Improving the Interpretation of Noun Phrases with Cross-linguistic Information

Roxana Girju
University of Illinois at Urbana-Champaign
girju@uiuc.edu

Abstract

This paper addresses the automatic classification of semantic relations in noun phrases based on cross-linguistic evidence from a set of five Romance languages. A set of novel semantic and contextual English–Romance NP features is derived based on empirical observations on the distribution of the syntax and meaning of noun phrases on two corpora of different genre (Europarl and CLUVI). The features were employed in a Support Vector Machines algorithm which achieved an accuracy of 77.9% (Europarl) and 74.31% (CLUVI), an improvement compared with two state-of-the-art models reported in the literature.

1 Introduction

Semantic knowledge is very important for any application that requires a deep understanding of natural language. The automatic acquisition of semantic information in text has become increasingly important in ontology development, information extraction, question answering, and other advanced natural language processing applications.

In this paper we present a model for the automatic semantic interpretation of noun phrases (NPs), which is the task of determining the semantic relation among the noun constituents. For example, *family estate* encodes a POSSESSION relation, while *dress of silk* refers to PART-WHOLE. The problem, while simple to state is hard to solve. The reason is that the meaning of these constructions is most of the time ambiguous or implicit. Interpreting NPs correctly requires various types of information from world knowledge to complex context features. Moreover, the extension of this task to other natural languages brings forward new issues and problems. For instance, *beer glass* translates into *tarro de cerveza* in Spanish, *bicchiere da birra* in Italian, *verre à bière* in French, and *pahar de bere* in Romanian. Thus, an important research question is how do the syntactic constructions in the target language contribute to the preservation of meaning in context.

In this paper we investigate noun phrases based on cross-linguistic evidence and present a domain independent model for their semantic interpretation. We aim at uncovering the general aspects that govern the semantics of NPs in English based on a set of five Romance languages: Spanish, Italian, French, Portuguese, and Romanian. The focus on Romance languages is well motivated. It is mostly true that English noun phrases translate into constructions of the form *N P N* in Romance languages where, as we will show below, the *P* (preposition) varies in ways that correlate with the semantics. Thus Romance languages will give us another source of evidence for disambiguating the semantic relations in English NPs. We also present empirical observations on the distribution of the syntax and meaning of noun phrases on two different corpora based on two state-of-the-art classification tag sets: Lauer's set of 8 prepositions (Lauer, 1995) and our list of 22 semantic relations. We show that various crosslingual cues can help in the NP interpretation task when employed in an SVM model. The results are compared against two state of the art approaches: a su-

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 568–575,
Prague, Czech Republic, June 2007. ©2007 Association for Computational Linguistics

pervised machine learning model, Semantic Scattering (Moldovan and Badulescu, 2005), and a web-based probabilistic model (Lapata and Keller, 2004).

The paper is organized as follows. In Section 2 we present a summary of the previous work. Section 3 lists the syntactic and semantic interpretation categories used along with observations regarding their distribution on the two different cross-lingual corpora. Sections 4 and 5 present a learning model and results for the interpretation of English noun phrases. Finally, in Section 6 we offer some discussion and conclusions.

2 Related Work

Currently, the best-performing NP interpretation methods in computational linguistics focus mostly on two consecutive noun instances (noun compounds) and rely either on rather ad-hoc, domain-specific semantic taxonomies, or on statistical models on large collections of unlabeled data. Recent results have shown that symbolic noun compound interpretation systems using machine learning techniques coupled with a large lexical hierarchy perform with very good accuracy, but they are most of the time tailored to a specific domain (Rosario and Hearst, 2001). On the other hand, the majority of corpus statistics approaches to noun compound interpretation collect statistics on the occurrence frequency of the noun constituents and use them in a probabilistic model (Lauer, 1995). More recently, (Lapata and Keller, 2004) showed that simple unsupervised models perform significantly better when the frequencies are obtained from the web, rather than from a large standard corpus. Other researchers (Pantel and Pennacchiotti, 2006), (Snow et al., 2006) use clustering techniques coupled with syntactic dependency features to identify IS-A relations in large text collections. (Kim and Baldwin, 2006) and (Turney, 2006) focus on the lexical similarity of unseen noun compounds with those found in training.

However, although the web-based solution might overcome the data sparseness problem, the current probabilistic models are limited by the lack of deep linguistic information. In this paper we investigate the role of cross-linguistic information in the task of English NP semantic interpretation and show the importance of a set of novel linguistic features.

3 Corpus Analysis

For a better understanding of the meaning of the N N and N P N instances, we analyzed the semantic behavior of these constructions on a large cross-linguistic corpora of examples. We are interested in what syntactic constructions are used to translate the English instances to the target Romance languages and vice-versa, what semantic relations do these constructions encode, and what is the corpus distribution of the semantic relations.

3.1 Lists of semantic classification relations

Although the NP interpretation problem has been studied for a long time, researchers haven't agreed on the number and the level of abstraction of these semantic categories. They can vary from a few prepositions (Lauer, 1995) to hundreds or thousands specific semantic relations (Finin, 1980). The more abstract the categories, the more noun phrases are covered, but also the more room for variation as to which category a phrase should be assigned.

