Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia, July 2002, pp. 1-8. Association for Computational Linguistics.

A Multilingual Approach to Disambiguate Prepositions and Case Suffixes

Eneko Agirre, Mikel Lersundi, David Martinez

IxA NLP group
University of the Basque Country
649 pk. - 20.080 Donostia (Spain)
{eneko, jialeaym, jibmaird}@si.ehu.es

Abstract

This paper presents preliminary experiments in the use of translation equivalences to disambiguate prepositions or case suffixes. The core of the method is to find translations of the occurrence of the target preposition or case suffix, and assign the intersection of their set of interpretations. Given a table with prepositions and their possible interpretations, the method is fully automatic. We have tested this method on the occurrences of the Basque instrumental case *-z* in the definitions of a Basque dictionary, looking for the translations in the definitions from 3 Spanish and 3 English dictionaries. The results have been that we are able to disambiguate with 94.5% accuracy 2.3% of those occurrences (up to 91). The ambiguity is reduced from 7 readings down to 3.1. The results are very encouraging given the simple techniques used, and show great potential for improvement.

1 Introduction

This paper presents some preliminary experiments in the use of translation equivalences to disambiguate the interpretations of case suffixes in Basque. Basque is an agglutinative language, and its case suffixes are more or less equivalent to prepositions, but are also used to mark the subject and objects of verbs. The method is general, and could be as easily applied to prepositions in any other language. The core of the method is to find a preposition in the translation of an occurrence of the target case suffix, and select the interpretation(s) in the intersection of both as the valid interpretation(s). At this point, we have not used additional sources for the disambiguation, e.g. governing verbs, nouns, etc., but they could complement the technique here presented.

In this particular experiment, the method was tested on the definitions of a Basque monolingual dictionary, using the *-z* instrumental as the target case suffix. The main reason is that we are in the process of building a Lexical Knowledge Base out of dictionary definitions, and the disambiguation of case suffixes and other semantic dependencies is of great interest.

The method searches for the respective definitions in English and Spanish monolingual dictionaries and tries to find a preposition that is the translation of the target case suffix. Once the preposition is found, the intersection of the set of interpretations of both the source case suffix and the translated preposition is taken, and the outcome is stored.

The resources needed to perform this task are the following: lemmatizers, bilingual dictionaries and monolingual dictionaries, as well as a table of possible interpretations of prepositions and case suffixes. In our case, we have used Basque, English and Spanish lemmatizers, Basque/English and Basque/Spanish bilingual dictionaries, a target Basque monolingual dictionary, 3 Spanish and 3 English monolingual dictionaries.

The method is fully automatic; the Spanish and English monolingual dictionaries are accessed from the Internet, and the rest are local, installed in our machines. The manual work has been to build the table with possible interpretations of the prepositions and case suffixes.

The paper is structured as follows. Section 2 presents the method for disambiguation in detail. Section 3 introduces the interpretations for the case suffix and the prepositions. The results are shown in Section 4, which are further discussed in Section 5. Finally, section 6 presents the conclusions and future work.

2 Method for disambiguation

The goal of the method is to disambiguate between the possible interpretations of a case suffix appearing in any text. We have taken as the target text the definitions from a monolingual Basque dictionary *Euskal Hiztegia*, *EH* in short (Sarasola, 1996). The method consists on five steps:

- Extraction of the definitions in *EH* where the target case suffix occurs.
- Search of on-line Spanish and English dictionaries to obtain the translation equivalent of the definitions.
- Extraction of the target preposition from the translation definitions.
- Disambiguation based on the intersection of the interpretations of case suffix and prepositions.

We will explain each step in turn.

2.1 Extraction of relations from *EH*

Given a case suffix, in this step we will search the *EH* dictionary for occurrences of the case suffix. We first lemmatize and perform morphological analysis of the definitions (Aduriz et. al, 1996). The definitions that contain the target case suffix in a morphological analysis are extracted, storing the following information: the Basque dictionary entry of the definition, the lemma that has the case suffix, the case suffix, and the following lemma.

