

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.

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 programa⁴
- EMNLP 2015: *call for papers*⁵ y programa⁶
- COLING 2014: *call for papers*⁷ y programa⁸
- SEPLN 2015: *call for papers*⁹ y programa¹⁰

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

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

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 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

Dan Jurafsky



Language Technology

making good progress

mostly solved

- Spam detection
 - Let's go to Agral ✓
 - Buy VIAGRA ... ✗
- Part-of-speech (POS) tagging
 - ADJ ADJ NOUN VERB ADV
 - Colorless green ideas sleep furiously...
- Named entity recognition (NER)
 - PERSON ORG LOC
 - Einstein met with UN officials in Princeton

still really hard

- Sentiment analysis
 - Best roast chicken in San Francisco! 👍
 - The waiter ignored us for 20 minutes. 🗨️
- Coreference resolution
 - Carter told Mubarak he shouldn't run again.
- Word sense disambiguation (WSD)
 - I need new batteries for my *mouse*.
- Parsing
 - I can see Alcatraz from the window!
- Machine translation (MT)
 - 第13届上海国际电影节开幕...
 - The 13th Shanghai International Film Festival...
- Information extraction (IE)
 - You're invited to our dinner party, Friday May 27 at 8:30
 - Party May 27 add
- Question answering (QA)
 - Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?
- Paraphrase
 - XYZ acquired ABC yesterday
 - ABC has been taken over by XYZ
- Summarization
 - The Dow Jones is up
 - The S&P500 jumped
 - Housing prices rose
 - Economy is good
- Dialog
 - Where is Citizen Kane playing in SF?
 - Castro Theatre at 7:30. Do you want a ticket?

¿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♥	segmentation issues the New York-New Haven Railroad the New York-New Haven Railroad	idioms dark horse get cold feet lose face throw in the towel
neologisms unfriend Retweet bromance	world knowledge Mary and Sue are sisters. Mary and Sue are mothers.	tricky entity names Where is <i>A Bug's Life</i> playing ... <i>Let It Be</i> was recorded a mutation on the <i>for</i> gene ...

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.

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

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

¹² <http://www.xkcd.org/114/>

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

