Procesamiento de Lenguaje Natural (NLP)

Dr. Gaddiel Desirena López

Maestría en Inteligencia Artificial

Descripción de la asignatura

Objetivo General. Introducir al análisis y transformación del lenguaje mediante técnicas computacionales. Representar al texto y discurso que pueden ayudar a la predicción, la extracción y el razonamiento semántico sobre el lenguaje. Comprender el lenguaje natural para su aplicación en los hardwares y software.

Descripción de la asignatura

- Esta disciplina estudia el diseño de métodos y algoritmos que reciben como entrada y/o producen como salida datos en forma de lenguaje natural (e.g., texto, voz).
- El curso se centra en el procesamiento de texto aunque se mencionan aplicaciones en procesamiento de voz.

Bibliografía

Mucho del contenido que veremos en el curso es tomado de la siguiente bibliografía:

1. Dan Jurasfky and James H, Martin. Speech and Language Processing, (2nd Edition). Pearson, 2014.
2. Joav Goldberg. Neural Network Methods for Natural Language Processing, Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2017.
3. Christopher Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT press, 1999.

Procesamiento de Lenguaje Natural

- Una cantidad enorme de datos de texto en los medios digitales se generan diariamente (por ejemplo, la Web, las redes sociales, los registros médicos, los libros digitalizados).
- Esto conduce a la necesidad de traducir, analizar y gestionar esta gran cantidad de palabras y texto.
- El procesamiento del lenguaje natural (NLP) es el área que se encarga del diseño de métodos y algoritmos que toman como entrada o producen como salida, sin estructura, textbf datos en lenguaje natural. [Goldberg, 2017]
- El procesamiento del lenguaje natural se centra en el diseño y análisis de algoritmos computacionales y representaciones para procesar el lenguaje humano [Eisenstein, 2018].

Procesamiento de Lenguaje Natural

- Ejemplo de una tarea de NLP: Named Entity Recognition (NER):

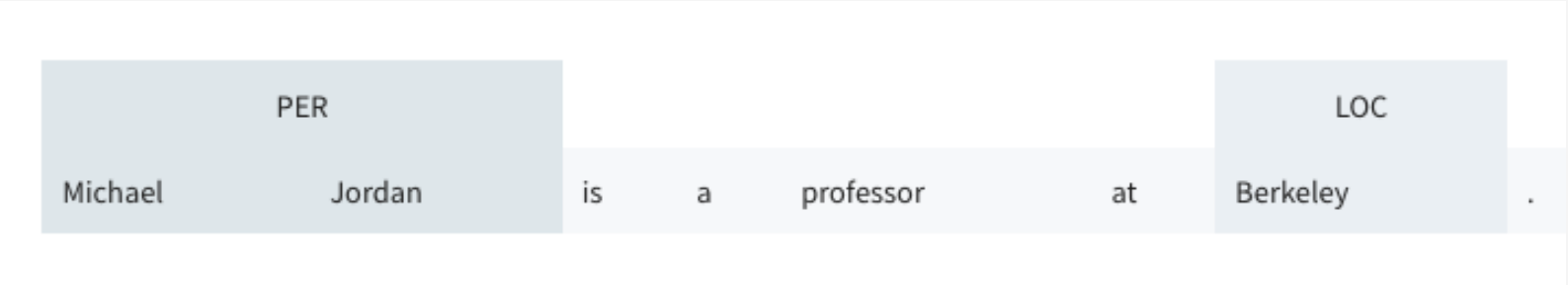


Figura 1: Named Entity Recognition

- El lenguaje humano es muy ambiguo: *I ate pizza with friends* vs. *I ate pizza with olives* vs. *I ate pizza with a fork*.
- También cambia y evoluciona constantemente (por ejemplo, Hashtags en Twitter).

Procesamiento del lenguaje natural y lingüística computacional

El procesamiento del lenguaje natural (PNL) desarrolla métodos para resolver problemas prácticos relacionados con el lenguaje [Johnson, 2014].

