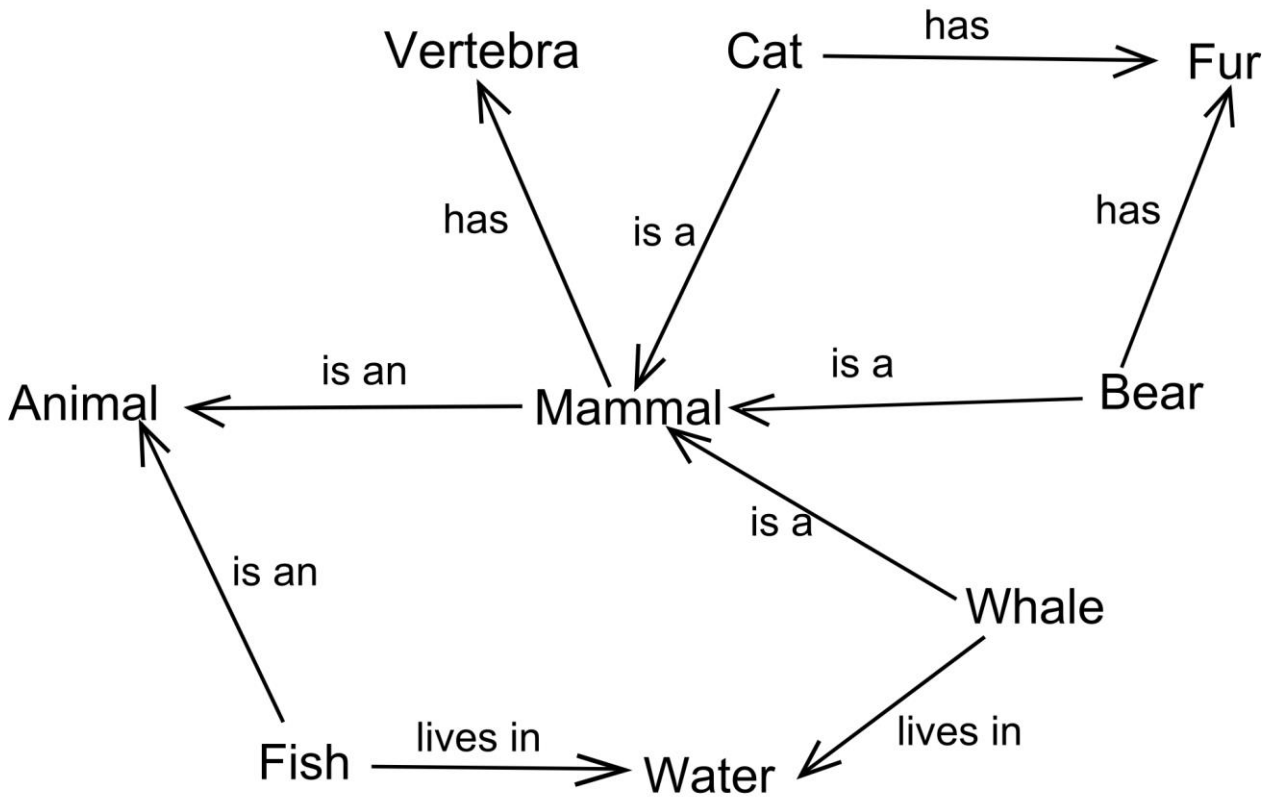
Representación del Conocimiento y Razonamiento

Redes Semánticas

Una Red Semántica o Red de Marcos es un modelo conceptual que se incluye en los métodos de representación del conocimiento (Knowledge Representation - KR) asociados a IA. Consiste en una red que representa relaciones semánticas entre conceptos. La representación es un grafo, generalmente dirigido en el que los nodos representan conceptos y los arcos relaciones entre los nodos.
Las redes semánticas pueden interpretarse también como una representación lógica, lo que se puede interpretar observando que cada **relación** (rama) entre dos conceptos (nodos) se puede expresar como una sentencia lógica, y a su vez una o varias de estas relaciones pueden ser vistas como antecedentes que soportan a una tercera (hipótesis). En definitiva representan conocimiento a través de conceptos relacionados entre sí. Los arcos y nodos permiten crear una jerarquía de conceptos, que en realidad no solo describe una taxonomía sino que más bien describe una ontología. Estos modelos pueden luego codificarse en algún lenguaje o base de datos, preferentemente orientada a Marcos y Ranuras.



