Introducción al NLProc | III Máster en Data Science | KSchool

Víctor Peinado

3-4 de marzo de 2017

Referencias

- *Intro to NLP*.[1]
- *What is Computational Linguistics?*[2]
- *Perspectives in Computational Linguistics*.[3]
- *The Stupidity of Computers*.[4]
- *An Inside Update on NLP*.[5]

¿Qué es el PLN?

El Procesamiento del Lenguaje Natural (PLN)[6] es el estudio científico del lenguaje desde un punto de vista computacional.

Es un área claramente multidisciplinar que aglutina lingüística, ingeniería, inteligencia artificial, informática, estadística, psicología, interacción hombre-máquina, ciencias ognitivas, etc.

Nace como disciplina en los años 1950s con un objetivo inicial claro: construir sistemas de traducción automática.

Objetivos

El PLN se interesa en proporcionar modelos computacionales para describir, modelar o reproducir distintos fenómenos lingüísticos. Tiene como principal objetivo el desarrollo de herramientas y soluciones que permitan:

- procesar automáticamente lenguaje natural.
- comprender el lenguaje natural.
- interaccionar de manera eficaz con ordenadores (o máquinas) de manera natural a través del habla.

Enfoques

Tradicionalmente, el PLN ha trabajado utilizando dos aproximaciones diferentes:

[1] Introduction to NLP `http://futurewavewebdevelopment.com/wp/2016/08/brucemwhealton/introduction-to-natural-language-processing-nlp-`

[2] What is Computational Linguistics `http://www.coli.uni-saarland.de/~hansu/what_is_cl.html`

[3] Perspectives in Computational Linguistics `http://www.linguisticsociety.org/content/computers-and-languages`

[4] The Stupidity of Computers `https://nplusonemag.com/issue-13/essays/stupidity-of-computers/`

[5] An Inside Update on NLP `https://breakthroughanalysis.com/2016/06/23/jbnlp/`

[6] En inglés, *Natural Language Processing* (NLP) o mejor #NLProc. La disciplina recibe otros nombres, como *Human Language Technologies* (HLT), tecnologías de la lengua, ingeniería lingüística, lingüística computacional, etc.

1. sistemas basados en conocimiento: en problemas que podemos modelar, proporcionamos conocimiento lingüístico formalizado y las máquinas actúan aplicando reglas.

2. sistemas basados en estadística: en problemas que son costosos o 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.

Progresos

La disciplina nace a partir de los 1950s y al inicio de la Guerra Fría con el objetivo principal de construir sistemas de **traducción automática**.

En la decada siguiente aparecen los llamados **sistemas expertos** que asistían en la toma de decisiones: sistemas de diálogo que trataban de imitar conversaciones humanas, creación de ontologías para capturar conocimiento del mundo.

Hasta los 1980s, la mayor parte de los sistemas de PLN estaban basados en conocimiento y manejaban complejas reglas diseñadas a mano. Se deja sentir la influencia de la Lingüística Generativa de Noam Chomsky.

A partir de esa década, irrumplen las aproximaciones estadísticas basadas en sistemas de **aprendizaje automático**, que requieren grandes colecciones de datos anotados manualmente. Este desarrollo discurre paralelo al aumento de potencia de los ordenadores.

Actualmente, vivimos un auge de los sistemas de aprendizaje automático *supervisados* (anotados manualemente) y *no supervisados* (sin anotaciones de ningún tipo), con especial énfasis en el uso de la Web. Se aprovecha la explosión de datos disponibles en formato electrónico.

En la década de los 2010s hemos visto el resurgir de los sistemas de aprendizaje automático que utilizan **redes neuronales** para procesar ingentes cantidades de datos (*deep learning*).