Below we can see a sample definition, its lemmatized version, and the two triples extracted from this definition. The occurrences of the instrumental *-z* are shown in bold.

Ildo iz. A1 **Goldeaz** lurra **irauliz** egiten den irekidura luzea[1]

```
/<@@lema    ildo>/<ID>/
/<@@Adiera_string A1.>/<ID>/
/<@@Kategoria     iz. >/<ID>/
"<Goldeaz>"
  "golde"  IZE ARR DEK INS NUMS MUGM
"<lurra>"
  "lur"  IZE ARR DEK ABS NUMS MUGM
"<irauliz>"
  "irauli"  ADI SIN AMM PART DEK INS MG
"<egiten>"
  "egin"  ADI SIN AMM ADOIN ASP EZBU
"<den>"
  "izan"  ADL A1 NOR NR_HU ERL MEN ERLT
"<irekidura>"
  "irekidura"  IZE ARR DEK ABS MG
"<luzea>"
  "luze"  ADJ IZO DEK ABS NUMS MUGM
"<$.>"
  PUNT_PUNT

golde#INS#lur[2]
irauli#INS#egin
```

Extracting lemma-suffix-lemma triples in this simple way leads to some errors (cf. section 5.1). For instance, the first triple should rather be the dependency *golde#INS#irauli* (*plow#with#turn*, to be read in reverse order). We will see that even in this case we will be able to obtain correct translations and disambiguate the preposition correctly. Nevertheless, in the future we plan to use a syntactic parser to identify better the lemmas that are related by the case suffix.

2.2 Search for Spanish/English translations

After we have a list of entries in the Basque dictionary that contain the lemma-suffix-lemma triple, we search for their equivalent definitions in Spanish and English. We first look up the entry in the bilingual dictionary, and then retrieve the

[1] The literal translation of the definition is the following : *furrow, a long trench produced turning over the ground with a plow*.

[2] The translation of the first triple is *plow#with#ground*, to be read on reverse. The translation of the second is *turn#NULL#produce*, to be also read on reverse. In this second triple the instrumental case suffix is not translated explicitly by a preposition, but by a syntactic construct.

definitions for each of the possible translations from the monolingual dictionaries.

We use two bilingual and 6 monolingual Machine Readable Dictionaries: *Morris* Basque/English dictionary (Morris, 1998) *Elhuyar* Basque/Spanish dictionary (Elhuyar, 1996); English monolingual on-line dictionaries are: Cambridge (online), Heritage (online), and Wordsmyth (online); and Spanish monolingual on-line dictionaries are: Colmex (online), Rae (online), and Vox (online). The Basque dictionary and the bilingual dictionaries are stored in a local server, while the monolingual dictionaries are accessed from the Internet using a wrapper.

The incomplete list of the translation of *ildo* (*furrow* in English, *surco* in Spanish) is shown below. Note that we got two different definitions for *surco,* coming from different Spanish dictionaries.

```
furrow#A long , narrow , shallow
  trench made in the ground by a
  plow

surco#Excavación alargada , angosta y
  poco profunda que se hace
  paralelamente en la tierra con el
  arado , para sembrarla después

surco#Hendedura que se hace en la
  tierra con el arado
```

2.3 Extraction of Spanish/English equivalent relations

Given a list of definitions in Spanish and English, we search in the definition the translation of the Basque triple found in step 2.1, that is, we look for a triple of consecutive words where the first word is the translation of the last word in the Basque triple, the second word is a preposition (which corresponds to the Basque suffix) and the third word is the translation of the first word in the Basque triple. Between the preposition and the last word in the triple we allow for the presence of a determiner or an adjective in the text. More complex patterns could be allowed, up to full syntactic analyses, but at this point we follow this simple scheme.

