Exploring portability of syntactic information from English to Basque

Eneko Agirre, Aitziber Atutxa, Koldo Gojenola, Kepa Sarasola[1]

IXA NLP Group
University of the Basque Country
{eneko,jibatsaa,jipgogak,jipsagak}@si.ehu.es

Abstract

This paper explores a crosslingual approach to the PP attachment problem. We built a large dependency database for English based on an automatic parse of the BNC, and Reuters (sports and finances sections). The Basque attachment decisions are taken based on the occurrence frequency of the translations of the Basque (verb-noun) pairs in the English syntactic database. The results show that with this simple technique it is possible to transfer syntactic information from a language like English in order to make PP attachment decisions in another language, in this case Basque.

Introduction & Motivation

This work is comprised in a broader endeavor in the context of the MEANING project (Rigau et al., 2002), with the goal of exploring the possibility of porting linguistic knowledge acquired in one language to another. This portability issue could be especially relevant for minority languages with few resources like Basque. Hence the main motivation underlying this experiment is to explore ways to overcome the limitations originated by the lack of resources. If we were able to transfer some of the linguistic knowledge available for English to other languages we would effectively reduce some of the restrictions in these languages (small corpora, lack of hand annotated corpora, etc.).

Cross-language information transfer is not something new, however most of the work done relies on the usage of parallel corpora (Hwa et al 2002), which are difficult to find, specially for lesser studied languages. This is one of the reasons that lead us to consider the usage of comparable corpora, since it is easier to obtain.

Another noteworthy aspect is the pair of languages selected for the experiment: English and Basque. Hypothetically, these two languages are linguistically distant enough to make this work extensible to any other language pair. The following could be a short characterization of the most relevant differences between the two languages:

- English is a head initial language with an SVO word order, while Basque is a head final free word order language.
- English does not show strong morphology, while Basque does.
- English is not a pro-drop language, and Basque is a three-way pro-drop language.
- English and Basque do not belong to the same typological family.

We chose the PP attachment problem in order to explore the portability issue. This problem is especially hard for free word order languages like Basque. Our current partial parser makes attachment decisions based on certain rules and heuristics.

Our experiment has been devised to transfer attachment information coming from English parsed data making the attachment decisions for Basque based on this transferred information. The basic idea behind the system presented here is that verbs show certain preferences on the nouns they appear with. Therefore, if we have a sentence with two verbs, and some noun phrases, one of the verbs will show higher preference for some of the noun phrases while the other verb will show higher preference for the others. We will make one assumption beyond this basic idea, the assumption being that these preferences happen and to some extent can be transferred cross-linguistically (Agirre et al. 2003). Note that this is a preliminary work so at this point we aim to keep the system as simple as possible. Thus, higher co-occurrence of the verb and a noun will be taken to be higher preference of that verb over that noun.

The results obtained suggest that cross language transferring of knowledge acquired from comparable corpora, is worth pursuing. Even employing a very simple machinery, results seem very promising.

Outline of the method

Our starting point was the Basque parser described in (Aldezabal et al 2000). This parser uses a unification grammar to build syntactic structures. Having a sentence it chunks it into phrases, finds the head of each phrase and then applying certain rules and heuristics tries to link those heads to the different verbs belonging to the sentence.

To test our attachment system, we selected sentences with two verbs, and used the Basque parser to obtain information about the chunks in the sentences. The attachment information provided by the parser is discarded, maintaining only the chunking information. The heads of the noun groups are extracted, and a set of all possible syntactically dependent (verb-noun) pairs are constructed. The goal was to select for each noun which verb should it be attached to from the two possibilities.

[1] Authors listed in alphabetical order.

The method works as follows. We first obtain from the Basque sentence the verbs and surrounding heads. We translate them into English using a bilingual dictionary, and for each (verb-noun) Basque pair we search all possible translation combinations in the dependency database built from an automatically parsed English corpus.
Take for example this Basque sentence,

Lendakariak hautezkundeak irabazi zituen botoen %60 lortuz inbersoreen artean.
The president won the election obtaining 60% of the votes among the investors.

The verbs and heads obtained by the Basque parser/chunker are the following:
NP-ergative(**lendakaria**) NP-absolutive(**elections**)
PPabsolutive(**boto**) PP-distributive(**Inbertsore**)
V1(**irabazi**) V2(**lortu**)

We translate all the nouns and verbs.

