IÑAKI ALEGRIA
MARIA JESUS ARANZABE
AITZOL EZEIZA
NEREA EZEIZA (jipecran@si.ehu.es)
RUBEN URIZAR

Informatika Fakultatea, Universidad del País Vasco.
649 P.K. 20080 Donostia. Euskal Herria.

Robustez y Flexibilidad de un Lematizador/Etiquetador

Abstract

This paper presents the lemmatiser/tagger for Basque EUSLEM. The initial design of the lemmatiser has been adjusted to face two main goals: on the one hand, the improvement of the morphological analyser, reducing the amount of errors and adding the option of using specialised lexicons, and, on the other hand, improving the recall of the disambiguation task.

On the one hand, we have added a module to treat multiword units, which processes lexical units, such as locutions and collocations, and identifies and treats some named entities, such as dates and numerical expressions, to obtain more precise results for standard words. On the other one, the average number of interpretations in non-standard words is significantly higher than in standard words. Given that some of these interpretations might be discarded using local non-contextual information, we have designed some heuristics to enhance the precision of morphological analysis adding a small amount of error to the result.

All of these adaptations have an effect on the disambiguation task, avoiding around 20% of the errors made when these improvements are not present. The enhancement is particularly meaningful in the case of non-standard word, for which 40%-60% of the errors is avoided.

Thus, based on all these proposals, we obtain a robust and high coverage lemmatiser/tagger, and, furthermore, it is flexible, because it gives the user the option of applying the necessary improvements to appropriately process the texts.

1 Introducción

En este artículo se presentan las mejoras realizadas en el lematizador/etiquetador de euskara EUSLEM. Consta de cuatro módulos principales: un analizador morfológico que incluye tratamiento de variantes dialectales y desviaciones de competencia —debidos al desconocimiento del idioma estándar— además de lematización de palabras desconocidas (Alegria *et al*., 2001); procesamiento de palabras no estándar (Alegria *et al.*, 2002); tratamiento de unidades multipalabra (Aduriz *et al*., 1996a); y por último, desambiguación morfológica combinando conocimiento lingüístico e información estadística (Ezeiza *et al*., 1998).

El diseño inicial del etiquetador se ha ido adaptando para afrontar dos objetivos, por un lado, mejorar el analizador morfológico y, por otro, mejorar la precisión para disminuir el error del módulo de desambiguación. La adaptación del analizador morfológico tiene como objetivo disminuir el número de errores cometidos en esta fase, puesto que dichos errores repercuten en todas las herramientas basadas en morfología, como desambiguación, análisis sintáctico, extracción de terminología, etc.

En cuanto a la mejora de la precisión[1] de la desambiguación, se ha realizado en dos sentidos. Por un lado, se ha integrado el procesamiento de unidades multipalabra, que además de tratar algunas unidades léxicas como locuciones y colocaciones, identifica y procesa entidades con nombre propio del tipo de fechas y expresiones numéricas, proceso que ayuda a disminuir la ambigüedad del análisis morfológico. Por otro, se ha diseñado una serie de heurísticos que, introduciendo un mínimo de error en el resultado, también ayudan a aumentar la precisión. Dichos heurísticos procesan las interpretaciones

[1] Precisión = número de interpretaciones correctas / número de total de interpretaciones. En el caso de los resultados de desambiguación presentados, es equivalente a la tasa de corrección, ya que el texto está completamente desambiguado.

morfológicas de las palabras no estándar, descartando algunas de ellas basándose en información local no contextual.

En el siguiente apartado se presenta la adaptación realizada en el analizador morfológico, a continuación, en el apartado 3 se describen los procesos incluidos para mejorar la precisión del análisis morfológico, el apartado 4 presenta los resultados globales de la desambiguación, y, por último, se evalúa la repercusión de las adaptaciones realizadas en la desambiguación y se presentan las conclusiones del trabajo.

