Semiautomatic labelling of semantic features

Arantza Díaz de Ilarraza, Aingeru Mayor and Kepa Sarasola
IXA Group. Computer Science Faculty. University of the Basque Country
Donostia/San Sebastian. The Basque Country
jipdisaa/jibmamaa/jipsagak@si.ehu.es

Abstract

This paper presents the strategy and design of a highly efficient semiautomatic method for labelling the semantic features of common nouns, using semantic relationships between words, and based on the information extracted from an electronic monolingual dictionary. The method, that uses genus data, specific relators and synonymy information, obtains an accuracy of over 99% and a scope of 68,2% with regard to all the common nouns contained in a real corpus of over 1 million words, after the manual labelling of only 100 nouns.

1 Introduction

Semantic information is essential in a lot of NLP applications. In our case, the feature [±animate] is necessary to disambiguate between the possible Basque translations for the English preposition "of" and the Spanish preposition "de", when referring to location or possession. This ambiguity appears very often when translating to Basque [Díaz de Ilarraza et al., 2000]. A complete manual labelling of semantic information would prove extremely expensive.

This study aims to outline the strategy and design of a semiautomatic method for labelling semantic features of common nouns in Basque, expanding and improving the idea outlined in [Díaz de Ilarraza et al. 2000]. Due to the poor results obtained, this study dismissed the possibility of an initial approach aimed at extracting the information corresponding to the (±animate) feature automatically from corpus. Instead, an alternative idea was proposed, i.e. that of using semantic relationships between words extracted from the Basque monolingual dictionary *Euskal Hiztegia* (Sarasola 1996). In this context, we used genus data and specific relators, together with a few words manually labelled, to extract the information corresponding to the (±animate) feature. The results obtained were very promising: 8,439 common nouns were labelled automatically after the manual labelling of just 100.

This paper describes the work carried out with the aim of expanding this idea this idea through the inclusion of information about synonymy, repeating the automatic process iteratively in order to obtain better results and, monitoring the reliability of the labelling of each individual noun. After studying the ideal relationship between the manual part of the operation and the scope of the automatic process, we generalised the process in order to adapt it to other semantic features. We obtained very satisfactory results considering the labelling of common nouns contained in the dictionary: for the [±animate] feature, we labelled 12,308 nouns with an accuracy of 99.2%, after the manual labelling of only 100.

This paper is organised as follows: section 2 presents the semantic relationships between words extracted from the Basque monolingual dictionary, and used by our semiautomatic labelling method. The method itself is described in section 3. The experiments carried out with the aim of optimising the efficiency of the method are described in section 4, and section 5 outlines the accuracy and scope of the labelling process for the [±animate] semantic feature. Finally, section 6 describes how the method was generalised to cover other semantic features. The study finishes by underlining the results obtained and suggesting future research.

2 Superficial semantic relationships between words in dictionaries

According to Smith and Maxwell, there are three basic methods for defining a lexical entry [Smith and Maxwell., 1980]:

- By means of a synonym: a word with the same sense as the lexical entry.
  finish. **conclude**(*sin*), **terminate**(*sin*)
- By means of a classical definition: 'genus + differentia'. The genus is the generic term or

hyperonym, and the lexical entry a more specific term or hyponym.

aeroplane. **vehicle** (*genus*) that **can fly** (*differentia)*

- By means of specific relators, that will often determine the semantic relationship between the lexical entry and the core of the definition.

horsefly. **Name given to** (*relator*) certain **insects** (related term) of the Tabanidae family

One method for identifying the semantic relationship that exists between different words is to extract the information from monolingual dictionaries.

Agirre et al. (2000) applied it for Basque, using the definitions contained in the monolingual dictionary *Euskal Hiztegia*. We use for our research the information about genus, specific relators and synonymy extracted by them.

3 Semiautomatic labelling using genus, specific relators and synonymy

In order to label the common nouns that appear in the dictionary, we used the definitions of the 26,461 senses of the 16,380 common nouns defined by means of genus/relators (14,569) or synonyms (11,892).

The experiment was carried out as follows: firstly, we used the information relative to genus and specific relators to extract the information regarding the [±animate] feature (3.1). Subsequently, we also incorporated the information relative to synonymy (3.2). Finally, we repeated the automatic process iteratively in order to obtain better results (3.3). An example of the whole process is given in section 3.4.

