Seminario Permanente de Formación en Inteligencia Artificial Aplicada a la Defensa

Noviembre, 2020

Procesamiento del Lenguaje Natural

Eugenio Martínez Cámara

Instituto Andaluz de Investigación en Data Science and Computational Intelligence (DaSCI)

Dpto. Ciencias de la Computación e I.A.
Universidad de Granada

emcamara@decsai.ugr.es
http://sci2s.ugr.es

DaSCI

UNIVERSIDAD DE GRANADA

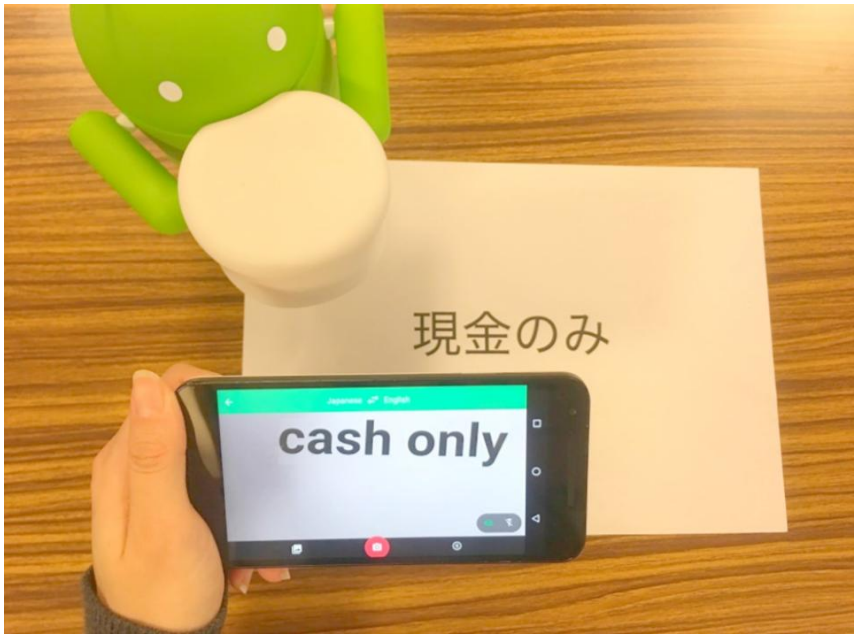
Índice

- Introducción al Procesamiento del Lenguaje Natural.
  - Motivación.
  - Definición y Objetivos.
  - Retos.
  - Orígenes.
    - Lingüistas vs Ingenieros.
  - Dónde encuadrarlo.
  - Tendencias.
- El lenguaje como tipo de dato.
  - Bolsa de palabras.
  - Cálculo de pesos.
  - Modelo de Espacio Vectorial.
  - *Word Embeddings*.

Índice

- Niveles de análisis del lenguaje.
  - *Tokenización* y Segmentación.
  - Análisis Léxico.
  - Análisis Sintáctico.
  - Análisis Semántico.
- Aplicaciones
  - Extracción de Información.
- Análisis de opiniones.
  - Niveles de Análisis de Opinión.
  - Corpora.
  - Recursos.
  - Retos.
  - Análisis de Opiniones en Twitter.
  - Negación.
  - Ejemplo de clasificación de la opinión.
  - Ejemplo clasificación de la opinión a nivel de aspecto.

INT. Motivación



INT. Motivación

INT. Motivación

INT. Motivación

- Satya Nadella (CEO Microsoft) afirmó en 2016 en la conferencia de desarrolladores de Microsoft (Build 2016) lo siguiente:

Human language is the new user interface layer.

To do that you have to infuse (intelligence) into the computers around us, you have to bring forth these technologies of artificial intelligence in machine learning so that we can teach computers to learn the human language, have conversational understanding, teach them about the broad contexts of personal preferences and knowledge so that they can help you with your everyday task.

Cortana

Hey Siri

Google Assistant

INT. Definición y Objetivos

El Procesamiento del Lenguaje Natural (PLN) es área de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos que faciliten la interrelación hombre/máquina y permitan una comunicación mucho más fluida y menos rígida que los lenguajes formales.

INT. Definición y Objetivos

Disciplina centrada en el diseño e implementación de aplicaciones informáticas que se comunican con personas mediante el uso de lenguaje natural [1].

Conjunto de métodos que hacen que el lenguaje humano sea accesible a los ordenadores [2].

[1] Dale, R., Somers, H. L., & Moisl, H., editores (2000). Handbook of Natural Language Processing. Marcel Dekker, Inc., New York, NY, USA, Primera Edición. ISBN 0824790006.

[2] Eisenstein, J. (2019). Introduction to Natural Language Processing. The MIT Press Cambridge, Massachusetts London, Englad. ISBN 9780262042840.

INT. Definición y Objetivos

De la definición se colige que el PLN persigue:

- Diseñar formalismos para la representación del lenguaje y del conocimiento subyacente en el mismo.
- Diseñar técnicas computacionales para el tratamiento del lenguaje.
- Diseñar técnicas computacionales para inferir conocimiento a partir de un fragmento de texto.
- Diseñar técnicas computacionales para el **entendimiento** del lenguaje.
- Diseñar técnicas computacionales que permitan la **generación** de lenguaje.
- Diseñar sistemas que se comuniquen con las personas en su propia lengua, es decir, en lenguaje natural.

Inmediato

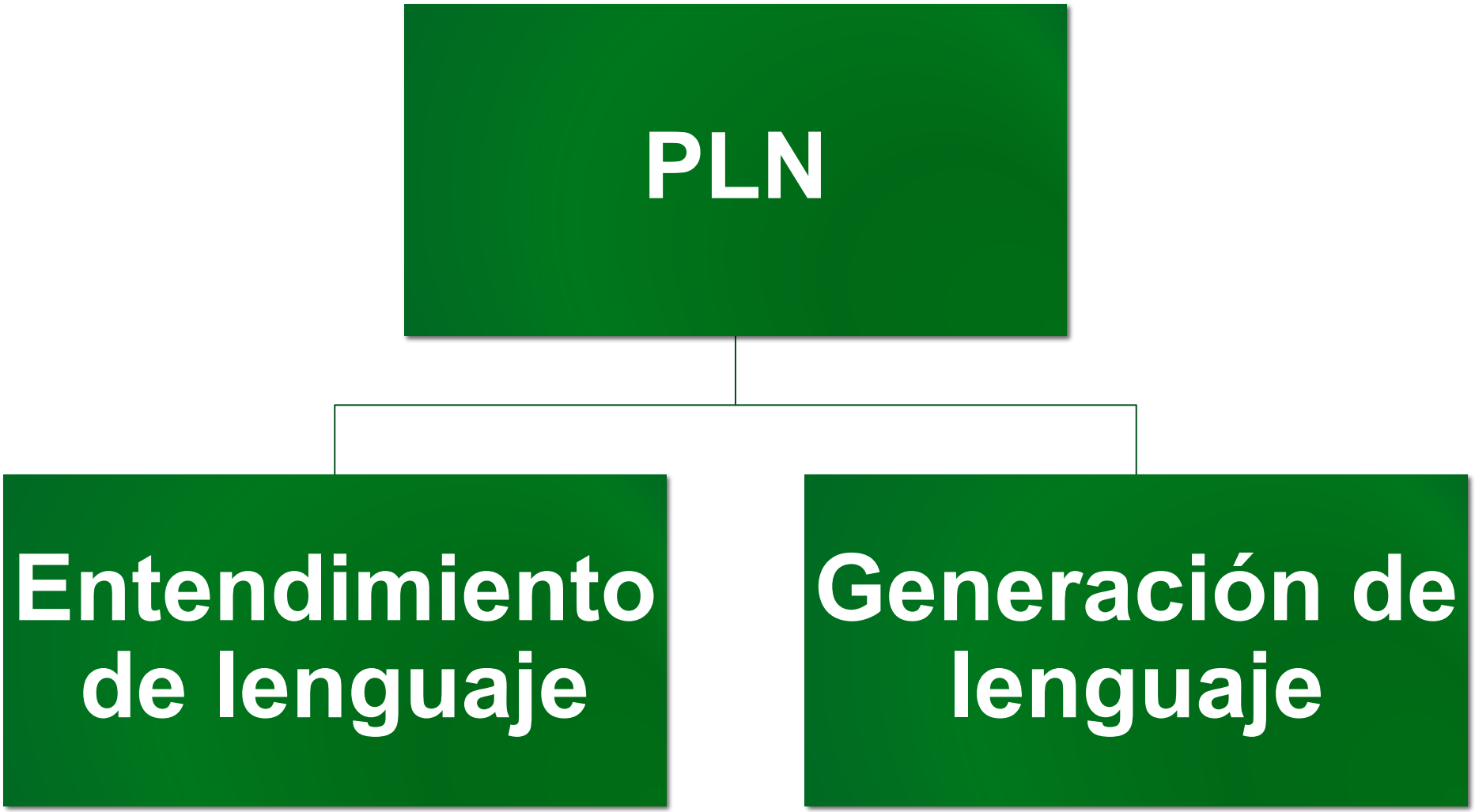
- Construir sistemas que puedan procesar texto y habla más eficientemente.

Final

- Construir sistemas computacionales que sean capaces de comprender y generar el lenguaje natural de la misma manera que lo hacen los humanos.

INT. Definición y Objetivos

INT. Definición y Objetivos



INT. Retos


Ambigüedad
Fonológica y fonética
Léxica
Sintáctica
Semántica
Referencial

INT. Ambigüedad Fonológica y Fonética

- Este tipo de ambigüedad se produce en sistemas de procesamiento del habla.

- El reto se encuentra en asociar correctamente el sonido de una palabra con la palabra correspondiente.

Beodos vs Veo dos.

***I scream* vs *Ice scream*.**

INT. Ambigüedad léxica

- Problema clásico en PLN → *Word Sense Disambiguation*.
- Reconocer el significado correcto de cada palabra en función de su contexto.
- Se enfrenta a la polisemia.

Me suelo sentar en el banco de al lado de mi banco.

Vaya banco de peces se ve en el estanque sentado desde el banco del parque.

Te espero en el banco de en frente del banco.

INT. Ambigüedad sintáctica

- Se refieren a que una oración se puede interpretar de diferente forma debido a al ambigüedad de su estructura.

María estaba en la clase completamente limpia.

¿Quién estaba limpia, la clase o María?

Pedro vio a Juan en lo alto de la montaña con los prismáticos

¿Quién tiene los prismáticos Pedro o Juan?

INT. Ambigüedad Semántica

- Dificultad en identificar el sentido adecuado de una oración, dado que puede tener varios significados.
- Se diferencia de la ambigüedad léxica, en que ésta debe elegir entre un conjunto finito de significados dependientes del contexto, mientras que la ambigüedad semántica entre un conjunto finito de interpretaciones.

Juan dio un pastel a los niños.
¿Uno para todos?
¿Uno para cada uno?
Le compró flores
¿Se refiere al tendero?
¿Se refiere a la persona que las recibe?

INT. Ambigüedad Referencial

- Ocurre cuando no es posible identificar con certeza como oraciones anteriores afectan a la interpretación de las siguientes oraciones.

Juan le dijo a Pepe que pusiera aquello allí, pero como no le hizo caso, él no lo encontró luego.

¿Qué se tenía que cambiar de lugar?

¿Quién no encontró aquello?

¿Quién no hizo caso?

Él le dijo, después, que lo pusiera encima

¿Quién dijo?

¿A quién?

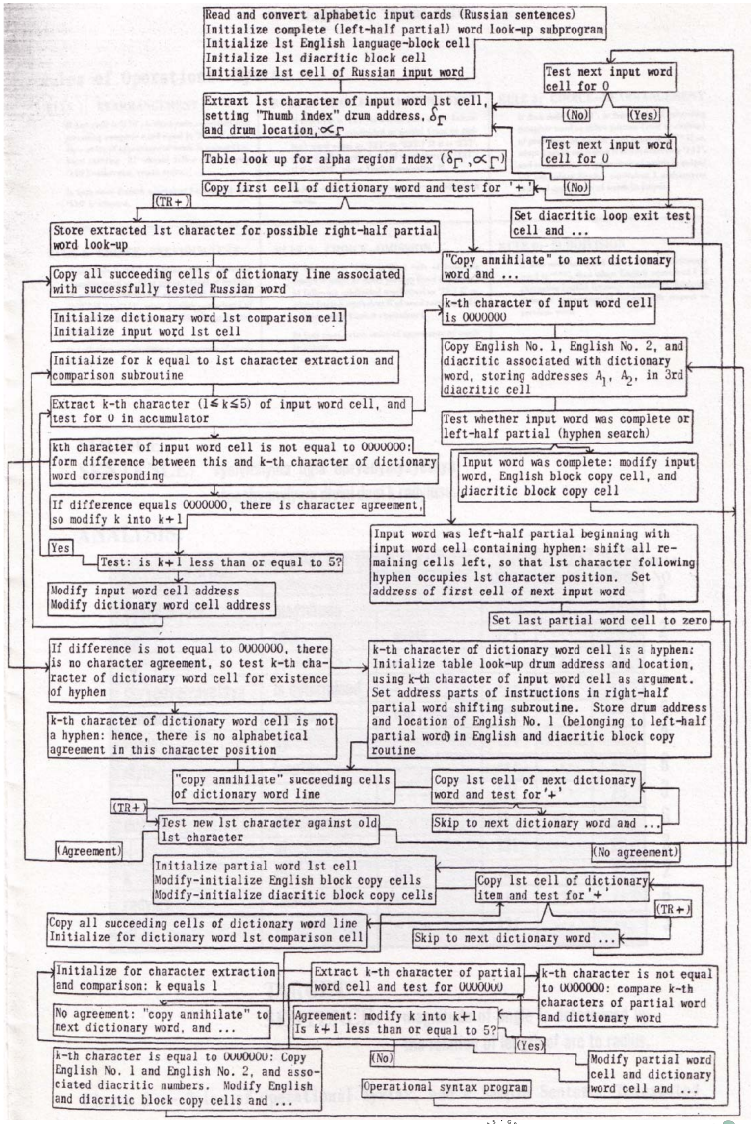
¿Cuándo, después de qué?

