A pilot study of English selectional preferences and their cross-lingual compatibility with Basque

Eneko Agirre, Izaskun Aldezabal, and Eli Pociello

IxA NLP group, University of the Basque Country,
649 pk., E-20080 Donostia, Basque Country
{eneko,jibalroi,elisabete}@si.ehu.es,
WWW home page:http://ixa.si.ehu.es

Abstract. The specific goals of this experiment are to study automatically acquired English selectional preferences from a number of sources, and to assess portability and compatibility issues with regard to selectional preferences acquired for Basque. We decided to study a wide-range of techniques and issues, with the aim of providing an analysis of the interplay of selectional-learning techniques, domain and multilinguality. The overall goal is the acquisition of complex lexical information for verbs (both syntactic and semantic) using multilingual sources.

1 Introduction

Selectional preferences try to capture the fact that linguistic elements prefer arguments of a certain semantic class, e.g. a verb like '*eat*' prefers as object edible things, and as subject animate entities, as in, (1) "*She was eating an apple*". Selectional preferences get more complex than it might seem: (2) "*The acid ate the metal*", (3) "*We ate our savings*", etc.

In their inception, selectional preferences were devised for verbs senses [8], but automatic efforts to acquire them have focused on verbs [6]. More recently, proposals for the acquisition of selectional preferences for specific senses of verbs have been published [1, 5]. Alternatively, [2] proposes to acquire selectional preferences from domain-related corpora, which might allow acquiring selectional preferences for the senses of the verb related to the domain.

The specific goals of this experiment are to study and compare automatically acquired English selectional preferences from a number of sources using different techniques, and to assess crosslingual relations with regard to selectional preferences acquired for Basque. We decided to study a wide-range of techniques and issues, with the aim of providing an analysis of the interplay of selectional-learning techniques, verb senses and multilinguality.

Cross-language comparison involves complex interactions among diverse linguistic phenomena such as word senses, domain of the word senses, syntactic relations and thematic roles. A useful resource is EuroWordNet [7], which can be used to link directly any English word sense with the equivalent word sense in Basque.

This article is structured as follows. In the next section, a short review of the different approaches to selectional preference learning is presented, alongside the corpora used on this experiment. Sections 3 and 4 show the results of the English and Basque acquisition. Section 5 presents the sources of errors. Finally, the conclusions are drawn.

2 Selectional Preference Learning

Selectional preferences have been acquired for all the synonyms of 8 English verb synsets,which are predominant in the sports domain.And for Basque, we studied exactly the same synsets (marked in bold in the list below):

<*play_1*, **jokatu_2**> <*encounter_5, meet_10, play_24, take_on_5*, **jokatu_3**>
<*draw_25, tie_2*, **berdindu_15**> <*equalize_1, get_even_1*, **berdindu_16**>
<*coach_2, train_7*, **entrenatu_1**> <*train_8*, **entrenatu_3**>
<*lose_2*, **galdu_9**> <*win_1*, **irabazi_3**>

The sources for the acquired selectional preference are the following: Semcor, a manually sense-tagged corpus for English, the British National Corpus (BNC), English news corpora from EFE (which is organized in different domains, e.g. finance or sports), and Basque news corpora from Egunkaria, also organized in different domains.

Selectional preferences have been acquired using different techniques, as presented below: word-2-class (w2c), class-to-class (c2c), sense-to-class (s2c) and word-to-semantic-file-domain (w2semf). Not all combinations of methods and source corpora were studied. Tables 1 and 2 show which selectional preference source and methods were tried.

2.1 Word-to-class model (w2c)

For each verb this method acquires a selectional preference given as a weighted list of classes of nouns, where the classes are taken from the hierarchy in WordNet [4]. The selectional preferences that we compute in this model are obtained from relations extracted from the target corpus. The first step is to apply the Minipar parser [3], and to extract [noun, relation, verb] triples for the relations "object" and "subject" from each occurrence of the target verb. The nouns in the triples are checked in WordNet in order to retrieve the corresponding word senses, which are returned as a list of synsets[1]. The triples are then converted into [noun-synset, relation, verb] triples, where one triple is obtained for each word sense of the original noun. In the case of Semcor, it is possible to retrieve [noun-synset, relation, verb] triples directly, as each occurrence of the nouns has been disambiguated by hand.

[1] We have to note that, the same as for the rest of the acquisition methods, no Name Entity Recognition system was used. Words not found in WordNet are all ignored. In the case of pronouns and prodrop arguments (in Basque) they are marked accordingly.

