Sesión 2:
Componentes e infraestructuras de tecnologías del lenguaje para sanidad

Marta Villegas
Jordi Armengol
Alejandro Asensio
(BSC-CNS)

Infoday sobre tecnologías del lenguaje,
Barcelona Supercomputing Center - BSC, Barcelona

02/12/2019

1. **Acciones y líneas de trabajo para fomentar el desarrollo y adaptación de componentes de software en el ámbito de la salud.**

2. **Desarrollo de componentes de PLN médico basados en Inteligencia Artificial (IA) y computación de alto rendimiento (HPC)**

Acciones y líneas de trabajo

1 Estudios

2 Subcon-tratas

3 Campañas

4 Casos de uso

5 Datos

- Identificar components básicos y ‘transversales’ de alto impacto en el dominio de la salud.
- Favorecer el desarrollo/adaptación de components.
- Garantizar interoperabilidad, escalabilidad, estandarización.
- Disponer de entornos de benchmarking que permitan comparativas.
- Disponer de datos suficientes para entrenar y evaluar.

Acciones y líneas de trabajo

1 Estudios

Subcon-tratas

Campañas

Casos de uso

Datos

https://www.plantl.gob.es/

- Diseño de plataformas de procesamiento del lenguaje natural y traducción automática en el área de la Biomedicina
- Estándares de Interoperabilidad
- Informe sobre el estado de infraestructuras lingüísticas en el área de la Biomedicina
- Estudio para la ampliación de UMLS
- Estudio y propuesta de técnicas reconocimiento de secciones en informes clínicos
- Protección de datos personales
- Modelos de licencias y Políticas de licencias
- Modelos de sostenibilidad y Reutilización de recursos y procesadores
- Política de reutilización de la información de interés lingüístico del sector público
- Estudio y propuesta de técnicas reconocimiento de secciones en informes clínicos
- Campañas de evaluación
- ….

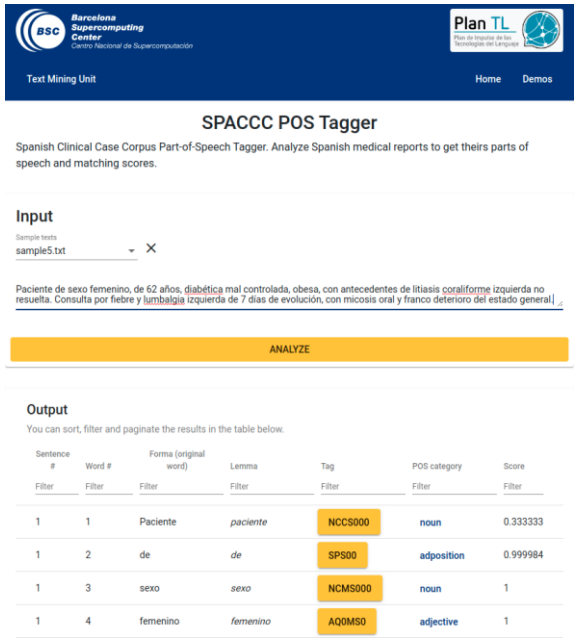
Acciones y líneas de trabajo

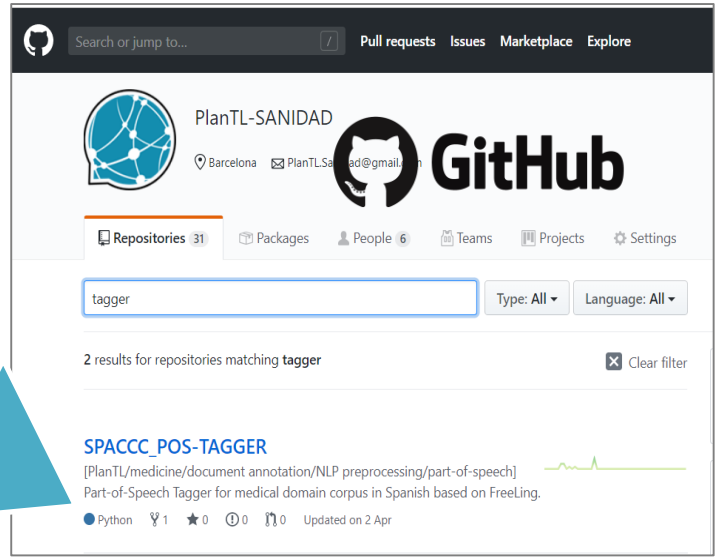
Estudios | **2 Subcon-tratas** | Campa-ñas | Casos de uso | Datos

FreeLing

instituto de ingeniería del conocimiento

demo





Guías de anotación
textos médicos en español (segmentacion,tokenización, anotación morfosintáctica.

