de sistema, transita a un nuevo estado de sistema en el sDM, actualiza el UHR con los datos suministrados por el sistema, transita a un nuevo estado de usuario en el sDM, y genera los frames de usuario.

## 7.2. EVALUACIÓN DE LAS TÉCNICAS DE SIMULACIÓN

La evaluación de las técnicas de modelado de usuarios es todavía un campo en desarrollo. Usualmente se utilizan métodos adoptados de otros campos de investigación como la recuperación de información y el aprendizaje automático. Una primera clasificación consiste en dividir las técnicas en métodos directos de evaluación que utilizan medidas como la Precisión y la Cobertura, y métodos indirectos que emplean medidas como la Utilidad.

Los métodos directos evalúan el modelo de usuario midiendo la calidad de sus predicciones. Con la precisión y la cobertura se evalúa si las respuestas generadas por el modelo de usuario concuerdan con las que proporcionaría un usuario real en la misma situación del diálogo [133]. Para ello, se extrae una partición de entrenamiento y otra de test a partir de un corpus de diálogos usuario real-máquina, se realiza el entrenamiento del modelo de usuario a partir de la partición correspondiente y se evalúan las respuestas generadas por el simulador para el conjunto de muestras en la partición de test, teniendo en cuenta la historia del diálogo y el objetivo del usuario.

La Cobertura ($C$) mide cuántas acciones de la respuesta del usuario real se predicen correctamente en la respuesta simulada:

$$C = 100 * \frac{Acciones\ predichas\ correctamente}{Total\ acciones\ en\ la\ respuesta\ del\ usuario\ real} \quad (33)$$

La Precisión ($P$) mide la proporción de acciones correctas entre todas las disponibles en la respuesta dada por el usuario simulado. Una respuesta simulada suele considerarse correcta si contiene al menos una de las acciones proporcionadas en la respuesta del usuario real.

$$P = 100 * \frac{Acciones\ predichas\ correctamente}{Total\ acciones\ en\ la\ respuesta\ simulada} \quad (34)$$

Para realizar estudios comparativos de sistemas entre diferentes aproximaciones utilizando para ello una única estadística, suele calcularse la medida $f$:

$$f = \frac{2PR}{P + R} \quad (35)$$

En [133] se recogen resultados de la precisión y la cobertura de diferentes modelos de usuario. Los valores máximos se sitúan alrededor del 35 %, mientras que la medida para el modelo de bigramas es del 20 %. Una crítica de estas medidas es que realizan una alta penalización a las acciones no vistas en la respuesta simulada, aunque las respuestas generadas sean aceptables y compatibles perfectamente con las que podría proporcionar un usuario real.

De este modo, dado que la estrategia del diálogo debe funcionar correctamente para todos los tipos posibles de respuestas de usuario, y no sólo para el usuario más probable, la evaluación del modelo únicamente teniendo en cuenta la Precisión y la Cobertura es insuficiente.

Una solución a esta problemática consiste en la adquisición de un mayor número de diálogos mediante la interacción del modelo de usuario simulado preferiblemente con el gestor de diálogo utilizado para la adquisición (misma estrategia de gestión) y la definición de medidas estadísticas que permitan realizar un estudio comparativo del corpus simulado con respecto al adquirido con usuarios reales. En [133] y [190] se proponen un conjunto de medidas estadísticas para llevar a cabo este tipo de evaluación:

- Características del diálogo: Número medio de turnos de diálogo, número medio de acciones por turno de diálogo, ratio de acciones de sistema frente a las del usuario, etc.
- Estilo del diálogo: Frecuencia de los diferentes actos de diálogo, grado de cooperación del

usuario (ratio de valores de slots proporcionados cuando se le solicita al usuario), proporción de acciones encaminadas a conseguir el objetivo del diálogo frente al resto de acciones, etc.

- Tasa de éxito y eficiencia de los diálogos: Tasa de consecución de los diferentes objetivos, tiempos transcurridos para cada uno de los objetivos, etc.

De nuevo, cabe destacarse que mediante estas medidas únicamente puede esbozarse el comportamiento del simulador, no existiendo rangos de medidas que indiquen si el modelo de usuario es lo suficientemente realista. Además, no hay ninguna garantía de que el diálogo simulado sea realista, aunque el valor de estas medidas coincida con el caso del corpus real.

En [192] se introduce el uso de la Perplejidad ($PP$) para la evaluación del modelo de usuario, midiendo si los diálogos simulados contienen secuencias de acciones similares a las contenidas en los diálogos usuario real-máquina. La definición de la Perplejidad se basa en la entropía $H$:

$$PP = 2^{\hat{H}} \tag{36}$$

representando la entropía la cantidad de información no redundante proporcionada por cada nueva acción (estado) en media:

$$\hat{H} = -\frac{1}{m} log_2 P(a_1, a_2, \ldots, a_m) \tag{37}$$

donde $P(a_1, a_2, \ldots, a_m)$ es un estimador de la probabilidad de que el simulador lleve a cabo la secuencia de acciones $a_1, a_2, \ldots, a_m$.

En [195] la comparación entre el corpus simulado y el adquirido con usuarios reales se lleva a cabo entrenando un HMM con cada uno de los corpus y midiendo la semejanza entre ambos corpus en base a la distancia entre los dos HMM. La distancia definida se rige por la desigualdad de Kullback-Leibler:

$$D(P, Q) = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2} \tag{38}$$

donde $D_{KL}$ es la distancia entre las distribuciones de probabilidad $P$ y $Q$.

En cuanto a los métodos indirectos de evaluación, el principal objetivo es medir la *Utilidad* del modelo de usuario en el contexto del funcionamiento del sistema completo. Usualmente, se trata de evaluar el funcionamiento de la estrategia de diálogo aprendida mediante el simulador. Esta evaluación, en la práctica, en lugar de realizarse comprobando el funcionamiento de la estrategia con usuarios reales, suele llevarse a cabo verificando el funcionamiento de la nueva estrategia mediante la nueva interacción con el simulador de usuario. De este modo, se compara la estrategia inicial (normalmente fijada a mano) con la aprendida con el simulador, reaprendiendo el modelo de usuario a partir de una partición del corpus que no haya sido utilizada en el aprendizaje del simulador utilizado en la obtención de la nueva estrategia. El principal problema de esta evaluación reside en la dependencia del corpus adquirido con respecto al modelo de usuario, lo que no permite detectar si la nueva estrategia se adapta únicamente al tipo de usuario disponible en el modelo.

## 7.3. ADAPTACIÓN AL USUARIO Y AL CONTEXTO DE LA INTERACCIÓN

Un tema de especial interés actualmente es el desarrollo de sistemas de diálogo capaces de interactuar en diferentes dominios, diferentes entornos o con la posibilidad de adaptarse a los diferentes perfiles o preferencias de usuarios heterogéneos.

Para tratar este aspecto se han desarrollado diferentes metodologías, que tienen como principal finalidad la adaptación de las estrategias utilizadas por el sistema para la interacción con el usuario y la confirmación de los datos aportados. A continuación, se comentan algunos trabajos realizados en esta temática:

- En [197] se presenta un sistema para crear y gestionar aplicaciones adaptadas al usuario que requieran un interfaz vocal. El sistema está compuesto por cuatro módulos: el generador automático de diálogo (ADG, *Automatic Dialogue Generator*), el gestor del perfil (PM, *Profile Manager*), el gestor de información y servicios (ISM, *Information Services Manager*) y el gestor de diálogo (DM, *Dialogue Manager*).

  El PM codifica las preferencias del usuario utilizando una representación de los servicios e informaciones en los que el usuario está interesado. De este modo, el usuario define una aplicación personalizada, en cuanto a los contenidos como al formato de presentación de la información. El módulo ADG genera un modelo de diálogo basado en estados finitos a partir de la información sobre la tarea descrita en un conjunto de tablas. El módulo DM

utiliza comandos en lenguaje VIL (*Voice Interface Language*) para llevar a cabo la gestión de diálogo.

En trabajos posteriores [198] se presentan ampliaciones del módulo. Este módulo permite modificar la estructura del diálogo siguiendo las preferencias anotadas por el usuario en su fichero correspondiente, facilita los cambios en la gestión de la tarea seleccionada e, incluso, la migración a nuevas tareas (dada la centralización de la información correspondiente a la misma en una serie de ficheros que consultan los diversos módulos del sistema).

- Los investigadores del MIT han desarrollado la utilidad Speech-Builder [199], diseñada para especificar la información lingüística específica del dominio, de manera que se facilite la creación o adaptación de sistemas para nuevas tareas. Las funciones que pueden llevarse a cabo utilizando esta herramienta son:

  - Editar las pronunciaciones en el módulo de reconocimiento del habla.
  - Configurar un fichero de gramática y un fichero de conversión de análisis a representaciones semánticas.
  - Diseñar un gestor de diálogo genérico, enfocado únicamente a tratar las situaciones que se producen en las consultas a las bases de datos.

  En [200] se resumen distintos trabajos llevados a cabo por el MIT para el desarrollo de gestores de diálogo genéricos capaces de realizar las funciones esenciales del diálogo y adaptables a un dominio específico mediante ficheros externos. Para facilitar la adaptación, se describe un método para la organización de la información específica del dominio, mediante la definición automática de categorías semánticas.

- En el sistema SENECA se desarrollaron dos tipos de interacción con el usuario, en las que se tiene en cuenta si se trata de un usuario novato o experto en el sistema. Estos modos de interacción influyen en las respuestas generadas por el sistema, de modo que al usuario experto se le proporcionan *prompts* menos detallados y concisos. Para la elección de un modo de interacción u otro se tiene en cuenta el número de errores de reconocimiento contabilizados en las intervenciones previas del usuario.

- En el gestor de diálogo desarrollado para AT&T Communicator se realiza un control de la iniciativa del diálogo teniendo en cuenta las intervenciones que va realizando el usuario. El sistema asume inicialmente condiciones de diálogo “normales”, concediendo la mayor iniciativa posible al usuario. Si el sistema detecta problemas (por ejemplo, solicitudes repeti-

das de un mismo atributo), va concretando las preguntas utilizando mensajes más concisos. Tras un número prefijado de intentos, el sistema selecciona un modelo de interacción en el que se pide la confirmación de toda la información que haya aportado el usuario hasta el momento actual del diálogo y seguidamente se siguen solicitando atributos.

- El gestor de diálogo desarrollado para el sistema TOOT [201] implementa tres posibles iniciativas del diálogo (iniciativa del usuario, del sistema o mixta) y tres posibles tipos de estrategias de confirmación (explícita, implícita o sin posibilidad de confirmación).

  Se desarrollaron dos versiones del sistema, una de ellas adaptativa. Para llevar a cabo la adaptación al usuario durante el transcurso del diálogo se tienen en cuenta las estadísticas de reconocimiento anotadas en un corpus de 120 diálogos adquirido en experimentaciones previas con usuarios no expertos. Cada uno de los turnos de usuario de este corpus se etiquetó utilizando una medida acústica (entre 0 y 1) que simboliza la semejanza de la frase reconocida con la verdaderamente mencionada por el usuario.

  A partir de esta medida, se clasificaron los diálogos en “buenos” o “malos” según el funcionamiento global del reconocedor. Además, se extrajeron 23 características que evalúan el diálogo teniendo en cuenta cinco categorías: confianza acústica, eficiencia del diálogo (por ejemplo, número de turnos), calidad o naturalidad del diálogo (por ejemplo, número de veces que el usuario solicita ayuda), parámetros experimentales (por ejemplo, la estrategia inicial de diálogo seleccionada) y léxicos (por ejemplo, léxico relativo al reconocedor). A partir de estas características se realizó el aprendizaje de un clasificador automático de los diálogos cuyo funcionamiento se rige por reglas. Mediante el uso de estas reglas, el gestor de diálogo decide de forma automática qué estrategia de diálogo es más conveniente para el usuario actual:

  - El sistema se inicializa siempre con una iniciativa de usuario sin confirmaciones.
  - Si aplicando las reglas comentadas a los últimos cuatro turnos del usuario se clasifica el diálogo como “malo”, la estrategia del diálogo pasa a ser iniciativa mixta con confirmaciones implícitas.
  - Si tras realizar esta primera adaptación, el diálogo vuelve a clasificarse como “malo” tras 4 turnos, la estrategia vuelve a restringirse adoptando una iniciativa del sistema con confirmaciones explícitas.
  - Mientras el diálogo se clasifique como “bueno”, no se realizan cambios en la estrategia del diálogo.

El sistema de diálogo desarrollado se evaluó comparativamente con una versión no adaptativa del sistema, midiéndose parámetros definidos en el modelo de PARADISE. Participaron 12 usuarios (6 en cada una de las versiones del sistema). Los porcentajes de éxito del diálogo son del 23 % en el sistema no adaptativo y del 65 % en la versión adaptativa, explicado en gran parte por los mejores porcentajes de reconocimiento. La opinión sobre el funcionamiento del sistema suministrados por los usuarios también es más favorable para la versión adaptativa.

- En el sistema HMIHY se realizó un estudio sobre la predicción automática de las situaciones problemáticas del diálogo a partir de la clasificación del diálogo teniendo en cuenta una serie de características [202]:

  - Características acústicas: gramáticas utilizadas, número de palabras reconocidas, etc.
  - Características de comprensión del lenguaje: medidas de confianza, cobertura, inconsistencia, etc.
  - Características de gestión de diálogo: tipo de tarea, número de confirmaciones, número de repeticiones de preguntas, número de subdiálogos, etc.
  - Características etiquetadas a mano: código del usuario, sexo, edad, modalidad utilizada, porcentajes de reconocimiento, etc.
  - Características globales del diálogo: Número total de turnos, duración de las llamadas, etc.

  En total se definieron 240 características, utilizadas para el etiquetado de un corpus de 4.774 diálogos cuya adquisición se realizó de forma supervisada por un Mago de Oz, a partir de las cuales se utilizaron técnicas de aprendizaje automático (programa RIPPER) para el aprendizaje de un modelo de clasificación. Las problemáticas a detectar eran que el usuario cuelgue antes de cumplir el objetivo del diálogo, situaciones en las que ha sido necesaria la intervención del Mago para reconducir el diálogo y fallos en la estrategia que imposibilitaron cumplir el objetivo.

  La evaluación de esta metodología se llevó acabo utilización validación cruzada. Mediante estas características se mejoraba un 23 % la identificación de las problemáticas respecto al baseline (64 %).

- El gestor de diálogo del sistema MIMIC dispone de una estrategia de iniciativa mixta adaptativa. MIMIC adapta automáticamente la estrategia del diálogo basándose en las carac-

terísticas del diálogo, modelando su comportamiento teniendo en cuenta el contexto del diálogo. Además, el módulo de iniciativa de MIMIC está desacoplado de los objetivos del gestor y de los procesos de selección de las estrategias, lo que permite una adaptación más sencilla a otros dominios. Este módulo determina el grado de iniciativa basándose en el comportamiento del usuario, indicaciones que puedan existir en su turno actual y la historia del diálogo.

El comportamiento o papel del usuario hace que el gestor emplee iniciativas basadas en el sistema para usuarios no familiarizados con el sistema. En este grado de iniciativa, el sistema va solicitando al usuario un único dato en cada turno y acompaña su respuesta con una explicación de las acciones que puede llevar a cabo el usuario tras cada respuesta del sistema. Teniendo en cuenta la historia del diálogo, el gestor utiliza el comportamiento global del usuario para decidir si es necesario realizar un cambio en la estrategia del diálogo.

En [203] se presentan los resultados de la evaluación de la versión adaptativa del sistema de diálogo con respecto a dos versiones no adaptativas del mismo sistema (una con iniciativa del sistema y otra con iniciativa por parte del usuario). Los resultados, obtenidos mediante encuestas al usuario y medición de las estadísticas de los diálogos adquiridos, muestran que la versión adaptativa posee un mejor comportamiento en términos de satisfacción del usuario, eficiencia del diálogo (número de turnos) y calidad del mismo (mejores tasas de reconocimiento).

- El sistema CU FOREX [150] permite dos modalidades de interacción dependiendo de la destreza del usuario:

  - *Directed dialog* (DD): Está diseñado para usuarios novatos. El sistema se encarga de guiar al usuario de forma detallada, indicándole incluso en algunos estados qué valores concretos debe aportar. En cada turno del diálogo, el sistema solicita un único atributo.
  - *Natural Language Shortcut* (NLS): Está diseñado para usuarios expertos que desean realizar toda la consulta en un único turno, aportando toda la información necesaria en dicho turno.

- El sistema AthosMail [204], desarrollado en el ámbito del proyecto europeo DUMAS y cuya tarea es la consulta del correo electrónico, contiene un módulo de modelado del usuario, encargado de realizar la adaptación de las respuestas del sistema a los diferentes niveles

de destreza de los usuarios. Los objetivos establecidos para el desarrollo de este módulo fueron:

- Proveer flexibilidad y variabilidad en las respuestas del sistema.
- Permitir al usuario la interacción con el sistema de una forma más natural.
- Posibilitar a los desarrolladores la implementación y evaluación de técnicas de aprendizaje automático.

Para cumplir estos objetivos, en el diseño del modelo de usuario se tuvieron en cuenta las siguientes directrices:

- Utilizar una representación flexible para codificar los turnos de sistema, utilizada para su generación.
- El sistema posee la funcionalidad de almacenar las acciones del usuario y su comportamiento, estimando así los niveles de decisión que se utilizan para proporcionar recomendaciones al usuario.
- El sistema, además, incorpora un módulo aprendido automáticamente que realiza la clasificación de los mensajes de correo del usuario en función del contenido y preferencias anotadas del usuario.

Los componentes que conforman el módulo de modelado de usuario son:

- Priorizador de mensajes: Clasifica los mensajes de correo utilizando una lista en la que los detectados como más interesantes o prioritarios para el usuario se sitúan en la parte superior. Para ello, se tienen en cuenta las acciones llevadas a cabo previamente por el usuario para asignar unas medidas de prioridad a los nuevos mensajes.
- Averiguador del objetivo: Este componente sugiere al gestor qué objetivo posible persigue el usuario en aquellas situaciones en las que existe incertidumbre en los datos proporcionados por el reconocedor.
- Modelo cooperativo: Este modelo permite al sistema variar la iniciativa y complejidad de sus respuestas dependiendo del nivel de destreza anotado para el mismo, así como de las estadísticas de reconocimiento obtenidas de sus sesiones anteriores. Se definieron cuatro niveles de iniciativa para realizar el control del diálogo: directiva (control por parte del sistema), modo sugerencia (el sistema lleva el control del diálogo, pero está preparado para cambiar el curso del diálogo de acuerdo con las preferencias

indicadas por el usuario), modo declarativo (el usuario tiene la iniciativa, pero el sistema puede solicitarle información requerida), modo pasivo (el usuario tiene el control completo del diálogo, no realizando ningún tipo de sugerencias el sistema).

- Categorización de mensajes: Este componente se utiliza para comparar los mensajes entrantes con los existentes, y categorizarlos utilizando palabras claves detectadas en los mismos.
- Preferencias del usuario: Se almacenan características relativas al estilo de habla preferido por el usuario, velocidad y tipo de voz, remitentes y temas preferidos, etc.

La evaluación realizada del sistema se expone en [205]. Se llevaron a cabo dos tipos de evaluaciones. En la primera de ellas, cinco expertos del diseño de sistemas interactivos proporcionaron su opinión sobre las diferentes modalidades de adaptación al usuario contempladas en el sistema. Para el segundo estudio se recogen las opiniones y estadísticas obtenidas tras el análisis de 104 diálogos adquiridos por 26 usuarios no familiarizados con el sistema, a partir de los cuales se identificaron una serie de errores que cometía el sistema (no detección de frases fuera de la tarea, solapamientos entre las respuestas del sistema e intervenciones del usuario, etc.).

- En [206] se describe un gestor de diálogo desarrollado para un robot doméstico con la tarea de proporcionar recetas de cocina. La implementación del gestor se ha realizado mediante las herramientas TAPAS [207], creadas para la elaboración del gestor de diálogo independiente del dominio y del idioma diseñado para el proyecto ARIADNE. Mediante estas herramientas se facilita la implementación rápida de prototipos dado que sólo es necesario desarrollar aquellos componentes dependientes del dominio. De este modo, permiten el desarrollo de gestores de diálogo independientes de la tarea y del idioma, separando los contenidos propios del dominio de los independientes del mismo.

Parte III

# SISTEMAS CONVERSACIONALES: DISEÑO, IMPLEMENTACIÓN Y EVALUACIÓN

# 8. PLATAFORMAS, ARQUITECTURAS Y HERRAMIENTAS

Una vez descritos los enfoques y métodos relacionados con el diseño de los diversos componentes de los sistemas de diálogo, en el presente capítulo se describen las principales arquitecturas, plataformas y herramientas disponibles para su implementación.

## 8.1. ARQUITECTURAS SOFTWARE

Las arquitecturas software más estrechamente relacionadas con los sistemas conversacionales actuales se pueden clasificar en dos bloques las orientadas a servicios y las orientadas a eventos.

### 8.1.1. *Arquitectura orientada a servicios*

Frente al modelo tradicional de "silos" en las que las aplicaciones software eran monolíticas, el modelo de arquitectura orientada a servicios (Service-Oriented Architecture, SOA), descompone la funcionalidad de un sistema en unidades denominadas servicios que cumplen objetivos muy concretos, de forma que éstos se comunican entre sí para alcanzar la funcionalidad del sistema.

Esta descomposición permite que los servicios estén desacoplados y puedan ser reutilizados para construir muchos otros sistemas. Como se explica con detalle en [208], el protocolo de servicios web engloba una serie de estándares que permiten a estos servicios ser descubiertos, intercambiar datos e invocar a otros independientemente de su localización, sistema operativo o lenguaje en el que estén implementados.

### 8.1.2. *Arquitectura orientada a eventos*

Una arquitectura basada en eventos (Event-Driven Architecture, EDA), se centra en recibir, propagar y dar respuesta a eventos que se producen en el sistema. Este patrón promueve la existencia de servicios desacoplados que se ejecutan sólo en caso de que se produzcan ciertos eventos. Un posible evento relacionado con los sistemas conversacionales sería una palabra clave que despierte un servicio (wake-up word), por ejemplo "Alexa" u "OK Google".

### 8.1.3. *Computación sin servidor*

Los dos modelos anteriores están estrechamente relacionados con la computación sin servidor (Serverless Computing). La mayoría de los servicios relacionados con las tecnologías conversacionales que se describen en la Sección 8.3 están en la nube. Esto no significa que no se ejecuten en una máquina (servidor), sino que el proveedor de la nube realiza la gestión completa de la ejecución de las funciones o servicios. Esto permite a los clientes escribir código, subirlo a la nube y ejecutarlo sin preocuparse de la arquitectura subyacente o el mantenimiento del sistema (olvidándose efectivamente del servidor), lo que abarata costes y acorta tiempo de desarrollo de las aplicaciones. En [209] hay una introducción reciente y detallada a este concepto.

La computación sin servidor está estrechamente relacionada con el concepto de *función como servicio* (function as a service, FAAS), en el que los eventos se responden con funciones simples, como ocurre por ejemplo con Amazon AWS Lambda.

Como ejemplo de todo lo anterior mostramos la arquitectura propuesta para un chabot con Amazon Lex en la Figura 10[20]. Lex es un servicio de Amazon para desarrollar chatbots (ver sección 8.3). En la figura se observa que el evento de una llamada del usuario inicia el servicio Amazon Connect, que invoca a Amazon Lex, quien se encarga de resolver la consulta del usuario invocando a funciones de AWS Lambda (FAAS) que permiten obtener la información requerida para responderle.

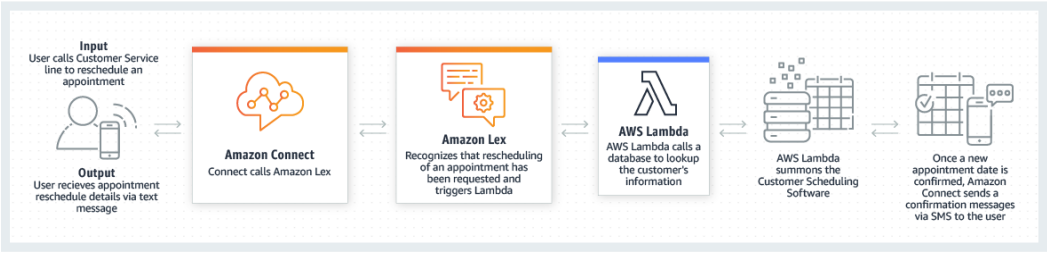


*Figura 10: Ilustración de los servicios empleados por una aplicación que utiliza Amazon Lex*

## 8.2. ENFOQUES DE IMPLEMENTACIÓN EN EL ÁMBITO INDUSTRIAL

Aunque existen ciertas diferencias entre unas plataformas y otras, la mayoría de los proveedores de servicios para la implementación de sistemas conversacionales (ver sección 8.3) manejan los términos que se describen a continuación en sus lenguajes y herramientas. Éstos difieren de

[20] *Fuente: https://aws.amazon.com/es/lex/*

los empleados habitualmente en el ámbito académico y científico ya descritos en las secciones anteriores, aunque presentan similitudes que analizaremos a continuación.

