Word Embeddings

Fabián Villena

Introducción

El lenguaje natural es complejo y se necesitan formas para poder representar significado en un espacio vectorial.

Los Word Embeddings son una técnica para capturar información semántica y sintáctica desde datos de texto.

Con esta técnica podremos mapear el significado de una palabra hacia un espacio vectorial conocido.

Aprendizaje de representaciones

Los Word Embeddings se enmarcan en un tópico del aprendizaje automático llamado aprendizaje de representaciones.

El aprendizaje de representaciones es un conjunto de técnicas para automáticamente descubrir las representaciones necesarias para poder modelar un problema sin tener que calcular las características de manera manual.

Sinonimia

Un componente importante del significado de una palabra es la relación entre significados de palabras. Una palabra puede tener un significado idéntico o similar a otra.

La definición formal de sinonimia es que dos palabras son sinónimas si ambas son sustituibles en cualquier oración sin cambiar las condiciones de verdad de una oración.

Similaridad de palabras

Si bien las palabras no tienen muchos sinónimos, muchas palabras sí tienen muchas palabras similares. Gato y perro no son sinónimos, pero sí son palabras similares. Debemos cambiar el foco desde hablar de relaciones entre significados hacia relaciones entre palabras.

Palabra 1	Palabra 2	Similaridad
viejo	nuevo	0,0
listo	inteligente	9,77
feliz	alegre	9,31
malo	culpable	4,23
largo	estrecho	2,15

Relación (relatedness) de palabras

El significado entre palabras puede estar relacionado de otras maneras distintas a la similaridad.

La relación, también llamada asociación en psicología, se puede definir con el ejemplo de las palabras *café* y *taza*, en donde estas palabras no son similares pero están asociadas por una coparticipación en un evento de la vida diaria.

Las palabras pueden estar asociadas al mismo espacio semántico el cual es un conjunto de palabras que cubren un dominio semántico particular, por ejemplo las palabras *cirujano*, *bisturí* y *enfermera* pertenecen al espacio semántico *hospital*.

Embeddings

Los vectores que representan palabras son llamados embeddings, estos vectores pertenecientes a un espacio vectorial semántico derivan de la distribución de los vecinos de las palabras.

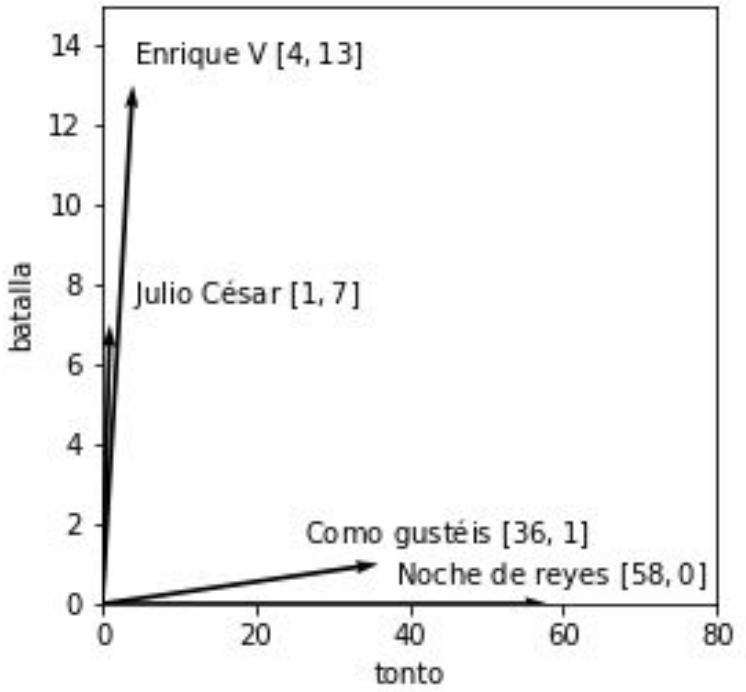
La palabra embedding deriva de su significado matemático como un mapeo desde un espacio hacia otro.

Vector

Un vector es una lista o arreglo ordenado de números llamados componentes.

Los vectores pueden ser operados aritméticamente.

Un espacio vectorial es una colección de vectores caracterizada por sus dimensiones y por sus propiedades.



La hipótesis distribucional y las representaciones

La hipótesis distribucional nos dice que palabras que ocurren en los mismos contextos tienden a tener significados similares. Esta idea ayudó a acuñar el el área de la semántica distribucional, que busca cuantificar las similaridades semánticas de acuerdo a sus propiedades distribucionales en grandes corpora.

Las palabras como vectores

Los vectores o modelos distribucionales de significado están basados en una matriz de coocurrencia que nos comunica la frecuencia en la cual las palabras coocurren en un contexto, este contexto puede ser un documento o una ventana de palabras vecinas.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Matriz palabra-palabra

Alternativamente a utilizar la matriz término-documento en donde las columnas y filas son palabras del vocabulario.