00228990	0.148	activity "any specific activity or pursuit;"	ACCEPTABLE
00004865	0.105	person, individual, someone, somebody, mortal, human, soul	
00009469	0.040	object, physical_object "a physical (tangible and visible) entity;"	
00017008	0.031	group, grouping, "any number of entities (members) ..."	
00018599	0.029	communication "something that is communicated between ..."	
00021098	0.028	action "something done (usually as opposed to something said);"	
00018966	0.008	measure, quantity, amount, quantum "how much there is of ..."	
00015437	0.007	state "the way something is with respect to its main attributes;"	
00017586	0.007	attribute "an abstraction belonging to or characteristic of ..."	
04771851	0.006	contest competition "an occasion on which a..."	CORRECT

Fig. 1. Ten synsets with highest probailities for the objects of play. Acquired using Semcor and the w2c method. Lines in bold correspond to correct or acceptable synsets for the first sense of play. The WN1.6 synset number, the estimate of the probaility for the synset and the words in the synset are provided.

In order to compute the probabilities of the triples, we use the probabilities of all the concepts above the target noun in the WordNet hierarchy. The formula to obtain these probabilities is based on estimated frequencies acquired from the target corpus. For each occurrence of a synset or a triple in the corpus, we distribute the frequency among its ancestors. The formulas, and a complete description of this work can be found in [1]. The approach used here is comparable to [6]. Figure 1 shows an example of this approach.

We want to note that in this model, the selectional preferences for each sense of the verb are mixed into a single selectional preference model for the whole verb.

2.2 Class-to-class model (c2c)

In contrast to word-to-class methods, this method tries to factor out the selectional preferences of each sense of the verb. For this, it mixes the selectional preferences of all synonyms and hyponyms of the verbal synset into a single selectional preference. This model was shown to yield better results than the word-to-class model on a WSD task [1]. The method works as in word-to-class methods, but in order to compute the probabilities of the triples, we use the probabilities of all the concepts above the target noun and verb synsets in the WordNet hierarchy. The outcome is a set of triples with the form [noun-synset, relation, verb-synset] together with their probabilities. The formulas, and a complete description of this work can be found in [1].

2.3 Sense-to-class model (s2c)

In this model selectional preferences are acquired for each sense of a verb, in a similar way to the class-to-class model. This work is described in [5].

2.4 Class-to-semantic-file-domain model (w2semf)

The above methods use the hierarchy in WordNet to defined classes of nouns and verbs. A different strategy would be to define classes of words based on

play-act	50.013	CORRECT
factotum-act	30.390	ACCEPTABLE
time_period-time	29.009	CORRECT
zoology-animal	25.2	
factotum-artifact	25.026	
sport-event	23.514	CORRECT
sport-act	23.038	CORRECT
number-quantity	22.957	
geography-location	16.918	

Fig. 2. Ten generalizations with highest weights for the objects of play. Acquired using EFE and the w2semf method. Lines in bold correspond to correct or acceptable generalizations for the first sense of play. The semantic file and domain pair and the weight are provided.

00238878 diversion, recreation	Play-act Sport-act
04771851 contest, competition	Sport-event
09065837 amount of time, period...	Sport-event

Fig. 3. Gold standard for the objects of the first sense of play. The gols standard is given for the two possible generalizations: in the left we give the WN1.6 synset number, alongside the words in the synset, and in the right we have the respective WordNet Domains and Semantic File pair generalizations.

domains. In this approach, each verb has a selectional preference for each argument, given as a weighted list of classes of nouns based on domain-semantic file pairs [2]. These pairs are formed combining WordNet Domains and the classification of nouns in Semantic Files from WordNet.Figure 2 shows an example of this approach.

3 Results for the English Selectional Preferences

In order to judge the quality of the English selectional preferences obtained, a linguist produced a gold standard for each of the target verb senses which is class based on WordNet hyperonyms.A sample of the gold standard for the first sense of *play* is shown in Fig. 3.

The study aims to evaluate whether the acquisition methods correctly capture the selectional preference for each target verb sense. In order to facilitate the linguistic analysis we focus on the 10 top classes in each case, i.e. the 10 classes with highest weights. Depending on this gold standard, the evaluation is given in the form of **correctly** acquired selectional preferences, **acceptable** selectional preferences (e.g. too general or specific in the hierarchy) and selectional preferences that are **missing** (i.e. not found by the acquisition method). For instance, Fig. 1 and Fig. 2 show the evaluation for two methods.

Table 3 shows the evaluation of the acquired selectional preferences for each corpus, algorithm and syntactic relation. The most important figure is that of missing, where 0 means that all relevant selectional preferences were acquired.

