Cross-Language Acquisition of Semantic Models for Verbal Predicates

Jordi Atserias*, **Bernardo Magnini**, **Octavian Popescu**, **Eneko Agirre**†,
Aitziber Atutxa†, **German Rigau**†, **John Carroll****, **Rob Koeling****

ITC-IRST Italy {magnini, popescu}@itc.it
* TALP Research Center Universitat Politécnica de Catalunya. Catalonia
{batalla, luisv}@talp.upc.es
†IXA Group, University of the Basque Country
{eneko,rigau,jibatsaa}@si.ehu.es
University of Sussex, Cognitive and Computing Sciences. UK
**{J.A.Carroll, robk}@sussex.ac.uk

Abstract

This paper presents a semantic-driven methodology for the automatic acquisition of verbal models. Our approach relies strongly on the semantic generalizations allowed by already existing resources (e.g. Domain labels, Named Entity categories, concepts in the SUMO ontology, etc). Several experiments have been carried out using comparable corpora in four languages (Italian, Spanish, Basque and English) and two domains (FINANCE and SPORT) showing that the semantic patterns acquired can be general enough to be ported from one language to the other language.

1. Introduction

Being a multidimensional problem, predicate knowledge is one of the most complex types of information to acquire. Predicates (verbs and their corresponding nominalizations) are essential for the development of robust and accurate parsing technology capable of recovering predicate-argument relations and logical forms. Without it, resolving most structural ambiguities of sentences is difficult, and understanding language impossible.

Full account of predicate information requires specifying the number and type of arguments, predicate sense under consideration, semantic representation of the particular predicate-argument structure, mapping between the syntactic and semantic levels of representation, semantic selectional restrictions/preferences on participants, control of the omitted participants and possible diathesis alternations. Unfortunately, all these kinds of knowledge are interdependent.

However, (Korhonen, 2002) showed that in terms of SCF distributions, individual verbs correlate more closely with syntactically similar verbs and clearly more closely with semantically similar verbs, than with all verbs in general. Moreover, her results show that verb semantic generalisations can successfully be used to guide and structure the acquisition of SCFs from corpus data.

Thus, it is possible to devise alternative acquisition schemes going top-down from semantics to syntax. If we identify specific associations between participants and predicates (selectional preferences), we can also gather information from corpus data about their particular syntactic behaviour in relation to a predicate, helping the acquisition of SCFs, diathesis alternations, etc. However, this new approach requires to work directly at a sense level, having predicates and associations to participants semantically disambiguated.

Furthermore, in a multilingual semantic scenario, it seems possible to devise ways to acquire from a particular language and using a bottom-up approach some predicate-argument knowledge, and then, following a top-down fashion, to acquire or validate some knowledge in another language.

Two different and complementary dimensions can help to minimise the WSD problem: multilingualism and domains. Although, working in parallel with comparable corpora in several languages will increase the complexity of the process, we believe that language translation discrepancies among word forms can help the selection of the correct word senses (Habash and Dorr, 2002). Moreover, further reduction of the search space among sense candidates can be obtained by processing domain corpora (Gale et al., 1992).

A set of empirical tests have been designed to evaluate the feasibility of the semantic-driven approach. These experiments have been carried out in the framework of the MEANING project (Developing Multilingual Web-scale Language Technologies[1]). Inside MEANING several wordnets from different European Languages have been aligned and integrated into a common semantic knowledge base: the Multilingual Central Repository (MCR (Atserias et al., 2004)). In MEANING the MCR acts as a multilingual interface for integrating and distributing all the knowledge acquired in the project. The resulting MCR has been also enriched with new semantic information coming from different sources and methods including an updated version of the EuroWordNet Top Concept Ontology (Vossen, 1998), the Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001), a general ontology, WordNet Domains (Magnini and Cavagli, 2000), etc.

This paper presents the first steps towards testing the validity of this new approach for the acquisition of predicate knowledge (SCFs, Selectional Restrictions, diathesis alternations, etc). The work here presented explores some basic issues in the acquisition of semantic models. First, how the current technology and the knowledge available can help large-scale acquisition tasks, mainly subcategorization

[1] `http://www.lsi.upc.es/~nlp/meaning`

frames (SCFs) and selectional restrictions or preferences (SPs) for Spanish, Italian, Basque and English. Second, the impact in the acquisition process when using several languages at the same time and third, when using domain corpus instead of a general corpus.

