Lingüística Computacional

Víctor Mijangos de la Cruz

I. Introducción a la Lingüística Computacional

Objetivo del curso

El curso de Procesamiento de Lenguaje Natural busca que lxs alumnxs **adquieran los conocimientos teóricos básicos** para desarrollar herramientas orientadas al lenguaje natural. Por tanto, el curso se enfocará en los siguientes puntos:

- Conocer los conceptos básicos sobre el estudio del lenguaje natural (**herramientas lingüísticas**).
- Manejar las herramientas **matemáticas** necesarias para la resolución de problemas en el procesamiento del lenguaje natural.
- Desarrollar en el alumno la capacidad de **comprender** y **resolver** problemas que tengan que ver con el lenguaje natural.
- Conocer las **áreas más prominentes** de la lingüística computacional y el procesamiento del lenguaje natural.

Prerrequisitos del curso

Para entender de manera amplia los temas que se expondrán en el curso, se considera un conocimientos básicos en los temas:

- Álgebra básica
- Autómatas y lenguajes formales
- Probabilidad y estadística
- Álgebra lineal
- Cálculo multivariable

Temario

1. Introducción
   1. Introducción
   2. Marco histórico
   3. Herramientas lingüísticas
2. Perspectivas de lenguajes formales
   1. Lenguajes formales
   2. Lenguajes regulares
   3. Lenguajes libres de contexto
3. Perspectivas estadísticas
   1. Ley de Zipf y leyes empíricas del lenguaje
   2. Modelo del canal ruidoso y lenguaje natural
   3. Técnicas de tokenización
4. Modelos del lenguaje
   1. Modelos del lenguaje de n-gramas
   2. Modelos del lenguaje neuronales
   3. Generación de lenguaje
5. Word embeddings
   1. Embeddings estáticos (Word2Vec, FastText, GloVe)
   2. Embeddings contextualizados (ELMo, CoVe)
   3. Composicionalidad y similitud
6. Modelos atencionales
   1. Modelos sequence-to-sequence
   2. Transformers

Bibliografía básica

- Jurafsky, D. & JH. Martin (2018). Speech and Language Processing (3rd Edition). Pearson.
- Goldber, Y. (2015) A Primer on Neural Netwrok Models for Natural Language Processing. https://u.cs.biu.ac.il/~yogo/nnlp.pdf
- Manning, C. & H. Schutze (1999). Foundations on Statistical Natural Language Processing. MIT Press.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press. https://www.deeplearningbook.org/
- Clark, A., S. Fox & S. Lappin (2003). The Handbook of Computational Linguistics and Natural Language Processing. John Wiley.
- Kornai, A. (2008). Mathematical Linguistics. Springer.

Evaluación

La calificación mínima es de **8**, sujeta a la entrega de las tareas. **Cada tarea faltante contará como un punto menos**.

- Se realizarán 3 tareas prácticas, orientadas a un **proyecto**, que se subirán en respectivas carpetas a través de Google Collaboratory o bien por archivos *.ipynb. Además, se contará con 2 o 3 cuestionarios a lo largo del curso.
- Se dejarán tareas teóricas como ejercicios extras. Estos tendrán un valor sobre la calificación final dependiendo del promedio obtenido en éstas.

Concepto	Porcentaje
Tareas	+1.5 pts
Asistencia y participación	+0.5 pts
Problemas extras	+1 pt

Introducción a la Lingüística Computacional

Inteligencia artificial y Lingüística Computacional

- La Lingüística Computacional (LC) se puede considerar una rama de la **Inteligencia Artificial**.
- La inteligencia artificial se encarga de los **sistemas (artificiales) inteligentes**, es decir, que emulen la inteligencia humana.
- El **lenguaje humano** es parte de los procesos cognitivos que permiten hablar de una inteligencia humana.
- La LC busca integrar el lenguaje humano para que sea procesado, comprendido y emulado por agentes artificiales (computadoras).

PLN y Lingüística computacional

Se pueden distinguir dos formas de aproximar el lenguaje y la computación:

Lingüística computacional: Desarrollo de métodos computacionales (principalmente teóricos) para entender los problemas del lenguaje.

