PP Attachment for Basque Based on English parses

Eneko Agirre, Aitziber Atutxa, Koldo Gojenola,
Kepa Sarasola, David Terrón
IXA NLP Group
University of the Basque Country
{jibatsaa}@si.ehu.es

Abstract[1]

This paper explores the viability of porting lexico-syntactic information from English to Basque in order to make PP attachment decisions. Basque is a free constituent order language where PPs in a multiple-verb sentence can be attached to any of the verbs. We compared a system trained in non-ambiguous (single verb) Basque sentences with another system trained on a parsed English corpus (BNC). The results show that crosslingual learning is possible though the performance is lower than in the monolingual version. But when combined, an increase in more than 15% of recall is attained, showing that porting information from a language with more resources can be helpful to overcome resource restrictions in some other language.

1 Introduction and Motivation

In the last decades a vast number of resources has been developed for some languages, specially for English, in the syntactic and semantic levels (PennTreebank, Wordnet, FrameNet, Propbank, VerbNet). This is not the case for most of the languages, including lesser-used languages like Basque, for which it is difficult to get raw or annotated corpora in significant amounts. This work explores the crosslingual portability of linguistic knowledge, more precisely of lexico-syntactic knowledge, as a way to reduce the lower performance attained by syntactic parsers in such languages, in this case Basque.

We chose to focus on the PP attachment problem to evaluate the portability hypothesis. The specific problem is that of deciding to which verb attach each phrase. We chose this kind of PP attachment problem because it is a specially hard problem for free word order languages like Basque; In principle PP attachment ambiguity in multiple-verb sentences is high as PPs can appear in any position. In contrast PP attachment ambiguities relating noun vs. verb anchors is more rare in Basque.

The experiment we present here was divided in two parts. The first part consisted of using Basque corpora to construct a monolingual classifier capable of attaching correctly Basque ambiguous PPs. The second part of the experiment corresponds to the construction of a crosslingual classifier. Both classifiers were evaluated over the same Basque PP attachment ambiguities.

The distribution of the paper is the following. Section 2 describes in more detail PP attachment ambiguities in Basque. Section 3 presents the monolingual approach for PP attachment. In section 4 we outline the crosslingual method. Section 5 describes the corpora used. The results are presented in Section 6. Finally, we draw the conclusions.

1 Authors listed in alphabetical order

2 PP Attachment ambiguities in Basque

As said before, Basque is a free constituent order language. PPs can attach in any order with respect to their corresponding verb. The only restriction being that crossing cannot occur between PPs corresponding to two different verbs. The following sentence illustrates the existing ambiguities to the sentence level:

[Lehendakaria][azkar]bildu zen [ministroekin]
[President] [without delay]met [ministers-with]
[berriak] kontatu zizkiotenean
[news] told aux-when.

The president met with the ministers right away when someone told him the news .

In Basque, the PPs[2] *with the ministers* and *news*, can be attached to either the first verb *meet* or to the second verb *tell*. Due to the subcategorization and selectional preferences of these verbs we know that the correct attachment is the one where *with the minister* attaches to the first verb (*to meet*) and *news* attaches to the second verb (*to tell*). The following example shows a parallel construction but with different attachments:

[Lehendakaria] [atzo]gaixotu zen [ministroekin]
[President] [yesterday] got-sick [ministers-with]
[kongresuan] bildu zenean.
[congress-in] met aux-when.

The president got sick yesterday when he met in the congress with the ministers.

Being the position of *ministroekin* (*with the ministers*) the same as in the previous example, the attachment varies, since *ministroekin* (with the minister) should attach to the second verb (to meet), again, because of subcategorization and semantic reasons.

It is well known that verbs show certain syntactic and semantic preferences on the prepositional phrases and noun phrases they appear with. Therefore, in a sentence with two verbs, and some prepositional or noun phrases, one of the verbs will show higher preference for some of the noun phrases than the other verb. We will make one assumption beyond this idea, the assumption being that these preferences happen and to some extent can be transferred crosslingually (Agirre et al., 2003, Agirre et al., 2004). As mentioned before, the preferences a verb shows with respect to a noun (inside a PP) are both syntactic and semantic. Syntactic in terms of subcategorization and semantic in terms of selectional preferences. So for instance *to meet* tends to subcategorize a PP where the preposition is *with*, and it tends to semantically select *persons* or *people to meet with* as opposed to, for example, *things*.