NP-ergative(**lendakaria**): President, chairman (ncsubj)
NP-absolutive(**elections**): poll, election
Ppabsolutive(**boto**): vote,vow
PP-distributive(**Inbertsore**): investor, shareholder
V1(**irabazi**): to win, to earn, to gain
V2(**lortu**): to get, to obtain, to attain

All possible English noun-verb pairs are created with the corresponding English relation or preposition for each Basque case, for example for *lendakari-irabazi* vs. *lendakari lortu*:

$$\left\{\begin{array}{l}\text{win-President-ncsubj}\\ \text{earn-President-ncsubj}\\ \text{gain-President-ncsubj}\\ \text{win-President-ncsubj}\\ \text{earn-President-ncsubj}\\ \text{gain-President-ncsubj}\end{array}\right\} \text{vs.} \left\{\begin{array}{l}\text{get-President-ncsubj}\\ \text{obtain-President-ncsubj}\\ \text{attain-President-ncsubj}\\ \text{get-President-ncsubj}\\ \text{obtain-President-ncsubj}\\ \text{attain-President-ncsubj}\end{array}\right\}$$

Note that we only search for the English verb and noun translations occurring in a direct syntactic dependency (moreover, we search for an English syntactic dependency equivalent to the Basque one). We collect and add the frequencies of all translated English pairs for each (verb-noun) Basque pair. In order to select the correct attachment for each noun, the mutual information of the two (verb-noun) pairs are compared. This way we normalize over the amount of translations, and also over the occurrences of the English translations in the target corpus.

$$MI(BV,BN)= \log \left(\frac{P(\text{any-EVT, any-ENT})}{P(\text{any-EVT})*P(\text{any-ENT})} \right)$$

P(any-EVT, any-ENT) corresponds to the probability of finding any translation of the Basque verb with any translation of the Basque noun in the English corpus.
P(any-EVT) corresponds to the probability of finding any translation of the Basque verb in the English corpus, and P(any-ENT) corresponds to the probability of finding any translation of the Basque noun in the English corpus.

A higher Mutual information value (maintaining the same syntactic relation in both languages) is taken as an indicative of a stronger preference between the head and one of the verbs, the one that will be selected.

As mentioned above, we intended to keep the same syntactic relation across both languages when searching. For that, we employed the information provided by the Basque morphological case attached to each noun as an indicative of this relation. There is an equivalence between Basque morphological cases and English prepositions. This equivalence is not one-to-one, thus each Basque case will have several English prepositions as possible translations, and the opposite. Bilingual dictionaries do not contain such information, so we used the equivalence table described in (Lersundi et al 2002). In this equivalence table all possibilities are listed, even low frequency and rare ones.

The RASP parser does not incorporate exhaustive information about multiwords, and therefore we included a heuristic method to search for them. So for example, the Basque verb *bilatu* is translated as "*look up*" in English. In "look up in the dictionary", we would like to have a dependency between "*look up*" and *dictionary*. The parser will find that *dictionary* is a dependent of *look*, through the preposition *in*, and *up* will appear as a particle of look in another relation. The heuristic applied consists of searching for the pair *look-dictionary* related through the preposition *in*, and checking that *up* also appears as a particle of look in the same sentence. Still, certain multiwords need more complex processing. For instance the Basque verb *garestitu* is a result of an incorporation process and it is translated as "*to make more expensive*". At this point we are not treating such multiword translations, and they would return a 0 frequency on the search.

The corpora

The English corpora are conformed by the BNC (http://www.natcorp.ox.ac.uk) and the Reuters newswire finances and sports corpus. The English parser used is the RASP dependency parser developed by Carroll and Briscoe (2001). Notice that a dependency parser links the head of the phrases to the verbs in contrast to what constituency parsers do, linking whole phrases to the verbs. This is a relevant feature because it will facilitate the searches.

The information we obtain from the English RASP parser for the following sentence is displayed in Figure1 a); *1 in 500 Londoners are believed to be infected*

The first dependency in Figure 1 a) represents the subject syntactic relation between the head *believe* and the dependent *Londoner (ncsubj)*. Being this a passive sentence RASP also provides information about the internal relation between *believe* and *Londoner*, where *Londoner* is not the subject but the object[2] of *believe (obj)*. The second dependency corresponds to the clausal relation between *believe* as a head and *infected*, that is, what we *believe* is that [there is somebody to be *infected*].The third