3.1 Labelling using information relative to genus and specific relators

Our strategy consisted of manually labelling the semantic feature for a small number of words that appear most frequently in the dictionary as genus/relators. We used these words to infer the value of this feature for as many other words as possible.

This inference is possible because in the hyperonymy/hyponymy relationship, that characterises the genus, semantic attributes are inherited. For example, if 'langile' (*worker*) has the [+animate] feature, all its hyponyms (or in other words, all the words whose hyperonym is 'langile') will have the same [+animate] feature.

Certain genus are ambiguous, since they contain senses with opposing semantic features. For example 'buru' (*head/boss*) has the [-animate] feature when it means '*head*' and the [+animate] feature when it means '*boss*'. The semantic feature of the sense defined can also be deduced from some specific relators. In this way, the semantic feature of words whose relator is 'nolakotasuna' *(quality)* would be [-animate], such as in the case of 'aitatasuna' *(paternity)*, for example. There are also certain relators that offer no information, such as 'mota' (*type*), 'izena' (*name*), and 'banako' (*unit, individual*).

We used four types of labels during the manual operation: [+], [-], [?] and [x]. [?] for ambiguous cases; and [x] for relators that do not offer information regarding this semantic feature.

```
procedure Labelling_of_the_dictionary {
  foreach (common Noun of the dictionary) {
       (Label, Reliability) = Find_its_label (Noun)   }
}
procedure Find_its_label (Noun) {
  foreach (Sense with Noun Genus/Relator) {
       if (Genus/Relator labelled){ Sense.Label       = Genus/Relator.Label
                                    Sense.Reliability = Genus/Relator.Reliability
                                  }
         else {( Sense.Label,
                 Sense.Reliability) = Find_its_label(Genus) } #recursion
       if (Noun.Label != Sense.Label)       { Noun.Label = [?] }
         else                               { Noun.Label =  Sense.Label }
  } # end foreach
  Noun.Reliability = Σ Reliability labelled senses / number of senses
  return (Noun.Label, Noun.Reliability)
}
```

Figure 1*. Implementation of the automatic process using genus and relater information*

In order to establish the reliability of the automatic labelling process for a particular noun, we considered the number of senses labelled, taking into account the reliability of the labels of the genus (or relator) that provided the information. The result was calculated as follows:

$$Rel\_noun = \sum Rel\_genus\_per\_sense / n\_senses$$

During manual labelling, we assigned reliability value 1 to all labels, since all the senses of these nouns are taken into account.

Figure 1 shows the algorithm used. For each common noun defined in the dictionary, we take, one by one, all their senses containing genus or relator, assigning in each case the first label associated to a genus or relator in the hierarchy of hyperonyms. When the sign of all the labels are coincident we use it to label the entry, in other case, we use the label [?]. In all cases, their reliability is calculated.

When we detect a cycle, the search is interrupted and the sense to be tagged remains unlabelled.

3.2 Labelling using synonymy information

Labelling using genus and relators can be expanded by using synonymy. Since the synonymy relationship shares semantic features, we can deduce the semantic label of a sense if we know the label of its synonymes.

Therefore, the information obtained during the previous phase can now be used to label new nouns. It also serves to increase the reliability of nouns already been labelled thanks to the genus information of some of their senses. If the synonymy information provided corroborates the genus information, the noun's reliability rating increases. If, on the other hand, the new label does not coincide with the previous one, a special label: [?] is assigned to the noun indicating this ambiguity.

The automatic process using synonymy was implemented in the same way as in the previous process.

3.3 Iterative repetition of the automatic process

Our next idea was to repeat the process; since the information gathered so far using synonymy may also be applied hereditarily through the genus' hyperonymy relationship.

We therefore repeated the process from the beginning, trying to label all the senses of the nouns that had not been fully labelled during the initial operations, by using the information contained in the senses of the nouns that had been fully labelled (reliability 1).

As with the initial operation, we first used information about genus and relators, and then, synonymy.

This process can be repeated any number of times, thereby labelling more and more words while increasing the reliability of the labelling itself. However, repetition of the process also increases the number of words labelled as ambiguous [?], since more senses are labelled during each iteration, thereby increasing the chances of inconsistencies. As we shall see, this iterative process improves the results logarithmically up to a certain number of repetitions, after which it has no further advantageous effects.