Procesamiento del Lenguaje Natural: Desarrollo de ingeniería para solucionar los problemas (prácticos) referentes al lenguaje.

Muchas veces, se utilizan ambos términos indistintamente.

El lenguaje

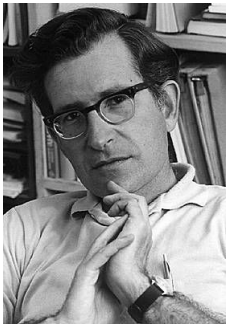
El **lenguaje** es el objeto que el PLN busca procesar de manera artificial. Para esto se apoya de la **lingüística**, que puede entenderse como el estudio del lenguaje en todas sus manifestaciones. Se pueden dar dos enfoques a la lingüística y al PLN:

Enfoque racionalista: Asume que el lenguaje es un proceso que debe ser estudiado desde los procesos cognitivos (Chomsky, 1957). Bajo este enfoque se hace uso de la teoría de los lenguajes formales para implementar un sistema de reglas que pueda interpretar un lenguaje específico.

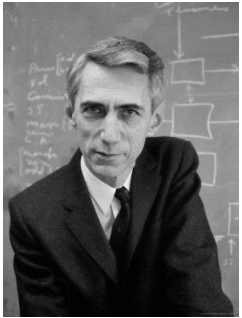
Enfoque empírico: Se centra en los datos observables de la lengua y suele tomar métodos estadísticos (Shannon, 1948). Bajo este enfoque se aplican métodos estadísticos y de aprendizaje automático para inferir generalizaciones

Matemáticas y Lenguaje

Perspectivas formales: Basadas en el enfoque racionalista. Hacen uso de la teoría de lenguajes formales, buscando formalizar al lenguaje en conceptos algebraico.



Perspectivas estadísticas: Basadas en el enfoque empirista. Hacen uso de la teoría de la información, técnicas estadísticas y el aprendizaje automático.



Los inicios de la Lingüística Computacional

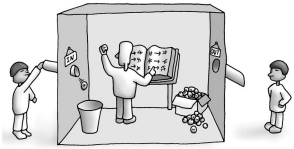
- En 1949, empieza a circular la idea de una **máquina de traducir**, gracias al matemático Warren Weaver.
- Esta máquina pensaba estar dividida en dos módulos:
  1. Vocabulario: Palabras y sus traducciones.
  2. Gramática: Reglas que las lenguas a traducir siguen para generar lenguaje.
- La traducción automática quedo relegada debido a los malos resultados que se obtuvieron:

"'Lo que significa esta idea, en caso de que se la tome seriamente, es que una máquina de traducir no se le debe proporcionar sólo un diccionario, sino una enciclopedia universal. Y, puesto que esto no es más que una quimera, no tiene sentido discutirlo más"' (Bar-Hillel, The current status of automatic translation of languages, 1960).

La habitación china

Un juego mental que resume los problemas de la traducción automática es la **habitación china**:

- Se encierra a una persona (que no habla chino) en un cuarto.
- El cuarto cuenta con diccionarios español-chino y libros de gramática.
- Por debajo de la puerta se le pasan notas que contienen preguntas en chino.
- La persona debe contestar a las preguntas con las herramientas disponibles.



Las preguntas centrales de este juego mental son:

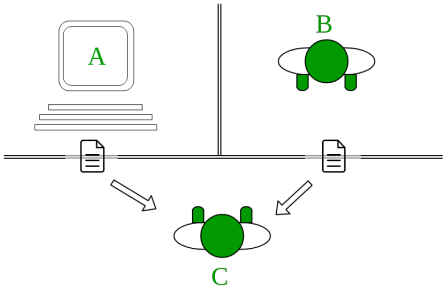
- ¿Se puede decir que la persona sabe o ha entendido el chino?
- ¿Existe una comprensión real de la lengua, o sólo es un proceso mecánico?

La prueba de Turing

En 1950, Alan Turing plantea un experimento mental: el juego de la imitación o **prueba de Turing**:

- Una máquina y un ser humano responden preguntas planteadas en lenguaje natural.
- Un juez (que realiza las preguntas) tiene que determinar quién es el humano.

¿Qué se requiere para que una máquina sea capaz de confundir al juez?