Ranuras y Marcos

El método posiblemente más difundido para almacenar conocimiento es conocido como Ranuras y Marcos (Frames & Slots). Un **marco** (frame) es una estructura de datos empleada para dividir el conocimiento en subestructuras que representan situaciones estereotipadas (tomadas como modelos simplificados). Se las considera derivadas de las Redes Semánticas y brindan un camino para codificar a éstas. Ensamblan hechos sobre un objeto particular enmarcándolo en una gran jerarquía taxonómica análogamente a la taxonomía biológica.

Estructura de un Marco: Un marco contiene información acerca de cómo usarlo, que esperar como información asociada, y que hacer si esa expectativa no se cumple. Alguna información en el marco es estable mientras que otra (terminales) puede ir variando. Diferentes marcos pueden compartir algunos terminales.
Cada pieza de información sobre un marco particular se almacena en una ranura. La información podría contener:

Hechos datos y valores.
Procedimientos (Anexos procedimentales).
Valores por default para datos y procedimientos.
Otros marcos y submarcos.

Se suele considerar una ventaja el hecho de que los terminales de un marco se completan con valores por defecto. Esto se basa en el modo de trabajo de la mente biológica, en el sentido de que si una persona oye _el chico pateó la pelota_ las personas imaginan una pelota específica y no una entidad abstracta general sin atributos. Otra fortaleza de este modo de representar el conocimiento es que a diferencia de las Redes Semánticas, **permiten excepciones** en ciertos casos particulares. Esto le da a los marcos una flexibilidad que permite una más precisa representación de los fenómenos del mundo. Los lenguajes de Marcos incluyen la función de herencia. Finalmente es posible generar a partir de los marcos una Red Semántica, aun faltando algún vinculo (arco).

Recursos lexicográficos creados por carga manual de Frames

Si bien existen una variedad de dispositivos de almacenamiento de conocimiento con diferentes posibilidades a los efectos de fundamentar temas de NLP (Procesamiento del Lenguaje Natural) nos referiremos aquí en particular al almacenamiento de **conocimiento lexicográfico**.
En ese sentido, varios diccionarios existentes para uso en NLP han sido creados manualmente, destacándose WordNet, FrameNet y VerbNet.

WordNet, es un diccionario (base de datos léxica) para **lengua inglesa** orientado a NLP, empleado por ejemplo para el desarrollo del sistema Watson. Este diccionario agrupa palabras en inglés en conjuntos de sinónimos llamados synsets (precisamente por conjuntos de sinónimos). Para cada una de estas palabras (conceptos) sinónimos, se encuentran definidas un conjunto de variables y sus valores. Esto permite proveer definiciones cortas y generales, además de registros de relaciones semánticas entre dichos synsets.

VerbNet es una base léxica que mapea verbos en sus clases correspondientes, e incluye información **sintáctica y semántica** de los verbos, tal como las secuencias sintácticas del Marco y restricciones a los argumentos.

VerbNet es el mayor diccionario on-line disponible en Inglés. Es una base jerárquica independiente del dominio, un diccionario con mapeos a otros recursos semejantes como WordNet, FrameNet y otros.
VerbNet está organizado en clases de verbos que a través del refinamiento y la adición de subclases alcanza coherencia sintáctica y semántica entre los miembros de una clase. Cada clase verbal es completamente descripta por roles temáticos, restricciones sobre argumentos, y marcos consistentes en una descripción de predicados sintácticos y semánticos.
Cada clase contiene un conjunto de descripciones sintácticas, o marcos sintácticos, mostrando el posible ordenamiento de la estructura sintáctica superficial, tales como transitivo, intransitivo, oraciones preposicionales y otras alternativas.
Las restricciones semánticas se usan para restringir los tipos de los roles temáticos seguidos por los argumentos, y se pueden añadir restricciones adicionales para indicar la naturaleza sintáctica del constituyente.
Cada marco está asociado con información semántica explícita, expresada como una conjunción de predicados semánticos booleanos tales como 'movimiento' contacto,' o 'causa.' Existen además de estas otras informaciones asociadas a los predicados semánticos.
http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

