Clasificación de documentos escritos en euskara: impacto de la lematización

O. Arregi e I. Fernández
Universidad del País Vasco
Euskal Herriko Unibertsitatea
acparuro@si.ehu.es idoiazum@si.ehu.es

Resumen La clasificación de documentos escritos en euskara es un área en la que podríamos decir, está todo por hacer. Este trabajo pretende establecer las bases de la categorización para, en adelante, ir mejorando los algoritmos y las técnicas a aplicar teniendo en cuenta las características propias de la lengua[1]. El corpus utilizado corresponde a los artículos de prensa publicados durante dos meses de 1999 en el diario "Euskaldunon Egunkaria"; y los algoritmos aplicados han sido Naive Bayes y Winnow. A la vista de los resultados y para intentar mejorarlos, se ha utilizado la técnica de lematización. Los resultados obtenidos nos han demostrado que en general es importante aplicar alguna técnica de reducción de la representación de los documentos a clasificar. Además se observa que la técnica de lematización mejora sensiblemente los resultados. Por otra parte, en cuanto a los algoritmos se refiere, se puede decir que *naive bayes* responde mejor con el corpus lematizado, es decir, con la información más concentrada y winnow al contrario, no se ve afectado por el ruido en su respuesta.

1 Introducción

El rápido desarrollo de las nuevas tecnologías y en particular de Internet, nos lleva a disponer de gran cantidad de información en soporte electrónico. El uso y sobre todo el acceso a esta información es a menudo un proceso tedioso e incluso en algunos casos infructuoso. Existen diversas técnicas que facilitan el acceso a semejantes volúmenes de información y que han dado lugar a diferentes aplicaciones con el fin de ordenar, clasificar, buscar, y en general tratar esta información.

De entre las diferentes técnicas existentes, la que aquí nos ocupa es la clasificación o categorización de documentos, que consiste en etiquetar los textos escritos en lenguaje natural con una categoría elegida de entre un conjunto de categorías temáticas previamente establecido.

Durante años la técnica más utilizada para clasificar documentos y extraer la información necesaria de los mismos ha sido el trabajo humano. Era labor de documentalistas estructurar y ordenar la información para poder acceder a la misma de manera racional. El transcurso del tiempo y sobre todo la rápida expansión de Internet ha llevado a desarrollar y mejorar las técnicas básicas de procesamiento semiautomático de la información, y entre ellas las técnicas de categorización de documentos.

Por otra parte, la realidad de las lenguas minorizadas como es el caso del euskara, es diferente a la de otras lenguas más extendidas como pueden ser el inglés o el castellano.

Los trabajos de clasificación de documentos escritos en euskara son mínimos por no decir que no existen. No hay un solo trabajo publicado al respecto y hasta ahora al menos, el único diario escrito íntegramente en euskara realiza la categorización de documentos de manera manual. Del mismo modo, los pocos portales de Internet que ofrecen la información en euskara clasifican las páginas a mano.

Esto nos lleva a plantearnos el problema desde el inicio, es decir, no hay un corpus básico preparado para poder trabajar sobre él y tampoco hay estudios previos que permitan vislumbrar por dónde atacar el problema para poder mejorarlo, o simplemente para poder comparar los resultados obtenidos.

Debido a esto, debemos plantearnos el trabajo de manera gradual. Comenzaremos con la conformación del corpus y el establecimiento

[1] Este trabajo ha sido parcialmente subvencionado por el MCyT (Proyecto Hermes, 8/DG00141.226-14247/200).

de las categorías, para así poder realizar un primer análisis con los algoritmos básicos y establecer los resultados iniciales. Una vez finalizada esta primera fase, se tratará de analizar las especificidades del euskara para ver como mejorar los resultados que se vayan obteniendo.

2 Antecedentes y estado actual

Las técnicas de clasificación o categorización de documentos se han encarado desde diferentes aproximaciones (modelos probabilísticos, modelos simbólicos u otros como "boosting" o SVM) obteniendo resultados diversos. En [1], Sebastiani presenta un interesante tutorial sobre clasificación automática de documentos que proporciona un panorama de los diferentes métodos, así como algunos resultados obtenidos. La mayoría de estos métodos se basan en las técnicas de aprendizaje automático, en las que se distinguen dos fases: la de *entrenamiento* en la que se obtiene una generalización inductiva del conjunto de documentos que se utilizan para el aprendizaje del sistema, es decir, la fase donde se genera el clasificador; y la de *test* que se encargará de evaluar la efectividad del mismo. Para la primera fase es necesario un conjunto de textos clasificados manualmente con el que poder realizar el entrenamiento. El sistema, analiza los documentos clasificados y extrae las características correspondientes a cada una de las categorías a las que los textos pertenecen. En concreto, en la mayoría de los algoritmos clásicos de clasificación, el sistema determina la importancia de cada palabra para cada categoría, dándole un valor o peso que será utilizado en la fase de test para determinar la clase a la que los documentos pertenecen.

