Etiquetado semiautomático del rasgo semántico de animicidad para su uso en un sistema de traducción automática.

Arantza Díaz de Ilarraza, Mikel Lersundi, Aingeru Mayor, Kepa Sarasola
{jipdisaa, jialeaym, jibmamaa, jipsagak}@si.ehu.es
Euskal Herriko Unibertsitatea

Abstract The ambiguity related to the use of movement and localisation declension cases in Basque is a serious problem in the morphological generation phase in Machine Translation. We present the approach we have developed to solve this ambiguity. Information about the [±animate] semantic feature of the lemma we want to decline is necessary to choose the appropriate suffix. The lexical database used does not contain such information. Besides, it would be very hard to add it manually to the 28.000 substantives contained in it. For this reason, we made two experiments to obtain the [±animate] feature from other resources. First, our aim was to get automatically the required knowledge from corpora, but the results were not good. Secondly, after a minimal manual tagging and using semantic relations between words extracted from definitions of a monolingual dictionary, we tagged more than half of the words in real texts with the [±animate] feature.

1 Introducción

El grupo IXA está acometiendo la construcción de un sistema de traducción automática multilingüe basado en transferencia. Han sido ya implementados los módulos de un primer prototipo para la traducción de sintagmas nominales y sintagmas preposicionales de inglés a euskara que opera interactivamente para resolver ambigüedades. En la construcción del prototipo se han reutilizado diferentes herramientas y recursos lingüísticos de amplia cobertura, que usan información morfológica y sintáctica [Diaz de Ilarraza et al., 1999].

En este artículo presentamos un trabajo complementario realizado para ese sistema en aras de resolver la ambigüedad que se crea en la generación morfológica al elegir el sufijo adecuado para los casos de declinación de movimiento y localización. La elección del sufijo necesita la información sobre el rasgo semántico [±animado] del lema que se vaya a declinar. Es importante resaltar que en la generación, resolvemos las ambigüedades buscando una sola de las alternativas posibles y asegurando un resultado correcto, sin analizar si son o no correctas el resto de las alternativas.

En la base de datos léxica del euskara EDBL que se usa en este prototipo no está incluido, por el momento, ningún rasgo semántico. La base de datos léxica contiene 28.000 nombres comunes por lo que resultaría muy costoso un etiquetado manual completo. Por ello realizamos dos aproximaciones diferentes para extraer la información del rasgo [±animado] de manera automática a partir de otros recursos lingüísticos: en la primera intentamos conseguir la información usando corpus, pero conseguimos escasos resultados; la siguiente aproximación la hicimos usando el estudio sobre las relaciones semánticas entre palabras a partir de las definiciones del diccionario monolingüe en euskara *Euskal Hiztegia* [Sarasola, 1996], consiguiendo resultados muy satisfactorios.

Este artículo está organizado de la siguiente manera: en el apartado 2 presentamos las características del rasgo semántico [±animado] y la declinación en euskara. El apartado 3 explica los experimentos realizados para el etiquetado automático usando corpus. Tras el apartado 4, dónde se detallan los estudios realizados sobre las relaciones semánticas entre palabras, describimos, en el apartado 5, los experimentos realizados usando éstas para el etiquetado semiautomático. El artículo termina subrayando los resultados obtenidos y planteando próximas líneas de trabajo.

2 El rasgo semántico [±animado] y la declinación en euskara

Los nombres que tienen el rasgo semántico [+animado] son aquellos nombres comunes que designan seres humanos o animales; también tendrán este mismo rasgo los grupos que quedan a la izquierda de los citados en la escala

de animicidad[1] de Silverstein [Silverstein, 1976]. Además de todos ellos, pueden ser incluidos en este grupo los así tomados, a pesar de no serlo, como son: Dios, ángel, diablo...

En la declinación vasca hay que diferenciar dos partes importantes:

- el elemento declinado
- la desinencia que toma

El euskara tiene una única declinación y no varias, ya que a pesar de que con una misma marca de caso puedan aparecer formas como gizoni (a hombre), gizonari (al hombre) y gizonei (a los hombres) (como es el caso del dativo), ésto no se debe a que sean diferentes marcas de caso, sino variantes de la misma marca, ya que lo que varía es debido a la presencia o ausencia del determinante y la marca de plural [Euskaltzaindia, 1985].

Vamos a analizar los casos usados en la descripción de conceptos espaciales. Por una parte los casos inesivo, ablativo, adlativo y adlativo final, y por otra el uso del genitivo, pues tienen problemáticas bien diferentes en relación con el rasgo [±animado].