FrameNet es una base de datos léxica que describe palabras mediante estructuras de marcos semánticos. **Un Marco semántico se utiliza para describir un objeto, estado o evento**. Cada Marco representa un predicado (p ej. comer, remover) con una lista de argumentos que constituyen los argumentos semánticos del predicado. Se denomina **Unidad Léxica** al emparejamiento de una palabra con un significado.

La base de datos léxica FrameNet contiene alrededor de 10.000 unidades léxicas y más de 120.000 frases de ejemplo.
Las palabras con varios significados están representadas por varias unidades léxicas.
Algunos ejemplos de nombres de marcos en FrameNet son Being_born (nacer) y Locative_relation (ubicación).
Junto con el nombre, los marcos incluyen una descripción textual del concepto que representan.
Cada marco tiene un cierto número de elementos básicos, secundarios y suplementarios (que describen papeles semánticos). En el caso de Being_born el único elemento básico del marco es Child, y los elementos secundarios son tiempo, lugar, familiares, etc.
Cada unidad léxica está asociada a una serie de elementos del marco a través de anotaciones.

FrameNet ofrece en las frases de ejemplo algunos datos sobre las funciones sintácticas que los elementos de los marcos desempeñan.
En **FrameNet**, además, los marcos se asocian con frases de ejemplo y los elementos de marco están marcados dentro de las frases. Así, la frase She was BORN about AC 460 (Ella nació aproximadamente en 460 DC) se asocia con el marco Being_born, mientras _She_ (ella) se marca como elemento básico Child y about AC 460 está marcado como tiempo. También incluye información estadística sobre sus propios marcos.
Existe un índice de las unidades léxicas definidas (https://framenet.icsi.berkeley.edu)
Entre otros elementos un frame en FrameNet puede constar de los siguientes campos:

Instancia. (Ejemplo de instancia del tipo definido),
Definición textual de la palabra,
Relación con otros marcos,
Hereda de, Heredada por, Escenario,
Escenificada en, Usos,
Usada por, Subframe de,
Tiene Subframes,
Precede a,
Es Precedida por,
Es Causa de,
Unidad léxica (Nombre del archivo de la unidad,por ej. _entity.n_), etc.

Procesamiento del Lenguaje Natural.

Reglas de Sustitución

Estas reglas (También llamadas Rewrite Rules o Context Free Grammar -CFG- Rules) capturan las regularidades en el orden en que se combinan las palabras y partes en la oración. Ellas describen como una categoría sintáctica puede ser reescrita como una o más categorías sintácticas o palabras, lo que significa que el símbolo de la izquierda puede reescribirse como una secuencia de símbolos (de la misma familia) a la derecha.
Podemos interpretarlas como que permiten expandir paso a paso los constituyentes de la oración y construir así el árbol sintáctico.
Estas reglas no dependen del contexto, sino que solo dependen de la categoría del antecedente y por eso la gramática que definen se suele denominar Context Free Grammar (CFG).

En general una oración puede descomponerse en sujeto y predicado y estos sucesivamente ir desglosándose en otros constituyentes. Según estas reglas se puede reescribir una categoría sintáctica en una palabra (o más) que corresponde a esa categoría. Es decir que la naturaleza de estas reglas es tal que una categoría sintáctica puede ser reescrita como una o más categorías sintácticas o palabras. Para producir la estructura de una oración se suele comenzar por simbolizar la oración por S (Sentence). Esta parte de la gramática se encuentra por lo general anotada también en un diccionario.