3.4 Example of semiautomatic labelling for the [±animate] feature

The 100 words that are most frequently used as genus (g) or relators (r) were labelled manually for the [±animate] feature, as shown in table 2 (tables 3, 4 and 5 contain the Basque words processed during the explained operation, along with their English translation in italics).

Noun		±anim	Freq	Gen/rel
nolakotasun	*(quality)*	**-**	531	Relator
pertsona	*(person)*	+	377	Genus
multzo	*(collection)*	**-**	362	Relator
txikigarri	*(collection)*	**x**	213	Relator
zati	*(part)*	**-**	230	Relator
gai	*(material)*	**-**	202	Genus
tresna	*(instrument)*	**-**	188	Genus
…				
buru	*(head)*	**?**	54	Genus

Table 2. *Manual labelling*

We shall now trace the implementation of the automatic labelling process for certain nouns.

Table 3 shows the results of the first labelling process using information about genus and relators. The words printed in bold in the results column are nouns that were labelled during the manual labelling process. We can see how the noun 'babesgarri' *(protector)* is labelled as [-] thanks to the information provided by the relator of its only sense, which was manually labelled.

Noun	N. sense	N. genus	Result of process using genus and relators	Lab	Rel.
babesgarri *(protector)*	1	1	(**zer**[-]1) *(thing)*	[-]	1
armadura *(armour)*	3	3	(**multzo**[-]1) (**babesgarri**[-]1)(soineko[]) *(collection) (protector) (garment)*	[-]	0.66
ama *(mother)*	5	3	(**emakume**[+]1)(**animalia**[+]1)(eme[]) *(woman) (animal) (female)*	[+]	0.4
iturburu *(spring)*	3	1	(aterabide[]) *(outlet)*	[]	0
gertaera *(event)*	1	1	(**gauza**[-]1) *(thing)*	[-]	1
giltzape *(prison)*	2	1	(**toki**[-]1) *(place)*	[-]	0.5
espetxe *(jail)*	2	2	(**eraikuntza**[-]1)(**leku**[-]1) *(construction) (place)*	[-]	1
adiskide *(friend)*	1	1	(**pertsona**[+]1) *(person)*	[+]	1
adiskidetzako *(friend)*	1	1	(lagun[]) *(companion)*	[]	0
apio *(celery)*	2	2	(jateko[]) (**landare**[-]1) *(food) (plant)*	[-]	0.5
filosofia *(philosophy)*	2	2	(**jakintza**[-]1)(**multzo**[-]1) *(knowledge) (collection)*	[-]	1
ikusgune *(viewpoint)*	2	1	(**gune**[-]1) *(point)*	[-]	0.5
jarrera *(attitude)*	2	2	(**era**[-]1)(ikusgune[-]0.5) *(way) (viewpoint)*	[-]	0.75
zinismo *(cynicism)*	2	2	(filosofia[-]1)(jarrera[-]0.75) *(philosophy) (attitude)*	[-]	0.87

Table 3. *Result of automatic labelling using genus and relator information*

The noun therefore has a reliability rating of 1. In the same way, 2 of the 3 senses of 'armadura' (*armour*) had coincident labels, thereby giving a reliability rating of 0.66 (f=(1+1)/3=0.66). The noun 'ama' *(mother)* was labelled as [+], thanks to the information about genus and relator of 2 of its 3 senses, out of a total of 5 (the remaining two have synonymy information). The reliability rating was therefore calculated as 0.4 (f=(1+1)/5=0.4). The word 'zinismo' *(cynicism)* was labelled as [-] thanks to the fact that the genus of its 2 senses were both labelled as such, although one did not have a reliability rating of 1. The reliability rating obtained for 'zinismo' was therefore 0.87 (f=(1+0.75)/2=0.87).

Table 4 shows some examples of the process using synonym information.