Lingüística Computacional en la actualidad

Actualmente el desarrollo de herramientas de PLN ha mostrado un crecimiento importante. La Lingüística Computacional se ha vuelto un área de espacial relevancia.

- Los métodos de traducción automática han mejorado notablemente.
- Sistemas de reconocimiento de voz están integrados en dispositivos móviles.
- Sistemas de etiquetado y clasificación han mostrado desempeños altos.
- Se han creado modelos de generación del lenguaje de grandes capacidades.

Lingüística Computacional en la actualidad

Lingüística y herramientas lingüística

¿Qué es la lingüística?

La **Lingüística Computacional** se interesa por procesar el lenguaje humano por medio de métodos computacionales.

Lenguaje

El lenguaje es un hecho social que permite comunicar ideas, pensamientos, hechos del mundo, etc.

Por tanto, se vuelve necesario el estudio del lenguaje. De esto se encarga la lingüística.

Lingüística

La lingüística es el estudio de todas las manifestaciones del lenguaje humano.

Tareas de la lingüística

En su libro *Curso de Lingüística General* (De Saussure, 1916), el lingüista Ferdinand De Saussure asigna las siguientes tareas a la lingüística:

- Descripción de las lenguas naturales.
- Deducir leyes generales del lenguaje natural.
- Delimitarse y definirse ella misma.

Estudio de la lingüística según los hablantes

El lingüísta Noam Chosmky (1965) ha propuesto una distinción del lenguaje en tanto sistema general (abstracto) o como sistema empírico llevado a cabo por los habalantes:

Competencia: Es el sistema lingüístico en abstracto, similar al concepto de *lengua*.

Actuación: Es la producción lingüísticas de hablantes particulares, también conocida como *habla*.

En PLN se busca pasar de la actuación (datos empíricos) a la competencia (modelo abstracto).

Estudio de la lengua en el tiempo

De Saussure (1916) distingue dos formas de estudiar los fenómenos del lenguaje según la temporalidad del análisis:

Diacronía: Estudio del lenguaje a través de diferentes períodos históricos.

Sincronía: El estudio del lenguaje sin tomar en cuenta cambios o influencias históricas.

Niveles de análisis del lenguaje

Para estudiar el lenguaje es común dividirlo en **niveles** o módulos que representan procesos particulares. Estos niveles pueden ser:

Fonética y fonología. Estudia los sonidos de los lenguaje y los cambios que pueden sufrir.

Morfología. Estudia la estructura interna de palabras.

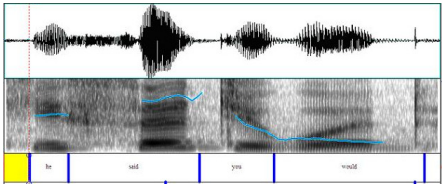
Morfosintaxis. Estudio de las funciones de las palabras.

Sintaxis. Estudio de las frases oraciones que se producen en una lengua.

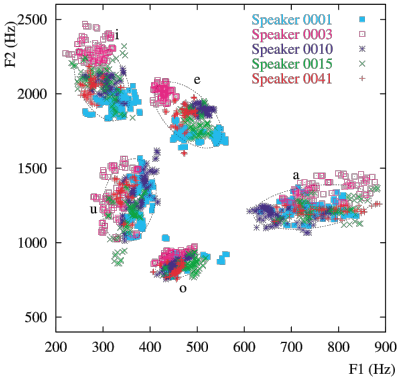
Semántica. Estudia el significado en todos sus niveles: en el léxico, en las oraciones o en los discursos.

Fonética y fonología

La **fonética** estudia las producciones acústicas y su procesamiento, así como características de la señal acústica.



La **fonología** busca representaciones abstractas de los sonidos.



Morfología, morfosintaxis

La **morfología** le interesa ver cómo se pueden formar nuevas palabras y que relaciones existen con respecto a la estructura interna de las palabras.