Extracción automática de conocimiento y llenado de Frames

Históricamente, la información orientada a constituir Bases de Datos de Conocimiento (KDB) ha sido almacenada _manualmente_ (interactivamente) en Marcos y Slots, a veces por ejemplo, empleando software específico para el desarrollo de Ontologías, pero siempre de manera interactiva y siguiendo la interpretación de personas sobre las interrelaciones entre los conceptos. En años recientes se han desarrollado programas para cumplir esta tarea con miras a aplicar la información obtenida en el análisis sintáctico de textos no estructurados.
Existen aplicaciones para extraer automáticamente información de grandes bases de texto. Entre las varias existentes pueden mencionarse las denominadas TextRunner, DIRT y Prismatic.

TextRunner extrae relaciones en forma de tuplas. Contiene más de 800 mil extracciones y se empleó en varias tareas especificas de NLP. Identifica y extrae relaciones automáticamente usando el procedimiento Conditional Random Field sobre texto en LN. Por ser una técnica veloz permite procesamiento en gran escala (web).

DIRT (Discovering Inference Rules from Text) Identifica reglas de inferencia en caminos de dependencia que tienden a vincular los mismos argumentos. Se aplica un análisis de dependencias sobre texto (se ha reportado su aplicación sobre 1 GB de texto), se colectan los caminos entre argumentos y se calcula la similitud entre caminos. Se empleó para reconocimiento de vinculación de textos (Recognition of TextualEntailment RTE).

El procedimiento para generar la base PRISMATIC se diferencia en varios aspectos de los anteriores pero comparte el hecho de extraer información de grandes bases de datos. Algunas de sus características son:

* Los Marcos no se plantean como conceptos generales abstractos sino que están definidos por las palabras y relaciones en el discurso.

* Dos marcos no iguales son considerados diferentes a partir de dos frases distintas aunque el significado sea casi el mismo: _Juan ama a María_ and _Juan adora a María_.

* Prefiriendo palabras a conceptos se elimina la necesidad de determinar a qué concepto corresponde a cada palabra cuando se hace el procesamiento automático. La redundancia hace el resto. Antes que conocimiento descriptivo ofrece solo conocimiento más bien numérico de las frecuencias de aparición de los predicados y sus argumentos en un corpus.

* Determina perfiles estadísticos automáticamente.

* Provee adicionalmente la posibilidad de inferir la restricción de tipos.

Uno de los cuellos de botella más significativos en NLP ha sido la falta de recursos informáticos que contengan información sobre los predicados del discurso. El software Prismatic ha sido capaz de obtener esta información y almacenarla en forma de marcos y ranuras. Muchas aplicaciones de NLP se benefician de la interpretación de relaciones léxicas en el texto procesado.

Por ejemplo, para seleccionar opciones de significados de **verbos y sustantivos**. Si se sabe que las cosas que pueden _anexarse_ son mayormente entidades geopolíticas, entonces, de la frase _Napoleón anexo el Piemonte_ se puede inferir que el Piemonte debe ser una entidad geopolítica (al menos con alta probabilidad).
Si bien existen algunos sistemas (diccionarios) semejantes, tales como Verbnet y Framenet, y que estos proveen algún tipo de la información deseada a través del empleo de Frames, están construidos manualmente y es natural concluir que las especificaciones que pueden almacenar sobre tipos léxicos solo pueden serlo a muy alto nivel, lo que los hace inviables para resolver el ejemplo citado de Napoleón. Un sistema de llenado automático por el contrario, es de esperar que pueda almacenar información de predicados grano_no, es decir más detallada y general y a partir de procesar grandes cantidades de datos.