As we can see, 'iturburu' *(spring)*, which the previous process had not managed to tag, is now labelled as [-] thanks to the synonymy information associated to one of the two senses. The resulting reliability rating is 0.06 (f=0.2/3=0.06). If we look at the term 'ama', which had previously been labelled as [+] on the basis of genus information, we see that the synonyms of the two senses that use synonymy

Noun	Genus lab.	N. sens	N. syn	Results of the process using synonymy	Lab.	Relia.
iturburu *(spring)*	[]	3	2	(etorki[]) (hasiera[-]0.20) *(origin) (start)*	[-]	0.06
ama *(mother)*	[+]	5	2	(iturburu[-])(jatorri[-]) *(spring) (origin)*	[?]	1
gertakuntza *(event)*		1	1	(gertaera[-]1) *(happening)*	[-]	1
lagun *(companion)*		1	1	(adiskide[+]1) *(friend)*	[+]	1
jateko *(food)*		1	1	(janari[-]1) *(food)*	[-]	1
giltzape *(prison)*	[-]	2	1	(espetxe[-]1) *(jail)*	[-]	1
ikusgune *(viewpoint)*	[-]	2	1	(ikuspen[-]0.33) *(view)*	[-]	0.66

Table 4. *Results of automatic labelling using synonymy information*

Noun	N. sense	N. genus	Result of process using genus and relators	Lab.	Relia.
armadura *(armour)*	3	3	(**multzo**[-]1)(babesgarri[-]1)(soineko[-]1) *(collection) (protector) (garment)*	[-]	1
adiskidetzako (friend)	1	1	(lagun[+]1) *(companion)*	[+]	1
apio *(celery)*	2	2	(jateko[-]1)(**landare**[-]1) *(food) (plant)*	[-]	1
ikusgune *(viewpoint)*	2	2	(**gune**[-]1) *(point)*	[-]	0.5
jarrera *(attitude)*	2	2	(**era**[-]1)(ikusgune[-]0.5) *(way) (viewpoint)*	[-]	0.75
zinismo *(cynicism)*	2	2	(filosofia[-]1)(jarrera[-]0.75) *(philosophy) (attitude)*	[-]	0.87

Table 5. *Results of the 2nd iteration of automatic labelling using genus and relator information*

information are labelled as [-]. Due to this inconsistency, the word is now labelled as [?]. The terms 'gertakuntza' *(event)*, 'lagun' *(companion)* and 'jateko' *(food)*, which previously only had one sense, are now labelled thanks to synonym information. The words 'giltzape' *(prison)* and 'ikusgune' *(viewpoint),* which had had one sense labelled on the basis of genus, now have both senses labelled. The reliability rating for 'ikusgune' is calculated as f=(1+0.33)/2=0.66.

We then repeated the process using first the genus/relator information (table 4) followed by the synonymy information (table 5).

The aim of this repetition was to label only those words that had not been fully labelled, using the information provided by the terms that had been and that had a reliability rating of 1, such as 'babesgarri', 'gertaera', 'espetxe', 'adiskide', 'filosofia', 'ama', 'gertakuntza', 'lagun', 'jateko' and 'giltzape' (tables 4 and 5).

This process succeeded in labelling the senses of 'armadura' *(protector)*, 'adiskidetzako' *(friend)* and 'apio' *(celery),* previously left unlabelled, since their genus 'soineko' *(garment)*, 'lagun' *(friend)* and 'jateko' *(food)* had been fully labelled using the synonym information. On the other hand, 'ikusgune' *(viewpoint)*, 'jarrera' *(attitude)* and 'zinismo' *(cynicism)*, did not benefit from this repetition.

Following this process, we applied the synonymy information, thus completing the second iteration. The process may be repeated as many times as you wish.

4 Experiments for optimising the efficiency of the method

We carried out a number of different tests for the [±animate] semantic feature labelling the 2, 5, 10, 50, 100, 125 and 150 words most frequently used as genus/relators, and repeating the whole process (using both genus and relator and synonymy information) 1, 2 and 3 times.