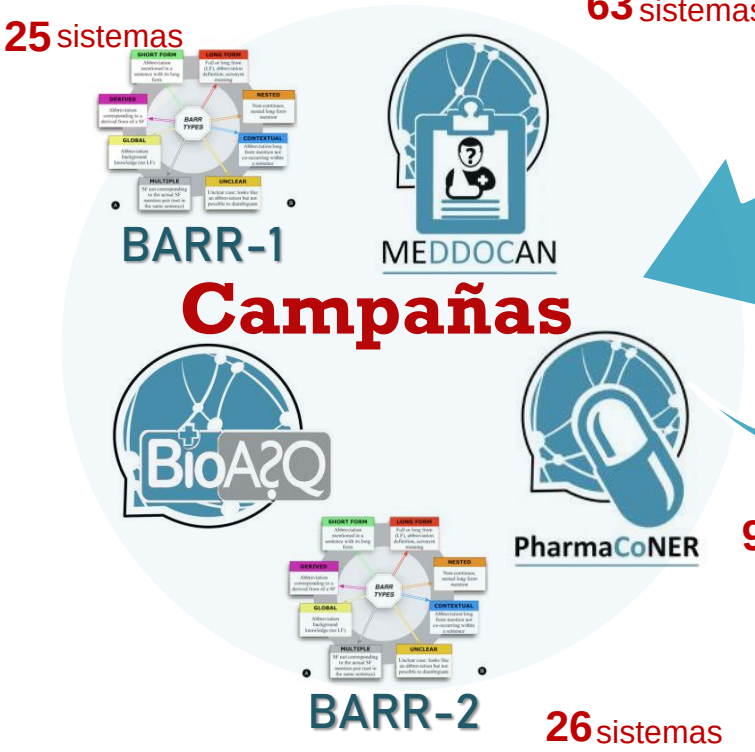
Gold Standard
corpus anotado de 1000 casos clínicos SPACCC

	Split	Token	POS
GS vs FrL	99.52%	99.97%	98.71%
Acuerdo Mínimo Requerido	99%	98%	96%

Tabla 8.Porcentaje de acuerdo sobre el 10% de validación

Acciones y líneas de trabajo

Estudios | Subcon-tratas | **3 Campañas** | Casos de uso | Datos

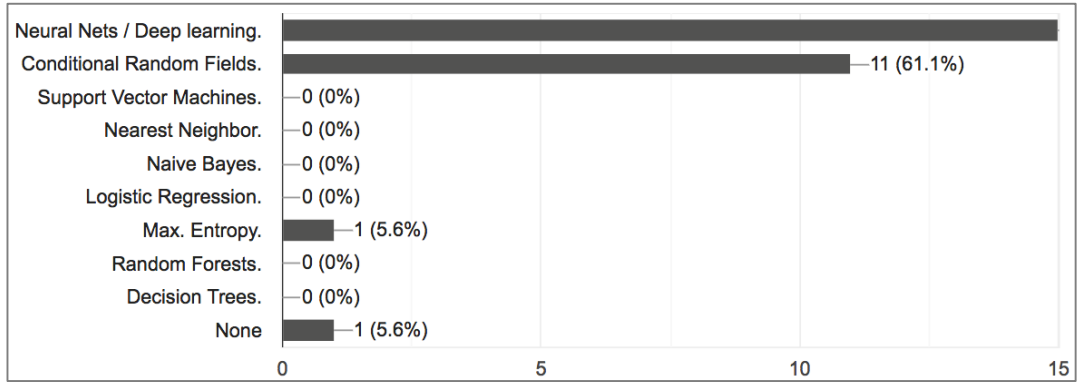


Evaluación

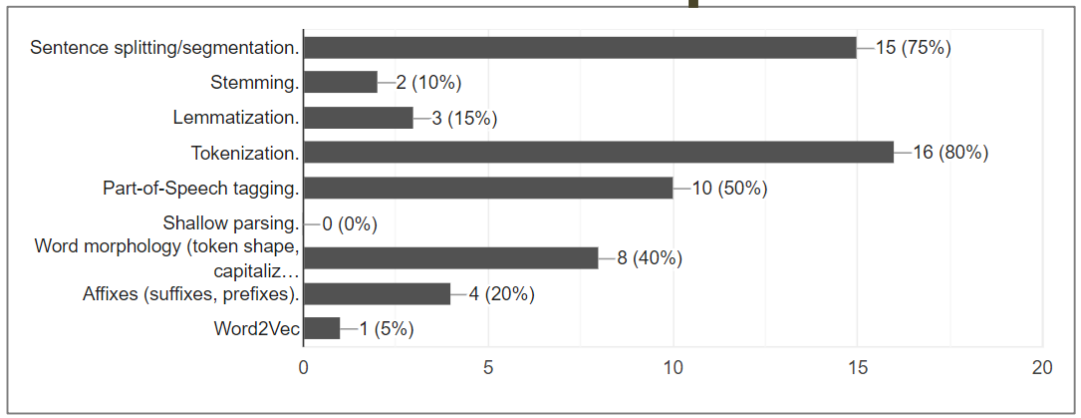
Precision	Recall	F1
0.91226	0.90879	0.91052
0.91589	0.90445	0.91013
0.91008	0.90662	0.90835
0.90751	0.90554	0.90652
0.90205	0.90988	0.90595
0.90625	0.91314	0.90968
0.90708	0.89082	0.89888
0.89297	0.89685	0.89491
0.88839	0.86369	0.87586

