Descubrimiento de información en textos

Tarea del tema 2: Estudio de distintos corpus

Gabriel Vázquez Torres

Descubrimiento de información en textos

Tarea del tema 2: Estudio de distintos corpus

Contenido

Descripción de la práctica 2
Descripción de los corpus 3
Brown Corpus 3
Sussane Corpus 4
Penn Treebank 5
Comparativa de distintos aspectos 6
Tipo de etiquetado 6
Tamaño del corpus 6
Tamaño del conjunto de etiquetas 6
Temáticas incluidas 7
Procedencia 8
Breve análisis de qué corpus es el más apropiado diferenciando entre el corpus de Brown y el de Sussane. 9
Bibliografía 10

Descripción de la práctica

En esta práctica vamos a realizar un análisis y estudio de diferentes corpus. Estos son los de BROWN, SUSANNE Y PENN TREEBANK.

En este documento se hablará sobre diferentes apartados:

1. Descripción de los corpus
2. Comparativa concisa de distintos aspectos que se consideren relevantes entre ellos:
    a. Tipo de etiquetado: etiquetado léxico (POS tagging), sintáctico, etc.
    b. Tamaño del corpus.
    c. Tamaño del conjunto de etiquetas.
    d. Temáticas incluidas.
    e. Procedencia de los textos: periódicos, transcripciones de habla, etc.
3. Breve análisis de cinco líneas como máximo sobre qué Corpus, si el de Susanne o el de Brown, es más apropiado en función de sus características para extraer información estadística significativa referente a cuales son las etiquetas y las parejas de etiquetas consecutivas que aparecen más frecuentemente en los textos.

Descripción de los corpus

A continuación se hará una descripción de los corpus previamente comentados:

Brown Corpus

La historia de la lingüística de corpus es muy corta. El punto de arranque obligado es la aparición en 1964 del "Brown University Estándar Corpus of Present-Day American English (**Brown Corpus**), de Francis y Kučera, que es el primer corpus concebido y construido para residir en una computadora y ser explotado mediante programación informática.

Tiene aproximadamente medio siglo, una historia corta y difícil en sus primeros tiempos y que coinciden casi exactamente con el primer período de expansión de la lingüística generativa, que en aquellos años mantenía una oposición radical a todo lo relacionado con hechos lingüísticos concretos, frecuencias, estadística, variación, etc. [1]

Francis y Kučera publican en 1967 su obra "Computational Analysis of Present-Day American English", que proporciona estadísticas de lo que hoy se conoce simplemente como "**Brown Corpus**". El Brown Corpus es una selección cuidadosamente compilada del actual inglés americano con un total aproximado de un millón de palabras extraídas de una amplia variedad de fuentes. [2].

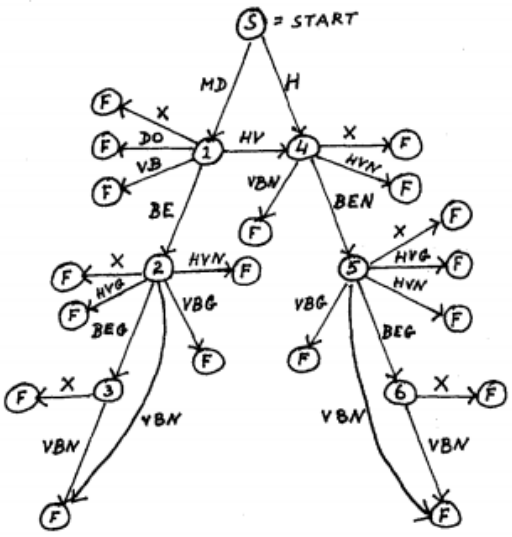

Ilustración 1: Representación en diagrama del corpus. Cada línea está etiquetada con un símbolo apropiado para la clase del nodo. [2]

Brown Corpus ha sido la base de otros corpus posteriores a él como el Lancaster-Oslo-Bergen Corpus o SUSANNE. En un principio contenía 1.014.312 palabras y 15 categorías de texto. Es un corpus de tipo "POS Tagging" (Part-Of-Speech tagging, de análisis léxico) y posee 82 etiquetas distintas. [3]

Actualmente sirve como conjunto de entrenamiento y de test en diferentes campos de la ciencia, e.g. PLN o biomedicina. La herramienta NLTK lo incorpora. [4]

En conclusión podríamos decir que Brown Corpus ha servido de inspiración para muchos investigadores. Fue la primera representación electrónica de un lenguaje y actualmente sigue siendo utilizado en diferentes ámbitos de la ciencia. [5]

Sussane Corpus

Susanne es la abrevación de *Surface and Underlying Structural ANalysis of Natural English*. Es un corpus procedente del Brown Corpus, que originariamente procedía de 64 de las 500 muestras del Brown Corpus y su versión inicial está formado por unas 130.000 palabras. [6]

Al igual que el Brown Corpus, es de tipo *"POS Tagging"* (Análisis Léxico), y está formado por **353 etiquetas**, y las temáticas que contiene son las A,G,J y N del Brown Corpus (Prensa, Bellas Letras, Aprendidas y Ficción: Aventura y Occidental). [7]

El Corpus SUSANNE se creó, con el patrocinio del Comité Económico y Social Research Council (Reino Unido), como parte del proceso de elaboración de una taxonomía completa del lenguaje de la ingeniería orientada y el esquema de anotación de la gramática (lógica y de la superficie) de Inglés. El Corpus SUSANNE, a pesar de haberse creado con el patrocinio de TEI, no cumple la normativa.