- Reconocimiento automático de voz.
- Traducción.
- Extracción de información de documentos.

La lingüística computacional (CL) estudia los procesos computacionales subyacentes al lenguaje (humano).

- ¿Cómo entendemos el lenguaje?
- ¿Cómo producimos el lenguaje?
- ¿Cómo aprendemos el idioma?

Se utilizan métodos y modelos similares en NLP y CL.

Procesamiento del lenguaje natural y lingüística computacional

- La mayoría de las reuniones y revistas que albergan investigaciones sobre el procesamiento del lenguaje natural llevan el nombre de " lingüística computacional " (por ejemplo, ACL, NACL). [Eisenstein, 2018]
- NLP y CL pueden considerarse esencialmente sinónimos.
- Si bien existe una superposición sustancial, existe una diferencia importante en el enfoque.
- CL es esencialmente lingüística respaldada por métodos computacionales (similar a la biología computacional, astronomía computacional).
- En lingüística, el lenguaje es objeto de estudio.
- La NLP se enfoca en resolver tareas bien definidas que involucran el lenguaje humano (por ejemplo, traducción, respuesta a consultas, mantener conversaciones).
- Los conocimientos lingüísticos fundamentales pueden ser cruciales para lograr estas tareas, pero el éxito se mide en última instancia por como se hace la tarea y qué tan bien se hace (de acuerdo con una métrica de evaluación) [Eisenstein, 2018].

Niveles de descripción de la Lingüistica

El campo de la **lingüística** incluye subcampos que se ocupan de diferentes niveles o aspectos de la estructura del **lenguaje**, así como subcampos dedicados a estudiar cómo la estructura lingüística interactúa con la cognición humana y la sociedad.
[Bender, 2013].

1. **Fonética**: El estudio de los sonidos del lenguaje humano.
2. **Fonología**: el estudio de los sistemas de sonido en los lenguajes humanos.
3. **Morfología**: Estudio de la formación y estructura interna de las palabras.
4. **Sintaxis**: El estudio de la formación y estructura interna de las oraciones.
5. **Semántica**: el estudio del significado de las oraciones
6. **Pragmática**: El estudio de la forma en que las oraciones con sus significados semánticos se utilizan para objetivos comunicativos particulares.

Fonética

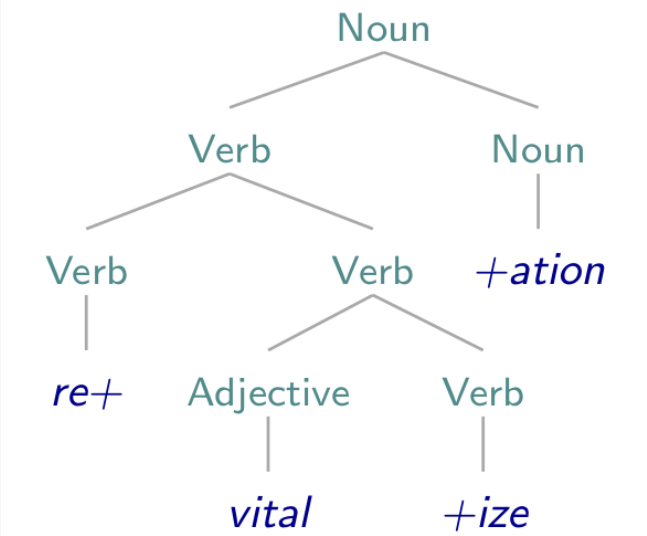
- La fonética estudia los sonidos de un idioma. [Johnson, 2014]
- Se ocupa de los órganos de producción de sonido (por ejemplo, boca, lengua, garganta, nariz, labios, paladar).
- Vocales vs consonantes.
- Las vocales se producen con poca restricción del flujo de aire desde los pulmones hacia la boca y/o la nariz. [Fromkin et al., 2018]
- Las consonantes se producen con alguna restricción o cierre en el tracto vocal que impide el flujo de aire de los pulmones. [Fromkin et al., 2018]
- Alfabeto Fonético Internacional (IPA): sistema alfabético de notación fonética.