Silver Standards

Tecnologías

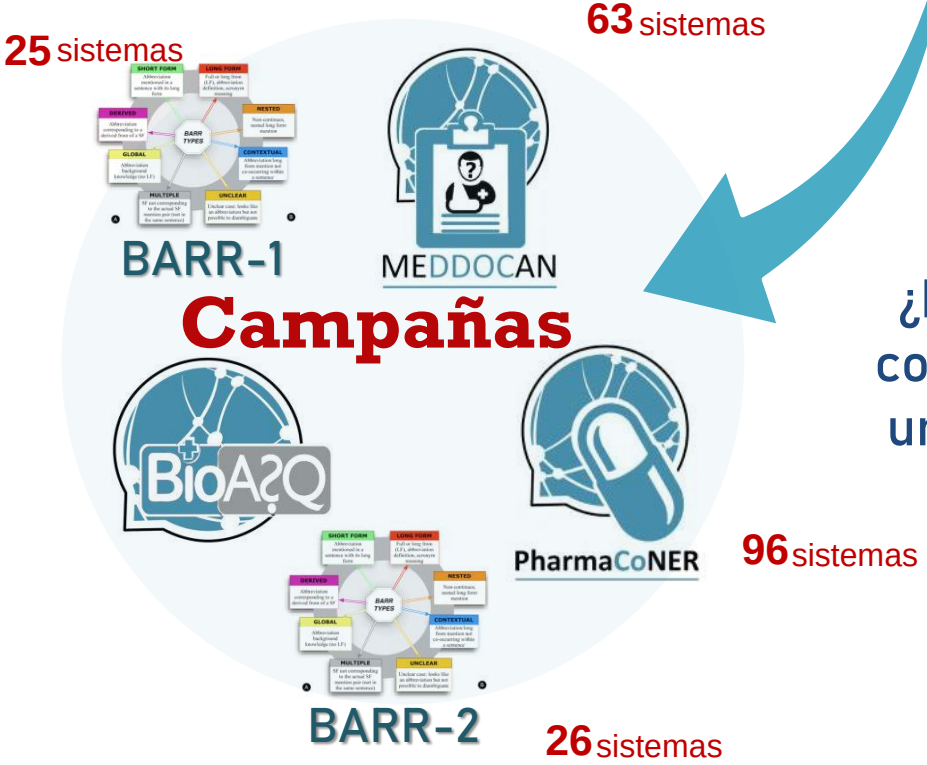


Componentes PLN

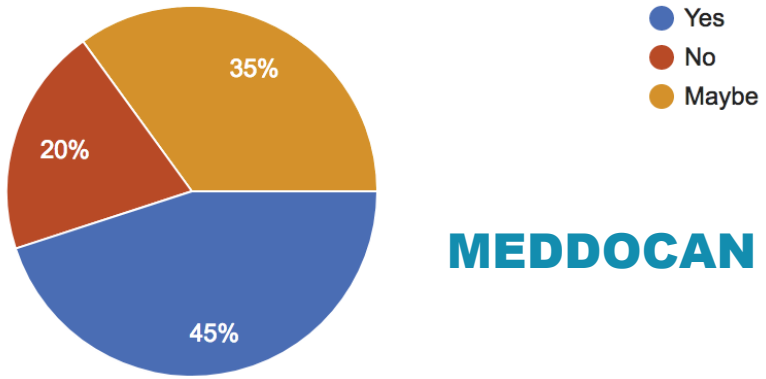


Acciones y líneas de trabajo

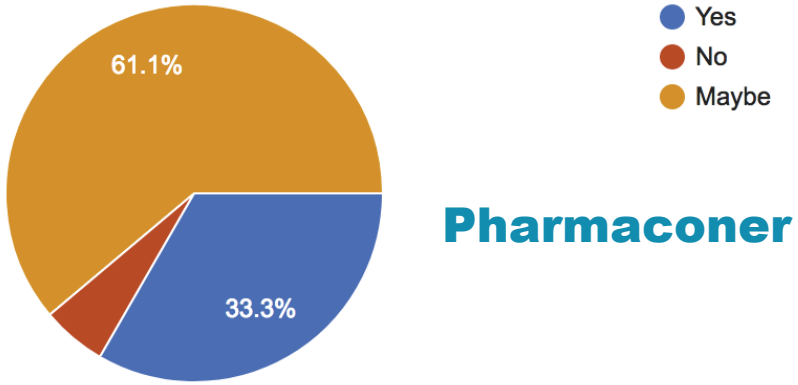
Estudios | Subcon -tratas | **3 Campañas** | Casos de uso | Datos



¿Estarías interesad@ convertir tu sistema en un producto /start up?