2.1 Inesivo, ablativo, adlativo y adlativo final

El caso inesivo en euskara expresa la localización inmóvil:

"behia pentze**an** da"
pentze+inesivo
(la vaca esta en el prado)

y los casos ablativo, adlativo y adlativo final expresan la localización móvil:

"Baiona**tik** etorri da"
Baiona + ablativo
(ha venido de Baiona)
"medikua**rengana** joan da"
medikua + adlativo
(ha ido al médico)
"zuhaitz**etaraino** heldu da"
zuhaitz + adlativo final
(ha llegado hasta los árboles)

Pero el morfema usado en estos casos es diferente según si el lema al que se une expresa un ente animado o no.

[1] escala de animicidad: pronombres de primera y segunda persona > pronombres de tercera persona > nombres propios > nombres comunes de seres humanos > nombres comunes animados no humanos > nombres comunes inanimados.

Cuando el lema al que se una el sufijo de declinación sea no animado el morfema usado será uno de los siguientes en función de su determinación y de su número:

	det. sg.	**det./det. pl.**
Inesivo	(e)an	(e)tan
Ablativo	(e)tik	(e)tatik
Adlativo	(e)ra	(e)tara
Adl. Final	(e)raino	(e)taraino

Cuando el ente asociado al lema es animado, se usan los siguientes morfemas, en los que siempre aparece la particula 'gan':

	det. sg.	**det./det.pl.**
Inesivo	a(ren)gan	(en)gan
Ablativo	a(ren)gandik	(en)gandik
Adlativo	a(ren)gana	(en)gana
Adl. final	a(ren)ganaino	(en)ganaino

Una excepción es el caso de usos literarios cuando a un ente no animado se le da metafóricamente el rasgo [±animado], usándose los morfemas de éstos:

"itsaso**arengan** dut esperantza"
(tengo esperanza en el mar).

Por otra parte, muy excepcionalmente se usan los sufijos de los no animados con lemas asociados a entes animados:

"hiru emakume**tatik** bat"
(una de cada tres mujeres)
"gizon**etik** gizon**era** alde handia dago"
(hay mucha diferencia de un hombre a otro)

2.2 Genitivo

A diferencia de otras lenguas como el inglés, el francés o el castellano en las que se usa la misma estructura de genitivo para expresar localización y posesión, en euskara cada una de ellas se indica con un caso de declinación diferente.

El locativo o genitivo locativo materializado con el caso de declinación -(e)ko (N1ko N2) expresa que N2 está localizado en N1:

"mendi**ko** etxea"
(la casa del monte)

El genitivo posesivo toma la forma -(r)en (N1ren N2) y expresa que N1 posee N2:

"aitona**ren** etxea"
(la casa del abuelo)

Cuando el término que vamos a declinar es un ente animado solamente aceptará el caso

genitivo posesivo, y en caso de ser no animado aceptará uno u otro dependiendo de si se quiere expresar localización o posesión, pero no resulta fácil dirimir cuál es el tipo de relación entre dos palabras de un texto. Este problema lo podemos ver claramente en la expresión de las relaciones parte-todo. ¿Cuándo la intención es expresar la localización de la parte en el todo, o en cambio subrayar algún tipo de posesión de la parte por el todo? [Aurnague, 98]. Para hacer frente a este problema habría que tener en cuenta las relaciones meronímicas (parte-todo) entre las palabras, siendo éste conocimiento demasiado caro de construir manualmente. Por ello, se está trabajando para obtener esta relación automáticamente a partir de las definiciones del diccionario.

Hemos visto lo importante que resulta el rasgo [±animado]. Nos encontramos ahora con el problema de que la base de datos léxica que se usa no incluye este rasgo. ¿Podríamos conseguir de modo más o menos automático el etiquetado de este rasgo para los nombres comunes en euskara?

3 Etiquetado automático a partir del Corpus

En nuestra primera aproximación pretendimos deducir el conocimiento semántico a partir de un corpus, para lo cual usamos el recogido del periódico *Euskaldunon Egunkaria*.

El corpus tiene:

1.267.453 palabras

1.187.781 palabras standard

311.901 nombres comunes

7.219 nombres comunes diferentes

Nuestra idea fue recoger todas las apariciones de palabras declinadas con los casos de localización: inesivo, ablativo, adlativo y adlativo final. Si en las apariciones de una palabra declinada en estos casos, el morfema de caso contenía la partícula "gan" con una frecuencia lo suficientemente alta, era etiquetado como [+animado]. Si no, era etiquetado como [-amimado].

