Procesamiento del lenguaje Natural, nº 41 (2008), pp. 65-71 recibido 7-05-2008; aceptado 16-06-2008

Aproximación a la Categorización Textual en español basada en la Semántica de Marcos

Frame Semantics-based Approach to Spanish Textual Categorization

Mario Crespo Miguel
University of Cadiz
Avda. Gómez Ulla, s/n
mario.crespo@uca.es

Antonio Frías Delgado
University of Cádiz
Avda. Gómez Ulla, s/n
antonio.frias@uca.es

Resumen: FrameNet es un recurso basado en la Semántica de Marcos que trata de representar el modo por el que diferentes lenguas dan cuenta lingüísticamente de situaciones cotidianas. Los marcos funcionan al modo de paquetes de información sobre cómo hablar de una determinada situación. Este trabajo presenta un procedimiento para categorizar documentos a partir del análisis de las situaciones de FrameNet que concurren en un texto determinado. El conjunto de marcos situacionales es usado como un vector de rasgos en el que la presencia o ausencia de determinados marcos situacionales en un texto sirve para establecer su categoría. Los resultados muestran cómo nuestro sistema fue capaz de categorizar textos en español con gran precisión.
Palabras clave: FrameNet, Categorización textual, Recuperación de información.

Abstract: FrameNet is a resource based on Frame Semantics that comprises how languages account for daily situations linguistically. Frames represent information packets about how to convey information about a certain situation. This paper presents an approach to categorize texts by analysing the range of FrameNet situations that co-occur in a particular text. The set of FrameNet situations is used as a feature vector where the presence or absence of certain frames in a text is used to determine its category. Results show how our system was able to categorize texts in Spanish with high accuracy.
Keywords: FrameNet, Textual Categorization, Information Retrieval**.**

1. Introducción.

FrameNet (Ruppenhofer et al., 2006) es un proyecto de semántica léxica concebido para dar cuentra de cómo las lenguas son capaces de describir situaciones diarias por medio de sus unidades léxicas y de cómo los hablantes son capaces de expresar y entender información a través de ellas. De esta manera, los marcos situacionales funcionarían al modo de paquetes lingüísticos con la información necesaria para hablar sobre una situación determinada.

Fillmore (1982,1985) afirma que las personas entienden cosas realizando operaciones mentales sobre lo que ya saben. Este conocimiento se puede describir mediante marcos situacionales, los cuales están formados por un conjunto de palabras que evocan tales marcos cuando aparecen en el discurso. Si asumimos que la lista de palabras de un determinado lenguaje es limitado, entonces los marcos que les dan soporte deben ser finitos también. Sin embargo, como sabemos, el número de temas de los que podemos hablar es ilimitado. Por lo tanto, los marcos situacionales deben combinarse unos con otros en el discurso para expresar información sobre cualquier tema cotidiano: medicina, política, familia, etc. Si para hablar sobre un tema, se usa un cierto número de marcos situacionales, entonces la categorización de un texto debe ser factible desde las situaciones que se le asocian.

A continuación se presenta un procedimiento capaz de determinar el tema de un determinado documento a partir del análisis de los diferentes marcos situacionales que resultan estadísticamente significativos al analizarlo.

ISSN: 1135-5948

1.1. Categorización textual mediante conceptos vinculados al texto.

Los enfoques actuales sobre Categorización Textual han estado normalmente basados en técnicas de aprendizaje automático (Sebastiani, 2002), orientadas al aprendizaje de las categorías en las que se divide la clasificación de un conjunto de documentos. Estas técnicas suelen llevar a cabo un análisis estadístico de la frecuencia de los términos del documento y determinar así cuáles son los que poseen una mayor relevancia. Estos términos suelen aparecer dispuestos en un vector de rasgos usado para comprobar su peso (basado en su frecuencia) en un determinado documento. La frecuencia es el indicador principal de la pertenencia de un documento a una categoría específica.