¿Que pusiera qué?

¿Encima de dónde?

INT. Orígenes

- El PLN encuentra su origen en los sistemas de traducción automática (TA).
- El primer sistema de TA fue Georgetown Automatic Translator (GAT) desarrollado conjuntamente por IBM y la Universidad de Georgetown.
- GAT era un sistema con 6 reglas lexicográficas para la traducción de ruso a inglés.
- La demostración de experimentación se realizó en 1954 con un pequeño conjunto de oraciones en ruso seleccionadas.
- La propia demostración controlada evidenciaba los problemas del procesamiento de lenguaje natural.



INT. Orígenes

Años 60.

- El PLN consistió principalmente en métodos de análisis de palabras clave o *pattern matching*, dando lugar a sistemas como:
- **Eliza** [3]. Primer sistema de diálogo que simulaba una conversación entre un psicoanalista (sistema) y un paciente (usuario).
- **STUDENT** [4] Sistema de traducción de enunciados de problemas en lenguaje natural a sistemas de ecuaciones con capacidad de cálculo algebraico.
- **SIR** [5]. Primer sistema de inferencia semántica para la respuesta de preguntas cuya respuesta explícitamente no tenía que estar en la base de conocimiento del sistema.

[3] Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1), 36–45.

[4] Bobrow, D. G. (1964). Natural language input for a computer problem solving system.

[5] Raphael, B. 'SIR: Semantic information retrieval', in *Semantic information processing*, Ed. M. Minsky, Cambridge MA: MIT Press, 1968.

INT. Orígenes

Años 70.

- **LUNAR**. Sistema que permite interrogar a una Base de Datos sobre las muestras recogidas en misiones espaciales.

¿Hay alguna muestra que presente más del 13% de aluminio?

- Aparecen diversos analizadores que usan gramáticas independientes del contexto, como SAD-SAM de Lindsay.
- Aparecen las Redes de Transición Aumentadas (ATN), desarrollo que mejora la potencia de las expresiones regulares y de las gramáticas independientes del contexto incorporando restricciones funcionales a un autómata de estados finitos variables.

INT. Orígenes

Años 80.

- Gran interés por los métodos lingüísticos.
- Gramáticas de Cláusulas Definidas.
- Gramáticas de Estructura de Frase Generalizadas.
- Gramáticas Léxico Funcionales.
- Gramáticas de Unificación Funcionales de Kay.
- Mayor aplicabilidad, se desarrollan sistemas cada vez mas sofisticados:
  - Traducción Automática: Ariane-78, EUROTRA o ATLAS.
  - Interfaces con bases de datos: TEAM, CHAT-80, ORBI.

INT. Orígenes

Años 90.

- Resurge el interés por los métodos empíricos.
- Comienzan a desarrollarse recursos lingüísticos para poder aplicar los métodos empíricos:
  - Grandes conjuntos de datos o *corpora*.
  - Ontologías
  - Diccionarios.
- Surgen nuevas necesidades de acceso a la información motivadas por la aparición de Internet, lo cual supuso un impulso a la investigación en PLN.

INT. Orígenes. Lingüistas vs Ingenieros

- La lingüística tiene como objetivo la caracterización de las diversas figuras lingüísticas, lo cual implica:
  - El estudio de la adquisición, generación y compresión del lenguaje.
  - El análisis de las relaciones entres las expresiones lingüísticas y el contexto en el que se producen.
  - La interpretación de las estructuras lingüísticas

- El PLN se sitúa en la interpretación o entendimiento de las estructuras lingüísticas.

INT. Orígenes. Lingüistas vs Ingenieros

INT. Orígenes. Lingüistas vs Ingenieros

Lingüística generativa.

- Noam Chomsky es su principal precursor.
- También conocida como teoría racionalista y Chonskiana.
- Máximo predicamento entre 1960-1985.
- Fundamento: El lenguaje es un mecanismo tan complejo que no puede ser adquirido por los sentidos, por lo que su estructura básica debe estar definida en el cerebro.
- Objetivo: Describir las estructuras del lenguaje humano definidas en el cerebro (Lenguaje-I) a partir de las derivaciones de dicho lenguaje, las cuales se encuentran impresas en los textos.

INT. Orígenes. Lingüistas vs Ingenieros

Tecnología de la Lengua.

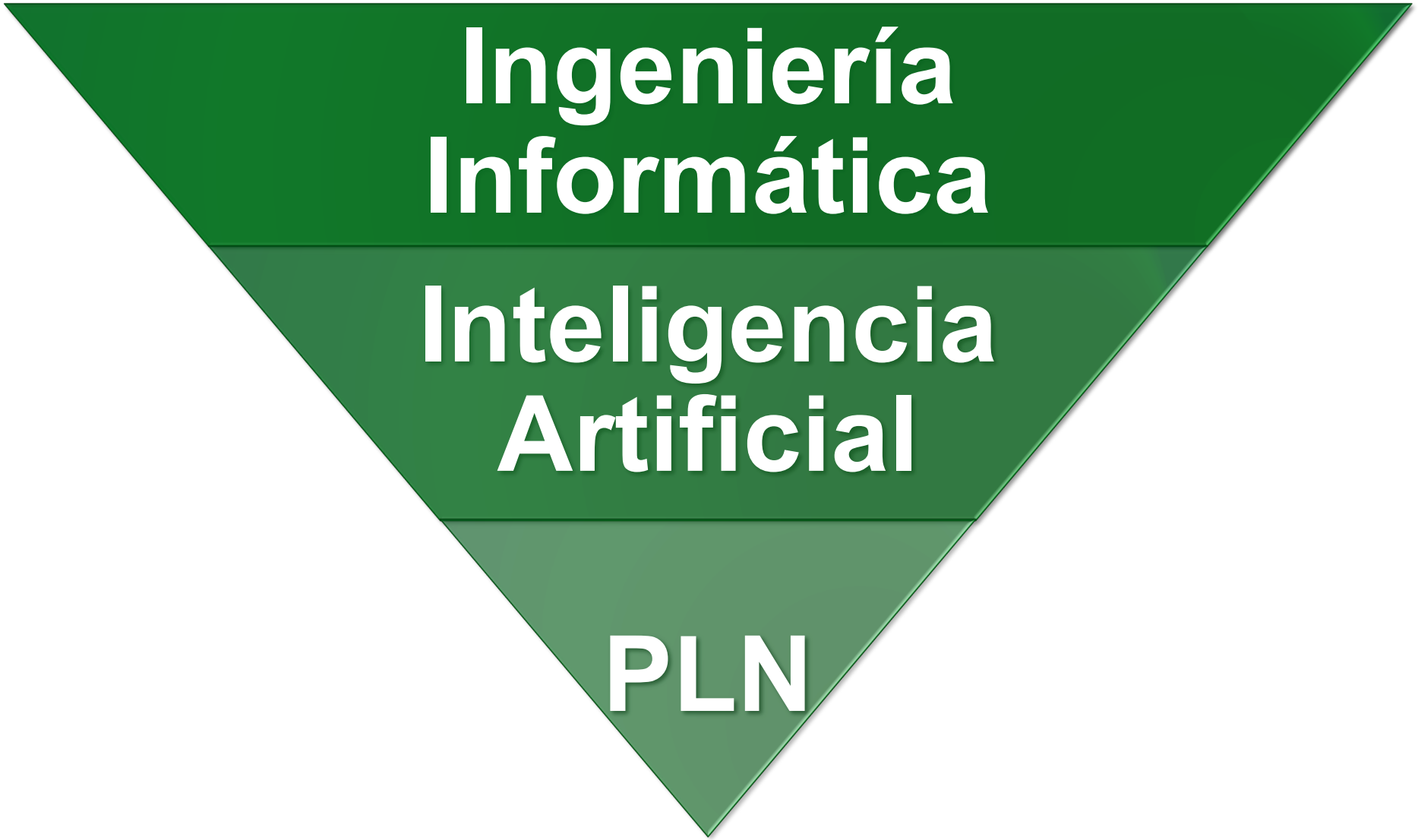
- Conocida también como Ingeniería Lingüística o PLN estadístico.
- Fundamento: La mente humana tiene la capacidad innata de establecer asociaciones, reconocer patrones y de generalizar ocurrencias de eventos que son percibidos a través de los sentidos. Progresivamente el cerebro humano interioriza la estructura del lenguaje hasta que es capaz de entenderlo y generarlo de forma natural.
- Objetivo: A partir de las reproducción física del lenguaje (Lenguaje-E) se persigue modelar e identificar los parámetros del lenguaje mediante el análisis y procesamiento exhaustivo del Lenguaje-E.

INT. Orígenes. Lingüistas vs Ingenieros

- Actualmente el paradigma que tiene un mayor predicamento es el de ingeniería del lenguaje, por:

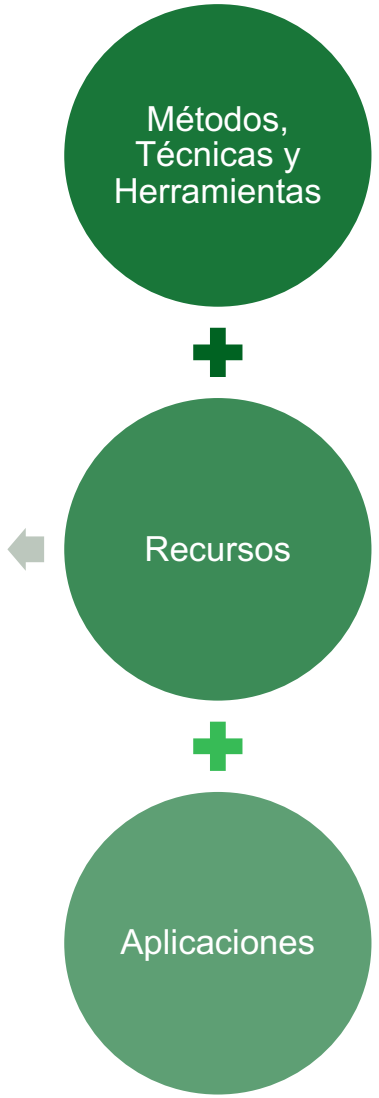
1. El desarrollo métodos computacionales con capacidad de reconocimiento de patrones.
2. La disponibilidad de recursos lingüísticos representativos del uso del lenguaje.
3. La disponibilidad de una mayor capacidad de computo posibilita el procesamiento de grandes cantidades de datos.
4. Aprendizaje basado en ejemplos (recursos) permite una aprendizaje más ajustado a los usos reales del lenguaje que los basados en modelos teóricos.

INT. Dónde encuadrar al PLN



INT. Dónde encuadrar al PLN

El PLN (Tecnologías de la Lengua) es la aplicación del conocimiento de la lengua al desarrollo de sistemas informáticos capaces de reconocer, comprender, interpretar y generar lenguaje humano en todas sus formas



INT. Tendencias

- La investigación y la aplicación de técnicas de PLN se caracteriza actualmente por:

1. Aplicar los últimos métodos de aprendizaje: *deep learning*, *reinforcement learning*, *transfer learning*.
2. Mejora en la representación continua de palabras: *word embeddings*, modelos pre-entrenados como BERT.
3. Desarrollo de sistemas *cross-lingües*: entrenamiento en un idioma y evaluación en un idioma distinto.
4. Nuevas aplicaciones dirigidas al análisis del significado de un mensaje, por ej.: minería de la argumentación.
5. Mejora en la anotación de datos, prestando atención en evitar sesgos en su anotación.
6. Desarrollo de bases de conocimiento: BabelNet.

EL LENGUAJE COMO TIPO DE DATO

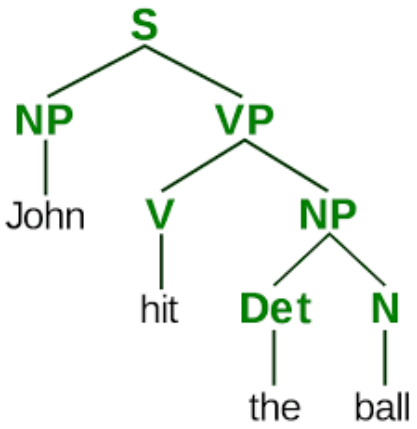
El Lenguaje como Tipo de Dato

- El texto es un dato no estructurado.
- Su procesamiento requiere su estructuración.
- Una máquina no puede procesar texto de manera directa.
- Hay que transformar el texto en números que puedan ser procesados por un algoritmo de aprendizaje automático.

Raw text	Structured facts
Marie Curie was born on November 1867, 7. She was a Polish and naturalized-French physicist and chemist who conducted pioneering research on radioactivity...	(Marie Curie, date_of_birth, 11/07/1867)
	(Marie Curie, nationality_at_birth, Polish)
	(Marie Curie, nationality_at_death, French)
	(Marie Curie, job, physicist)
	(Marie Curie, job, chemist)
	(Marie Curie, job, researcher)
	(Marie Curie, research_field, radiocativity)
	...