After this introduction, section 2. presents the resources used in this exploration. Section 3. describes the methodology used to acquire large–scale monolingual Semantic Models for predicates. Section 4. provides some qualitative views with about the domain and multilingual exploration and finally, in Section 5. we conclude with some prospects for future work.

2. Experimental Setting

Summarising, this paper presents new ways for restricting the search space when performing acquisition tasks, in order to obtain more accurate knowledge for some languages and balancing the coverage of such knowledge across languages.

Thus, this experiment can be also seen as a common framework to study productive paths to exploit appropriately:

- available semantic knowledge (wordnets, Semantic Files, MultiWordNet Domains, EuroWordNet Top Ontology, SUMO, etc. already present into the MCR (Atserias et al., 2004))
- cross language discrepancies/agreements through the EuroWordNet Interlingual Index
- available comparable domain corpora in several languages
- large-scale selectional preferences already acquired from this multilingual corpora (Atserias et al., 2003; Agirre et al., 2003)

3. The multilingual Adquisition

We carried out the experiment for particular verbal synsets which have common senses in the considered languages. For instance, the following verbal synsets belongs to the same ILI–record: English verbal synset 01564908–v <gain,clear,make,earn,realize>, Italian <reallizare,guadagnare>, Spanish <ganar> and Basque <irabazi>.

First, we collect sentences containing those verbs in comparable corpora for both domains FINANCE and SPORT. For each sentence, depending of the current capabilities of the Linguistic Processors used, we obtained the heads of the verb–slots acting possibly as subjects and objects. Only the English linguistic processor RASP (Bricoe and Carroll, 2002) performs high accurate dependency analysis.

For the rest of languages, in the pre-processing phase the sentences are PoS tagged and parsed into non-recursive phrasal units. The quality of parsing, especially with respect of NP chunks, is a crucial factor in the success of analysis. For Basque, we used a chuncker based on an unification grammar. For Italian and Spanish, in order to extract subject/object groups, three simple heuristics are applied: first, consider NP groups directly at the left hand side and at the right hand of the VP, second, identify passives and the postponed subject, and, finally, the VP NP NP case. As Italian and Spanish are subject-drop languages, we also use simple heuristics, based on barriers phrases, to detect the subject/object-drop cases.

	lemma	Sport	Finance
Spanish	*empatar*	1580	2
Italian	*pareggiare*	4551	80
Basque	*berlindu*	6	-
English	*draw*	120	60
	tie	500	48

Table 1: Verb occurrences for synset 00756166-v in both Sport and Finance corpora

Finally, once the subject/object pairs are extracted we associated a Named Entity category (or Semantic File from WN) and a Domain label to each head of the nominal groups. We also implemented a very simple generalization procedure associating to each verb one or more semantic patterns of type Name_Entity+WN_Domain on the base of their frequency.

In order to work with compatible representations across languages, we obtained for each verb–slot filler all their synsets. We also mapped the Named Entities types (PERSON, ORGANIZATION, AMOUNT, PERCENTAGE, DATE, etc.) to a common semantic representation. In the next tables, NONE stands for words that doesn't appear in the local wordnets and NO_SUBJECT/NO_OBJECT represents sentences where the subject/object was not detected.

4. Preliminary Results

Being this a preliminary and exploratory study (with many, hard and biased simplifications) we have performed only a preliminary qualitative evaluation. We have compared several semantic patterns coming from translation equivalent verbs selected from different languages and domains. The analysis of these results provide an initial characterization of the different cross–lingual behaviours.

4.1. Monosemous verbs

First, we analyze a very simple case. The word *empatar* is monosemous in Spanish while its English translations *tie* and *draw* are highly ambiguous (9 and 33 senses respectively). Table 1 shows some frecuency figures for the verbal synset 00756166-v. This sense belongs to the SPORT WN domain, but we obtain verb sentences in both domain corpora. Obviously, the number of verbal occurrences is different due to the different origin, domain, language and nature of the corpus processed.

Now, we can merge all these pairs in comparable representacions. Table 2 shows the first subject+object pairs of an ordered list resulting from merging Spanish and English, when performing some simple sense frequency counting.