### 8.2.1. *Intents, campos y nodos*

En el capítulo 4, se describió que la gestión del diálogo se realiza en torno al concepto de acto del diálogo, una representación abstracta del propósito del turno actual del usuario (p.ej. reservar una mesa en un restaurante, confirmar una fecha para conseguir un vuelo, etc.).

En los sistemas comerciales, el propósito del turno actual del usuario se suele denominar intent (*intent*). No obstante, los intents no son simplemente una denominación de lo que pretende el usuario, sino que usualmente llevan asociadas las distintas entradas válidas del usuario, los datos relevantes que contienen (ver sección 8.2.2) y las posibles respuestas o forma de calcularlas.

La figura 11 muestra la estructura básica de un intent. Como puede observarse, el intent tiene asociadas habitualmente una lista de posibles frases del usuario que activan dicho intent. En el ejemplo de la figura, cuando el usuario dice *Dime la hora* o *¿Tienes hora?* se activa el intent *PEDIR HORA* y cuando dice *Dime tu nombre* se activa el intent *PEDIR NOMBRE*. Opcionalmente, es posible indicar qué datos relevantes se pueden extraer de la frase de entrada, en el ejemplo de pedir la hora se puede observar que se puede obtener la ciudad para la cual se quiere saber la hora. Finalmente, se indican qué posibles respuestas se pueden dar o cómo calcularlas. En el ejemplo de la figura, al pedir el nombre se puede responder con cualquiera de las frases previstas (*Me llamo Asistente Virtual*, *Soy Asistente Virtual*, etc.), mientras que al pedir la hora, se indica una lógica para calcularla (ver sección 8.2.4).

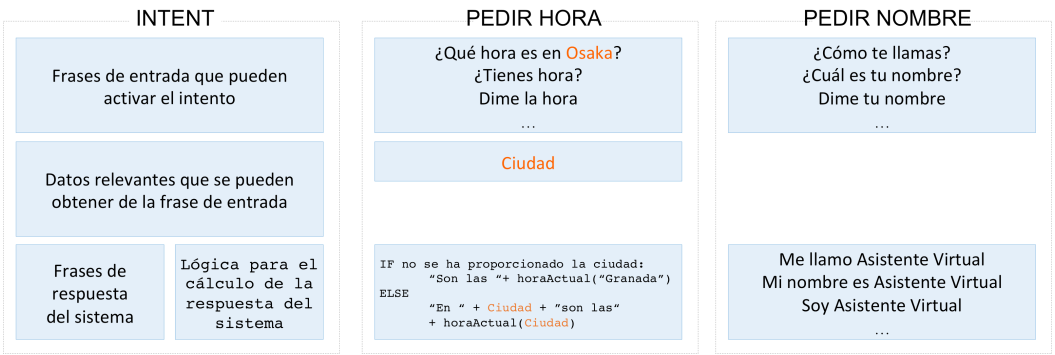


*Figura 11: Representación esquemática de la información asociada habitualmente a un intent o intención detectada por un sistema conversacional a partir de una intervención del usuario.*

Esta idea es la que siguen la mayoría de las plataformas actuales, que además proveen bibliotecas de *intents* predefinidos. Algunos ejemplos comunes son las predicciones meteorológicas, control de música y vídeo o búsqueda de comercios[21]. También es muy común tener intents predefinidos como opciones de recuperación (fall-back) cuando la entrada del usuario no casa con las frases de entrada previstas para ninguno de los intents definidos por el desarrollador[22] (p.ej. con frases como “no lo he entendido” o “¿podrías repetir?”).

Los intents son herederos del concepto de campo (*field*) de los lenguajes estándar tradicionales para el desarrollo de sistemas de diálogo como es el caso de VoiceXML[23]. La diferencia fundamental entre los intents y los campos es que los segundos están preparados para ser ejecutados de forma secuencial, de forma que su contenido básico es la frase con la que el sistema comenzará el turno y la gramática que indicará las posibles respuestas válidas del usuario y su interpretación. Así, los campos parten de una pregunta del sistema y prevén las posibles entradas válidas del usuario y cómo procesarlas, mientras que los intents se activan partiendo de una intervención del usuario e indican cómo sacar información relevante y construir la respuesta del sistema.

La interacción se gestiona por tanto como una secuencia de campos o intents. En el caso de los campos, están pensados para ejecutarse de forma secuencial. Por ejemplo en VoiceXML el algoritmo de interpretación de los campos se denomina FIA (Form Interpretation Algorithm) y visita cada campo de forma secuencial en el orden en el que se hayan indicado. Puesto que cada campo indica una pregunta del sistema y espera una cierta respuesta procesable del usuario, el diálogo es una serie de pares *[Pregunta del sistema, Respuesta del usuario]*. Existen varias formas de saltarse ese orden secuencial: incluyendo un campo inicial que permita al usuario aportar varios datos en un mismo turno, estableciendo condiciones de guarda en los campos (de forma que éstos sólo se visiten si la condición es verdadera), usando comandos tipo *goto* para iniciar un campo explícitamente o activando o desactivando campos.

En VoiceXML los campos son tratados como variables que toman el valor del dato proporcionado por el usuario como respuesta a cada pregunta del sistema. Los campos están activos mientras ese valor esté indefinido y se volverán inactivos una vez la variable tenga un valor definido (cuando el usuario haya dado una respuesta válida para ese campo o se asigne explícitamente un valor a la variable correspondiente a ese campo).

[21] *Por ejemplo con Amazon: https://developer.amazon.com/es/docs/custom-skills/built-in-intent-library.html#built-in-intent-categories*

[22] *Por ejemplo con DialogFlow: https://dialogflow.com/docs/intents/default-intents*

[23] *http://www.voicexml.org/*

En el caso de los intents, éstos se activan cuando se encuentra una frase de usuario que corresponde con una de las esperadas dentro del intent. Por tanto en principio se podrían ejecutar en cualquier orden formando finalmente una secuencia de pares *[Entrada del usuario, Respuesta del sistema]*. Esto dota de una gran flexibilidad para programar sistemas conversacionales que respondan a comandos o preguntas aisladas del usuario (p.ej. “enciende la luz”, “pon la radio”, “¿cuántos euros son 80 dólares?”), pero no para la programación de interfaces donde haya una conversación que involucre varios turnos.

Para poder abordar este último caso, las plataformas basadas en intents proveen de varias alternativas que permiten crear secuencias. Entre ellas destaca el uso de contextos que veremos en la sección 8.2.3.

### 8.2.2. *Slots y entidades*

Las entidades representan los datos relevantes para el sistema. Éstas se deben identificar procesando la entrada del usuario y se almacenan como variables con las que después se puede trabajar para estimar cuál es la mejor respuesta del sistema.

La figura 12[24] muestra un ejemplo de entidades con DialogFlow. Como puede observarse, hay dos entidades: la fecha @$sys.date$ y la ciudad @$sys.geo-city$. Así, en la frase de entrada se ha entendido que *mañana* y *el viernes* son fechas y *Sevilla* y *Madrid* ciudades.

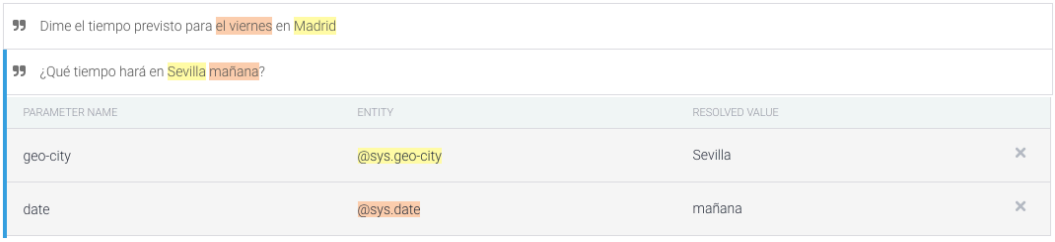


*Figura 12: Ejemplo de entidades en DialogFlow*

En el ejemplo anterior el prefijo *@sys* indica que se trata de entidades del sistema. Como ocurría con cada intents, las plataformas actuales proveen de bibliotecas de entidades comunes ya predefinidas como números, direcciones, lugares, etc. El desarrollador puede crear sus propias entidades y también suele haber opciones para crear vocabulario relacionado con la entidad mediante mecanismos de aprendizaje automático. Por ejemplo, si se ha definido la entidad *fruta* y se han introducido las palabras *manzana* y *plátano* con algunos servicios el sistema podría aprender

[24] *Fuente: https://dialogflow.com/docs/intents/actions-parameters*

que *fresa* también es una fruta si se usa en contextos similares. De igual forma, la mayoría de las plataformas permiten importar bases de datos de entidades.

Algunos servicios, como es el caso de IBM Watson[25], ofrecen también la posibilidad de usar expresiones regulares para que distintas variaciones de una misma palabra se identifiquen como la misma entidad (p.ej. *cuaderno*, *cuadernillo*, ...).

### 8.2.3. *Contextos, diálogos y subdiálogos*

Como se ha comentado anteriormente, la mayoría de las plataformas actuales basan la gestión del diálogo en el procesamiento de intents, si bien éstos en un principio no están diseñados para seguir una secuencia predefinida a priori. Por ello, cuando es preciso crear diálogos con varios turnos las plataformas basadas en intents emplean distintas alternativas.

En DialogFlow se emplean los denominados **contextos**. Cada intent puede llevar aparejados varios contextos de entrada y salida que pueden estar activos durante varios turnos. Los contextos de entrada deben estar activos en el momento en que el usuario pronuncia una frase para que ésta inicie el intent, de tal forma que actúan como las condiciones de guarda de VoiceXML. Por otra parte, los contextos de salida permiten indicar que se ha iniciado una conversación sobre una temática en particular y pueden servir para activar contextos que sean necesarios a la entrada de otros intents en los siguientes turnos de la conversación.

De esta forma, los contextos permiten ampliar el ámbito de las variables a varios turnos de forma que se puedan compartir en varios intents[26]. En Amazon Lex esta forma de compartir variables entre intents se explicita aún más mediante el establecimiento de atributos de sesión[27].

Una forma abreviada de usar contextos son los denominados *follow-up intents* de DialogFlow. Por ejemplo, un intent de reserva de hotel puede tener como frases de entrada del usuario "**quiero reservar una habitación para el viernes**" o "**quiero una doble para tres noches a partir del 3 de diciembre**". Ambos activarían el intent, pero en el primer caso habría que seguir pidiendo datos al usuario (número de personas y número de noches) manteniendo asímismo los datos que ya se tienen (la fecha de entrada). Esas preguntas se pueden implementar como *follow-up intents*, es

[25] *https://console.bluemix.net/docs/services/assistant/entities.html#defining-entities*

[26] *Más información en: https://dialogflow.com/docs/contexts/input-output-contexts#lifespan*

[27] *Más información en: https://docs.aws.amazon.com/lex/latest/dg/context-mgmt.html#context-mgmt-cross-intent*

decir, intents que se activan justo a continuación de la reserva del hotel.

En otros sistemas, el intent representa un **subdiálogo** que engloba toda la información aportada anteriormente pero con la posibilidad de que sea multiturno. Es decir, no sólo se indica la frase que dispara el intent y la respuesta a proporcionar, sino también qué datos sería necesario recopilar para proporcionar la respuesta y cómo. Este es el modelo que sigue por ejemplo Amazon Lex.

La Figura 13[28] muestra un intent llamado *ReservarRestaurante* con Amazon Lex. Todo el diálogo que se muestra en la figura se gestiona con ese intent donde se indican qué datos es necesario recopilar para llevarlo a cabo con éxito.

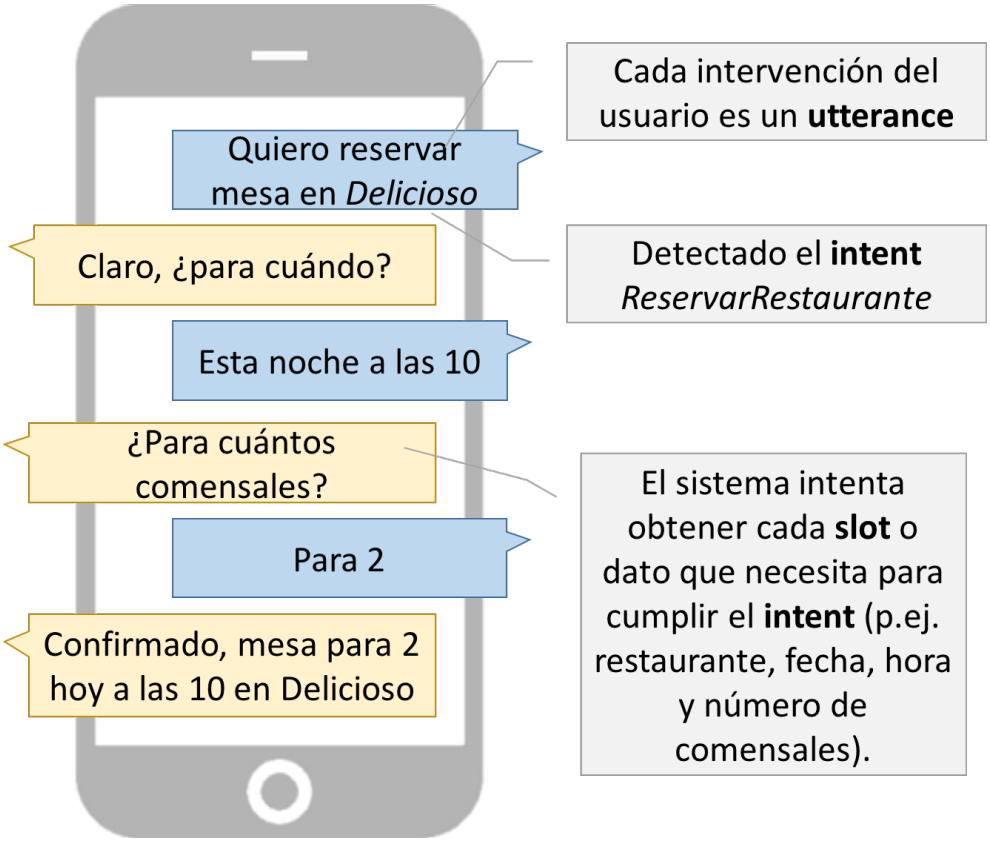


*Figura 13: Conceptos clave en un Intent con Amazon Lex*

En el caso de IBM Watson, se trabaja con el concepto de **nodo** y **diálogo**. Un diálogo es un árbol de nodos interconectados por ciertas reglas. Esta estructura hace que el flujo de la conversación sea más legible para el programador. Por ejemplo, existen entre otros nodos de tipo "event_handler", "frame" y "slot", donde un frame es un nodo con uno o más hijos de tipo slot[29]. La Figura 14[30]

[28] *Fuente: https://aws.amazon.com/es/lex/details/*

[29] *Más información en: https://console.bluemix.net/docs/services/conversation/dialog-api.html#modifying-a-dialog-using-the-api*

[30] *Fuente: https://console.bluemix.net/docs/services/conversation/dialog-overview.html#dialog-overview*

muestra un ejemplo.

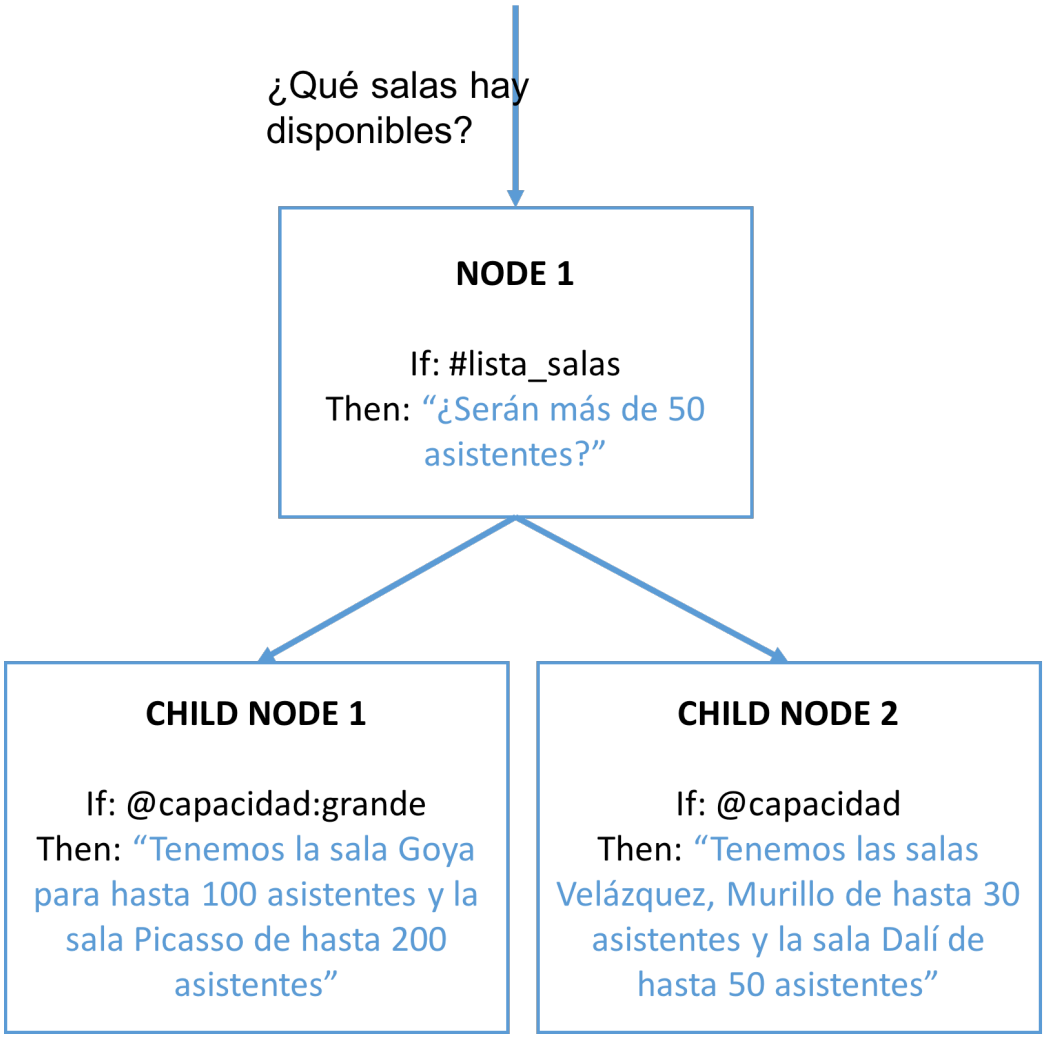


*Figura 14: Ejemplo de nodos con IBM Watson*

#### 8.2.4. *Fullfilments y acciones*

Se trata de trozos de código que contienen una lógica necesaria en el sistema para calcular la respuesta que se debe ofrecer al usuario. Por ejemplo, en un sistema de consulta meteorológica, en el **fullfilment** estaría el código que conecte y consulte al servicio meteorológico y construya la frase de salida en función de la respuesta obtenida.

Cuando es necesario invocarlos usualmente reciben el nombre del intent que los invoca con los parámetros y metadatos pertinentes (p.ej. valor de las entidades). Usualmente la comunicación se hace en un formato determinado, por ejemplo, JSON en DialogFlow. En otros casos, como IBM

Watson a este concepto se le denomina **acción**[31].

## 8.3. PROVEEDORES DE SERVICIOS Y PLATAFORMAS SOFTWARE

Las empresas relacionadas con el desarrollo de sistemas conversacionales se pueden clasificar según si sus clientes son desarrolladores, usuarios finales o ambos. Así, algunas empresas ofrecen servicios que permiten a sus clientes crear su propio sistema conversacional mientras que otras venden soluciones cerradas y funcionales, aunque usualmente entrenables o personalizables con nuevos datos proporcionados por los clientes.

En el primer caso, las empresas desarrollan uno o varios de los componentes necesarios para crear sistemas conversacionales de forma que lo que se ofrece al cliente es la herramienta que le permitirá desarrollar sus propios sistemas conversacionales de una forma sencilla, permitiéndole que se centre en dar sentido a la interacción y en el diseño de los posibles diálogos sin que se tenga que preocupar de los detalles de implementación a bajo nivel.

Usualmente, el foco de atención está en la parte de procesamiento del lenguaje y la gestión del diálogo, mientras que el reconocedor y sintetizador de habla se ofrecen como servicios que se pueden emplear como una caja negra. Así, el cliente introduce el vocabulario, su interpretación y cómo tendrá lugar el diálogo, habitualmente aportando pares de preguntas y respuestas. El cliente define el tipo de interacción que necesita creando un *agente* accesible en la web a través de una API. Estas APIs suelen permitir enviar frases y recibir respuestas que contenga una interpretación de la frase, la fase del diálogo en la que se encuentra y una respuesta en modo texto.

El modelo de negocio de estas empresas está en el pago por uso, por ejemplo, cobrando por frase procesada. En el caso de que además provean de un servicio de reconocimiento y síntesis, el cobro se puede realizar además por caracteres (p.ej. número de caracteres sintetizados). Además, como se comentó en la sección 8.1, usualmente estos servicios suelen estar en la nube y en algunos casos existen además cobros por el tiempo de computación o el almacenamiento de datos, entre otros. Más adelante en esta sección se presentan ejemplos relevantes de Amazon, Google, IBM o Microsoft.

En cuanto a las empresas que presentan productos acabados (sistemas conversacionales completos), usualmente éstos están preparados para trabajar en dominios de aplicación concretos (p.ej.

[31] *Descripción muy detallada de las acciones con Watson: https://console.bluemix.net/docs/services/conversation/dialog-actions.html#dialog-actions*

un servicio de atención al cliente) y se suelen hacer a la medida concreta de las necesidades de los clientes. En este caso, el cliente debe informar de los requisitos del sistema y el equipo de la empresa diseña el agente para cumplirlos, estando por tanto al cargo del desarrollo completo del sistema. En algunos casos, la solución presentada al cliente está dotada de capacidad de aprendizaje de forma que el cliente pueda valorar los diálogos y que el sistema aprenda a evitar los valorados negativamente y producir un mayor porcentaje de diálogos exitosos.

A continuación, describimos algunas de las que según Gartner son las empresas más relevantes en plataformas conversacionales en 2018 [210], explicando los servicios que facilitan.

### 8.3.1. *Los grandes actores*

Entre los grandes proveedores de servicios relacionados con los sistemas conversacionales, está extendida la filosofía de uso de intents, entidades, contextos y fullfilments descrita en la Sección 8.2. Cabe destacar que, a pesar de que estas tecnologías se suelen presentar como grandes innovaciones, es difícil usualmente generar diálogos complejos en estas plataformas sin que ello requiera de mucha programación adicional. Es por esto que en los últimos tiempos, algunas de estas empresas están acudiendo a la comunidad científica para intentar integrar sus últimos avances (por ejemplo, la sesión especial Speech and Language Technology for Alexa Conversational AI organizada en el congreso Interspeech 2019).A continuación, mencionamos los más relevantes.

**Amazon Web Services (AWS)** integra *Amazon Lex*, un servicio que permite crear interfaces conversacionales con entrada y salida tanto oral como textual. Presenta APIs adaptadas para el desarrollo de aplicaciones en dispositivos móviles, web y de escritorio en distintos sistemas operativos y cobra a los desarrolladores a razón del número de turnos de usuario procesados y con distintos precios según consista en procesar una entrada oral (más cara) o escrita.

Amazon Lex está basado en las tecnologías de Amazon Alexa para la comprensión del lenguaje natural y para la interacción oral puede conectarse con Amazon Polly (el servicio TTS de Amazon). Además, Lex está integrado con otros servicios de forma que éste puede emplearse para consultar de forma eficiente bases de datos y servicios externos como Skype, ASWS Lambda y API Gateway para realizar todo tipo de aplicaciones, por ejemplo, de control domótico e Internet de las Cosas. Todos estos servicios tienen también un modelo de pago por uso. Por ejemplo, Amazon Polly se cobra actualmente a 4,00 USD por millón de caracteres sintetizados.

En la web de Amazon AWS[32] pueden verse vídeos acerca de cómo implementar este tipo de interacciones siguiendo la estructura de intents y entidades descrita en la sección 8.2. **Google**[33] basa la interacción conversacional en DialogFlow cuyos diálogos se construyen en base a intents, entidades, contextos y fulfillments.

**Facebook**[34] provee la platafoma Messenger donde se pueden crear conversaciones en modo texto. Estas conversaciones pueden incluir texto, los denominados activos (audio, vídeos imagen y archivo) y botones y respuestas rápidas. Las respuestas del sistema se pueden crear usando plantillas. El procesamiento del lenguaje se hace con un sistema integrado basado en el reconocimiento de entidades que se puede extender con wit.ai, que también se usa para la gestión de la interacción. Esta gestión también está basada en entidades.