Recogimos del Corpus 38.836 apariciones de nombres declinados en estos casos, siendo 2.366 nombres diferentes. Observamos que el número de nombres que se pudieron etiquetar con este método es muy limitado, más aún si le exigimos un mínimo de fiabilidad (por ejemplo un mínimo de 5 palabras para ese lema declinadas con los casos de localización)

	todos	**> 5 apariciones**
Con "gan" [+animado]	75	3
Sin "gan" [-animado]	2.291	764

Otra aproximación fue etiquetar como [-animado] las palabras que aparecen en el corpus declinadas con el genitivo locativo. En este caso, aparece una fuente de error proveniente del análisis del corpus, ya que al no realizarse desambiguación alguna, en las apariciones de palabras declinadas con el caso distributivo, cuyo sufijo es también "ko", una de las posibilidades que da el analizador morfosintáctico es la declinación de caso genitivo locativo, que no es correcta.

A pesar de obviar esta fuente de error, tan sólo recogimos 1.544 palabras declinadas con el genitivo locativo, que seguía resultando un número muy escaso.

Visto que usando corpus, no conseguimos los resultados deseados, planteamos la posibilidad de usar estudios de las relaciones semánticas entre palabras para intentar otra estrategia.

4 Estudio de las relaciones semánticas superficiales entre palabras

Unas palabras están relacionadas semánticamente con otras. Para conocer el tipo de relación que existe entre ellas utilizamos la extracción de información a través de diccionarios monolingües –en nuestro caso el diccionario monolingüe en euskara *Euskal Hiztegia* de Ibon Sarasola–.

Las definiciones en cuestión tienen diferentes patrones, los cuales nos dan información sobre si las palabras que aparecen en la definición son sinónimos, genus, o relatores específicos con respecto a la entrada léxica que definen.

Según Smith y Maxwell hay básicamente tres métodos para definir una entrada léxica [Smith et al., 1980]:

- por medio de un sinónimo: una palabra que tendrá el mismo significado que su entrada léxica. La entrada léxica y el sinónimo formarán parte de la misma categoría léxica.
  akabatu. **Bukatu**(*sin*),**amaitu**(*sin*).
  [acabar. **Finalizar**(*sin*), **terminar**(*sin*).]
- por medio de una definición clásica: 'genus + differentia'. El significado de la entrada léxica es definido por medio de un término

genérico (genus) y una descripción de los rasgos que distinguen a la entrada léxica de las demás que están situadas bajo el mismo término genérico. La relación existente entre la entrada léxica y el genus es la de hiperonimia; esto es, el genus es el término genérico o hiperónimo, y la entrada léxica un término más específico o hipónimo.

aireontzi. **Hegalda daitekeen** (*differentia*) zernahi ibilgailu (*genus*).

[aeronave.**Vehículo** (*genus*) que **puede volar** (*differentia*).]

- por medio de relatores específicos, que son ciertas vías sintácticas para unir una palabra de la definición con la entrada léxica. El relator específico utilizado en la definición determina, a menudo, la relación semántica existente entre la entrada léxica y el núcleo de la definición.

  ezpara. Tabanidae familiako **intsektuei** (*término relacionado*) **ematen zaien izena** (*relator*).

  [tábano. **nombre que se da a**(*relator*) ciertos **insectos**(*término relacionado*) de la familia Tabanidae.]

Los resultados que obtenemos al intentar reconocer estos elementos automáticamente son bastante satisfactorios en lo que respecta a la sinonimia e hiperonimia/hiponimia. Para calibrar estos resultados hemos tomado muestras de 100 entradas léxicas por cada categoría, y los resultados que obtenemos son:

- cobertura: 96,85%
- precisión: 96,47%

Actualmente estamos trabajando con el tema de los relatores específicos. Hemos definido 17 relatores específicos, y con ellos obtenemos estos resultados:

- cobertura: 87,24%
- precisión: 100%

Entendemos por cobertura la relación entre el número de elementos que son bien marcados y el número total. La precisión señala la relación entre el número de elementos bien marcados y el total de los elementos marcados.

5 Etiquetado semiautomático usando el Genus y los Relatores específicos

La idea fue etiquetar los nombres comunes del diccionario monolingüe *Euskal Hiztegia* ya que en un alto número de las definiciones de los mismos se usa el genus o los relatores de la entrada léxica y, como veremos a continuación, esta información se puede utilizar para nuestros fines.