To show the potentiality of this approach, we can also perform some basic generalizations, choosing the combination of Wordnet Semantic Field and MultiWordNet Domains as the semantic representation for each synset. Table 3 shows the initial part of an ordered list of subject+object generalized pairs resulting from Spanish and English, when

Subject	Object
NONE	NOOBJECT
PERSON	NOOBJECT
NONE	NONE
ORGANIZATION	NONE
PERSON	NONE
ORGANIZATION	NOOBJECT
ORGANIZATION	ORGANIZATION
ORGANIZATION	PERSON
<team_1>	NOOBJECT
LOCATION	NOOBJECT
<club_2>	NOOBJECT
PERSON	<match_2>

Table 2: Merging the Spanish/English Subject+Object pairs of 00756166-v from Sport corpora

Subject	Object
person-factotum	act-sport
person-factotum	quantity-sport
group-factotum	event-sport
person-factotum	event-sport
group-sport	NOOBJECT
person-factotum	act-sport
person-factotum	quantity-sport
group-factotum	event-sport
person-factotum	event-sport

Table 3: Merging the Spanish/English Subject+Object generalized pairs of 00756166-v from Sport corpora

performing some simple sense frequency counting and filtering out pairs not belonging to the SPORT WN domain.

Although some of the verbs are monosemous in the sports domain, an in-depth analysis of the data obtained, provides two different but related (by a causality relation) semantic patterns for <tie_2> that applies in all the languages:

<team_1> <tie_2> <score_3>
<team_1> <tie_2> <match_2>

Recall that, initially, all this data has been obtained without any kind of WSD preprocess. The existing combinations of cross–lingual correspondences in a restricted domain helps to corpus produce the final semantic patterns.

4.2. Multilingualy restricted polysemous verbs

Table 4 shows some volume figures for the verbs connected to the ILI–record 00756166-v. This ILI–record belongs also to SPORT WN domain. Although in this case, none of the verbs is monosemous, they are mutually restrictive.

Table 5 shows an ordered list of generalized patterns when processing this data and combining Spanish, English and Italian languages.

	lemma	Sport	Finance
Spanish	*entrenar*	2880	18
Italian	*allenare*	7243	16
Basque	*entrenatu*	23	-
English	*train*	114	12
	coach	346	-

Table 4: Verb occurrences for synset 00565367-v in both Sport and Finance corpora

Subject	Object
group-factotum	person-factotum
person-factotum	person-factotum
group-factotum	time-time_period
person-factotum	group-factotum

Table 5: Merging the Spanish/English/Italian Subject+Object generalized pairs of 00565367-v from Sport corpora

	lemma	Sport	Finance
Spanish	*ganar*	24268	1618
Italian	*realizare*	5421	5615
	guadagnare	3701	1618
Basque	*irabazi*	132	14
English	*earn*	117	143
	realize	17	8
	clear	22	23
	gain	35	794
	make	789	1695

Table 6: Verb occurrences for synset 01564908-v in both Sport and Finance corpora

In general, however, we will obtain all kinds of polysemous combinations of verbal senses.

Table 6 shows some frecuency figures of the verbal synset 01564908-v in both Sport and Finance corpora. However, this ILI–record is labeled with the WN domain ECONOMY.

Table 8 presents an ordered list of the first generalized semantic patterns of 01564908-v acquired from the Sport corpus when filtering out all non ECONOMY related domains.

Table 9 shows some figures of the verbal synset 01564238-v in both Sport and Finance domains. This ILI–record has no specific domain assigned (FACTOTUM).

Table 10 presents the first generalized semantic patterns of 01564908-v from the Sport corpus. This example shows that without filtering erroneous semantic patterns are also obtained.