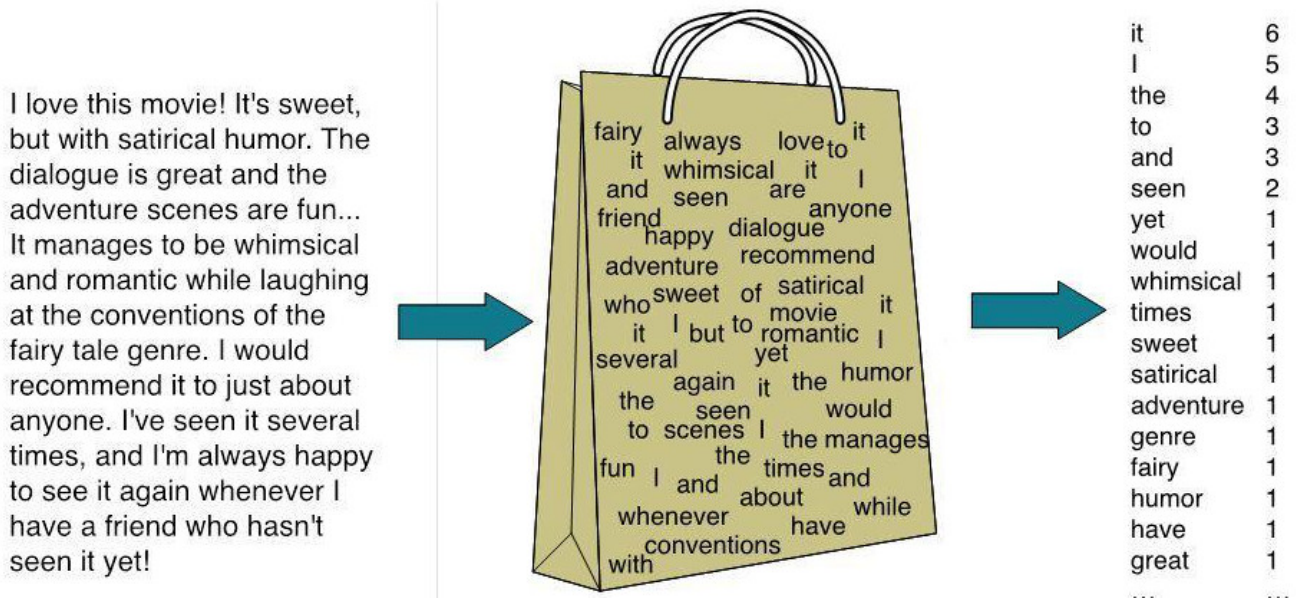
El Lenguaje como Tipo de Dato

- El texto requiere ser transformado en una representación estructurada para su procesamiento.
- Convertir el texto en un conjunto de **características** representativas.
- Características:
  - Carácter.
  - Subpalabras: morfemas, lexemas, *stems*, prefijos, sufijos.
  - Palabras: unigramas.
  - Conjuntos de palabras: bigramas, trigramas, *n-gramas*.
  - Categorías morfológicas (*pos-tags*): verbo, artículo, sustantivo…
  - Árboles sintácticos.
  - Representación semántica.



El Lenguaje como Tipo de Dato. Bolsa de palabras

- La representación más simple se conoce como **bolsa de palabras** (*bag of words*).
- No tiene en cuenta el orden de las palabras.
- Sólo presta atención a la aparición o frecuencia de las palabras en un documento.



El Lenguaje como Tipo de Dato. Cálculo de pesos

- Dado un conjunto de documentos, y el vocabulario conformado por todas las palabras únicas que aparece en tal conjunto, la representación por bolsa de palabras ofrece una representación de la distribución de tal vocabulario en el conjunto de documentos.
- ¿Se puede medir la importancia de las palabras?
- Sí se puede → Cálculo de pesos.
- El resultado del proceso de representación es una matriz término-documento o documento término.

	T_1	T_2	....	T_t
D_1	w_{11}	w_{21}	...	w_{t1}
D_2	w_{12}	w_{22}	...	w_{t2}
:	:	:		:
:	:	:		:
D_n	w_{1n}	w_{2n}	...	w_{tn}

El Lenguaje como Tipo de Dato. Cálculo de pesos

- El cálculo del peso/importancia/relevancia de cada palabra en cada documento ($w_{i,j}$) es una forma de representar la información que aporta al proceso de aprendizaje.
- **Pesos Binario**. El valor de $w_{i,j} \in \{0, 1\}$, e indica si una palabra aparece o no en un documento. Proporciona una información local.
- **Frecuencia absoluta/relativa**. El valor de $w_{i,j}$ es el nº de repeticiones absolutas o relativas al:
  - Número de palabras por documento: Sólo aporta información local.
  - Número de palabras del conjunto de datos. Aporta información global de todo el conjunto de datos.

El Lenguaje como Tipo de Dato. Cálculo de pesos

- **Peso TF-IDF**. El objetivo de asignar un peso a una palabra es medir su capacidad de discriminación entre clases.
- Sería recomendable que palabras poco frecuentes a nivel de todo el corpus pero muy frecuentes en un subconjunto del corpus, tuvieran más peso que otras palabras muy frecuentes a nivel de todo el corpus.
- Este tipo de peso facilitaría la recuperación de información y la clasificación de temas.
- Ej. Dominio: fútbol.
  - Palabras frecuentes a nivel de corpus: noticia, enfermedad, evento
  - Palabras frecuentes a nivel de subconjunto: balón, portería, árbitro, partido, gol, lateral…

El Lenguaje como Tipo de Dato. Cálculo de pesos

- **Peso TF-IDF**.

$$w_{i,j} = tf_{i,j} * idf_i$$

$tf_{i,j}$ → Frecuencia absoluta del término *i* en el documento *j*.

idf_i → Frecuencia inversa del término *i* a nivel global de corpus.

$$idfi = log \frac{N}{n_i}$$

N → Número total de documentos.

n_i → Número de documentos en los que aparece el término *i*.

TF-IDF Normalizado

$$w_{i,j} = \frac{tf_{i,j} * idf_i}{\sum_{k=1}^{|d_j|} w_{k,j}}$$

El Lenguaje como Tipo de Dato. Modelo de Espacio Vectorial

- La representación matricial del esquema de representación bolsa de palabras origina un espacio vectorial [6]:
  - Cada palabra se representa por un vector de pesos de longitud igual al número de documentos.
  - Cada documento se representa por un vector con los pesos de las palabras que aparecen en él, y una longitud igual al número de palabras del vocabulario.
- Resultado: Un conjunto de documentos se puede representar como un espacio vectorial, de forma que los datos no estructurados ya se han estructurado matemáticamente, y por ende pueden ser procesados.

[6] Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141-188.

El Lenguaje como Tipo de Dato. Modelo de Espacio Vectorial

- El Modelo de Espacio Vectorial es muy versátil.
- En lugar de términos individuales, *unigramas*, se puede ponderar grupos de palabras para representar información estructural:

  - Bigramas: Parejas de términos.

    El perro tiene hambre → (el, perro), (perro, tiene), (tiene, hambre)
  - Trigramas: Tripletas de términos.

    El perro tiene hambre → (el, perro, tiene), (perro, tiene, hambre)
  - N-gramas: N-grupos de palabras.

El Lenguaje como Tipo de Dato. Modelo de Espacio Vectorial

- Dependiendo de la información que se quiera representar se pueden construir otro tipo de espacios vectoriales.
- Hipótesis distribucional del lenguaje. Palabras que aparecen en contextos similares, tienen significados similares.
  - Se construyen mediante matrices término-término o término-contexto.
  - Contexto: ventanas de palabras, dependencias gramaticales u otra estructura lingüística de interés.
  - Son la base de las vectores de representación continúa de palabras.
- **Matrices de patrones**. En el caso de que se quiera representar patrones semánticos, se puede definir espacios vectoriales del tipo:
  - Profesional-material: “X usa Y en su trabajo”. X en las filas e Y en las columnas, respondiendo a patrón “X usa Y en su trabajo”. Albañil:ladrillo, carpintero:madera.

El Lenguaje como Tipo de Dato. Modelo de Espacio Vectorial

- **Peso Punto de Información Mutua *(Pointwise Mutual Information*, PMI*)***. Mide la cantidad de información que un par de términos o patrones tienen entre ellos. A mayor coocurrencia en términos o patrones mayor PMI.
- Variante: Punto de Información Mutua Positiva. Los valores negativos se sustituyen por un valor cero.

$$pmi_{i,j} = \log(\frac{p_{i,j}}{p_i * p_j})$$

$$p_i = \frac{\sum_{j=1}^{n_c} f_{i,j}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{i,j}}$$

$$p_i = \frac{\sum_{i=1}^{n_r} f_{i,j}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{i,j}}$$

$$p_{i,j} = \frac{f_{i,j}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{i,j}}$$

El Lenguaje como Tipo de Dato. *Word Embeddings*

- Hasta ahora las características generadas sólo proporcionan información a nivel de corpus.

- ¿Sería posible obtener la representación semántica de una palabra e independiente de un conjunto de entrenamiento?

- Sí es posible, y es lo que se conoce como *word embeddings*.

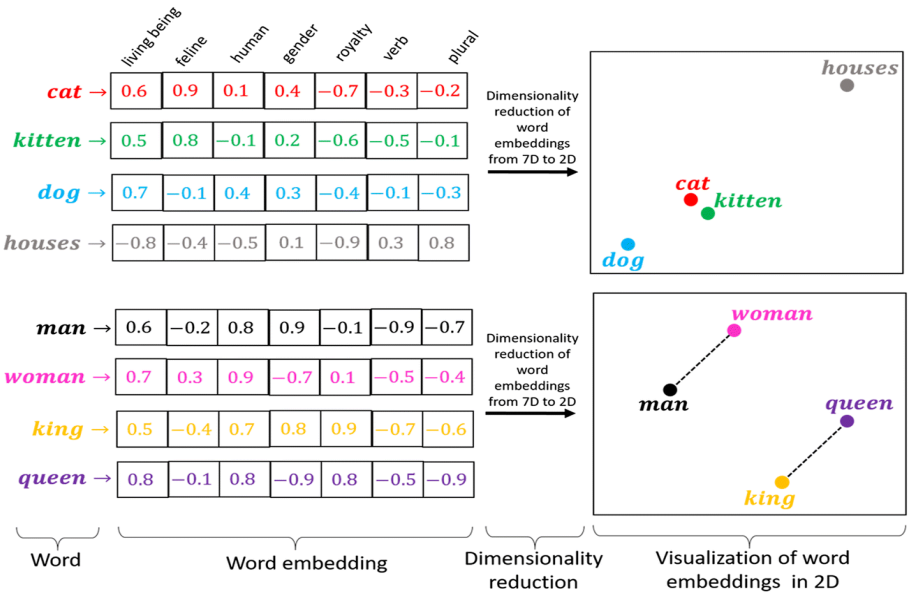
- También se les conoce como representación continua de palabras.

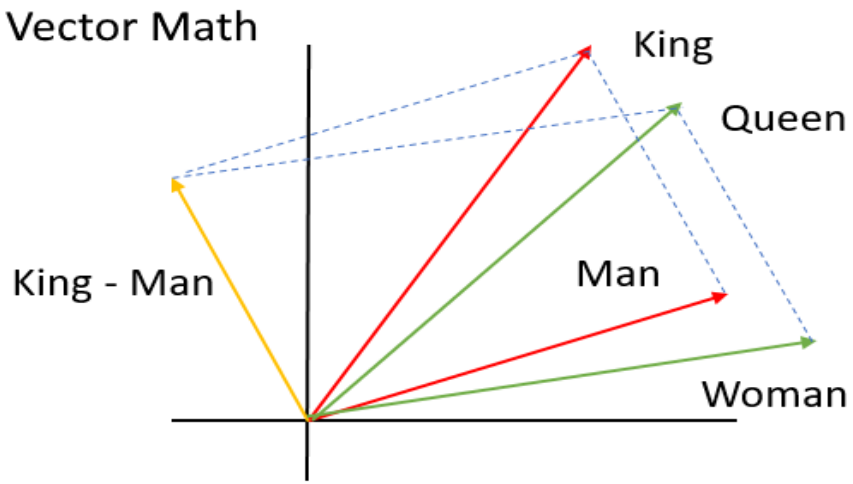
El Lenguaje como Tipo de Dato. *Word Embeddings*

- Los *embeddings* son espacios vectoriales de palabras.
- Cada palabra tiene asociado un vector con *n* componentes, que representa toda su información semántica.
- Dicho espacio vectorial se construye para un idioma (mono-lingües) o para varios (*cross-lingües*) aplicando un modelo de lenguaje, red neuronal, o método basado en co-ocurrencias sobre una gran cantidad de texto representativo del uso del lenguaje en un idioma.
- Ejemplos de conjuntos de *embeddings* pre-entrenados: Word2Vect, Glove, FastText, BERT, Dependency Embeddings…
- Estos modelos suelen estar pre-entrenados para inglés. Para español comienzan a estar disponibles algunos conjuntos de vectores.

El Lenguaje como Tipo de Dato. *Word Embeddings*

- Permiten la construcción no supervisada de características.
- Aumentan la capacidad de representación al no depender las características del conjunto de entrenamiento.
- Los modelos de *embeddings* muestran propiedades algebraicas.
- Palabras similares tienen vectores cercanos.
- Permiten operaciones del estilo: (Rey-Hombre)+Mujer = Reina.





El Lenguaje como Tipo de Dato. *Word Embeddings*

- Word2Vec [6]: https://code.google.com/archive/p/word2vec/
- Glove [7]: https://nlp.stanford.edu/projects/glove/
- FastText [8]: https://fasttext.cc/

[6] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

[7] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[8] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

El Lenguaje como Tipo de Dato. Conclusión.

- La lengua sí se puede procesar.
- Sólo es preciso transformar un texto a una representación matemática.
- Una vez que se obtiene dicha representación matemática ya se puede aplicar cualquier tipo de algoritmo de aprendizaje.
- El proceso de transformación matemática no es más que un proceso d generación de características.
- Las naturaleza de las características dependerá del problema en cuestión.
- Hasta ahora sólo se han mostrado las características más usadas o estándar.

NIVELES DE ANÁLISIS DEL LENGUAJE

Niveles de Análisis del Lenguaje

- Para el procesamiento de la lengua es necesario una representación abstracta de la misma.

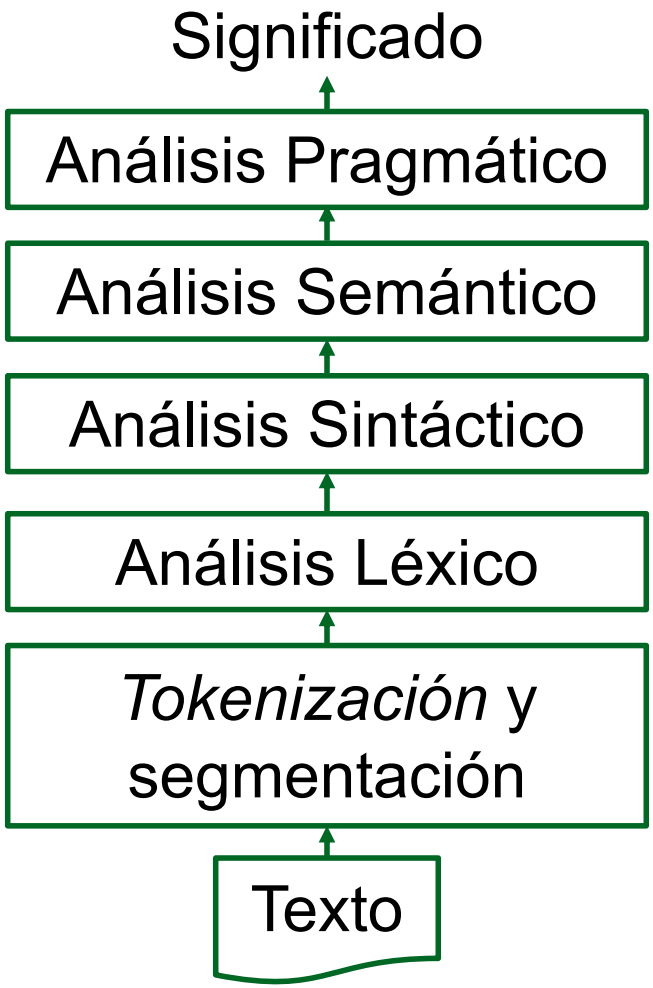
- Para generar dicha representación es necesario analizar el lenguaje que constituye un mensaje.

- Según la lingüística se identifican tres niveles de análisis:
  - Análisis Sintáctico: Se analiza la estructura de un mensaje y si se adecua a la gramática de una lengua.
  - Análisis Semántico: Se identifica el significado de las palabras y de las figuras lingüísticas presentes en un mensaje.
  - Análisis Pragmático: Se trata de determinar el significado del discursos subyacente en el mensaje.

Niveles de Análisis del Lenguaje

- Desde un punto de vista práctico, los 3 niveles mencionados son insuficientes para el procesamiento de un mensaje.
- Antes del análisis sintáctico se requiere identificar las unidades mínimas del lenguaje: las palabras y *tokens*, así como toda la información relativa a ellas.
- Esto es propio del Análisis Léxico.
- A su vez, el análisis léxico requiere identificar las unidades mínimas que constituyen un mensaje, así como su organización en oraciones.
- Por tanto se debe añadir una nueva fase de procesamiento conocida como *Tokenización* y Segmentación.

Niveles de Análisis del Lenguaje



NAL. Tokenización y Segmentación

- Objetivo: Dado un fragmento de texto identificar los *tokens* y palabras, así como el conjunto de oraciones en el que se organiza el texto.

- Entrada: Un texto de origen a priori indeterminado:
  - Texto procedente de una página web.
  - Texto procedente de la transcripción léxica de una grabación de voz.
  - Texto procedente de un reconocedor de caracteres OCR.
  - …

- Salida: Un conjunto de *tokens* y palabras organizados en oraciones.

NAL. Tokenización y Segmentación

Ejemplo

<p>Lo sabe todo, absolutamente todo. Figúrense lo tonto que será.<p>

[[‘Lo’, ‘sabe’, ‘todo’, ‘,’, ‘absolutamente’, ‘todo’, ‘.’][‘Figúrense’, ‘lo’, ‘tonto’, ’que’, ‘será’, ‘.’]]

Frase de Miguel de Unamuno

NAL. Tokenización y Segmentación

Dependencias

Codificación de caracteres	Idioma	Corpus	Aplicación

NAL. Tokenización y Segmentación

Técnicas de *tokenización* y segmentación.

- Se suelen seguir enfoques basados en reglas, los cuales se fundamentan en el uso de gramáticas y expresiones regulares:
  - Alembic [9]: Sistema de extracción de información basado en expresiones regulares a través del uso de Flex (Nicol 1993).
  - Kiss and Strunk [10]: Sistema heurístico que se concentra principalmente en la resolución de las abreviaturas, y posteriormente en la identificación de los límites de las oraciones.
  - *Tokenizador* de Twitter orientado a la detección de Opinión: *Tokenizador* y detector de oraciones para Twitter basado en la definición de expresiones regulares, y está determinado por ser la entrada de un sistema clasificación de la opinión, por lo que está preparado para la identificación de *tokens* con significado de opinión, emoción y sentimiento (emoticonos, onomatopeyas de risa). http://sentiment.christopherpotts.net/tokenizing.html

[9] Nicol, G. T. (1993). Flex – The Lexical Scanner Generator. Cambridge, MA: The Free SotwareFoundation.

[10] Kiss, T. Y J. Strunk (2006). Unsupervised Multilingual Sentence Boundary Detection. Computational Linguistics 32(4), 485-525.

NAL. Análisis Léxico

- Las unidades mínimas de información de todo sistema de PLN son las palabras.
- La fase de *Tokenización* y Segmentación ha identificado las palabras y su organización en oraciones, pero no ha añadido ninguna información relacionada con dichas palabras.
- Entrada: Conjunto de palabras y *tokens* organizados en oraciones.
- Objetivo: Asociar a los *tokens* información relacionada con su propia naturaleza y con la función que desempeñan en la oración.
- Salida: Conjunto de *tokens* organizados en oraciones con información fundamental para las siguientes fases de análisis.

NAL. Análisis Léxico

[[‘Lo’, ‘sabe’, ‘todo’, ‘absolutamente’, ‘todo’, ‘.’][‘Figúrense’, ‘lo’, ‘tonto’, ’que’, ‘será’, ‘.’]]

[[(‘Lo’, ‘lo’, ‘PP3MSA0’), (‘sabe’, ‘saber’, ‘VMIP3S0’), (‘todo’, ‘todo’, ‘PI0MS00’), (‘absolutamente’, ‘absolutamente’, ’RG’), (‘todo’, ‘todo’, ‘PI0MS00’), (‘.’, ’.’, ’Fp’)], [(‘Figuren’, ‘Figurar’, ‘VMM03P0’), (‘se’, ‘se’, ‘PP3CN000’), (‘lo’, ‘el’, ‘DA00S0’), (‘tonto’, ‘tonto’, ‘NCMS000’), (‘que’, ’que’, ’ PR0CN00’), (‘será’, ‘ser’, ‘VSIF3S0’), (‘.’, ‘.’, ‘Fp’)]]

NAL. Análisis Léxico

- Analizador léxico: Tiene la capacidad de identificar toda la información relacionada con un término:
  - Lema. Entrada en un diccionario.
  - Familia semántica.
  - Categoría morfológica.
  - En Recuperación de Información incluso identificar su *stem* o lema canónico.

NAL. Análisis Léxico. Categoría Morfológica

- La categoría morfológica marca el rol sintáctico que cada palabra ejerce en una oración.
- Se le suele conocer por el nombre de *pos-tags*.
- Los sistemas que los calculan se llaman *pos-taggers*.
- ¿Qué son los roles sintácticos?
  - Si una palabra es un verbo, y ya que es un verbo su tiempo, número y género.
  - Si una palabra es nombre propio o común, así como su número y género.
  - Si una palabra es adverbio, adjetivo, signo de puntuación, cifra…
- La categoría morfológica de una palabra depende del contexto en el que se encuentra. Por ej.: la palabra “casa” puede ser:
  - Nombre común: La casa.
  - Tercera persona del presente del verbo casar: Él se casa.

NAL. Análisis Léxico. Categoría Morfológica

- La identificación de las categorías morfológicas dependen del conjunto de etiquetas morfológicas.
  - *Tagset* de EAGLE. Conjunto de etiquetas morfológicas para el español definido por el grupo EAGLE, el cual está promovido por la comisión europea. Este *tagset* es usado por el software de PLN Freeling: http://blade10.cs.upc.edu/freeling-old/doc/tagsets/tagset-es.html.
  - *Tagest* de *Universal Dependencies*. Conjunto de etiquetas promovido por el proyecto *Universal Dependencies*, el cual tiene como fin promover un conjunto que pueda usarse para cualquier idioma: https://universaldependencies.org/

[[('Lo', 'lo', 'PP3MSA0'), ('sabe', 'saber', 'VMIP3S0'), ('todo', 'todo', 'PI0MS00'), ('absolutamente', 'absolutamente', 'RG'), ('todo', 'todo', 'PI0MS00'), ('.', '.', 'Fp')], [('Figuren', 'Figurar', 'VMM03P0'), ('se', 'se', 'PP3CN000'), ('lo', 'el', 'DA00S0'), ('tonto', 'tonto', 'NCMS000'), ('que', 'que', ' PR0CN00'), ('será', 'ser', 'VSIF3S0'), ('.', '.', 'Fp')]]

NAL. Análisis Sintáctico

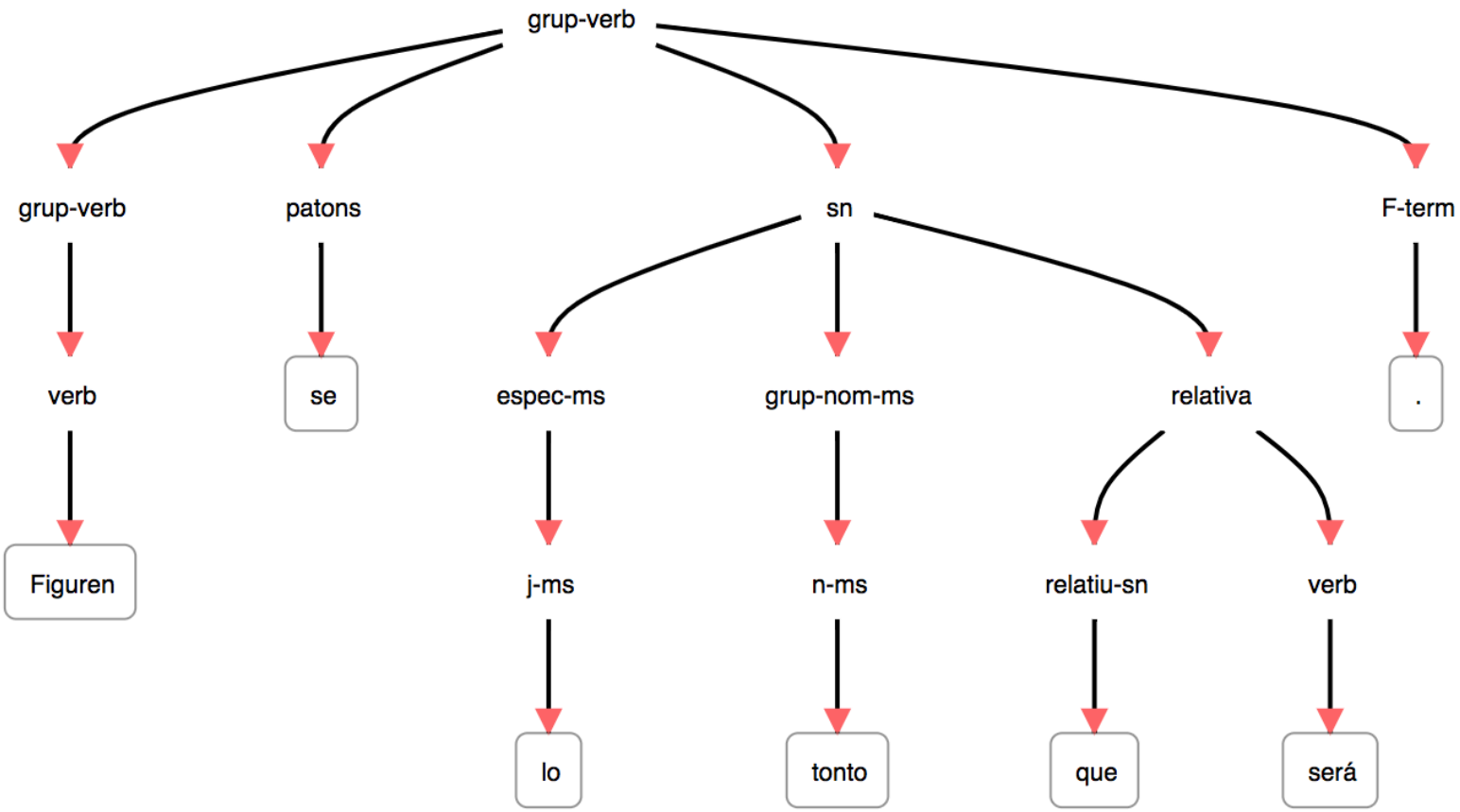
- Entrada: La salida del análisis léxico.

- Objetivo: Analizar cada una de las oraciones de la entrada para determinar su estructura de acuerdo a una gramática formal.

- Salida: Depende del analizador, pero generalmente el árbol sintáctico en el que se indica la estructura, y por extensión las relaciones existentes entre cada una de las palabras de la oración que se está analizando.

NAL. Análisis Sintáctico

Figuren	**se**	**lo**	**tonto**	**que**	**será**	**.**
figurar	se	el	tonto	que	ser	.
VMM03P0	*PP3CN00*	*DA00S0*	*NCMS000*	*PR0CN00*	*VSIF3S0*	*Fp*



NAL. Análisis Sintáctico

- La teoría relacionada con el Análisis Sintáctico está relacionada con la teoría de autómatas y la definición de gramáticas.
- Se asemeja al análisis sintáctico de un lenguaje de programación, pero:
  - Potencia de la gramática: Las gramáticas que determinan los lenguajes de programación suelen ser gramáticas independientes del contexto. El lenguaje natural requiere de gramáticas que tengan en cuenta el contexto e incluso de trabajar con dependencias de ámbito mayor.
  - Ambigüedad estructural: En el lenguaje natural puede darse el caso de que varias reglas gramaticales se puedan aplicar a una misma oración, situación que no sucede con los lenguajes de programación.
  - Ruido: Los elementos léxicos y estructuras sintácticas de los lenguajes de programación se encuentran bien definidas, mientras que el lenguaje natural puede presentar construcciones mal formadas.