El sistema referencial para almacenamiento automático de Conocimiento resulta ser el diseñado para la KDB Prismatic. El proceso se compone de tres módulos principales:

1. Corpus Processing

El Corpus o Cuerpo de búsqueda se puede entender, entre otros significados, como la fuente primaria de datos a partir de la cual se extrae el conocimiento. Así un Corpus sería una colección de textos, a veces constituido por solo un conglomerado de texto y otras veces estructurado de algún modo (Ontologías, Bases de Conocimiento, lenguajes de marcado). Consiste en:

a. Parsing (Análisis de dependencias, ejecutado por un software **Analizador Sintáctico** por Dependencias. Se conviene que una palabra depende de otra si existe una relación de concordancia o conexión entre ellas, o si el desplazamiento en la frase o el borrado de una implica hacer lo mismo con la otra.

b. Reconocimiento de entidades mencionadas (Named Entity Recognition-NER) Anotaciones de términos restrictivos para las entidades nombradas en en los documentos.

c. Resolución de Coreferencias (Coreference Resolution Component). Módulo encargado de resolver _coreferencias_, es decir palabras en una oración que se refieren a la misma entidad.

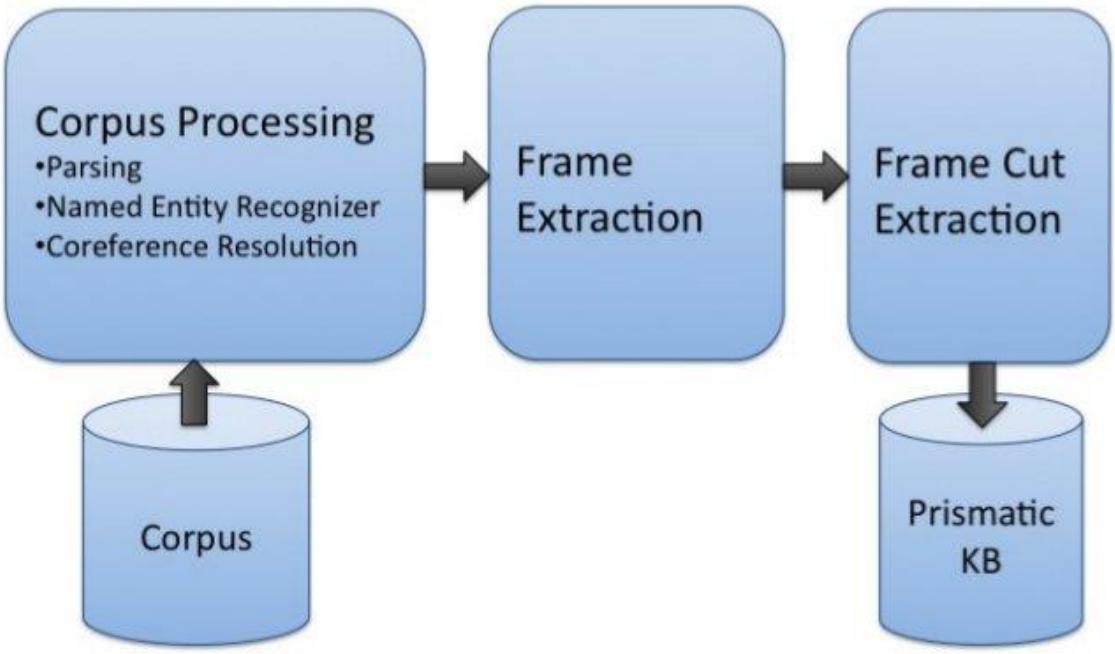
Por ejemplo: Juan empujo a quien estaba delante de él. (_Juan_ y _él_ se refieren al mismo componente)

2. Extraccion de Frames

El resultado es almacenado en Frames, las que constituyen la representación básica de relaciones léxicas y contexto. Para 30 gb de texto resultan aproximadamente un millón de Marcos.

3. Podado o recorte (Frame-Cut)

Consiste en la extracción de Frames de interés sesgadas. Por ejemplo para una frame-cut S-V-O, esta relación es detectada sobre todas las frames junto con información estadística (frecuencia de ocurrencias). Se reduce y generaliza el conjunto de relaciones descartando detalles menos relevantes. En el caso de la frase de Napoleón se computa cual es el tipo más frecuente a ser anexado a partir de la información hallada en el texto no estructurado original. Esto se hace para poder aplicar reglas sobre la base de conocimientos.