Tareas típicas del PLN

Una buena manera de conocer los temas que trata un área de investigación es revisar el calendario de los congresos más importantes:[7]

- ACL 2016: *call for papers*[8] y programa[9].
- EMNLP 2016: *call for papers*[10] y programa[11].
- COLING 2016: *call for papers*[12] y programa[13]

[7] NLP Conferences Calendar `http://cs.rochester.edu/~omidb/nlpcalendar/`

[8] ACL 2016 CFP `http://acl2016.org/index.php?article%20id=9`

[9] ACL 2016 CFP `http://acl2016.org/index.php?article%20id=9`

[10] EMNLP 2016 CFP `http://www.emnlp2016.net/call.html`

[11] EMNLP 2016 Program `http://www.emnlp2016.net`

[12] COLING 2016 CFP `http://coling2016.anlp.jp/cfp/`

[13] COLING 2016 Program `http://coling2016.anlp.jp/`

- SEPLN 2016: *call for papers*[14]y programa[15]

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 y los bundles de Inbox
- 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*): tímidos intentos de Google o Bing y, sbre todo, Watson
- Análisis de opiniones (*sentiment analysis*): Bitext y Atribus
- Comprensión del lenguaje natural (*natural language understanding*): Siri, Google Now y Cortana

[14] SEPLN 2016 CFP `http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=51713&copyownerid=85257`

[15] SEPLN 2016 Program `http://www.congresocedi.es/es/sepln#tabs7`

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*, etc.
- ambigüedad de discurso: correferencia, resolución de anáforas.

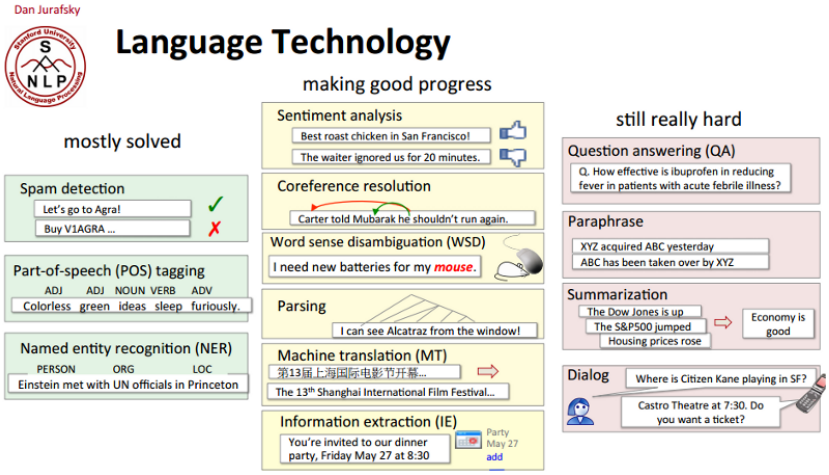

Figure 1: Language Technologies Progress, according to Stanford NLP

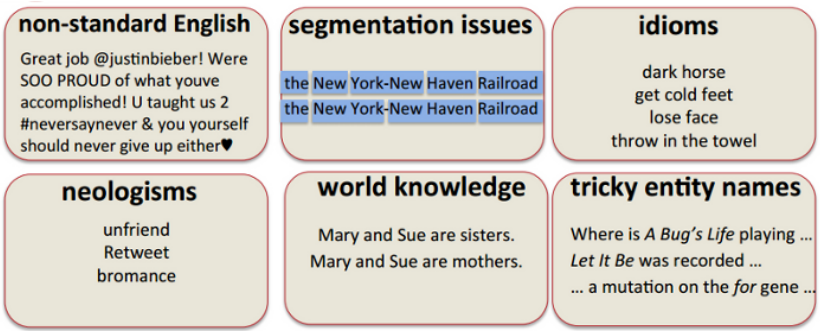

Figure 2: Language Technologies Difficulties, according to Stanford NLP

Según la ACL (*Association for Computational Linguistics*): *Computational Linguistics, or Natural Language Processing (NLP), is not a new field*[16], 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 problemas, técnicas, ideas y soluciones de distinto tipo, dependiendo del origen del investigador.

[16] ACL FAQ `http://www.aclweb.org/aclwiki/index.php?title=Frequently_asked_questions_about_Computational_Linguistics`