Note that the most studied PP attachment ambiguity, namely the one occurring between attaching to a verb or to a noun, stays out of the scope of this paper. Despite of being a frequent ambiguity in English (*I [bought books] for children, I bought [books for children]*) it does not occur as often in Basque. This is due to the fact that Basque sometimes shows the possibility of using a different case marking for attachments to the verb or to the noun. For instance in the example above the ambiguous English preposition *for* corresponding to the noun attachment site (*books meant for children)* and the beneficiary *for* corresponding to the verb attachment site (*buy books to give them to children)* has two non ambiguous Basque equivalents; the *–tzako* case marking equivalent to noun attachment site *for*, and the *–tzat* case marking equivalent to the verb attachment site *for*. Hence being frequent to use a different case marking depending on the attachment site (verb or noun) reduces this kind of PP attachment ambiguity in Basque as opposed to English.

3 The monolingual approach

We generated training data starting from raw Basque text, using a shallow parser (Aranzabe et al. 2004) and a simple extraction heuristic to select non ambiguous examples (those with a single verb).The shallow parser used is based on constraint grammar (Karlson et al. 1995). It chunks the text and gets syntactic structures to the level of phrases (PPs and VPs). The phrases obtained as output of the shallow parser were replaced with their head and the relevant case or

[2] Basque is a head final agglutinative language thus it does not have prepositions but grammatical cases/postpositions attached at the final word of each noun phrase. In the example above (*ministroekin*/ with the ministers) corresponds to *ministro+ekin*, where the associative case marking (*ekin*) would be equivalent to the English preposition *with*. Therefore it would be more appropriate to call them CPs(Case Phrases) as opposed to PPs, but for clarity reasons we will keep calling them PPs.

postposition (equivalent to an English preposition) that links them to the verb.

The extraction heuristic consisted of selecting monoverbal sentences, since they are unambiguous in terms of the PP attachment task we were concerned with.

We built a table with verbs and the head-case pairs obtained from the monoverbal sentences. The goal was to solve ambiguous attachments in the two-verb sentences using as training data the non ambiguous information contained in our table. The classifiers are based on Ratnaparki´s algorithm (Ratnaparki 1993), but we tried two variations: in the standard case (mono1 below) we only considered the case/postposition and the two main verbs, while in a second case (mono2) we also considered the head noun of the PP being attached. The attachment decisions are formalized as follows:

$$mono1(p,v1,v2) = arg\_max_{att \in \{v1,v2\}} Pr(att, p, v1, v2)$$

$$mono2(p,n,v1,v2) = arg\_max_{att \in \{v1,v2\}} Pr(att, p, n, v1, v2)$$

$$Pr(att, p, v1, v2) = Pr(v1)\,Pr(v2)\,Pr(att \mid p, v1, v2)$$

$$Pr(att, p, n, v1, v2) = Pr(v1)\,Pr(v2)\,Pr(att \mid p, n, v1, v2)$$

p corresponds to the case/postposition of the PP phrase, *n* to its head noun, *att* to the attachment decision, and *v1* and *v2* to the two verbs in the sentence.

To be able to estimate the probabilities as they stand we should have unambiguous training examples where *v1* and *v2* occur simultaneously. As our training examples come from monoverbal sentences, we can make the assumption (as in Ratnaparki's work) that the attachment of a preposition to a given verb relates to the strength the verb alone selects that preposition, and we end up with the following approximations:

For *mono1*:

$$Pr(att = v1 \mid p, v1, v2) \sim Pr(att = v1 \mid p, v1)$$
$$Pr(att = v2 \mid p, v1, v2) \sim Pr(att = v2 \mid p, v2)$$

$$Pr(att = vx \mid p, vx) = \begin{cases} \frac{\#(att = vx, p, vx)}{\#(p, vx)} & \text{if } \#(p, vx) > 0 \\ \frac{1}{|P|} & \text{otherwise} \end{cases} \quad vx \in \{v1, v2\}$$

where $\#(att=vx,p,vx)$ is the frequency of *p* appearing with *vx* in our unambiguous table, and $\#(p,vx)$ is the frequency of p appearing with *vx* in the whole corpus. |P| is the number of different cases/prepositions found in the whole corpus, and is used to smooth 0 probabilities.