In this paper we experiment with two state of the art classification sets used in NP interpretation. The first is a core set of 22 semantic relations (22 SRs) identified by us from the computational linguistics literature. This list, presented in Table 1 along with examples is general enough to cover a large majority of text semantics while keeping the semantic relations to a manageable number. The second set is Lauer's list of 8 prepositions (8 PP) and can be applied only to noun compounds (*of, for, with, in, on, at, about*, and *from* – e.g., according to this classification, *love story* can be classified as *story* `about` *love*). We selected these sets as they are of different size and contain semantic classification categories at different levels of abstraction. Lauer's list is more abstract and, thus capable of encoding a large number of noun compound instances, while the 22-SR list contains finer grained semantic categories. We show below the coverage of these semantic lists on two different corpora and how well they solve the interpretation problem of noun phrases.

3.2 The data

The data was collected from two text collections with different distributions and of different genre,

POSSESSION (family estate); KINSHIP (sister of the boy); PROPERTY (lubricant viscosity); AGENT (return of the natives); THEME (acquisition of stock); TEMPORAL (morning news); DEPICTION-DEPICTED (a picture of my niece); PART-WHOLE (brush hut); HYPERNYMY (IS-A) (daisy flower); CAUSE (scream of pain); MAKE/PRODUCE (chocolate factory); INSTRUMENT (laser treatment); LOCATION (castle in the desert); PURPOSE (cough syrup); SOURCE (grapefruit oil); TOPIC (weather report); MANNER (performance with passion); beneficiary (rights of citizens); MEANS (bus service); EXPERIENCER (fear of the girl); MEASURE (cup of sugar); TYPE (framework law);

Table 1: The list of 22 semantic relations (22-SRs).

Europarl[1] and CLUVI[2]. The Europarl data was assembled by combining the Spanish-English, Italian-English, French-English and Portuguese-English corpora which were automatically aligned based on exact matches of English translations. Then, we considered only the English sentences which appeared verbatim in all four language pairs. The resulting English corpus contained 10,000 sentences which were syntactically parsed (Charniak, 2000). From these we extracted the first 3,000 NP instances (N N: 48.82% and N P N: 51.18%).

CLUVI is an open text repository of parallel corpora of contemporary oral and written texts in some of the Romance languages. Here, we focused only on the English-Portuguese and English-Spanish parallel texts from the works of John Steinbeck, H. G. Wells, J. Salinger, and others. Using the CLUVI search interface we created a sentence-aligned parallel corpus of 2,800 English-Spanish and English-Portuguese sentences. The English versions were automatically parsed after which each N N and N P N instance thus identified was manually mapped to the corresponding translations. The resulting corpus contains 2,200 English instances with a distribution of 26.77% N N and 73.23% N P N.

3.3 Corpus Annotation

For each corpus, each NP instance was presented separately to two experienced annotators in a web interface in context along with the English sentence and its translations. Since the corpora do not cover some of the languages (Romanian in Europarl and CLUVI, and Italian and French in CLUVI), three other native speakers of these languages and fluent in English provided the translations which were added to the list. The two computational semantics annotators had to tag each English constituent noun with its corresponding WordNet sense and each instance with the corresponding semantic category. If the word was not found in WordNet the instance was not considered. Whenever the annotators found an example encoding a semantic category other than those provided or they didn't know what interpretation to give, they had to tag it as "OTHER-SR", and respectively "OTHER-PP"[3]. The details of the annotation task and the observations drawn from there are presented in a companion paper (Girju, 2007).

The corpus instances used in the corpus analysis phase have the following format: $<$NP$_{En}$;NP$_{Es}$; NP$_{It}$; NP$_{Fr}$; NP$_{Port}$; NP$_{Ro}$; target$>$. The word *target* is one of the 23 (22 + OTHER-SR) semantic relations and one of the eight prepositions considered or OTHER-PP (with the exception of those N P N instances that already contain a preposition). For example, $<$*development cooperation; cooperación para el desarrollo; cooperazione allo sviluppo; coopération au développement; cooperare pentru dezvoltare;* PURPOSE / FOR$>$.

The annotators' agreement was measured using Kappa statistics: $K = \frac{Pr(A)-Pr(E)}{1-Pr(E)}$, where $Pr(A)$ is the proportion of times the annotators agree and $Pr(E)$ is the probability of agreement by chance. The Kappa values were obtained on Europarl (N N: 0.80 for 8-PP and 0.61 for 22-SR; N P N: 0.67 for 22-SR) and CLUVI (N N: 0.77 for 8-PP and 0.56 for 22-SR; N P N: 0.68 for 22-SR). We also computed the number of pairs that were tagged with OTHER by both annotators for each semantic relation and preposition paraphrase, over the number of examples classified in that category by at least one of the judges (in Europarl: 91% for 8-PP and 78% for 22-SR; in CLUVI: 86% for 8-PP and 69% for 22-SR).