4.3. Comparing Domains

Table 7 presents an ordered list of the first generalized semantic patterns of 01564908-v acquired from the Sport

Subject	Object
act-economy	NOOBJECT-NOOBJECT
NONE-NONE	possession-money
possession-money	NONE-NONE
possession-economy	NONE-NONE
group-factotum	possession-money
possession-money	state-factotum
cognition-factotum	possession-economy
possession-money	cognition-factotum
cognition-factotum	possession-money
quantity-money	cognition-factotum
possession-money	location-military
group-factotum	possession-economy

Table 7: Merging and filtering the Spanish/English/Italian Subject+Object generalized pairs of 01564908-v from Finance corpus

Sport corpus	
Subject	**Object**
possession-economy	NOOBJECT-NOOBJECT
NONE-NONE	possession-money
person-factotum	possession-money
possession-money	NOOBJECT-NOOBJECT
person-factotum	possession-economy
possession-money	NONE-NONE
possession-economy	act-factotum
group-factotum	possession-economy

Table 8: Merging and filtering the Spanish/English/Italian Subject+Object generalized pairs of 01564908-v

	lemma	Sport	Finance
Spanish	*ganar*	24268	1618
Italian	*ottenere*	10762	4929
	guadagnare	3701	3110
	raccogliere	3774	1880
	riportare	4074	2019
	conquistare	10233	1173
	conseguire	461	1057
	vincere	54927	1836
Basque	*irabazi*	132	14
English	*gain*	35	794
	win	913	51

Table 9: Verb occurrences for synset 01564238-v in both Sport and Finance corpora

corpus when filtering out all non ECONOMY related domains. In this case, we are obtaining similar results to those obtained from the Sport corpus (see table 8).

5. Conclusions

Automatic acquisition of semantic patterns for predicate structures (verbs and their corresponding nominalizations) is one of the most complex task for lexical acquisition. Verbs show multidimensional and interdependent features (selectional preferences, diathesis alternations, subcategorization frames) and their behavior may vary not only across languages, but also across corpus domains and genre. These facts are problematic for any syntax-driven approach (Atserias et al., 2001).

We proposed a cross-language methodology of acquiring semantic patterns for predicates. The pilot study we have conducted shows that it is possible to obtain promising results using this framework, if we consier the high level of polysemy degree we are dealing with. We used very simple criteria together with large collections of comparable corpora and already existing semantic resources to acquire large amounts of semantic patterns that can be very useful for a number of applications based on shallow semantics. Obviously, the whole process can be widely improved in several steps, in particular the semantic generalization process.

Subject	**Object**
person-factotum	act-factotum
person-factotum	event-sport
group-factotum	event-sport
group-factotum	act-factotum
person-factotum	group-factotum
person-factotum	time-time_period

Table 10: Merging the Spanish/English/Italian Subject+Object generalized pairs of 01564238-v from Sport corpus

Finally, we also plan to evaluate the application of the acquired cross-lingual models in particular NLP tasks, such as PP–attachment (Agirre et al., 2004), detection of subject/objects or WSD.

6. References

Agirre, E., I. Aldezabal, and E. Pociello, 2003. A pilot study of english selectional preferences and their cross-lingual compatibility with basque. In *Procceeding of TSD*. Czech Republic.

Agirre, E., A. Atutxa, K. Gojenola, and K. Sarasola, 2004. Exploring portability of syntactic information from english to basque. In *Procceeding of LREC*. Lisbon, Portugal.

Atserias, J., L. Padró, and G. Rigau, 2001. Integrating multiple knowledge sources for robust semantic parsing. In *Proceedings of the International Conference, Recent Advances on Natural Language Processing RANLP'01*. Tzigov Chark, Bulgaria.

Atserias, Jordi, Mauro Castillo, Francis Real, Horacio Rodríguez, and German Rigau, 2003. Exploring large-scale acquisition of multilingual semantic models for predicates. In *SEPLN'03*. Alcala de Henares, Spain. ISSN 1136-5948.

Atserias, Jordi, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen, 2004. The meaning multilingual central repository. In *Second International WordNet Conference-GWC 2004*. Brno, Czech Republic. ISBN 80-210-3302-9.

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

Gale, W., K. Church, and D. Yarowsky, 1992. One sense per discourse. In *Proceedings of of DARPA speech and Natural Language Workshop*. Harriman, NY.

Habash, N. and B. Dorr, 2002. Handling translation divergences: Combining statistical and symbolic techniques in generation-heavy machine translation. In *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002*. Tiburon, CA.

Korhonen, A., 2002. *Subcategorization acquisition*. Ph.D. thesis, University of Cambridge.

Magnini, B. and G. Cavagli, 2000. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*. Athens. Greece.

Niles, I. and A. Pease, 2001. Towards a standard upper ontology. In *In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*. Chris Welty and Barry Smith, eds.

Vossen, P. (ed.), 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers .