Servicio Andaluz de Salud
CONSEJERÍA DE SALUD

INSTITUTODEBIOMEDICINADESEVILLA

Tecnologías de lenguaje en Sanidad y en Biomedicina

Carlos Luis Parra Calderón
Director del Grupo de Investigación e Innovación en Informática Biomédica,
Ingeniería biomédica y Economía de la Salud.
Instituto de Biomedicina de Sevilla / Hospital Universitario Virgen del Rocío
carlos.parra.sspa@juntadeandalucia.es
Tlf.- 955 01 36 62

Sociedad Española para el
Procesamiento del Lenguaje Natural

Plan TL
Plan de Impulso de las
Tecnologías del Lenguaje

¿Por qué necesitamos la tecnología del Lenguaje?

- Porque es necesario acceder al conocimiento implícito de los textos de las Historias de Salud Electrónica.
- Porque la capacidad de resolver el reto desde la tecnología del mercado de la TIC convencional (**reconocimiento del lenguaje basado en reglas**) resuelve algo pero es *de corto alcance*, **fácil de entender por el usuario pero muy limitado en la capacidad de resolver el problema.**
- En este estado de la tecnología encontramos PRACTICAMENTE TODA LAS SOLUCIONES QUE LA INDUSTRIA HOY OFRECE AL SISTEMA SANITARIO.
- EJEMPLO, LA MAYORÍA DE LOS CLASIFICADORES CIE 10 APLICADOS ACTUALMENTE EN EL SNS.

A vista de pajaro del PLN, algunas pinceladas…

- Adquirir conciencia de la dificultad del reto.
- Hay que adquirir cierta familiaridad con los conceptos del Procesamiento del Lenguaje Natural (PLN, NLP en inglés)
- Identificar el gran retraso y a la vez la gran oportunidad, respecto a la tecnología desarrollada y funcionando en inglés desde hace decadas.
- Separar el grano de la paja:
  - **Aplicar tecnología Big Data facilita el pipeline del PLN pero no necesariamente ni automáticamente aplica PLN de manera coherente ni adecuada, y menos en un/OS dominio/S del lenguaje tan complejo como es la biomedicina y la sanidad!!!.**

Antecedentes

- En los registros médicos electrónicos modernos, la mayoría de los datos clínicamente importantes, signos y síntomas, gravedad de los síntomas, estado de la enfermedad, etc., no se proporcionan en campos de datos estructurados, sino que están almacenados en textos narrativos generados por los clínicos.
- El Procesamiento del Lenguaje Natural (PLN) proporciona un medio para "desbloquear" esta importante fuente de datos, convirtiendo el texto no estructurado en datos estructurados y procesables para su uso en aplicaciones como:
  - **Soporte a la decisión clínica,**
  - Seguimiento de la calidad asistencial o
  - Vigilancia epidemiológica en salud pública.
- En el mundo anglosajón, existen muchos sistemas de PLN que se han aplicado con éxito en textos biomédicos.

Los orígenes del PLN en Medicina: Linguistic String Project (LSP)

- El “Linguistic String Project (LSP)” fue un proyecto inicial que comenzó **en 1965** y que se centró en el procesamiento del lenguaje médico.
- El proyecto creó un nuevo esquema para representar el texto clínico y un diccionario de términos médicos, además de abordar varios problemas clínicos claves de PLN, tales como:
  - la anonimización,
  - el análisis sintáctico,
  - el mapeo y normalización.
- La metodología y la arquitectura de LSP han influido sustancialmente en muchos sistemas clínicos de PLN posteriores.

Claves sobre la evolución de la tecnología PLN fundamentales para la innovación.