The first 5 terms that appear most frequently

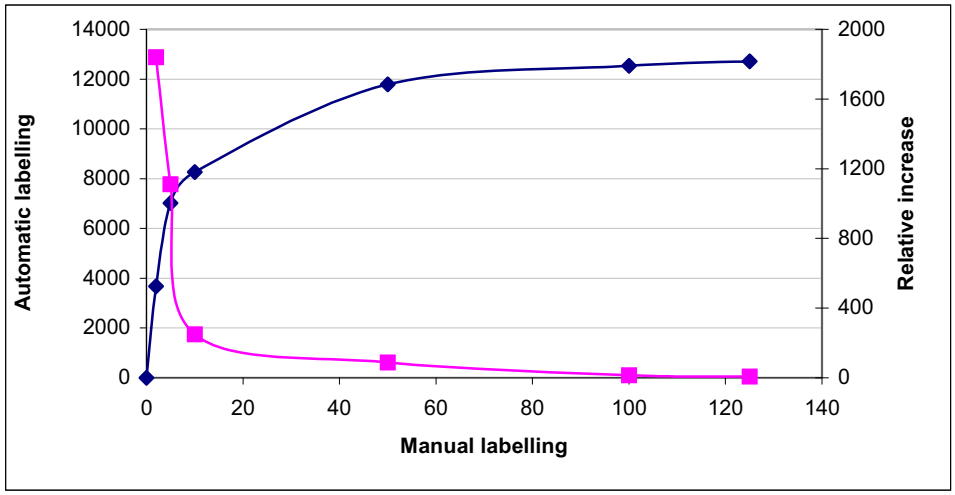

Fig. 2. *Automatic labelling and relative increase*

as genus/relators are also the most productive during the automatic labelling process. From here on, the rate of increase gradually falls, until only 7 terms are labelled automatically for every noun labelled manually.

On average, the first 2 nouns each enabled 1840 terms to be labelled, the next 3 enabled 1112 while the next 5 enabled only 250. After the hundredth noun, this average dropped to just 7 new terms labelled automatically for every term labelled manually. These results are illustrated in figure 2.

For efficiency reasons, we decided that when labelling other semantic features, we will label manually the 100 nouns most frequently used as genus/relators.

In order to decide the number of iterations required for optimum results, we compared the results obtained after 1 to 10 iterations after manually labelling 100 nouns (Figure 3).

Although no increase was recorded for the number of nouns with reliability rating 1 (i.e. with all senses labelled) after the 3rd iteration, the results for other reliability ratings continued to increase up until the 8th iteration, since as more and more information is gathered, new contradictions are generated and the number of ambiguous labels increases. When the results stabilise, we can affirm that all the available information has been used and the most accurate results possible with this manual labelling operation have been obtained. It is important to check that the process does indeed stabilise, and that it does so after a fairly low number of iterations (in this case, after 8).

The repetition of the process does not significantly increase execution time. 10 iterations of the automatic labelling process for the [±animate] feature takes just 11 minutes 33 seconds using the total capacity of the CPU of a Sun Sparc 10 machine with 512 Megabytes of memory running at 360 MHz.

We can therefore conclude that the method is viable and that, in the automatic process for other semantic features, the necessary iterations should be carried out until the results are totally stabilised.

5 Accuracy and scope of the labelling process for the [±animate] feature

In order to calculate the accuracy of the automatic labelling process, we took 1% of the labelled words as a sample and checked them manually. The results are shown in table 6.

	Reliability			
	f=1	1>f>0.5	0.5>f>0	Total
Accuracy	100%	100%	94%	99.2%

Table 6. *Accuracy of automatic labelling*

Although we initially planned to use only the labels with a reliability rating of 1, after seeing the accuracy of the others, we decided to use all the labels obtained during the process, thereby achieving an overall accuracy rating of 99.2%. We can affirm that the semiautomatic process designed and implemented here is very efficient.

The scope for the automatic labelling of the [±animate] feature (table 7) was 75.14% of all the nouns contained in the dictionary (12,308 of 16,380), having manually labelled 100 nouns and