Otra cuestión a tener en cuenta es si la clasificación es multi-etiqueta (cada documento puede pertenecer a más de una categoría) o no (sólo pertenece a una de ellas). En [2] se observa que los resultados obtenidos en los sistemas multi-etiqueta son mejores que aquellos correspondientes a una sola etiqueta.

La representación de los documentos es otro apartado que conviene apuntar; la mayoría de los estudios efectuados hasta ahora realizan una clasificación semántica basada en las palabras que componen el documento a clasificar, *bag of words*, [3], [4] y [5], si bien existen algunos trabajos [6] y [7] que tratan de analizar el impacto que puede producir la utilización de otro tipo de características para representar el texto que se pretende categorizar, como pueden ser los lemas, los términos multipalabras o las frases. En [8] y [9] en concreto, se presenta cómo puede influir la utilización de algoritmos de stemming en la clasificación de documentos.

Por último, el gran tamaño de la información a procesar es un problema en los sistemas de clasificación y es habitual aplicar alguna técnica de reducción de la representación del corpus. Este aspecto se analiza con más detalle en el apartado 4. En cualquier caso, una reducción básica consiste en representar los documentos con las palabras diferentes que lo componen pero sin repetir cada una de las ocurrencias de las mismas, o bien adjuntándole a cada palabra el número de ocurrencias si es mayor que uno.

3 Especificidad del euskara

El euskara es una lengua de rica flexión y aglutinante. Debido a esto, en los textos escritos en esta lengua encontramos muchas veces formas similares con un lema o raíz común. El sufijo de este lema es el que va a hacer que dos palabras sean diferentes, bien en número, en determinación y en caso, para los sustantivos, o bien en modo, tiempo, aspecto, persona y número, en el caso de las formas verbales.

En este sentido podríamos conseguir flexionar cientos de miles de formas para un mismo lema. Por ejemplo, del lema *txori* (pájaro) podemos obtener:

txoriA (el pájaro),
txoriAK (los pájaros),
txoriArI (al pájaro),
txoriArEN habiA (el nido del pajaro),
txoriArEN-A (el (nido) del pajaro),
txoriArEN-A-rI (al (nido) del pajaro), ...

Como se muestra en [10] combinando número, determinación y caso, y aglutinando hasta dos genitivos, se pueden obtener hasta cerca de medio millón de formas a partir de un solo lema, si bien las formas habituales son unas decenas.

Esta cuestión resulta importante para el problema de clasificación que nos ocupa; lo que realmente contiene información semántica no es tanto la palabra sino el lema que le corresponde. Si consideramos cada palabra como la secuencia de caracteres entre dos espacios, la relación de esa palabra con la categoría a la que pertenece el documento, obtendrá un peso según las veces que se repita en el texto. Cuanto mayor sea la presencia de esa palabra en un texto, mayor será el peso que se le asignará. Por

tanto, en el caso de basarnos en las palabras, aunque un mismo lema se repita en numerosas ocasiones, al estar presente en diferentes formas, puede llegar a no considerarse importante.

Es por tanto que, a priori, la lematización parece un proceso interesante en cuanto que reduce la dimensión de la información a tratar e incluso puede producir una mejora en la eficiencia del sistema.

4 Corpus utilizado

El corpus que se ha utilizado en este trabajo, proviene del diario *"Euskaldunon Egunkaria"* http://www.egunkaria.com/ único diario escrito totalmente en euskara. Los textos que utilizamos corresponden a los artículos de prensa publicados durante los meses de enero y febrero de 1999. Una característica de este corpus es que todos los documentos que contiene son de la misma época y del mismo estilo, es decir, escritos por las mismas personas. Esto puede ser un inconveniente a la hora de generalizar los resultados obtenidos dada la homogeneidad del corpus.