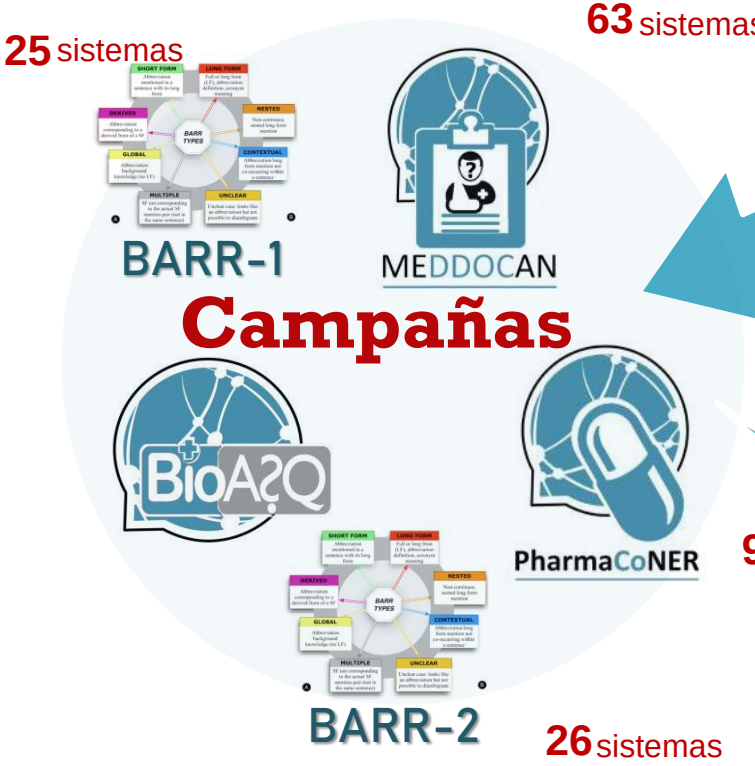
MEDDOCAN



Pharmaconer

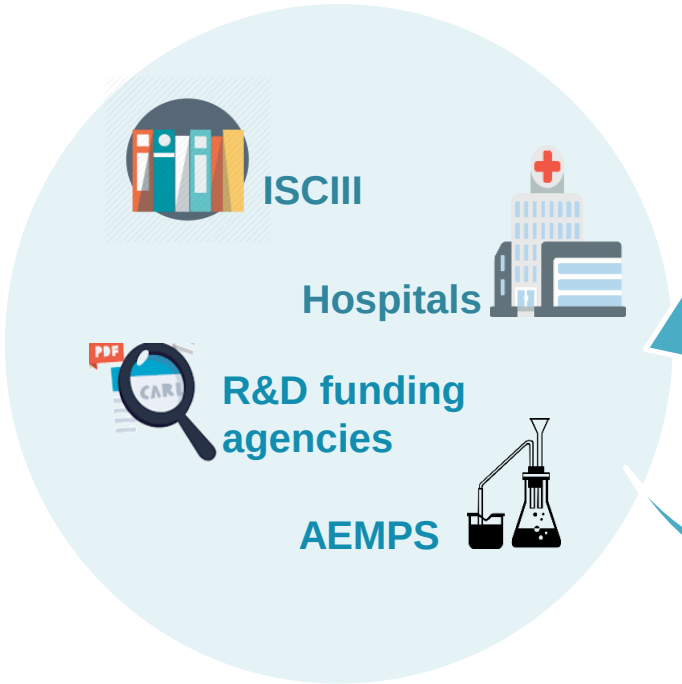
Acciones y líneas de trabajo

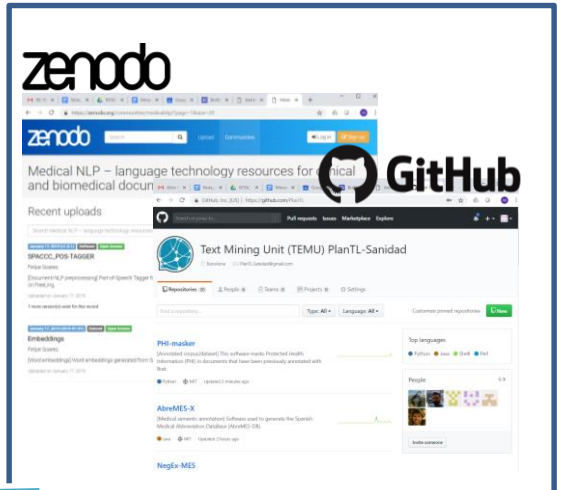
Estudios | Subcon -tratas | **3 Campañas** | Casos de uso | Datos