Similarly for *mono2*:

$$Pr(att = v1 \mid p, n, v1, v2) \sim Pr(att = v1 \mid p, n, v1)$$
$$Pr(att = v2 \mid p, n, v1, v2) \sim Pr(att = v2 \mid p, n, v2)$$

$$Pr(att = vx \mid p, n, vx) = \begin{cases} \frac{\#(att = vx, p, n, vx)}{\#(p, n, vx)} & \text{if } \#(p, vx) > 0 \\ \frac{1}{|N|} & \text{otherwise} \end{cases} \quad vx \in \{v1, v2\}$$

where $\#(att=vx,p,n,vx)$ is the frequency of *p* appearing with *n*, and *vx* in our unambiguous table, and $\#(p,n,vx)$ is the frequency of *p* appearing with *n* and *vx* in the whole corpus. |*N*| is the number of different head nouns found in the whole corpus.

We used |N| for smoothing because $\#(att=vx,p,n,vx)/\#(p,n,vx)$ the ratio ends up being very small since the counts are few, and it is smaller than the one between $\#(att=vx,p,vx)/\#(p,vx)$, so we used |*N*| as a way to represent this fact in smoothing.

4 The crosslingual approach

For the crosslingual approach, we trained the classifier over unambiguous English examples acquired from automatically parsed English data. The method is based on (Agirre et al. 2004). The crosslingual classifier was then applied to disambiguate the same test examples as the monolingual classifiers.

The method used performs the following steps:

1.-obtain the head-case/preposition from the test Basque data.
2.-translate these Basque heads and cases to English
3.-build all possible English VP(head)-PP(head-case) translation combinations.
4.-collect English combinations frequencies in the English corpus and assign a weight to each frequency.

We will describe each of the steps in turn. The first step consists in obtaining the verbs and the nominal head (with its respective case/preposition) from each of the PPs in the Basque test sentences.

The second step consists in translating the heads, cases and the verbs into English using a bilingual dictionary. For each (verb-noun-case)

Basque triplet, build all possible translation combinations and then search them in the dependency database built from an automatically parsed English corpus (using the RASP parser (Carroll and Briscoe (2001)). Take for example this Basque sentence:

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

The verbs, heads and cases/prepositions are the following:

PP-ergative (*lehendakaria*)
PP-absolutive(*hauteskundeak*)
PP-absolutive(*boto*)
PP-distributive(*inbertsore*)
V1(*irabazi*) V2(*lortu*)

We translate the nouns, cases and verbs:

Basque	English equivalents
lehendakaria ergative	president, chairman, subj
hauteskundeak absolutive[3]	poll, election subj,obj
boto absolutive	vote,vow subj,obj
inbertsore distributive	investor, shareholder among, between
irabazi	to win /earn/gain
lortu	to get/obtain/attain

All possible English noun-verb pairs are created with the corresponding English relations or prepositions for each Basque case, for example, for the first Basque PP *lehendakari-ergative*:

***lehendakari-irabazi* vs. *lehendakari lortu*:**	
win-President-ncsubj	*get-President-ncsubj*
earn-President-ncsubj	*obtain-President-ncsubj*
gain-President-ncsubj	*attain-President-ncsubj*
win-Chairman-ncsubj	*get-Chairman-ncsubj*
earn-Chairman-ncsubj	*obtain-Chairman-ncsubj*
gain-Chairman-ncsubj	*attain-Chairman-ncsubj*

Note that we only search for the English verb and noun translations occurring in a direct syntactic dependency. Moreover, the English and Basque syntactic dependencies need to be compatible (see below).

The third step consists in training the classifier over the English data with the English equivalents:

$$cross(batt,bp,bn,bv1,bv2) = arg\_max_{eatt?\{ev1,ev2\}} assoc(eatt,ep,en,ev1,ev2)$$

where *b* stands for Basque and *e* for English.