Fonología

- Fonología: el estudio de cómo los sonidos del habla forman patrones. [Fromkin et al., 2018].
- Los fonemas son la forma básica de un sonido (por ejemplo, el fonema /p/)
- Example: ¿Por qué **g** no se escucha al pronunciar la palabra *sign* pero si se escucha en *signature*?
- Ejemplo: Los que hablan inglés pronuncian / t / de manera diferente (por ejemplo, *water*)
- En español /z/ se pronuncia de manera diferente en España y América Latina.
- Fonética vs Fonología: `http://www.phon.ox.ac.uk/jcoleman/PHONOLOGY1.htm`.

Morfología

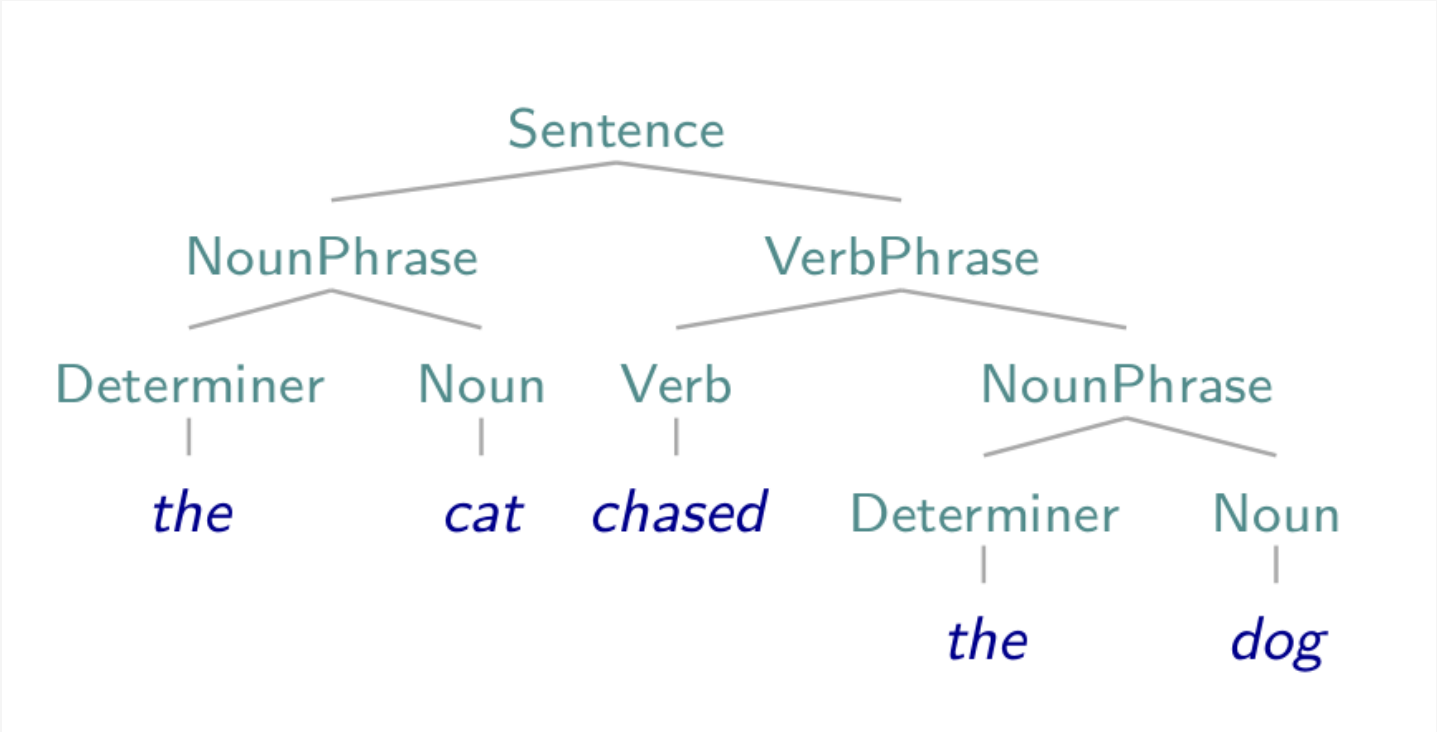
- La morfología estudia la estructura de las palabras (p. Ej.„re+structur+ing,un+remark+able) [Johnson, 2014]
- Morfema: El término lingüístico para la unidad más elemental de forma gramatical [Fromkin et al., 2018]. Ejemplo morphology= morph + ology (la ciencia de).
- Morfología derivada: proceso de formar una nueva palabra a partir de una palabra existente, a menudo agregando un prefijo o sufijo.
- La morfología derivada exhibe una estructura jerárquica. Example: re+vital+ize+ation