[2] One believes Londoners (obj) to be infected.

```
a)(1_MC |in_II| 500_MC |Londoner+s_NN2| |be+_VBR| |believe+ed_VVN| |to_TO| |be_VB0| |infect+ed_VVN|)

 (|ncsubj| |believe+ed:6_VVN| |Londoner+s:4_NN2| |obj|)
 (|clausal| |believe+ed:6_VVN| |infect+ed:9_VVN|)
 (|ncsubj| |infect+ed:9_VVN| |Londoner+s:4_NN2| |obj|)
 (|ncmod| |in:2_II| |1:1_MC| |500:3_MC|)
 (|ncmod| _ |Londoner+s:4_NN2| |1:1_MC|)
 (|aux| _ |infect+ed:9_VVN| |be:8_VB0|)
 (|aux| _ |believe+ed:6_VVN| |be+:5_VBR|)

b) code lem1     sufx1 posit PoS1 w_pos1 lem2    sufx2 posit2 PoS2  w_pos2  rel    int_pos prep sent parsed  txt
  | 70 | believe | ed |  6  |  V  | VVN| Londoner| s   |  4  |  N  | NN2| ncsubj  | obj  | -   | 22 | 01#22 | A00#22|
  | 71 | believe | ed |  6  |  V  | VVN| infect  | ed  |  9  |  V  | VVN| clausal | -    | -   | 22 | 01#22 | A00#22|
  | 72 | infect  | ed |  9  |  V  | VVN| Londoner| s   |  4  |  N  | NN2| ncsubj  | obj  | -   | 22 | 01#22 | A00#22|
  | 73 | 1       | -  |  1  |  M  | MC | 500     | -   |  3  |  M  | MC | ncmod   | -    | in  | 22 | 01#22 | A00#22|
  | 74 | Londoner| s  |  4  |  N  | NN2| 1       | -   |  1  |  M  | MC | ncmod   | -    | -   | 22 | 01#22 | A00#22|
  | 75 | infect  | ed |  9  |  V  | VVN| be      | -   |  8  |  V  | VB0| aux     | -    | -   | 22 | 01#22 | A00#22|
  | 76 | believe | ed |  6  |  V  | VVN| be      | -   |  5  |  V  | VBR| aux     | -    | -   | 22 | 01#22 | A00#22|
```

Figure 1: Output of the RASP parser, and the way we code it in the database

provided dependency expresses the subject relation between *infected* as a head and *Londoner* as a dependent. The *obj* value inside that dependency reveals *Londoner* as the *internal object* of *infected*. The fourth dependency represents information about a modification relation. According to the parser *500* is a modifier *(ncmod)* of *Londoners*, and this relation materializes through the preposition *in*.

In order to make efficient searches in the English parsed corpora we created a database for each corpus (BNC, Reuters_sports, Reuters_finances), where each tuple represents a dependency syntactic relation between a verb and a dependent (head of a noun or prepositional phrase). Figure 1b) illustrates the encoding of the example above in our database. For instance take the first dependency in Figures 1 a) and b). The specific syntactic relation between the head and the dependent (ncsubj in this case) is stored in the twelfth field of the tuple. The second and sixth fields maintain information about the head and the dependent respectively (believe and Londoner). The database also stores information relative to the PoS of each words (fifth and tenth fields respectively, V for believe and N for Londoner). This way we can select from all the relations, just those where the first lemma is a verb. Sentential index information (the last three fields, 22, 01#22,A00##22) is encoded in order to relate the parsed sentences to the original text sentences. The fourth and ninth fields store the positional information of the words in the sentence (sixth position in the sentence for believe and fourth for Londoners). This information is currently being used to check that two words conform a multiword or not (as in the "look up" example in the previous section).

All in all, the database contains 47,145,584 syntactic relations from BNC, 1,439,445 from Reuters Sports and 9,858,633 from Reuters Finances. From these relations, 10,447,129 relations are verb-noun dependencies in BNC, 366,805 in Reuters Sports, and 2,547,843 in Reuters Finances.

Design of the experiment

The Basque corpus used comes from a newspaper and it refers to news from several months of the year 2000 (33.669 sentences) in different domains (culture, sports, finances, politics, etc.). From this corpus we chose sentences having two verbs, where one of the verbs belonged to a list of 12 verbs, which are roughly equivalent to the following: *win, loose, increase, decrease, tie, train, play, sign up, run, injure, reduce, classify*.

The criteria to select these verbs was their high relation either to sports or to both sports and finances domains. This correlation was estimated by looking at the target sections of the Basque news paper where these verbs appear more frequently. The selection of 7 of these verbs was done in coordination to the other partners in the MEANING project, as described in (Magnini et al. 2004) to be used in several experiments which included Basque, Italian, English, Spanish and Catalan. The domain-related frequencies mentioned where calculated for each of these languages and the selection of 7 verbs corresponds to the top 7 verbs after merging the top lists of these languages. The other 5 verbs belong to the top Basque list for verbs with high relation to the sports domain (to sign up, to run, to injure, to reduce, to classify)[3].

The number of sentences obtained containing two verbs including one of these 12 verbs is 386, where we found 1,278 syntactic relations. Over these 1,278 relations, 400 were manually tagged, which constituted the gold-standard for this task. From these relations 91 occurred within the finances section of the Basque news paper, 190 within the sports section. That is, from 400, 281 belong to either sports or finances sections.

The results

Table 1 presents the results obtained in terms of precision and coverage. The table is divided in two regions with respect to the first column. The first region (in white) corresponds to the values obtained using Mutual Information (MI). The second region (in gray) corresponds to using the plain frequencies, that is, the hits on the English database (freq). Each of this regions comprises 3 rows, which correspond to the sections of the Basque newspaper where the target example was found:

[3] Some of the top verbs had to be excluded because they show incorporation in Basque, and their equivalent in English corresponds to a multiword.

all (all sections), fin (finance section) and spo (sports section).