Below we can find the triples for *golde#INS#lur*, obtained from the three definitions above. One triple is obtained twice from two different definitions.

```
furrow#ground#by#plow
surco#tierra#con#arado
surco#tierra#con#arado
```

Definitions that do not have a matching triple are discarded, leaving Basque triples without matching triple ambiguous. For instance we could not find triples for *irauli#INS#egin*(cf. example in section 2.1). The instrumental suffix is sometimes translated without prepositions (in this case "... *made turning* ...").

Looking up the bilingual dictionaries for translation requires lemmatization and Part of Speech tagging. For English we use the TnT PoS tagger (Brants, 2000) and WordNet for lemmatization (Miller et al., 1990). For Spanish we use (Atserias et al., 1998).

2.4 Disambiguation

For each Basque case suffix, Spanish preposition and English preposition we have a list of interpretations (cf. Table 1). We assign the interpretations of the preposition to each Spanish/English triple. The intersection of all the interpretations is assigned to it.

Continuing with out example, we can see that the intersection between the interpretations of the English *by* preposition (three interpretations) and the interpretations of the Spanish *con* preposition (four interpretations) are **manner** and **instrument**. Therefore, we can say that the Basque *instrumental* case interpretation in this case will be **manner** or **instrument**.

```
furrow#ground#by a#plow#
  manner instrument during-time
surco#tierra#con el#arado#
  manner instrument cause containing

golde#INS#lur#instrument manner
```

3 Interpretations for the instrumental case suffix and equivalent prepositions

The method explained in the previous section is

fully automatic, and it only requires the list of interpretations for each case suffix and preposition. In this work, we want to evaluate if the overall approach is feasible, so we selected Basque as the target language and a single case suffix, *-z* the instrumental case. Table 1 shows the list of possible interpretations and Table 2 and 3 examples for each interpretation.

The sources for the interpretations of the instrumental case have been a grammar of Basque (Euskaltzaindia, 1985) and a bilingual dictionary (Elhuyar, 1996). Possible interpretations for Spanish and English prepositions have been taken from an English dictionary (Cambridge, online), a Spanish dictionary (Vox, online) and a Spanish grammar (Bosque & Demonte, 1999).

For this work we have taken a descriptive approach, but other more theoretically committed approaches are also possible. The overall method is independent of the set of interpretations, as it only needs a table of possible interpretations in the style of Table 1. Section 5.4 further discusses other alternatives.

In order to disambiguate the occurrences of the instrumental case suffix we have taken the Spanish and English translations for this case suffix. The list of possible translations is preliminary and covers what we found necessary to make this experiment. Table 1 shows the list of prepositions and interpretations for Spanish and English. Examples of the interpretations can be found in Table 2. The Spanish preposition *de* had the same interpretations as the instrumental case suffix (cf. Table1), so it was discarded.

4 Results

The instrumental case occurs in 4,004 different definitions in the *EH* dictionary. The algorithm in Section 2 was applied to all these definitions, yielding a result for 125 triples, 3.1% of the total. The triples for which we had an answer were tagged by hand independently, i.e. not consulting the results output by the algorithm. The hand-tagged set constitutes what we call the gold standard.

A single linguist made the tagging, consulting other teammates when in doubt. Apart from marking the interpretation, there were some other special cases.

1. In some of the examples, the instrumental case was part of a more complex scheme, and was tagged accordingly:
   - Part of a postposition (XPOST), e.g. *-en bidez* (by means of) or *-en ordez* (instead of).
   - Part of a conjunction (XLOK), e.g. *batez ere* (specially).
   - Part of a compounded suffix *–zko* (XZKO), which results from the aggregation of the instrumental *–z* with the location genitive *-ko*.
2. There were three errors in the lemmatization process (XLEM), due to lexicalized items, e.g. *gizonezko* (meaning male person).
3. Finally, the relation in the definition was sometimes wrongly retrieved, e.g.
   - The triple would contain the determiner or an adjective instead of the dependencies. We thought that the algorithm would be able to work well even with those cases, so we decided to keep them.
   - The triple contains a conjunction (X): these were tagged as incorrect.