	Entradas	**Acepciones**
Total	16.380	26.461
Con genus o relator	10.517	14.569

Nuestra estrategia consistió en etiquetar manualmente el rasgo [±animado] en un pequeño número de las palabras que son las que con más frecuencia aparecen como genus o como relatores en el diccionario y a partir de ellas heredar o deducir el valor en el rasgo [±animado] para el máximo número posible del resto de palabras del diccionario.

La relación hiperonimia/hiponimia que marca el genus, es una relación en la cual se hereda el atributo [±animado]. Es decir, si "langile" (trabajadora) posee el rasgo [+animado], todos sus hipónimos, o lo que es lo mismo, todas las palabras cuyo hiperónimo sea "langile", tendrán su mismo rasgo [+animado].

De algunos relatores específicos se puede deducir el rasgo [±animado] de la palabra que lo tiene. Por ejemplo la etiqueta de todas las palabras cuyo relator sea "nolakotasuna" (cualidad) será [-animado].

Describamos más a fondo el proceso que fue implementado:

```
procedure Etiquetado_del_diccionario {
   para cada (Nombre común del diccionario) {
      Buscar_su_etiqueta (Nombre)
    }
}
procedure Buscar_su_etiqueta (Nombre) {
   si ( Nombre tiene Genus/Relator)
      si (  Genus no esta etiquetado ){
         Buscar_su_etiqueta(Genus)   #recursividad
      }
      si ( Nombre está etiquetado){
         si ( Nombre.Etiq != Genus/Relator.Etiq){
            Nombre.Etiq = [?anim]
         }
      }
      sino {
         Nombre.Etiq =  Genus/Relator.Etiq
      }
   }
}
```

Es decir, se fue examinando cada nombre común, buscando en su jerarquía hiperonímica hasta encontrar una palabra etiquetada, o que tuviese un relator etiquetado, recogiendo esa etiqueta para el nombre. Diferentes acepciones de una misma entrada podían recoger etiquetas contradictorias. En ese caso les fue asignada una etiqueta especial [?anim] que señalaba este hecho. Como la jerarquía hiperonímica creada permite la posibilidad de ciclos, la búsqueda se truncaba en un determinado número de recursiones.

Las palabras más frecuentemente usadas como genus (g) o relatores (r) más frecuentemente fueron etiquetadas a mano:

g/r	frec	nombre		rasgo
r	531	Nolakotasun	(cualidad)	[-anim]
g	377	Pertsona	(persona)	[+anim]
r	362	Multzo	(conjunto)	[-anim]
r	230	Zati	(parte)	[-anim]
g	202	Gai	(materia)	[-anim]
g	188	Tresna	(instrumento)	[-anim]
g	187	Landare	(planta)	[-anim]
r	177	Egoera	(situación)	[-anim]
		...		

Resulta interesante observar que, entre estos relatores, "multzo" (conjunto) expresa holonimia y "zati" (parte) meronimia, relaciones que podrán ser usadas para hacer frente a la ambigüedad entre localización y posesión.

Realizamos pruebas diferentes, siendo los siguientes los datos obtenidos etiquetando las 2, 5, 10, 25 y 100 palabras más usadas como genus o relator:

Etiq. manual	Etiq. autom.	[+anim]	[-anim]	[?anim]
2	2.445	962	1.483	72
5	6.157	789	5.368	245
10	6.672	790	5.882	244
50	8.155	969	7.186	289
100	**8.439**	**1.195**	**7.244**	**324**

Etiquetando los 100 nombres con mayor frecuencia conseguimos etiquetar un 75% de los nombres con información de genus o relator del diccionario (8.439 de 10.517), y un 50% del total de nombres del diccionario (8.439 de 16.380).

Podemos ver como los 5 primeros términos que aparecen con mayor frecuencia como genus o relatores, son los más productivos en el proceso de etiquetado automático.

Se observa que al pasar de etiquetar 2 a etiquetar 5 términos, el número de términos que se etiquetaron como [+animado] descendió de 962 a 789. Esto se explica porque se creó contradicción en algunas palabras para las cuales alguna acepción había sido etiquetada como [+animado], y al ir incluyendo más genus y relatores etiquetados manualmente, que son la mayoría [-animado], a otras acepciones de esa palabra les correspondió la etiqueta [-animado], siéndole asignada entonces la etiqueta especial [?anim].