- El sufijo generalmente determina la categoría sintáctica (parte del discurso) de la palabra derivada.

Syntaxis

- La sintaxis estudia las formas en que las palabras se combinan para formar frases y oraciones. [Johnson, 2014]



- El análisis sintáctico ayuda a identificar **quién hizo qué a quién**, un paso clave para comprender una oración.

Semántica

- La semántica estudia el significado de palabras, frases y oraciones.[Johnson, 2014].
- Roles semánticos: indican el papel que juega cada entidad en una oración.
- Ejemplos de roles semánticos: **agente** (la entidad que realiza la acción), **tema** (la entidad involucrada en la acción), o **instrumento** (otra entidad utilizada por el agente para realizar la acción).
- Oración anotada: **The boy** cut **the rope** with **a razor**.
- Relaciones léxicas: relación entre diferentes palabras. [Yule, 2016].
- Ejemplos de relaciones léxicas: sinonimia (ocultar / esconder), antonimia (superficial / profunda) e hiponimia (perro / animal).

Pragmática

- **Pragmática**: el estudio de cómo el contexto afecta el significado en determinadas situaciones[Fromkin et al., 2018].
- Ejemplo: como la oración“It’s cold in here” viene a ser interpretado como “close the windows”.
- Ejemplo 2: Can you pass the salt?

Procesamiento de lenguaje natural y aprendizaje automático

- Si bien los seres humanos somos grandes usuarios del lenguaje, también somos muy pobres en comprender y describir formalmente las reglas que gobiernan el lenguaje.
- Comprender y producir un lenguaje usando computadoras es un gran desafío.
- El conjunto de métodos más conocido para tratar con datos lingüísticos se basa en el aprendizaje automático supervisado.
- Aprendizaje automático supervisado: intente inferir patrones de uso y regularidades a partir de un conjunto de pares de entrada y salida previamente anotados (también conocido como conjunto de datos de entrenamiento).

Desafíos del lenguaje

- Tres propiedades desafiantes del lenguaje: discreción, composicionalidad y escasez.
- **Discreción**: no podemos inferir la relación entre dos palabras a partir de las letras que las componen (por ejemplo, hamburguesa y pizza).
- **Composicionalidad**: el significado de una oración va más allá del significado individual de sus palabras.
- **Escasez**: La forma en que las palabras (símbolos discretos) se pueden combinar para formar significados es prácticamente infinito.

Ejemplo de tarea de NLP: clasificación de temas

- Clasificación un documento en una de cuatro categorías: deportes, política, chismes y economía.
- Las palabras de los documentos proporcionan pistas para la identificación de las clases.
- ¿Qué palabras dan qué pistas para la identificación?
- Escribir reglas para esta tarea es bastante desafiante.
- Sin embargo, los lectores pueden clasificar fácilmente varios documentos con respecto a diferentes áreas.
- Un algoritmo de aprendizaje automático supervisado puede crear patrones que toman en cuenta cada una de las palabras que ayudan a categorizar los documentos.