**Microsoft Bot Framework**[35] tiene una filosofía parecida a la de Amazon, donde se pueden conjugar muchos otros servicios de Microsoft. El Azure Bot Service[36] permite crear chabots que tienen gran capacidad de comprension del habla (usando LUIS) y permiten entrada y salida oral. Además, se pueden unir a los Microsoft Cognitive Services para hacer diálogos más sofisticados

**IBM**. Para crear sistemas conversacionales con tecnología IBM, se ofrece la tecnología IBM Watson Assistant[37]. Como se ha descrito en la sección 8.2, con Watson los diálogos se crean en base a nodos a los que se llega a través de las entradas del usuario[38]. Dentro de los nodos se pueden definir “slots”, que son los datos que se requiere recoger para completar la interacción en sistemas orientados a la tarea[39]. **Oracle**[40] también presenta una solución para generar sistemas de diálogo que incluye estados, intents y entidades. En su caso además permiten indicar explícitamente el flujo del diálogo.

[32] *https://aws.amazon.com/es/lex/*

[33] *https://developers.google.com/actions/dialogflow*

[34] *https://developers.facebook.com/docs/messenger-platform/built-in-nlp*

[35] *https://dev.botframework.com*

[36] *https://docs.microsoft.com/en-us/azure/bot-service/bot-service-overview-introduction?view=azure-bot-service-3.0*

[37] *https://www.ibm.com/watson/ai-assistant/*

[38] *https://console.bluemix.net/docs/services/conversation/dialog-overview.html#dialog-overview*

[39] *https://console.bluemix.net/docs/services/conversation/dialog-slots.html#dialog-slots*

[40] *http://www.oracle.com/us/technologies/mobile/chatbots-primer-3899595.pdf*

### 8.3.2. *Lekta.ai: Un framework industrial para el desarrollo de sistemas conversacionales híbridos*

El desarrollo de sistemas conversacionales es actualmente una de las aplicaciones más complejas y exigentes de entre el conjunto de ámbitos de investigación y desarrollo de las tecnologías del lenguaje y la inteligencia artificial.

La plataforma Lekta2 fue desarrollada inicialmente por Jose F Quesada (Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Sevilla) entre 2013 y 2016, siendo utilizado en el desarrollo de una aplicación (ELEna) para la enseñanza de idiomas. Posteriormente se ha convertido en el núcleo funcional de la empresa Lekta.ai[41], con sedes actuales en España y Polonia.

Lekta se puede describir resumidamente como un framework tecnológico para el diseño de sistemas conversacionales (sistemas de diálogo tanto escrito como hablado). Utiliza un enfoque híbrido en el ámbito de las tecnologías del lenguaje, permitiendo la implementación de modelos de diálogo que integren distintas características utilizadas en modelado de diálogo, tales como enfoques basados en scripts conversacionales, basados en agenda o utilizando estrategias del enfoque ISU (*Information State Update approach*).

**La arquitectura Lekta.** El núcleo tecnológico de Lekta está basada en un modelo que integra funcionalmente varios módulos, cada uno de ellos diseñado para una parte específica del diálogo. Estos módulos están interconectados entre sí para cubrir una conversación desde el reconocimiento de la expresión del usuario hasta la generación de la respuesta del sistema.

**MindBoard.** El módulo Mindboard asume la gestión de la memoria del sistema. Este módulo está conectado a todos los demás de la arquitectura, ya que la mayoría de ellos pueden enviar o recuperar información desde el Mindboard en cualquier momento y cambiar el valor de los campos. Este es un módulo fundamental capaz de mantener un contexto para el diálogo, ya que puede recordar"lo que el usuario ha dicho. En el rol de una memoria a corto y largo plazo, permite lidiar con múltiples cambios de parámetros o tareas en medio de una conversación que genera diálogos altamente flexibles. Mediante un lenguaje de alto nivel es posible especificar el modelo de memoria de MindBoard, permitiendo estructuras complejas de datos para modelar la conversación, la propia información relativa al usuario, los intercambios con el backoffice, así como toda la información relevante para la generación de eventos, etc.

[41] *https://lekta.ai/*

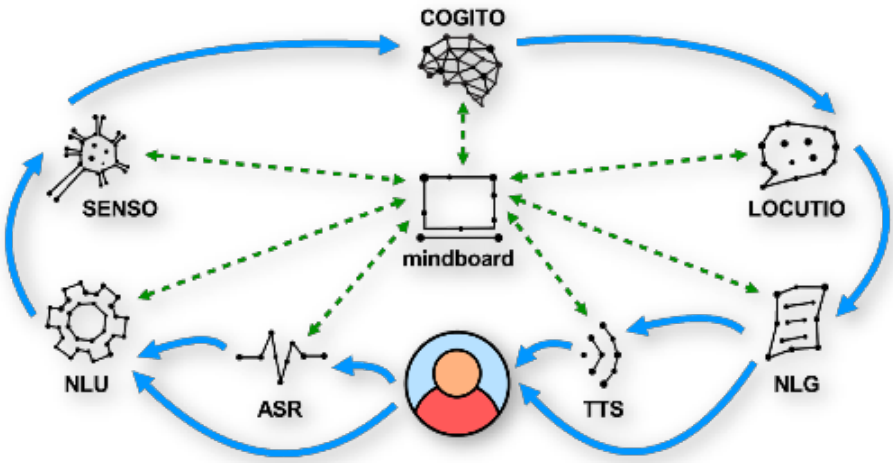


*Figura 15: Arquitectura Lekta*

**NLU.** Este módulo se centra en la comprensión del lenguaje. Su objetivo principal es convertir las preferencias de los usuarios en estructuras sintácticas y semánticas. Las capacidades del módulo van más allá del simple reconocimiento de intención / entidad y aplican un modelo híbrido para la comprensión y el procesamiento del lenguaje. Aquí coexiste una amplia gama de técnicas de PNL, y todos los resultados posibles se pasan a los siguientes módulos para mejorar la comprensión aplicando distintas heurísticas, el nivel de análisis pragmático y el contexto e historia completa de la conversación.

**Senso.** El módulo Senso se corresponde con la parte sensorial implicada en el uso del lenguaje. Este módulo funciona como un filtro tanto para integrar distintas modalidades de interacción como para analizar a un nivel superior la información recibida desde el módulo de comprensión.

Consecuentemente, podemos usar las estrategias de esta etapa para inicializar el contexto y las estructuras de datos usadas en MindBoard. Este módulo también representa la proactividad del sistema, ya que puede interactuar con el usuario sin una interacción previa del usuario, por ejemplo, si el usuario ha configurado una alarma en el momento en que se activan las alarmas, Senso puede configurarse para recibir como entrada la señal o evento disparado por la alarma, iniciando a continuación la conversación con el usuario.

**Cogito.** Cogito es el módulo más importante en la arquitectura, ya que es donde se encuentra el gestor de diálogo (DM o Dialogue Manager). El DM se ocupa de las decisiones que debe tomar el sistema para obtener una respuesta e interacción adecuadas con el usuario. El módulo Cogito es el cerebro del sistema. Simula el pensamiento humano, la pragmática dependiente del contexto

y el razonamiento lógico. Decide estrategias de diálogo y dimensiones de diálogo.

El módulo Cogito de Lekta permite diseños complejos de estrategias de diálogo de negociación. Este es el punto clave para establecer la robustez y flexibilidad del sistema, ya que, por un lado, gestiona el enfoque de Lekta en la toma de decisiones, teniendo en cuenta todo lo que los módulos han enviado a Mindboard.

**Locutio.** El módulo Locutio se encarga de estructurar y ordenar los nodos de alto nivel general definidos por Cogito como respuesta al usuario. Es importante tener en cuenta que el resultado de la fase de comprensión (NLU) se representa en un modelo de representación semántica abstracta, que es utilizado incluso por Cogito para generar la respuesta. Estos nodos de alto nivel son finalmente analizados por Locutio, antes de enviar al módulo NLG para realizar la generación final del mensaje.

**NLG.** Este módulo se corresponde con la parte motora del uso del lenguaje. Como resultado de ese proceso, se crea la respuesta final que se comunicará al usuario.

La salida generada por el módulo NLG se puede dividir en varios canales dependiendo del propósito, puede generar un texto con enlaces a páginas web para pantallas de chat, o puede generar texto marcado con características prosódicas para ayudar a la pronunciación correcta del TTS.

**Principales componentes y características funcionales.** Resumiendo, Lekta proporciona un marco completo que integra los siguientes componentes funcionales principales, metodologías y marcos:

- Entendimiento del lenguaje natural,
- Interpretación y Generación del Lenguaje Natural,
- Integración sensorial,
- Modelado de usuario,
- Adaptación a tareas y dominios,
- Modelado y gestión avanzada del diálogo,
- Componentes funcionales basados en Inteligencia Artificial y Aprendizaje Automático,
- Desarrollo tecnológico y marcos de despliegue.

**Entendimiento del lenguaje natural.**

- Estrategias híbridas basadas en modelos formales y conocimiento de la tarea, mejoradas mediante técnicas basadas en aprendizaje automático para NLU,
- La comprensión de la intención, objetivos y parámetros obtenidos durante una conversación,
- Esquema de anotación de orientación semántica,
- Anotaciones semánticas y pragmáticas,
- Arquitectura nativa multilingüe,
- El módulo ParlanceSensor de Lekta detecta el cambio del idioma de entrada incluso durante una conversación, y adapta automáticamente los módulos de comprensión y generación,
- Comprensión, anotación y representación semántica de fenómenos lingüísticos complejos.
    - Interacción social,
    - Orden flexible de comandos, parámetros,
    - Tareas / dominios múltiples simultáneos,
    - Repeticiones, disfluencias, construcciones idiomáticas,
    - Corrección ortográfica,
    - Ambigüedad, polisemia, fenómenos lógicos (negación, disyunción, etc.).

**Interpretación y generación del lenguaje natural.**

- Extracción eficiente y representación de contenidos semánticos,
- Entendimiento y fusión multimodal,
- Lekta proporciona soporte para el reconocimiento y la síntesis del habla emocional multilingüe,
- Interpretación enriquecida, teniendo en cuenta diferentes factores como:
    - Estado de diálogo actual,
    - Preguntas activas (Question under discussion) y expectativas,

- Historia del diálogo,
- Perfil y modelo de usuario

**Integración sensorial.**

- Fusión sensorial multimodal
  - Una conversación requiere la integración de múltiples fuentes de información. El módulo Senso de Lekta permite una integración en tiempo real de toda esta información,
  - Síntesis sensorial multimodal.

**Modelado de usuario.**

- Enfoque centrado en la experiencia del usuario,
- Modelado humano cognitivo (basado en modelos cognitivos del comportamiento interactivo humano).

**Tarea y adaptación de dominio.**

- Adaptación de dominio flexible,
- Adaptación entre dominios e idiomas,
- Reutilización de módulos entre tareas y dominios,
- Enfoque general de dominio abierto (Open domain).

**Modelado y gestión avanzada del diálogo.**

- Gestión paralela del diálogo multitarea,
- Gestión del diálogo multilingüe,
- Estrategias de diálogo colaborativo,
- Resolución de tareas compartidas,
- Diálogo multilingüe y resolución de problemas,

- Robustez,
- Sistemas de diálogo orientados a negociaciones complejas,
- Sub-diálogos,
- Reutilización de datos durante una conversación y en diferentes diálogos,
- Integración de back-office mediante el uso de un administrador de recursos genérico,
- Soporte avanzado para fenómenos de diálogo complejos:
    - Peticiones / comandos incompletos,
    - Parámetros incoherentes,
    - Validación de parámetros,
    - Múltiples tareas simultáneas,
    - Control de prioridades entre tareas simultáneas.
    - Detección de parámetros durante múltiples iteraciones (para parámetros especialmente largos y complejos como identificadores, ...)
    - Diferentes estrategias de confirmación.

**Inteligencia Artificial y Aprendizaje Automático.**

- Capacidades de razonamiento:
    - Gestión colaborativa y proactiva del diálogo,
    - Acomodación dinámica de diferentes estrategias de diálogo.
- Entendimiento del lenguaje natural mejorado mediante técnicas de aprendizaje automático,
- Estrategias de diálogo declarativo.
    - Gestión del diálogo híbrido,
    - Business intelligence,
    - Patrones conversacionales.

**Desarrollo tecnológico y marcos de despliegue.**

- Integración entre diversas tecnologías multimodales,
- Integración a través de diferentes medios y sensores,
- Métricas de evaluación,
- Escalabilidad,
- Kernel en tiempo real.

### 8.3.3. *Herramientas drag and drop*

Con los primeros entornos de desarrollo aparecieron herramientas que seguían una filosofía *drag and drop*, en la que el programador simplemente arrastraba cajas a un panel y las conectaba de forma gráfica. Cada una de esas cajas representaba un campo con un dato que había que pedir al usuario y las líneas que las unían representaban la secuencia en la que se iban a visitar dichos campos. Este es el caso del extinto CSLU Toolkit [211].

En la actualidad, algunas empresas dan servicios para la generación de diálogos siguiendo esta filosofía en la que se intenta hacer que el desarrollo sea lo más sencillo posible ocultando los detalles de programación a más bajo nivel.

La diferencia con las herramientas clásicas que existían hace décadas es que las nuevas herramientas explotan la potencialidad proporcionada por la nube. Por lo general, permiten incluir funciones y usar servicios de terceros para la comunicación con el usuario (p.ej. Facebook) e incluso para la propia gestión de la interacción (p.ej. DialogFlow). También pueden ligarse a servicios como Alexa o Google Home y realizar aprendizaje automático beneficiándose de lo aprendido con cada nueva interacción.

Por ejemplo, Conversable[42] es una plataforma para desarrollar chatbots que incluye el editor ICE (Interactive Conversation Editor) que permite diseñar el flujo del diálogo con un modelo drag&drop.

**Gupshup**[43] también provee de una interfaz gráfica con la que diseñar conversaciones para los no programadores y un IDE con el que realizar un diseño más detallado para los programadores.

[42] *http://conversable.com/platform/*
[43] *https://www.gupshup.io/developer/bot-builder-tools*

**Kore.ai**[44] presenta otro editor gráfico, cada elemento que se arrastra a la pantalla contiene los conceptos de entidad e intent discutidos con anterioridad. En la mayoría de los servicios descritos es posible comprobar sobre la marcha el diálogo que se está diseñando mediante una ventana tipo chat.

### 8.3.4. *Otras soluciones*

Como alternativa a los grandes actores, **Rasa**[45] propone una solución de código abierto. En este caso el desarrollador tiene que ser más explícito que en otras plataformas a la hora de especificar cómo se realiza el procesamiento del lenguaje natural indicando la representación semántica de múltiples frases. Por otra parte, algunas empresas ofrecen soluciones ya creadas siguiendo los requisitos planteados por el cliente. Por ejemplo **Nuance**, que es una empresa muy importante en el sector de las tecnologías del habla, ofrece soluciones preparadas para entornos específicos[46] (relación con el cliente, automoción, sanidad...), pero no provee de herramientas para que los desarrolladores implementen sus propios sistemas de diálogo.

Otros proveedores interesantes son:

- **Artificial Solutions** ofrece la plataforma Teneo[47].
- **Creative Virtual** ofrece V-Person[48].
- **OneReach**[49], una solución para evitar humanos en el call center.
- **SmartBotHub**[50] da soluciones ya hechas en distintos sectores.

### 8.3.5. *Ámbito académico*

En el ámbito académico existen distintas herramientas disponibles para resolver problemas específicos. Por ejemplo, la Universidad de Cambridge publicó un toolkit de código abierto denominado **PyDial**[51] para el desarrolo de sistemas de diálogo con gestión estadística del diálogo. El sistema

[44] *https://info.kore.ai/bot-building-simplified*
[45] *https://rasa.com*
[46] *https://www.nuance.com/omni-channel-customer-engagement.html*
[47] *https://www.artificial-solutions.com*
[48] *https://www.creativevirtual.com/solutions/*
[49] *https://onereach.com/action-desk*
[50] *https://smartbothub.com/solutions/*
[51] *http://www.camdial.org/pydial/*

provee de modelos independientes para cada una de las fases (semantic parsing, belief tracking, dialog policy estimation...), que se pueden sustituir por el código desarrollado por el usuario.

Para el desarrollo de un sistema de este tipo hay que tener conocimiento acerca de cada una de estas fases y la programación se debe hacer en Python. La ventaja que aporta frente a las soluciones comerciales comentadas estriba es su valor científico al posibilitar un entorno en el que comparar diferentes algoritmos en condiciones similares. De hecho, también contiene un benchmark, lo que hace que las contribuciones científicas en el área sean más reproducibles y fácilmente comparables.

Otra iniciativa reseñable es **OpenDial**[52], de la Universidad de Oslo. En este caso desarrollada como una pizarra con módulos conectados, de forma que se puedan incluir distintos servicios para el reconocimiento del habla, etc. El énfasis de la herramienta es en la gestión del diálogo, que en este caso se puede hacer tanto de forma estadística como incluyendo reglas y conocimiento experto.

# 9. CORPUS DE DATOS Y EVALUACIÓN

Este capítulo se centra en la presentación de los corpus de datos disponibles tanto para tareas de entrenamiento y modelado de sistemas de diálogo como para tareas y retos (challenges) de evaluación y control de calidad.

## 9.1. PRINCIPALES BASES DE RECURSOS

En esta sección recopilamos los principales repositorios de datos relacionados con las tecnologías del lenguaje en el ámbito de los sistemas conversacionales, así como los principales retos organizados utilizando dichos datos.

### 9.1.1. *Repositorios de datos*

La principal iniciativa para la compartición de datos relacionados con tecnologías del lenguaje en Europa es la asociación europea de recursos de lenguaje (European Language Resources Association, ELRA[53]). Entre sus misiones está la identificación de recursos interesantes, la promoción,

[52] *http://www.opendial-toolkit.net/home*
[53] *http://www.elra.info*

producción y validación de los mismos, así como su evaluación, distribución y estandarización.

En Estados Unidos es de gran relevancia el repositorio del Linguistic Data Consortium[54] (LDC), que también se encarga de favorecer la creación y compartición de recursos lingüísticos (datos, herramientas y estándares).

Estos catálogos están centrados principalmente en datos de carácter lingüístico o acústico como pueden ser corpus de textos, grabaciones de audio en distintos idiomas, etc. Aunque no es su foco principal, también contienen corpus de diálogos, como por ejemplo los paradigmáticos ATIS, DARPA Communicator o TRAINS. Así, en el LCD se puede filtrar la búsqueda[55] por aplicaciones recomendadas *spoken dialogue modeling* y *spoken dialogue systems* y en el LRE Map[56] de ELRA se puede filtrar por modalidad *speech* y uso del recurso *dialogue*. No obstante, ambas búsquedas devuelven un número de recursos limitados frente al tamaño de los catálogos (p.ej. en ELRA solo 30 recursos son de diálogo frente a los más de 6.000 catalogados).

Estos recursos son de gran valor para el desarrollo de sistemas de diálogo pues para la implementación de la gestión del diálogo o sistemas extremo-a-extremo (end-to-end) es necesario contar con corpus de conversaciones donde se puedan observar fenómenos de alto nivel como los actos de diálogo empleados, la toma de turnos y fenómenos como las confirmaciones, negaciones, etc.

Con el auge de los enfoques basados en aprendizaje automático para el desarrollo de los sistemas conversacionales, ha aumentado la necesidad de compilar y compartir este tipo de corpus de diálogos. En [212] se puede encontrar un amplio estudio de los corpus existentes actualmente para el aprendizaje de sistemas de diálogo, donde se discuten además sus principales características y las tareas para las que pueden emplearse. Remitimos a los lectores interesados a dicho estudio para una lista detallada de corpus humano-humano y humano-máquina.

Estos mismos autores mantienen una versión online en github[57] a la que se puede contribuir con nuevos corpus y donde se pueden encontrar enlaces para la descarga de los corpus listados, así como información acerca de su dominio de aplicación, tipo de datos (oral o escrito), número de turnos, número de diálogos y una pequeña descripción.

[54] *https://catalog.ldc.upenn.edu/*
[55] *https://catalog.ldc.upenn.edu/search*
[56] *http://lremap.elra.info*
[57] *https://breakend.github.io/DialogDatasets/*

### 9.1.2. *Limitaciones derivadas de la inexistencia de un estándar anotación común*

A continuación, quisiéramos destacar algunas cuestiones comunes a los corpus y repositorios comentados. En primer lugar, el que haya una mayor cantidad de información disponible permite que los resultados de investigación sean más replicables, así como compartir los enfoques y algoritmos empleados para el entrenamiento. Sin embargo, existe una carencia de esquemas de anotación comunes que permitan sacar el máximo partido al uso de estos corpus. En el Capítulo 6 se argumentó la necesidad del uso de estándares. Este reto se está abordando de diversas formas. Por ejemplo, algunos autores han elaborado herramientas para convertir las anotaciones de los corpus disponibles a un formato estándar [213].

### 9.1.3. *Challenges en el sector de los sistemas conversacionales*

Por otra parte, el auge de los métodos de aprendizaje automático ha fomentado la aparición de retos (challenges), en ellos, los participantes deben trabajar con el mismo corpus para entrenamiento, generar un sistema o parte de un sistema y evaluarlo con otro corpus proporcionado. Esto hace que los resultados obtenidos sean directamente comparables y se pueda avanzar en la identificación de los mejores métodos para resolver tareas concretas.

En el ámbito de los sistemas de diálogo destacan los retos *Dialog State Tracking Challenges*, corpora[58] con diversas tareas relacionadas con predecir el estado del diaĺogo, el aprendizaje extremo a extremo de la gestión del diálogo, modelado de la conversación y predicción de errores fatales en la comunicación (dialogue breakdown). Los datos empleados para el entrenamiento y la evaluación en los distintos retos han sido liberados para ser utilizados por la comunidad.

Por otra parte, también han aparecido recientemente retos promovidos por la industria. Por ejemplo, los concursos Alexa Prize en los que en cada año se propone un reto para el desarrollo de un sistema conversacional con la tecnología de Amazon[59] y diversos equipos pueden aplicar para ser seleccionados y subvencionados para participar [214].

Este concurso, así como otros retos recientes como el *Conversational Intelligence Challenge* han generado corpus de interacciones de usuarios con chatbots, algunas con sistemas sociales no orientados a la tarea. Entre este tipo de corpus se encuentran el *Common Alexa Prize Chats* y los

[58] *https://www.microsoft.com/en-us/research/event/dialog-state-tracking-challenge/*
[59] *https://developer.amazon.com/alexaprize*

listados en el repositorio de ParlAI[60].

#### 9.1.4. *Generación de nuevos datos*

Dada la necesidad de datos en distintos dominios de aplicación, muchos equipos de investigación están optando por medidas de crowdsourcing para la generación de corpus. Esto se suele hacer empleado plataformas que permiten pagar pequeñas cantidades de dinero a sujetos para que realicen las tareas estipuladas (p.ej. Amazon Mechanical Turk). En estas plataformas es posible establecer el perfil del trabajador que realizará la tarea, aunque se pierde mucho control sobre los sujetos frente a la grabación en laboratorio.

Por esto, el uso de crowdsourcing exige prestar atención a medidas para verificar la calidad del trabajo realizado (p.ej. la calidad de anotaciones de actos del diálogo). Para ello, los diferentes equipos que utilizan este método de recolección de datos han compartido diversas estrategias y protocolos a seguir, que incluyen pautas relativas el tiempo dedicado a la tarea, la consistencia de un mismo sujeto, el acuerdo entre sujetos y heurísticas de diversa índole [215]. También existen medidas en el ámbito industrial.

Por ejemplo, Mozilla ha iniciado el proyecto de código abierto **Common Voice**[61] con el objetivo de generar el mayor repositorio de datos vocales mundial. La idea es tener una gran colección con la que poder entrenar sistemas de reconocimiento y síntesis de habla sin depender de las grandes corporaciones. Además, se basa en un sistema colaborativo para recabar voces también de idiomas minoritarios[62]. Los datos generados se compartirán con licencia CC0 y se están usando actualmente en los proyectos de Mozilla de reconocimiento y síntesis de habla.

### 9.2. ENFOQUES PARA LA EVALUACIÓN DE LOS SISTEMAS CONVERSACIONALES

A medida que se producen avances en el estudio y desarrollo de sistemas de diálogo, se hace necesario desarrollar nuevas medidas de evaluación para comprobar si estos sistemas son efectivos o no. La tarea de fijar nuevas medidas, es decir, de plantear nuevos procedimientos y medidas que sean aceptadas unívocamente por la comunidad científica para la evaluación de este tipo de sistemas presenta muchas dificultades. Puede considerarse que el campo de las técnicas y medidas

[60] *http://parl.ai/*
[61] *https://voice.mozilla.org*
[62] *https://voice.mozilla.org/en/record*

de evaluación de este tipo de sistemas se encuentra en una fase inicial de desarrollo.