Procesamiento del Corpus

El Corpus o Cuerpo de búsqueda se puede entender, entre otros significados, como la fuente primaria de datos a partir de la cual se extrae el conocimiento. Así un Corpus sería una colección de textos, a veces constituido por solo un conglomerado de texto y otras veces estructurado de algún modo (Ontologías, Bases de Conocimiento, lenguajes de marcado).
El corpus empleado para generar la base **PRISMATIC** comprende fuentes tales como la Wikipedia completa, archivos del New York Times y párrafos de la web sobre tópicos listados en wikipedia. Luego de una limpieza y desetiquetado de la codificación html (demarcado) resulta un total de 30 GB de texto. De esto se extrae aproximadamente 1 millón de frames, y de estas se producen los frame-cuts más comunes tales como S-V-O, S-V-P-O y S-V-O-IO. (S sujeto V verbo P predicado O objeto IO objeto indirecto).
El paso principal en el procesamiento del cuerpo de documentos (en el caso de Prismatic) consiste en la aplicación del análisis sintáctico de dependencias. Este se emplea para identificar un grupo reducido de slots que son los que utiliza el método.

En Prismatic y Watson se emplea el software ESG -English Slot Grammar- como analizador sintáctico Este paso permite resolver la etapa siguiente que consiste en la Extraccion de Frames.

Es importante detectar las coreferencias, ya mencionadas, para identificar con precisión la entidad participante y para no perder información. Para esto se aplica un componente de resolución de correferencias basado en reglas.

Como también se mencionó, para identificar los tipos de argumentos en los slots, se emplea un NER (Named Entity Recognition) basado en reglas (Nombres de cosas, por ejemplo el token _Juan_ seria anotado como _Persona_).
El paso siguiente es extraer las frames del análisis ya realizado. Un frame constituye la estructura básica que representa un conjunto de entidades y sus relaciones en un párrafo. El frame está compuesto por un conjunto de pares slot-valor donde los slots son relaciones de dependencia extraídas por el parser y los valores son los términos de la sentencia o tipos anotados por el NER (Named Entity Recognition).

Con el fin de capturar solo la información más relevante, cada frame es restringida a dos niveles de profundidad como máximo. Por este motivo un gran árbol sintáctico podrá generar varias frames. La restricción de profundidad es por dos razones. Primero ningún parser es perfecto y sentencias más largas tienen más probabilidad de error. En segundo lugar, aislando un subárbol, cada frame se focaliza en los participantes inmediatos del predicado.

Aprendizaje automático en PLN

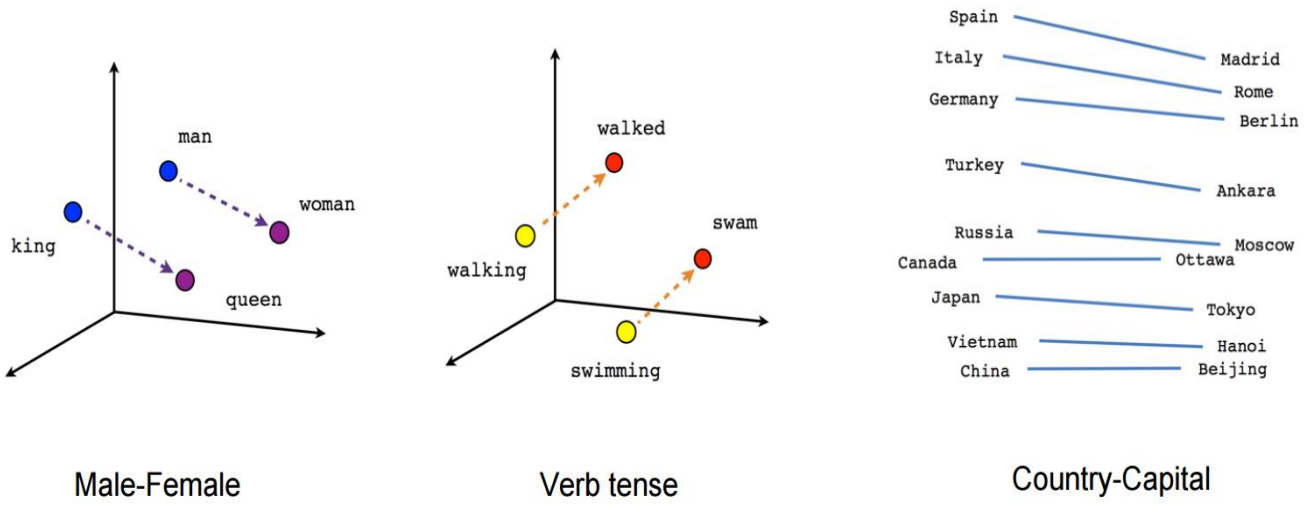
¿Qué es Word Embeddings en el Natural Language Processing?