Esta clasificación, basada en espacios vectoriales, podría tener un rendimiento deficiente si los documentos relevantes no contienen uno de los términos que conducen a que el documento sea recuperado. Además, la recuperación basada exclusivamente en términos puede ser un método vago y ruidoso. Esto ha hecho que ciertas líneas de investigación exploren la recuperación de información basada más en conceptos e ideas que se reflejan en el texto, que en términos usados como índices. La siguiente figura ilustra el proceso de recuperación de documentos basada en conceptos en vez de términos:

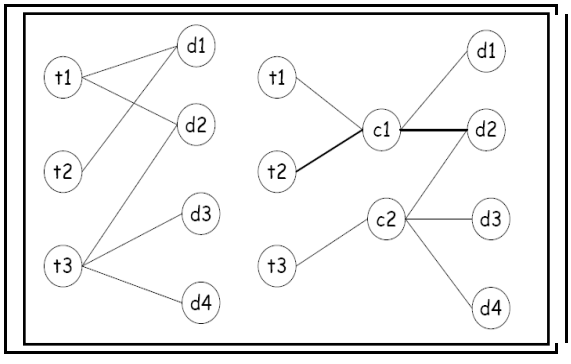

Figura 1. Recuperación de documentos mediante términos y mediante conceptos.

Como se puede apreciar en la figura anterior, los términos (t1, t2, etc,) apuntan directamente a un determinado dominio temático (d1, d1, etc) en el caso de la izquierda. A su derecha, los términos de un documento apuntan a determinados conceptos y de ahí a un dominio temático.

El problema principal es cómo obtener el espacio de conceptos sobre el que se fundamenta la clasificación. El modelo *latent semantic* de recuperación de información surge como una de las alternativas operativas que solucionan este problema. El fundamento director que conduce el desarrollo del modelo *latent semantic* (Dumais et al. 1988) (Deerwester et al. 1990) de recuperación de información se basa en la idea de que el tema general del texto (la semántica latente) se vincula con mayor profundidad con los conceptos volcados en el documento que con los términos de indización utilizados en la descripción. La propuesta se resuelve en un intento por desentrañar la semántica latente en los documentos a través de la identificación de los conceptos concretos vinculados; de este modo, el proceso de correspondencia entre documentos y consultas se establece a nivel de conceptos y no de términos, buscando minimizar el impacto del ruido y el silencio en la recuperación, y posibilitando recuperar documentos que no habían sido representados por los términos de consulta y excluir de la recuperación documentos con tales términos pero no asociados a los conceptos que expresan la necesidad de información.

El análisis que propone este modelo se sustenta sobre la técnica de descomposición en valores singulares, consistente en descomponer automáticamente la matriz de ocurrencias documentos-términos en varias matrices asociativas que definan la correspondencia entre documentos y conceptos, y entre términos y conceptos.

Entre los trabajos en este ámbito destacamos Huang (2003) el cual propone la clasificación textual mediante *máquinas de soporte vectorial* (Support Vector Machines) basadas en el modelo latent semantic.

2. Metodología.

En este trabajo se utilizan los marcos situacionales como unidades conceptuales al que se asocian una serie de términos del vocabulario. De esta manera, el conjunto de 795 marcos situacionales que componen FrameNet son usados como un vector de rasgos para computar la pertenencia o no de un documento

a una determinada categoría. Como veremos, la ausencia o presencia de uno de estos rasgos se estima analizando cuáles son los marcos estadísticamente significativos de un determinado documento o texto. Emparentados se encuentran los trabajos de Petridis et al. (2001) o Gómez et al. (2004) los cuales proponen un modelo de clasificación textual usando los synsets de WordNet como índices y aplicando las técnicas de máquinas de soporte vectorial. En el ámbito de la utilización de recursos léxico-semánticos para la categorización textual destaca Shehata et al. (2007), que entrena un etiquetador de roles semánticos basado en PropBank para anotar y así determinar la información relevante de la oración que será usada posteriormente para clasificar el documento.

2.1. Corpus

El corpus a analizar fue extraído de *medlineplus.gov*, un sitio web sobre salud de la biblioteca médica más grande del mundo, the United States National Library of Medicine[1] y umm.edu, el dominio web del centro medico de la Universidad de Maryland, ya que ambas proveen información en español para la comunidad hispana de Estados Unidos. A esto hay que sumarle las fuentes y secciones disponibles en español de la web del *elmundo.es, ecosumer.es y 100cia.es* lo que nos hizo contar con un corpus de 7730 documentos formado por 25674497 palabras y 76616 lemas diferentes y 4 áreas temáticas diferentes:

DOMINIO	DOCUMENTOS
Medicina y salud	3623
El Mundo	4107
Ciencia	950
Productos y consumo	462

Tabla 1. Número de documentos para cada ámbito temático.

Cada documento fue analizado de nuevo usando el analizador TreeTagger de la Universidad de Stuttgart[2] para el español. De este análisis sólo se tomó la información relativa a los lemas, de los que se computó su frecuencia en cada documento.