In this case, we could probably find unambiguous training examples where *v1* and *v2* occur simultaneously, since the training corpus is deeply parsed. But to be consistent with the monolingual approach we maintained the same assumption as before, namely that the verb alone selects a given preposition no matter which are the verbs appearing around. So we end up with the same approximations:

$$assoc(att = v1,p,n,v1,v2) \sim assoc(att = v1,p,n,v1)$$
$$assoc(att = v2,p,n,v1,v2) \sim assoc(att = v2,p,n,v2)$$

We used mutual information as the association measure

$$MI(att=v1,n,p,v1) = log \frac{P(v1,n,p)}{P(v1)P(n,p)}$$

To estimate *P(v,n,p)*, *#(v,n,p)* was computed as:

$$? \#(v\ translations, n\ translations, English\ equivalent\ prep).$$

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. This way we would be able to maintain the syntactic subcategorization information, so for example in the sentence from the first section, "*The president met urgently with the ministers when someone told him the news*", we do not only search for *meet* and *minister*, we include *with* in the search, as being equivalent to the Basque associative case (*-kin*) attached to *ministro(minister)*. The equivalence between Basque morphological cases and English prepositions was taken from (Lersundi et al., 2002). In this equivalence table all possibilities were listed, but we discarded the ones marked as marginal.

[3] It is important to mention that there are cases associated to the subject and object functions, and sometimes they are phonologically null (as if they were just an NP).

5 The Corpora

The English corpus is the BNC (http://www.natcorp.ox.ac.uk). The English parser used is the RASP parser (Carroll and Briscoe 2001). The output parses follow the dependency formalism where syntactic relations at the sentence level are represented as links between the heads of the phrases and the verbs in contrast to what constituency parsers do, linking whole phrases to verbs. This is a relevant feature because it will facilitate the searches.

In order to make efficient searches in the English parsed corpora we created a database (Agirre et al. 2004). All in all, the database contains 47,145,584 syntactic relations from BNC. From these relations, 10,447,129 relations are verb-noun dependencies in BNC.

The Basque corpus comes from a newspaper and it refers to news from several months of the years 2000 and 2001 (~1.3M sentences) in different domains (culture, sports, finances, etc.). The sentences comprising the corpus were automatically chunked (see Section 3) and 400748 monoverbal sentences were extracted, and used for training. From this corpus a small part belongs to a treebank of 3092 sentences built by our group (Aduriz et al. 2003). From this treebank sentences with two verbs were selected (732 in total) and set aside for testing purposes.

6 The results

Table 1 shows the results for the different classifiers. Mono1 and mono2 are the monolingual classifiers (cf. Section 3). Mono1 represents the monolingual classifier where information provided by the noun head of each phrase was not used, but the information supplied by the grammatical case. Mono2 stands for the monolingual classifier for which this nominal information was used at training. Cross symbolizes the crosslingual classifier trained over English data and tested over Basque data (cf. Section 4).

	#rel	Prec	Cov	Rec
Mono1	2032	62%	70%	43.1%
Mono2	2032	69%	53%	36.6%
Cross	2032	58%	77%	44.6%
Mono1+2	2032	67%	70%	46.9%
Mono1+2+Cross	2032	66%	91%	60%

Table1 Results for the classifiers over Basque test data

The results show that Mono2 is the classifier that performs best in terms of precision but the worst in terms of coverage and recall. This was expected, as this classifier requires also information about the head noun, and thus suffers from heavier sparse data problems. Mono1 has lower precision but higher coverage and recall. The best recall is obtained by the crosslingual classifier, at the cost of lower precision. The better recall is explained by the fact that Cross has much more data available (in the English corpus) than in the monolingual methods (in the Basque corpus). The lower precision could be due to the loss of information in the translation process or because syntactic information in one language does not always have a direct equivalent in the other language.

Table 1 also shows the results of combining the systems. Mono1+2 corresponds to applying Mono1 when Mono2 does not offer any answer. This way the coverage and recall are improved, but at the cost of some precision. Mono1+2+Cross corresponds to applying Mono2, Mono1 and Cross in cascade, that is, using Mono1 when Mono2 cannot decide, and Cross when neither Mono1 or Mono2 can decide. This combination attains the best recall (by 14%) with a slight decrease in precision (a single point). These combined results show that the monolingual and crosslingual methods are complementary, and that crosslingual approaches can help overcome data restrictions in the target language.

7 Related work

In the literature there is a vast number of papers dealing with the disambiguation of the PP attachment decisions. These papers focus on disambiguating the traditional PP attachment ambiguity between attaching to a verb or to a noun. Some use supervised methods training over examples coming from a Treebank (Brill and Resnik 1994, Collins and Brooks 1995, Merlo et al. 1997, Stenina and Nagao 1997), while others use unsupervised methods (Hindle and Rooth 1993, Ratnaparkhi 1998) learning from (v,n,p) tuples obtained from large amounts of raw texts. For example, Hindle and Rooth used a 13M words corpus, Ratnaparkhi used 970K sentence corpus.