Example 2: Análisis de Sentimientos

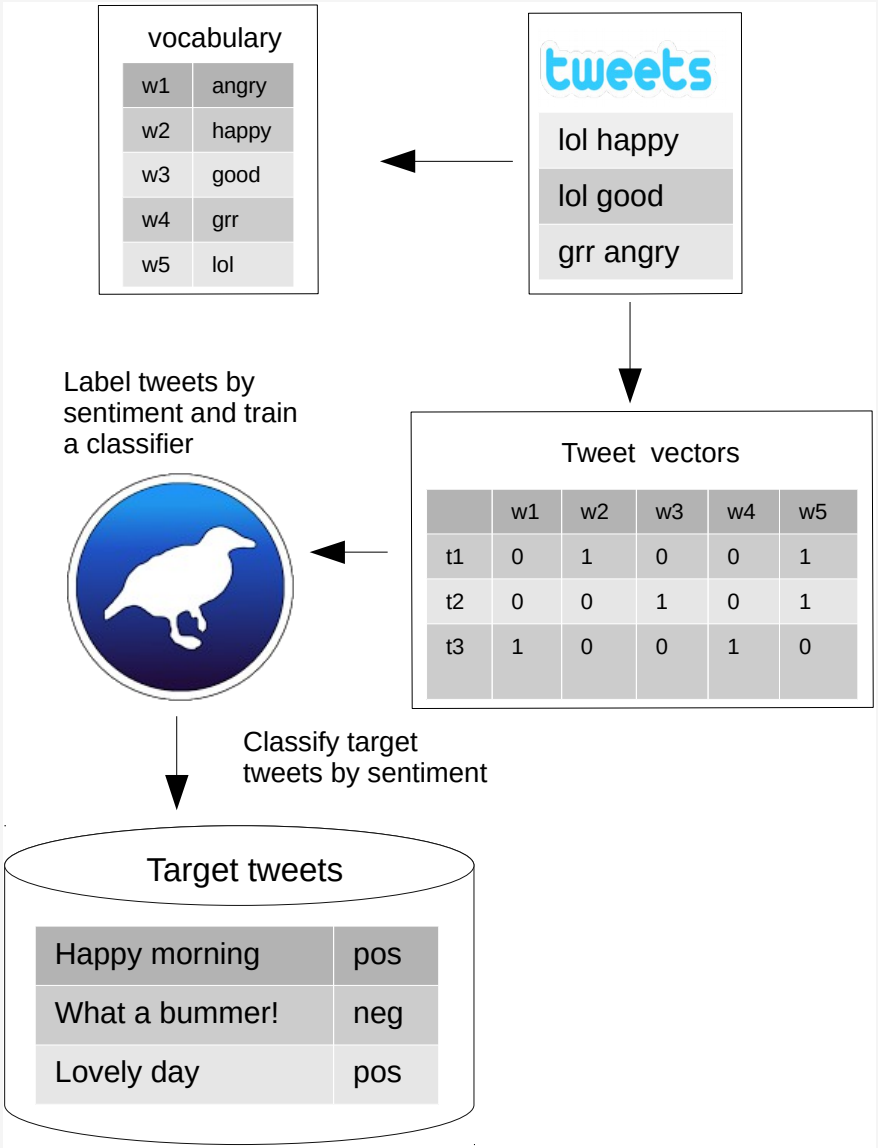
- Aplicación de técnicas de **NLP** para identificar y extraer información subjetiva de conjuntos de datos textuales.

Problema principal: Clasificación de polaridad a nivel de mensaje (MPC)