Cada celda almacena la cantidad de veces que la palabra objetivo y la palabra contexto coocurrieron en el mismo contexto de un conjunto de entrenamiento.

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

Coseno para medir similitud

Para medir la similaridad entre vectores (que están representando palabras) necesitamos una métrica que tome dos vectores de las mismas dimensiones y retorne una medida de similaridad.

La similaridad coseno es el producto punto entre dos vectores normalizado por el largo de los vectores, para que el largo del vector no sesgue la similaridad.

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

Representaciones distribuidas

En contraste a las representaciones locales en donde las entidades son representadas como símbolos discretos y las interacciones entre entidades son codificadas como un conjunto de relaciones discretas formando un grafo.

Las representaciones distribuidas cada entidad es representada como un vector de valores y el significado de una entidad y sus relaciones son capturadas por este vector y sus similaridades con otros vectores.

Word2vec

Ya sabemos cómo representar una palabra como un vector poco denso en donde cada dimensión corresponde a las palabras del vocabulario o a documentos en un corpus.

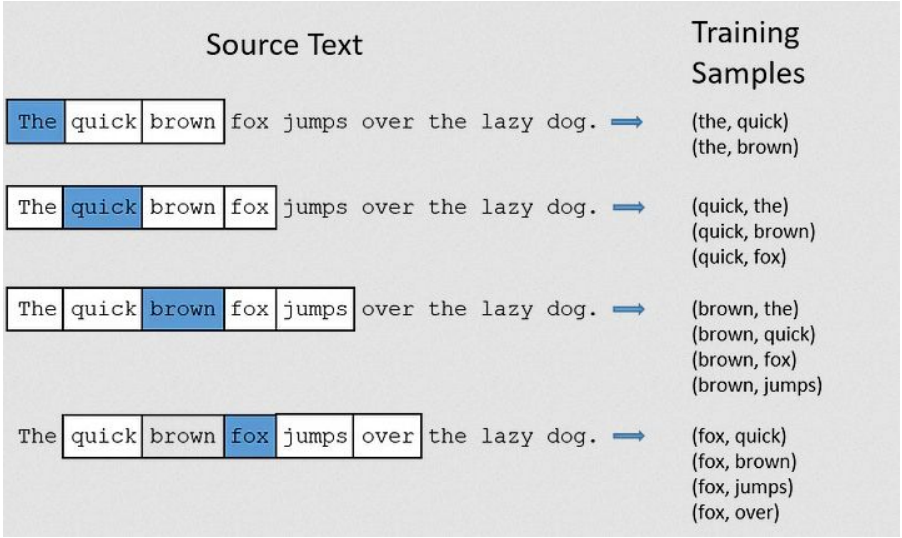
Ahora representaremos las palabras como un vector denso de bajas dimensiones en donde cada dimensión no tiene una interpretación clara. Estos vectores pueden tomar valores en el conjunto de los números reales.

Skip-gram

Skip-gram es un método para calcular embeddings y normalmente se refiere como a un sinónimo de Word2vec.

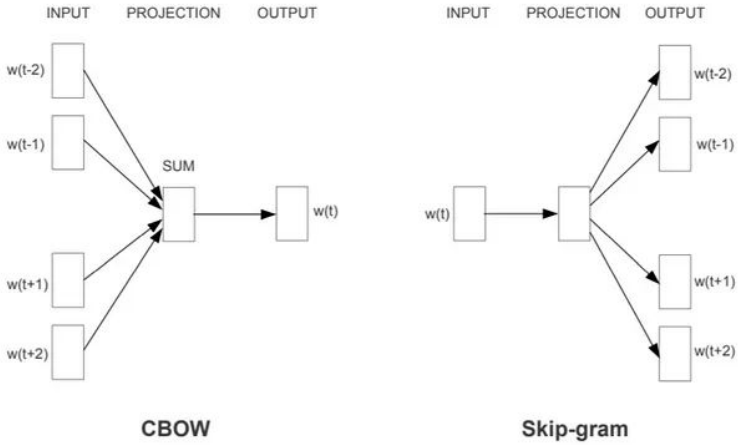
Es un método basado en redes neuronales para predecir palabras de contexto desde una palabra central.

Los embeddings serán los parámetros aprendidos de la red para cada palabra del vocabulario.



CBOW

Continuous Bag of Words es otro método implementado en Word2vec para calcular embeddings en donde también se entrena una red neuronal pero busca predecir la palabra central dado las palabras de contexto.



Embeddings estáticos

Al resolver las tareas expuestas en los algoritmos anteriores, los parámetros ajustados para cada palabra en las redes neuronales capturan las propiedades que necesitamos para representar las palabras como vectores.