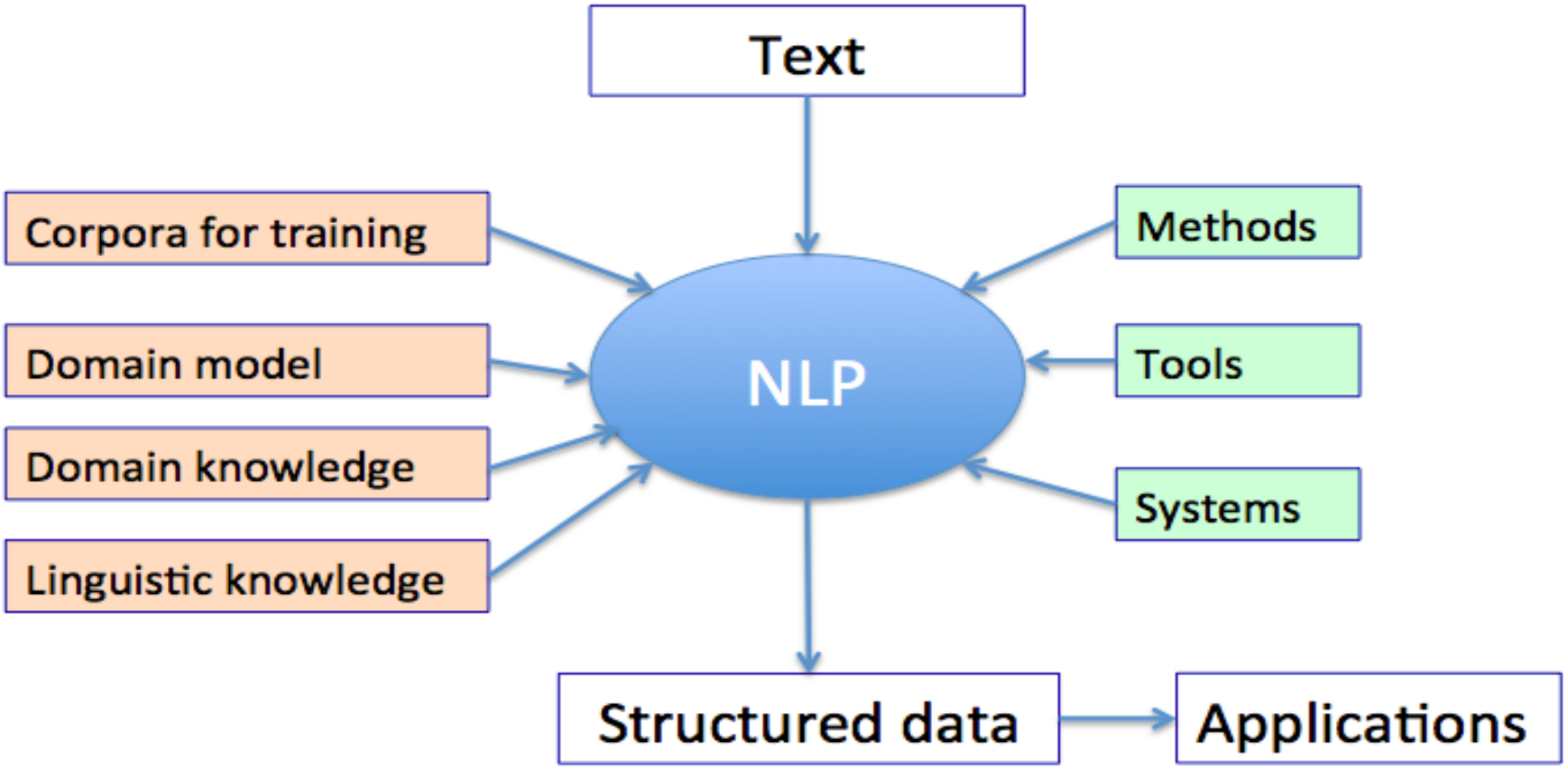
- Se evidencia una evolución desde software monolítico construido sobre plataformas que estaban disponibles en el momento en que fueron desarrollados a sistemas contemporáneos basados en componentes construidos en base a frameworks generales.
- El rendimiento de estos sistemas está estrechamente asociado con sus "ingredientes" (es decir, los módulos que se utilizan para formar su conocimiento previo), y cómo estos módulos se combinan a partir del framework general.

Frameworks

- Se han desarrollado varios frameworks de software que facilitan la integración de diferentes herramientas en un único canal:
  - GATE (General Architecture for Text Engineering)
  - UIMA (Unstructured Information Management Architecture)

Componentes básicos de un Sistema de PLN

Fig. 1. Los rectángulos en el lado izquierdo representan el conocimiento previo, y los componentes en el lado derecho representan el framework (es decir, algoritmos y herramientas). El conocimiento previo y el framework son los principales componentes de un sistema de PLN.



Doan S, Conway M, Phuong TM, Ohno-Machado L. Natural language processing in biomedicine: a unified system architecture overview. Methods Mol Biol.2014;1168:275-94. doi: 10.1007/978-1-4939-0847-9_16. Review. PubMed PMID:24870142.