NAL. Análisis Sintáctico

- Dependiendo de la información que proporcione un analizador sintáctico se identifican:
- *Shallow parsing*: Proporciona un análisis limitado de la estructura sintáctica, en el que como máximo se identifican las relaciones existentes entre las palabras de una oración.
- Análisis parcial: Se trata de un análisis superficial en el que se identifican elementos constituidos por varios términos como oraciones subordinadas. Un ejemplo es el *chunk parsing*.
- *Deep parsing*: Identifica dependencias entre los elementos de una oración e incluso las relaciones entre los sintagmas.
- Análisis completo: De manera similar al *deep parsing* trata de extraer la máxima información sintáctica de las oraciones.

NAL. Análisis Semántico

- Entrada: La información procedente del análisis sintáctico.
- Objetivo: Determinar el significado de una oración o de un enunciado.
- Salida: Representación conceptual del significado subyacente en el enunciado: identificación de los agentes implicados; de las acciones en las que están involucrados; las relaciones existentes entre ellos.
- Aplicación relacionada con la salida: La propia salida del análisis semántico puede llevar implícita la inferencia de nuevo conocimiento. Esta aplicación está estrechamente relacionada con la tarea de implicación textual.

NAL. Análisis Semántico

- No existe una especificación estándar para representar el significado de un enunciado. Dependiendo del enfoque que se siga, se representará una cantidad de información diferente.
- Tradicionalmente existen dos grandes enfoques:
  - Semántica léxica: Se centra en el estudio de las palabras y algunas combinaciones entre ellas.
  - Semántica supraléxica o composicional: Se centra en el análisis del significado de combinaciones más amplias de palabras como sintagmas, enunciados y oraciones completas. Se aplica un análisis abajo-arriba partiendo del significado de las palabras hasta llegar al significado de la composición de los términos.
- Estos están muy relacionados, de manera que actualmente se prefiere hablar de semántica lexicogramática, debido a que muchas combinaciones de términos están basadas en el significado individual de las palabras y en la construcción gramatical resultante.
- Otro problema está en determinar el grado de dependencia existente entre el significado de un enunciado y el grado de conocimiento que se tiene del mundo.

NAL. Análisis Semántico

En el ámbito del PLN el Análisis Semántico está relacionado con la resolución de la ambigüedad:

- Ambigüedad léxica: Determinación del significado de las palabras. Tipos:
  - Homonimia: Palabras que se escriben igual (homógrafas) o se pronuncian igual (homófonas) pero su significado depende del contexto.
  - Polisemia: El significado de una palabra depende del contexto en el que se esté empleando.
- Ambigüedad de ámbito: Identificación del ámbito de actuación de una palabra. Por ej.: las partículas negativas, las cuales modifican el significado de todas aquellas que pertenecen a su ámbito.
- Ambigüedad referencial: No se puede identificar claramente la palabra o concepto al que se hace referencia. Por ej.: El uso de pronombres para referenciar una entidad mencionada anteriormente. En PLN existe una tarea específica para resolver este problema: Resolución de la Anáfora.

APLICACIONES

Aplicaciones

Aplicaciones PLN

Clásicas

- Traducción automática
- Desambiguación
- Recuperación de información
- Resolución de preguntas
- Extracción de información
- Generación de LN

"Nuevas"

- Análisis de Opiniones
- Generación de resúmenes
- Simplificación de texto
- Minería de la argumentación
- Perfilado de usuarios
- Fake news

Aplicaciones

Aplicaciones PLN

Clásicas

- Traducción automática
- Desambiguación
- Recuperación de información
- Resolución de preguntas
- Extracción de información
- Generación de LN

"Nuevas"

- Análisis de Opiniones
- Generación de resúmenes
- Simplificación de texto
- Minería de la argumentación
- Perfilado de usuarios
- Fake news

Aplicaciones. Extracción de Información

- El objetivo es identificar las entidades de interés en una determinada aplicación.
- Cada entidad tiene una diversidad léxica enorme.
- Las entidades están relacionadas entre sí, dependiendo estas relaciones del contexto.

Aplicaciones. Extracción de Información

- Reconocimiento de entidades. Reconocer todas las entidades referenciadas en un documento.
- Dificultades:
  - Reconocimiento explícito e implícito de entidades: entidades nombradas, enfermedades, síntomas, fármacos, o referencias no canónicas de términos médicos.
    - Al paciente se le ha suministrado Metformina para combatir los niveles de glucosuria con los que ha ingresado.
  - Reconocimiento de entidades multipalabra:
    - El paciente ha ingresado con un trauma craneoencefálico en el Hospital 12 de Octubre.
    - Ayer me caí y me duele mucho la cabeza.

Aplicaciones. Extracción de Información

- Agrupamiento y clasificación de entidades.
  - Agrupar todas las referencias de una misma entidad.
  - Clasificar la categoría de interés en el dominio médico: fármaco, síntoma, enfermedad, paciente…
- Dificultades:
  - Agrupar expresiones médicas con dolencias descritas por personas legas en la materia:
    - Cefalea → Dolor de cabeza.
    - Neumonía → Pulmonía.
  - Correferencia y anáfora.
  - Asociar con documentación médica. Encontrar documentación interna o de referencia y asociar con el historial médico del paciente.

Aplicaciones. Extracción de Información

- Relación semántica entre entidades: Identificación de las relaciones semánticas entre las entidades identificadas.
- Dificultades: Existen distintos tipos de relaciones:
  - Relaciones entre conceptos: Es_parte_de; Es_propiedad_de.
  - Relaciones entre actores:
    - Sujeto: Quien realiza la acción: El asma afecta las vías respiratorias.
    - Objeto: Quien recibe la acción: Diagnosticar el cáncer es difícil.
  - Relaciones semánticas:
    - Temporal: Identificación de las referencias temporales.
    - Causa: Relaciones entre síntomas y diagnósticos; efectos de fármacos.

Aplicaciones. Reconocimiento de Entidades Nombradas

- Aplicación similar a la extracción de información.
- Se fija en la determinación de conjuntos de palabras que pueden ser una entidad en un determinado contexto.
- Clasificación del tipo de entidad.
- Ejemplo: OTAN TIDE Hackathon 2020.
- Objetivo: Identificación de nuevos Productos de Información (Information Products, IP) en los documentos de doctrina de la OTAN a partir de IP conocidos.
- UGR representó al Ejército de Tierra consiguiendo el primer puesto.
- Ejemplo.

ANÁLISIS DE OPINIONES

Aplicaciones

Aplicaciones PLN

Clásicas

- Traducción automática
- Desambiguación
- Recuperación de información
- Resolución de preguntas
- Extracción de información
- Generación de LN

"Nuevas"

- Análisis de Opiniones
- Generación de resúmenes
- Simplificación de texto
- Minería de la argumentación
- Perfilado de usarios
- Fake news

Análisis de Opiniones

Tratamiento computacional de la opinión, el sentimiento y la subjetividad en un texto [9].

Conjunto de técnicas computacionales para la extracción, clasificación, entendimiento y evaluación de opiniones expresadas en fuentes de noticias, comentarios en redes sociales y otros contenidos generados por usuarios [10].

[9] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.

[10] Cambria, E., & Hussain, A. (2012). Sentic computing. *marketing*, *59*(2), 557-577.

Análisis de Opiniones

- Objeto de estudio: **la opinión**

- ¿Qué es una opinión? La evidencia o expresión de un estado de ánimo [11].

- Las opiniones, las emociones y las evaluaciones son expresiones de estados de ánimo.

- [11] Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). A Comprehensive Grammar of the English Language. Longman, London.

Análisis de Opiniones

Autor: Pedro Reyes Fecha: 20/11/2019

La semana anterior pasé unos días en Jaén con mi esposa. Elegimos una habitación bonita en el centro de la ciudad. Nos ha quedado un buen recuerdo del personal del hotel. Debo resaltar que la cama era muy confortable. Por el contrario, mi mujer no pudo dormir porque, según ella, la almohada era muy mala. No nos gustó el baño porque encontramos algunos desagradables pelos cuando nosotros entramos en la habitación. Para terminar, debo decir que lo mejor del hotel fue el desayuno, un día más y reviento.

Análisis de Opiniones

Autor: Pedro Reyes

Fecha: 20/11/2019

La semana anterior pasé unos días en Jaén con mi esposa. Elegimos una habitación bonita en el centro de la ciudad. Nos ha quedado un buen recuerdo del personal del hotel. Debo resaltar que la cama era muy confortable. Por el contrario, mi mujer no pudo dormir porque, según ella, la almohada era muy mala. No nos gustó el baño porque encontramos algunos desagradables pelos cuando nosotros entramos en la habitación. Para terminar, debo decir que lo mejor del hotel fue el desayuno, un día más y reviento

Análisis de Opiniones

Autor: Pedro Reyes

Fecha: 20/11/2019

La semana anterior pasé unos días en Jaén con mi esposa. Elegimos una habitación bonita en el centro de la ciudad. Nos ha quedado un buen recuerdo del personal del hotel. Debo resaltar que la cama era muy confortable. Por el contrario, mi mujer no pudo dormir porque, según ella, la almohada era muy mala. No nos gustó el baño porque encontramos algunos desagradables pelos cuando nosotros entramos en la habitación. Para terminar, debo decir que lo mejor del hotel fue el desayuno, un día más y reviento

Análisis de Opiniones

Autor: Pedro Reyes

Fecha: 20/11/2019

La semana anterior pasé unos días en Jaén con mi esposa. Elegimos una habitación bonita en el centro de la ciudad. Nos ha quedado un buen recuerdo del personal del hotel. Debo resaltar que la cama era muy confortable. Por el contrario, mi mujer no pudo dormir porque, según ella, la almohada era muy mala. No nos gustó el baño porque encontramos algunos desagradables pelos cuando nosotros entramos en la habitación. Para terminar, debo decir que lo mejor del hotel fue el desayuno, un día más y reviento

Análisis de Opiniones

Autor: Pedro Reyes

Fecha: 20/11/2019

La semana anterior pasé unos días en Jaén con mi esposa. Elegimos una habitación bonita en el centro de la ciudad. Nos ha quedado un buen recuerdo del personal del hotel. Debo resaltar que la cama era muy confortable. Por el contrario, mi mujer no pudo dormir porque, según ella, la almohada era muy mala. No nos gustó el baño porque encontramos algunos desagradables pelos cuando nosotros entramos en la habitación. Para terminar, debo decir que lo mejor del hotel fue el desayuno, un día más y reviento

$$(e_i, a_{ij}, p_{ijhl}, h_h, t_l)$$

[12] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Análisis de Opiniones

Tareas propias del AO:

1. Extracción y categorización de entidades.
2. Extracción y categorización de aspectos.
3. Extracción del opinador.
4. Extracción del momento en el que tiene lugar la opinión.
5. Clasificación de la polaridad de la opinión: Opiniones explícitas e implícitas.
6. Generación de la quíntupla de la opinión.

Análisis de Opiniones

(hotel, estilo, positivo, Pedro Reyes y esposa, 20/11/2019)

(hotel, personal, positivo, Pedro Reyes y esposa, 20/11/2019)

(hotel, cama, positivo, Pedro Reyes, 20/11/2019)

(hotel, almohada, negativo, esposa de Pedro Reyes, 20/11/2019)

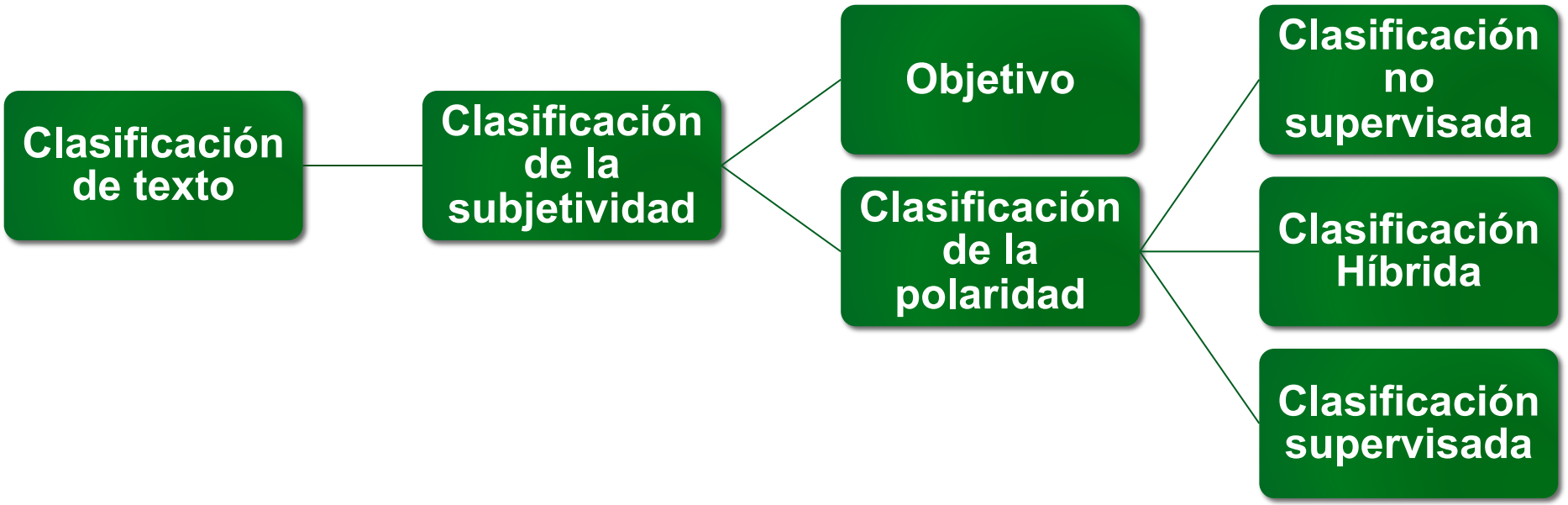
(hotel, desayuno, positivo, Pedro Reyes, 20/11/2019)

Análisis de Opiniones

Tareas propias del AO

1. Extracción y categorización de entidades.
2. Extracción y categorización de aspectos.
3. Extracción del opinador.
4. Extracción del momento en el que tiene lugar la opinión.
5. **Clasificación de la polaridad de la opinión: Opiniones explícitas e implícitas.**
6. Generación de la quíntupla de la opinión.