Se han desarrollado diferentes iniciativas para definir marcos generales en los que englobar el diseño y evaluación de sistemas de diálogo. Algunos ejemplos en Europa de los primeros intents en esta dirección son EAGLES (Expert Advisory Group on Language Engineering Standards) 96, ELSE99 y DISC99.

Otras instituciones europeas centradas en el estudio y definición de técnicas de evaluación son: COSCODA (*Coordinating Committee on Speech Databases and Speech I/O Systems*), dedicada a aspectos relacionados con la creación de bases de datos multilingüe, ELRA (*European Language Resources Association*), centrada en la colección y distribución de recursos lingüísticos, SQUALE (*Speech Recognition Quality Assessment for Linguistic Engineering*) centrada en la adaptación a contextos multilingües del paradigma LVCSR (*Large Vocabulary Continuous Speech Recognition*) de ARPA.

En la bibliografía sobre evaluación de sistemas puede encontrarse un gran número de tipos de evaluación. A continuación, se resumen diferentes criterios reseñados en [216]:

- Teniendo en cuenta el tipo de medidas utilizadas para la evaluación, ésta puede clasificarse como objetiva o subjetiva, dependiendo del número de medidas que predominen:
  - Medidas objetivas: Son aquellas que se obtienen directamente del funcionamiento del sistema, no incluyendo ningún tipo de valoración subjetiva por parte de desarrolladores o usuarios del sistema.
  - Medidas subjetivas: Denotan un proceso de valoración subjetiva, normalmente llevado a cabo por los usuarios finales del sistema. Un ejemplo de evaluación utilizando este tipo de medidas es la definida en el proyecto europeo Trindi.

  Estas medidas también pueden clasificarse teniendo en cuenta la manera de realizar su cómputo (medidas automáticas o manuales) o su influencia en la calidad global del sistema (mediadas positivas o negativas)
- Teniendo en cuenta el objetivo de la evaluación, puede distinguirse entre:
  - Evaluación general: Se analiza el funcionamiento global del sistema teniendo en cuenta entradas y salidas a nivel general.
  - Evaluación por componentes: Se analiza de forma separada el funcionamiento de cada

uno de los módulos del sistema, teniendo en cuenta las entradas y salidas parciales del mismo.

- Según el patrón de referencia tomado una vez realizada la evaluación del sistema, puede distinguirse entre:
  - Evaluación comparativa. Se realizan diferentes sistemas en paralelo con las mismas especificaciones, pero desarrollados por centros diferentes. Esta evaluación se ha utilizado usualmente en proyectos financiados por DARPA, como DARPA Communicator.
  - Evaluación temporal. El patrón de referencia es el propio sistema desarrollado, realizando comparaciones de su funcionamiento en las diversas fases temporales de su desarrollo.
  - Evaluación sustitutiva. Se evalúa el sistema con respecto a otro con las mismas funcionalidades desarrollado previamente, normalmente con diferente tecnología.
  - Evaluación inicial. Una alternativa cuando no hay disponible un sistema de referencia con el que comparar consiste en la estimación a priori de su funcionamiento durante la fase de especificaciones y la posterior evaluación de la desviación con respecto al comportamiento previsto.

A la hora de evaluar un sistema conversacional de manera global, la propuesta con mayor repercusión a nivel internacional es PARADISE [217] [218]. El modelo PARADISE (*PARAdigm for DIalogue System Evaluation*) combina diferentes medidas en una única función que mide el rendimiento del sistema, en correlación directa con la satisfacción de los usuarios del sistema. Los supuestos principales del modelo PARADISE son:

- el objetivo a maximizar es la satisfacción del usuario;
- el éxito en la tarea y varios costes asociados a la interacción (medidas objetivas) pueden usarse para predecir la satisfacción del usuario.

Estos dos supuestos pueden formalizarse en la siguiente ecuación:

$$Satisfaccion\ del\ usuario = (\alpha N(Exito\ de\ la\ tarea)) - \sum_{i=1}^{N} w_i N(Costes\ del\ dialogo)$$

donde las medidas del éxito de la tarea y los costes del diálogo se utilizan normalizando su distribución $N()$ a una distribuciónnormal de media cero y varianza unidad.

Esta formulación se basa en el modelo mostrado en la Figura 16. La maximización de la satisfacción del usuario se lleva a cabo minimizando los costes del diálogo y maximizando el éxito de la tarea. Los costes del diálogo se cuantifican mediante medidas de eficiencia y de calidad.

La utilización del modelo de PARADISE requiere de un corpus de diálogos obtenidos en experimentos controlados en los que los diferentes usuarios evalúan subjetivamente su satisfacción sobre una escala.

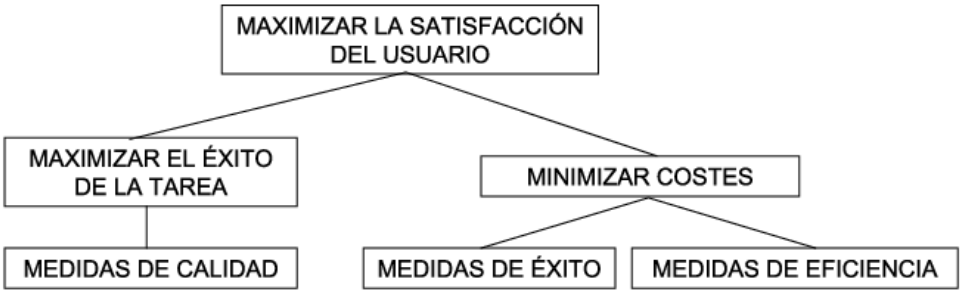


*Figura 16: Modelo de desarrollo del paradigma PARADISE*

Las medidas de éxito de la tarea más comúnmente utilizadas son:

- El Factor Kappa (K) se propuso en la formulación inicial del modelo PARADISE. Se calcula a partir de una matriz de confusión de los valores de atributos intercambiados entre el usuario y el sistema, de forma que la diagonal principal de la matriz indica los casos en el que el sistema reconoce y comprende correctamente la información del turno(s) de usuario. Se utiliza la siguiente expresión:

  $$K = \frac{P(A) - P(E)}{1 - P(E)}$$

  donde $P(A)$ es la probabilidad de que el sistema reconozca y comprenda correctamente la información aportada por el usuario; y $P(E)$ es una estimación de las veces que el sistema podría acertar por casualidad, calculada mediante:

  $$P(E) = \sum_{i=1}^{n} \left(\frac{t_i}{T}\right)^2$$

  donde $t_i$ es la suma de intents en la columna $i$ y $T$ es la suma total de intents.

- Tasa de tarea completada: Es el porcentaje de veces que el sistema satisface correctamente la solicitud realizada por el usuario. Para obtener esta tasa, es necesario revisar manualmente los diálogos transcritos y observar los casos en los que el sistema responde correctamente

Las medidas de eficiencia más importantes son:

- Tiempo medio necesario para completar una tarea (los diferentes tiempos para cada tarea pueden promediarse mediante diferentes criterios: considerar todos los diálogos o sólo aquellos que han terminado con éxito, etc.).
- Tiempo medio por turno (se pueden aplicar los mismos criterios que en el caso anterior para realizar los promedios).
- Número medio de turnos por tarea (mismos promedios que en los dos casos anteriores).
- Mínimo número de turnos o tiempo mínimo para completar una tarea (sirven para establecer una medida de la complejidad del sistema de diálogo independientemente de las pruebas realizadas por los usuarios).
- Tipos de confirmaciones de datos utilizadas (un estudio del número de confirmaciones y de su tipo, explícitas o implícitas, revela el funcionamiento del sistema).
- Número de palabras reconocidas correctamente por turno (a mayor cantidad de palabras correctamente reconocidas, lógicamente, mayor será la calidad del sistema y menor número de turnos se necesitarán para completar la tarea).

Entre las medidas de calidad del sistema, cabe referenciar:

- Tasa de reconocimiento (porcentaje de palabras reconocidas correctamente).
- Tasa de conceptos semánticos correctos (porcentaje de conceptos semánticos correctamente generados por el módulo de comprensión).
- Porcentaje de errores corregidos con éxito (medida de la eficiencia de las técnicas utilizadas para la detección y corrección de errores).
- Tiempo de respuesta del sistema (tiempo que tarda el sistema en reconocer y comprender la frase pronunciada por el usuario).
- Tiempo de respuesta del usuario (pueden detectarse las situaciones del diálogo en los que el usuario tarda en proporcionar una respuesta).
- Número de veces que el usuario no contesta (medida del número de veces que el usuario supera el tiempo máximo establecido para realizar una intervención).

- Número de veces que el usuario solicita repetición (indica el número de veces que el usuario no entiende la información formulada por el sistema).
- Número de veces que el usuario solicita ayuda (indica el entendimiento que tiene el usuario sobre el funcionamiento del sistema).
- Número de veces que el usuario interrumpe al sistema (puede revelar el grado de destreza o entrenamiento que posee el usuario).

El modelo PARADISE se presenta como un marco general para la evaluación de los agentes de diálogo, dado que permite desacoplar los requisitos de la tarea y los comportamientos de los agentes, comparar estrategias de diálogo, calcular el rendimiento de diálogos completos y de subdiálogos, especificar las contribuciones relativas al rendimiento, y comparar agentes que aborden diferentes tareas, mediante una normalización de la complejidad de la tarea. En cuanto a sus inconvenientes, cabe destacar:

- excesivo acoplamiento entre satisfacción de usuario yusabilidad;
- dudas acerca de la posibilidad de predecir la satisfacción de usuario a partir de la información registrada en los ficheros log del sistema;
- dificultad en la interpretación de los cuestionarios, y ausencia de fundamentos teóricos acerca de las cuestiones aincluir en los mismos;
- uso del modelo, por el momento, limitado a experimentos controlados (y no con usuarios reales).

El método de evaluación utilizado por diversos grupos de investigación (por ejemplo, los adscritos al programa ARPA Spoken Language Systems en el dominio de ATIS) consiste en utilizar el protocolo CAS (*Common Answer Specification*). Este protocolo compara el funcionamiento del sistema con una respuesta canónica de la base de datos, presentando la ventaja de poder realizar la evaluación de forma automática, una vez que los principios para generar las respuestas de referencia se han establecido y un corpus de datos ha sido convenientemente etiquetado.

Además, permite la comparación directa entre sistemas de forma sencilla. Sin embargo, el procedimiento de evaluación CAS está bastante limitado. La evaluación se realiza sólo en el nivel de frase, es decir, se compara frase a frase con su respuesta canónica. Tampoco se realiza ninguna

distinción entre una respuesta parcialmente correcta y una completamente incorrecta. Por tanto, este método no es completamente eficaz en la evaluación de este tipo de sistemas interactivos, no permite detectar ni corregir errores, no permite evaluar la calidad de las respuestas, etc.

Una posible alternativa consiste en plantear una serie de escenarios donde se le indica al usuario qué información deberá solicitar al sistema. Estos escenarios tienen una respuesta bien definida. El usuario realiza una o varias preguntas al sistema, manteniendo un cierto diálogo con el mismo, y el sistema le envía una respuesta para cada una de ellas, o una nueva pregunta/mensaje de error. Toda esta información queda registrada en un fichero, es decir, cada par pregunta-respuesta. Además, se mide el tiempo consumido por el usuario y el sistema en llevar a cabo los distintos procesos (reconocimiento/transcripción del Mago de Oz de la pregunta, envío de la respuesta del sistema al usuario, tiempo de almacenamiento de toda la información que registra el sistema de medida, tiempo que tarda el usuario en pensar y realizar cada pregunta, etc.).

Se suele incluir algún tipo de cuestionario para el usuario, preguntando por su aceptación del sistema, qué le gusta y qué no le gusta, si considera que el sistema le ha "entendido" y con qué frecuencia, si le ha parecido ágil el diálogo, si ha entendido las respuestas enviadas por el sistema, etc. A partir del fichero generado y del cuestionario se realizan una serie de medidas que permiten evaluar cuantitativamente el sistema interactivo. Entre ellas:

- Tiempo en finalizar el escenario (el usuario no sabe que le están midiendo).
- La existencia o no de una solución.
- Si la solución encontrada por el sistema es correcta.
- Número de preguntas que ha necesitado hacer el usuario.
- Número de preguntas del usuario que ha contestado el sistema.
- Número de preguntas que se consideran (por parte de un evaluador experto) que han sido contestadas correctamente, incorrectamente o parcialmente correctas.
- Satisfacción del usuario sobre la base del cuestionario.
- Medida de comprensión de acuerdo a otros criterios como el CAS.

Otras medidas comúnmente utilizadas en la evaluación de los diferentes módulos de un sistema de diálogo son:

- **Reconocimiento Acústico** (Tasa de Acierto o Precisión de Palabras (*Word Accuracy, WA*), Tasa de Inserción de Palabras (Word Insertions Rate), Tasa de Sustitución de Palabras (*Word Substitutions Rate*), Tasa de Frases Correctas (*Sentence Accuracy*), etc.

  La Tasa de Error (*Word Error rate: WER*) y el *Word Accuracy* pueden calcularse a partir del número total de substituciones en las frases de evaluación ($N_S$), el número total de borrados ($N_B$), el número total de inserciones ($N_I$) y el número total de palabras en dichas frases ($N_T$):

$$Sus(\%) = 100 * \frac{N_S}{N_T}$$

$$Borr(\%) = 100 * \frac{N_B}{N_T}$$

$$Inser(\%) = 100 * \frac{N_I}{N_T}$$

$$WER(\%) = Subs(\%) + Inser(\%) + Borr(\%)$$

$$WA(\%) = 100\,\% - WER(\%)$$

  El cálculo de estos porcentajes requiere disponer de la frase de referencia, lo que implica la necesidad de etiquetar manualmente las frases mencionadas por el usuario

- **Comprensión del habla**: Porcentaje de frases comprendidas correctamente, no comprendidas o parcialmente comprendidas, y un análisis de los errores; porcentaje de frases correctamente analizadas; porcentaje de palabras fuera del diccionario; porcentaje de frases sin cobertura lingüística; etc.

  Al igual que en el caso del reconocedor, es necesario etiquetar cada concepto semántico generado por el módulo de comprensión como correcto o incorrecto (conceptos correctos, insertados, borrados y substituidos). Para la evaluación de la traducción semántica final generada por los módulos de comprensión desarrollados en los proyectos BASURDE y DIHANA se definieron las siguientes medidas [219]:

  - $fc$. Porcentaje de frases cuya representación semántica final es igual que la de referencia;

$$\%fc = 100 * \frac{num.\ de\ frases\ con\ representacion\ en\ frame\ correcta\ en\ la\ hipotesis}{numero\ de\ frases}$$

  - $ufc$. Porcentaje de unidades de frame correctas, considerando unidades de frame el propio tipo de frame y cada uno de los valores de sus atributos;

$$\%ufc = 100 * \left(1 - \frac{num.\ inserciones\ +\ num.\ substituciones\ +\ num.\ borrados}{numero\ de\ unidades\ de\ frame\ en\ la\ referencia}\right)$$

- $P_f$. Precisión a nivel de frame;

$$\%P_f = 100 * \left( \frac{numero\ de\ unidades\ de\ frame\ correctas\ en\ la\ hipotesis}{numero\ de\ unidades\ de\ frame\ en\ la\ hipotesis} \right)$$

- $C_f$. Cobertura a nivel de frame;

$$\%C_f = 100 * \left( \frac{numero\ de\ unidades\ de\ frame\ correctas\ en\ la\ hipotesis}{numero\ de\ unidades\ de\ frame\ en\ la\ referencia} \right)$$

- **Gestión del Diálogo**. Análisis de las estrategias para recuperarse de errores, para corregir/dirigir la interacción del usuario, manejo del contexto cuando se producen múltiples preguntas-respuestas asociadas a un escenario % de respuestas correctas, % de respuestas incorrectas, % de respuestas a medias, % de veces que el sistema actúa intentando solucionar un problema, % de veces que el usuario actúa intentando solucionar un problema, etc.

- **Generación de respuestas**. Las medidas deben tratar de evaluar si el usuario entiende y asimila correctamente las informaciones que le ofrece el sistema. Ejemplos de medidas utilizadas son:

  - Número de veces que el usuario solicita la repetición de la respuesta facilitada por el sistema.
  - Tiempo de respuesta del usuario: cuanto peor se formulen las preguntas del sistema o mayor información suministre, mayor será este tiempo.
  - Número de veces que el usuario no contesta.
  - Tasa de palabras fuera de vocabulario: indica si debe aumentarse el vocabulario del reconocedor o reformular las preguntas.

  Esta evaluación suele realizarse habitualmente mediante el uso de opiniones subjetivas de los usuarios, utilizando preguntas que miden la relevancia, calidad, satisfacción global del usuario, etc.

- **Síntesis de voz**. En la evaluación de la síntesis de voz suelen considerarse dos aspectos fundamentales: la inteligibilidad de la voz sintética y la naturalidad de la voz. La inteligibilidad es un factor fundamental para asegurar el éxito del sistema. Hoy en día existen sintetizadores capaces de suministrar altas prestaciones en ambos parámetros. El estándar SSML (*Speech Synthesis Markup Language*) recoge especificaciones relativas a la síntesis, como el control de la prosodia, el énfasis en palabras o frases, etc.

En [220] se recogen 15 criterios que deben evaluarse para garantizar la usabilidad del sistema:

- Uso adecuado de las modalidades.
- Reconocimiento de las entradas de usuarios.
- Naturalidad del habla del usuario, en relación con la cobertura del vocabulario y las gramáticas utilizadas.
- Calidad de voz del sistema.
- Generación de respuestas adecuada.
- Realimentación adecuadas.
- Uso adecuado de la iniciativa del diálogo relativa a la tarea o tareas del sistema.
- naturalidad de la estructura del diálogo para la tarea(s).
- Cobertura suficiente del dominio.
- Capacidades de razonamiento del sistema suficientes.
- Guía o ayuda durante la interacción suficiente.
- Tratamiento de errores adecuado.
- Adaptación suficiente a las diferencias entre usuarios.
- Número de problemas durante la interacción.
- Satisfacción del usuario.

El grupo de evaluación EAGLES (*Expert Advisory Group on Language Engineering Standards*) [221] propone medidas cuantitativas (tasa de finalización de la tarea, tasa de éxito de la transacción, tiempo de respuesta del sistema, concisión de las respuestas del sistema...) y cualitativas (satisfacción del usuario, capacidad de adaptación a nuevos usuarios, capacidad para manejar la multimodalidad...). Además de proponer qué evaluar, se busca también establecer cómo evaluar e informar de los resultados, fijando un conjunto de parámetros (del sistema, de condiciones del test, de resultados del test) que permitan una comparación homogénea.

En el proyecto DISC (Spoken Language Dialogue Systems and Components) [222], en la misma línea que EAGLES, se propone qué evaluar (el conjunto de propiedades) y cómo evaluar (el criterio a aplicar). La metodología se basa en el uso de plantillas y ciclos de vida.

La evaluación propuesta en el proyecto francés EVALDA [223], orientado a la evaluación de las tecnologías del lenguaje desarrolladas para ese idioma, consiste en la utilización de conjuntos de test procedentes de corpus reales, una representación semántica del diálogo y métricas de evaluación comunes. No se han publicado todavía sus resultados.

En [224] se describe LINTEST, una herramienta para la evaluación de sistemas de diálogo mediante la utilización de un corpus basada en JUNIT (para la evaluación del sistema se definen ficheros con casos que se ejecutan para comprobar el funcionamiento correcto del sistema). Permite dos modos de funcionamiento: *batch mode* (se realizan tests de evaluación que generan como salida un fichero log de resultados) y *interactive mode* (permiten test más exhaustivos a modo de traza del sistema). En el trabajo se describe la utilización de la herramienta para el desarrollo y la evaluación del sistema BIRDQUEST (información sobre las aves de los países nórdicos).

### 9.2.1. *Calidad de los sistemas conversacionales y diálogos hablados*

**Calidad de los sistemas conversacionales y diálogos hablados**

**Benjamin Weiss, Stefan Hillmann, Sebastian Möller (Technical University Berlin, Alemania)**

Con los últimos avances tecnológicos en el reconocimiento automático del habla (ASR) y comprensión del lenguaje natural (NLU), nuevos sistemas basados en la voz emergen constantemente. No sólo los porcentajes del reconocimiento de voz aumentan con este desarrollo, sino también nuevas aplicaciones y dominios son logrados por las interfaces para usuarios basadas en el habla. Por lo tanto, nuevos retos y expectativas por parte de los usuarios deben ser consideradas en el proceso de diseño y evaluación.

Un desarrollo actual es la invasión creciente de la voz en contextos personales y profesionales. Introducido por los smartphones capaces de manipular la voz, continuando por los portátiles actuales, pero especialmente por los dispositivos inteligentes diseñados para el hogar y otros aparatos diseñados para propósitos específicos, están todos ellos continuarán, entrando en nuestras vidas privadas, así como en los coches (semi) autónomos o espacios inteligentes de trabajo en la

industria.

En contraste con la última generación de módulos de interacción, que han sido principalmente sujetos de uso puntual sin capacidad conversacional, enfocados en tareas muy específicas, y ejecutados por usuarios individuales, nuevos servicios están apareciendo en situaciones sociales involucrando múltiples usuarios. Por lo tanto, aspectos tales como, la privacidad y seguridad, adaptación e inteligencia, así como la sociabilidad ganan relevancia para las experiencias de los usuarios, y para la aceptación de los sistemas. Esto requiere avances científicos y tecnológicos en varias áreas:

1. Algunos sistemas son capaces de acumular datos acústicos que muestran situaciones de usuarios, estados personales, emociones, preferencias y demás. De estos datos se pueden construir perfiles de usuarios y predecir sus preferencias y comportamientos. Mientras que aprender sobre el usuario es favorable para la personalización y adaptación (ver el siguiente punto), debe ser asegurado por el diseño que esos datos no se escaparán del control directo del usuario. Permitiendo el control a los usuarios es un prerrequisito, sin embargo, no es suficiente ya que una seguridad y privacidad adecuadas están muy lejos aún, en particular en relación a sistemas interactivos que no contienen un interfaz para el usuario.

2. Los usuarios individuales esperan que los sistemas aprendan y se adapten a sus preferencias con el paso del tiempo, así como tener en cuenta el conocimiento sobre el mundo. Mientras que este tipo de comportamiento está surgiendo en asistentes personales (como los smartphones) mediante la incorporación información de fuentes adicionales a la aplicación, sistemas interactivos no contienen habitualmente ni el conocimiento más básico sobre el mundo.

   Algunos aspectos, sin embargo, están relacionados con una comprensión profunda del lenguaje, por ejemplo, comprender frases incompletas o actos de habla indirectos, o la detección de cambios de tema dentro del dominio en uso, pero también la detección de tópicos fuera de las capacidades de los sistemas y la capacidad de reconocerlos en vez de simplemente mostrar un malentendido. La adaptación a las preferencias del usuario, sin embargo, no es sólo difícil a la hora del diseño, sino también de cara a la evaluación. Mientras que los usuarios rechazan una configuración manual, las adaptaciones automáticas pueden ser percibidas como momentos positivos, pero también como altamente inadecuadas, y por lo tanto resultan en un rechazo cognitivo, indicado por la frustración y el rechazo. Estos momentos de frustración, sin embargo, pueden no aparecer durante las fases de evaluación

posteriores, y son muy difíciles de anticipar.

3. Los requisitos en la sociabilidad del sistema incluyen capacidades para lidiar con múltiples usuarios, así como con ambientes acústicos cambiantes y complejos (cambios de sala o simplemente la dirección del usuario y el ruido de fondo). Para soportar las conversaciones entre usuarios y máquina en ese tipo de ambientes, un sistema debe ser capaz de identificar las situaciones sociales, por ejemplo, quienes están dirigiéndose a un sistema en un contexto con múltiples usuarios.

   Incluso para interacciones de un sólo usuario, como en un coche, la situación social puede cambiar si el usuario está estresado porque se encuentra en una situación de tráfico desfavorable comparado con un usuario aburrido que tiene por delante un largo viaje. No sólo un comportamiento proactivo por parte del sistema, nuevas soluciones deben ser identificadas, así como modelos para estudiar las interacciones más frecuentes y cortas, así como largas conversaciones (cuando los sistemas sean capaces de lidiar con ellas).

Estos cambios esperados y en desarrollo en la tecnología y el comportamiento del usuario han hecho que los antiguos sistemas telefónicos de respuesta de voz interactiva (IVR) estén ya obsoletos, ya que no son capaces de generalizar dominios o grupos de usuarios. Adicionalmente, requieren nuevos métodos de evaluación y nuevos resultados (guiados por la evaluación). Por ejemplo, aún no es conocido, si las habilidades parecidas a las humanas de los sistemas conversacionales (interrupciones por parte del sistema, predictibilidad en los turnos, dudas, tratamiento de errores avanzados), que están actualmente siendo desarrolladas mejorarán la experiencia del usuario. A día de hoy también esperamos más requisitos por parte del usuario para las conversaciones con coches inteligentes, especialmente debido a la falta de control en la conducción, y para sistemas colaborativos en el trabajo (Industria 4.0), para los que la información y la comunicación es muy diversa.

