Introducción al PLN | ProgPLN

Víctor Peinado v.peinado@filol.ucm.es

8-9 de octubre de 2015

¿Qué es el PLN?

El Procesamiento del Lenguaje Natural (PLN)[1] 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.

[1] En inglés, *Natural Language Processing* (NLP) o mejor #NLProc.

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:[2]

- ACL 2015: *call for papers*[3] y programa[4]
- EMNLP 2015: *call for papers*[5] y programa[6]
- COLING 2014: *call for papers*[7] y programa[8]
- SEPLN 2015: *call for papers*[9] y programa[10]

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

[2] http://www.cs.rochester.edu/~tetreaul/conferences.htm

[3] http://acl2015.org/call_for_papers.html

[4] http://acl2015.org/program.html

[5] http://www.emnlp2015.org/call.html

[6] http://www.emnlp2015.org/program.html

[7] http://www.coling-2014.org/call-for-papers.php

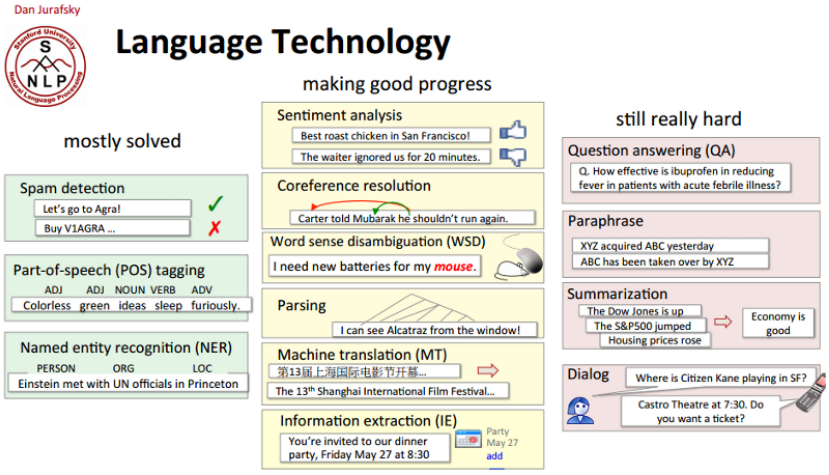
[8] http://www.coling-2014.org/schedule.php

[9] http://gplsi.dlsi.ua.es/sepln15/es/2-convocatoria-de-comunicaciones

[10] 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

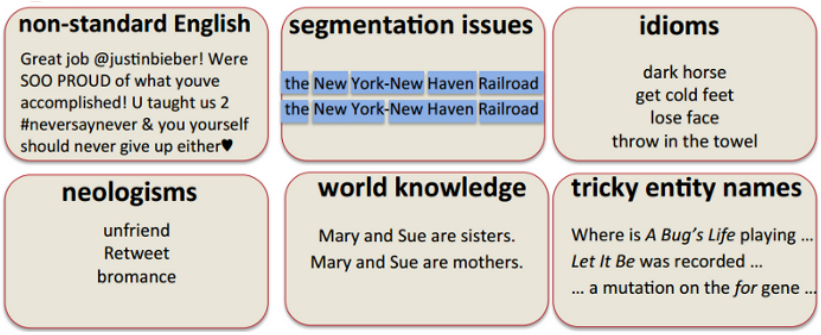


¿Por qué es tan difícil el PLN?

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



Según la ACL (*Association for Computational Linguistics*): *Computational Linguistics, or Natural Language Processing (NLP), is not a new field*.[11], 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 investigador.

[11] http://www.aclweb.org/aclwiki/index.php?title=Frequently_asked_questions_about_Computational_Linguistics

Según xkcd,[12] los lingüistas computacionales han vivido muy bien hasta ahora vendiendo motos, así que no se merecen más que nos metamos con ellos :-) [13]

[12] http://www.xkcd.org/114/

[13] http://www.explainxkcd.com/wiki/index.php/114:_Computational_Linguists