Análisis de Opiniones



Análisis de Opiniones

Clasificación

- Supervisada:
  - Requiere de un corpus/conjunto de datos anotado.
  - Principalmente basado en el uso de algoritmos de aprendizaje automático: SVM, KNN, Naïve Bayes, Random Forest, Neural Networks, Logistic Regression…
  - Tendencia actual: Redes neuronales (*Deep Learning*).
- No supervisada:
  - No requiere de un corpus/conjunto de datos anotado.
  - Métodos: Lista de palabras de opinión, patrones sintácticos, *clustering*.
- Híbrida:
  - Combinación de los métodos anteriores. Por ejemplo:
    - Sistema no supervisado para anotar un pequeño conjunto de elementos con el que construir un clasificador supervisado.
  - Multi-clasificador de sistema de clasificación supervisado o no supervisado.

AO. Niveles de Análisis de Opinión

Niveles de análisis

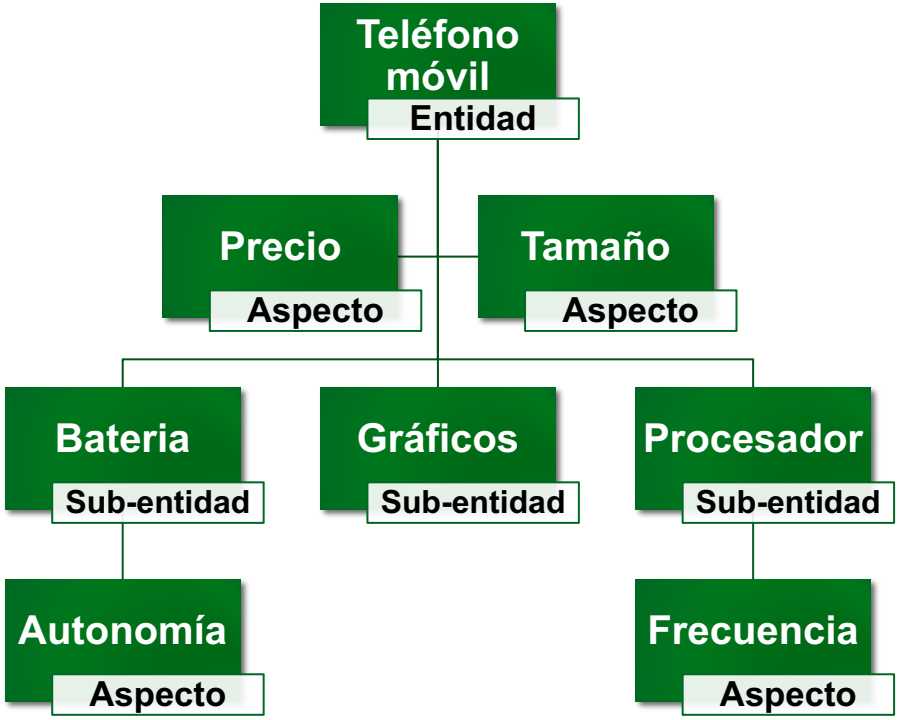
- Documento
- Oración
- Entidad/Aspecto

AO. Niveles de Análisis de Opinión

- A nivel de documento:
  - Objetivo: La clasificación de la orientación de la opinión general expresada en un documento.
  - Entrada: Un documento (largo / corto (tuit)).
  - Asunción: El documento está formado por opiniones explícitas o implícitas.
  - Salida: La opinión del documento.
- A nivel de oración:
  - Objetivo: La clasificación de la opinión expresada en una oración.
  - Entrada: Un documento / una oración.
  - Asunción: El documento está formado por opiniones explícitas o implícitas.
  - Salida:
    - Documento: Combinación de la orientación de las opinión de las oraciones que componen el documento.
    - Oración: La orientación de la opinión expresada en la oración.

AO. Niveles de Análisis de Opinión

- A nivel de aspecto:
  - Objetivo: La clasificación de la opinión sobre una entidad o aspecto.
  - Entrada: Una oración/un conjunto de oraciones/un documento.
  - Asunción: Las entidades/aspectos están dados.
  - Salida:
    - El conjunto de entidades y aspectos de las entidades.
    - La polaridad de la opinión sobre cada entidad y aspecto.



AO. Corpora

Textos largos

- Customer Review Dataset (2004)
  - Documentos: 500.
  - Anotación: A nivel de documento.
  - Dominio: Productos electrónicos (5).
- SFU Review Corpus (2004)
  - Documentos : 400.
  - Anotación: A nivel de documento.
  - Dominio: Opiniones de productos (8).
  - Versión en español: 2009.
- MPQA (2005):
  - Documentos: 535.
  - Anotación: A nivel de oración.
  - Dominio: Noticias.
  - Última versión: 2015.

Textos cortos

- Stanford Twitter Corpus (2009)
  - Documentos:
    - Entrenamiento: 1.600.000 (balanceado). Supervisión distante (*distant supervision*).
    - Test: 182 positivas 177 negativas. Anotado manualmente.
  - Anotación: A nivel de documento.
  - Dominio: General.
  - Idioma: Inglés.
- SemEval Twitter corpus:
  - Desde 2013, la organización publica un nuevo corpus.
  - Idioma: Inglés.
- General Corpus of TASS
  - Documentos:
    - Entrenamiento : 7.219.
    - Test: 60.798.
  - Idioma: Español.
  - Anotación: A nivel de documento. 6 clases.
- InterTASS (2017; 2018, 2019)
  - Documentos:
    - Entrenamiento: 1.008.
    - Desarrollo: 506.
    - Test: 1.899.
  - Idioma: Español.
  - Anotación: A nivel de documento. 4 clases.

AO. Recursos

- **Listas de palabras de opinión**: Listas de palabras positivas y negativas
  - Lista de Bing Liu:
    - Idioma: Inglés
    - Palabras: sobre 6.800 palabras
  - iSOL:
    - Idioma: Español
    - Palabras: 2.509 positivas; 5.626 negativas
  - GermanPolarityWords
    - Idioma: Alemán
    - Palabras: 10.000

AO. Recursos

- Lexicones: Conjunto de palabras relacionadas por propiedades léxicas
  - SentiWordNet
    - Basado en WordNet.
    - Cada significado/sentido (*synet*) tiene asociado tres valores de polaridad: positivo, negativo y neutro.
    - La suma de los tres valores es igual a 1.
  - Q-WordNet
    - Basado en WordNet.
    - Cada significado/sentido (*synet*) tiene asociado dos valores de polaridad: positivo y negativo.

AO. Retos

- Extracción de información.
- Extracción de relaciones semánticas.
- Resolución de la co-referencia y anáfora.
- Desambiguación.
- Tratamiento de la negación.
- Conocimiento implícito.
- Adaptación al dominio.
- Tratamiento de ironía y sarcasmo.

AO en Twitter

- La investigación en AO en Twitter avanzó en paralelo al aumento del uso de *microblogs*.

- Los primeros estudios estuvieron más relacionados con la sociología que con la informática.

- Boca a boca → Boca a boca electrónico o en línea (EWOM / OWOM).

[12] Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L. A. & Montejo-Ráez A. (2014) Sentiment Analysis in Twitter. Journal of Natural Language Engineering 20(1):1-28.

AO en Twitter

Textos largos

- La longitud no está limitada.
- Compuestos por varias oraciones.
- Pueden expresar varios mensajes o ideas.
- El estilo formal es el más común.
- Suele haber un uso correcto de la gramática.
- Hay contexto suficiente para determinar la información del discurso.

Textos cortos

- La longitud es limitada (tuits: 280 caracteres).
- Normalmente compuesto por un máximo de 3 oraciones.
- Usualmente sólo se expresa una oración.
- Prevalece el estilo informal.
- Uso pobre de la gramática.
- Dispersión de los datos. Hay que buscar redundancia con técnicas lingüísticas (implicación textual).
- No suele haber contexto.

AO en Twitter

- Primer trabajo sobre AO en Twitter [14].
- Corpus de tuits escritos en inglés:
  - Entrenamiento: 1.600.000 (800.000 positivos y 800.000 negativos).
    - Supervisión imprecisa: :) (positivo) :((negativo) de acuerdo a (Read, 2005).
  - Test: 182 positivos y 177 negativos. Etiquetado manual.
  - Algoritmo: SVM, Naïve Bayes, Máxima entropía.
  - Características: Unigramas, bigramas, unigramas+bigramas, pos-tags (categoría morfológica)
  - Mejor configuración:
    - Algoritmo: Máxima entropía.
    - Características: unigrams-bigrams.

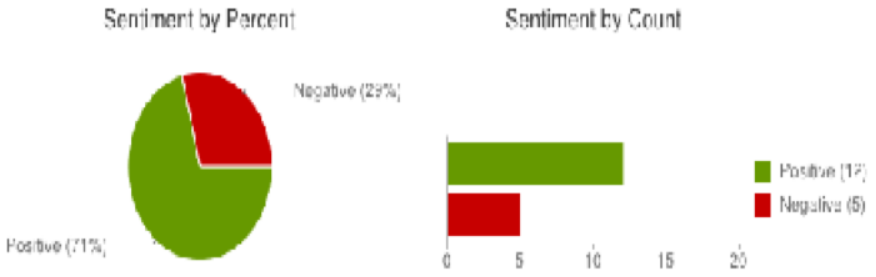
[14] Go, A., Bhayani, R and Huang, Lei (2009). Twitter Sentiment Classification using Distant Supervision. CS224N Project Report, Stanford, 1, 12

Sentiment140

Tweet 601 | Me gusta | +1 157

Cocacola | Spanish | Search

Sentiment analysis for Cocacola



Tweets about: Cocacola

MQMtumblr: Photoset: Cuando tu madre pide una **cocacola**. http://t.co/5lXQgHQ8r2
Posted: 1 minute ago

MasterSMMAST: RT @ChuscSoria: **@CocaCola_**es crea nuevas formas de atención al cliente, Consumer Care. conversar con **cocacola** 2.0 sea una **sensación social** ?
Posted: 3 minutes ago

RosiFernandez7: MI papi es malo. Siempre me deja con intriga, y aún por encima hoy, que le compré una lata de **CocaCola** con su nombre
Posted: 5 minutes ago

Beerto10: RT @raullja: Comida favorita y bebida favorita ? Patatas fritas con filete y **Cocacola** http://t.co/5iu3Df15PW
Posted: 5 minutes ago

ijuliansuaza: Le metian perico hasta a la **cocacola** (8)
Posted: 5 minutes ago

The results for this query are: Accurate | Inaccurate

AO. Negación

- ¿Afecta la negación a la clasificación de la opinión?

- Objetivo: Determinar si la identificación del ámbito de la negación es beneficioso para la clasificación de la opinión en español.

[15]Jiménez-Zafra, S., Martín-Valdivia, M. T., Martínez-Cámara, E., Ureña-López, L. A. (2019). Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter. IEEE Transactions on Affective Computing.

AO. Negación.

Metodología:

- Corpus: General Corpus de TASS
  - Entrenamiento: 7.219.
  - Test: 60.798.
  - Clases: P+, P, NEU, N, N+, NONE.
  - Uso:
    - Sólo se usa el conjunto de test; se descarta la clase NONE; se agrupan las clases P+ y P, y N y N+.

P	22.233
NEU	1.305
N	15.844

AO. Negación.

Metodología:

- Sistema: Clasificación no supervisada basada en el uso de:
  - La lista de palabras de opinión iSOL.
  - Lista de *hashtags* anotados a nivel de opinión.
  - Anotación de la opinión/emoción de emoticonos.
  - Identificación del ámbito de partículas negativas.
  - Clasificación basada en reglas.

$$opinión(tuit) = \begin{cases} P & si\ pv > nv \\ NEU & si\ pv = nv \\ N & si\ pv < nv \end{cases}$$

AO. Negación.

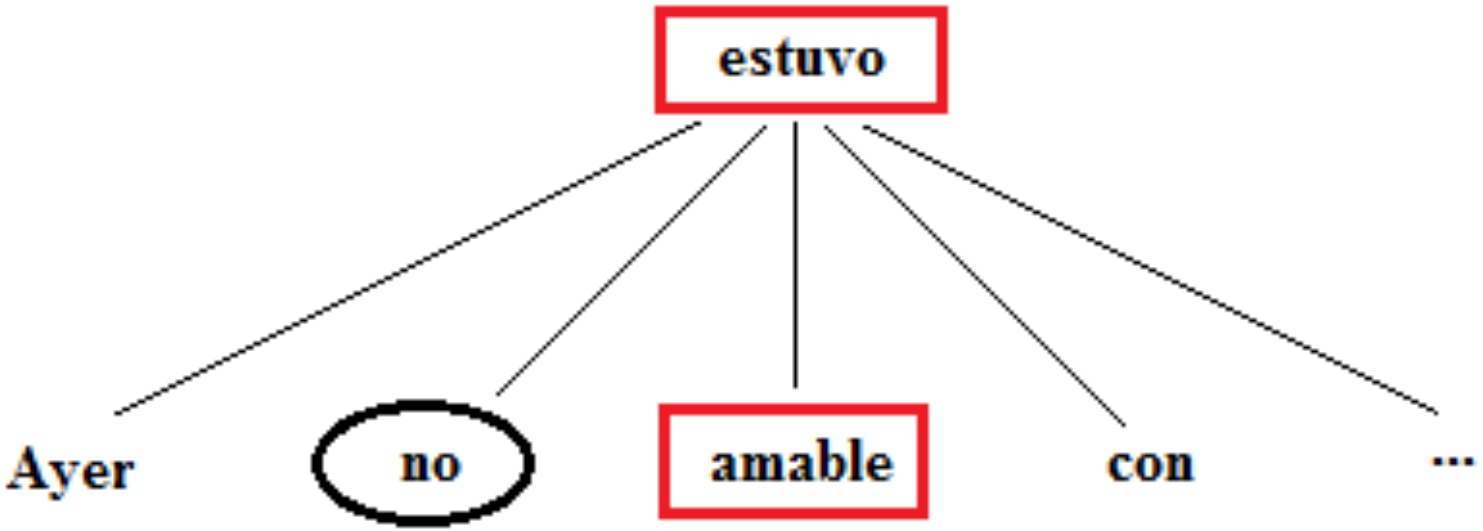
Identificación de la opinión:

- Partículas negativas usadas en español: no, tampoco, nadie, jamás, ni, sin, nada, nunca y ninguno.
- Tres grupos de partículas negativas según su construcción sintáctica.
- Cálculo árbol de dependencias sintácticas. Uso de Freeling.

