Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA

Tecnologías del Lenguaje para el sector I+D

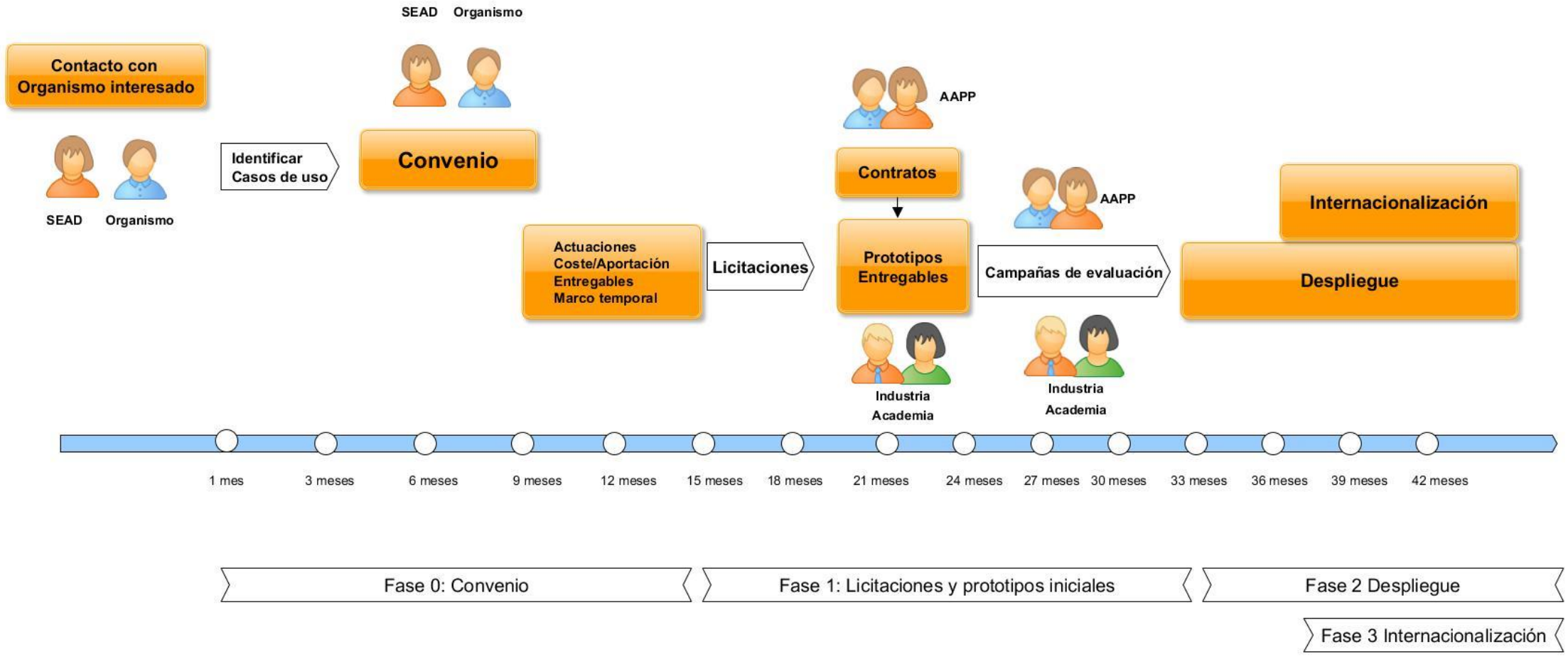
Doaa Samy
Oficina Técnica Genera del Plan TL- Personal Externo
Advanced Lingüista Computacional
Instituto de Ingeniería del Conocimiento-IIC

Líneas generales

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

1. Del Plan TL al Corpus Viewer: ¿Cómo hemos llegado hasta aquí?
2. ¿Por qué utilizar TL en I+D+i?
    - 2.1. El ámbito I+D+i y sus características
    - 2.2. ¿Qué ventajas ofrecen las TL?
    - 2.3. ¿Cómo aplicamos TL?
3. ¿Qué retos y limitaciones supone?
4. Otros aspectos

Del Plan TL al Proyecto Faro de Inteligencia Competitiva



Líneas generales

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

1. Del Plan TL al Corpus Viewer: ¿Cómo hemos llegado hasta aquí?
2. ¿Por qué utilizar TL en I+D+i?
    - 2.1. El ámbito I+D+i y sus características
    - 2.2. ¿Qué ventajas ofrece TL?
    - 2.3. ¿Cómo aplicamos TL?
3. ¿Qué retos y limitaciones supone?
4. Otros aspectos

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

2.1. El ámbito I+D+i y sus características

¿Cuál es el espacio de la información?

¿Qué dinámicas y qué características tiene este espacio de información?

¿Cuáles el estado de la cuestión? ¿Existen otras iniciativas?

¿Cuál es el espacio de la información?

- **El espacio de la información:** El ámbito I+D+i desde el punto de vista de este proyecto implica una serie de datos directamente relacionados con I+D+i (Publicaciones, Ayudas I+D+i y Patentes) y otros indirectamente relacionados (Formación, Empleo, Marcas y Contratación Pública).

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Características y dinámicas de este espacio:

-Volumen: Los volúmenes de datos en los repositorios son cada vez mayor

- Variedad: Las fuentes presentan una variedad de formatos y metadatos

- Velocidad: Las publicaciones y las patentes crecen con una velocidad marcada.