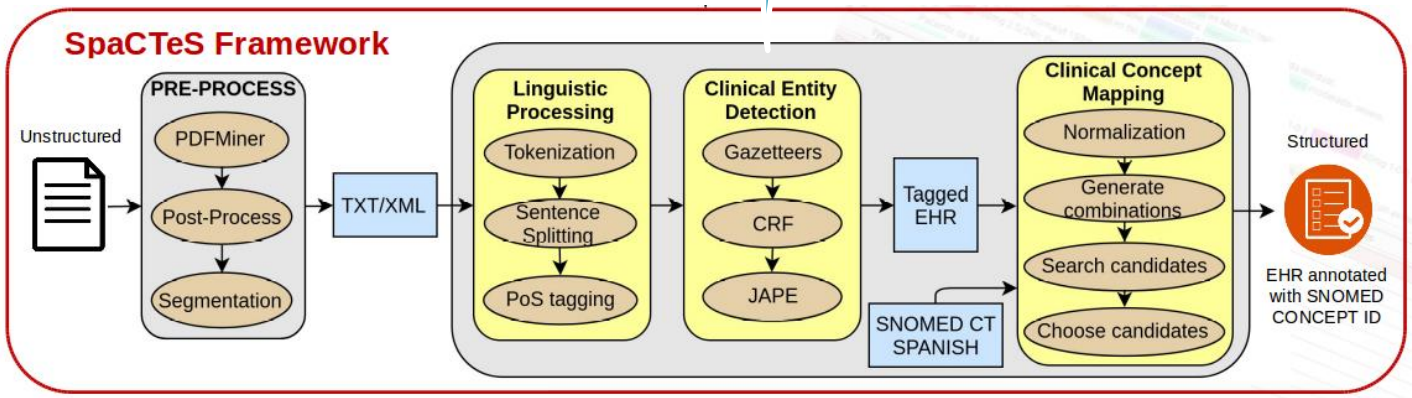
Acciones y líneas de trabajo

Estudios | Subcon-tratas | Campa-ñas | **4 Casos de uso** | Datos





Publicados y accessibles con licencias permisivas



https://zenodo.org/communities/medicalnlp https://github.com/PlanTL-SANIDAD

Acciones y líneas de trabajo

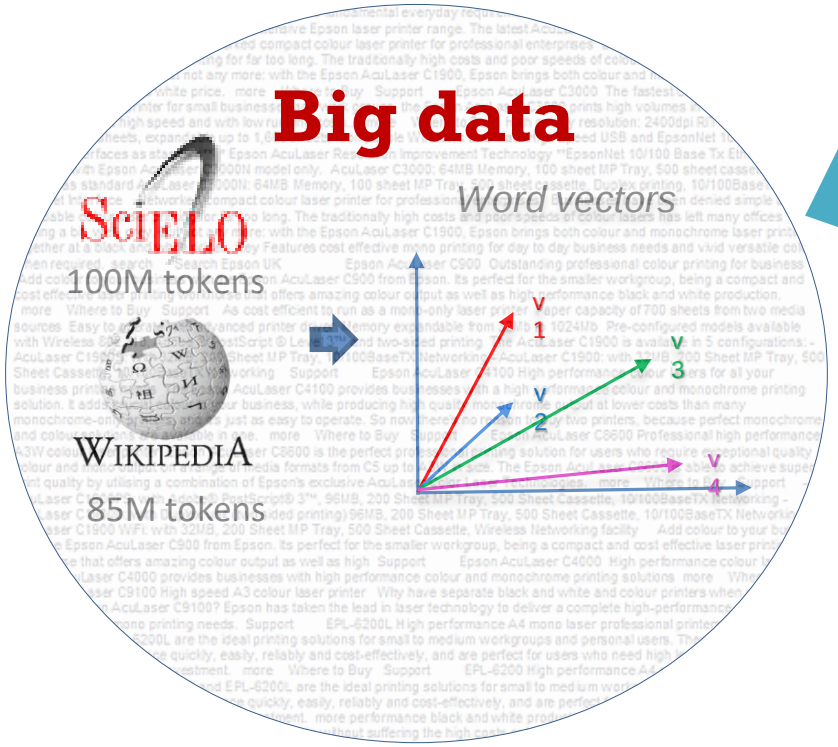
Estudios

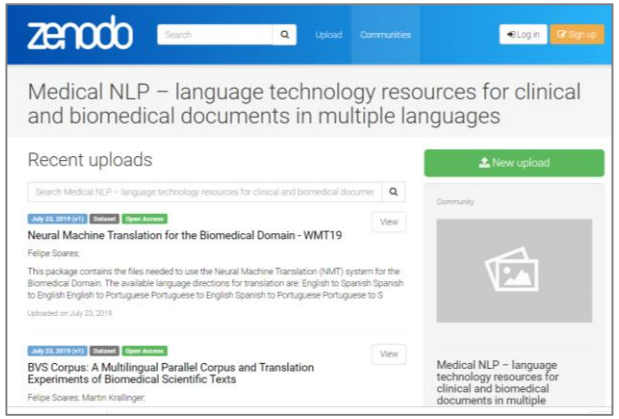
Subcon-tratas

Campa-ñas

Casos de uso

5 Datos





CC-BY

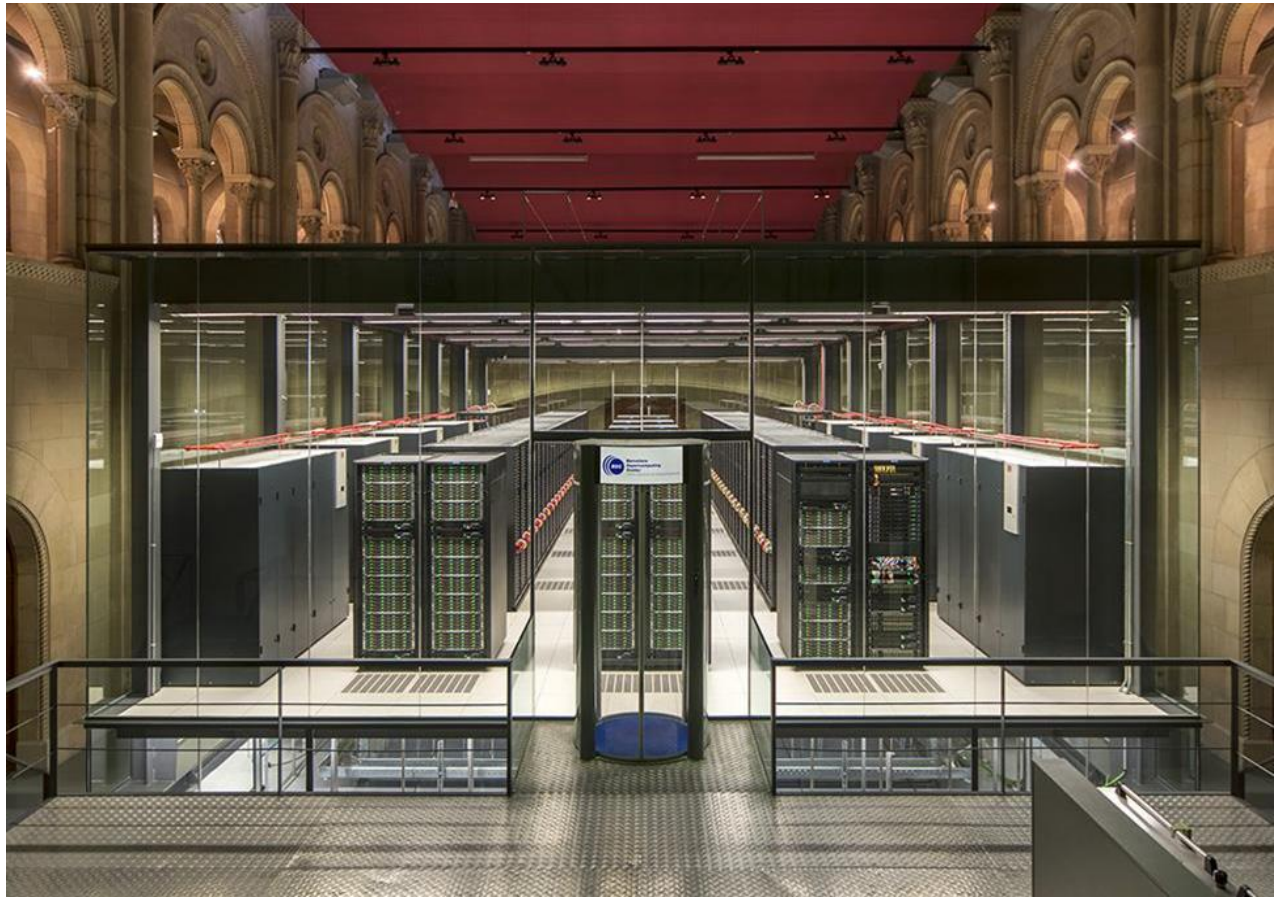
IA, HPC y textos médicos

Las tecnologías basadas en el aprendizaje profundo están revolucionando todo el panorama de IA…

- … especialmente en el caso del procesamiento de lenguaje natural.
- Pero:
  - Necesidad de **muchos datos** (Big Data)
  - Computacionalmente exigente: IA/deep learning, **HPC**.
- Componentes disponibles de manera abierta en Github:
  - github.com/PlanTL-SANIDAD
- Veremos algunos ejemplos.

IA, HPC y textos médicos

Computación: BSC



Datos: ¿Proveedores?

- Los datos, especialmente si están anotados, son el petróleo del siglo XXI.