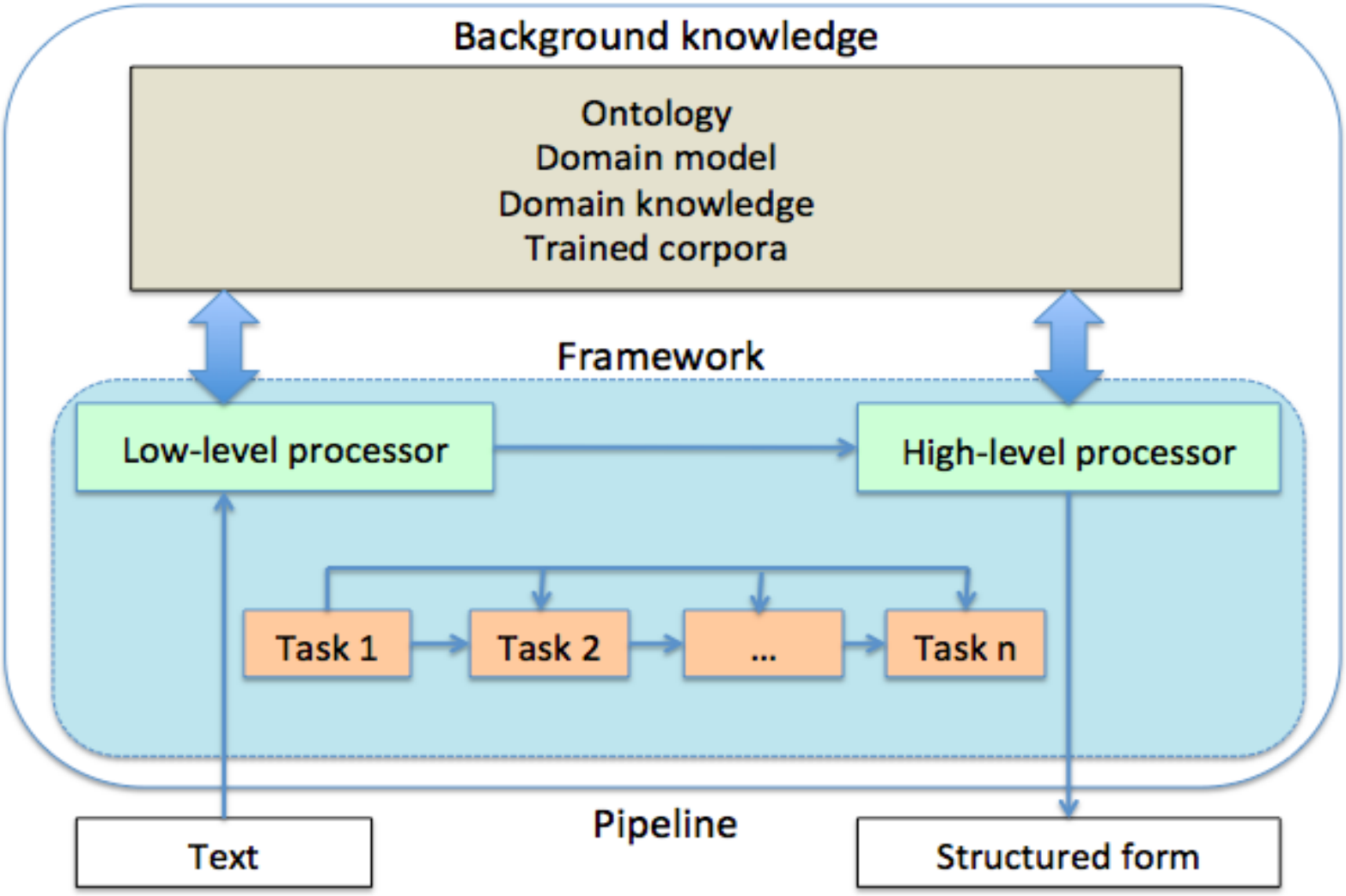
Arquitectura de un Sistema de PLN

Una arquitectura general de un sistema de PLN clínico contiene dos componentes principales:

- **El conocimiento previo** contiene ontologías, un modelo de dominio, conocimiento de dominio y corpus entrenados.
- **El framework** incluye:
  - Un procesador de bajo nivel para tareas como la tokenización (segmentación de palabras) y el etiquetado de categorías gramaticales.
  - Un procesador de alto nivel se utiliza para tareas tales como el reconocimiento de entidades nombradas y la extracción de relaciones.
  - Las tareas o módulos en el framework pueden ser dependientes o independientes y están organizados secuencial o jerárquicamente.

