**Colocaciones**

**Jordy Joaquin Cuan Robledo [1], Eduardo Cantorán Flores [1],**
**Rafael Pérez Aguirre [1]**
**[1] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, México**

## Colocaciones

La colocación es una expresión que consiste en dos o más palabras que corresponden a una forma particular de decir las cosas. Las palabras juntas pueden significar mas que la suma de sus partes.

Algunos ejemplos de colocaciones:

Frases nominales: Té fuerte y armas de destrucción masiva

Verbos compuestos:

Criterios típicos para colocaciones:

- No composicionalidad.
- No sustituibilidad.
- No modificabilidad.

Las colocaciones no pueden ser traducidas palabra por palabra a otro lenguaje.

No compuestas

Una frase puede es compuesta si se puede predecir por el significado de su contenido

Por ejemplo: Nuevas compañías

Una frase no es compuesta si no se puede predecir por el significado de sus partes.

Por ejemplo: Hot dog

Los modismos es uno de los mejores ejemplos .

No sustituibles.

No podemos sustituir los componentes de una colocación por sinónimos. Muchas colocaciones no pueden ser modificadas libremente con material léxico adicional o a través de transformaciones gramaticales, estas últimas se les conoce como no modificables.
Por ejemplo: vino blanco y blanco vino.

Para poder identificar las colocaciones fácilmente contamos con diferentes métodos:

- Selección de colocaciones por frecuencia
- Selección de colocaciones basada en media y varianza
- Prueba de hipótesis
- Información mutua

**Selección de colocaciones por frecuencia**

Para encontrar las colocaciones se hace un conteo del número de repeticiones, se define un tamaño de ventana máximo y utilizando PoS (“Part of Speech”, partes del habla en español) se hace un filtrado de las frases para obtener las más parecidas.

Ventana de colocaciones
Es necesario definir un tamaño de ventana para las colocaciones, ya que estas pueden aparecer en distancias variables. Esto quiere decir que las palabras aparecen no consecutivamente, en este caso no podemos utilizar la selección por frecuencia.

**Media y Varianza**

En este método la media es la promedio de la distancia que existe entre dos palabras del corpus. La fórmula de la varianza es:

$$s^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$

Donde n es el número de veces que dos palabras co-ocurren, di es el conjunto de palabras de la co-ocurrencia i, y $\mu$ es la media.

La media y la varianza caracterizan la distribución de distancias entre dos palabras del corpus. Una varianza alta significa que la co-ocurrencia es muy baja. Una varianza baja significa que la co-ocurrencia es casi a la misma distancia.

**Descartando la probabilidad**

Dos palabras pueden co-ocurrir por casualidad. Una alta frecuencia y una baja varianza pueden ocurrir por accidente. En este tipo de casos podemos utilizar el método de prueba de Hipótesis. Con este método podemos descartar que la co-ocurrencia sea por casualidad y en verdad se trate de una asociación.

Formular una hipótesis nula H0 de que no existe una asociación entre las dos palabras más allá de la ocurrencia. La hipótesis nula será verdadera solo si dos palabras no forman una colocación. Si la hipótesis es rechazada significa que las dos palabras no co-ocurren por accidente y estas forman una colocación. La hipotesis nula sera rechazada solo cuando tenga valores muy pequeños, normalmente valores menores a 0.05.

T-Test

El T-Test se enfoca en la media y la varianza de una muestra de mediciones, donde la hipótesis nula es la muestra que se extrae de una distribución con media $\mu$. El T-Test observa la diferencia entre la media esperada y la observada, escalada por la varianza de datos, nos dice que tan probable es obtener un ejemplo de media y varianza, asumiendo que el ejemplo es tomado de una distribución normal con media $\mu$.

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

Donde x es la media real de los datos (observada), S es la varianza, N es el tamaño del ejemplo y $\mu$ es la media de distribución esperada.

**Prueba de hipótesis de diferencias**

Nos sirve para encontrar la mejor distinción entre dos palabras con patrones de co-ocurrencia. Por ejemplo, queremos encontrar la diferencia de significado entre las palabras fuerte y poderoso.

El T-Test se extiende a la comparación de las medias de dos poblaciones normales.

Aquí el promedio de la prueba de hipótesis es 0. En el denominador agregamos la varianza de las dos palabras como se muestra en la siguiente fórmula.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

$x^2\ Test$

A diferencia del T-Test, aquí no se asume que la la probabilidad es una distribución normal. La esencia de este test es que compara las frecuencias observadas con las frecuencias esperadas por la independencia. Si la diferencia entre la frecuencia observada y la esperada es grande, podemos rechazar la hipótesis nula de independencia.

Esto se puede realizar con la siguiente fórmula.

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

**Información Mutua**

Es aproximadamente una medida de lo mucho que nos dice una palabra acerca de otra. En otras palabras, nos da la cantidad de dependencia que existe entre dos palabras.

$$\begin{aligned} I(x', y') &= \log_2 \frac{P(x'y')}{P(x')P(y')} \\ &= \log_2 \frac{P(x'|y')}{P(x')} \\ &= \log_2 \frac{P(y'|x')}{P(y')} \end{aligned}$$