# Introducción al PLN | ProgPLN

Víctor Peinado v.peinado@filol.ucm.es 8-14 de octubre de 2015

## ¿Qué es el PLN?

El Procesamiento del Lenguaje Natural (PLN)<sup>1</sup> es el estudio científico del lenguaje desde un punto de vista computacional.

Es un área claramente multidisciplinar: lingüística, ingeniería, inteligencia artificial, informática, psicología, etc.

El PLN se interesa en proporcionar modelos computacionales para describir, modelar o reproducir distintos fenómenos lingüísticos. Tradicionalmente, estos modelos han tenido dos aproximaciones diferentes:

- 1. sistemas basados en conocimiento: en problemas que podemos modelar, proporcionamos conocimiento lingüístico formalizado.
- 2. sistemas basados en estadística: en problemas que no podemos modelar, proporcionamos ingentes cantidades de datos (colecciones de documentos) y dejamos que la máquina cree el modelo a partir del cálculo de probabilidades y la detección de patrones de uso.

#### Tareas típicas del PLN

Una buena manera de conocer los temas que trata un área de investigación es revisar el programa de los congresos más importantes:<sup>2</sup>

- ACL 2015: *call for papers*<sup>3</sup> y programa<sup>4</sup>
- EMNLP 2015: call for papers<sup>5</sup> y programa<sup>6</sup>
- COLING 2014: lista de artículos<sup>7</sup>
- SEPLN 2015: call for papers<sup>8</sup> y programa<sup>9</sup>

De este modo, podemos identificar algunas de las tareas más comunes del área:

- Desambiguación semántica (*word sense disambiguation*) y reconocimiento de entidades (*named entities recognition*).
- Análisis morfo-sintáctico (PoS tagging/parsing)
- Traducción automática (machine translation): Google Translate
- Extracción de información (information extraction): TripIt

<sup>1</sup> En inglés, *Natural Language Processing* (NLP) o mejor #NLProc.

<sup>&</sup>lt;sup>2</sup> http://www.cs.rochester.edu/~tetreaul/conferences.htm

<sup>&</sup>lt;sup>3</sup> http://acl2015.org/call\_for\_papers.html

<sup>4</sup> http://acl2015.org/program.html

<sup>&</sup>lt;sup>5</sup> http://www.emnlp2015.org/call.html

<sup>&</sup>lt;sup>6</sup> http://www.emnlp2015.org/program.html

<sup>7</sup> http://anthology.aclweb.org/C/C14/

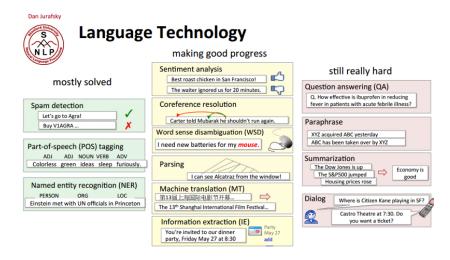
<sup>8</sup> http://gplsi.dlsi.ua.es/sepln15/es/2-

convocatoria-de-comunicaciones

<sup>9</sup> http://gplsi.dlsi.ua.es/sepln15/es/node/52

- Reconocimiento del habla (automatic speech reconition) y síntesis de voz (speech synthesis): Google Voice Search
- Recuperación de información (information retrieval): Google Search, Bing y Wolfram | Alpha
- Resumen automático (automatic summarization) y generación automática de textos: Quakebot y Automated Insights
- Búsqueda de respuestas (question answering): Ask.com, Watson
- Análisis de opiniones (sentiment analysis) NaturalOpinions
- Comprensión del lenguaje natural (natural language understanding): Siri, Ok Google y Cortana

Problemas resueltos y cuestiones abiertas



### ¿Por qué es tan difícil el PLN?

El lenguaje natural es eminentemente ambiguo:. Esta es la principal diferencia entre lenguas naturales y lenguajes artificiales.

Esta ambigüedad existe a varios niveles:

- ambigüedad fonética y fonológica: vaca/baca, casa/caza, has sido tú/has ido tú
- ambigüedad morfológica: casa, beso, río, bajo
- ambigüedad sintáctica: Ayer me encontré a tu padre corriendo
- ambigüedad semántica: banco, pie,
- ambigüedad de discurso: correferencia, resolución de anáforas

#### non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

### neologisms

unfriend Retweet bromance

### segmentation issues

the New York-New Haven Railroad the New York-New Haven Railroad

#### world knowledge

Mary and Sue are sisters. Mary and Sue are mothers.

#### idioms

dark horse get cold feet lose face throw in the towel

#### tricky entity names

Where is A Bug's Life playing Let It Be was recorded ... ... a mutation on the for gene

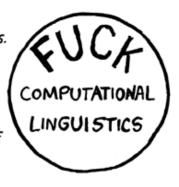
Según la ACL (Association for Computational Linguistics): Computational Linguistics, or Natural Language Processing (NLP), is not a new field. 10, sin embargo no es sencillo definir los límites de la disciplina. Así que podemos considerarla como un conjunto de problemas relacionados con fenómenos lingüísticos y una amalgama de soluciones computacionales, de distinto tipo dependiendo del origen del investi-

Según xkcd, 11 los lingüistas computacionales han vivido muy bien hasta ahora vendiendo motos, así que no se merecen más que nos metamos con ellos :-) 12

AND THE DUMBEST THING ABOUT EMO KIDS ISTHAT ... I ... YOU KNOW, I'M SICK OF EASY TARGETS. ANYONE CAN MAKE FUN OF EMO KIDS. YOU KNOW WHO'S HAD IT TOO EASY? COMPUTATIONAL LINGUISTS.



"OOH, LOOK AT ME! MY FIELD IS SO 14-DEFINED I CAN SUBSCRIBE TO ANY OF DOZENS OF CONTRADICTORY MODELS AND STILL BE TAKEN SERIOUSLY!



10 http://www.aclweb.org/aclwiki/index.php ?title=Frequently\_asked\_questions \_about\_Computational\_Linguistics

11 http://www.xkcd.org/114/

12 http://www.explainxkcd.com/wiki/index.php/ 114:\_Computational\_Linguists