Parte IV

# SISTEMAS CONVERSACIONALES: INVESTIGACIÓN Y RETOS

# 10. PANORAMA ACTUAL, TENDENCIAS Y OPORTUNIDADES

El informe de Gartner de las 10 tendencias tecnológicas estratégicas más importantes para 2018[63], incluye las platformas conversacionales como un cambio de paradigma. Entre los retos de futuro señalados, se indica que los usuarios aún deben comunicarse de una forma muy estructurada, lo que en ocasiones causa frustración. De esta forma, Gartner establece como elemento diferenciador la robustez de los modelos conversacionales y la forma en que las APIs integren servicios de terceros que permitan elaborar respuestas complejas.

## 10.1. PANORAMA ACTUAL

### 10.1.1. *La industria de las tecnologías del lenguaje y los sistemas conversacionales en Europa*

| Interfaces Conversacionales para Usuarios: Pasado y Futuro |
| --- |
| **Michael F. McTear, University of Ulster** |

Las interfaces conversacionales para usuarios están rápidamente convirtiéndose en una alternativa a las interfaces gráficas. También conocidos como asistentes personales o chatbots, las interfaces conversacionales para usuarios proveen una manera intuitiva de interactuar con servicios a través de internet a través de dispositivos inteligentes basados en voz tales como Amazon Echo y Google Home, asistentes personales como Siri, Cortana, Google Assisstant, and Bixby en smartphones, así como otro tipo de dispositivos inteligentes.

Las interfaces conversacionales para usuarios pueden estar enfocadas en tareas específicas, proveyendo información o realizando transacciones, o basadas en conversaciones, proveyendo un chat social que no aborda especialmente ninguna tarea. Las interfaces para usuarios basadas en tareas pueden ser a su vez divididas en distintos asistentes personales que proveen servicios para usuarios individuales, o asistentes de negocios que dan servicios para distintos distintas operaciones y transacciones. Varios avances tecnológicos han contribuido para mantener la viabilidad de interfaces conversacionales para usuarios, incluyendo los avances en inteligencia artificial, particularmente en aprendizaje profundo; el aumento de la potencia y capacidad de procesamiento; el aumento en la conectividad y el acceso a fuentes en la nube; la disponibilidad de una gran cantidad de datos disponibles para el entrenamiento del software; y las mejoras en la exactitud en

[63] *https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018/*

reconocimiento del habla y otras tecnologías del procesamiento del lenguaje natural.

Realmente las interfaces conversacionales para usuarios no son nuevas. Hay una larga tradición de investigación y desarrollo en sistema de diálogo desde los 1960 que ha investigado como los humanos pueden interactuar con los ordenadores usando lenguaje natural. Las interfaces de voz para usuarios, la contrapartida comercial a los sistemas de diálogo, se vienen usando 1999 en centros de contacto (contact centers) para automatizar la atención al cliente. También hay una larga tradición de chatbots comenzando con ELIZA en 1960 que simulaba conversaciones humanas llegando a la actividad con algunas interfaces conversacionales socialmente centradas, como por ejemplo sistemas dirigidos para ancianos.

Dos perspectivas diferentes pueden ser identificadas en el desarrollo de interfaces conversacionales: enfoques basados en reglas, y sistemas basados en aprendizaje automático o dirigidos por datos. Hasta el año 2000 los sistemas de diálogo hablados eran creados a mano usando guías para las mejores prácticas y (normalmente complejas) reglas para controlar el flujo del diálogo. Generalmente este acercamiento sigue siendo preferido en la industria debido a que el desarrollador es capaz de mantener el control en las capacidades del sistema. Alternativamente, y especialmente en investigaciones actuales, las estrategias basadas en datos están siendo usadas para incluir en el sistema la capacidad de aprender automáticamente cómo interactuar, esquivando la complejidad y la falta de escalabilidad y de creación de reglas.

El aprendizaje por refuerzo (Reinforcement Learning) puede ser usado para permitir a los sistemas aprender una política óptima para determinar las mejores opciones de interacción a través de una secuencia de estados con varias opciones en cada turno. Recientemente, el aprendizaje profundo está siendo aplicado para construir sistemas conversacionales end-to-end entrenables usando tecnologías que han sido desplegadas en respuesta automática de emails y traducción automática. Tanto el aprendizaje por refuerzo y como el aprendizaje profundo ambos requieren una cantidad inmensa de datos para el entrenamiento, además aún hay algunos problemas de manejabilidad para RL y la falta de una memoria a largo plazo explícita para sistemas conversacionales basados en aprendizaje profundo que están siendo tratadas en las investigaciones actuales.

The Sound of Language Intelligence

**Philippe Wacker, LT-Innovate**

LT-Innovate[64] es la Asociación de la Industria de las Tecnologías del Lenguaje. Engloba más de 200 compañías, principalmente PYMES, Europeas y de otros lugares. Sus campos de actividad cubren todas las áreas de la tecnología del lenguaje, esto es, aprendizaje automático, análisis de datos multilingües e interacciones conversacionales. Esta última es de una importancia cada vez más creciente: un informe Gartner sobre las 10 tendencias tecnológicas estratégicas para 2018[65] incluían *sistemas y plataformas conversacionales* como la clave para las interacciones en auge entre humanos y máquinas. En su predicción para 2019[66], Gartner ve está tendencia confirmada y ampliada por las experiencias multimodales como la realidad virtual.

Un estudio de mercado reciente[67] prevee un aumento significativo en el sector global de reconocimiento de voz y habla, con predicciones de crecimiento hasta los $22.3 miles de millones para 2024. Esta predicción está basada en tecnologías (identidad del hablante y verificación, reconocimiento del habla y texto a voz) que encuentran su uso en segmentos del mercado (automovilística, salud, seguros, gubernamental, etc.) y regiones (US, UK, Francia, Alemania, Brazil, India, China, Japón, África). Esta tendencia implica un crecimiento de más del 20 % anual durante los siguientes 5 años.

En el 2005, LT-Innovate y sus colaboradores en la Alianza de la Innovación de la Tecnología de la Interacción Conversacional - CITIA[68] propuso un *Mapa para las Tecnologías de la Interacción Conversacional*[69]. Apuntó a 5 escenarios para la investigación e innovación: 1. Interfaces adaptables para todo; 2. Asistentes personales inteligentes; 3. Acceso a información activa; 4. Robots comunicativos; 5. Colaboración y creatividad compartida. Tres años después, esta agenda está siendo implementada y muchas de sus predicciones están cumpliéndose a un ritmo mucho más alto que el esperado.

En Europa, una llamada [70] del Parlamento Europeo a las administraciones a todos los niveles *para mejorar un acceso a los servicios online e información en todos los idiomas, [...] incluyendo tra-*

[64] *www.lt-innovate.org*
[65] *https://www.youtube.com/watch?v=TPbKyD2bAR4*
[66] *https://www.youtube.com/watch?v=nRTRyfIDp4k*
[67] *https://www.zionmarketresearch.com/report/speech-and-voice-recognition-technologies-market*
[68] *http://www.lt-innovate.org/citia*
[69] *http://www.lt-innovate.org/sites/default/files/citia_files/rockit-scenarios-whitepaper-v2.1.pdf*
[70] *https://goo.gl/Sn8VBA*

*ducción automática, reconocimiento del habla y texto a voz y sistemas lingüísticos inteligentes, tales como sistemas capaces de realizar recuperación de información plurilingüe, y comprensión del habla...*. Extrañamente, la Comisión Europea ha prestado poca atención a la tecnología del lenguaje en sus versiones previas del plan de 7 años para la Innovación e Investigación, Horizonte Europeo[71], y su contrapartida digital Europa Digital[72], la cual realizo el doblaje de su *primer programa digital* [73].

Mientras, el reciente Artículo de Visión[74] de LT-Innovate posiciona la comprensión del lenguaje natural en el mismo corazón de la inteligencia artificial y aclama: *mientras el Mercado Único Digital no sea plurilingüe, Europa estará compuesta de 20 mercados separados*. Cada vez más figuras de la industria comparten esta visión. Recientemente, Wolfgang Blau[75], Presidente de Condé Nast International, hizo una llamada a la inversión de mil millones de euros en traducción automática neural y declaró la *traducción de internet* como la siguiente gran tendencia en su discurso en la conferencia 'Challenging the Content' en Vienna el 8 de Octubre de 2018. Durante el discurso, señaló lo obvio: en Europa, la fragmentación lingüística del Mercado Único Digital es un cuello de botella... que puede ser superado, pero al coste de voluntad política y una inversión sustancial.

El evento anual de LT-Innovate, el Language Technology Industry Summit, celebrará su 8ª edición el 24-25 de junio 2019 en Bruselas. Es el punto de encuentro ideal para la industria de tecnologías del lenguaje, usuarios e integradores de la tecnología, investigadores y desarrolladores y representantes del sector público. Ha sido bautizado bajo el eslogan *El sonido de la Inteligencia Lingüística*[76].

### 10.1.2. *Grupos de investigación y empresas internacionales*

Puede encontrarse un listado muy completo de grupos de investigación y empresas internacionales en los campos de Procesamiento de Lenguaje Natural, Tecnologías del Habla e Interfaces Conversacionales en la web de la European Network in Language and Speech (ELSNET) [77]. Entre los grupos de investigación internacionales dentro del campo de las tecnologías de diálogo hablado, en Estados Unidos cabe destacar:

[71] *https://goo.gl/1LpjmN*
[72] *https://goo.gl/7LA5fg*
[73] *http://europa.eu/rapid/press-release_IP-18-4043_en.html*
[74] *http://www.lt-innovate.org/sites/default/files/Mission\%20Paper.pdf*
[75] *https://www.youtube.com/watch?v=pNoYGyGHu3o*
[76] *https://www.conferize.com/the-sound-of-intelligence*
[77] *www.elsnet.org/lsorglist_a.html*

- el Grupo Spoken Language Systems del MIT[78],
- el grupo Center for Spoken Languaje Undesrtanding del Institute of Science and Technology de Oregon[79],
- el Sphinx Group de la Carnegie Mellon University [80],
- los grupos Computational Linguistics [81] y Computational Language and EducAtion Research Center (CLEAR) [82] de la Universidad de Colorado,
- el grupo Conversational Interaction and Spoken Dialogue Research de la Universidad de Rochester[83],
- el grupo Natural Language Dialog Group en University of Southern California[84].

Entre las compañías estadounidenses que participaban activamente en este campo de investigación cabe destacar AT&T (*www.att.com*), los laboratorios Bell (*www.bell-labs.com*), Microsoft (*www.microsoft.com*), IBM (*www.ibm.com*), etc. (ver sección 8.3. Algunas de ellas tienen grupos de investigación específicos, como es el caso del grupo Conversational Systems Research Group[85] de Microsoft.

En algunos casos, los grupos de investigación en el ámbito de los sistemas de diálogo dentro de la empresa surgen por la relación estrecha de ésta con un grupo académico de especial relevancia. Este es el caso por ejemplo de la alianza entre Toshiba y el grupo de sistemas de diálogo de la Universidad de Cambridge[86]. En el ámbito europeo existen también un gran número de centros de investigación:

- el grupo Centre for Speech Technology Research[87] de la Universidad de Edimburgo (Reino Unido),
- el grupo Human-Robot Interaction (HRI)[88] de la Universidad Heriot-Watt (Reino Unido),

[78] *groups.csail.mit.edu/sls*
[79] *cslu.cse.ogi.edu*
[80] *www.speech.cs.cmu.edu*
[81] *https://www.colorado.edu/linguistics/research/computational-linguistics*
[82] *https://www.colorado.edu/lab/clear/*
[83] *www.cs.rochester.edu/research/cisd*
[84] *http://nld.ict.usc.edu/group/*
[85] *https://www.microsoft.com/en-us/research/group/conversational-systems-research-group/*
[86] *https://www.toshiba.eu/eu/Cambridge-Research-Laboratory/Speech-Technology/Dialogue-Systems/*
[87] *http://www.cstr.ed.ac.uk/*
[88] *http://www.cstr.ed.ac.uk/*

- el LIMSI Spoken Language Processing Group en Paris[89],
- el Centre for Speech Technology de la Universidad de Edimburgo [90],
- el grupo Speech Communication and Technology[91] del KTH en Estocolmo,
- el grupo Language Technology del DKFI[92] en Alemania,
- CSELT[93] en Italia,
- el Speech Research Group de la Universidad de Cambridge[94],
- el grupo Dialogue Systems Group[95] de la Universidad de Ulm (Alemania),
- el grupo de Dialogue Systems Group[96] de la Universidad de Bielefeld (Alemania),
- los grupos Human-Computer Dialogue Systems[97] , Speech and Hearing Research [98] y Machine Intelligence for Natural Interfaces [99] de la Universidad de Sheffield (Reino Unido).

### 10.1.3. *Grupos de investigación y empresas nacionales*

A nivel nacional cabe destacar los siguientes grupos:

- Pattern Recognition and Human Language Technology (PRHLT) de la Universitat Politècnica de València[100],
- Center for Language and Speech Technologies and Applications (TALP)[101] de la Universitat Politècnica de Catalunya[102],
- Research Group Spoken and Multimodal Dialogue Systems (SISDIAL) de la Universidad de Granada[103],

[89] *www.limsi.fr/TLP*
[90] *www.cstr.ed.ac.uk*
[91] *www.speech.kth.se/speech*
[92] *www.dfki.de*
[93] *www.cselt.it*
[94] *mi.eng.cam.ac.uk/research/speech*
[95] *www.dialogue-systems.org*
[96] *http://www.dsg-bielefeld.de/dsg_wp/*
[97] *https://www.sheffield.ac.uk/dcs/research/groups/nlp/dialogue*
[98] *http://spandh.dcs.shef.ac.uk/*
[99] *https://mini.dcs.shef.ac.uk/*
[100] *https://www.prhlt.upv.es/wp/*
[101] *www.talp.upc.es*
[102] *www.talp.upc.es*
[103] *https://sisdial.ugr.es/*

- Multimedia Technology Group (GTM) de la Universidad de Vigo[104],
- Grupo de Aplicaciones del Procesado de Señal (GAPS) de la Universidad Politécnica de Madrid[105],
- el Laboratorio de Comunicación Oral ROBERT WAYNE NEWCOMB del Departamento de Arquitectura y Tecnología de Sistemas Informáticos de la Universidad Politécnica de Madrid[106],
- el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Sevilla [107],
- el Grupo de Tecnologías de las Comunicaciones del Instituto de Investigación en Ingeniería de Aragón en el Centro Politécnico Superior de la Universidad de Zaragoza[108],
- el Speech Interactive Research Group de la Universidad del País Vasco[109],
- el Grupo de Tecnología del Habla del Departamento de Ingeniería Electrónica de la ETSI de Telecomunicación en la Universidad Politécnica de Madrid[110],
- el Grup d'Enginyeria del Llenguatge Natural i Reconeiximent de Formes (ELiRF)[111] del Departamento de Sistemas Informáticos y Computación de la Universidad Politécnica de Valencia,

### 10.1.4. *Organismos, redes, congresos e iniciativas*

**Red Temática en Tecnologías del Habla**

| Presentación de la RTTH |
| --- |
| **Carlos Martínez Hinarejos (Universidad de Valencia), Coordinador de la RTTH)** |

La Red Temática en Tecnologías del Habla (RTTH)[112] se fundó en el año 2000. Desde entonces

[104] *http://gtm.uvigo.es/*
[105] *https://goo.gl/ooQRsT*
[106] *labaudio.datsi.fi.upm.es*
[107] *www.cs.us.es*
[108] *i3a.unizar.es*
[109] *https://www.ehu.eus/en/web/speech-interactive/about-us*
[110] *www-gth.die.upm.es*
[111] *users.dsic.upv.es/grupos/rfia/webelirf2/*
[112] *http://www.rthabla.es/*

ha realizado una importante labor por la promoción de las tecnologías del habla en el estado español, centrándose en el fomento de la investigación en las Tecnologías del Habla. Para ello ha involucrado en sus actividades a agentes académicos y empresariales, ha atraído a talento joven hacia estas tecnologías y ha apoyado la internacionalización de sus grupos miembro (con resultados notables en la participación de los mismos en proyectos financiados en el programa europeo H2020). En estos años se ha favorecido la colaboración e integración de las personas investigadoras de los distintos grupos miembro, creando una comunidad cohesionada y con una identidad definida, articulada alrededor de distintos foros y lazos establecidos entre los grupos. A su vez, esto ha permitido en enlace con otros grupos internacionales y ha supuesto el afianzamiento del sector de la investigación en estas tecnologías en España, consiguiendo en varios casos el liderazgo en investigación en el tratamiento tanto del español como del resto de lenguas oficiales del territorio estatal.

Las principales actividades realizadas en el marco de la RTTH han sido:

- La organización bianual de las Jornadas en Tecnologías del Habla desde el año 2000. Las cinco últimas ediciones se han celebrado de manera conjunta con el Iberian SLTech Workshop, organizado por el SIG-IL (Special Interest Group on Iberian Languages) de ISCA (International Speech Communication Association), pasando a tener así un carácter internacional y conociéndose a partir de 2012 como IberSpeech.

  Estas jornadas se han consolidado como un foro preferente para establecer y consolidar contactos entre personas y grupos de investigación, para la integración de talento joven, como foro para el contacto con empresas, para mostrar prototipos y para mostrar los últimos proyectos y tesis desarrollados en estas tecnologías. La asistencia a estas jornadas ha sido variable pero suele superar el centenar de participantes, y suelen contar con la participación invitada de personalidades científicas relevantes en las Tecnologías del Habla.

- La generación de los libros de actas de las distintas Jornadas. En las ediciones de 2012, 2014 y 2016 se llegó a un acuerdo con la editorial Springer para publicar aquellos trabajos con mejor valoración en volúmenes de la serie Lecture Notes in Computer Science.

- La organización bianual de la Escuela de Verano de la RTTH desde 2011, eventos que permiten a las personas que están iniciando sus estudios de doctorado tener un primer contacto con otras personas en su misma situación, así como recibir formación de primera mano de personas expertas en Tecnologías del Habla.

- La organización de evaluaciones tecnológicas competitivas, llamadas Evaluaciones Albayzín. Estas evaluaciones, abiertas a toda la comunidad, han sido un excelente instrumento para el intercambio de ideas, metodologías y técnicas entre distintos grupos que trabajan en temas semejantes. Las evaluaciones se han realizado en diversas temáticas: identificación de idioma, síntesis de habla, segmentación de audio, búsqueda en documentos hablados, traducción automática, reconocimiento automático de habla patológica, diarización y reconocimiento multimodal. La participación en estas evaluaciones ha contado con notable éxito.

- La convocatoria anual, desde 2006, del premio al mejor artículo publicado en una revista con índice de impacto (según el Journal Citation Reports) donde un o una estudiante de doctorado en una institución española sea autor o autora principal, en las temática de Tecnologías del Habla. Esta convocatoria anual tuvo una excelente acogida inicial, la cual se ha ido reafirmando en sucesivas ediciones en las que la cantidad y calidad de los artículos presentados no ha hecho más que incrementarse.

- La convocatoria de premios asociados a los mejores artículos con primer firmante estudiante, tesis, demostraciones y proyectos presentados dentro de cada edición de las Jornadas organizadas bianualmente. Estos premios en general han contado con el patrocinio de empresas con intereses de investigación y desarrollo en Tecnologías del Habla (como Telefónica I+D o Cirrus Logic).

- La colaboración con el SIG-IL de ISCA en la organización conjunta de las Jornadas y el Iberian SLTech Workshop, así como en la edición de un número especial de la revista Speech Communication dedicada a investigación en lenguas ibéricas.

Por otra parte, las actuaciones realizadas por la RTTH hasta ahora y las que tiene previsto realizar se encuadran en los siguientes objetivos:

1. Potenciar la internacionalización de la RTTH: se trata de mejorar la vinculación de sus grupos con otros grupos europeos e iberoamericanos y con su tejido empresarial; para ello se realizan una serie de acciones:

    - La búsqueda de socios europeos fuera del estado español para optar a proyectos del programa H2020 y continuaciones de dicho programa.
    - La difusión internacional de los eventos de la RTTH (Jornadas y la Escuela de Verano).

- La atracción del talento joven internacional, apoyando al establecimiento de titulaciones internacionales de máster y doctorado en Tecnologías del Habla.
- La difusión y exportación de los desarrollos y avances en Tecnologías del Habla en España.
- La promoción de la transferencia de tecnología a empresas.
- La colaboración con otras entidades no españolas con temática similar (ISCA, IEEE) y para la distribución de recursos (LDC, ELRA/ELDA).
- La movilidad de recursos humanos a entornos internacionales.

2. Fomentar la investigación a nivel estatal en las Tecnologías del Habla, en particular en la atracción de talento joven hacia estudios de máster y doctorado relacionados; para ello, se establecen las siguientes acciones:

   - La promoción de la participación en las Jornadas y en la Escuela de Verano.
   - La realización de eventos específicos para el intercambio de experiencias entre estudiantes en Tecnologías del Habla.
   - La constitución de un fondo de ayuda a la asistencia a los eventos de la Red para estudiantes que tengan una carrera prometedora sin financiación alternativa.
   - La generación de material bibliográfico de apoyo.
   - La promoción del intercambio de alumnado entre los distintos programas de doctorado asociados a los grupos miembro.
   - El mantenimiento y actualización del Centro Investigador y Docente virtual.
   - La organización de los premios a mejor artículo de estudiante en revista de impacto, a los mejores artículos y demostradores en las Jornadas, y a personas investigadoras consolidadas cuyas aportaciones a las Tecnologías del Habla en España hayan sido de gran relevancia.

3. Mantener los foros de reunión y las colaboraciones entre los grupos miembro, y ampliarlas a otros grupos estatales, tanto académicos como del sector empresarial, estableciendo las siguientes acciones:

   - La organización de las Jornadas en el marco del evento internacional IberSpeech, involucrando en las mismas a organizaciones tanto académicas como empresariales.
   - La organización de las Evaluaciones Albayzín.

- La organización de la Escuela de Verano.
- La búsqueda de nuevas fuentes de financiación fruto de la colaboración entre los distintos grupos de la Red y de los mismos con otros grupos.
- La coordinación en el uso de infraestructuras científico-tecnológicas por parte de grupos miembros.

4. Incrementar la transferencia científica y tecnológica entre academia y empresa, en particular por proyectos conjuntos y trasvase de recursos humanos; para ello se realizan las siguientes acciones:
   - La promoción de la realización de prototipos y demostradores por parte de los grupos de la Red y su difusión a agentes empresariales y a la sociedad en general.
   - El establecimiento de foros de encuentro de empresas y academia en el marco de los eventos de la Red, en la forma de mesas redondas.
   - La promoción de la realización de prácticas en empresa en los marcos de las titulaciones de máster y doctorado, por la difusión de ofertas y la creación de una bolsa de personal investigador.
   - La institución de patrocinios por parte de las empresas de los eventos de la Red.
5. Resaltar la importancia transversal de las Tecnologías del Habla en temáticas estratégicas (manejo de grandes volúmenes de datos y su uso inteligente), mediante las siguientes acciones:
   - El establecimiento de relaciones y alianzas colaborativas con otros grupos y redes en temáticas de multimodalidad, *big data*, *Smart Cities* e *Internet of Things*.
   - La promoción de la formación en tecnologías complementarias del alumnado de máster y doctorado en Tecnologías del Habla.
   - La organización de sesiones específicas en los eventos de la Red sobre las aplicaciones en estos ámbitos.
   - El contacto con empresas de servicios *big data* e *Internet of Things* para establecer sus necesidades en Tecnologías del Habla y formar vínculos con los grupos apropiados de la Red.
   - El contacto con administraciones públicas que necesiten Tecnologías del Habla a fin de establecer posibles colaboraciones con los grupos de la Red.

**Sociedad Española para el Procesamiento del Lenguaje Natural**

Presentación de la SEPLN

**Luis Alfonso Ureña López (Universidad de Jaén), presidente de la SEPLN)**

La Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) es una asociación sin ánimo de lucro formada por socios numerarios e instituciones que se creó en el año 1983 con el objeto de promover y difundir todo tipo de actividades referentes a la enseñanza, investigación y desarrollo en el ámbito del procesamiento del lenguaje natural, tanto a nivel nacional como internacional.

Los objetivos de la SEPLN son establecer canales de intercambio de información y materiales científicos, la organización de seminarios, simposios y conferencias, la promoción de publicaciones y la colaboración con otras instituciones relacionadas con su campo de actuación.

Entre las actividades más destacadas de la SEPLN figuran la celebración de un Congreso Internacional, un referente anual que sirve de punto de encuentro para los distintos investigadores y grupos de investigación que trabajan en el área del procesamiento del lenguaje natural.