Table 4 shows the amount of such cases, alongside the frequency of each interpretation. The most frequent interpretation is *instrument*. In seven examples, the linguist decided to keep two interpretations: *instrument* and *manner*. In a single example, the linguist was unable to select an interpretation, so this example was discarded.

The output of the algorithm was compared with the gold standard, yielding the accuracy figures in Table 5. An output was considered correct if it yielded at least one interpretation in common with the gold standard. The accuracy is given for each dictionary in isolation, or merging all the results (as mentioned in section 2, when two dictionaries propose interpretations for the same triple, their intersection is taken). The remaining ambiguity is 3.1 overall.

	Basque	English				Spanish			
	-z (ins.)	of	by	with	in	**de**	con	a	en
theme	x	x			x	**x**		x	
during-time	x	x	x			**x**			
instrument	x		x	x	x	**x**	x		x
manner	x		x		x	**x**	x	x	x
cause	x	x		x	x	**x**	x		
containing	x	x		x	x	**x**	x		
matter	x	x				**x**			

Table 1: interpretations for the instrumental case in Basque and its equivalents in English and Spanish.

	Basque	**English**
theme	Seguru nago horretaz Matematikaz asko daki	I'm sure of that He's an expert in maths
during-time	Arratsaldez lasai egon nahi dut Gauez egin dut	I like to relax of an evening I did it by night
instrument	Autobusez etorri naiz Belarra segaz moztu Euskaraz hitz egin	I have come by bus To cut grass with a scythe To speak in Basque
manner	Animali baten hestea betez egindako haragia Ahots ozen batez	A meat preparation made by filling an animal intestine In a loud voice
cause	Haren aitzakiez nekatuta nago Beldurrez zurbildu Kanpoan lan egitea baztertu zuenez, lan-aukera ederra galdu zuen	Sick of his excuses To turn white of fear In refusing to work abroad, she missed an excellent job opportunity
containing	Edalontzia ardoz beteta dago Txapelaz dagoen gizona Ilez estalia	The glass is full of wine The man with the beret on Cover in hair
matter	Armairua egurrez egina dago	The wardrobe is made of wood

Table 2: examples in Basque and English for the set of possible interpretations.

	Basque	**Spanish**
theme	Mariaz aritu dira Honetaz ziur naiz	Han mencionado a Maria Estoy seguro de esto
during-time	Gauez egin dut	Lo he hecho de noche
instrument	Belarra segaz moztu Euskaraz hitz egin Hiria harresiz inguratu dute	Cortar la hierba con la guadaña Hablar en vasco Han cubierto la ciudad de murallas
manner	Oinez etorri zen Ahots ozen batez Bere familiaren laguntzaz erosi zuen Berdez margotzen ari dira	Vino a pie En voz alta Lo compró con la ayuda de su familia Lo estan pintando de verde
cause	Beldurrez zurbildu Maitasunez hil	Con el miedo me quedé pálido Morir de amor
containing	Edalontzia ardoz beteta dago Txapelaz dagoen gizona ikusi dut	El baso esta lleno de vino He visto a un hombre con boina
matter	Armairua egurrez egina dago	El armario está hecho de madera

Table 3: examples in Basque and Spanish for the set of possible interpretations.

Table 4 also shows the most frequent baseline (MF), constructed as follows: for each occurrence of the suffix, the three most frequent interpretations are chosen. The accuracy of this baseline is practically equal to that of the algorithm. Note that the frequency is computed on the same sample where it is applied, yielding better results than it should.

5 Discussion

The obtained results show a very good accuracy, leaving a remaining ambiguity of 3.1 results per example. This means that we were able to discard an average of 4 readings for each of the examples, introducing only 5.5% of error. The results are practically equal to the most frequent baseline, which is usually hard to beat using knowledge-based techniques.

Coverage of the method is very low, only 2.3%, but this was not an issue for us, as we plan to couple this method with other Machine Learning techniques in a bootstrapping framework. Nevertheless, we are still interested in increasing the coverage, in order to obtain more training data.

Next, we will analyze more in depth the causes of the low coverage, the sources of the errors and ambiguity and the interpretations of case suffixes and prepositions.