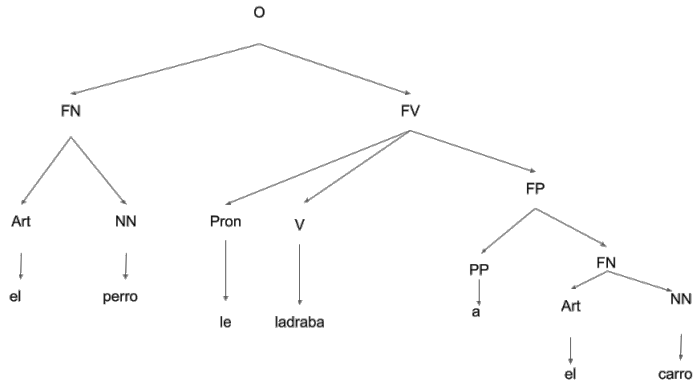
Palabra	Estructura
gato	gat-o
gatos	gat-o-s
perros	perr-o-s

La **morfosintaxis** le interesa la función que una palabra puede desempeñar dentro de una oración.

Palabra	Función	Categoría
comer	Denota una acción	Verbo
gato	Puede servir como sujeto de una acción o como objeto de ésta	Sustantivo
rápidamente	Modifica a un verbo	Adverbio

Sintaxis

A la **sintaxis** le interesa como se combinan las palabras para conformar frases y oraciones, así como sus estructuras.



Semántica

El estudio del signifcado o **semántica** es quizá uno de los más complicados para el PLN. No hay una noción precisa de cómo representar el significado. Algunas propuestas son:

- Desde la lógica formal, como operaciones: "John ve a Maria" $\mapsto$ *see*(*John*, *Mara*)
- A partir de relaciones (ontologías, taxonomías).
- Por medio de representar las palabras en espacios vectoriales (word embeddings).

Palabras

El concepto de palabra es esencial, una definición práctica dentro de PLN es la siguiente:

Palabra

Una palabra es una cadena de caracteres entre dos espacios en blancos.

Surgen diversos **problemas** de esta definición:

- No todas las lenguas usan espacios en su escritura (chino, japonés).
- Sólo se aplica a lenguaje escrito. ¿Qué pasa con grabaciones acústicas?
- Casos conflictivos como "'dormirse"' contra "'se duerme"'.

Palabras

Otra definición de palabra puede darse a partir de los elementos que buscamos en éstas:

Palabra

Una palabra es una unidad lingüística con estructura, significado y función.

Forma	Significado	Función
gato	"Mamífero carnívoro de la familia de los félidos, digitígrado, doméstico, de unos 50 cm de largo desde la cabeza hasta el arranque de la cola, que por sí sola mide unos 20 cm, de cabeza redonda, lengua muy áspera, patas cortas y pelaje espeso, suave, de color blanco, gris, pardo, rojizo o negro, que se empleaba en algunos lugares para cazar ratones"	Sustantivo

Lemas y lexicón

Una forma común de representar un lexema es a partir de su **lema**.

Lema

Un lema es la forma de diccionario de una palabra.

Determinar un lema se vuelve complicado cuando trabajamos con lenguas que no necesariamente tienen un diccionario.

Lexicón

Diccionario mental del hablante que contiene las palabras de una lengua. También podemos considerarlo como un diccionario (electrónico) que se utiliza para diferentes tareas de procesamiento del lenguaje natural.

Palabras

Un ejemplo de estos conceptos aplicado a ciertas palabras se presenta a continuación:

Palabra	**Lema**
gato	GATO
gatitos	GATO
amamos	AMAR
comen	COMER

El lexicón se conforomaría por el conjunto $\{$ GATO, AMAR, COMER $\}$.

Tipos y tokens

En NLP, la computadora no es capaz de reconocer cuando dos formas de palabra distintas pertenecen a un mismo lexema.
Por tanto, en NLP se suele hablar de **tipos** y **tokens**:

Token

Un token es la ocurrencia individual de una palabra dentro de un documento.

Tipo

Los tipos son los diferentes elementos lingüísticos que existen en un documento.

El **tamaño de corpus** se mide en número de tokens.

Tipos y tokens

Considérese el siguiente texto:

> *El perro negro le quitó el hueso al perro amarillo.*

En este texto encontramos tipos y tokens distribuidos de la siguiente forma:

- Número de tokens $= 10$ (palabras en el texto, sin importar repetición).
- Número de tipos $= 8$ (palabras únicas en el texto).