1. Clasifica automáticamente una oración en clases. **positive**, **negative**, or **neutral**.

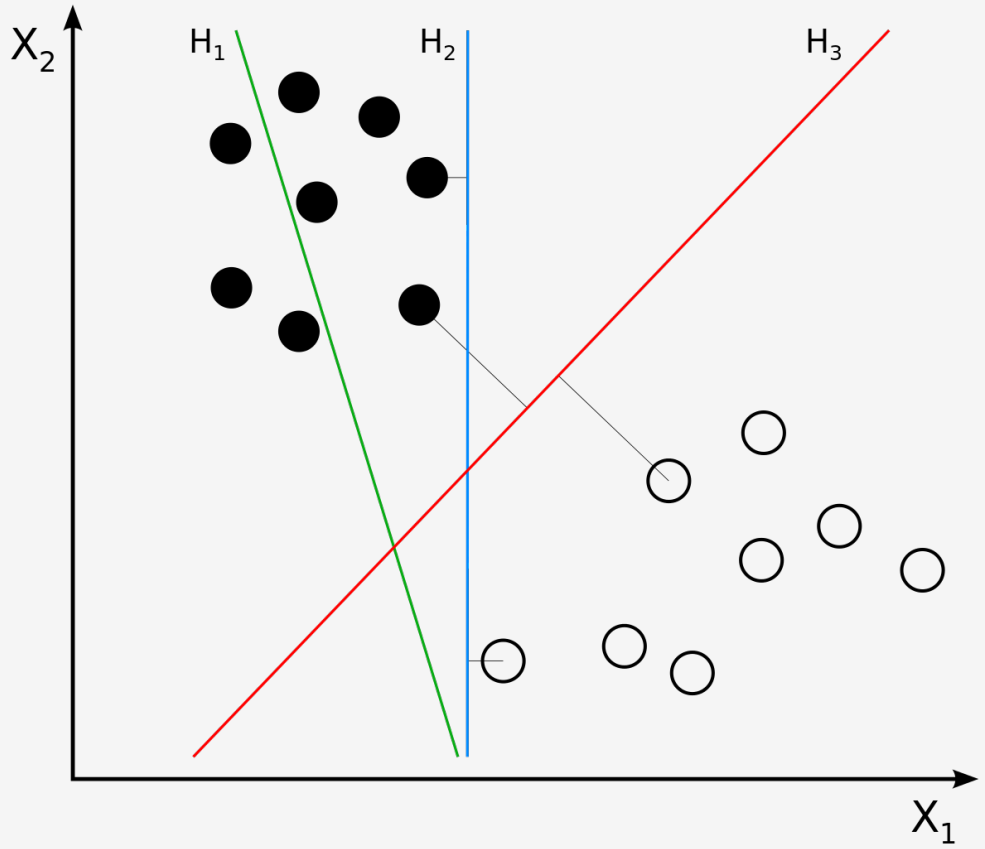
2. Las soluciones que se muestran en el estado del arte usan modelos entrenados de algoritmos **supervizados** entrenados con datasets creados de manera **manual** [Mohammad et al., 2013].

Clasificación de sentimientos a través de aprendizaje supervisado y vectores BoWs



Aprendizaje supervisado: máquinas vectoriales de soporte (SVMs)

- Idea: Encuentre un hiperplano que separe las clases con el margen máximo (separación más grande).



- H_3 separa las clases con el margen máximo.

[1] Image source: Wikipedia

Lingüística y NLP

- Conocer la estructura lingüística es importante para el diseño de características y análisis de errores en NLP [Bender, 2013].
- Los enfoques de aprendizaje automático para NLP requieren características que puedan describir y generalizar en instancias particulares el uso del lenguaje.
- Objetivo: guiar al algoritmo de aprendizaje automático para encontrar correlaciones entre el uso del lenguaje y su conjunto de etiquetas de salida.
- El conocimiento sobre las estructuras lingüísticas puede ayuda en el diseño de algoritmos de aprndizaje automático en NLP.

Desafíos en NLP

- **Costo de anotación**: la anotación manual es **laboriosa** y **requiere mucho tiempo**.
- **Variaciones de dominio**: Los patrones que queremos aprender en los textos pueden variar de un corpus a otro (por ejemplo, deportes, política).
- Un modelo entrenado a partir de datos anotados para un dominio **no necesariamente** funcionará en otro.
- Los modelos entrenados pueden quedar obsoletos con el tiempo (por ejemplo, nuevos hashtags).

Variación de dominio en el sentimiento

1. Ejemplo: For me the queue was pretty **small** and it was only a 20 minute wait I think but was so worth it!!! :D @raynwise
2. Odd spatiality in Stuttgart. Hotel room is so **small** I can barely turn around but surroundings are inhumanly vast & long under construction.