El corpus se compone de 5.809 documentos clasificados en 7 categorías (economía, europa, sociedad, deporte, cultura, mundo y política), correspondientes a otros tantos artículos de prensa. Del corpus se han obtenido dos conjuntos de documentos, uno para el entrenamiento (4357 textos, el 75%) y otro para test (1452 textos, el 25%), respetando esta proporción en cada una de las siete categorías. La tabla 4.1 presenta las características del corpus.

Categoría	Nº docum. Aprendizaje	Nº docum. Test	Palabras diferentes
economía	330	109	13.067
europa	367	123	15.205
sociedad	694	231	24.411
deporte	1294	431	27.908
cultura	568	189	28.102
mundo	314	105	16.122
política	790	264	23.108
Total	**4357**	**1452**	**85.364**

Tabla 4.1. Características del corpus

5 Reducción de la dimensión

La representación de los documentos del corpus es un tema a tener en cuenta en el problema de clasificación. Si los textos se representan mediante las palabras que lo componen, la dimensión del corpus suele ser habitualmente alta. Esto puede ser problemático en la clasificación inductiva, por lo que es importante aplicar alguna técnica para reducir la dimensión de la representación. De esta forma se obtienen dos ventajas: por un lado, se reduce el tiempo de ejecución tanto en la fase de aprendizaje como en la de test, y por otro se elimina el ruido de los documentos, es decir, las palabras que no ofrecen información útil para la clasificación y pueden distorsionar los resultados. En cualquier caso, el objetivo consiste en incrementar la eficiencia sin disminuir la precisión e incluso en algunos casos mejorarla.

Esta reducción puede realizarse de dos formas diferentes: bien eliminando palabras o características que se consideren de poco valor semántico (dado que la clasificación que se va a realizar es semántica) o bien reparametrizando el texto, es decir, sustituyendo algunas palabras por otras que las representen (lematización, sinonimia, hiperonimia, ...).

En este trabajo se ha optado por llevar a cabo los dos tipos de reducción de la dimensionalidad. Por una parte, se ha diseñado una lista de palabras a eliminar por su bajo valor semántico (*stopword list*), y por otra, se han sustituido todas las palabras por sus lemas (lematización), reduciéndose así el número total de características (lemas) diferentes en cada documento. En este caso, es posible aplicar también un *stopword list* de lemas para evitar aquellos con bajo valor semántico.

5.1 Stopword list

La primera lista de palabras a eliminar se ha generado teniendo en cuenta la frecuencia de las mismas en los documentos de entrenamiento. Se ha probado con diferentes listas y finalmente se ha optado incluir una palabra en la lista si aparece en más de 30 documentos (en más de 50 para el caso de los lemas) y en menos de tres; en estos casos se considera que la palabra no tiene valor discriminatorio.

Otra segunda opción tenida en cuenta es la frecuencia de las palabras en cada una de las categorías, en este caso se consideran características no validas aquellas palabras (o lemas) que aparecen en todas o casi todas las categorías. Para las pruebas que presentamos, se han eliminado las palabras (lemas) que se encuentran en más de una o dos categorías.

5.2 Lematización

Como hipótesis de trabajo, hemos supuesto que la técnica de lematización nos permite mantener la misma información semántica de los textos a tratar, disminuyendo el tamaño de los documentos a procesar. Además, suponemos, que al sustituir una palabra por su lema, estamos concentrando la información semántica dándole el peso real a cada uno de los lemas que aparecen, de manera que podríamos mejorar la eficiencia del clasificador. En [11] se utiliza el lematizador para desarrollar un buscador de textos en euskara en la web, y la mejora obtenida gracias a él es importante.

Para obtener los lemas correspondientes a cada palabra, se ha utilizado la herramienta diseñada por el grupo IXA (http://ixa.si.ehu.es) que obtiene para cada palabra del documento, el lema que le corresponde así como la categoría morfosintáctica de la misma. En [12]se presentan las características de este lematizador.

Para conseguir una reducción del tamaño del corpus utilizando esta técnica, no basta sustituir una palabra por su lema o raíz (ocuparía lo mismo en número de características), sino que se debe modificar la representación del documento. La forma de hacerlo consiste en representar cada lema junto con el número de apariciones del mismo a lo largo del documento y no repetir el lema de la palabra en cada ocurrencia. Esta apreciación no es exclusiva para los lemas, puesto que ocurre lo mismo con las palabras.