5.1 Sources of low coverage

As soon as we started devising this method, it was clear to us that the coverage will be rather low. The main reason is that different dictionaries tend to give different details in their definitions, or use differing paraphrases. This fact is intrinsic to our method, and accounts for the large majority of missing answers.

On the other hand, the simple method used to find triples means that a change in the order of the complements will cause our method to fail looking for a translation triple. Syntactic analysis, even shallow parsing methods, will help increase the coverage.

Another source of discarded triples are the cases where the suffix is not translated by a preposition, e.g. the relation is carried out by a subject or direct object. When syntactic analysis is performed, we also plan to incorporate the interpretations of the other syntactic relations.

#	interpretation
8	XPOST
1	XLOK
12	XZKO
3	XLEM
9	X
1	No interpretation
34	**Total discarded**
37	instrument
35	containing
7	instrument manner
6	manner
5	theme
1	cause
0	matter
0	during-time
91	**Total kept**

Table 4: frequency of tags in gold standard.

Dictionary	total	correct	accur.	ambig.
cambridge	16	15	0.938	4.0
Am. heritage	34	32	0.941	3.2
wordsmith	26	26	1.000	3.7
Colmex	10	9	0.900	2.6
vox_ya	7	7	1.000	2.8
Rae	26	25	0.962	2.8
overall	91	86	0.945	3.1
MF baseline	91	85	0.934	3.0

Table 5: results for each of the dictionaries, overall combination for all and the most frequent baseline.

5.2 Sources of error

Only five errors we made by the algorithm, which were caused by the wrong triple pairings, especially when the Basque triple contained a determiner instead of the related word. Examples:

- xixta/prick: punta **batez** osatua/made by **a** needle
- luma/feedle: odi **batez** osatua/wake made by **a** submarine

There errors could be avoided using a syntactic parser. Other wrong pairings were caused by

errors in the English PoS tagger, or chance made the algorithm find an unrelated definition.

5.3 Remaining ambiguity

The amount of readings left by our method in this experiment is rather high, around 3.1 readings compared to 7 possible readings for the instrumental. This is a strong reduction but we would like to make it even smaller.

We plan to study which is the source of the residual ambiguity. Alternative sets of interpretations (cf. Section 5.4) with coarser grained differences and smaller ambiguity, could yield better results. Another alternative is to explore more infrequent translations of the case suffixes, which might yield a narrower overlap. This is the case for the instrumental case suffix being translated with *from*, *up*, etc.

5.4 Interpretations of case suffixes and prepositions

Different authors give differing interpretations for prepositions. It has been our choice to take a descriptive list of possible interpretations from a set of sources, mainly dictionaries and grammar books.

This work covers only the instrumental case suffix and its translations to English and Spanish. If tables for all case suffixes and prepositions were built, the method could be applied to all case suffixes and prepositions, yielding disambiguated relations in all three languages.

More theoretically committed lists of interpretations (Dorr et al., 1998; Civit et al., 2000; Sowa, 2000) should also be considered, but unfortunately we have not found a full account for all prepositions. If such a full table of interpretations existed, it could be very easy to apply our method, and obtain the outcome in terms of these other interpretations.

6 Conclusion and further work

This paper presents preliminary experiments in the use of translation equivalences to disambiguate prepositions or case suffixes. The core of the method is to find translations of the occurrence of the target preposition or case suffix, and assign the intersection of their set of interpretations. The method is fully automatic, given a table with prepositions and their possible interpretations.

We have tested this method on the occurrences of the Basque instrumental case *-z* in the definitions of a Basque dictionary. We have searched the translations in the definitions from 3 Spanish and 3 English dictionaries.

The results have been that we are able to disambiguate with 94.5% accuracy 2.3% of those occurrences (up to 91). The ambiguity is reduced from 7 readings down to 3.1. We think that these are very good results, especially seeing that there is room for improvement.