Word Embeddings es una técnica del Natural Language Processing que consiste, básicamente, en asignar un vector a cada palabra. Este vector guarda información semántica, lo que permite que pueda ser asociado o disociado a otros vectores (palabras) según distintos contextos gramaticales.

En este sentido, Word Embeddings se convierte en una solución efectiva para **codificar tanto la semántica como la relación de las palabras entre sí.** Dicha codificación es generalizable, lo que significa que el algoritmo creado puede ser utilizado para resolver distintos tipos de problemas. Tales como de traducción, de generación de textos, entre otros.

Además, **los vectores creados mediante Word Embeddings pueden ser ingresados en redes neuronales artificiales.** Esto les facilita a dichas redes establecer relaciones complejas entre las palabras gracias a que ya conoce su semántica.

¿Cómo funciona Word Embeddings en el Natural Language Processing?



Fuente: TensorFlow

Las palabras, en sí, no pueden ser procesadas por los sistemas computacionales. Por lo tanto, estas deben se convertidas en formatos que sean digitalmente procesables. Aquí es donde entran en juego los vectores de Word Embeddings como representaciones matemáticas de las palabras. **Ya que la matemática es un lenguaje natural para las computadoras y permite ejecutar el Natural Language Processing.**

¿Pero qué son los vectores?

Los vectores en el Natural Language Processing son elementos matemáticos que poseen 2 características: longitud y orientación, y están ubicados en planos multidimensionales. Esto significa que un vector puede ser analizado tanto por lo que mide de largo como por hacia donde está apuntando.

Los vectores que representan palabras con significados similares se ubican más cerca entre sí, y el significado de cada palabra viene dado por su respectivo entorno.

Al ser elementos matemáticos, los vectores pueden ser sometidos a operaciones matemáticas como suma, resta, entre otras. Además, se les puede modificar sus dimensiones y sus perspectivas.

Ejemplo del funcionamiento de Word Embeddings

Sobre estos fundamentos, un ejemplo del funcionamiento de **Word Embeddings** como **Word2vec** sería:

Se tiene el vector correspondiente a la palabra "Rey". Este está asociado al vector de la palabra "Hombre". Así, si se le resta el vector "Hombre" y se le suma el vector "Mujer", quedaría entonces el vector "Reina".

Por otra parte, en los **Word Embeddings** más avanzados, como ELMo, Embeddings from Language Models **el vector de cada palabra se genera según el contexto de esta palabra dentro de una frase concreta**. Así, el **vector** para "**banco**" se genera de forma distinta según si se refiere a la institución bancaria o a un banco para sentarse.

En este panorama, **los algoritmos de Natural Language Processing más avanzados pueden comprender y procesar contextos de ironía, sarcasmo, humor, entre otros.** El análisis de *datasets* con cantidades enormes de contenidos permite este tipo de capacidades que, hasta el momento, parecían ser solo de humanos.

Limitaciones del Word Embeddings en el Natural Language Processing

El Word Embeddings es excelente para convertir las palabras en vectores. Sin embargo, no es suficientemente potente para comprender relaciones entre ellas en una misma frase. **Por lo tanto, no logra resolver los problemas de continuidad o de completado de frases dentro del Natural Language Processing.**

Por ejemplo, el modelo de Word Embeddings no puede completar frases como: "Estoy armando las maletas porque me voy de _____".