Doan S, Conway M, Phuong TM, Ohno-Machado L. Natural language processing in biomedicine: a unified system architecture overview. Methods Mol Biol.2014;1168:275-94. doi: 10.1007/978-1-4939-0847-9_16. Review. PubMed PMID:24870142.

Arquitectura de un Sistema de PLN



Doan S, Conway M, Phuong TM, Ohno-Machado L. Natural language processing in biomedicine: a unified system architecture overview. Methods Mol Biol.2014;1168:275-94. doi: 10.1007/978-1-4939-0847-9_16. Review. PubMed PMID:24870142.

Algunos Sistemas de PLN y sus características

Table 1. Summary of characteristic features of some representative clinical NLP systems.

System	Programming language	Creator	Framework	Open/Closed source and License	Background knowledge resource	Clinical domain or source of information	Encoding
LSP-MLP	Fortran C++	New York University		Software provided by Medical Language Processing LLC corporation	Developed its own medical lexicons and terminologies	Progress note, clinical note, X-ray report, discharge summary	SNOMED
MedLEE	Prolog	Columbia University		Closed source Commercialized by Columbia University and Health Fidelity, Inc.	Developed its own medical lexicons (MED) and terminologies	Radiology Mammography, discharge summary	UMLS's CUI
SPRUS/ SymText/ MPLUS	LISP, C++	University of Utah		Closed source	UMLS	Radiology Concepts from findings in radiology reports	ICD-9
MetaMap	Perl, C, Java, Prolog	National Library of Medicine		Not open source but free available under UMLS Metathesaurus License Agreement	UMLS	Biomedical text Candidate and mapping concepts from UMLS	UMLS's CUI
HITEx	Java	Harvard University	GATE	Open source i2b2 software license	UMLS	Clinical narrative Family history concept, temporal concepts, smoking status, principal diagnosis, co-morbidity, negation	UMLS's CUI
cTAKES	Java	Mayo clinic and IBM	UIMA	Open source Apache 2.0	UMLS + Trained models	Discharge summary, clinical note Clinical named entities (diseases/disorders, signs/symptoms, anatomical sites, procedures, medications), relation, co-reference, smoking status classifier, side effect annotator	UMLS's CUI and RxNorm

Doan S, Conway M, Phuong TM, Ohno-Machado L. Natural language processing in biomedicine: a unified system architecture overview. Methods Mol Biol.2014;1168:275-94. doi: 10.1007/978-1-4939-0847-9_16. Review. PubMed PMID:24870142.

¿Pero que hace la tecnología PLN???

cTAKES, Sistema de Análisis de Textos Clínicos y Generación de Conocimiento.

- Conjunto de componentes ejecutados secuencialmente (PIPELINE) para procesar textos clínicos.
- Cada componente añade anotaciones respecto a los anteriores.
- Construido sobre tecnologías de código abierto (UIMA y OpenNLP).
- Combina técnicas basadas en reglas y aprendizaje automático.
- Los conjuntos de datos “Gold-Standard” para las anotaciones lingüísticas y los conceptos clínicos se obtienen de un subconjunto de notas clínicas de la Clínica Mayo (Rochester, Minnesota).

Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA. 2010;17(5):507-513. doi:10.1136/jamia.2009.001560.

Componentes de cTAKES (I)

- Segmentador de frases.
- Segmentador de palabras.
- Normalizador.
  - Homogeneiza el texto (ej. consideración de los términos en mayúscula o minúscula; control de abreviaturas y acrónimos, eliminación de palabras vacías, etc.)
- Etiquetador gramatical.
  - Asigna a cada palabras su categoría gramatical.

Componentes de cTAKES (y II)

- Análizador sintáctico superficial.
  - Identifica los elementos constituyentes de una frase (grupos nominales, verbos, etc.).
- Reconocedor de entidades nombradas.
  - Clasifica las entidades nombradas en categorías predefinidas (personas, organizaciones, lugares, expresiones de tiempo, cantidades, etc.).
  - Incluye detección de la negación.