2 El analizador morfológico

MORFEUS es un analizador morfológico robusto de euskara. Es una herramienta básica en el desarrollo actual y futuro del procesamiento de euskara. Es una herramienta básica en el desarrollo actual y futuro del procesamiento de euskara. Algunas de las aplicaciones basadas en análisis morfológico son el lematizador/etiquetador (Ezeiza *et al*., 1998), un motor de búsqueda en Intranet (Aizpurua *et al*., 2000) y un asistente para realizar versos (Arrieta *et al*., 2000).

El analizador morfológico (Alegria *et al*., 2001) consta de tres módulos: el analizador de palabras estándar, el de variantes dialectales y desviaciones de competencia y, por último, un módulo de lematización de palabras desconocidas. El diseño inicial integra dichos módulos de forma incremental, es decir, primero se analizan las palabras que siguen las reglas estándar del euskara, a continuación se intentan analizar como variantes las palabras que no han sido tratadas en el anterior paso, y por último, las palabras restantes se tratan como desconocidas.

Otra característica incorporada al analizador es la posibilidad de incorporar de forma sencilla léxicos de usuario, dando opción a la creación de léxicos especializados. Esta característica es especialmente interesante a la hora de procesar textos técnicos, y permite mantener el léxico general utilizado por el analizador independiente de las aplicaciones basadas en morfología.

El primer módulo procesa en torno al 70%-80% de los tokenes[2], con un error en torno al 0.2%-0.3%. El segundo módulo procesa, en función del tipo de texto, el 0.5%-3% de las palabras, aunque en textos escritos en alguno de los dialectos, esta proporción puede alcanzar hasta el 10% de palabras. La precisión de este módulo, por tanto, varía bastante en función de si se procesan textos dialectales o en euskara estándar, obteniendo resultados bastante más pobres en el caso de corpus estándar, como puede observarse en la tabla 1. Por último, el módulo de análisis de palabras desconocidas procesa el resto, en torno al 3%-6%.

	DT	AR	I/A	I/T	C
estándar	78.76%	81.13%	3.82	3.29	**99.75%**
variantes	0.70%	74.00%	4.14	3.32	**70.00%**
desconocidas	3.04%	100%	19.42	19.42	**99.54%**
total palabras	82.50%	81.76%	4.53	3.89	**99.51%**
total tokenes	100%	67.45%	4.53	3.38	**99.60%**

Tabla 1.- Medidas de ambigüedad y corrección del analizador incremental[3].

Realizado un estudio de los resultados del analizador, se observó que gran parte de los errores (50%-75%) ocurrían al procesar nombres propios. A continuación se proponen algunas soluciones para evitar alrededor del 50% de los errores, cuyo causante principal es el carácter incremental de la arquitectura del analizador.

Algunos de los errores se pueden evitar enriqueciendo el léxico del usuario, pero muchos de los nombres son coyunturales, principalmente en textos periodísticos, y no merece la pena incluirlos todos en el léxico. Además, si se quiere una herramienta robusta, es necesario plantear soluciones más generales para evitar este tipo de errores.

---

[2] En estos porcentajes se han incluido signos de puntuación y otros símbolos, que conforman el 8-15% de los tokenes

[3] **D** = Distribución de tokenes;
AR = Tasa de ambigüedad (Ambiguity Rate);
I/A = Interpretaciones por token ambiguo;
I/T = Interpretaciones por token;
C = Tasa de corrección (número de interpretaciones correctas / número de palabras).

En el primer paso del proceso, el analizador estándar asigna interpretaciones erróneas, básicamente cuando se obtienen interpretaciones de lemas muy cortos o de uso poco común. No obstante, al obtener (al menos) una interpretación, el proceso de análisis se detiene.