2.2. Procesamiento de los datos

2.2.1. Selección de marcos situacionales.

Desde la óptica de FrameNet, los disparadores son aquellas unidades léxicas que "disparan" o activan el marco en la mente de los hablantes cuando aparecen en el discurso. De esta manera, es lógico asumir que la selección de marcos situacionales ha de hacerse partiendo de las unidades del vocabulario. Cada marco situacional está compuesto de una lista de disparadores que serán usados para determinar si el marco situacional al que pertenecen es representativo del corpus. Este procedimiento usa una traducción para el español de los disparadores de cada marco situacional del inglés planteada por Crespo y Buitelaar (2008) en LREC'08. De esta manera, la Figura 2 muestra los disparadores de los marcos *Fall_asleep* y *Medical_instruments*:

Fall asleep	**Medical instruments**
A Sleeper goes from wakefulness to the altered state of conciousness called sleep.	*It includes words for medical instruments.*
Lexical Units: adormecerse, dormirse	**Lexical Units:** broncoscopio, algalia, catéter, endoscopio, ...

Figura 2. Vista de los disparadores de dos marcos situacionales diferentes.

Como vemos, bajo '*Lexical Units*' se definen aquellas unidades o disparadores que sirven para activar lingüísticamente el marco situacional. El número de disparadores varía tanto en FrameNet inglés como en nuestra traducción de marco a marco.

Existen diferentes métodos que pueden aplicarse sobre cada conjunto de disparadores y

[1] http://en.wikipedia.org/wiki/MedlinePlus

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger

tratar así de determinar si el marco situacional al que pertenecen debe ser interpretado como representativo del texto analizado. Un análisis de frecuencias de los disparadores de cada marco situacional podría ayudarnos a determinar si el marco debe ser seleccionado o no. En este sentido, se puede comparar la distribución de frecuencias relativas de los disparadores en un corpus general y la distribución de las mismas unidades en el texto que está siendo analizado. Se asume que las palabras orientadas al tema del corpus o del documento van a tener un frecuencia proporcionalmente mayor.

La tabla 2 compara las frecuencias relativas de las unidades de dos marcos situacionales diferentes en un texto médico y en la proveniente de un corpus de 5,5 millones de palabras de la Universidad Politécnica de Cataluña[3] usado como corpus de referencia. Se puede apreciar como existen diferencias entre ambos. En el caso del marco *economy* las frecuencias de sus disparadores en el corpus de referencia son mucho más altas que en el médico, lo que conduce a una media final más alta. En el caso de *Active substance* ocurre el caso contrario, las frecuencias de estos disparadores en el corpus médico son más altas que en el corpus de referencia, lo que lleva asociada una media más elevada.

ECONOMY	**TEXTO MÉDICO**	**CORPUS DE REFERENCIA**
económico.a,	0.3e-06	27e-06
economía.n	0	140.e-06
MEDIA	0.15e-06	97e-06
ACTIVE SUBSTANCE	**TEXTO MÉDICO**	**CORPUS DE REFERENCIA**
medicina.n	931e-06	107.4e-06
químico.a	173e-06	60e-06
irritante.n	15.5e-06	0
MEDIA	373.1e-06	55.8e-06

Tabla 2. Comparación de las frecuencias y medias de dos marcos situacionales.

[3] http://www.lsi.upc.edu/%7Epadro/index.php?page=nlp

Esta metodología de análisis podría ser utilizada para seleccionar los marcos más representativos de un determinado documento. El problema derivado de tal metodología es establecer los límites en los que una diferencia entre medias y frecuencias es lo suficientemente representativa como para seleccionar el marco. El hecho de que los valores de una serie de unidades léxicas en el corpus médico sea superior a la media de los mismos valores en el corpus de referencia no va a ser suficiente en muchos casos, para determinar que un marco determinado es representativo.