		Object			Subject		
Origin	Method	Correct	Acceptable	Missing	Correct	Acceptable	Missing
Semcor	w2c	0.122	0.264	0.305	0.324	0.062	0.125
Semcor	c2c	0.069	0.244	0.44	0.38	0.02	**0.071**
Semcor	s2semf	0.16	0.47	0.607	0.157	0.006	0.6
BNC	w2c	0.08	0.15	0.135	0.112	0.06	0.15
BNC	c2c	0.015	0	0.96	0	0	1
BNC	s2c	0.083	0.111	0.375	0.107	0	0.5
EFE (sports)	w2semf	0.148	0.155	**0.009**	0.027	0.384	**0.045**

Table 1. Results for selectioal preferences acquired from different origins using several methods.

The best results for objects are for w2semf from EFE. The best results for subjects are also for w2semf from EFE, and c2c from Semcor. Note that c2c from BNC misses nearly all relevant selectional preferences.

4 Analysis of the errors

Acquiring selectional preferences from running text, without sense tags, involves dealing with a great amount of noise. Apart from this we found the following sources of errors:

Using hyponyms in class-to-class selectional preferences: in c2c selectional preferences mix the linguistic information of all synonyms and hyponyms of the verbal synset into a single selectional preference, therefore, most of the errors in this kind of selectional preferences are due to this inclusion.

Tagging errors: this applies to Semcor, and refers to the words being tagged with the wrong sense. For instance, *person* and *group* object selectional preferences do not belong to *play_1* but to *play_24*.

Senses missing in WordNet: Selectional preferences are based on WordNet1.5, and as consequence, it depends on its hierarchy. It could be the case that some synsets are not represented on the ontology and therefore, they affect our final result. For instance, *Argentina* appears as a possible object of *play* which is taken to be location instead of sports teams. This is because WordNet1.5 has no synset for this sense of *Argentina*¸ and for that reason it has been tagged in Semcor, with the most similar sense found in WordNet1.5, that is, *location*.

Sense ambiguity: This is the main source of noise on selection preference learning. Even if the acquisition methods claim that the signal spread across the different words will allow to filter out the noise it is not always the case. For instance, the animal interpretation of *game* (which is the most frequent object of *play* in the sports domain) makes *animal* a highly placed selectional preference for *play*.

5 Comparison with Basque Selectional Preferences

One of the objectives of this experiment is to compare the selectional preferences of a verb or verb class in English with the selectional preferences of the

```
ABSOLUTIVE 04771851 contest, competition
ABSOLUTIVE 09065837 amount of time, period, period of time, time period
INESSIVE   00238878 diversion, recreation
ADLATIVE   00238878 diversion, recreation
```

Fig. 4. Gold standard for the selectional preferences acquired from Basque corpora for jokatu in the sports domain. Selectional preferences have been coded for the main grammatical cases, but only the ones related to the English object are shown here.

translations into Basque. In principle, c2c and s2c selectional preferences for English can be directly ported into Basque, as they are linked to a certain sense of the verb that yields a translation into Basque. The same can be tried for w2semf selectional preferences acquired from the sports domain, assuming that the domain of the corpus narrows the senses of the target verb and its possible translations.Alternatively, we can directly apply the selectional learning algorithms to the Basque data from Egunkaria. In this section we provide an analysis of the c2c and w2semf from Semcor, w2semf from English EFE, and w2semf from Egunkaria.

Arguments for English verbs do not directly translate into Basque. Roughly speaking we can say that subjects in English can be reflected by the ergative case in Basque, and that objects in English can be reflected by the absolutive. Unfortunately this is not always the case. For instance, Basque verb *jokatu* (*play*) does not take activities such as *football* or *golf* like objects. In fact, it does allow them as arguments, but in the inessive or adlative form (*-n* and *-ra* respectively, literally *play in football* and *play to football*) rather than in the absolutive one:

Futbolean <inessive> **jokatzen** badakitela erakutsi zuten.
They showed they know how to ***play football.***

The gold standard (cf. Fig 5) for the second sense of *jokatu* (equivalent to *play_1* in EuroWN) shows that part of the nouns that go with objects in English appear with the absolutive, but also with the inessive or adlative cases in Basque.

Table 5 summarizes the quality of the acquired selectional preferences for the verb *jokatu*. Separate figures are given for each case suffixes. In the case of English selectional preferences, the subject and object are evaluated supposing that a correct mapping to the corresponding case-suffix is possible.

Selectional restrictions extracted from Egunkaria are of lower quality than those from EFE. Several factors could be responsible: the amount of data is smaller, the parsing is of lower quality (e.g. the absolutive case is wrongly marked for a number of proper nouns), and *jokatu* in Basque might be more polysemous, even in the sports domain.

All in all, selectional preferences acquired for *play_1* are perfectly portable to *jokatu_2*. The only problem is that of mapping subject and object functions in English with case suffixes in Basque.