El esquema analítico SUSANNE ha sido desarrollado sobre la base de las muestras de ambos Inglés británico y americano. Fue inicialmente orientado hacia la lengua escrita solamente, y de hecho contiene muestras exclusivamente del lenguaje escrito. Sin embargo, en trabajos posteriores patrocinado por primera vez por el *Royal Signals and Radar Establishment*, se produjeron extensiones al sistema para anotar los fenómenos distintivos estructurales del lenguaje hablado, y ha aplicado a estas muestras de los últimos Inglés hablado espontáneo (modificación mostrada en el Corpus CHRISTINE). La primera etapa del Corpus CHRISTINE, que incluye análisis de una equilibrada sección del Inglés hablado en todas partes del Reino Unido en la última década, fue lanzado en agosto de 1999 y es uno de los corpus orales para poder analizar el lenguaje hablado. [8]

El Corpus SUSANNE abarca un subconjunto de aproximadamente 130.000 palabras del Brown Corpus de Inglés Americano, anotado, de acuerdo con el esquema de Susanne. Los motivos originales para la producción de esta base de datos incluye el de proporcionar mejores estadísticas para el análisis probabilístico, pero en este sentido, el Proyecto SUSANNE fue alcanzado después de su creación por los proyectos (en particular, Mitchell Marcus Pennsylvania proyecto Treebank) que han utilizado métodos cuasi-industrial para generar cuerpos mucho más grandes de material a analizar gramaticalmente. Sin embargo, el Corpus Susanne sí que mejora notablemente el análisis probabilístico en comparación con el Brown Corpus, aunque de esto hablaremos en el último punto del trabajo.

Penn Treebank

El Treebank Penn, es un corpus de más de 4,5 millones de palabras de Inglés Americano. Durante la primera fase de tres años del Proyecto Penn Treebank (1989 - 1992), este corpus posee 2 tipos de etiquetado de, de tipo léxico y de tipo sintáctico. Sus orígenes están en la Universidad de Pennsylvania.

Existen diferentes tipos de sintaxis para las diferentes lenguas. Para el español nos encontramos con dos. [9] [10]

La estructura sintáctica se ha representado generalmente como una estructura arbórea que recibe la denominación de TreeBank.1 En la mayoría de los casos se ha empleado etiquetado gramatical. La denominación alternativa corpus parseado se emplea a menudo con el Treebank: realizando énfasis en la primacía de las frases en lugar de las estructuras arbóreas. Los corpus Treebanks se pueden crear a mano mediante un grupo de lingüistas que anotan cada frase con una estructura sintáctica, o mediante procedimientos semi-automáticos, donde un analizador sintáctico (parser) asigna la estructura bajo la supervisión de un lingüista. En la práctica, el completo control del parseado del lenguaje natural con el objeto de establecer diferentes corpus es una labor intensiva que dedica el tiempo de varios equipos de lingüistas, pudiendo alcanzar varios años. [11]

El conjunto de muestras del que está compuesto, procede de varios corpus, entre ellos, el Brown Corpus.

1. Tiene 36 etiquetas de análisis léxico, además de 12 etiquetas para puntuaciones y símbolos. Y 14 etiquetas de tipo sintáctico además de 4 elementos nulos. De los 3 Corpus que analizamos en este trabajo, este es el más nuevo y el mejor en el sentido del análisis probabilístico del lenguaje, ya que es más sencillo de analizar que los dos anteriores. [12]

```
Au cours de             P+PNP
la                      Dfs
conférence_de_presse    NCfs+NPN
qui                     PROR3fs
a                       VP3s
clos                    VKms
cette                   Dfs
rencontre               NCfs
,
le                      Dms
premier_ministre        NCms+AN
est-allemand            Ams+XA
est                     VP3s
revenu                  VKms
sur                     P
les                     Dmp
incidents               NCmp
de                      P
lundi                   NCms
soir                    NCms
```

Ilustración 2: Ejemplo de etiquetado en tree bank para el francés. [12]

Comparativa de distintos aspectos

Tipo de etiquetado

Corpus	*Etiquetado*
Brown	Léxico
Sussane	Léxico
Penn Treebank	Léxico y sintáctico

En el corpus **Brown** y en el corpus **Susanne**, el tipo de etiquetado que tienen es de POS Tagging (Part Of Speech Tagging o Etiquetado gramatical en español) o lo que es lo mismo, etiquetado léxico. Sin embargo el corpus **Penn Treebank** posee un etiquetado mixto, de tipo léxico y de tipo sintáctico [13].