Nuestro problema se asemeja al que se presenta en muchos otros estudios donde es necesario comparar ciertas características de dos o más grupos de sujetos para determinar si las diferencias que se aprecian entre ambos son aparentes, o por el contrario, se debe ciertamente a diferencias significativas. Normalmente estos análisis tratan de establecer una hipótesis de partida (hipótesis nula), por ejemplo, en nuestro caso, que los valores de las frecuencias relativas de las palabras en dos corpus diferentes realmente son iguales. Entre las diferentes técnicas de evaluación, *el test t* o *test de student* analiza si las medias de dos grupos son estadísticamente diferentes la una de la otra en relación a la variabilidad de los valores de cada uno de los individuos. La metodología es diferente dependiendo del caso con el que nos encontremos. El nuestro se trata de uno de los análisis estadísticos más comunes en la práctica científica, pues es el utilizado para comparar dos muestras de grupos independientes respecto a una variable numérica. La formula aplicada en este caso es la del Test de Student para dos muestras independientes :

$$t = \frac{X_1 - X_2}{s_{X1-X2}} \qquad [1]$$

donde

$$S_{X1-X2} = \frac{s_1^2 - s_2^2}{n} \qquad [2]$$

El numerador de la fórmula [1] es fácil de computar ya que se trata de una diferencia entre las medias (la proveniente del texto a analizar y el corpus usado como referencia). El

denominador calcula la varianza de cada grupo y lo divide por el número de individuos de cada grupo. El número de individuos en cada grupo va a ser el mismo ya que tomamos aquellos disparadores con frecuencia mayor a cero en el texto a analizar rechazando aquellos disparadores con frecuencia cero.

Una vez que se ha calculado el *valor-t* se comprueba en una tabla de significación si su ratio supera a los indicados en la tabla, lo que nos llevaría a afirmar que la diferencia que existe entre los valores del grupo de disparadores en ambos corpus no es debida a la casualidad y realmente ambos grupos se diferencian. Esta diferencia es debida a que el grupo de palabras analizado está orientado significativamente al tema del documento y no sigue lo que se esperaría en un corpus general.

La aplicación del *test t* o *test de Student* exige que las observaciones en cada grupo provengan de una distribución normal con una variabilidad semejante. Realmente nuestro caso se trata de una distribución binomial, pero por el Teorema Central del Limite podemos aproximarlo a una normal, es decir, una distribución binomial converge hacia una distribución normal cuanto más grande es el número de observaciones (frecuencias de las palabras) extraídas del corpus. Este hecho nos permite la utilización de un método paramétrico en la selección de marcos situacionales.

2.1.2. Fase de entrenamiento.

De los 7730 documentos de nuestro corpus, aproximadamente un 75% fueron seleccionados aleatoriamente para entrenar nuestro sistema: 2654 textos médicos, 2105 textos sobre noticias, 708 documentos sobre ciencia y 322 sobre artículos de productos de alimentación y consumo en general. Esta fase contempla el análisis de los documentos tal como ya se ha explicado en el punto anterior. A partir de la distribuciones de sus palabras se computaron los marcos situacionales que resultaban significativos tras aplicar el test-t. Esto proporciona una lista de marcos situacionales relativo a cada uno de los textos. El conjunto de marcos situacionales será usado como rasgos indicadores de la pertenencia del documento a un determinado dominio temático o no.

Una vez nos hicimos con la lista de marcos situacionales asociados cada documento, se computo el *clasificador bayesiano Naive.* Este clasificador se usa cuando queremos clasificar una instancia descrita por un conjunto de atributos ($a_i's$), en nuestro caso, el conjunto de marcos situacionales asociados a cada uno de los documentos, en un conjunto finito de clases (V). Este clasificador asume que los valores de los atributos son condicionalmente independientes dado el valor de la clase, por lo que:

$$V_{MN}=\text{argmax}_{vj} \; v(P(v_i)\prod_i P(a_i|v_j)) \quad [3]$$

Los valores P(ai |vj) se estiman con la frecuencia de los datos observados. Aquella categoría que maximice la formula será tomada como la más apropiada.

3. Resultados y evaluación.

De esta manera el sistema fue evaluado utilizando el 25% restante de los documentos extraídos de Internet: 969 médicos, 242 de ciencia, 590 de prensa y 140 sobre productos de consumo. Los resultados sobre el conjunto de documentos son los siguientes:

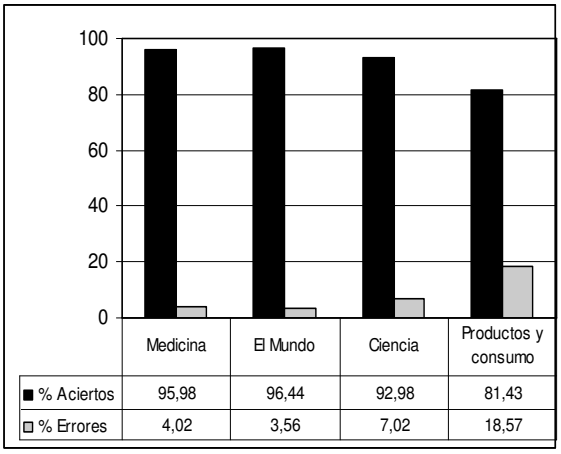

Tabla 3. Porcentaje de aciertos y errores en cada ámbito temático.