Origin	Sel.Prefs	Case	Correct	Acceptable	Missing
Egunkaria sports	w2semf	abs	1 out of 10	0	1 out of 2
		ine	2 out of 10	1 out of 10	0
		ala	0	2 out of 10	0
		erg	0	7 out of 10	1 out of 2
Semcor	c2c	obj	1 out of 8	1 out of 8	1 out of 3
		subj	2 out of 5	0	0
Semcor	s2semf	obj	2 out of 10	2 out of 10	2 out of 3
		subj	2 out of 7	1 out of 70	0
EFE sports	w2semf	obj	4 out of 10	1 out of 10	0
		subj	2 out of 10	1 out of 10	0

Table 2. Results for selectional preferences directly acquired for Basque (Egunkaria) and ported for English (Semcor, EFE).

6 Conclusions and further work

The pilot study described in this paper has two goals: To compare selectional preferences acquired using different techniques (word to class, class to class, sense to class, word to semantic file-domain) from different corpora (SemCor, BNC, EFE) for English, and to study the phenomena involved when selectional preferences from different languages are compared.

From the study of the English selectional preferences we concluded the following:

- Each corpus has its own idiosyncrasies, which can affect the results. Being **Semcor** a hand annotated corpus, the acquired selectional preferences are of better quality than those from the **BNC** (good for w2c, very bad for c2c). Results for Semcor are lower than expected, especially due to hand tagging errors (*play_1* where it should have been *play_24*) or missing senses (*Argentina*). Limiting to a domain like EFE sports seems to be highly satisfactory in the case of *play*: it focuses on just two senses.
- The **class-to-class** and **sense-to-class** methods do not seem to do better than **word-to-class** methods. In the former, the data from hypernyms and **hyponyms** does not seem to help acquire selectional preferences for the verb class. On the latter, even if the quality seems to be better, the acquired selectional preference applies to **all senses** of the verb, and is of limited use. Restricting the domain coupled with word-to-class techniques seems to be a promising option. We expected c2c and s2c methods to be a good platform for crosslinguistic porting and cross-fertilizing the languages, but the poor results are worrying.
- The output of **word-to-semantic-file** selectional preferences is more difficult to interpret than those of hierarchical classes. Still they provide simple means to get the selectional preferences, and applied on EFE they provide the best quality selectional preferences.
- Focusing on texts from one **domain** provides the best quality selectional preferences, and it might allow narrowing the acquired selectional preferences for a target verb sense. Being this true for *play* it might well be that generalization to other verbs is not possible.

Regarding the comparison across Basque and English selectional preferences, the results are still preliminary, as we need to analyse more Basque verbs. Selectional preferences for *play* can be all **translated** to Basque, but one of the references changes the argument from object position into the inesive instead of the absolutive (from *play football* into *play in football*).

In summary, it seems that domains and cross-linguistic overlap might allow getting better quality selectional preferences for verb senses or verb classes. We are currently studying whether it is possible to devise an algorithm to port or cross-fertilize selectional preferences coming from different languages, based on the preliminary results from the relations between Basque and English selectioal preferences.

Acknowledgements

We want to thank the reviewers for their insightful comments. This work is partially funded by the European Comission (MEANING IST-2001-34460) and MCYT (HERMES TIC-2000-0335). Elisabete Pociello has a PhD grant from the Governmet of the Basque Country.

References

1. Agirre, E., Martínez, D.: Integrating Selectional Preferences in WordNet. Proceedings of First International WordNet Conference. Mysore (India). 2002.
2. Agirre E., Atserias J., McCarthy, D., Real, F., Rigau, G., & Rodriguez, H.: MEANING: Developing Multilingual Web-scale Language Technologies. Working paper 5.2a.
3. Lin, D.: Principle Based parsing without Overgeneration. In 31st Annual Meeting of the Association for Computational Linguistics. Columbus, Ohio. pp 112-120. (1993)
4. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Five Papers on WordNet. Special Issue of International Journal of Lexicography, 3 (4). (1990)
5. McCarthy, D.: Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. Ph.D. University of Sussex (2001)
6. Resnik, P.: Selection and Information: A Class-Based Approach to Lexical Relationships. Ph.D. University of Pennsylvania (1993)
7. Vossen, P., L. Bloksma, S. Climent, M.A. Marti, M. Taule, J. Gonzalo, I. Chugur, M. F. Verdejo, G. Escudero, G. Rigau, H. Rodríguez, A. Alongué, F. Bertagna, R. Marinelli, A. Roventini, L. Tarasi, W. Peters: Final Wordnets for Dutch, Spanish, Italian and English, EuroWordNet (LE2-4003) Deliverable D032/D033, University of Amsterdam. (2001)
8. Wilks, Y.: Y. Preference Semantics. In E. Keenan, (ed.) The Formal Semantics of Natural Language. Cambridge: Cambridge U. P. (1973)