Un claro ejemplo es este tipo de interpretaciones erróneas es *Barak*. Este nombre, cuando aparece en forma básica, se interpreta como *bara*, un nombre común de baja frecuencia. Sin embargo, cuando aparece declinado, por ejemplo *Barakek* (*Barak* en caso ergativo), el analizador estándar no asigna ninguna interpretación, por lo que el analizador de palabras desconocidas lo interpreta correctamente como nombre propio, entre otras opciones.

Por lo tanto, hay que evitar análisis raros o improbables cuando la palabra comienza con mayúsculas. Para ello, hemos marcado los lemas cortos o conflictivos de baja frecuencia como raros en la base de datos. Cuando todas las interpretaciones de una palabra están marcadas como raras, el proceso continuará aplicando el siguiente módulo de análisis. Si en el siguiente paso no encuentra ninguna interpretación sin marca de rareza, entonces, se asignarán sólo las interpretaciones del analizador estándar. En el caso de lemas de baja frecuencia, las palabras escritas en mayúsculas serán analizadas por el último módulo como desconocidas y se añadirán las interpretaciones de nombres propios.

	DT	AR	I/A	I/T	C
estándar	78.65%	81.54%	3.92	3.38	**99.91%**
variantes	0.67%	72.92%	4.83	3.79	**91.67%**
desconocidas	3.18%	100%	19.46	19.46	**99.56%**
total palabras	82.50%	82.19%	4.66	4.00	**99.83%**
total tokenes	100%	67.80%	4.66	3.48	**99.86%**

Tabla 2.- Medidas de ambigüedad y corrección del analizador mejorado.

Para mejorar los resultados del analizador de variantes, hemos limitado el número de reglas morfofonológicas aplicadas para obtener una interpretación. Si en todas las interpretaciones se supera el umbral de reglas aplicables, la palabra será tratada como desconocida, descartando el resto de interpretaciones.

Tras implementar estas propuestas, creemos que los resultados son satisfactorios (tabla 2), pues, como resultado de la relajación de las restricciones del analizador, se evita el 50% de los errores.

3 Mejora de la precisión

Como se observa en la tabla 2, la ambigüedad es bastante alta, especialmente en el caso de las palabras desconocidas, pero también para las palabras estándar. Se han añadido dos procesos para disminuir dicha ambigüedad y aumentar la precisión del análisis morfológico: el primero consiste en identificar términos multipalabra, y el segundo procesa las palabras no estándar para descartar algunas de sus interpretaciones utilizando información local no contextual.

El tratamiento de unidades multipalabra identifica algunas unidades léxicas como locuciones, colocaciones y palabras compuestas. Además, se identifican y tratan entidades con nombre propio del tipo de fechas y expresiones numéricas. Las unidades tratadas son seguras, con lo cual se consigue descartar algunas interpretaciones de sus componentes y se reduce en cierta medida tanto la tasa de ambigüedad como el número de interpretaciones por palabra.

Por otro lado, en el caso de las palabras no estándar, el número de interpretaciones es superior al de las estándar. Esta ambigüedad es, de alguna forma, artificial por la forma en la que se obtienen las interpretaciones, ya que para una misma interpretación morfológica el analizador puede proponer más de un lema. Dado que el desambiguador no tiene por objetivo decidir entre diferentes lemas, conviene resolver esta ambigüedad cuanto antes. Además, puede proponer interpretaciones morfológicas muy similares variando mínimamente el lema. Esto implica mayor dificultad para resolver la ambigüedad.

Para evitar esto, se han diseñado una serie de heurísticos que, combinados, reducen el número de interpretaciones añadiendo un error mínimo. En el caso de las variantes dialectales, a la hora de descartar análisis, se tiene en cuenta el número de reglas morfofonológicas aplicadas, dando prioridad a las interpretaciones con menor número de reglas para cada categoría-subcategoría.