Conforme a estos resultados, obtenemos un 94.6% de precisión en la identificación general de estos cuatro tipos de documentos.

4. Discusión.

Los resultados demuestran la plausibilidad de nuestro procedimiento. El tema de un determinado documento se puede determinar desde los marcos situacionales que concurren en él. La aplicación de uno de los clasificadores más simples como es el clasificador bayesiano *naive* ofrece buenos resultados y buen rendimiento.

Es de destacar que el tema *Productos y consumo* es el que proporcionalmente da más errores (81.43% de acierto). No obstante, hay que tener en cuenta que es aquel al que se le ha dedicado menos recursos de entrenamiento y que puede confundirse con los documentos de *elmundo.es* ya que éste no sólo incluye noticias, sino reportajes y secciones variadas.

Este procedimiento podría usarse en vez de la clasificación que propone *Latent Semantic Indexing.* Al ser FrameNet un recurso creado manualmente, solventa los errores de crear grupos conceptuales automáticamente.

5. Trabajo futuro.

Nuestro trabajo contempla la extensión de los temas más allá de los cuatro propuestos aquí, lo que implica una ampliación del corpus y la investigación con otros clasificadores lineales más sofisticados que el clasificador bayesiano *naive*.

En esta línea, también sería oportuno probar el grado de precisión a la hora de clasificar textos dentro una misma temática usando marcos situacionales. Quizá esta nueva metodología permita la discriminación de subtemas dentro de un mismo dominio o la clasificación entre diferentes géneros o estilos de lenguaje como coloquial frente a formal o prensa o informativo frente a literario, etc. Diferentes estilos de lenguaje se valen de recursos lingüísticos diferentes por lo que quizá un análisis mediante FrameNet sea viable para diferenciar estilos aunque traten un mismo tema general.

6. Conclusiones.

FrameNet ofrece la posibilidad de poder usarse en tareas semánticas como la categorización textual. Este recurso ofrece un análisis del lenguaje basado en cómo los hablantes entienden y usan el lenguaje para hablar sobre el mundo. De esta manera, la categorización textual es factible desde el análisis de los marcos situacionales que aparecen en un texto. Los marcos situaciones sirven de enlace entre los términos que aparecen en un documento y el tema general. La única limitación es que FrameNet no se encuentra disponible para todas las lenguas, por lo que las técnicas de categorización como *Latent Semantic Index* que agrupa las palabras por conceptos automáticamente siguen siendo las más factibles.

Bibliografía

Crespo Miguel, M. y Buitelaar, P. 2008. "Domain-Specific English-To-Spanish Translation of FrameNet". *Proceedings of LREC (Language Ressources and Evaluation Conference).*

Fillmore, Charles J. 1982. Frame semantics. En Linguistics in the Morning Calm, Seúl: Hanshin Publishing Co., págs.111-137.

Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di Semántica* 6.2:222-254.

Gómez Hidalgo, J.M., Cortizo Pérez, J.C., Puertas Sanz, E., Buenaga Rodríguez, M. de. *Experimentos en indexación conceptual*. In Gutiérrez, J.M., Martínez, J.J., Isaías, P. (Eds) Actas de la Conferencia Ibero-Americana WWW/Internet 2004, Madrid, Spain, October, 7-8, 2004, pp. 251-258.

Huang, Y. (2003). *Support vector machines for text categorization based on latent semantic indexing*. Technical report, Electrical and Computer Engineering Department, The Johns Hopkins University.

Petridis, V., V.G. Kaburlasos, P. Fragkou, y A. Kehagias, 2001. Text classification using the s-FLNMAP neural network. *Proceedings of the 2001 International Joint Conference on Neural Networks*.

Ruppenhofer, J., Ellsworth M., Petruck, M. R. L., Christopher R. Johnson, Jan Scheffczyk. 2006., *FrameNet II: Extended Theory and Practice*.

Sebastiani. F. 2002. Machine Learning in automated text categorization. *ACM Comput. Surv*. 34(1): 1-47.

Shehata, S., Karray, F. and Kamel, M., "A concept-based model for enhancing text categorization", 13th, ACM KDD, August, 2007, pp. 629-637.