Bag of Words

El "bag of words" se refiere a la representación de un documento o fragmento de texto como una lista de palabras. Esta técnica permite analizar el contenido de un texto y extraer información valiosa, como las palabras más frecuentes, las frases comunes y los patrones temáticos.

¿Qué es el Bag of Words?

El Bag of Words (BoW) es una **técnica de procesamiento de texto** utilizada en procesamiento de lenguaje natural (NLP) y minería de texto. La técnica consiste en representar un documento de texto como un conjunto (bag) de palabras, ignorando el orden y la estructura gramatical de las palabras en el texto.
En el modelo BoW, se crea un vocabulario de todas las palabras únicas en un conjunto de documentos de texto, y cada documento se representa como un vector de tamaño igual al tamaño del vocabulario.
Cada posición en el vector representa una palabra en el vocabulario, y el valor en la posición indica la frecuencia de la palabra en el documento.

Por ejemplo, si el vocabulario contiene las palabras "gato", "perro" y "juguete", y un documento tiene una frecuencia de dos para la palabra "gato", una frecuencia de tres para la palabra "perro" y una frecuencia de cero para la palabra "juguete", entonces el vector de representación BoW para este documento sería [2,3,0].

La técnica BoW es útil para la clasificación y agrupación de documentos basados en su contenido textual. Es ampliamente utilizado en aplicaciones de procesamiento de texto como análisis de sentimientos, clasificación de texto, etiquetado automático y recomendación de contenidos.

¿Cómo funciona el Bag of Words?

El proceso de creación de la matriz BoW se lleva a cabo en varios pasos:

1. **Creación del vocabulario**: Se crea un conjunto de palabras único a partir de todos los documentos de texto que se van a analizar.

2. **Vectorización del texto**: Cada documento de texto se convierte en un vector de igual tamaño que el vocabulario creado. Las palabras que aparecen en un documento se convierten en un valor de uno en el vector, mientras que las palabras que no aparecen en el documento se convierten en un valor cero.

3. **Normalización del peso**: Se normalizan los vectores para tener una magnitud consistente. Esto se realiza dividiendo cada valor del vector por la suma de los valores en el vector. Una vez que se ha creado la matriz BoW, se puede utilizar en una variedad de aplicaciones de inteligencia artificial, como la clasificación de textos, el análisis de sentimientos, la agrupación de documentos y la recomendación de contenidos. Por ejemplo, un algoritmo de clasificación podría utilizar la matriz BoW para comparar las similitudes entre documentos y asignar una etiqueta de clasificación a cada documento en función de estas similitudes.

En resumen, el proceso de creación de la matriz implica la creación del vocabulario, la vectorización del texto y la normalización del peso. La matriz BoW puede ser utilizada para una variedad de aplicaciones de inteligencia artificial, incluyendo la clasificación de textos, el análisis de sentimientos y la agrupación de documentos.

El Bag of Words es una técnica utilizada en inteligencia artificial para procesar y analizar texto. En el ámbito de Content marketing y SEO optimización de los motores de búsqueda , se utiliza para analizar el contenido de los sitios web e identificar las palabras clave más relevantes. Además, el Bag of Words también puede ser utilizado para entender el lenguaje natural que utilizan los usuarios en sus búsquedas, lo que permite a los especialistas en marketing digital crear contenido más relevante y atractivo. De esta manera, el Bag of Words ayuda a optimizar la estrategia de contenido digital, mejorar la experiencia del usuario y aumentar el tráfico orgánico de un sitio web.

Referencias:

- Jaiswal, M., & Tiwari, M. (2018). Bag of words (BoW) model based recommender system using collaborative filtering technique. Computer Science Review, 28, 45-54. https://doi.org/10.1016/j.cosrev.2018.03.001
- Jurafsky, D., & Martin, J. H. (2019). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed.). Pearson Education.