Ejemplo obtenido de cTAKES

An example of a sentence discovered by the sentence boundary detector:
Fx of obesity but no fx of coronary artery diseases.

Tokenizer output – 11 tokens found:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  diseases  .
```

Normalizer output:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  disease   .
```

Part-of-speech tagger output:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  diseases  .
NN  IN  NN       CC   DT  NN  IN  JJ         NN       NNS       .
```

Shallow parser output:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  diseases  .
NP  PP  NP       O    NP  NP  PP  ---------- NP -------------
```

Named Entity Recognition – 5 Named Entities found:

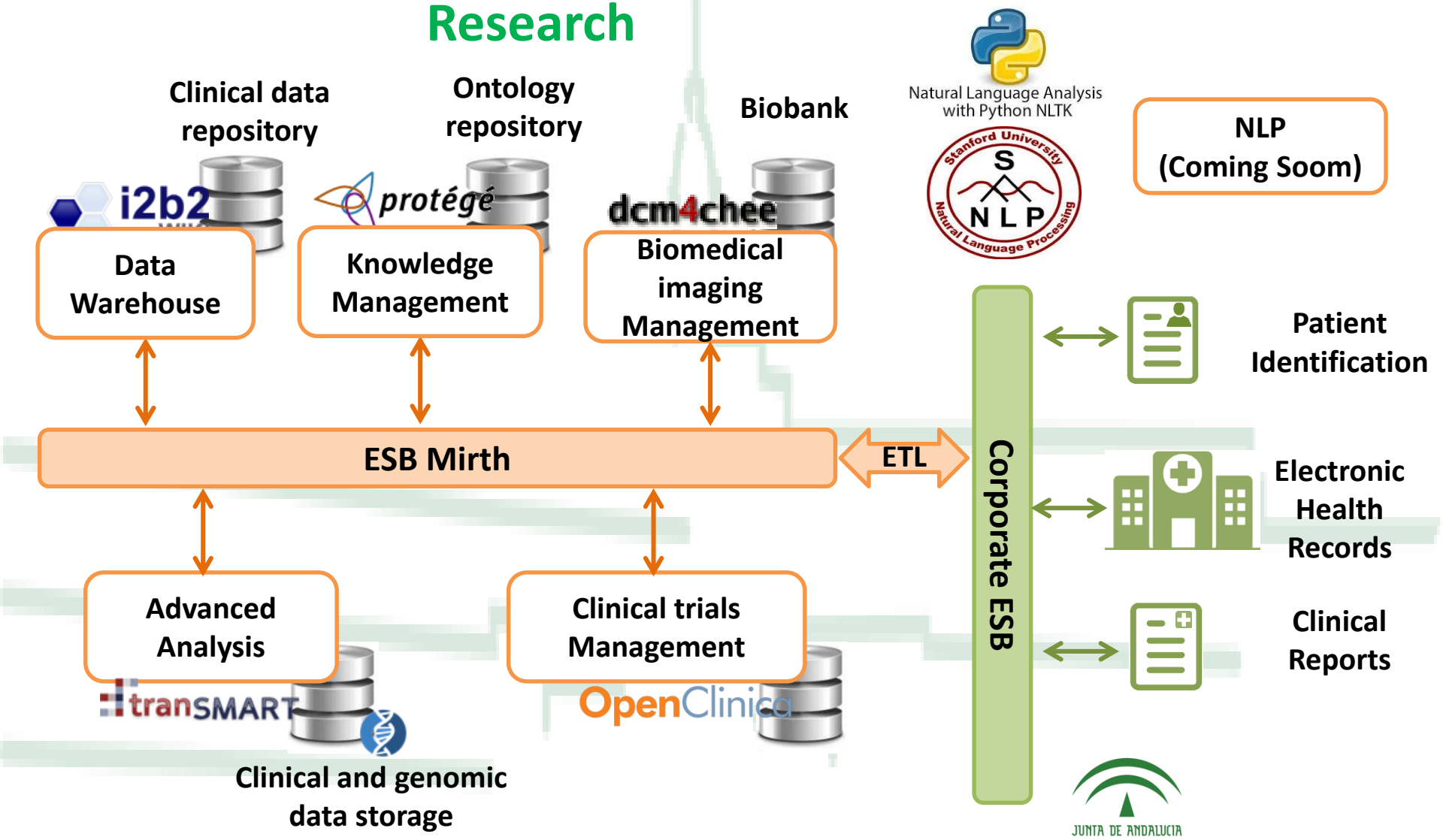
```
Fx of obesity but no fx of coronary artery diseases .
      obesity  (type=diseases/disorders, UMLS CUI=C0028754, SNOMED-CT codes=308124008 and 5476005)
                        coronary artery diseases (type=diseases/disorders, CUI=C0010054, SNOMED-CT=8957000)
                        coronary artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
                                artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
                                diseases (type=diseases/disorders, CUI = C0010054)
```

