

Introducción al NLPProc | Máster en Data Science | KSchool Madrid

Víctor Peinado

enero de 2021

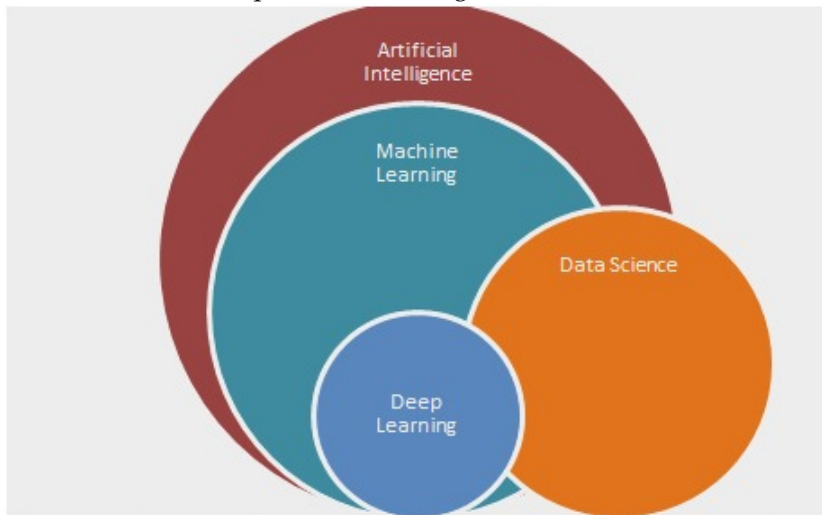
Referencias

- *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 cognitivas, 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.

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 Computa-

tional 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 #NLPProc. La disciplina recibe otros nombres, como *Human Language Technologies* (HLT), tecnologías de la lengua, ingeniería lingüística, lingüística computacional, etc.

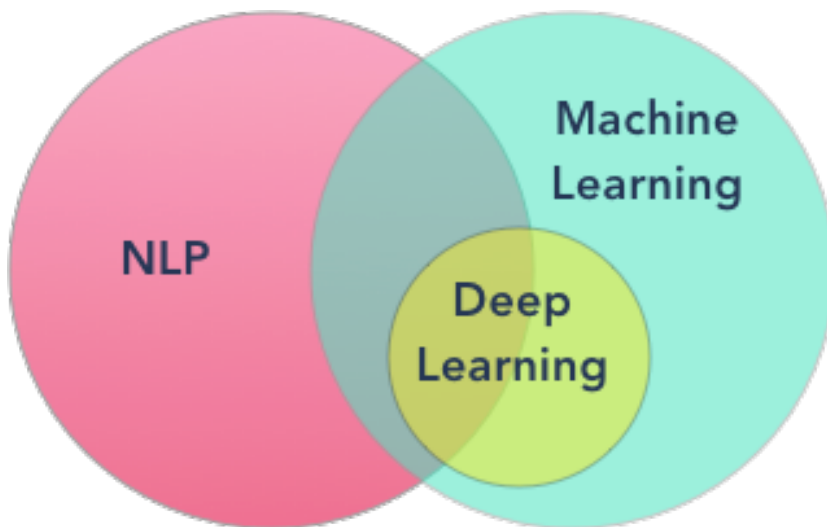
Tiene como principal objetivo el desarrollo de herramientas y soluciones que permitan:

- 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. 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 década 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, irrumpen 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 manualmente) 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*). La gran pega reside en que es necesario contar con muchos datos y mucha potencia de cálculo (centenares de GPUs)... Hasta que llegaron los modelos pre-entrenados y la aplicación de *Transfer Learning* al mundo del procesamiento del lenguaje. Y es que... somos afortunados, porque justo ahora estamos viviendo el momento ImageNet del NLP.⁷

⁷ NLP's ImageNet moment has arrived
<http://ruder.io/nlp-imagenet/>

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:⁸

- ACL: <https://acl2020.org>
- EMNLP: <https://2020.emnlp.org/>
- COLING: <https://coling2020.org/>
- SEPLN: <http://sepln2020.sepln.org/>

⁸ NLP Conferences Calendar
<http://cs.rochester.edu/~omidb/nlpcalendar/>

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*): Natural Language API y Meaning Cloud
- 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 recognition*) y síntesis de voz (*speech synthesis*): Google Speech API
- 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, sobre todo, Watson
- Análisis de opiniones (*sentiment analysis*) y gestión de la reputación online (*online reputation management*): Bitext y Atribus.
- Comprensión del lenguaje natural (*natural language understanding*): Siri, Google Assistant, Cortana y Aura.

Problemas resueltos y cuestiones abiertas

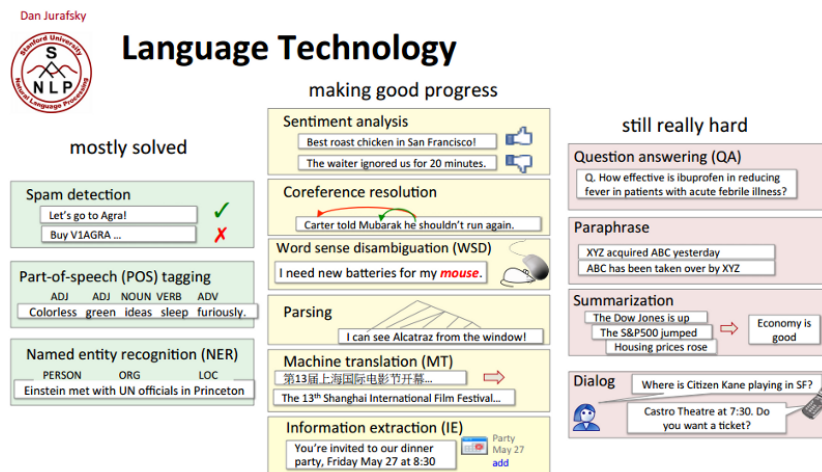


Figure 1: Language Technologies Progress, according to Stanford NLP

¿Por qué es tan difícil el PLN? ⁹

El lenguaje natural es eminentemente **ambiguo**. Esta es la principal diferencia entre lenguas naturales y lenguajes artificiales.

Esta ambigüedad existe a varios niveles:

⁹ The 4 Biggest Open Problems in NLP <http://ruder.io/4-biggest-open-problems-in-nlp/>

- 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.

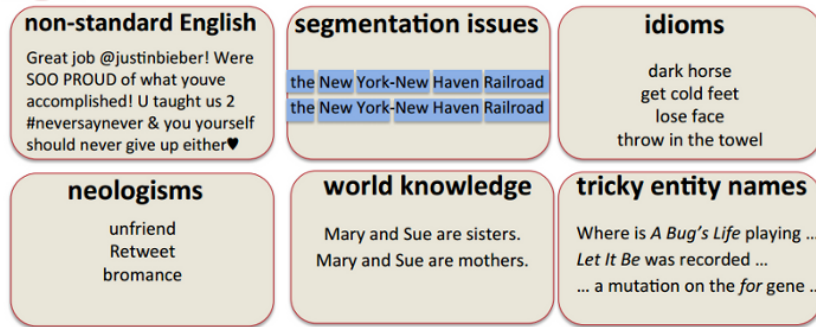


Figure 2: Language Technologies Difficulties, according to Stanford NLP

Según la ACL (*Association for Computational Linguistics*): *Computational Linguistics, or Natural Language Processing (NLP), is not a new field*¹⁰, 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.

¹⁰ ACL FAQ about Computational Linguistics https://aclweb.org/aclwiki/Frequently_asked_questions_about_Computational_Linguistics