→ Estos rasgos requieren unas herramientas capaces de abordarlo complementándose con los enfoques estadísticos para ofrecer una visión integral que reúna las ventajas de los diferentes enfoques.

¿Cuáles el estado de la cuestión? ¿Existen otras iniciativas?

Estudios interdisciplinarios:

- Sciencemetrics
- Innovation policies

Herramientas:

- CoreText
- ScienceMiner
- Intelligo

Iniciativas y proyectos europeos en Inteligencia Competitiva

- RISIS2
- Data4Impact

Líneas generales

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

1. Del Plan TL al Corpus Viewer: ¿Cómo hemos llegado hasta aquí?
2. ¿Por qué utilizar TL en I+D+i?
    - 2.1. El ámbito I+D+i y sus características
    - 2.2. ¿Qué ventajas ofrecen las TL?
    - 2.3. ¿Cómo aplicamos TL?
3. ¿Qué retos y limitaciones supone?
4. Otros aspectos

¿Por qué TL? ¿Qué ventajas ofrece el uso de TL?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

En cuanto al espacio de la información, las Tecnologías del Lenguaje nos permite procesar los datos no-estructurados, y por tanto nos permite:

- Reducir el espacio heterogéneo de los datos
- Identificar y extraer la información relevante
- Generar una **representación semántica** de un espacio de información dotada de significado
- Inferir **agrupaciones temáticas** y calcular **comunidades**
- Conseguir un grado de **granularidad** en cuanto a temáticas subyacentes
- Ofrecer **transversalidad** para mover entre los diferentes espacios

¿Por qué TL? ¿Qué funcionalidades ofrece el uso de TL?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

En cuanto a funcionalidades, las TL nos permite

- Cartografiar el espacio temático o **el paisaje científico** "Landscape"
- Detectar **duplicidades**
- Detectar **evolución temática a través del tiempo**
- Detectar las **hibridaciones** o **emergencias temáticas**
- **Clasificar** los datos basándose en el contenido
- **Perfilar** agentes por sus productos científicos en el espacio de información
- Aportar algunas indicaciones para ayudar a medir **el impacto** por ejemplo a través de *Lead Lag*

Líneas generales

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

1. Del Plan TL al Corpus Viewer: ¿Cómo hemos llegado hasta aquí?
2. ¿Por qué utilizar TL en I+D+i?
    - 2.1. El ámbito I+D+i y sus características
    - 2.2. ¿Qué ventajas ofrecen las TL?
    - 2.3. ¿Cómo aplicamos TL?
3. ¿Qué retos y limitaciones supone?
4. Otros aspectos

¿Cómo aplicamos las TL?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

El uso de TL y ML sirve fundamentalmente para:

1. **Reducir el espacio de los datos heterogéneos**
2. Para **la representación semántica** de los documentos y el cálculo de las **distancias entre documentos**
3. **Para la detección de comunidades y distancias entre comunidades**

¿Cómo aplicamos las TL?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Para reducir el espacio de los datos:

- Se filtran por un conjunto de palabras claves por dominio
- NLP Pipeline [POS y Lematizar]
- Identificar Entidades Nombradas NER utilizando DBPedia
- Seleccionar las categorías: Sustantivo, verbo, adjetivo
- Filtrar los Stopwords
- TF-IDF
- MT para algunos datos

¿Cómo aplicamos las TL?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Para **la representación semántica** de los documentos y las **distancias entre documentos** se aplica lo siguiente:

- Técnicas de Recuperación de Información (IR) para similitud entre documentos se calculas las distancias inter-documentales y los modelos de tópicos basándose en:
  - Representación de BoW y distancias de BM25 para indexación.
  - Representación de W2V + Word Moving Distance y distancias de Coseno para detección de alarmas.
- Representación de tópicos LDA con distancias JS para detectar alarmas y como base para detectar comunidades.

¿Cómo aplicamos las TL?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

- **Para la detección de comunidades y distancias entre comunidades se aplica:**
- Se detectan las comunidades utilizando Louvaine
- Las distancias entre comunidades se calculan con *Personalized Page Rank*

Líneas generales

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

1. Del Plan TL al Corpus Viewer: ¿Cómo hemos llegado hasta aquí?
2. ¿Por qué utilizar TL en I+D+i?
    - 2.1. El ámbito I+D+i y sus características
    - 2.2. ¿Qué ventajas ofrecen las TL?
    - 2.3. ¿Cómo aplicamos TL?
3. ¿Qué retos y limitaciones supone?
4. Otros aspectos

3. ¿Cuáles son los retos y las limitaciones?

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

- Aspectos metodológicos:
  - Metodología de seleccionar el espacio de información
  - Validación heterogeneidad de las fuentes
  - Las técnicas basadas en datos no estructurados no son 100% precisas
- Aspectos legales:
  - Apertura de datos
  - Datos personales

Líneas generales

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

1. Del Plan TL al Corpus Viewer: ¿Cómo hemos llegado hasta aquí?
2. ¿Por qué utilizar TL en I+D+i?
    - 2.1. El ámbito I+D+i y sus características
    - 2.2. ¿Qué ventajas ofrecen las TL?
    - 2.3. ¿Cómo aplicamos TL?
3. ¿Qué retos y limitaciones supone?
4. Otros aspectos

4. Otros aspectos

- Medición de impacto
- Ciencia abierta
- Cantidad vs. calidad

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

¡Gracias!

www.PlanTL.gob.es

plantecnologiaslenguaje@mineco.es

Plan TL - Tecnologías Lenguaje