Stopwords

En los textos suelen encontrarse palabras que aportan poca información y que muchas veces son irrelevantes para las aplicaciones de PLN. A estas palabras se les conoce como **stopwords**. Las stopwords suelen ser preposiciones, conjunciones o artículos. Estas se eliminan a partir de una lista de paro.

- Texto con stopwords:

  Homo sapiens es una especie del orden de los primates perteneciente a la familia de los homínidos. También son conocidos bajo la denominación genérica de "'humanos"'.

- Texto sin stopwords:

  Homo sapiens es especie orden primates perteneciente familia homínidos. También son conocidos denominación genérica "'humanos"'.

Corpus

La LC busca tomar datos empíricos para crear un modelo computacional que represente la lengua. Estos datos empíricos se presentan dentro de lo que llamamos corpus:

Corpus

Un corpus es una recopilación bien organizada de muestras del lenguaje a partir de materiales escritos o hablados, agrupados bajo criterios mínimos.

Algunas distinciones que se hacen entre tipos de corpus son:

- Corpus textuales y corpus orales
- Corpus anotado y no anotado
- Corpus monolingüie y multilingüe

Lingüística computacional

Lingüística computacional

Lingüística computacional

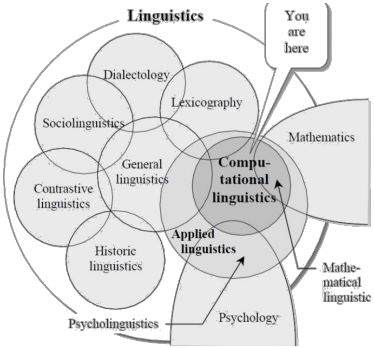
La lingüística computacional se enfoca en entender y emular los procesos humanos que conlleva la producción lingüística. Busca establecer modelos computables que sean capaces de procesar el lenguaje natural.

La lingüística computacional busca tener las capacidades de:

- Manejar grandes cantidades de información en lenguaje natural.
- Generar conocimiento a partir de datos en lenguaje natural.

Multidiscplina

La lingüística computacional es un área multidisciplinaria, conlleva el conocimiento de otras áreas.



Muy ligado a la lingüística computacional están la **lingüística matemática** y la**lingüística cuantitativa**.

Problemas de lingüística computacional

A. Gelbukh y G. Sidorov (2006) proponen una división de los **problemas** del PLN en dos grandes paradigmas:

Problemas conceptuales: Tratan de problemas más teóricos que involucran la comprensión del lenguaje para generar procesos o reglas formales que sean computables.

Problemas técnicos: Cuestiones más prácticas que involucran la representación del lenguaje dentro de una computadora, su codificación y decodificación, su estructura y la identificación de ésta.

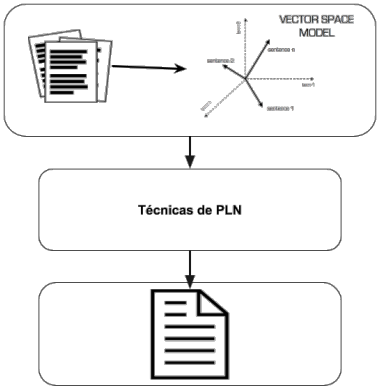
Estructura de una aplicación de LC

En una aplicación de LC, generalmente se asumen 3 módulos:

Codificación: Llevar los datos a una representación comprensible por una computadora.

Procesamiento: A partir de la representación, hacer los procesos necesarios para *entender* lo que está plasmado en lenguaje natural.

Decodificación: Una vez procesados los datos, transformar éstos a una representación que pueda ser entendida por un ser humano (usuario).



Textos recomendados

Chomsky, N. (1957). *Estructuras sintácticas*. Siglo XXI.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press

De Saussure, F. (1916). *Curso de Lingüística General*. Fontamara.

De Saussure, F. (1916). *Curso de Lingüística General*. Fontamara.

Gelbukh, A. y Sidorov, G. (2006). *Procesamiento automático del español con enfoque en recursos léxicos grandes*. IPN.

Shannon, C. (1948). "A mathematical theory of communication". *Bell Systems Technical Journal*, 27(3), 379-423.