También, la edición de una revista, Revista de Procesamiento del Lenguaje Natural, con carácter semestral, dotada de comité de asesor y de redacción que garantiza unos criterios estables de calidad y periodicidad. Esta revista editada por la SEPLN, ocupa los primeros puestos entre las revistas científicas españolas, tanto del área de ingeniería como de lingüística y humanidades. Está presente en los principales índices y rankings, como el ESCI, concretamente se encuentra en Máster Journal List de la Web of Science (Thomson Reuters), así como en los principales repositorios. Tiene el Sello de Calidad FECYT. La revista es de carácter abierto, acceso on-line a todo el catálogo (revistas desde 1983…), donde se incluyen artículos originales, presentaciones de proyectos en curso, reseñas bibliográficas y resúmenes de tesis doctorales. Asimismo, un portal con información sobre temas relacionados con el procesamiento del lenguaje natural y un servicio de correo electrónico que informa sobre las cuestiones de actualidad y se utiliza como espacio de discusión para los socios.

Por otra parte, la sociedad también ha consolidado el Premio SEPLN a la investigación en procesamiento del lenguaje natural. La finalidad del premio es la promoción y divulgación de la investigación en el campo de las tecnologías del lenguaje humano.

Finalmente destacar que la Sociedad Española para el Procesamiento del Lenguaje Natural es un

referente y agente activo en el Plan de Impulso de la Tecnologías del Lenguaje (Plan TL) puesto en marcha por el Gobierno de España.

**ACL, ISCA y SIGDIAL.** Entre las asociaciones científicas internacionales más relevantes en el ámbito de las tecnologías del habla destacan:

- La Asociación para la Lingüística Computacional (Association for Computational Linguistics[113], ACL). La ACL organiza varios congresos de reconocido prestigio en el área como son ACL, EACL o NAACL y edita las revistas Computational Linguistics y Transactions of the ACL.
- La Asociación Internacional de Comunicación Oral (International Speech Communication Association, ISCA[114]). ISCA organiza uno de los congresos más relevantes del área: INTERSPEECH y edita dos revistas de referencia: Speech Communications y Computer Speech and Language.

Ambas asociaciones tienen varios grupos de interés relacionados con los sistemas conversacionales, entre los cuales destaca el Grupo de Interés Especial en Discurso y Diálogo (Special Interest Group on Discourse and Dialogue, SIGDIAL[115]). Cada año se celebra un congreso, el SIGDIAL Workshop y junto con él una mesa redonda donde los jóvenes investigadores pueden presentar sus ideas y trabajo en curso: el Young Researchers Round Table.

## 10.2. TENDENCIAS

### 10.2.1. *Proyectos de investigación nacionales*

**GENIOVOX: GENeración computacIOnal de VOz eXpresiva**

*Financiación:* Ministerio de Economía, Industria y Competitividad (TEC2016-81107-P)

*Periodo*: 30-12-2016 - 29-12-2019.

*Investigadores principales:* Oriol Guasch y Francesc Alías, La Salle, Universitat Ramon Llull

El proyecto GENIOVOX

**Francesc Alías (La Salle, Universitat Ramon Llull)**

[113] *https://www.aclweb.org/portal/*

[114] *https://www.isca-speech.org/iscaweb/*

[115] *https://www.sigdial.org/content/about-sigdial*

En este proyecto se pretende abordar por primera vez la generación computacional de voz expresiva mediante un enfoque híbrido al problema. Esto nos permitirá eludir las limitaciones de los corpus de voz aprovechando los desarrollos en técnicas de transformación de voz. La idea principal consiste en aplicar las modificaciones de parámetros identificadas en las grabaciones a los modelos de pulsos glotales que sirven como condiciones de contorno para las simulaciones numéricas de la acústica del tracto vocal. A la vez, la geometría de estos últimos también se modificará dinámicamente para conseguir los efectos expresivos deseados.

### 10.2.2. *Proyectos de investigación internacionales*

**LIHLITH: Learning to Interact with Humans by Lifelong Interaction with Humans**

*Financiación:* EU CHIST-ERA

*Periodo*: 2018 a 2020.

*Investigador principal:* Eneko Aguirre, Universidad del País Vasco

Mejorando los sistemas de aprendizaje a través del aprendizaje continuo

**Eneko Aguirre (Universidad del País Vasco), coordinador del proyecto LIHLITH**

Los sistemas de diálogo suelen construirse de forma que su comportamiento sea constante en el tiempo. Para mejorar el rendimiento a lo largo del tiempo, el sistema de dialogo debe poder aprender de su experiencia, sus errores y los comentarios del usuario. Este proceso, apropiadamente llamado aprendizaje continuo, es el enfoque de LIHLITH. LIHLITH (*Learning to Interact with Humans by Lifelong Interaction with Humans*) es un proyecto de tres años de alto riesgo / alto impacto financiado por CHIST-ERA[116] que comenzó en enero de 2018. El proyecto está liderado por investigadores de la Universidad del País Vasco, con participantes del Laboratorio de Ciencias de la Computación en Mecánica e Ingeniería en Francia, la Universidad Nacional de Educación a Distancia en España, la Universidad de Ciencias Aplicadas de Zurich en Suiza y Synapse Développement en Francia.

La inteligencia artificial es un campo que está progresando rápidamente en muchas áreas, incluidos los diálogos con máquinas y robots. Los ejemplos incluyen hablar con un dispositivo para solicitar tareas simples como encender la radio o preguntar por el tiempo, pero también confi-

[116] *http://www.chistera.eu/*

guraciones más complejas donde la máquina llama a un restaurante para hacer una reserva[117], o donde un robot ayuda a los clientes en una tienda. LIHLITH[118] es un proyecto que se centra en los diálogos persona-máquina. Su objetivo es mejorar las capacidades de autoaprendizaje de la inteligencia artificial. Más específicamente, LIHLITH diseñará sistemas de diálogo que aprendan a mejorarse a sí mismos en función de sus interacciones con los humanos.

Los chatbots industriales actuales se basan en reglas que deben elaborarse a mano cuidadosamente para cada dominio de aplicación. Alternativamente, los sistemas basados en aprendizaje automático utilizan datos anotados manualmente del dominio para entrenar el sistema de diálogo. En ambos casos, producir reglas o datos de entrenamiento para cada dominio de diálogo consume mucho tiempo, y limita la calidad y la adopción generalizada de los chatbots. Además, las empresas deben monitorear el rendimiento del sistema de diálogo después de ser implementado y volver a diseñarlo para responder a las necesidades del usuario. A lo largo del proyecto, LIHLITH explorará el paradigma del aprendizaje continuo en los sistemas de diálogo hombre-máquina con el objetivo de mejorar su calidad, reducir el costo de mantenimiento y reducir los esfuerzos para el despliegue en nuevos dominios.

**Objetivo principal:** mejora continua de los sistemas de diálogo.

El objetivo principal de los sistemas de aprendizaje continuo [225] es continuar aprendiendo después de implementarse. En el caso de LIHLITH, el sistema de diálogo se desarrollará como de costumbre, pero incluirá maquinaria para continuar mejorando sus capacidades en función de su interacción con los usuarios. La idea clave es que los diálogos se diseñarán para obtener retroalimentación de los usuarios, mientras que el sistema se diseñará para aprender de esta retroalimentación continua. Esto permitirá que el sistema siga mejorando durante su vida útil, adaptándose rápidamente a los cambios de dominio que se producen después de la implantación.

LIHLITH se centrará en diálogos de respuesta a preguntas basados en objetivos, donde el usuario tiene una necesidad de información y el sistema intentará satisfacer esta necesidad mientras conversa con el usuario. El proyecto se ha estructurado en tres áreas de investigación: aprendizaje continuo para el diálogo; aprendizaje continuo para la inducción del conocimiento y la respuesta a preguntas; y evaluación de la mejora del diálogo. Todos los módulos se diseñarán para aprender de los comentarios disponibles utilizando técnicas de aprendizaje profundo.

[117] *https://kwz.me/htg*
[118] *http://ixa2.si.ehu.es/lihlith/*

La Figura 17 muestra esta estructura donde el esquema de un sistema de diálogo estándar se muestra con casillas blancas y el innovador módulo de aprendizaje continuo se muestra en azul. Este módulo puede mejorar todos los módulos (en azul) según las interacciones pasadas y la interacción con el usuario actual, actualizando el conocimiento del dominio en consecuencia [226].

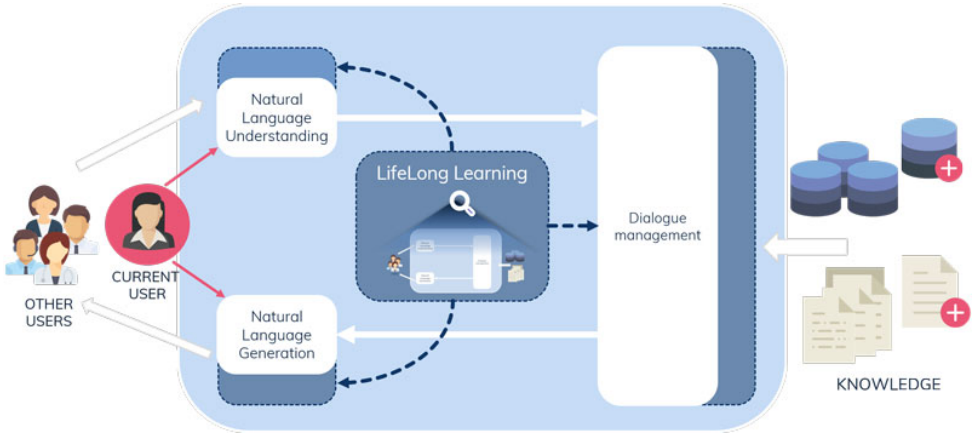


*Figura 17: Esquema de un sistema de diálogo estándar (blanco) y con LIHLITH (azul)*

El objetivo del aprendizaje continuo para el diálogo será obtener un método para producir un módulo de gestión de diálogos que aprenda de diálogos anteriores. El proyecto explorará la reconfiguración autónoma de estrategias de diálogo basadas en los comentarios de los usuarios. También proporcionara capacidades proactivas al sistema, que se utilizarán para solicitar al usuario nuevos conocimientos y comentarios sobre el rendimiento. Esto se activará, por ejemplo, cuando las reacciones pasadas hayan sido rechazadas, cuando la interacción del usuario sea demasiado ambigua, cuando las respuestas posibles sean demasiado numerosas o si tengan puntuaciones de confianza demasiado similares.

Con respecto a la inducción del conocimiento y la respuesta a las preguntas, el objetivo es mejorar el conocimiento del dominio, que incluye la representación de los enunciados y el rendimiento de la respuesta a las preguntas en función de los comentarios de diálogo obtenidos por el módulo de gestión del diálogo. La representación de los enunciados y la base de conocimientos se basará en representaciones en espacios de bajas dimensiones. El sistema de respuesta a preguntas aprovechará tanto la información de los textos de fondo como las ontologías de dominio. La retroalimentación se utilizará para proporcionar una señal supervisada en estos sistemas de aprendizaje y, por lo tanto, ajustar los parámetros de los sistemas de aprendizaje profundo subyacentes.

La evaluación de los sistemas de diálogo sigue siendo un reto, con problemas de reproducibilidad y comparabilidad. LIHLITH producirá puntos de referencia para el aprendizaje continuo en sistemas

de diálogo, que se aplicarán en una tarea compartida internacional para explorar las capacidades de las soluciones existentes. Además, la investigación en LIHLITH se transferirá al sistema de diálogo industrial de Synapse. Para llevar a cabo esta investigación, LIHLITH combina el aprendizaje automático, la representación del conocimiento y la experiencia lingüística. El proyecto se basará en los avances recientes en varias disciplinas de investigación, incluido el procesamiento del lenguaje natural, el aprendizaje profundo, la inducción del conocimiento, el aprendizaje por refuerzo y la evaluación del diálogo, para explorar su aplicabilidad al aprendizaje continuo.

**EUNISON: Extensive UNIfied-domain SimulatiON of the human voice**

*Financiación:* Future Emerging Technologies (FET) financiado por la Comisión Europea (308874)

*Periodo*: 01-03-2013 a 29-02-2016.

*Investigador principal:* Oriol Guasch, La Salle, Universitat Ramon Llull

El proyecto EUNISON

**Francesc Alías (La Salle, Universitat Ramon Llull)**

En el proyecto EUNISON, tratamos de construir un nuevo simulador de voz que se basa en los principios físicos básicos del habla. A partir de unos determinados datos de entrada, que representan la topología o las activaciones musculares o los fonemas, se generará la física 3D de la voz, incluyendo, por supuesto, su salida acústica. Esta simulación permitirá obtener información importante de cómo funciona la voz. El objetivo del proyecto no es desarrollar un sistema de síntesis de voz, sino un sistema de simulación de voz, con distintas aplicaciones. Tomando los controles adecuados y con la suficiente potencia de computación, el sistema podría hablar en cualquier idioma, o cantar en cualquier estilo. El KTH coordina este proyecto, con un presupuesto de 2,96 millones de euros.

**SENSEI: Interpretando Datos de Conversaciones Humano-Humano**

*Financiación:* EU FP7-ICT-2013-10

*Periodo*: 01-11-2013 a 31-10-2016.

*Investigador principal:* Giuseppe Riccardi, Universidad de Trento, Italia

## The SENSEI Project

**Giuseppe Riccardi (Universidad de Trento, Italia), coordinador del proyecto SENSEI**

Los ciudadanos, clientes y usuarios generan cantidades masivas de trazas digitales cuando interactúan con administraciones públicas, proveedores de servicios y compañías de medios de comunicación. La mayoría, si no todos, son ignoradas por las administraciones públicas, proveedores de servicios y en general cualquier empresa interesada en escuchar a los usuarios. Consecuentemente, los problemas y las preocupaciones señaladas por los ciudadanos no pueden ser tenidas en cuenta adecuadamente o, análogamente, las preocupaciones de la atención al cliente no pueden ser transmitidas.

SENSEI es un proyecto de investigación e innovación que apunta a la invención de tecnologías pensadas para la comprensión de una cantidad masiva de conversaciones simultáneas ocurriendo en distintos canales (centros de atención al cliente, emails, plataformas de blogging etc.). El objetivo principal del proyecto es el desarrollo de tecnologías de resumen y análisis para ayudar a los usuarios a comprender los flujos de conversaciones humanas en diversos canales; y para evaluar la tecnología desarrollada en ambientes ecológicos, con el objetivo de mejorar la actuación y productividad de cara a usuarios finales.

La mayoría de la tecnología de análisis de lenguaje está limitada en el sentido de que hace búsqueda por palabras clave, lo cual no permite hacer una descripción automática de lo que ha ocurrido, quién dijo qué, qué opiniones se han mostrado en qué tópico, de manera coherente, leíble y ejecutable. SENSEI ha ido más allá del estado del arte y ha diseñado prototipos que generan automáticamente documentos leíbles de análisis (resúmenes) y apoyo de cara a los usuarios finales en los contextos de tareas de análisis de grandes cantidades de datos. Estos prototipos han sido evaluados extrínsecamente con usuarios finales en situaciones reales.

SENSEI se ha puesto en contacto con la comunidad científica para una evaluación independiente y revisada de sus resultados y han sido publicados más de 50 artículos. SENSEI ha conectado con el sector industrial y ha organizado más de 40 eventos para difundir la visión y resultados de la tecnología SENSEI. Por último, los inversores de SENSEI han llevado su perspectiva, algoritmos y software más allá de la duración del proyecto y han publicado herramientas y datos en repositorios públicos.

Los logros del proyecto SENSEI son de dos tipos. Primero SENSEI ha ido más allá del estado del arte

y ha diseñado prototipos que generan automáticamente documentos leíbles de análisis (resúmenes) y apoyo a los usuarios finales en los contextos de tareas de análisis de grandes cantidades de datos. Segundo, SENSEI ha evaluado la tecnología desarrollada en ambientes ecológicos, con el objetivo de mejorar la actuación y productividad de cara a los usuarios finales. Las conversaciones generadas en diversos medios y canales son analizadas y elaboradas a través de un canal que incluye: 1) Análisis semántico de conversaciones, 2) Análisis para-semántico de conversaciones, 3) Análisis del discurso de conversaciones, 4) Resumen de conversaciones, y 5) Representación gráfica de los resúmenes incluyendo gráficos, imágenes, vídeos y textos.

Los grupos de usuarios utilizados en el análisis focalizado provienen de centros de atención al cliente y de dominios específicos de medios de comunicación. En el primero, este grupo está formado por analistas de datos, profesionales del control de calidad y directores de los grupos. En el segundo, el grupo está formado por lectores de comentarios de noticias, autores de dichos comentarios, periodistas y editores/analistas de medios de comunicación.

Uno de los logros más importantes ha sido el desarrollo final, así como la entrega y evaluación de tecnologías de parsing de conversaciones humanas. Dichas tecnologías han sido evaluadas dentro del proyecto en los casos de uso planificados, así como externamente en tareas internacionales compartidas, y finalmente testeado en eventos en vivo.

En particular, hemos desarrollado un componente fundamental para el consenso automático de computación: el parser de la relación acuerdo/desacuerdo. Una segunda innovación es el diseño de algoritmos para el análisis automático de dinámicas basadas en turnos y empatía en conversaciones habladas. El componente de parsing del discurso ha sido mejorado considerablemente para la extracción de relaciones en el discurso y adaptación de algoritmos para conferencias a los medios de comunicaciones sociales. El valor e impacto de dichos resultados ha sido confirmado por las numerosas publicaciones científicas.

Por último, el equipo de SENSEI tomó el desafío de probar en vivo sus datos y canales de procesamiento del lenguaje en la tarea de predecir la orientación de los medios de comunicación sociales en el día a día durante un mes completo previo al referendum del Brexit. El sistema SENSEI-Brexit fue el único, y en gran contraste con las predicciones, en sugerir el resultado final con proporciones muy similares al resultado final del voto (SALIR contra QUEDARSE).

El segundo logro más importante ha sido la evaluación extrínseca (de cara a usuarios finales) de los prototipos de resumen enfocados en habla y medios de comunicación sociales SENSEI.

Los experimentos de evaluación han sido producidos siguiendo el principio de participación de usuarios en tareas reales. En el caso del habla, el escenario fue real y estuvo llevado a cabo por analistas de conversaciones y profesionales de seguro de calidad. Se desarrolló en dos idiomas (francés e italiano) y en dos centros de atención al cliente en Francia e Italia.

El proceso de evaluación mostró resultados interesantes a la vez que prometedores en términos de ganancia de productividad y perspectivas sobre cómo influenciar la innovación de las herramientas de resumen del habla. La evaluación extrínseca de los resúmenes en los medios de comunicación sociales ha sido llevada a cabo en situaciones reales. La complejidad y la novedad de estas herramientas y la visualización del análisis han generado resultados distintos entre los usuarios participantes. Es relevante destacar que el análisis completo de las frases del usuario y las diferencias del sistema han producido puntos de vista para el desarrollo futuro de los requisitos para un sistema y su interfaz de usuario.

**KRISTINA: A Knowledge-based Information Agent with Social Competence and Human Interaction Capabilities**

*Financiación:* H2020 ICT theme Multimodal and Natural Computer Interaction

*Periodo*: 01-03-2015 a 28-02-2018

*Investigador principal:* Leo Wanner, Universitat Pompeu Fabra

KRISTINA: Un agente virtual socialmente competente en el ámbito de la salud

**Leo Wanner (Universitat Pompeu Fabra), coordinador del proyecto KRISTINA**

La migración en Europa es un fenómeno que forma parte de la misma esencia del continente y que ha llegado a convertirse en una seña de identidad. El gran crisol de culturas que forma la Unión Europea de hoy en día es el legado de la riqueza cultural y diversidad lingüística que ha venido caracterizando el viejo continente desde sus orígenes. Sin embargo, la integración en el ámbito de la salud y la atención geriátrica básica de algunos colectivos especialmente sensibles, como la comunidad árabe, aún es un reto para la Europa del siglo XXI.

El Proyecto KRISTINA, financiado por el programa Horizon 2020, es una de las apuestas más importantes de desarrollo tecnológico de la Unión Europea, un pilar estratégico para el progreso de una sociedad pluricultural en el plano tecnológico y social. KRISTINA es un agente virtual experto en cuestiones sanitarias y que está especialmente dirigido a las comunidades de inmigrantes árabes, turcos y polacos. En muchos casos, estos colectivos desconocen los sistemas sanitarios de sus

países de acogida y no hablan con soltura la lengua de destino. De ahí que el principal objetivo del Proyecto KRISTINA ha sido el de desarrollar un agente socialmente competente y comunicativo para facilitar la superación de las barreras lingüísticas y culturales de las personas migrantes en los servicios de atención sanitaria primaria y geriátrica de los países anfitriones.

El consorcio que ha formado parte en el proyecto KRISTINA está compuesto por seis instituciones técnicas: la Universidad Pompeu Fabra, que ha realizado funciones de dirección y técnicas, el Instituto CERTH, la Universidad de Augsburgo, la Universidad de Ulm, la empresa Almende y Vocapia Research; y tres socios clínicos: la Universidad de Tübingen, experta en atención geriátrica, la Cruz Roja alemana y la sociedad española de medicina de familia y comunitaria (semFYC).

Las comunidades de inmigrantes sobre las que ha trabajado durante los 3 años de duración del proyecto están formadas por cuidadoras de origen polaco en Alemania, ancianos de origen turco que viven en Alemania y pacientes de origen árabe que requieren de asistencia sanitaria primaria en España. KRISTINA puede dar información sobre el funcionamiento administrativo de la sanidad del país de residencia, proporcionar consejos y recursos informativos disponibles en la red y permitir a los pacientes comprender cuándo es imprescindible la visita a su médico de familia. El vehículo de comunicación es la conversación en el idioma nativo del usuario. KRISTINA es capaz de identificar el estado de ánimo del usuario y responder teniendo en cuenta esta situación. Así mismo, la apariencia del avatar está adaptada a cada una de las culturas del proyecto.

Desde el punto de vista técnico, la arquitectura del sistema se divide en tres grandes capas de integración que a su vez incluyen distintos módulos:

- Capa de procesamiento de señales de entrada. La gestión de la entrada de voz y vídeo se realiza bajo la plataforma de reconocimiento desarrollada por la Universidad de Augsburgo, “Social Signal interpretation” (SSI). Los módulos que componen esta capa cumplen dos funciones:
    - Análisis del contenido verbal
        - Reconocimiento automático de voz: transforma la señal de voz en texto.
        - Parsing sintáctico: analiza la estructura sintáctica del mensaje de texto y lo transforma en una representación semántica de su contenido.
        - Análisis semántico: convierte las estructuras semánticas en representaciones ontológicas para su procesamiento en la base de conocimiento central.
    - Análisis del contenido no verbal

- Reconocimiento de expresiones faciales: se procesa la señal de audio para identificar las emociones básicas exhibidas por el usuario mientras está hablando.
- Reconocimiento de gestos: se identifican gestos que puedan conllevar contenido semántico (por ejemplo, gestos simbólicos o deícticos).
- Fusión de rasgos sociales y emocionales: este módulo combina los datos recogidos por los demás módulos y los integra en una representación global de la modelización del comportamiento del usuario.

- Capa de procesamiento semántico. Esta capa está compuesta por los dos sistemas de gestión de contenido centrales del sistema, a saber, los repositorios de conocimiento y el gestor de diálogo.
  - Repositorios de conocimiento: incluyen la información proveniente del motor de búsquedas en la web y la información modelizada para cada caso de uso. En esta última categoría, se encuentran los modelos del usuario correspondientes a datos proporcionados por la entidad clínica que KRISTINA tiene permiso de gestionar (por ejemplo, hábitos alimenticios, costumbres y rutinas en el caso de cuidado de ancianos) así como datos contextuales y culturales de las distintas comunidades (costumbres religiosas, alimentos prohibidos, etc.). Por último, la base de conocimiento también incluye el historial de diálogo.
  - Gestor de diálogo: es el encargado de que los movimientos del diálogo sean naturales y sigan la estructura convencional de una conversación natural en los idiomas de los usuarios.
- Capa de generación de comunicación. La gestión de generación tanto de la apariencia física como de la voz del avatar se realiza bajo la plataforma Visual SceneMaker2. Un módulo intermedio de selección de modalidad es el que determina si el contenido proveniente de la base de conocimiento y gestor del diálogo ha de ser representado por una modalidad u otra, es decir, por el modo verbal, mediante técnicas de generación del lenguaje natural, o no verbal, a través de gestos o expresiones faciales.

El proyecto finalizó el pasado mes de marzo de 2017 después de pasar con éxito la revisión científico-técnica por parte del grupo de expertos asignado por la Comisión Europea. En estos momentos queda abierta la posibilidad de explotación de los resultados mediante estrategias de implantación en el mercado y desarrollo empresarial. El potencial de asistencia del agente con-

versacional KRISTINA va más allá de lo estrictamente técnico o económico. Es una herramienta de apoyo al equipo de asistencia sanitario europeo en su trato con personas que necesitan información básica en su lengua materna. Sin duda, un paso hacia adelante en la integración sociocultural de Europa en materia de salud.