- En el caso del dominio de textos clínicos, son especialmente difíciles de conseguir.
- Así que… ¡Necesitamos la colaboración de instituciones!

IA, HPC y textos médicos

Word embeddings

- La mayoría de algoritmos de aprendizaje automático trabajan con números/vectores, idealmente con ciertas propiedades.
- Word embeddings: Representación vectorial continua (“numérica”) de cada palabra semánticamente consistente.
- Aplicaciones: Una vez pre-entrenados (con grandes cantidades de texto), se pueden usar para representar las palabras de varios sistemas: clasificadores, NER/anotadores, traducción automática…

100M tokens

85M tokens

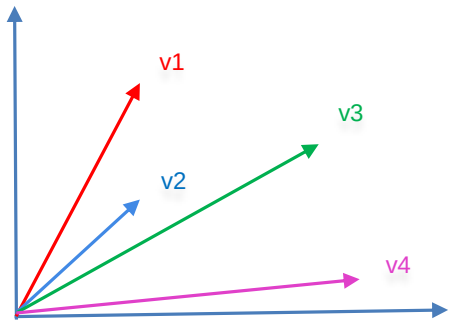
Word2Vec

fastText

|V|

v1	0,78	0,65	0,98
v2	0,23	0,12	0,32
v3	0,90	0,32	0,56
v4	0,08	0,43	0,65
v5	0,77	0,88	0,77

Word vectors



IA, HPC y textos médicos

Embeddings médicos en español

- Word embeddings pre-entrenados con grandes cantidades de texto de dominio: Wikipedia médica y SciELO (4 y 3.3M de frases).
- Algoritmo: FastText.
- Evaluación:
  - Intrínseca: Similitud entre términos sinónimos en SNOMED
  - Extrínseca: Usado en el anotador para fármacos.