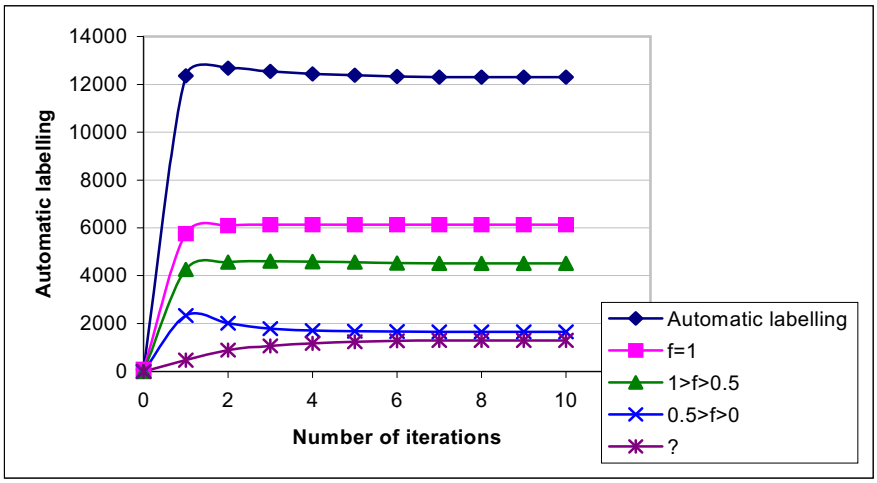

Fig. 3. *Automatic labelling according to number of iterations*

carried out 8 iterations.

	Labelling			?
	f=1	1>f>0.5	0.5>f>0	
Auto lab.	6132	4513	1663	
	12308 (75.14%)			1301

Table 7. *Scope of the dictionary*

We also calculated the scope of this labelling in a real context, using the corpus gathered from the newspaper *Euskaldunon Egunkaria*, which contains 1,267,453 words and 311,901 common nouns, of which 7,219 are different nouns. Table 8 shows the results – a scope of 69.2% with regard to the nouns that appear in the text (47.6% of the total number of different common nouns contained in the corpus). In other words, after carrying out a very minor manual operation, we managed to label two out of every three nouns that appear in the corpus. Similarly, we noted that of the 500 nouns that appear most frequently in the corpus, 348 (69.6%) were labelled.

	Appearances in the corpus		Different nouns	
Total		311,901		7,219
Labelled	(68.2%)	212,887	(47.6%)	3,434
[+]		17,408		356
[-]		195,479		3,078

Table 8. *Scope of labelling within the corpus*

6 Generalisation for use with other semantic features

Given the process's efficiency, it can be generalised for use with other semantic features. To this end, we have adapted its implementation to enable the automatic process to be carried out on the basis of the manual labelling of any semantic feature.

So far, we have carried out the labelling process for the [±animate], [±human] and [±concrete] semantic features. Table 12 shows the corresponding results.

Label	±animate	±human	±concrete
[+]	1,643	1,118	7,611
[-]	10,665	10,684	1,143
Total	12,308	11,802	8,754

Table 12. *Labelling data for different semantic features*

Conclusions

We have presented a highly efficient semiautomatic method for labelling the semantic features of common nouns, using the study of genus, relators and synonymy as contained in the *Euskal Hiztegia* dictionary. The results obtained have been excellent, with an accuracy of over 99% and a scope of 68,2% with regard to all the common nouns contained in a real corpus of over 1 million words, after the manual labelling of only 100 nouns.

As far as we know, no so method of semantic feature labelling has been described in the literature, although many authors [Pustejovsky, 2000; Sheremetyeva & Nirenburg, 2000] claim the significance of semantic features in general, and [animacy] in particular, for NLP systems.

One of the possible applications of these experiments is to enrich the Basque Lexical Database, EDBL, using the semantic information obtained.

Acknowledgements

The Basque Government Department of Education, Universities and Research sponsored this study.

Bibliography

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*.

Diaz de Ilarraza A., Lersundi M., Mayor A., Sarasola K., 2000. Etiquetado semiautomático del rasgo semántico de animicidad para su uso en un sistema de traducción automática. *SEPLN'2000.* Vigo..

Diaz de Ilarraza A., Mayor A., Sarasola K., 2000. "Reusability of Wide-Coverage Linguistic Resources in the Construction of a Multilingual MT System".*MT 2000*. Exeter. UK.

Pustejovsky J., 2000. "Syntagmatic Processes". Handbook of Lexicology and Lexicography. de Gruyter, 2000.

Sheremetyeva S. and Nirenburg S., 2000. "Towards A Universal Tool for NLP Resource Acquisition". *LREC2000*. Greece.

Smith, R.N., Maxwell, E., 1980, "An English dictionary for computerised syntactic and semantic processing systems", Proceedings of the International Conference on Computational Linguistics. 1980.