Status and Negation attributes assigned to Named Entities:

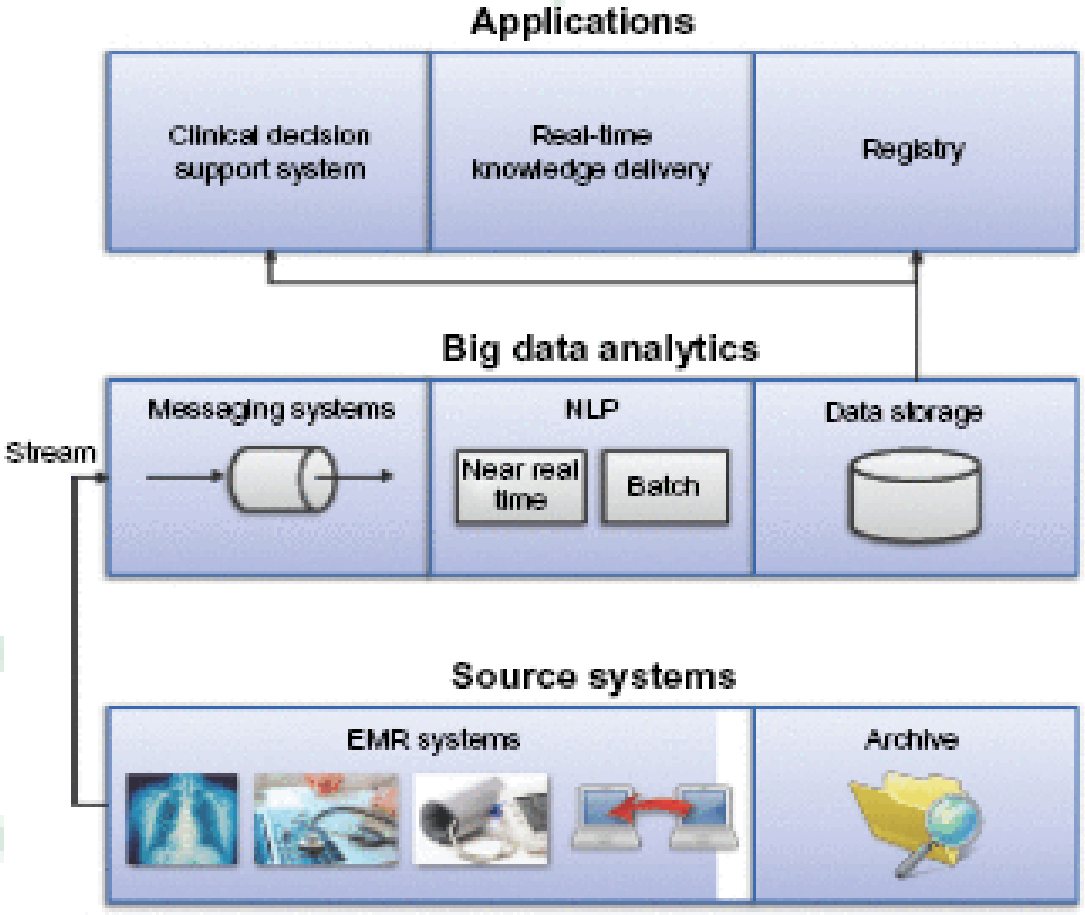
```
Fx of obesity but no fx of coronary artery diseases .
      obesity (status = family_history_of; negation = not_negated)
                        coronary artery diseases (status = family_history_of, negation = is_negated)
```

Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA. 2010;17(5):507-513. doi:10.1136/jamia.2009.001560.

Virgen del Rocío University Hospital IT Ecosystem for Translational & Clinical Research



Big Data combinado con tecnología PLN, permite el tiempo real: experiencia de la Clinica Mayo.