	SHE (our)	SBWC
Dataset	ρ	ρ
UMNSRS-sim	**0.5826***	0.4319*
UMNSRS-rel	**0.5239***	0.3947*
MayoSRS	**0.3174***	0.1237

- Listos para ser exportados y usados en otros sistemas.

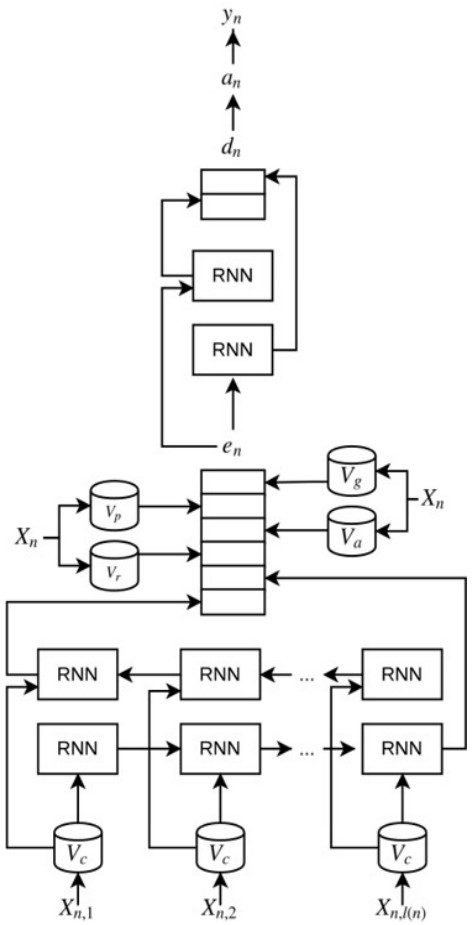
Medical Word Embeddings for Spanish: Development and Evaluation (BSC-TeMU, 2019):
github.com/PlanTL-SANIDAD/Embeddings

IA, HPC y textos médicos

PharmaCoNER Tagger

- Neural Named Entity Recognition (NER) para fármacos:
  - Proteínas
  - Normalizables
  - No-normalizables
- Basado en un sistema neuronal existente, incorpora la posibilidad de añadir información de dominio:
  - Part-Of-Speech
  - Diccionario de afijos
  - Gazetteer

PharmacoNER Tagger: a deep learning-based tool for automatically finding chemicals and drugs in Spanish medical texts (BSC-TeMU,2019)



Arquitectura de la red neuronal de PharmaCoNER Tagger

Componentes PlanTL: tagger

- Embeddings de dominio vs embeddings genéricos

	SHE (our)		SBWC	
	Val	Test	Val	Test
Overall				
Accuracy	**99.51**	**99.62**	99.45	99.57
Precision	**90.63**	90.42	90.30	**90.87**
Recall	**88.25**	**86.03**	86.12	84.45
F1	**89.42**	**88.17**	88.16	87.76
Normalizables				
Precision	**92.82**	93.18	91.87	**93.93**
Recall	**89.81**	88.09	88.89	**88.34**
F1	**91.29**	90.56	90.35	**91.05**
Proteins				
Precision	**87.86**	**86.94**	88.22	86.19
Recall	**87.86**	**84.52**	84.39	81.75
F1	**87.86**	**85.71**	86.26	83.91
Unclear				
Precision	**100**	84.21	92.86	**88.24**
Recall	**81.25**	**84.21**	81.25	78.95
F1	**89.66**	**84.21**	86.67	83.33

- La mejor configuración: Part-Of-Speech + gazetteer

Componentes PlanTL: traducción automática

- Sistemas de traducción automática para el dominio de textos biomédicos.
- Español – Inglés.
- Entrenados solamente con textos de dominio.
- WMT shared task 2018: Sistemas basados en Seq2seq.
- WMT shared task 2019: Sistemas basados en el Transformer.

IA, HPC y textos médicos

Translator

Translate clinical text using an open-source toolkit for neural machine translation (NMT).

ENGLISH | SPANISH | PORTUGUESE

Sample texts

en-sample2.txt

30-year-old male patient who referred to our clinic with the diagnosis of primary hyperparathyroidism. The patient had a history of stage IV-B non-Hodgkin lymphoma diagnosed in 2005 and treated with radiochemotherapy, currently in remission. Our patient presented with polydipsia and polyuria without associated bone pain. The laboratory tests showed a serum calcium of 12.7 mg/dl and preserved renal function. Cervical ultrasound performed preoperatively showed a hyperechogenic nodule of 9 mm adjacent to the left lower thyroid pole. A scintigraphy was performed in which a pathological hypercaptation was observed at the level of the left inferior thyroid.