Los embeddings obtenidos a través de Word2vec son estáticos, lo que significa que a cada palabra del vocabulario se le asigna un vector fijo, el problema es que una palabra puede tener múltiples significados según su contexto. Esto será resuelto posteriormente con los embeddings contextualizados.

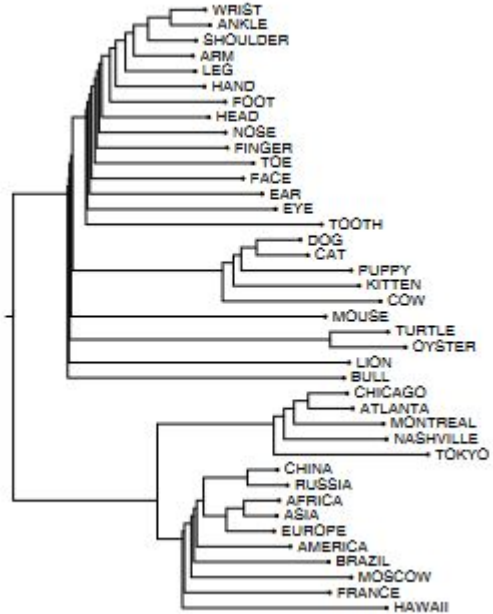
Autosupervisión

La idea que hemos expuesto es que podemos utilizar simplemente texto crudo como una tarea supervisada implícita. Este método se llama autosupervisión y evita la necesidad de algún tipo de etiquetado manual.

Visualizando embeddings

La forma más simple de visualizar el significado de una palabra en un espacio de embeddings es a través de listar las palabras más similares a la palabra consultada.

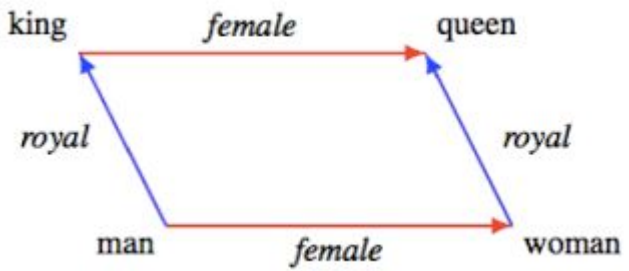
También podemos utilizar métodos de agrupamiento y métodos de reducción de dimnesionalidad.



Relaciones entre significados

Otra propiedad semántica de los embeddings es la habilidad para capturar significados relacionales.

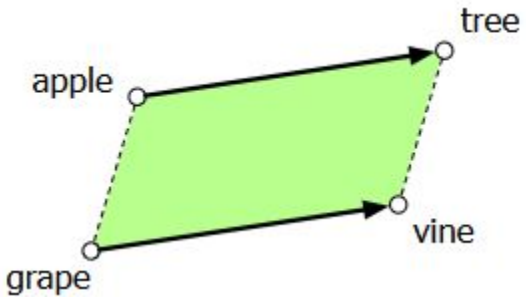
Se refiere al grado de similaridad entre dos o más relaciones entre diferentes conceptos. Se enfoca en la comparación y evaluación de relaciones.



El modelo del paralelogramo

El modelo del paralelogramo es una manera de resolver problemas simples de analogía en la forma **x** *es a* **y** *como* **x*** *es a* **?** .

En este tipo de problemas se le da al sistema un problema como **manzano**:**manzana**::**uva**:**?** y el sistema debe responder **parra**.



Analogías de palabras

Estas pruebas de analogía que pueden ser resueltas con el método del paralelogramo nos sirven para evaluar intrínsecamente el rendimiento de nuestro embedding.

Existen conjuntos de datos con estas pruebas para medir el rendimiento.

```
perro:ladrar      ::  gato:maullar
otoño:hojas       ::  invierno:nieve
libro:leer        ::  película:ver
fútbol:balón      ::  tenis:raqueta
lápiz:escribir    ::  pincel:pintar
sol:día           ::  luna:noche
pájaro:ala        ::  pez:aleta
cuchillo:cortar   ::  martillo:golpear
café:cafeína      ::  té:teína
nube:lluvia       ::  sol:brillo
```

fastText

fastText es otro algoritmo para poder calcular word embeddings derivado desde Word2vec. La diferencia de este algoritmo es que los embeddings están calculados a nivel de piezas de palabras, por lo que fastText es capaz de retornar una representación para palabras fuera de vocabulario al componer las representaciones de las piezas de palabra que componen la palabra fuera del vocabulario.

Representaciones preentrenadas

Estas representaciones calculadas son capaces de mapear el significado de una palabra hacia un espacio vectorial. Esto es útil porque estos vectores pueden ser utilizados como parámetros de inicialización en arquitecturas de redes neuronales para resolver tareas de procesamiento de lenguaje natural.

Existen múltiples representaciones ya preentrenadas libres para ser utilizadas por cualquier persona.