

Introducción al Procesamiento del Lenguaje Natural | ProgPLN

Víctor Peinado v.peinado@filol.ucm.es

3 de octubre de 2014

¿Qué es el PLN?

El Procesamiento del Lenguaje Natural (PLN) 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 de distintos fenómenos lingüísticos. Estos modelos pueden tener dos aproximaciones diferentes:

1. sistemas basados en conocimiento: en problemas que podemos modelar, proporcionamos conocimiento lingüístico formalizados
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 tratan un área de investigación es revisar el programa de los congresos más importantes del área:

- ACL 2014: *call for papers*¹ y programa²
- COLING 2014: *call for papers*³ y programa⁴
- SEPLN 2014: *call for papers*⁵ y programa⁶

De este modo, podemos identificar algunas de las tareas:

- Desambiguación semántica (*word sense disambiguation*)
- 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

¹ <http://www.acl2012.org/program/subo3.asp>

² <http://acl2014.org/Program.htm>

³ <http://www.coling-2014.org/call-for-papers.php>

⁴ <http://www.coling-2014.org/schedule.php>


⁵ <http://www.taln.upf.edu/pages/sepln2014/es/call-for-papers.html>

⁶ <http://taln.upf.edu/pages/sepln2014/es/program.html>

- Recuperación de información (*information retrieval*): Google Search, Bing y Wolfram | Alpha
- Resumen automático (*automatic summarization*)
- Búsqueda de respuestas (*question answering*): Ask.com, Watson
- Análisis de opiniones (*sentiment analysis*) NaturalOpinions
- Comprensión del lenguaje natural (*natural language understanding*): Siri y Ok Google

Problemas resueltos y cuestiones abiertas

Dan Jurafsky



Language Technology

making good progress

mostly solved

- Spam detection
 - Let's go to Agra! ✓
 - 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

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

still really hard

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**: 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, rio*
- ambigüedad sintáctica: *Ayer me encontré a tu padre corriendo*
- ambigüedad semántica: *banco, pie,*
- ambigüedad de discurso: *correferencia, resolución de anáforas*

<p>non-standard English</p> <p>Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥</p>	<p>segmentation issues</p> <p>the New York-New Haven Railroad the New York-New Haven Railroad</p>	<p>idioms</p> <p>dark horse get cold feet lose face throw in the towel</p>
<p>neologisms</p> <p>unfriend Retweet bromance</p>	<p>world knowledge</p> <p>Mary and Sue are sisters. Mary and Sue are mothers.</p>	<p>tricky entity names</p> <p>Where is <i>A Bug's Life</i> playing ... <i>Let It Be</i> was recorded a mutation on the <i>for</i> gene ...</p>

Según la ACL (*Association for Computational Linguistics*): *Computational Linguistics, or Natural Language Processing (NLP), is not a new field.*⁷, 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.

⁷ http://www.aclweb.org/aclwiki/index.php?title=Frequently_asked_questions_about_Computational_Linguistics

Según xkcd,⁸ los lingüistas computacionales han vivido muy bien hasta ahora vendiendo motos, así que metámonos con ellos.⁹

⁸ <http://www.xkcd.org/114/>

⁹ http://www.explainxkcd.com/wiki/index.php/114:_Computational_Linguists