More recently Volk proposed a combined method where he uses 10k sentences coming from the NEGRA treebank, and a raw corpus of

around 5.5M words. He introduces already the problem of small set of data for certain languages like German.

The precision obtained by these works ranges from 81 to 88%.

The work presented here is different in two main respects with the ones dealing with the traditional PP attachment, first the attachment decisions to be made differs. Our task does not consist in deciding whether to assign to a verb or a noun but deciding between two verbs. Second, and very important the amount of data available for learning is much smaller than in any of them. For all these reasons it is very difficult to make a comparison since the differences are substantial.

8 Conclusions and further work

This work aimed at exploring the portability of linguistic knowledge from one language to another, comparing the results with a monolingual approach on the same task. The results reported suggest that the transfer is possible and moreover that crosslingual and monolingual approaches can be complementary.

In contrast with other work in the literature, we use totally unrelated corpora in the two languages, which makes the method easily portable to new languages, requiring only a bilingual dictionary in order to reuse the English database.

The classifiers built for this experiment only exploited lexico-syntactic information to make attachment decisions.

For the future we plan to explore the use of comparable corpora (as in Agirre et al. 2004) together with positional information for PP attachment in Basque, and study the possibilities for combination with the mono and crosslingual lexico-syntactic information used here.

We also plan to explore selectional preferences (both monolingual and crosslingual) as a way to incorporate more semantics into the system.

References

I. Aduriz, M.J. Aranzabe, J.M. Arriola, A. Atutxa, A. Díaz de Ilarraza, A. Garmendia, M. Oronoz (2003) *Construction of a Basque Dependency Treebank.* Proceedings of the Second Workshop on Treebanks and Linguistic Theories "TLT 2003", Växjö University Press. Växjö, Suecia

E. Agirre, I. Aldezabal and E. Pociello. (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.

E. Agirre, A. Atutxa, K. Gojenola and K. Sarasola. (2004). *Exploring portability of syntactic information from English.* In Proceedings of LREC. Lisbon, Portugal.

I. Aduriz, M. Aranzabe, J. Arriola, A. Atutxa, A. Díaz de Ilarraza, Garmendia, M. Oronoz (2003).*Construction of a Basque Dependency Treebank*. In Proceedings of TLT. Second Workshop on Treebanks and Linguistic Theories, Vaxjo, Sweden.

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

M. Aranzabe,J. Arriola, A. Díaz de Ilarraza, (2004) *Towards a Dependency Parser of Basque.* In Proceedings of the Coling 2004 Workshop on Recent Advances in Dependency Grammar. Geneva, Switzerland.

E. Brill and P. Resnik. (1994). A rule-based approach to prepositional phrase attachment disambiguation. In Proceedings of COLING Kyoto. ACL.

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

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

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

M. Lersundi and E. Agirre, (2003). *Semantic interpretations of postpositions and prepositions: 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.

B. Magnini, O. Popescu, J. Atserias, E. Agirre, E. Pociello, G. Rigau, J. Carroll and R. Koeling (2004). *Cross-Language Acquisition of Semantic Models for Verbal Predicates.* In Proceedings of LREC. Lisbon, Portugal.

P. Merlo, M.W. Crocker,C. Berthouzoz (1997). *Attaching Multiple Prepositional Phrases: Generalized Backed off Estimation.* In Claire Cardie and Ralph Weischedel, editors, Second

Conference on Empirical Methods in Natural Language Processing, Providence, R.I.

A. Ratnaparkhi (1998). *Statistical Models for Unsupervised Prepositional Phrase Attachment*. In Proceedings of the 36th Annual Meeting of the ACL (ACL'98) - joint with Coling'98. Montreal, Canada.

Stenina J., Nagao M.,(1997). *Corpus Based PP attachment ambiguity resolution with a semantic dictionary*.In J. Zhou and K. Church, editors, Proc. of the 5th Workshop on Very Large Corpora, Beijin and Hongkong.

M. Volk (2002). *Combining Unsupervised and Supervised Methods for PP Attachment Disambiguation*. In Proc. of COLING-2002. Taipeh. 2002.