Tamaño del corpus

Corpus	*Tamaño de Corpus*
Brown	500 muestras de 2.000 o más palabras (1.014.312 palabras en total) [5]
Susanne	64 Muestras de 2.000 o + palabras cada una (130.000 palabras) [14]
Penn Treebank	4.885.798 palabras en total

El corpus **Susanne** posee 64 muestras obtenidas de las muestras del corpus **Brown**, y el corpus **Penn Treebank**, obtuvo esas palabras de distintos textos anteriormente numerados.

Tamaño del conjunto de etiquetas

Corpus	*Tamaño conjunto de etiquetas del Corpus*
Brown	82 divididas en 6 partes: A.Partes de la oración Nombre, común y propio, verbo, adjetivo.... B. Función de las palabras: determinantes, preposiciones, conjunciones... C. Palabras individuales importantes: no, infinito existencial, la forma del verbo. D. Las marcas de puntuación de importancia sintáctica. E. Morfemas flexivos. F. Dos etiquetas (FM y NC) PALABRAS extranjera o citada.

Susanne	353 wordtags (sin contar las etiquetas para expresiones gramaticales)
Penn Treebank	36 y 12 para puntuaciones y símbolos en el etiquetado léxico. Para este cas podemos ver todas las etiquetas utilizadas en la siguiente referencia [15] Para el etiquetado sintáctico 14 y además 4 más para elementos nulos. [16]

Temáticas incluidas

Corpus	*Temáticas incluidas*
Brown	A. DE PRENSA: Reportaje B. PRENSA: Editorial C. DE PRENSA: Comentarios D. RELIGIÓN E. HABILIDAD Y HOBBIES F. POPULAR LORE G. Bellas Letras H. VARIOS: Gobierno de los EE.UU. y los órganos de Lujo J. APRENDIDAS K. FICCIÓN: General L. FICCIÓN: Misterio y ficción de detectives M. FICCIÓN: Ciencia N. FICCIÓN: Aventura y Occidental P. FICCIÓN: Romance y Love Story R. HUMOR [5]
Susanne	Las temáticas que contiene son obtenidas del corpus Brown: A. De Prensa. G. Bellas Letras. J. Aprendidas N. Ficción : Aventura y Occidental.

Penn Treebank	Diarios Revistas [17]

Procedencia

Corpus	*Procedencias de textos*
Brown	R. HUMOR *(9 textos)*
Susanne	El Susanne procede del corpus Brown, de 64 de las 500 muestras que posee el corpus Brown, y fue creado para mejorar su análisis probabilístico.
Penn Treebank	Procede de los siguientes documentos: ▪ Dept. of Energy abstract ▪ Dow Jones Newswire stories ▪ Dept. of Agriculture bulletins ▪ Library of America texts ▪ MUC-3 messages ▪ IBM Manual sentences ▪ WBUR radio transcripts ▪ ATIS sentences ▪ Brown Corpus, retagged

Breve análisis de qué corpus es el más apropiado diferenciando entre el corpus de Brown y el de Sussane.

El corpus Sussane posee un conjunto de etiquetas más precisas, granulares y fáciles de interpretar por personas que las etiquetas del corpus Brown. Podemos poner muchos ejemplos. En el caso de la etiqueta CSN que se utiliza en el Corpus Sussane, es equivalente a usar la etiqeuta CS y la preposición IN en el corpus Brown lo que otorga mayor rapidez de entendimiento. En conclusión, podemos decir que las etiquetas del corpus Sussane aparecen más frecuentemente que las de Brown.

Bibliografía

[1] G. Rojo. [En línea]. Available: https://gramatica.usc.es/~grojo/Publicaciones/Lgca_corpus_lgca_espanol.p

[2] F. y. Kucera, «aclweb,» [En línea]. Available: http://www.aclweb.org/anthology/C80-1006.

[3] A. Lindebjerg. [En línea]. Available: http://www.hit.uib.no/icame/brown/bcm.html.

[4] O. KHOKOVSKAIA. [En línea]. Available: http://radio.feld.cvut.cz/conf/poster/proceedings/Poster_2017/

[5] [En línea]. Available: https://en.wikipedia.org/wiki/Brown_Corpus.

[6] [En línea]. Available: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpor

[7] [En línea]. Available: https://www.grsampson.net/SueDoc.html.

[8] [En línea]. Available: https://research.csc.fi/-/susanne-corpus.

[9] [En línea]. Available: http://www.lllf.uam.es/~sandoval/UAMTreebank.html.

[10] [En línea]. Available: http://clic.ub.edu/.

[11] [En línea]. Available: https://es.wikipedia.org/wiki/TreeBank.

[12] «books.google,» [En línea]. Available: https://books.google.es/books?id=r3xyBgAAQBAJ&pg=PA238&lpg=PA238&dq=thematics+penn+treeban

[13] [En línea]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.8216&rep=rep1&type=p

[14] [En línea]. Available: https://www.grsampson.net/SueDoc.html.

[15] [En línea]. Available: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

[16] [En línea]. Available: https://www.clips.uantwerpen.be/pages/mbsp-tags.

[17] [En línea]. Available: https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html.

[18] Ron Kohavi, George H. John, «Wrappers for feature subset selection,» Mountain View, 1995.

[19] José Hernández Orallo, M. José Ramírez Quintana César, Introducción a la Minería de Datos.