Etiquetando manualmente un número lo suficientemente grande de palabras, aseguramos en el proceso automático una buena fiabilidad de las palabras etiquetadas. Esto podemos verlo en la siguiente tabla, donde se muestra qué porcentaje de palabras etiquetadas tienen alguna acepción todavía no etiquetada, no siendo por lo tanto su etiquetado totalmente fiable

Etiq. Manual	Etiq. Autom.	No Fiable	
2	2.445	385	9,1%
5	6.157	677	10,9%
10	6.672	710	10,6%
50	8.155	634	7,7%
100	**8.439**	**577**	**6,8%**

El siguiente paso consistió en calcular la cobertura del etiquetado en texto real, para lo que usamos el mismo corpus del periódico *Euskaldunon Egunkaria* que hemos descrito anteriormente. En la tabla podemos observar los datos obtenidos que suponen una cobertura del 55% de las apariciones de nombres en el texto, es decir hemos conseguido una cobertura de más de la mitad habiendo realizado un etiquetado manual muy poco costoso.

	Nombres diferentes	Apariciones en el corpus
Total	7.219	311.901
Etiquetados	**2.618 (36%)**	**171.593 (55%)**
[+animado]	302	16.600
[-animado]	2.316	154.993

Asimismo pudimos observar que de los 100 nombres con mayor frecuencia de aparición en el corpus, 58 han sido etiquetados.

6 Conclusiones y líneas futuras

Hemos presentado la problemática que se da en euskara en la relación entre los casos de declinación de movimiento y localización y el rasgo semántico [±animado].

Es necesario tener etiquetados los nombres comunes con este rasgo para resolver la ambigüedad que se crea en la fase de generación morfológica de un sistema de traducción automática inglés-euskara que está siendo desarrollado en el grupo IXA.

El etiquetado manual de la base de datos léxica EDBL que contiene unos 28.000 nombres comunes supone un trabajo excesivamente caro. Hemos intentado diferentes estrategias para realizar un etiquetado de nombres comunes automático o semiautomático a partir de otros recursos lingüísticos.

El primer intento que realizamos con la intención de extraer de manera totalmente automática a partir de corpus la información semántica deseada parecía en principio más fácil y eficaz, pero los datos obtenidos resultaron ser muy pobres

El etiquetado semiautomático usando el estudio del genus y de los relatores sobre el diccionario *Euskal Hiztegia* nos ha proporcionado unos resultados muy buenos, ya que hemos conseguido una cobertura del 55% de las apariciones de nombres comunes en un corpus real de mas de 1 millón de palabras, habiendo etiquetado a mano únicamente 100 nombres. Esperamos aumentar la cobertura una vez completado el estudio del genus y de los relatores.

Cuando hayamos obtenido las relaciones meronímicas a partir de las definiciones del diccionario monolingüe podremos, a su vez, estudiar las posibilidades de reducción de ambigüedad en el uso del caso genitivo en las relaciones parte-todo.

Una de las aplicaciones de estos experimentos a plantear en el futuro consiste en enriquecer la base de datos léxica EDBL a partir de la información semántica obtenida.

7 Agradecimientos

Este trabajo está subvencionado por la Universidad del País Vasco (UPV 141.226-G19/99).

Queremos agradecer su colaboración a todas las personas del grupo de investigación IXA

8 Bibliografía

Agirre E., Ansa O., Arregi X., Artola X., Díaz de Ilarraza A., Lersundi M., Martinez D., Sarasola K., Urizak R., 2000, “Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar”, *EURALEX’2000*.

Aurnague M, 1998, "Basque genitives and part-whole relations: typical configurationsand dependences", *Carnets de Grammaire,* Rapporrt nº 1 – avril 1998, Equipe de Recherche en Sintaxe et Sémantique, Université de Toulouse-Le Mirail.

Diaz de Ilarraza A., Mayor A., Sarasola K., 1999, "Reusability of wide-coverage linguistic resources in the construction of an English-Basque Machine Translation System", Hybrid Approaches to Machine Translation, Institut of Applied Information Sciences, Saarbruecken http://rockey.iis.sinica.tw/oliver/iaiwp/

Euskaltzaindia, 1985, *Euskal Gramatika Lehen Urratsak-I*, Iruñea.

Sarasola, I., 1996, *Euskal Hiztegia*, Donostia, Gipuzkoako Kutxa.

Silverstein, M., 1976, “Hierarchy of Features and Ergativity”, *Grammatical categories in Australian languages*, R.M.W. Dixon (ed.), Canberra, Australian Institute of Aboriginal Studies, *Linguistic Series* 22 (1976) 112-171.

Smith, R.N., Maxwell, E., 1980, “An English dictionary for computerised syntactic and semantic processing systems”, *Proceedings of the International Conference on Computational Linguistics* 36 (1980) 303-322.