Kaggal VC, Elayavilli RK, Mehrabi S, Pankratz JJ, Sohn S, Wang Y, Li D,Rastegar MM, Murphy SP, Ross JL, Chaudhry R, Buntrock JD, Liu H. Toward a Learning Health-care System - Knowledge Delivery at the Point of Care Empowered by Big Data and NLP. Biomed Inform Insights. 2016 Jun 23;8(Suppl 1):13-22. doi:10.4137/BII.S37977. eCollection 2016. PubMed PMID: 27385912; PubMed Central PMCID: PMC4920204.

Dominios de aplicación actuals según IMIA

- Representación de enfermedades
  - Enfoques supervisados en la representación de enfermedades
  - Enfoques semi-supervisados para la generación de fenotipos
  - Más allá del modelado de una única enfermedad
- Selección de cohortes de pacientes
  - Extracción de características de cohortes a partir de textos clínicos
  - Más allá de las reglas de correspondencia
- Otros usos secundarios de los datos clínicos
  - Análisis predictivo
  - Farmacovigilancia y nuevas indicaciones de los fármacos
  - Caracterización de poblaciones y patrones de asistencia
- Apoyo a la gestión hospitalaria
  - Apoyo a la generación de indicadores de gestión y de generación automatizada de informes
    - Gestión de la calidad asistencial.
  - **Ayuda a la decisión clínica**
- Apoyo a las necesidades de los individuos y poblaciones
  - PLN en el lenguaje del usuario/paciente
  - Redes sociales como fuente para la evaluación de la calidad asistencial
  - Comprendiendo las preguntas online de los usuarios/pacientes

Demner-Fushman D, Elhadad N. Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing. Yearb Med Inform. 2016 Nov 10;(1):224-233.PubMed PMID: 27830255.

Taller en MEDINFO 2017 donde hemos contribuído con algunos líderes mundiales en PLN en Biomedicina y Sanidad

Unstructured Clinical Data Reuse for Precision Medicine: Current Status and Future Progress

Stéphane M. Meystre[a], Christian Lovis[b], Thomas Bürkle[c], Hua Xu[d], Guergana Savova[e], Hongfang Liu[f], Angus Roberts[g], Ronald Cornet[h], Carlos Luis Parra Calderón[i], Andrius Burdionis[j], Christoph U. Lehmann[k]

[a] *Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, USA*
[b] *Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland*
[c] *Bern University of Applied Sciences, Biel, Switzerland*
[d] *School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX, USA*
[e] *Harvard University and Boston Children's Hospital, Boston, MA, USA*
[f] *Section of Medical Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA*
[g] *Department of Computer Science, University of Sheffield, Sheffield, UK*
[h] *Department of Medical Informatics, University of Amsterdam, Amsterdam, Netherlands*
[i] *Institute of Biomedicine of Seville, Seville, Spain*
[j] *Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway*
[k] *Departments of Biomedical Informatics and Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA*

NCBI Resources How To clparrac My NCBI Sign Out

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | Expanding the Diversity of Texts and Applications: Findings from the Section on Clinical | Search

Create RSS Create alert Advanced Help

Format: Abstract Send to

See 1 citation found by title matching your search:

Yearb Med Inform. 2018 Aug;27(1):193-198. doi: 10.1055/s-0038-1667080. Epub 2018 Aug 29.

Expanding the Diversity of Texts and Applications: Findings from the Section on Clinical Natural Language Processing of the International Medical Informatics Association Yearbook.

Névéol A[1], Zweigenbaum P[1]; Section Editors for the IMIA Yearbook Section on Clinical Natural Language Processing.

Author information

1 LIMSI, CNRS, Université Paris-Saclay, Orsay, France.

Full text links

FREE Full text THIEME CONNECT

Save items

Favorite

Similar articles

Review Making Sense of Big Textual Data for Health Care: Fir [Yearb Med Inform. 2017]

Gracias por vuestra atención

Carlos Luis Parra Calderón
Director del Grupo de Investigación e Innovación en Informática Biomédica, Ingeniería biomédica y Economía de la Salud.
Instituto de Biomedicina de Sevilla / Hospital Universitario Virgen del Rocío
carlos.parra.sspa@juntadeandalucia.es
Tlf.- 955 01 36 62

Servicio Andaluz de Salud
CONSEJERÍA DE SALUD