Superar los costos de anotación de datos

Supervisión distante

- Automáticamente **label** datos sin etiquetar (**API de Twitter**) mediante un método heurístico.
- **Enfoque de anotación de emoticonos (EAA)**: tweets con emoticones positivos **:)** o negativos **:(** son etiquetados deacuerdo a la polaridad que indica cada emoticon [Read, 2005].
- El emoticon es **removido** del contenido.
- El mismo enfoque se ha extendido utilizando hashtags #anger y emojis.
- No es trivial encontrar técnicas de supervisión a distancia para todo tipo de problemas de NPL.

Crowdsourcing

- Depende de servicios de **Amazon Mechanical Turk** o **Crowdflower** para para recolectar datos.
- Esto puede ser muy costoso.
- Es difícil garantizar la calidad.

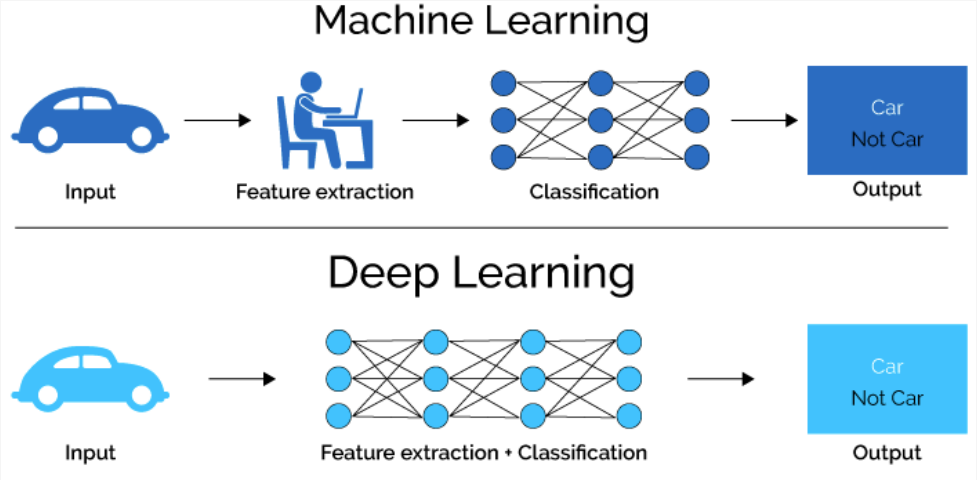
Clasificación de sentimientos de Tweets

- En 2013, el taller de Evaluación Semántica (SemEval) organizó el “ Análisis de sentimiento en la tarea de Twitter ” [Nakov et al., 2013].
- La tarea se dividió en dos subtareas: la expresión nivel y el nivel del mensaje.
- Nivel de expresión: enfocado en determinar la polaridad de sentimiento de un mensaje según una entidad marcada dentro de su contenido.
- Nivel de mensaje: la polaridad debe determinarse de acuerdo con el mensaje general.
- Los organizadores publicaron conjuntos de datos de capacitación y prueba para ambas tareas. [Nakov et al., 2013]

Feature Engineering y Deep Learning

- Hasta 2014, la mayoría de los sistemas de NPL de última generación se basaban en modelos de ingeniería de características + aprendizaje automático supervizado (por ejemplo, SVM, HMM).
- Diseñar las características de un buen sistema de NLP requiere mucho conocimiento específico del dominio.
- Los sistemas de aprendizaje profundo, por otro lado, se basan en redes neuronales para aprender automáticamente buenas representaciones.

Feature Engineering y Deep Learning



- El aprendizaje profundo produce excelentes resultados en la mayoría de las tareas de NLP.
- Grandes cantidades de datos de entrenamiento y máquinas con GPUs son la clave para del éxito del aprendizaje profundo.
- Las **Redes Neuronales** y técnicas de **word embeddings** desempeñan un papel clave en los modelos modernos de NLP.