Para las palabras desconocidas, ha sido necesario tener en cuenta diversos tipos de información, puesto que el número de interpretaciones es mucho mayor de la media (en torno a 20 interpretaciones por palabra, aunque algunas de ellas pueden tener incluso 100 interpretaciones). Los heurísticos utilizados son los siguientes (Alegria *et al*., 2002):

- Desambiguación tipográfica para descartar algunos lemas y/o interpretaciones morfológicas.
- Desambiguación de palabras derivadas, para contrarrestar la sobregeneración del analizador.
- Identificación y desambiguación de nombres propios no incluidos en el léxico.
- Desambiguación de lemas basada en información lingüística y estadística.

El resultado de aplicar estos procesos se puede apreciar en la tabla 3. Como se puede observar, la tasa de ambigüedad de las palabras estándar ha descendido en 2 puntos y el número medio de interpretaciones de las palabras desconocidas ha descendido de 19 a 4. Es cierto que se ha añadido un error del 11% pero teniendo en cuenta los diferentes niveles de etiquetado, que se describen a continuación, la tasa de corrección es del 95% al 99% en función del conjunto de etiquetas utilizado.

	AR	I/A	I/T	C
estándar	79.51%	3.87	3.28	**99.91%**
variantes	59.57%	3.46	2.47	**91.49%**
desconocidas	92.11%	3.98	3.74	**88.16%**
total palabras	79.83%	3.87	3.30	**99.39%**
total tokenes	65.86%	3.87	2.89	**99.43%**

Tabla 3.- Resultado de los tratamientos.

4 Desambiguación morfosintáctica

De cara a afrontar la desambiguación, se ha descrito un sistema de etiquetas multinivel general independiente de la tarea de desambiguación (Agirre *et al*., 1995). El primer nivel se define mediante la categoría morfológica de la palabra, que consta de 20 etiquetas, el segundo incluye también la subcategoría, que consta de 45 etiquetas, en el tercero se añade otro tipo de información interesante como el caso en que aparece declinada la palabra o el tipo de verbo que se está tratando y, por último, el cuarto nivel incluye toda la información dada por el analizador morfológico.

La complejidad de esta tarea varía en función del nivel de etiquetado que se utilice. Como se ve en la tabla 4, la tasa de ambigüedad es bastante alta, especialmente en los niveles 3 y 4, aunque para muchas aplicaciones el nivel 2 es el más adecuado. A la otra de presentar los resultados, nos centraremos exclusivamente en este último.

	AR	I/A	I/T	C
1º nivel	36.05%	2.32	1.48	**99.94%**
2º nivel	42.02%	2.34	1.56	**99.75%**
3º nivel	64.01%	3.13	2.36	**99.68%**
4º nivel	65.86%	3.87	2.89	**99.43%**

Tabla 4.- Ambigüedad en función del nivel de etiquetado.

Dado que los recursos con los que contamos no son tan amplios como se pueden obtener para otros idiomas como el inglés, la desambiguación por métodos estocásticos da unos resultados bastante peores que para otros idiomas, dado el tamaño reducido del corpus de entrenamiento (unas30000 palabras etiquetadas manualmente). Por ello, estimamos interesante la combinación de métodos.

A la hora de diseñar el desambiguador se han combinado, por un lado, una gramática de restricciones (Karlsson *et al*., 1995) para el euskara (Aduriz *et al*., 1996b), y, por otro, se ha utilizado una implementación de modelos ocultos de Markov de primer grado (Armstrong *et al*., 1995).

El primer método reduce ostensiblemente la ambigüedad introduciendo un mínimo error, con lo que los modelos de Markov obtienen mucho mejores resultados. Con ello, la precisión de la desambiguación en el segundo nivel es del 95.42%, lo cual, comparado con el 85.01% obtenido utilizando los modelos ocultos de Markov exclusivamente, supone una mejoría del 70%.