**EMPATHIC: Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly**

*Financiación:* H2020-EU.3.1.4. - Active ageing and self-management of health

*Periodo*: 01-11-2017 a 31-10-2020

*Investigadora principal:* María Inés Torres, Universidad del País Vasco

El proyecto de Investigación e Innovación EMPATHIC[119] investigará, innovará y validará nuevos paradigmas, asentando los fundamentos para generaciones futuras de Asesores Virtuales Personalizados para la ayuda de la tercera edad a vivir por su cuenta. El asesoramiento del bienestar recomendará hábitos y comportamientos saludables, mediante la estimulación del usuario para que transforme sus objetivos y necesidades personales en acciones. El asistente virtual EMPATHIC (EMPATHIC-VC) incitará a sus clientes de avanzada edad de tomar previsiones para evitar enfermedades crónicas potenciales, mantener una dieta sana, tener actividad física adecuada, así como a socializar, contribuyendo así a una vida independiente y satisfactoria para la tercera edad.

El EMPATHIC-VC motivará a los usuarios mediante un asistente virtual amistoso a conseguir beneficios predefinidos, cuya realización será estudiada mediante las métricas diseñadas para el seguimiento. Nuestra ambición es crear un ambiente amigable y familiar para los usuarios, evitando los efectos hostiles de los dispositivos centrados en supervisión de la salud. El proyecto mirará más allá de las necesidades médicas y físicas básicas de una persona, y se centrará en la conexión entre el bienestar emocional y la salud física. El EMPATHIC-VC será capaz de percibir el estado emocional y social de una persona, en el contexto aprendido de las necesidades y expectativas de usuarios de tercera edad, y su historia personal y responderá adaptativamente a sus necesidades.

[119] *http://www.empathic-project.eu/*

**MENHIR: Mental health monitoring through interactive conversation**

*Financiación:* H2020-MSCA-RISE-2018

*Periodo*: 01-02-2019 a 31-01-2023

*Investigadora principal:* Zoraida Callejas Carrión (Universidad de Granada)

La Organización Mundial de la Salud considera que la salud mental es un componente esencial de la salud y por tanto no debe entenderse únicamente desde el tratamiento de los síntomas, sino que cobra especial relevancia su comprensión, promoción y protección.

La depresión y la ansiedad son desórdenes mentales muy frecuentes en la Unión Europea. Los tratamientos como la terapia cognitiva y la medicación son efectivos para reducir los síntomas de ansiedad y depresión y gracias a ellos muchas personas experimentan una recuperación completa tras un episodio de enfermedad mental. Sin embargo, para otros, los desórdenes de depresión y ansiedad siguen una trayectoria diferente donde pueden producirse recurrencia de los síntomas o episodios con diferentes grados de severidad.

Para este segundo escenario se desconoce el momento idóneo para el tratamiento o cuándo es preferible una *espera vigilada*. Muchas personas viven con una enfermedad mental y la tratan de gestionar empleando sus propias estrategias y redes de soporte con la ayuda de servicios especializados y terapeutas. Los sistemas conversacionales pueden ayudar a marcar esta diferencia y facilitar la monitorización de síntomas durante dicha espera, iniciando un contacto frecuente con el usuario. El proyecto MENHIR tiene como objetivo investigar y desarrollar tecnologías conversacionales para promover la salud mental y asistir a las personas con problemas de ansiedad y depresión no severa a gestionar su situación mediante la monitorización de sus síntomas para prevenir recaídas.

La monitorización llevada a cabo por el agente desarrollado reconocerá y monitorizará los estados emocionales, comportamiento y síntomas de los usuarios pudiendo aportar a terapeutas información constante y ecológicamente válida. Por otra parte, la monitorización empoderará a los pacientes para gestionar sus propios síntomas, aportando además una retroalimentación positiva y adecuada.

El uso de sistemas conversacionales facilitará la adherencia de los usuarios proveyendo una interfaz más realista donde se podrá actuar de manera más fluida e inteligible para el usuario. Los miembros del equipo compartirán su experiencia para alcanzar este objetivo común intercam-

biando ideas y conocimiento entre España, Alemania, Italia y Reino Unido con socios académicos, empresas y asociaciones de salud mental. El desarrollo de este proyecto generará nuevas oportunidades para los sistemas conversacionales y el análisis multimodal del habla enraizados en un conocimiento más profundo de sus usuarios, a su vez que los psicólogos y expertos en salud mental conocerán las nuevas tecnologías a su disposición para ayudar a sus pacientes en su día a día.

## Referencias


[1] M. McTear, Z. Callejas, and D. Griol, *The Conversational Interface: Talking to Smart Devices*. Springer, 2016.

[2] AIMultiple, "Top 60 Chatbot Companies: In-depth Guide [2019 update]," tech. rep., 2019.

[3] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "Galaxy-II: A reference architecture for conversational system development," in *Proc. of International Conference on Spoken Language Processing (ICSLP'98)*, (Sydney (Australia)), pp. 931–934, 1998.

[4] S. Seneff, R. Lau, and J.Polifroni, "Organization, communication, and control in the galaxy-II conversational system," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'99)*, (Budapest (Hungria)), pp. 1271–1274, 1999.

[5] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual spoken-language understanding in the MIT Voyager system," in *Speech Communication*, vol. 17, pp. 1–18, 1995.

[6] V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goodine, D. Goddeau, and J. Glass, "PEGASUS: A spoken dialogue interface for on-line air travel planning," in *Speech Communication*, vol. 15, pp. 331–340, 1994.

[7] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," in *IEEE Transactions on Speech and Audio Processing*, vol. 8(1), pp. 85–96, 2000.

[8] D. Bohus, A. Raux, T. Harris, M. Eskenazi, and A. Rudnicky, "Olympus: an open-source framework for conversational spoken language interface research," in *Proc. of HLT-NAACL'07 Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology*, (Rochester, NY (Estados Unidos)), pp. 32–39, 2007.

[9] D. Bohus, S. Grau, D. Huggins-Daines, V. Keri, G. Krishna, R. Kumar, A. Raux, and S. Tomko, "Conquest - an Open-Source Dialog System for Conferences," in *Proc. of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, (Rochester, NY (Estados Unidos)), pp. 9–12, 2007.

[10] A. Raux, B. Langner, A. Black, and M. Eskenazi, "Let's Go Public! Taking a Spoken Dialog System to the Real World," in *Proc. of the International Conference on Spoken Language Processing (Interspeech'05)*, (Lisboa (Portugal)), pp. 885–888, 2005.

[11] R. Rosenfeld, X. Zhu, S. Shriver, A. Toth, K. Lenzo, and A. W. Black, "Towards a universal speech interface," in *Proc. of International Conference on Spoken Language Processing (ICSLP'00)*, vol. 2, (Beijing (China)), pp. 102–105, 2000.

[12] S. Tomko and R. Rosenfeld, "Speech Graffiti vs. Natural Language: Assessing the User Experience," in *Proc. of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL'04*, (Boston (Estados Unidos)), pp. 73–76, 2004.

[13] Varios Autores, "Carnegie mellon university's universal speech interface," 2019.

[14] M. Turunen, J. Hakulinen, K.-J. Räihä, E.-P. Salonen, A. Kainulainen, and P. Prusi, "An architecture and applications for speech-based accessibility systems," in *IBM Systems Journal*, vol. 44(3), pp. 485–504, 2005.

[15] M. Turunen and J. Hakulinen, "Mailman - a Multilingual Speech-only E-mail Client Based on an Adaptive Speech Application Framework," in *Proc. of Workshop on Multi-Lingual Speech Communication (MSC'00)*, (Kyoto (Japón)), pp. 7–12, 2000.

[16] K. Mäkelä, E. P. Salonen, M. Turunen, J. Hakulinen, and R. Raisamo, "Conducting a Wizard of Oz Experiment on a Ubiquitous Computing System Doorman," in *Proc. of the International Workshop on Information Presentation and Natural Multimodal Dialogue*, (Verona (Italia)), pp. 115–119, 2001.

[17] G. Ferguson and J. Allen, "TRIPS: An Intelligent Integrated Problem-Solving Assistant," in *Proc. of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, (Madison (Estados Unidos)), pp. 567–573, 1998.

[18] J. Alexandersson and T. Becker, "Overlay as the basic operation for discourse processing in a multimodal dialogue system," in *Proc. of 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, (Seattle (Estados Unidos)), pp. 376–383, 2001.

[19] M. González and M. Gatius, "Un Sistema de Diálogo Multicanal para Acceder a la Información y Servicios de las Administraciones Públicas," in *Revista de la Sociedad Española de Procesamiento del lenguaje natural*, vol. 35, pp. 285–292, 2005.

[20] C.-U. Shin and J.-W. Cha, "End-to-end task dependent recurrent entity network for goal-oriented dialog learning," *Computer Speech & Language*, vol. 53, pp. 12–24, 2019.

[21] L. F. D'Haro, R. E. Banchs, C. Hori, and H. Li, "Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics," *Computer Speech & Language*, vol. 55, pp. 200–215, 2019.

[22] B. Kim, K. Chung, J. Lee, J. Seo, and M.-W. Koo, "A bi-lstm memory network for end-to-end goal-oriented dialog learning," *Computer Speech & Language*, vol. 53, pp. 217–230, 2019.

[23] T.-H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. Rojas-Barahona, S. U. P.-H. Su, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system," in *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, vol. 16, pp. 438–449, 2017.

[24] X. Aubert and H. Ney, "Large Vocabulary Continuous Speech Recognition Using Word Graphs," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, (Detroit (Estados Unidos)), pp. 49–52, 1995.

[25] F. Wessel, K. Macherey, and R. Schlüler, "Using word probabilities as confidence measures," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, vol. 1, (Seattle, Washington (Estados Unidos)), pp. 225–228, 1998.

[26] R. Zhang and A. Rudnicky, "Word level confidence annotation using combinations of features," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'01)*, (Aalborg (Dinamarca)), pp. 2105–2108, 2001.

[27] R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, and J. Pardo, "Confidence measures for spoken dialogue systems," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, vol. 1, (Salt Lake City, Utah (Estados Unidos)), pp. 393–396, 2001.

[28] T. Hazen, S. Seneff, and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems," in *Computer Speech and Language*, vol. 16, pp. 49–67, 2002.

[29] R. López-Cózar, A. Rubio, P. García, and J. Segura, "Uso de Valores de Confianza y Expectativas en el Sistema de Diálogo SAPLEN," in *Revista de la Sociedad Española de Procesamiento del Lenguaje Natural*, vol. 24, pp. 37–41, 1999.

[30] D. Pérez-Piñar and C. García, "Application of confidence measures for dialogue systems through the use of parallel speech recognizers," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'05)*, (Lisboa (Portugal)), pp. 2785–2788, 2005.

[31] F. Torres, L. Hurtado, F. García, E. Sanchis, and E. Segarra, "Error handling in a stochastic dialog system through confidence measures," in *Speech Communication*, vol. 45(3), pp. 211–229, 2005.

[32] G. Bouwman and J. Hulstijn, "Dialogue strategy redesign with reliability measures," in *Proc. of the First International Conference on Language Resources and Evaluation (LREC'98)*, (Granada (España)), pp. 191–198, 1998.

[33] F. García, L. Hurtado, E. Sanchis, and E. Segarra, "The incorporation of Confidence Measures to Language Understanding," in *International Conference on Text Speech and Dialogue (TSD'03). Lecture Notes in Artificial Intelligence series 2807*, (Ceské Budejovice (República Checa)), pp. 165–172, 2003.

[34] V. Sama, J. Ferreiros, F. Fernández, R. S. Segundo, and J. Pardo, "Utilización de medidas de confianza en sistemas de comprensión del habla," in *Revista de la Sociedad Española de Procesamiento del Lenguaje Natural*, vol. 35, pp. 229–234, 2005.

[35] OECD, "Digital Government Strategies for Transforming Public Services in the Welfare Areas," tech. rep., 2016.

[36] R. Mendiola Ruiz, L. Gondra Sangroniz, V. Ormaechea Goiri, J. M. Martínez Eizaguirre, A. Tadeo Múgica, C. Bretos Paternain, and P. Daza Asumendi, "Triaje telefónico en Atención Primaria: análisis de la implantación de un modelo," *Pediatría Atención Primaria*, vol. 16, no. 63, pp. 205–210, 2014.

[37] K. Lee, J. K. Kim, M. W. Park, L. Kim, and K. Hsiao, "A Situation-Based Dialogue Classification Model for Emergency Calls," in *2017 International Conference on Platform Technology and Service (PlatCon)*, pp. 1–4, 2017.

[38] U. Nallasamy, A. W. Black, T. Schultz, R. Frederking, and J. Weltman, "Speech translation for triage of emergency phonecalls in minority languages," in *Proc. of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications (Coling'08)*, (Manchester), pp. 48–53, 2008.

[39] V. Young, E. Rochon, and A. Mihailidis, "Exploratory analysis of real personal emergency response call conversations: considerations for personal emergency response spoken dialogue systems," *Journal of NeuroEngineering and Rehabilitation*, vol. 13, no. 1, 2016.

[40] T. Bickmore and T. Giorgino, "Health dialog systems for patients and consumers," *Journal of Biomedical Informatics*, vol. 39, no. 5, pp. 556–571, 2006.

[41] M. Sillice, P. Morokoff, G. Ferszt, T. Bickmore, B. Bock, R. Lantini, and W. Velicer, "Using relational agents to promote exercise and sun protection: Assessment of participants' experiences with two interventions," *Journal of Medical Internet Research*, vol. 20, no. 2, 2018.

[42] A. Shamekhi, H. Trinh, T. Bickmore, T. Deangelis, T. Ellis, B. Houlihan, and N. Latham, "A virtual self-care coach for individuals with spinal cord injury," pp. 327–328, 2016.

[43] D. Griol and Z. Callejas, "Mobile Conversational Agents for Context-Aware Care Applications," *Cognitive Computation*, vol. 8, no. 2, pp. 336–356, 2016.

[44] H. Tanaka, H. Adachi, N. Ukita, M. Ikeda, H. Kazui, T. Kudo, and S. Nakamura, "Detecting Dementia Through Interactive Computer Avatars," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 5, 2017.

[45] I. Leite, C. Martinho, and A. Paiva, "Social Robots for Long-Term Interaction: A Survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[46] S. Turkle, *Reclaiming Conversation: The Power of Talk in a Digital Age*. New York: Penguin Books, 2016.

[47] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 236, pp. 433–460, 1950.

[48] J. Weizenbaum, "Eliza - a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, pp. 36–45, 1966.

[49] K. Colby, F. Hilf, S. Weber, and H. Kraemer, "Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes," *Artificial Intelligence*, vol. 3, pp. 199–221, 1972.

[50] B. F. Green, A. Wolf, and C. Chomsky, "Laughery. k. baseball: An automatic question answerer," *Computers and Thought*, pp. 207–216, 1963.

[51] T. Winograd, "Understanding natural language," *Cognitive psychology*, vol. 3, no. 1, pp. 1–191, 1972.

[52] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd, "Gus, a frame-driven dialog system," *Artificial intelligence*, vol. 8, no. 2, pp. 155–173, 1977.

[53] DARPA, "Speech and Natural Language Workshop," in *Book of Proceedings*, (San Mateo (Estados Unidos)), 1992.

[54] ARPA, "Speech and Natural Language Workshop," in *Book of Proceedings*, (San Mateo (Estados Unidos)), 1994.

[55] R. Pieraccini, E. Levin, and W. Eckert, "AMICA: The AT&T mixed initiative conversational architecture," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'97)*, (Rodas (Grecia)), pp. 1875–1878, 1997.

[56] W. Ward, "Evaluation of the CMU ATIS System," in *Proc. DARPA Speech and Natural Language Workshop*, (Pacific Grove (Estados Unidos)), pp. 101–105, 1991.

[57] S. Seneff, L. Hirschman, and V. Zue, "Interactive problem solving and dialogue in the ATIS domain," in *Proc. of the Fourth ARPA Speech and Natural Language Workshop*, (Pacific Grove, California (Estados Unidos)), pp. 354–359, 1991.

[58] V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goddeau, J. Glass, and E. Brill, "The MIT ATIS System: December 1993 Progress Report," in *Proc. of ARPA Spoken Language Technology Workshop*, (Princeton (Estados Unidos)), pp. 339–353, 1993.

[59] J. Peckham, "A new generation of spoken dialogue systems: results and lessons from the SUNDIAL project," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'93)*, vol. 1, (Berlín (Alemania)), pp. 33–42, 1993.

[60] J. Bos, E. Klein, O. Lemon, and T. Oka, "The verbmobil prototype system - a software engineering perspective," in *Journal of Natural Language Engineering*, vol. 5(1), pp. 95–112, 1999.

[61] N. Bernsen and L. Dybkjaer, "The DISC project," in *ELRA Newsletter*, vol. 2(2), pp. 6–8, 1997.

[62] E. Den, L. Boves, L. Lamel, and P. Baggia, "Overview of the ARISE project," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'99)*, vol. 3, (Budapest (Hungría)), pp. 1527–1530, 1999.

[63] H. Aust, M. Oerder, M. Seide, and V. Steinbiss, "The Philips automatic train timetable information system," in *Proc. of the Interactive Voice Technology for Telecommunications Applications (IVTTA'94)*, (Kyoto (Japón)), pp. 67–72, 1994.

[64] G. Castagneri, P. Baggia, and M. Danieli, "Field trials of the Italian ARISE train timetable system," in *Proc. of the Interactive Voice Technology for Telecommunications Applications Workshop (IVTTA'98)*, pp. 97–102, 1998.

[65] P. Baggia, G. Castagneri, and M. Danieli, "Field trials of the Italian ARISE train timetable system," in *Speech Communication*, vol. 31, pp. 355–367, 2000.

[66] L. Lamel, S. Rosset, J. Gauvain, S. Bennacef, M. Garnier-Rizet, and B. Prouts, "The LIMSI ARISE system," in *Speech Communication*, vol. 31, pp. 339–353, 2000.

[67] A. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Shern, K. Lenzo, W. Xu, and A. Oh, "Creating natural dialogs in the Carnegie Mellon Communicator system," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'99)*, vol. 1(4), pp. 1531–1534, 1999.

[68] W. Ward and B. Pellom, "The CU Communicator system," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'99)*, (Keystone, Colorado (Estados Unidos)), pp. 341–344, 1999.

[69] J. Gauvain, S. Bennacef, L. Devillers, and L. Lamel, "Spoken language system development for the Mask Kiosk," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'95)*, (Salt Lake City (Estados Unidos)), pp. 119–120, 1995.

[70] R. Billi and L. Lamel, "Railtel: Railway telephone services," in *Speech Communication*, vol. 23, pp. 63–82, 1997.

[71] J. Wilpon, L. Rabiner, C.-H. Lee, and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden markov models," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, no. 11, pp. 1870–1878, 1990.

[72] J. L. Brandt, B. Mamuzic, J. T. Miller, and S. M. Mueller, "System and method for creating and accessing outgoing telephone call log," May 6 2008. US Patent 7,369,651.

[73] A. Gorin, G. Riccardi, and J. Wright, "How may I help you?," in *Speech Communication*, vol. 23, pp. 113–127, 1997.

[74] H. Kamp and U. Reyle, *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer, 1993.

[75] L. Polanyi, "The Linguistic Structure of Discourse," tech. rep., Center for the Study of Language and Information, Stanford University, 1996.

[76] N. Asher, *Reference to Abstract Objects in Discourse*. Dordrecht, the Netherlands: Kluwer Academic Publishers, 1993.

[77] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text*, vol. 8, no. 3, pp. 243–281, 1988.

[78] M. Taboada and W. C. Mann, "Rhetorical structure theory: looking back and moving ahead," *Discourse Studies*, vol. 8, no. 3, pp. 423–459, 2006.

[79] J. Searle, "Speech acts: An essay in the philosophy of language," in *Cambridge University Press*, (Cambridge (Reino Unido)), 1969.

[80] J. Alexandersson, "Plan Recognition in Verbmobil," in *Proc. of IJCAI-95 Workshop on the Next Generation of Plan Recognition Systems*, (Montreal, Canada), pp. 1–10, 1995.

[81] M. Core and J. Allen, "Coding Dialogs with the DAMSL Annotation Scheme," in *Proc. of AAAI Fall Symposium on Communicative Action in Humans and Machines*, (Boston, MA, USA), pp. 1–8, 1997.

[82] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, "Meeting Recorder Project: Dialog Act Labeling Guide," tech. rep., Technical Report TR-04-002. International Computer Science Institute, 1998.

[83] J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson, *HCRC Dialogue Structure Coding Manual*. University of Edinburgh, 1997.

[84] M. E. Pollack, "The uses of plans," *Artificial Intelligence*, vol. 57, no. 1, pp. 43–68, 1992.

[85] D. R. Traum and E. A. Hinkelman, "Conversation acts in task-oriented spoken dialogue," *Computational Intelligence*, vol. 8, no. 3, pp. 575–599, 1992.

[86] B. Grosz and C. Sidner, "Attention, intentions and the structure of discourse," *Computational Linguistics*, vol. 12, no. 3, pp. 175–204, 1986.

[87] S. Larsson and D. R. Traum., "Information state and dialogue management in the TRINDI dialogue move engine toolkit," *Natural Language Engineering*, vol. 6, no. 3-4, pp. 323–340, 2000.

[88] Varios Autores, “The hidden markov model toolkit (htk).” urlhttp://htk.eng.cam.ac.uk/, 2019.

[89] Varios Autores, “Cmusphinx open source speech recognition toolkit.” urlhttps://cmusphinx.github.io/, 2019.

[90] L. Rabiner, B. Juang, and C. Lee, “An overview of automatic speech recognition,” in *Automatic Speech and speaker Recognition: Advanced Topic* (K. A. Publishers, ed.), pp. 1–30, 1996.

[91] L. R. Bahl, F. Jelinek, and R. L. Mercer, “A maximum likelihood approach to continuous speech recognition,” in *Readings in Speech recognition*, (San Francisco, CA (Estados Unidos)), pp. 308–319, Morgan Kaufmann Publishers Inc., 1990.

[92] F. Jelinek, “Self-organized language modeling for speech recognition,” in *Readings in speech recognition*, (San Francisco, CA (Estados Unidos)), pp. 450–506, Morgan Kaufmann Publishers Inc., 1990.

[93] E. Segarra, *Una aproximación inductiva a la comprensión del discurso continuo*. PhD thesis, DSIC - UPV, Valencia (España), 1993.

[94] F. Jelinek, J. Lafferty, and R. Mercer, “Basic methods of probabilistic context free grammars,” in *Speech Recognition and Understanding. Recent Advances, Trends and Applications. Springer Verlag*, pp. 345–360, 1992.

[95] K. Fu and T. Booth, “Grammatical Inference: Introduction and Survey. Parts I and II,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 3, pp. 343–359, 409–423, 1986.

[96] W. Minker, “Stochastic versus rule-based speech understanding for information retrieval,” in *Speech Communication*, vol. 25(4), pp. 223–247, 1998.

[97] E. Segarra, E. Sanchis, F. García, and L. Hurtado, “Extracting semantic information through automatic learning techniques,” in *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 16(3), (Salt Lake City (Estados Unidos)), pp. 301–307, 2002.

[98] A. Stolcke *et al.*, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[99] S. Keizer, R. op den Akker, and A. Nijholt, "Dialogue act recognition with bayesian networks for dutch dialogues," in *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, 2002.

[100] T. Klüwer, H. Uszkoreit, and F. Xu, "Using syntactic and semantic based relations for dialogue act recognition," in *Proceedings of the 23rd international conference on computational linguistics: Posters*, pp. 570–578, Association for Computational Linguistics, 2010.

[101] J. H. Martin and D. Jurafsky, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River, 2009.

[102] G. Tur and D. Z. Hakkani-Tür, "Human/human conversation understanding," 2011.

[103] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)*, (San Francisco), pp. 517–520, 1992.

[104] C. Lee, S. Jung, K. Kim, D. Lee, and G. Lee, "Recent approaches to dialog management for spoken dialog systems," *Journal of Computing Science and Engineering*, vol. 4, no. 1, pp. 1–22, 2010.

[105] Y. Wilks, R. Catizone, S. Worgan, and M. Turunen, "Some background on dialogue management and conversational speech for dialogue systems," *Computer Speech and Language*, vol. 25(2), pp. 128–139, 2011.

[106] A. H. Oh and A. I. Rudnicky, "Stochastic language generation for spoken dialogue systems," in *Proc. of ANLP/NAACL Workshop on Conversational Systems*, pp. 27–32, 2000.

[107] C. Müller and F. Runge, "Dialogue design principles - key for usability of voice processing," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'93)*, (Berlín (Alemania)), pp. 943–946, 1993.

[108] P. B. Nielsen and A. Baekgaard, "Experience with a dialogue description formalism for realistic applications," in *Proc. of International Conference on Spoken Language Processing (ICSLP'92)*, (Banff (Canadá)), pp. 719–722, 1992.

