

Introducción al PLN | ProgPLN

Víctor Peinado

30 de septiembre de 2016

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, 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 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.

El objetivo último del PLN reside en conseguir que los ordenadores analicen el lenguaje natural, lo comprendan y sean capaces de extraer significado de manera (que parezca) inteligente y resulte útil.

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

¹ Introduction to NLP

<http://futurewavewebdevelopment.com/wp/2016/08/brucewhealton/introduction-to-natural-language-processing-nlp-2016/>

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.

⁷ NLP Conferences Calendar <http://cs.rochester.edu/~omidb/nlpcalendar/>

- ACL 2016: *call for papers*⁸ y programa⁹.
- EMNLP 2016: *call for papers*¹⁰ y programa¹¹.
- COLING 2016: *call for papers*¹² y programa¹³
- SEPLN 2016: *call for papers*¹⁴ y programa¹⁵

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
- Reconocimiento del habla (*automatic speech recognition*) 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*, etc.

⁸ ACL 2016 CFP <http://acl2016.org/index.php?article%20id=9>

⁹ EMNLP 2016 CFP <http://www.emnlp2016.net/cfp.html>

¹⁰ ACL 2016 CFP <http://acl2016.org/index.php?article%20id=9>

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/>

¹⁴ SEPLN 2016 CFP <http://www.wikicfp.com/cfp/servlet/event.showcfp?eventId=51713©ownerid=85257>

¹⁵ SEPLN 2016 Program <http://www.congresocedi.es/es/sepln#tabs7>

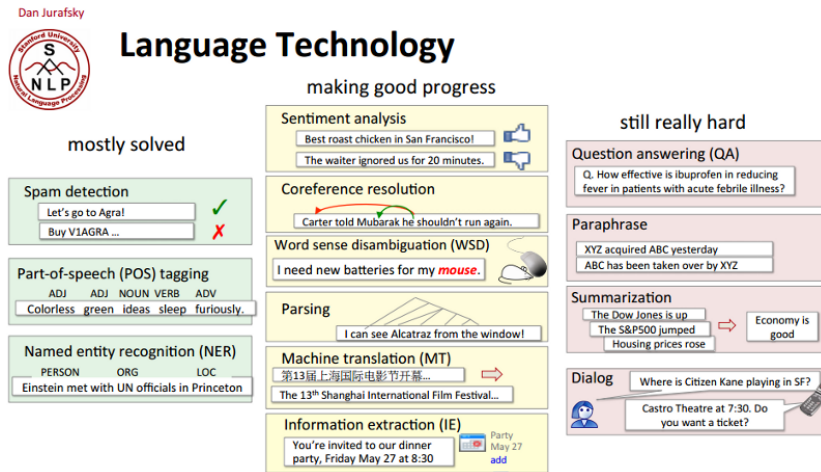


Figure 1: Language Technologies

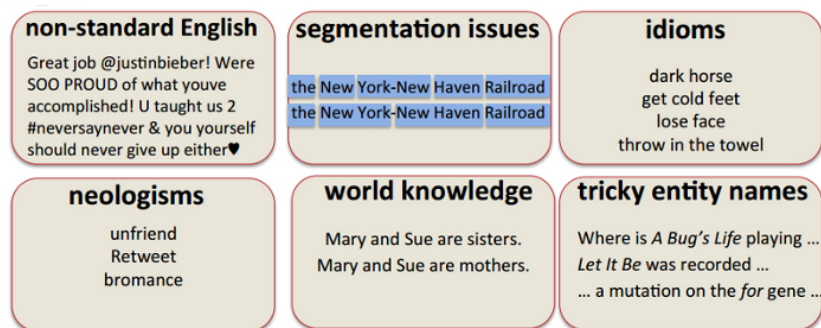


Figure 2: Language Technologies

- 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*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 soluciones computacionales, de distinto tipo dependiendo del origen del investigador.

Según xkcd,¹⁷ los lingüistas computacionales han vivido muy bien hasta ahora vendiendo la moto, así que no se merecen más que nos metamos con ellos :-)¹⁸

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

¹⁷ <http://www.xkcd.org/114/>

¹⁸ http://www.explainxkcd.com/wiki/index.php/114:_Computational_Linguists

Figure 3: Computational Linguistics