5 Evaluación y conclusiones

En el gráfico 1 se muestra la aportación de cada una de las mejoras propuestas a lo largo de este artículo[4]. Los resultados del analizador morfológico propuesto no mejoran significativamente los resultados del original, a pesar de que parte de mejores niveles de corrección. Esto es debido, principalmente a que no se ha modificado la gramática de restricciones para tratar adecuadamente las ambigüedades generadas por el nuevo diseño del analizador. Aún así, creemos que es interesante su incorporación a la arquitectura del etiquetador, puesto que da mayor margen de mejora de los resultados finales.

El tratamiento de unidades multipalabra, en cambio, al no introducir errores y disminuir la ambigüedad de las palabras estándar, ayuda a obtener resultados más cercanos a los esperados dado el conjunto de etiquetas y el tamaño del corpus de entrenamiento utilizados.

Por último, la aportación del tratamiento de palabras no estándar, resulta bastante evidente, aún más teniendo en cuenta que se le aplica únicamente al 4% de las palabras del corpus de evaluación. En el gráfico 2 se observa en detalle la evolución de los resultados en el caso particular de las palabras desconocidas, que son las que mayores niveles de ambigüedad presentan. Gracias a la aportación del analizador propuesto[5] y al tratamiento de palabras desconocidas, la precisión aumenta hasta el 78.6%, mientras que basándonos en el analizador morfológico incremental la precisión apenas supera el 50%.

Analizados los resultados en los diferentes niveles, el tratamiento de palabras no estándar resulta decisivo a la hora de mejorar los resultados, evitando en torno al %40-%60 de los errores cometidos al aplicar otros métodos para disminuir la ambigüedad (sea la gramática de restricciones o únicamente los modelos ocultos de Markov).

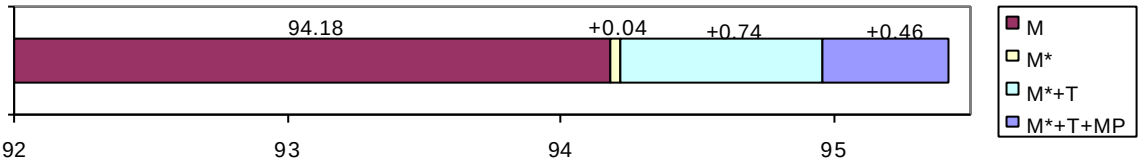

Gráfico 1.- Aportación de los procesos en la precisión media.

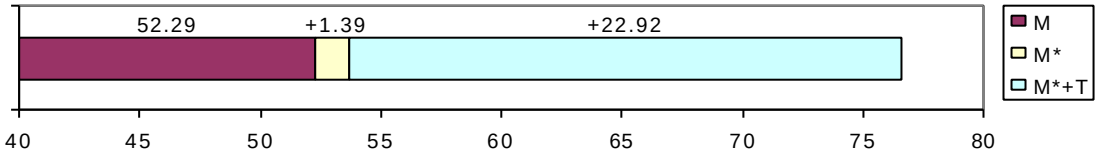

Gráfico 2.- Evolución de la precisión de las palabras desconocidas.

Por último, cabe destacar que todas las adaptaciones presentadas son opcionales, incluso se da la posibilidad de elegir los heurísticos a aplicar. Por otro lado, existe la posibilidad de enriquecer el léxico de usuario. Todo ello compone una herramienta robusta pero a la vez flexible, ya que el usuario, en función del tipo de texto a procesar, puede decidir qué tipo de procesamiento requiere y, si fuera necesario, aportar léxicos especializados, lo cual resulta especialmente interesante para el tratamiento de textos técnicos.

Bibliografía:

(Aduriz *et al*., 1996a) Aduriz I., Aldezabal I., Artola X., Ezeiza N. and Urizar R. Multiword lexical units in EUSLEM, a lemmatiser-tagger for Basque. *Proceedings of COMPLEX*. 1-8. 1996.

(Aduriz *et al*., 1996b) Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. A corpus based morphological disambiguation tool. *Proc. of SEPLN*. 41-50. 1996.