Regarding the columns, #rel shows the amount of relations in each row, BNC means that the English corpora used to search and acquire the English syntactic information was the BNC, while REU-SPO and REU-FIN indicate that the sports section or finance section of the Reuters corpus was used.

			BNC		REU-SPO		REU-FIN	
	Sect.	#rel	prec	cov	Prec	cov	prec	cov
MI	all	400	0,60	0,76	**0,67**	0,51	0,62	0,68
	fin	91	**0,64**	0,78	**0,64**	0,46	0,61	0,75
	spo	190	0,61	0,69	**0,72**	0,58	0,58	0,63
freq	all	400	0,57	0,76	0,59	0,51	0,58	0,68
	fin	91	0,55	0,78	0,60	0,46	0,57	0,75
	spo	190	0,52	0,69	0,52	0,58	0,56	0,63

Table 1: Results in terms of precision and coverage

The results in Table 1 show that the system performs with precisions well above the random baseline (0.5 in this case) for all combinations of source and target corpora. We can also see that in all cases MI attains better results than using the raw frequencies.

The best precision value is 0.72, and was obtained when searching over Reuters sports for deciding on (verb-noun pairs) relations belonging to the Basque sports section, showing that narrowing the domain of the texts is useful to improve the results. Searching in Reuters-sports also provides the best results even to make decisions over relations belonging to the finances and any section of the Basque news paper. The reason could be that the verbs selected are highly tied to the sports domain, and even within some other sections we still get better results searching on the sports corpus.

Conclusions and further work

This work aimed at exploring the portability of linguistic knowledge from one language to another The results reported suggest that the transfer is possible, as a very simple technique which searches on English dependencies is able to make valuable PP attachment decisions in Basque. This could be especially helpful to deal with the structural ambiguity problems in Basque that scrambling poses to an already difficult task like PP attachment.

The system we have developed uses comparable corpora, as opposed to parallel corpora, which makes it very suitable for languages where parallel corpora is not easy to find., and also allows us to get large amounts of corpora linked to any target domain.

For our future work we plan to improve the precision of the system with a better treatment of multiwords (including Named-Entities) and using other dependencies in the source sentence in order to narrow the search space in English. For the same reason, we would also like to include frequency information in the preposition-postposition equivalence tables. Besides we plan to combine this multilingual system with the heuristics already coded in the Basque parser.

We also plan to extend this work to the acquisition of selectional preferences in the target language, and use it in the source language. Finally, we would like to take the attachment decisions of all the phrases in the sentence at the same time, in order to take the best decision overall.

Acknowlegments

The authors would like to thank Philip Resnik for his comments and ideas. The work is partially funded by the European Commission (MEANING project, IST-2001-34460), and the Spanish MCYT (HERMES project, TIC-2000-0335-C03-03).

References

Agirre E., Aldezabal I., Pociello E. (2003). A pilot study of English Selectional Preferences and their Cross-Lingual Compatibility with Basque. In the proceedings of the International Conference on Text Speech and Dialogue . Czech Republic.

Hwa, R., Resnik, P., Kolak, O., Weinberg, A. (2002). Evaluating Translational Correspondence using Annotation Projection. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. Philadelphia, USA.

Aldezabal I., Gojenola K., Sarasola K. (2000) A Bootstrapping Approach to Parser Development In Proceedings of the International Workshop on Parsing Technologies (pp. 17-28). Trento, Italy.

Briscoe, E. and J. Carroll (2002). Robust accurate statistical annotation of general text. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, (pp. 1499-1504) Las Palmas, Canary Islands, Spain.

Lavoi, B., White, M., Koreslsky, T., (2001). Including Lexico-Structural Transfer Rules form Parsed Bi-texts. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. DDMT Workshop. Toulouse, France.

Lersundi, M., Agirre, E. (2003). Semantic interpretations of postpositions and prepositions: a multilingual inventory for Basque, English and Spanish. In proceedings of the workshop on The linguistic dimensions of prepositions and their use in computational linguistics formalisms and applications. Tolouse, France.

Magnini Bernardo Magnini, Octavian Popescu, Jordi Atserias, Eneko Agirre, Eli Pociello, German Rigau, John Carroll, Rob Koeling (2004). Cross-Language Acquisition of Semantic Models for Verbal Predicates. In Proceedings of LREC. Lisbon, Portugal.

Rigau G., Magnini B., Agirre E., Vossen P. and Carroll J (2002). MEANING: A Roadmap to Knowledge Technologies, in Proceedings of the COLING Workshop, A Roadmap for Computational Linguistics. Taipei, Taiwan.