Partículas Negativas	Regla
No, tampoco, nadie, jamás, ninguno	Nodo padre y todo el árbol formado por el hermano derecho
Ni, sin	Nodos hijos y todos sus descendientes hasta el nodo hoja
Nada, nunca	Nodo padre

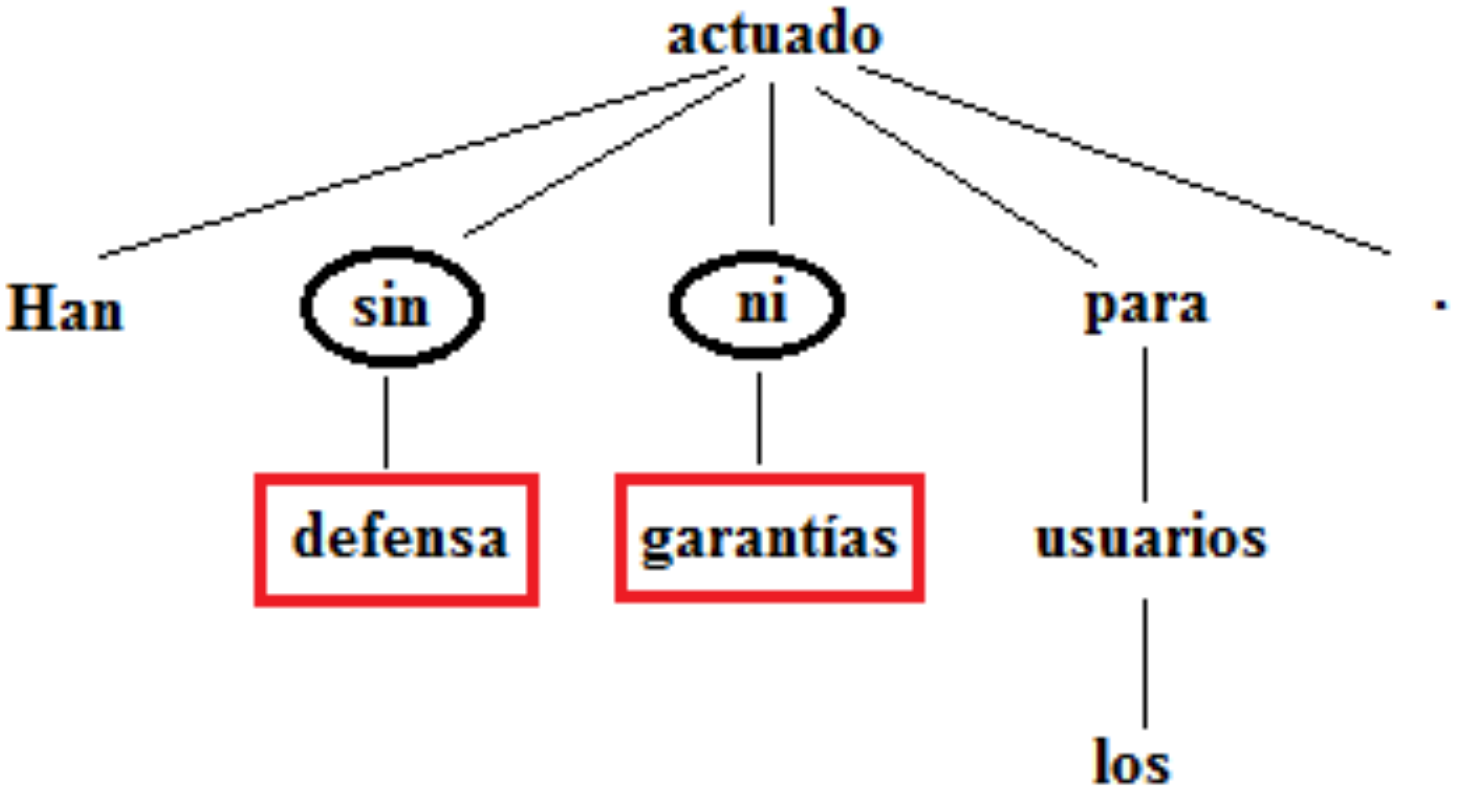
AO. Negación.

Grupo partícula NO.



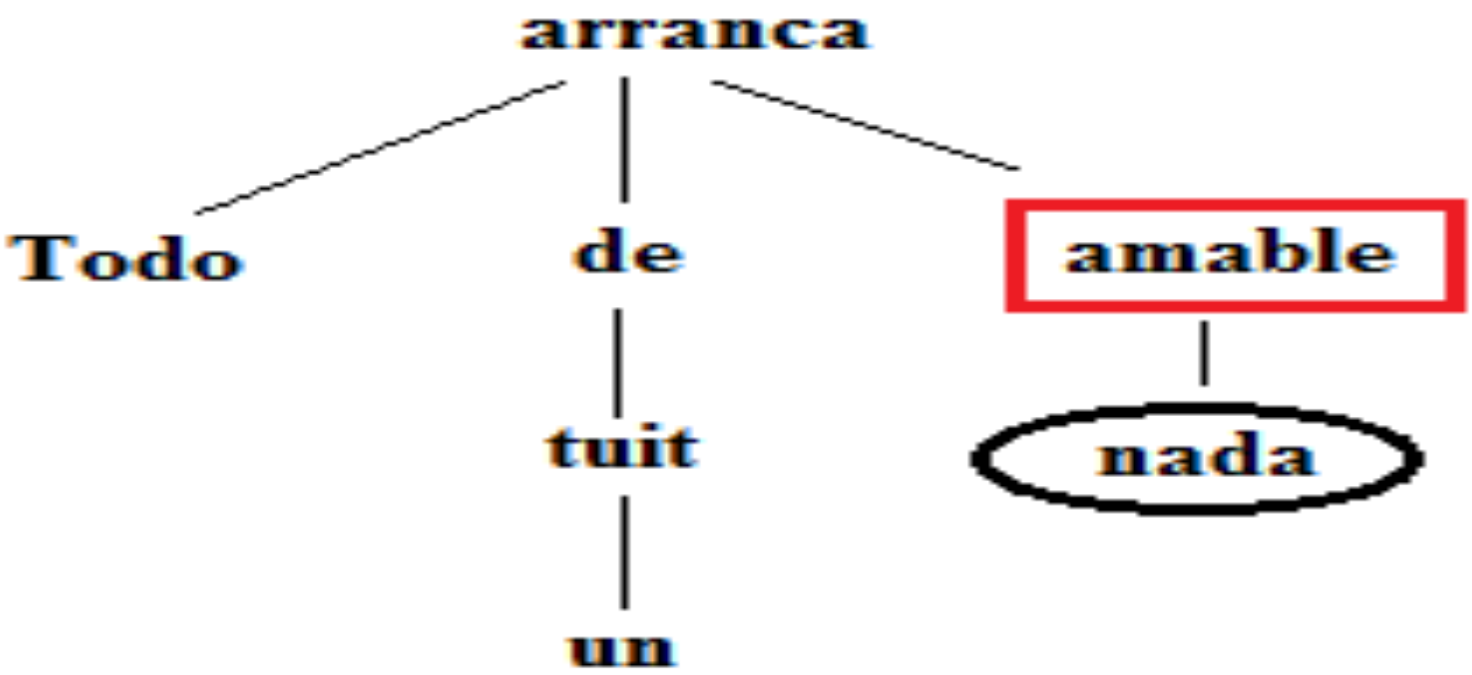
AO. Negación.

Grupo partícula NI.



AO. Negación.

Grupo partícula NADA.



AO. Negación.

Todo arranca de un tweet nada amaaable. #maldad =(

Tokenización

[Todo] [arranca] [de] [un] [tweet] [nada] [amaaable][.] [#maldad] [=(]

Normalización

[Todo] [arranca] [de] [un] [tweet] [nada] [amable][.] [#maldad] [=(]

Part-of-Speech Tagging y Lematización

[Todo] [arrancar] [de] [un] [tweet] [nada] [amable][.] [#maldad] [=(]

AO. Negación.

Todo arranca de un tweet nada amaaable. #maldad =(

Negación

[Todo] [arrancar] [de] [un] [tweet] **[nada] [amable]**[.] [#maldad] [=(]

Clasificación de la opinión

[Todo] [arrancar](-1) [de] [un] [tweet] {**[nada] [amable]** (+2)}(-2)[.] [#maldad] (-2)[=(] (-2)

Nv = 7

Pv = 0

Opinión = N

AO. Negación.

Casos base:

- No considerar la negación (BS)
- Considerar como ámbito de la negación todas las palabras desde la partícula hasta un signo de puntuación (BSN).

Evaluaciones:

- Conjunto completo de datos (Total).
- Tuits con partículas negativas (NegCue).
- Tuits con partículas negativas y términos de opinión (RuleAffect).

AO. Negación.

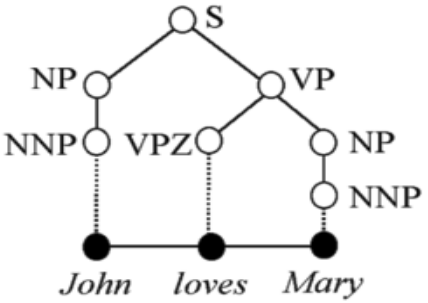
Corpus	Sistema	Macro-P	Macro-R	Macro-F1	Accuracy
	BS	0,5764	0,5235	0,5486	0,6258
Total	BSN	0,5705	0,5190	0,5435	0,6205
	Propuesta	0,5810	0,5296	0,5541	0,6308

Corpus	Sistema	Macro-P	Macro-R	Macro-F1	Accuracy
	BS	0,4861	0,4702	0,4780	0,4866
NegCue	BSN	0,4621	0,4514	0,4567	0,4622
	Propuesta	0,5060	0,4936	0,4997	0,5092

Corpus	Sistema	Macro-P	Macro-R	Macro-F1	Accuracy
	BS	0,3971	0,3949	0,3960	0,4463
RuleAffect	BSN	0,4431	0,4545	0,4487	0,5026
	Propuesta	0,4660	0,4792	0,4725	0,5292

AO. Clasificación de la Opinión

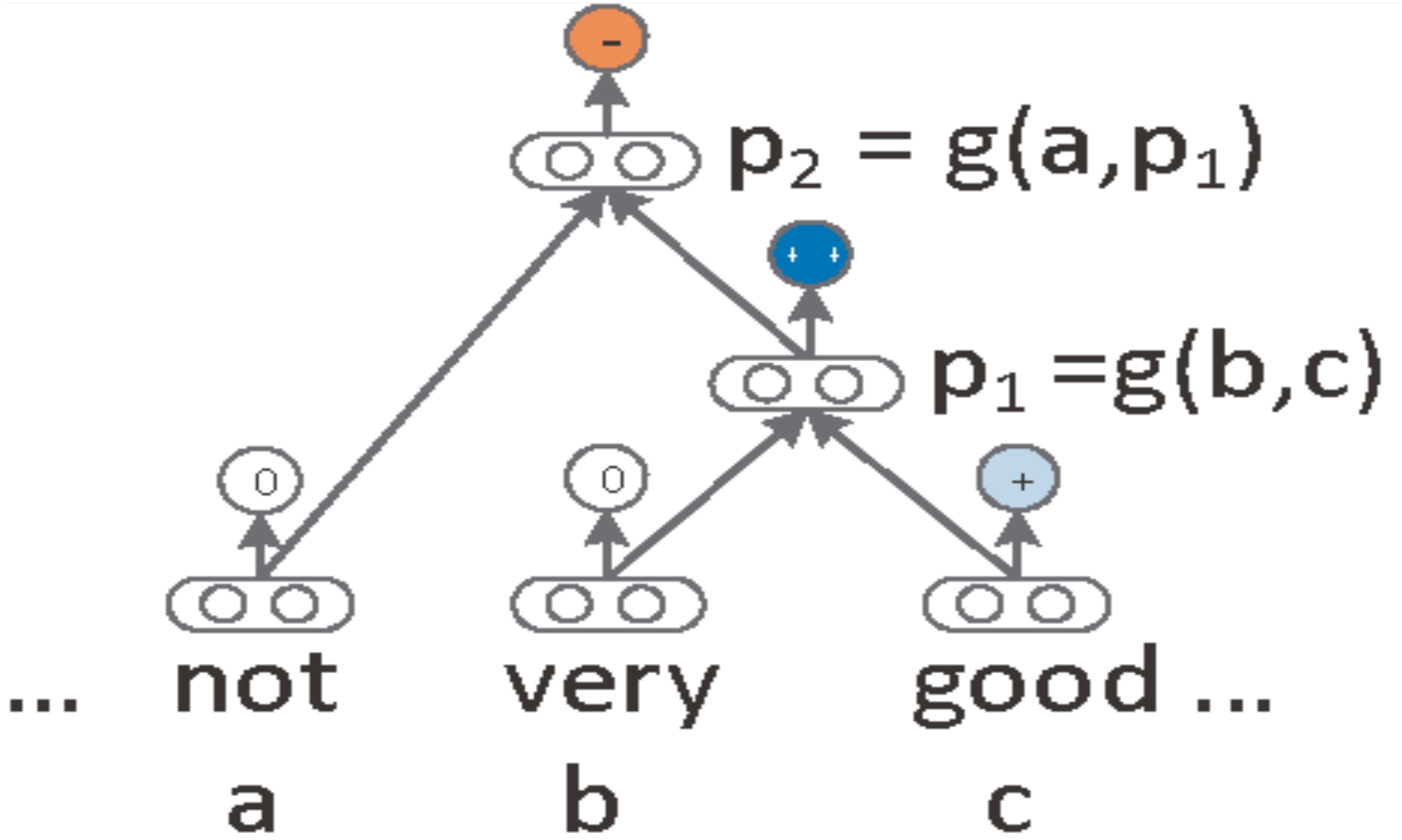
- Las redes neuronales (RN) representan ya el estado del arte en AO.
- El artículo de Socher et al. 2013 presenta una de las primeras. aplicaciones con éxito de las RN en AO.
- Los autores argumentan que la opinión de un documento no sólo depende de la polaridad individual de las palabras, sino del significado global de todas ellas → Carácter composicional del lenguaje.
- Generación del primer Treebank de opiniones: Sentiment Treebank.
- Un Treebank es un corpus en el que se encuentra anotada la estructura sintáctica de las oraciones y la categoría morfológica de las palabras.
- Un Treebank de opiniones es un Treebank donde además están anotadas la polaridad de los sintagmas.