---

[4] **M** = desambiguación sobre la salida del analizador morfológico incremental;
M* = desambiguación sobre la salida del analizador morfológico propuesto;
T = desambiguación incluyendo el tratamiento de palabras no estándar;
MP = desambiguación incluyendo el tratamiento de unidades multipalabra.

[5] A pesar de que en la salida del analizador propuesto el porcentaje de palabras desconocidas y el número medio de interpretaciones aumenta, al disminuir el porcentaje de errores, los resultados del aprendizaje supervisado ayudan a etiquetar mejor este conjunto de palabras.

(Agirre *et al*., 1995) Agirre E., Arregi X., Arriola J.M., Artola X., Diaz de Illarraza A., Insausti J.M., Sarasola K. Different issues in the design of a general-purpose Lexical Database for Basque. *First workshop on application of Natural Language to Data Bases, NLDB'95*. 1995.

(Aizpurua *et al*., 2000) Aizpurua I., Alegria I., Ezeiza N. 2000. GaIn: un buscador Internet/Intranet avanzado para textos en euskera. *Actas del XVI Congreso de la SEPLN*. Vigo 2000.

(Alegria *et al*., 2001) Alegria I., Aranzabe M., Ezeiza A., Ezeiza N, Urizar R. 2001. Using Finite State Technology in Natural Language Processing of Basque. *6th Conf. on Implementation and Applications of Automata. CIAA'2001.*

(Alegria *et al.,* 2002) Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., Urizar R. Robustness and customisation in an analyser/lemmatiser for Basque. *Proceedings of LREC-2002 Workshop Customizing knowledge in NLP applications*. 2002.

(Armstrong *et al*., 1995) Armstrong S., Russel G., Petitpierre D., Robert G. An open architecture for Multilingual Text Processing. *Proceedings of EACL'95.* v1, 101-106. 1995.

(Arrieta *et al*., 2000) Arrieta B., Arregi X., Alegria I. 2000. An Assistant Tool For Verse-Making In Basque Based On Two-Level Morphology. *Proceedings of ALLC/ACH*.

(Ezeiza *et al.*, 1998) Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. Combining stochastic and rule-based methods for desanbiguation in agglutinative languages. *Proc. of Coling-ACL'98*. 1998.

(Karlsson *et al.* 1995) Karlsson F., Voutilainen A., Heikkila J,. Anttila A. *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*.. Mouton de Gruyter. 1995.

Resumen

En este artículo se presenta el lematizador/etiquetador de euskara EUSLEM. El diseño inicial del lematizador se ha ido adaptando para afrontar dos objetivos: por un lado, la mejora del analizador morfológico, reduciendo el número de errores y dando la opción de incorporar léxicos especializados, y, por otro, la mejora de la precisión de la desambiguación.

Con el fin de obtener resultados más precisos, por un lado, en el caso de las palabras estándar, se incluye el módulo de procesamiento de unidades multipalabra, que además de tratar algunas unidades léxicas como locuciones y colocaciones, identifica y procesa entidades con nombre propio del tipo de fechas y expresiones numéricas. Por otro lado, las palabras no estándar, al ser analizadas, obtienen mayor número de interpretaciones morfológicas que las estándar. Por ello, y dado que algunas de las interpretaciones se pueden descartar basándose en información local no contextual, se ha diseñado una serie de heurísticos para aumentar la precisión del análisis morfológico introduciendo un mínimo de error en el resultado.

Todas estas adaptaciones repercuten en el proceso de desambiguación, consiguiendo evitar alrededor del 20% de los errores cometidos sin incorporar los procesos descritos. Especialmente significativa es la mejora en el caso de palabras desconocidas, para las que se evita el 40%-60% de los errores.

Con todo ello se consigue un lematizador/etiquetador robusto y de gran cobertura, pero, al mismo tiempo, flexible, dando al usuario opción de aplicar las mejoras que estime necesaria para un mejor procesamiento de sus textos.