[109] D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai, "A Form-Based Dialogue Manager for Spoken Language Applications," in *Proc. of International Conference on Spoken Language Processing (ICSLP'96)*, vol. 2, (Philadelphia (Estados Unidos)), pp. 701–704, 1996.

[110] S. McGlashan, D. Burnett, J. Carter, P. Danielsen, J. Ferrans, A. Hunt, B. Lucas, B. Porter, K. Rehor, and S. Tryphonas, "Voice Extensible Markup Language (VoiceXML) Version 2.0," in *Recomendación del W3C.* *www.w3.org/TR/voicexml20/* , 2004.

[111] J. Allen and C. Perault, "Analyzing intentions in dialogues," in *Artificial Intelligence*, vol. 15(3), pp. 143–178, 1980.

[112] D. Appelt, "Planning English Sentences," in *Cambridge, University Press*, 1985.

[113] P. Cohen and H. Levesque, "Rational interaction as the basis for communication," in *P. R. Cohen, J. Morgan, and M. E. Pollack, editors, Intentions in Communication. MIT Press*, 1990.

[114] J. Kowtko, S. Isard, and G. Doherty, "Conversational games within dialogue," in *Proc. of the ESPRIT Workshop on Discourse Coherence*, (Edimburgo (Escocia)), pp. 4–6, 1991.

[115] S. Williams, "Dialogue management in a mixed-initiative, cooperative, spoken language system," in *11th Twente Workshop on Language Technology (TWLT11)*, (Enschade (Países Bajos)), pp. 199–208, 1996.

[116] Website TRINDI. (Task Oriented Instructional Dialogue). www.ling.gu.se/projekt/trindi/ .

[117] D. Traum, J. Bos, R. Cooper, S. Larsson, I. Lewin, C. Matheson, and M. Poesio, "A model of dialogue moves and information state revision," in *Tech. rept. Deliverable D2.1. Trindi*, 1999.

[118] S. Larsson, A. Berman, J. Bos, L. Grönqvist, and P. LJunglöf, "A model of dialogue moves and information state revision," tech. rep., D5.1 Trindi (Task Oriented Instructional Dialogue), 1999.

[119] P. Bohlin, R. Cooper, E. Engdahl, and S. Larsson, "Information States and Dialogue Move Engines," in *Proc. of the IJCAI-99 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, (Murray Hill (Estados Unidos)), pp. 25–31, 1999.

[120] C. Matheson, M. Poesio, and D. Traum, "Modelling grounding and discourse obligations using update rules," in *Proc. of the 1st Annual Meeting of the North American Association for Computational Linguistics (NAACL'00)*, (Seattle (Estados Unidos)), pp. 1–8, 2000.

[121] S. J. Young, "Talking to machines (statistically speaking)," in *Proc. of 7th International Conference on Spoken Language Processing (ICSLP'02 - INTERSPEECH 2002)*, (Denver, Colorado, USA), pp. 1–8, 2002.

[122] E. Levin, R. Pieraccini, and W. Eckert, "A stochastic model of human-machine interaction for learning dialog strategies," in *IEEE Transactions on Speech and Audio Processing*, vol. 8(1), pp. 11–23, 2000.

[123] W. Biermann and P. Long, "The composition of messages in speech-graphics interactive systems," in *Proc. of the International Symposium on Spoken Dialogue*, (Philadelphia (Estados Unidos)), pp. 97–100, 1996.

[124] E. Levin and R. Pieraccini, "A stochastic model of human-machine interaction for learning dialog strategies," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'97)*, (Rodas (Grecia)), pp. 1883–1896, 1997.

[125] S. Singh, M. Kearns, D. Litman, and M. Walker, "Reinforcement learning for spoken dialogue systems," in *Proc. of Neural Information Processing Systems (NIPS'99)*, (Denver (Estados Unidos)), pp. 956–962, 1999.

[126] C. Watkins, *Learning from Delayed Rewards*. PhD thesis, Cambridge University, Cambridge (Reino Unido), 1989.

[127] H. Cuayáhuitl, S. Renals, O. Lemon, and H. Shimodaira, "Learning Multi-Goal Dialogue Strategies Using Reinforcement Learning with Reduced State-Action Spaces," in *Proc. of the 9th International Conference on Spoken Language Processing (Interspeech/ICSLP)*, (Pittsburgh (Estados Unidos)), pp. 469–472, 2006.

[128] S. Young, "Probabilistic Methods in Spoken Dialogue Systems," in *Philosophical Trans Royal Society (Series A)*, vol. v358 i1769, pp. 1389–1402, 2000.

[129] S. Young, "The Statistical Approach to the Design of Spoken Dialogue Systems," tech. rep., CUED/F-INFENG/TR.433, Cambridge University Engineering Department, Cambridge (Reino Unido), 2002.

[130] J. Williams and S. Young, "Partially observable markov decision processes for spoken dialog systems," *Computer Speech & Language*, vol. 21, no. 2, pp. 393–422, 2007.

[131] S. Young, J. Williams, J. Schatzmann, M. Stuttle, and K. Weilhammer, "The Hidden Information State Approach to Dialogue Management," tech. rep., Department of Engineering. University of Cambridge, Cambridge (Reino Unido), 2005.

[132] J. Williams and S. Young, "Scaling POMDPs for dialog management with composite summary point-based value iteration (CSPBVI)," in *Proc. of AAAI Workshop on Statistical and*

*Empirical Approaches for Spoken Dialogue Systems*, (Boston (Estados Unidos)), pp. 37–42, 2006.

[133] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young, "A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies," in *Knowledge Engineering Review*, vol. 21(2), pp. 97–126, 2006.

[134] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," in *Artificial Intelligence*, vol. 101, pp. 99–134, 1998.

[135] M. Littman, "The witness algorithm: Solving partially observable Markov decision processes," tech. rep., CS-94-40, Brown University, Department of Computer Science, Providence, RI, 1994.

[136] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," in *A Bradford Book. The MIT Press*, (Cambridge, Massachusetts (Estados Unidos)), 1998.

[137] J. Williams, P. Poupart, and S. Young, "Partially Observable Markov Decision Processes with Continuous Observations for Dialogue Management," in *Recent Trends in Discourse and Dialogue. Eds L. Dybkjaer and W. Minker, Springer*, pp. 191–217, 2006.

[138] J. Williams and S. Young, "Scaling POMDPs for Spoken Dialog Management," in *IEEE Audio, Speech and Language Processing*, vol. September 2007, 2007.

[139] J. Williams and S. Young, "Scaling POMDPs for dialog management with composite summary point-based value iteration (CSPBVI)," in *Proc. of AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems*, (Boston (Estados Unidos)), pp. 37–42, 2006.

[140] J. Pineau, G. Gordon, and S. Thrun, "Point-based value iteration: An anytime algorithm for POMDPs," in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, (Acapulco (México)), pp. 1025–1032, 2003.

[141] S. Young, "Using POMDPs for Dialog Management," in *Proc. of IEEE-ACL Workshop on Spoken Language Technology (SLT 2006)*, (Palm Beach (Aruba)), pp. 8–13, 2006.

[142] S. Young, J. Schatzmann, K. Weilhammer, and H. Ye, "The Hidden Information State Approach to Dialogue Management," in *Proc. of 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, (Honolulu, Haway (USA)), pp. 149–152, 2007.

[143] H. Cuayáhuitl, S. Renals, O. Lemon, and H. Shimodaira, "Reinforcement learning of dialogue strategies with hierarchical abstract machines," in *Proc. of IEEE/ACL Workshop on Spoken Language Technology (SLT)*, (Palm Beach (Aruba)), pp. 182–186, 2006.

[144] D. Griol, L. Hurtado, E. Segarra, and E. Sanchis, "A Statistical Approach to Spoken Dialog Systems Design and Evaluation," *Speech Communication*, vol. 50, no. 8–9, pp. 666–682, 2008.

[145] D. Griol, Z. Callejas, R. López-Cózar, and G. Riccardi, "A domain-independent statistical methodology for dialog management in spoken dialog systems," *Computer, Speech and Language*, vol. 28, no. 3, pp. 743–768, 2014.

[146] D. Griol, A. Sanchis, and J. Molina, "FRB-Dialog: A Toolkit for Automatic Learning of Fuzzy-Rule Based (FRB) Dialog Managers," in *Proc. of International Conference on Hybrid Artificial Intelligence Systems (HAIS 2017)*, (Logrono, Spain), pp. 306–317, 2017.

[147] D. Griol, J. Molina, and A. Sanchis, "Integration of context-aware conversational interfaces to develop practical applications for mobile devices," *Journal of Ambient Intelligence and Smart Environments*, vol. 9, no. 5, pp. 561–577, 2017.

[148] D. Griol and Z. Callejas, "A Neural Network Approach to Intention Modeling for User-Adapted Conversational Agents," *Computational Intelligence and Neuroscience*, vol. 8402127, pp. 1–11, 2016.

[149] Z. Callejas, D. Griol, and R. López-Cózar, "Predicting user mental states in spoken dialogue systems," *EURASIP Journal on Advances in Signal Processing*, vol. 6, no. 5, pp. 1–21, 2011.

[150] H. H. Meng, C. Wai, and R. Pieraccini, "The Use of Belief Networks for Mixed-Initiative Dialog Modeling," in *IEEE Transactions on Speech and Audio Processing*, vol. 11(6), pp. 757–773, 2003.

[151] H. Meng, S. Lee, and C. Wai, "CU FOREX: a bilingual spoken dialog system for foreign exchange enquiries," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, vol. 2, (Estambul (Turquía)), pp. 229–232, 2000.

[152] T. Schultz, A. W. Black, S. Vogel, and M. Woszczyna, "Flexible speech translation systems," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 403–411, 2006.

[153] R. Sahraeian and D. V. Compernolle, "A study of rank-constrained multilingual DNNS for low-resource ASR," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*, (Shanghai, China), pp. 5420–5424, 2016.

[154] J. Nouza, R. Safarik, and P. Cerva, "ASR for South Slavic Languages Developed in Almost Automated Way," in *Proc. of 17th Annual Conference of the International Speech Communication Association (InterSpeech'16)*, (San Francisco, CA, USA), pp. 3868–3872, 2016.

[155] R. Picard, *Affective computing*. The MIT Press, Cambridge, 1997.

[156] G. Johnson, *Theories of emotion*. Internet Encycl Philos, 2009.

[157] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. Wiley, 2013.

[158] R. Calvo, S. D'Mello, J. Gratch, and A. K. (eds), *The Oxford handbook of affective computing*. Oxford University Press, 2014.

[159] Z. Callejas and R. López-Cózar, "Influence of contextual information in emotion annotation for spoken dialogue systems," *Speech Communication*, vol. 50, no. 5, pp. 416–433, 2008.

[160] C. Nass and K. Lee, "Does computer-generated speech manifest personality? An experimental test of similarity-attractione," in *Proc. of SIGCHI Conference on Human Factors in Computing Systems (CHI'00)*.

[161] C. Nass and C. Yen, *The man who lied to his laptop: what we can learn about ourselves from our machines*. Penguin Group, New York, 2012.

[162] Z. Callejas, D. Griol, and R. López-Cózar, "A framework for the assessment of synthetic personalities according to user perception," *International Journal of Human-Computer Studies*, vol. 72, no. 7, pp. 567–583, 2014.

[163] D. C. Burnett, M. R. Walker, and A. Hunt, "Speech synthesis markup language (ssml) version 1.0," *W3C recommendation*, vol. 7, 2004.

[164] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Hypertext transfer protocol–http/1.1," tech. rep., 1999.

[165] I. Fette and A. Melnikov, "The websocket protocol," tech. rep., 2011.

[166] J. Barnett, R. Akolkar, R. Auburn, M. Bodell, D. Burnett, J. Carter, S. McGlashan, T. Lager, M. Helbing, R. Hosn, *et al.*, "State chart xml (scxml) state machine notation for control abstraction. w3c recommendation," 2015.

[167] J. Barnett, M. Bodell, D. Dahl, I. Kliche, J. Larson, B. Porter, *et al.*, "Multimodal architecture and interfaces. w3c recommendation," 2012.

[168] M. Johnston, D. Dahl, T. Denny, and N. Kharidi, "Emma: Extensible multimodal annotation markup language version 2.0," *World Wide Web Consortium. http://www. w3. org/TR/emma20/. Accessed*, vol. 16, 2015.

[169] J. Pustejovsky, K. Lee, H. Bunt, and L. Romary, "Iso-timeml: An international standard for semantic annotation.," in *LREC*, vol. 10, pp. 394–397, 2010.

[170] H. Bunt, M. Kipp, and V. Petukhova, "Using diaml and anvil for multimodal dialogue annotations.," in *LREC*, pp. 1301–1308, 2012.

[171] N. Almeida, S. Silva, A. Teixeira, and D. Vieira, *Multi-Device Applications Using the Multimodal Architecture*, pp. 367–383. Cham: Springer International Publishing, 2017.

[172] K. Ashimura, O. Nakamura, and M. Isshiki, "Accessible tv based on the w3c mmi architecture," in *2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE)*, pp. 157–158, IEEE, 2014.

[173] F. Cutugno, V. A. Leano, R. Rinaldi, and G. Mignini, "Multimodal framework for mobile interaction," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 197–203, ACM, 2012.

[174] P. Manchón, C. del Solar, G. Amores, and G. Pérez, "The mimus corpus," in *Proc. of LREC 2006 International Workshop on Multimodal Corpora From Multimodal Behaviour Theories to Usable Models*, pp. 56–59, 2006.

[175] R. Porzel, H.-P. Zorn, B. Loos, and R. Malaka, "Towards a separation of pragmatic knowledge and contextual information," in *Proceedings of ECAI 06 Workshop on Contexts and Ontologies*, pp. 5–9, 2006.

[176] M. Pous and L. Ceccaroni, "Multimodal interaction in distributed and ubiquitous computing," in *2010 Fifth International Conference on Internet and Web Applications and Services*, pp. 457–462, IEEE, 2010.

[177] D. Schnelle-Walka, S. Radomski, and M. Mühlhäuser, "Jvoicexml as a modality component in the w3c multimodal architecture," *Journal on Multimodal User Interfaces*, vol. 7, no. 3, pp. 183–194, 2013.

[178] J. L. N. Mesa, I. G. Moreno, E. H. Pérez, A. Ortega, P. Q. Morales, A. R. García, A. Teixeira, and D. T. Toledano, *Advances in Speech and Language Technologies for Iberian Languages: Second International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, November 19-21, 2014. Proceedings*. Springer International Publishing, 2014.

[179] D. A. Dahl, "Standard portals for intelligent services," in *Multimodal Interaction with W3C Standards*, pp. 257–269, Springer, 2017.

[180] I. Zukerman and D. Litman, "Natural language processing and uyser modeling: Synergies and limitations," in *User modeling and user-adapted interaction*, vol. 11, pp. 129–158, 2001.

[181] W. Eckert, E. Levin, and R. Pieraccini, "User modeling for spoken dialogue system evaluation," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'97)*, (Santa Barbara, California (Estados Unidos)), pp. 80–87, 1997.

[182] K. Scheffler and S. Young, "Simulation of human-machine dialogues," tech. rep., CUED/F-INFENG/TR 355, Cambridge University Engineering Dept., Cambridge (Reino Unido), 1999.

[183] O. Pietquin and R. Beaufort, "Comparing ASR modeling methods for spoken dialogue simulation and optimal strategy learning," in *Proc. of the 9th European Conference on Speech Communication and Technology (Interspeech/Eurospeech'05)*, (Lisboa (Portugal)), pp. 861–864, 2005.

[184] O. Pietquin and T. Dutoit, "A probabilistic framework for dialog simulation and optimal strategy learning," in *IEEE Transactions on Speech and Audio Processing, Special Issue on Data Mining of Speech, Audio and Dialog*, vol. 14, pp. 589–599, 2005.

[185] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young, "Effects of the User Model on Simulation-based Learning of Dialogue Strategies," in *Proc. of IEEE Workshop Automatic Speech Recognition and Understanding (ASRU'05)*, (San Juan (Puerto Rico)), pp. 220–225, 2005.

[186] J. Schatzmann, K. Georgila, and S. Young, "Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems," in *Proc. of 6th SIGdial Workshop on Discourse and Dialogue*, (Lisboa (Portugal)), pp. 45–54, 2005.

[187] F. Torres, E. Sanchis, and E. Segarra, "Learning of stochastic dialog models through a dialog simulation technique," in *Proc. of the 9th European Conference on Speech Communication and Technology (Eurospeech'05)*, (Lisboa (Portugal)), 2005.

[188] W. Eckert, E. Levin, and R. Pieraccini, "Automatic evaluation of spoken dialogue systems," tech. rep., TR98.9.1, ATT Labs Research, 1998.

[189] K. Scheffler and S. Young, "Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning," in *Proc. of Human Language Technology (HLT'02)*, (San Diego (Estados Unidos)), pp. 12–18, 2001.

[190] K. Scheffler and S. Young, "Corpus-based Dialogue Simulation for Automatic Strategy Learning and Evaluation," in *Proc. of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001). Workshop on Adaptation in Dialogue Systems*, (Pittsburgh (Estados Unidos)), 2001.

[191] K. Scheffler and S. Young, "Probabilistic simulation of human-machine dialogues," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, (Estambul (Turquía)), pp. 1217–1220, 2000.

[192] K. Georgila, J. Henderson, and O. Lemon, "Learning user simulations for information state update dialogue systems," in *Proc. of the 9th European Conference on Speech Communication and Technology (Eurospeech'05)*, (Lisboa (Portugal)), pp. 893–896, 2005.

[193] J. Bos, E. Klein, O. Lemon, and T. Oka, "DIPPER: Description and Formalisation of an Information-State Update Dialogue System Architecture," in *Proc. of 4th SIGdial Workshop on Discourse and Dialogue*, (Sapporo (Japón)), pp. 115–124, 2003.

[194] S. Singh, D. Litman, M. Kearns, and M. Walker, "Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system," in *Journal of Artificial Intelligence Research (JAIR)*, vol. 16, pp. 105–133, 2002.

[195] H. Cuayáhuitl, S. Renals, O. Lemon, and H. Shimodaira, "Human-Computer Dialogue Simulation Using Hidden Markov Models," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'05)*, pp. 290–295, 2005.

[196] F. Torres, *Sistemas de diálogo basados en modelos estocásticos*. PhD thesis, DSIC - UPV, Valencia (España), 2006.

[197] A. Pargellis, H. Kuo, and C. Lee, "Automatic dialogue generator creates user defined applications," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'99)*, (Budapest (Hungria)), pp. 1175–1178, 1999.

[198] A. Pargellis, H. Kuo, and C. Lee, "An automatic dialogue generation platform for personalized dialogue applications," in *Speech Communication*, vol. 42, pp. 329–351, 2004.

[199] J. Glass and E. Weinstein, "SPEECHBUILDER: Facilitating Spoken Dialogue System Development," in *Proc. of European Conference on Speech Communication and Technology*, (Aalborg (Dinamarca)), pp. 1335–1339, 2001.

[200] J. Polifroni, G. Chungand, and S. Seneff, "Towards the Automatic Generation of Mixed-Initiative Dialogue Systems from Web Content," in *Proc. of European Conference on Speech Communications and Technology (Eurospeech'03)*, (Ginebra (Suiza)), pp. 193–196, 2003.

[201] D. J. Litman and S. Pan, "Designing and Evaluating an Adaptive Spoken Dialogue System," in *User Modeling and User-Adapted Interaction*, vol. 12(2-3), pp. 111–137, 2002.

[202] M. Walker, I. Langkilde, J. Wright, A. Gorin, and D. Litman, "Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You," in *Proc. of the North American Meeting of the Association for Computational Linguistics*, (Seattle (Estados Unidos)), pp. 210–217, 2000.

[203] J. Chu-Carroll and J. S. Nickerson, "Evaluating Automatic Dialogue Strategy Adaptation for a Spoken Dialogue System," in *Proc. of the first conference on North American chapter of the Association for Computational Linguistics*, vol. 4, (Seattle (Estados Unidos)), pp. 202–209, 2000.

[204] K. Jokinen, K. Kanto, and J. Rissanen, "Adaptative User Modelling in AthosMail," in *Lecture Notes in Computer Science. Springer*, vol. 3196/2004, pp. 149–158, 2004.

[205] K. Jokinen, K. Kanto, A. Kerminen, and J. Rissanen, "Evaluation of Adaptivity and User Expertise in a Speech-Based E-Mail System," in *Proc. of the 20th ACL International Conference on Computational Linguistic*, (Ginebra (Suiza)), pp. 44–52, 2004.

[206] P. Gieselmann and A. Waibel, "Dynamic extension of a grammar-based dialogue system: Constructing an all-recipes knowing robot," in *Proc. of the 9th International Conference on Spoken Language Processing (Interspeech/ICSLP)*, (Pittsburgh (Estados Unidos)), pp. 449–452, 2006.

[207] H. Holzapfel, “Towards development of multilingual spoken dialogue systems,” in *Proc. of the 2nd Language and Technology Conference (L&T’05)*, (Poznan (Polonia)), 2005.

[208] H. Taylor, A. Yochem, L. Phillips, and F. Martinez, “The “New” Era of Interoperability Dawns,” in *Event-Driven Architecture: How SOA Enables the Real-Time Enterprise*, Addison-Wesley Professional, Feb. 2009.

[209] M. Stigler, “Understanding Serverless Computing,” in *Beginning Serverless Computing: Developing with Amazon Web Services, Microsoft Azure, and Google Cloud* (M. Stigler, ed.), pp. 1–14, Berkeley, CA: Apress, 2018.

[210] Gartner, “Market guide for conversational platforms,” 2018.

[211] S. Sutton and R. Cole, “The cslu toolkit: Rapid prototyping of spoken language systems,” in *Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology*, UIST ’97, (New York, NY, USA), pp. 85–86, ACM, 1997.

[212] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, “A survey of available corpora for building data-driven dialogue systems: The journal version,” *Dialogue and Discourse*, vol. 9, no. 1, pp. 1–49, 2018.

[213] S. Mezza, A. Cervone, E. Stepanov, G. Tortoreto, and G. Riccardi, “Iso-standard domain-independent dialogue act tagging for conversational agents,” in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3539–3551, Association for Computational Linguistics, 2018.

[214] A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, E. King, K. Bland, A. Wartick, Y. Pan, H. Song, S. Jayadevan, G. Hwang, and A. Pettigrue, “Conversational ai: The science behind the alexa prize,” *CoRR*, vol. abs/1801.03604, 2017.

[215] M. Eskénazi, G. Levow, H. M. G. Parent, and D. Suendermann, *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. John Wiley and Sons, 2013.

[216] R. San-Segundo, “La evaluación objetiva de sistemas de diálogo,” in *Curso de Tecnologías Linguísticas. Fundación Duques de Soria*, (Soria (España)), 2004.

[217] M. Walker, D. Litman, C. Kamm, and A. Abella, "Evaluating spoken dialogue agents with PARADISE: Two case studies," in *Computer Speech and Language*, vol. 12, pp. 317–347, 1998.

[218] L. Dybkjaer, N. Bernsen, and W. Minker, "Evaluation and usability of multimodal spoken language dialogue systems," in *Speech Communication*, vol. 43, pp. 33–54, 2004.

[219] W. Minker, A. Waibel, and J. Mariani, *Stochastically-Based Semantic Analysis*. Dordrecht (Holanda): Kluwer Academic Publishers, 1999.

[220] L. Dybkjaer and N. Bernsen, "Usability issues in spoken language dialogue systems," in *Natural Language Engineering (2000). Cambridge University Press*, vol. 6, pp. 243–271, 2000.

[221] EAGLES, "Evaluation of Natural Language Processing Systems," tech. rep., Final Report, EAGLES Document EAG-EWG-PR2. Center for Sprogteknologi, Copenhague (Dinamarca), 1996.

[222] K. Failenschmid, D. Williams, L. Dybkjaer, and N. Bernsen, "DISC Deliverable D3.6," in *www.disc2.dk*, 1999.

[223] L. Devillers, H. Maynard, and S. Rosset, "The French Media/Evalda project: the evaluation of the understanding capability of spoken language dialog systems," in *Proc. of International Conference on Language Resources and Evaluation (LREC'04)*, vol. 6, (Lisboa (Portugal)), pp. 2131–2134, 2004.

[224] L. Degerstedt and A. Jönsson, "LinTest, A development tool for testing dialogue systems," in *Proc. of the 9th International Conference on Spoken Language Processing (Interspeech/ICSLP)*, (Pittsburgh (Estados Unidos)), pp. 489–492, 2006.

[225] Z. Chen and B. Liu, *Lifelong Machine Learning*. Morgan Clayton, 2016.

[226] E. Agirre, S. Marchand, S. Rosset, A. Peñas, and M. Cieliebak, "Lihlith: Improving communication skills of robots through lifelong learning," in *ERCIM News*, 2018.

## Glosario y Términos de Búsqueda