Este sistema reduce el número de características diferentes de cada categoría a más de la mitad. La tabla 5.1 muestra esta reducción.

	Total palabras	**Palabras diferentes**	**Lemas diferentes**
Economía		13.067	5.664
Europa		15.205	6.818
Sociedad		24.411	9.947
Deporte		27.908	14.926
Cultura		28.102	13.628
Mundo		16.122	5.415
Política		23.108	9.262
Total	799.337	85.364	38.566

Tabla 5.1. Reducción del corpus

Una vez lematizado el corpus, existe la posibilidad de establecer alguna lista de lemas a eliminar. Para ello, se pueden utilizar los mismos parámetros usados en el caso de las palabras, es decir, la frecuencia de los lemas en los documentos o en las categorías, o bien se puede partir de la información sintáctica producida por el lematizador y eliminar aquellos lemas correspondientes a alguna categoría sintáctica concreta. En nuestro caso, se ha realizado una prueba teniendo sólo en cuenta los *nombres*, *verbos*, *siglas* y *adjetivos*, y dejando de lado el resto.

6 Algoritmo de clasificación

En esta sección presentaremos los algoritmos de clasificación que hemos utilizado para realizar el estudio comparativo sobre el corpus del diario *"Euskaldunon Egunkaria"*.

La herramienta en la que nos hemos basado es el clasificador de propósito general SNoW (Sparse Network of Winnows) [13].

De los tres posibles algoritmos que implementa la aplicación, hemos elegido dos de ellos, Winnow y Naive Bayes, por ser los que mejores resultados ofrecen.

6.1 Naive Bayes

En este método, la clasificación se efectúa en base a la probabilidad de que un documento pertenezca a una categoría dada.

$$P(c_i|d_j)=\frac{P(c_i)P(d_j|c_i)}{P(d_j)}$$

donde: c_i = categoría *i*; y d_j = documento *j*.

En concreto, en el algoritmo de Naive Bayes, se asume que las características de los documentos son independientes entre sí, de tal forma, que la estimación de la probabilidad $P(d_j|c_i)$ se puede obtener multiplicando las probabilidades de las palabras que pertenecen a dicho documento:

$$P(d_j|c_i) = \Pi_{k=1}{}^{r} P(w_{kj}|c_i)$$

siendo w_{kj} = palabra *k* del documento *j*

6.2 Positive Winnow

Este algoritmo utiliza el corpus de aprendizaje, donde cada documento está etiquetado con una o más categorías, para obtener una representación de las categorías, o lo que es lo mismo, para aprender un vector de pesos. Estos pesos se utilizarán más adelante, en la fase de test, para clasificar los nuevos documentos.

La actualización de pesos sólo se realiza cuando se comete un error en la predicción y para ello se utilizan dos parámetros: *promotion* ($\alpha > 0$) y *demotion* ($0 < \beta < 1$); además del *threshold* θ. Sea $A_t = \{i_1, ..., i_m\}$ el conjunto de características activas de una categoría t y w_i^t el peso de la característica i en la categoría t. Si el

algoritmo predice 0, es decir, $\sum_{i \in At} w_i^t \leq \theta_t$ y la etiqueta correspondiente al documento es 1, los pesos activos del mismo se actualizan de manera multiplicativa: $\forall i \in A_t, w_i^t \leftarrow \alpha . w_i^t$ Si la predicción es al contrario, el algoritmo predice 1 y la categoría era 0, $\sum_{i \in At} w_i^t > \theta_t$, los pesos activos se actualizan con el parámetro *demotion*: $\forall i \in A_t, w_i^t \leftarrow \beta . w_i^t$ El resto de los pesos no se modifican.

Este algoritmo se caracteriza por ser robusto frente a la presencia de ruido [2]

7 Experimentos

7.1 Descripción de los experimentos

Los experimentos que hemos realizado se agrupan en tres apartados dependiendo de los criterios utilizados para reducir la dimensión del corpus.

Las pruebas se han realizado sobre el corpus original formado por palabras y el corpus lematizado formado por lemas.

7.1.1. Criterio semántico

La elaboración de la lista de palabras a eliminar se ha realizado teniendo en cuenta la frecuencia de aparición de las mismas en el corpus de entrenamiento. Cada una de estas pruebas proporciona dos tipos de resultados, uno correspondiente a las formas (palabras) y otro a los lemas. Para cada una de ellas, se obtiene el valor que proporcionan los dos algoritmos utilizados (winnow y bayes).