More specifically, we plan to apply surface syntactic analysis to better extract the dependency relations, which is the main source of errors. We would like to study other inventories of preposition interpretations, both in order to have better theoretical foundations as well as to investigate whether coarser grained distinctions would lead to a reduction in the ambiguity.

In the future, we plan to explore the possibility to feed a Machine Learning algorithm with the automatically disambiguated examples, in order to construct a full-fledged disambiguation algorithm following a bootstrapping approach. On the other hand, we would like to apply the method to the set of all prepositions and case suffixes, and beyond that to all syntactic dependencies. The results will be directly loaded in a Lexical Knowledge Base extracted from the Basque dictionary (Ansa et al., in prep.).

We also plan to explore whether this method can be applied to free running text, removing the constraint that the translations have to be definitions of the equivalent word.

Finally, this technique could be coupled with techniques that make use of the semantic types of the words in the context.

Overall, we found the results are very encouraging given the simple techniques used, and we think that it shows great potential for improvement and interesting avenues for research.

Acknowledgments

Mikel Lersundi and David Martinez were supported by Basque Government grants AE-BFI:98.217 and AE-BFI:01.2485. This work was

partially funded by the MCYT HERMES project (TIC-2000-0335) and the EC MEANING project (IST-2001-34460).

References

Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N., Urizar R., 1996, "EUSLEM: A Lemmatiser / Tagger for Basque" *Proc. Of EURALEX'96*, Göteborg (Sweden) Part 1, 17-26.

Ansa O., Arregi X., Lersundi M., "A Conceptual Schema for a Basque Lexical-Semantic Framework" (in preparation)

Bosque, I., Demonte, V., 1999, *Gramatica descriptiva de la lengua Española*, Espasa, Madrid.

Brants, T. 2000. *TnT - A Statistical Part-of-Speech Tagger*. In Proceedings of the Sixth Applied Natural Language Processing Conference, Seattle, WA.

Cambridge, online. *Cambridge International Dictionary of English* http://dictionary.cambridge.org/

Civit, M., Castellón, I., Martí, M.A. and Taulé, M., 2000, "LEXPIR: a verb lexicon for Spanish" *Cuadernos de Filología Inglesa*, Vol. 9.1. *Corpus-based Research in English Language and Linguistics*, University of Granada.

Colmex, online. Diccionario del español usual en México (Colmex) http://mezcal.colmex.mx (also accessible from http://www.foreignword.com

Dorr, Bonnie J., Nizar Habash, and David Traum, 1998, "A Thematic Hierarchy for Efficient Generation from Lexical-Conceptual Structure," in *Proceedings of the Third Conference of the Association for MT in the America's*, Langhorne, PA, pp. 333--343

Elhuyar, 1996, *Elhuyar Hiztegia,* Elhuyar K.E., Usurbil.

Euskaltzaindia, 1985, *Euskal Gramatika Lehen Urratsak-I* (EGLU-I), Euskaltzaindia, Bilbo.

Heritage, online. The American Heritage® Dictionary of the English Language. http://www.bartleby.com/61

J. Atserias, J. Carmona, I. Castellon, S. Cervell, M. Civit, L. Marquez, M.A. Marti, L. Padro, R.Placer, H. Rodriguez, M. Taule & J. Turmo "Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text" First International Conference on Language Resources and Evaluation (LREC'98). Granada, Spain, 1998.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. *Five Papers on WordNet*. Special Issue of International Journal of Lexicography, 3(4).

Morris M., 1998, *Morris Student dictionary*, Klaudio Harluxet Fundazioa, Donostia.

Rae, online. *Diccionario de la Real Academia de la Lengua* http://buscon.rae.es/drae/drae.htm

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

John F. Sowa, 2000, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA

John F. Sowa, ed. (1992) *Knowledge-Based Systems*, Special Issue on Conceptual Graphs, vol. 5, no. 3, September 1992

Vox, online. *Diccionario General de la lengua española VOX* http://www.vox.es/consultar.html

Wordsmyth, online. The Wordsmyth Educational Dictionary-Thesaurus http://www.wordsmyth.net