AO. Clasificación de la Opinión

- Se propone una RN que clasifica la opinión de cada palabra.
- La RN es recursiva y recorre el árbol sintáctico de una oración.
- En cada nodo hoja, la RN calcula la polaridad del nodo del árbol teniendo en cuenta las palabras que dependen del nodo y del resultado de la anterior iteración.

AO. Clasificación de la Opinión



AO. Clasificación de la Opinión

- Los autores proponen el modelo Tensor Neuronal Recurrente (*Recurrent Neural Tensor*).

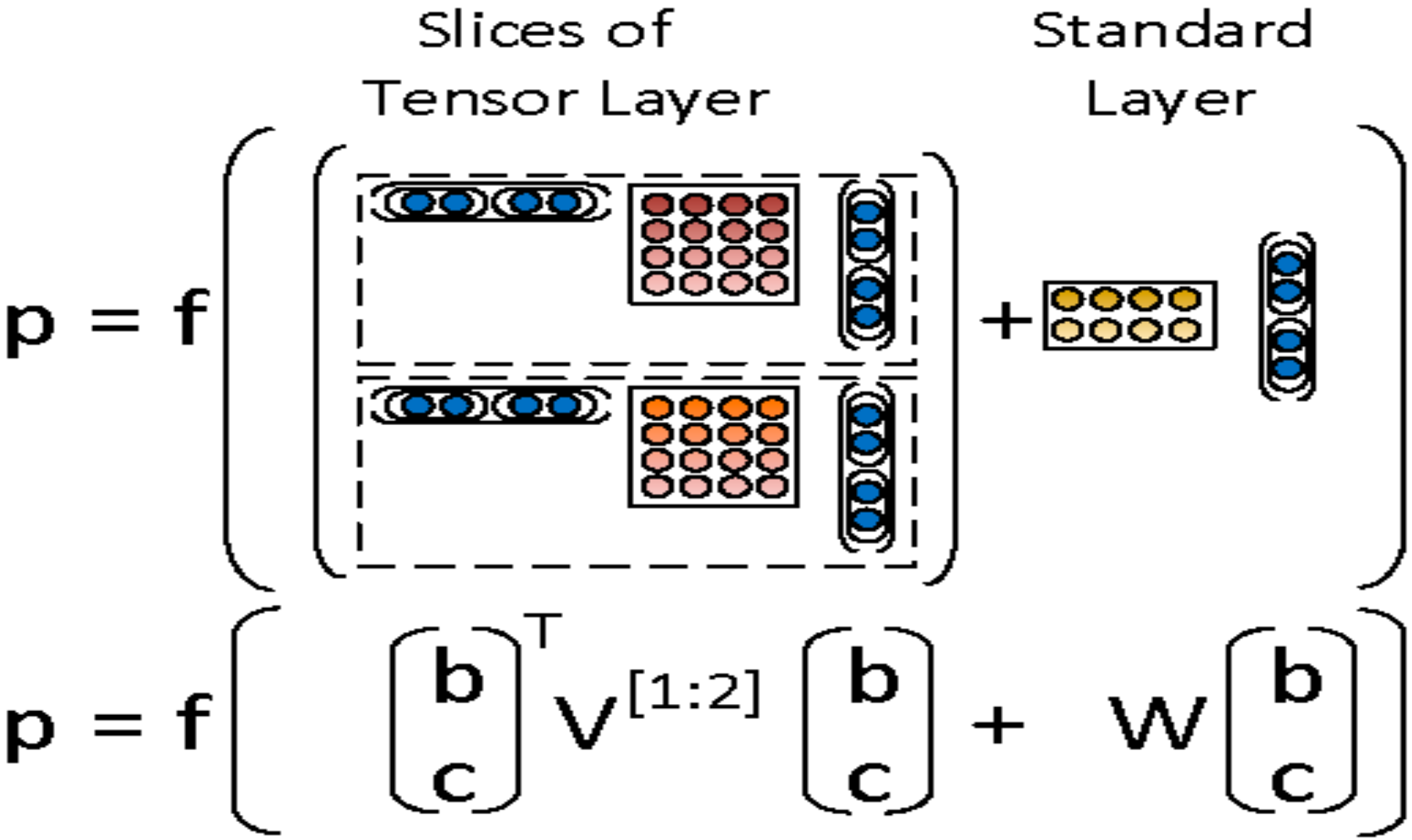
- La idea principal es el aprendizaje automático de la función de composición del significado de varias palabras.

$$h = \begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} b \\ c \end{bmatrix}$$

$$h_i = \begin{bmatrix} b \\ c \end{bmatrix}^T V^{[i]} \begin{bmatrix} b \\ c \end{bmatrix}$$

$$V^{[1:d]} \in \mathbb{R}^{2d \times 2d \times d}$$

AO. Clasificación de la Opinión



AO. Clasificación de la Opinión

- En el corpus Movie Review (Pang & Lee 2005) realizan varios experimentos:
  - Clasificación 5 clases: 80,7% Accuracy (No exp. Anterior).
  - Clasificación binaria: 85,4% Accuary. Mejoran Est. Arte ≈ 6,75%.
  - Opinión en oraciones positivas negadas: 71,4% de Accuary.
  - Opinión en oraciones negativas negadas: 81,8% de Accuary.

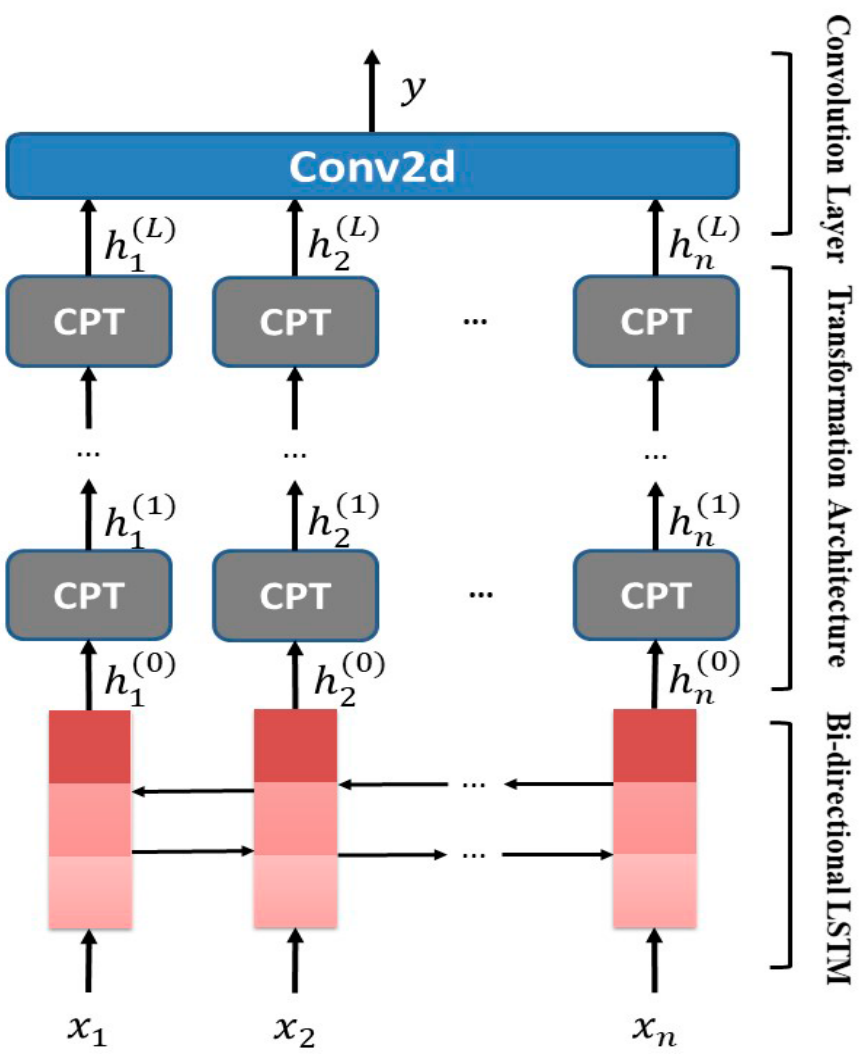
- [16] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

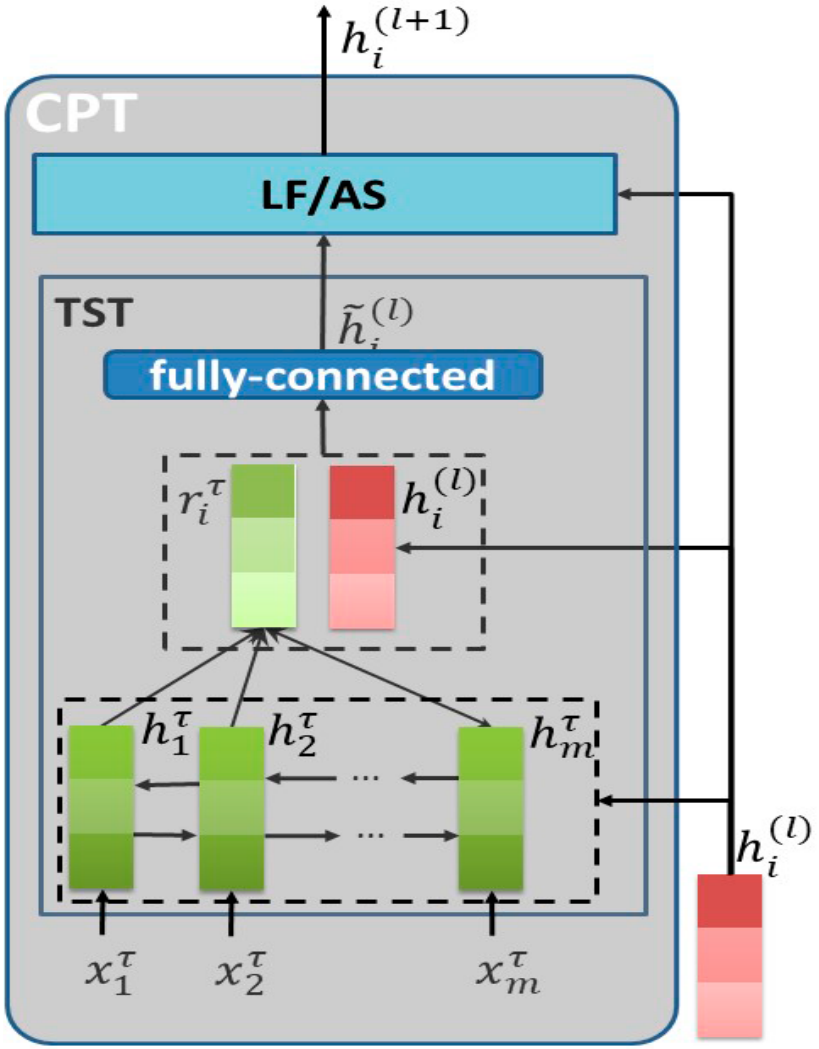
- Más información en: https://nlp.stanford.edu/sentiment/

AO. Clasificación Opinión a nivel de Aspecto

- La tendencia actual es la de aumentar la codificación de la información del contexto.
- Muy aconsejable para el AO a nivel de aspecto.
- [17] Li, X., Bing, L., Lam, W., & Shi, B. (2018). Transformation Networks for Target-Oriented Sentiment Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 946-956)
- Corpus: SemEval 2014
- Dominios:
  - Restaurantes: *All the **appetizers** and **salads** were **fabulous**, the **steak** was mouth watering and the **pasta** was delicious!!!*
  - Portátiles: *From the **build quality** to the **performance**, everything about it has been sub-par from what I would have expected from Apple*.

AO. Clasificación Opinión a nivel de Aspecto





AO. Clasificación Opinión a nivel de Aspecto

Sentence	BILSTM-ATT-G	RAM	TNet-LF	TNet-AS
1. Air has higher [resolution]$_P$ but the [fonts]$_N$ are small .	($N^✗$, N)	($N^✗$, N)	(P, N)	(P, N)
2. Great [food]$_P$ but the [service]$_N$ is dreadful .	(P, N)	(P, N)	(P, N)	(P, N)
3. Sure it ' s not light and slim but the [features]$_P$ make up for it 100% .	$N^✗$	$N^✗$	P	P
4. Not only did they have amazing , [sandwiches]$_P$, [soup]$_P$, [pizza]$_P$ etc , but their [homemade sorbets]$_P$ are out of this world !	(P, $O^✗$, $O^✗$, P)	(P, P, $O^✗$, P)	(P, P, P, P)	(P, P, P, P)
5. [startup times]$_N$ are incredibly long : over two minutes .	$P^✗$	$P^✗$	N	N
6. I am pleased with the fast [log on]$_P$, speedy [wifi connection]$_P$ and the long [battery life]$_P$ (> 6 hrs) .	(P, P, P)	(P, P, P)	(P, P, P)	(P, P, P)
7. The [staff]$_N$ should be a bit more friendly .	$P^✗$	$P^✗$	$P^✗$	$P^✗$

CASO PRÁCTICO

Caso práctico

- Problema: Clasificación de opiniones a nivel de tuit en español.
- Conjunto de datos: Corpus General de TASS.
- Entrenamiento: 7.218.
- Número de clases: 4. P, NEU, N, NONE
- Lenguaje de programación: Python 3.
- Librerías PLN: NLTK 3.0.
- Librería de aprendizaje automático: Scikit-learn.
- Se evaluará los algoritmos Naïve Bayes y SVM con representación basada en unigramas pesados por TF-IDF.

TASS
Taller de Análisis Semántico en la SEPLN