Prueba 1: Esta prueba se ha realizado tomando como base el corpus completo sin aplicar ninguna reducción.

Prueba 2: la clasificación de los documentos del corpus se ha realizado eliminando las características de mayor frecuencia y aquellas que aparecen en menos de tres documentos.

7.1.2. Criterio sintáctico

Los criterios sintácticos se aplican al corpus conformado por lemas, eliminando algunos lemas dependiendo de su categoría sintáctica. Cada prueba ofrecerá dos resultados dependientes del algoritmo utilizado.

Prueba 3: se tienen en cuenta las categorías sintácticas *nombre*, *verbo*, *adjetivo*, y *sigla* y se elimina el resto.

Prueba 4: se realizará la clasificación tomando en consideración solamente los nombres dada la importancia que esta categoría adquiere [14].

7.1.3. Criterio mixto

Este criterio conjuga la opción semántica con la sintáctica.

Prueba 5: partiendo del corpus utilizado en la prueba 3 (*nombre*, *verbo*, *adjetivo*, y *sigla)* se eliminan las características de mayor frecuencia y aquellas que aparecen en menos de tres documentos.

7.2 Evaluación

Los resultados obtenidos para las pruebas presentadas se resumen en la tabla 7.1

	formas		**lemas**	
	winnow	bayes	winnow	bayes
Prueba1	42.36	62.53	88.29	82.23
Prueba2	89.74	87.74	87.19	88.15
Prueba3			88.15	86.29
Prueba4			86.98	90.08
Prueba5			85.88	86.85

Tabla 7.1. Resultados de la clasificación. Recall

La figura 7.1. muestra la mejora obtenida al lematizar el corpus. En el caso del algoritmo winnow, la ganancia es sustancial. Esta mejora se debe a la concentración de características. El beneficio es superior con el algoritmo winnow, dado que éste es menos sensible al ruido que naive bayes.

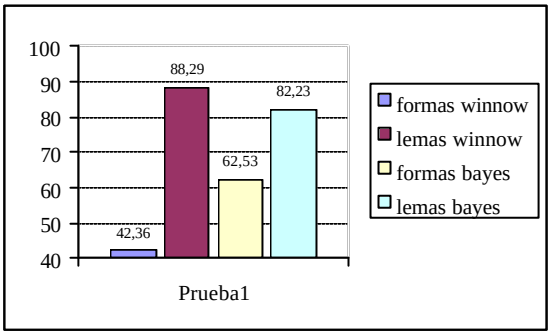

Figura 7.1. Mejora debida al lematizador.

Las figuras 7.2. y 7.3. muestran la influencia de la aplicación de una lista de palabras a eliminar en ambos algoritmos aplicados al corpus compuesto tanto por palabras (formas) como por lemas.

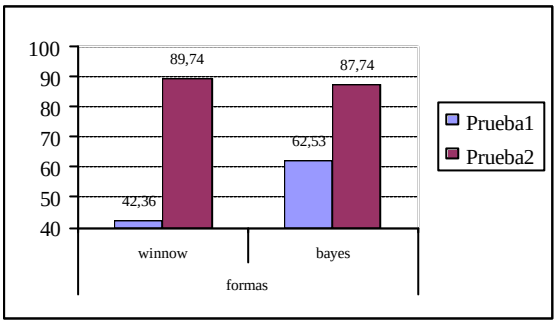

Figura 7.2.Influencia del stopword list (formas).

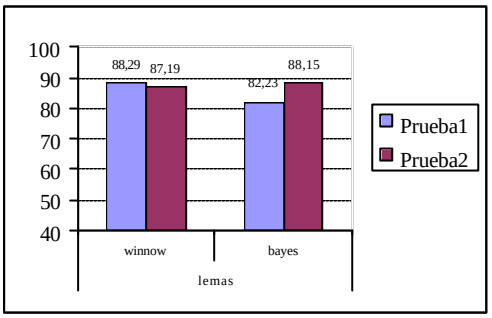

Figura 7.3.Influencia del stopword list (lemas).

La figura 7.4. muestra los diferentes resultados obtenidos con el corpus lematizado. Cabe destacar el valor logrado teniendo solamente en cuenta la categoría sintáctica nombre. La diferencia de una prueba a otra nos indica la importancia de reducir de una u otra manera el tamaño del corpus.