TRANSLATE

ENGLISH | SPANISH | PORTUGUESE

Paciente masculino de 30 años que acude a nuestra consulta con el diagnóstico de hiperparatiroidismo primario. El paciente tenía antecedente de linfoma no Hodgkin estadio IV-B diagnosticado en 2005 y tratado con radioquimioterapia, actualmente en remisión. Nuestro paciente presentó polidipsia y poliuria sin dolor óseo asociado. Las pruebas de laboratorio mostraron un calcio sérico de 12,7 mg/dl y una función renal preservada. La ecografía cervical realizada preoperatoriamente mostró un nódulo hiperecogénico de 9 mm adyacente al polo inferior izquierdo. Se realizó una gammagrafía en la que se observó hipercaptación patológica a nivel del tiroides inferior izquierdo.

Stats
Total sentences: 6
Average prediction score (always negative): -6.1794
Translation time: 8.9045 seconds

DOWNLOAD JSON RESULTS

Data retrieved successfully. OK

IA, HPC y textos médicos

Nuevos componentes

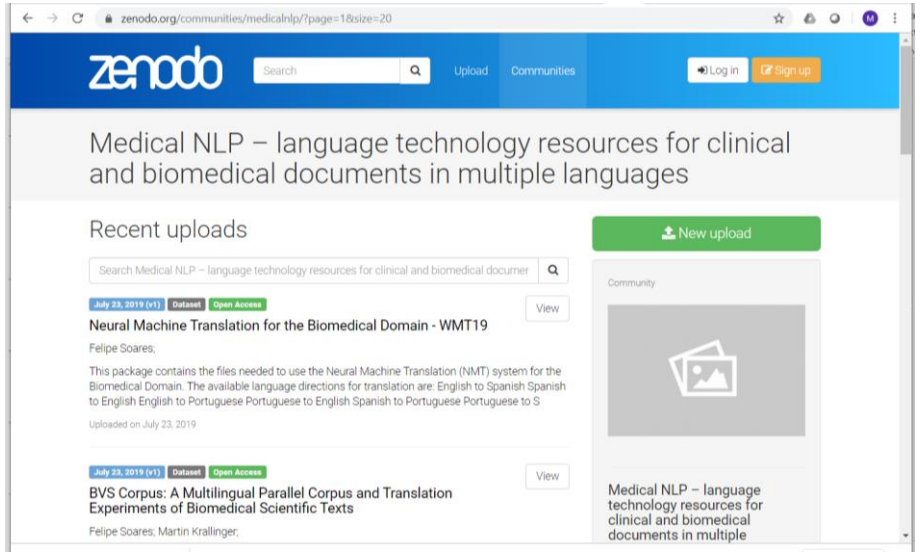
- Adaptación de dominio para traducción automática: En lugar de entrenar sólo con textos de dominio, partir de un sistema genérico y adaptarlo.
- Modelos de lenguaje: En lugar de los clásicos word embeddings, desarrollar un modelo de embeddings contextuales (BERT):
  - Español dominio general.
  - Español dominio biomédico.
  - Lenguas co-oficiales
- Usar dichos modelos de lenguaje para mejorar los sistemas de reconocimiento de conceptos médicos.

En resumen: las dos claves

- **Datos**: colaboración de instituciones
- **Cómputo**: Inteligencia Artificial y HPC (BSC)

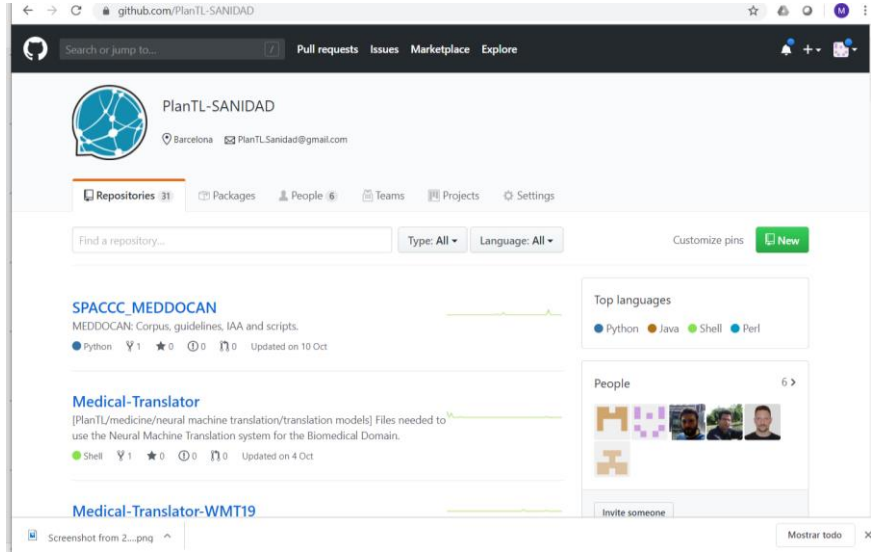
Muchas gracias!

zenodo



https://zenodo.org/communities/medicalnlp

GitHub



https://github.com/PlanTL-SANIDAD