Deep Learning y Conceptos lingüísticos

- Si los modelos de aprendizaje profundo pueden aprender representaciones automáticamente, ¿siguen siendo útiles los conceptos lingüísticos (p. Ej., Sintaxis, morfología)?
- Algunos defensores del aprendizaje profundo argumentan que tales propiedades lingüísticas inferidas y diseñadas manualmente no son necesarias, y que la red neuronal aprenderá estas representaciones intermedias (o equivalentes, o mejores) por sí mismas [Goldberg, 2016].
- Los científicos aún están deliberando sobre esto.
- Goldberg cree que muchos de estos conceptos lingüísticos pueden ser inferidos por la red por sí sola si se les proporcionan suficientes datos.
- Sin embargo, para muchos otros casos, no tenemos suficientes datos de entrenamiento disponibles para la tarea que nos interesa y, en estos casos, proporcionar a la red los conceptos generales más explícitos puede ser muy valioso.

History

El progreso de la NLP se puede dividir en tres oleadas principales: 1) racionalismo, 2) empirismo y 3) aprendizaje profundo[Deng and Liu, 2018].

1950 - 1990 Racionalismo: enfoques destinados a diseñar reglas hechas a mano para incorporar conocimientos y mecanismos de razonamiento en sistemas inteligentes de PNL (por ejemplo, ELIZA para simular un psicoterapeuta rogeriano, MARGIE para estructurar información del mundo real en ontologías de conceptos).

1991 - 2009 Empirismo: caracterizado por la explotación de corpus de datos y de aprendizaje automático (superficial) y modelos estadísticos (por ejemplo, Naive Bayes, HMM, modelos de traducción de IBM).

2010 - act. Aprendizaje profundo: la ingeniería de características (considerada un cuello de botella) se reemplaza con el aprendizaje de representación y/o redes neuronales profundas (por ejemplo, `https://www.deepl.com/translator`). Un artículo muy influyente en esta revolución: [Collobert et al., 2011].

Roadmap

En este curso introduciremos conceptos modernos en el procesamiento del lenguaje natural basados en **modelos estadísticos** y **redes neuronales**. Los principales conceptos a cubrir se enumeran a continuación:

1. Clasificación de texto.
2. Modelos lineales.
3. Naive Bayes.
4. Modelos ocultos de Markov (Hidden Markov Models).
5. Neural Networks.
6. Word embeddings.
7. Convolutional Neural Networks (CNNs)
8. Recurrent Neural Networks: LSTMs, GRUs.
9. Modelos de Atención.
10. Modelos Secuencia-Secuencia.

Links Importantes

1. Plataforma de ALINNCO: `https://digital.alinnco.mx/`
2. El link del curso: `https://github.com/gdesirena/NLP_Course/`

Referencias I


Bender, E. M. (2013).
Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax.
Synthesis lectures on human language technologies, 6(3):1–184.


Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011).
Natural language processing (almost) from scratch.
Journal of machine learning research, 12(Aug):2493–2537.


Deng, L. and Liu, Y. (2018).
Deep Learning in Natural Language Processing.
Springer.


Eisenstein, J. (2018).
Natural language processing.
Technical report, Georgia Tech.


Fromkin, V., Rodman, R., and Hyams, N. (2018).
An introduction to language.
Cengage Learning.

Referencias II


Goldberg, Y. (2016).
A primer on neural network models for natural language processing.
J. Artif. Intell. Res.(JAIR), 57:345–420.


Goldberg, Y. (2017).
Neural network methods for natural language processing.
Synthesis Lectures on Human Language Technologies, 10(1):1–309.


Johnson, M. (2014).
Introduction to computational linguistics and natural language processing (slides).
2014 Machine Learning Summer School.


Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013).
Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets.
Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013).

Referencias III

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013).
Semeval-2013 task 2: Sentiment analysis in twitter.
In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Read, J. (2005).
Using emoticons to reduce dependency in machine learning techniques for sentiment classification.
In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yule, G. (2016).
The study of language.
Cambridge university press.