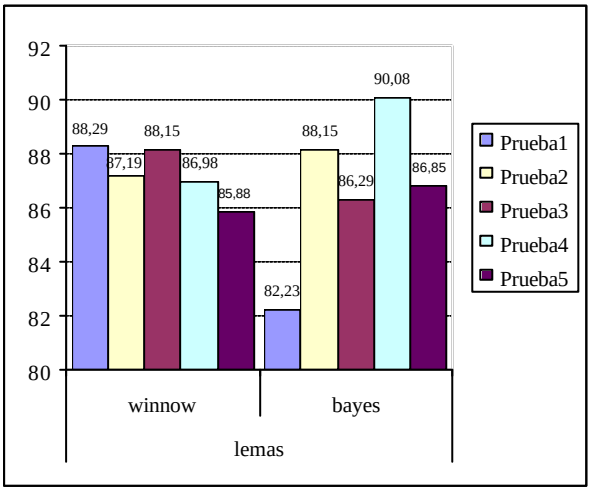

Figura 7.4. Resultados del corpus lematizado

8 Conclusiones

Este trabajo establece los resultados de la clasificación de documentos escritos en euskara utilizando dos algoritmos básicos como son naive bayes y winnow.

Como mejora del sistema se aplica la técnica de lematización para así obtener un corpus más reducido sin perder la información semántica necesaria para clasificar los documentos adecuadamente.

A la vista de los resultados, la hipótesis de trabajo resulta ser realista, en general la lematización mejora los resultados y permite aprovechar la información morfosintáctica que proporciona para reducir la dimensión de la representación del corpus.

Por otra parte cabe reseñar los buenos resultados que se obtienen con el algoritmo naive bayes al reducir el tamaño del corpus, bien por la lematización o bien por la aplicación de la stopword list.

Como dato a destacar nos fijamos en la capacidad de los nombres para aportar la información semántica necesaria al clasificador, obviando el resto de categorías sintácticas.

9 Referencias

[1] F. Sebastiani. 2000. Machine Learning in Automated Document Categorization. *The 18th International Conference on Computational Linguistics. Tutorials,* Nancy, Francia, 2000.

[2] R. E. Schapire y Y. Singer. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, vol. 39, nº 2/3, pag. 135-168. 2000.

[3] I. Dagan, Y. Karov y D. Roth. Mistake-Driven Learning in Text Categorization. *EMNLP '97, 2nd Conference on Empirical Methods in Natural Language Processing*, August 1997

[4] R. E. Schapire, Y. Singer y A. Singhal. Boosting and Rocchio applied to text filtering. *Proceedings of {SIGIR}-98, 21st {ACM} International Conference on Research and Development in Information Retrieval.* ACM Press, New York, US", pag. 215--223, 1998.

[5] P.P.T.M. van Mun. Text Classification in Information Retrieval using Winnow. url: citeseer.nj.nec.com/133034.html

[6] M. Grobelnik y D. Mladenic. Efficient text categorization. *Text Mining workshop on the 10th European Conference on Machine Learning ECML98*. 1998.

[7] D. Mladenic y M. Globelnik. Word sequences as features in text learning. *In*

Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98), Ljubljana, Slovenia, 1998.

[8] R. Krovetz. Viewing morphology as an inference process. *In Proceedings of ACM-SIGIR93*, pag. 191--203. 1993.

[9] E. Riloff. Little words can make a big difference for text classification. *In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pag. 130--136, 1997.

[10] Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Maritxalar M., Sarasola K. XUXEN: A Spelling Checker/Corrector for Basque Based on Two-Level Morphology. *Proceedings of ANLP'92*, pag. 119-125. Povo Trento. 1992

[11] I. Aizpurua, I. Alegria, N. Ezeiza. GaIn: un buscador Internet/ Intranet avanzado para textos en euskera. *Actas del XVI Congreso de la SEPLN Universidade de Vigo*, 26-28 septiembre de 2000.

[12] Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. *Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages*. COLING-ACL'98, Montreal (Canada). August 10-14, 1998.

[13] A. J. Carlson, C. M. Cumby, J. L. Rosen y D. Roth. SNoW. UIUC Tech report UIUC-DCS-R-99-210. University of Illinois, Urbana, Illinois, 1999.

[14] A. Chowdhury y M. C. McCabe. Improving Information Retrieval Systems using Part of Speech Tagging. 1997.
url: citeseer.nj.nec.com/256084.html