The agreement obtained on the Europarl corpus is

[1]http://www.isi.edu/koehn/europarl/. This corpus contains over 20 million words in eleven official languages of the European Union covering the proceedings of the European Parliament from 1996 to 2001.

[2]CLUVI - Linguistic Corpus of the University of Vigo - Parallel Corpus 2.1 - http://sli.uvigo.es/CLUVI/

[3]The annotated corpora resulted in this research is available at http://apfel.ai.uiuc.edu.

higher than the one on CLUVI on both classification sets. This is partially explained by the distribution of semantic relations in both corpora, as will be shown in the next subsection.

3.4 Cross-linguistic distribution of Syntactic Constructions

From the sets of 2,954 (Europarl) and 2,168 (CLUVI) instances resulted after annotation, the data show that over 83% of the translation patterns for both text corpora on all languages were of the type N N and N P N. However, while their distribution is balanced in the Europarl corpus (about 45%, with a 64% N P N – 26% N N ratio for Romanian), in CLUVI the N P N constructions occur in more than 85% of the cases (again, with the exception of Romanian – 50%). It is interesting to note here that some of the English NPs are translated into both noun–noun and noun–adjective compounds in the target languages. For example, *love affair* translates in Italian as *storia d'amore* or the noun–adjective compound *relazione amorosa*. There are also instances that have just one word correspondent in the target language (e.g., *ankle boot* is *bottine* in French). The rest of the data is encoded by other syntactic paraphrases (e.g., *bomb site* is *luogo dove è esplosa la bomba* (It.)). [4].

From the initial corpus we considered those English instances that had all the translations encoded only by N N and N P N. Out of these, we selected only 1,023 Europarl and 1,008 CLUVI instances encoded by N N and N P N in all languages considered and resulted after agreement.

4 Model

4.1 Feature space

We have identified and experimented with 13 NP features presented below. With the exceptions of features F1-F5 (Girju et al., 2005), all the other features are novel.

A. English Features

F1 and **F2.** *Noun semantic class* specifies the WordNet sense of the head (F1) and modifier noun (F2) and implicitly points to all its hypernyms. For example, the hypernyms of *car#1* are: {*motor vehicle*}, .. {*entity*}. This feature helps generalize over the semantic classes of the two nouns in the corpus.

F3 and **F4.** *WordNet derivationally related form* specifies if the head (F3) and the modifier (F4) nouns are related to a corresponding WordNet verb (e.g. *statement* derived from *to state*; *cry* from *to cry*).

F5. *Prepositional cues* that link the two nouns in an NP. These can be either simple or complex prepositions such as "*of*" or "*according to*". In case of N N instances, this feature is "–" (e.g., *framework law*).

F6 and F7. *Type of nominalized noun* indicates the specific class of nouns the head (F6) or modifier (F7) belongs to depending on the verb it derives from. First, we check if the noun is a nominalization. For English we used NomLex-Plus (Meyers et al., 2004) to map nouns to corresponding verbs.[5] For example, "*destruction of the city*", where *destruction* is a nominalization. F6 and F7 may overlap with features F3 and F4 which are used in case the noun to be checked does not have an entry in the NomLex-Plus dictionary. These features are of particular importance since they impose some constraints on the possible set of relations the instance can encode. They take the following values (identified based on list of verbs extracted from VerbNet (Kipper et al., 2000)):

a. Active form nouns which have an intrinsic active voice predicate-argument structure. (Giorgi and Longobardi, 1991) argue that in English this is a necessary restriction. Most of the time, they represent states of emotion, such as fear, desire, etc. These nouns mark their internal argument through *of* and require most of the time prepositions like *por* and not *de* when translated in Romance. Our observations on the Romanian translations (captured by features F12 and F13 below) show that the possible cases of ambiguity are solved by the type of syntactic construction used. For example, N N genitive-marked constructions are used for EXPERIENCER–encoding instances, while *N de N* or *N pentru N* (N for N) are used for other relations. Such examples are *the love of children* – THEME (and not *the love by the children*). (Giorgi and Longobardi, 1991) mention that with such nouns that resist passivisation,

[4] "*the place where the bomb is exploded*" (It.)

[5] NomLex-Plus is a hand-coded database of 5,000 verb nominalizations, de-adjectival, and de-adverbial nouns including the corresponding subcategorization frames (verb-argument structure information).

the preposition introducing the internal argument, even if it is *of*, has always a semantic content, and is not a bare case-marker realizing the genitive case.

b. Unaccusative (ergative) nouns which are derived from ergative verbs that take only internal arguments (e.g., not agentive ones). For example, the transitive verb *to disband* allows the subject to be deleted as in the following sentences (1) "*The lead singer disbanded the group in 1991.*" and (2) "*The group disbanded.*". Thus, the corresponding ergative nominalization *the disbandment of the group* encodes a THEME relation and not AGENT.

c. Unergative (intransitive) nouns are derived from intransitive verbs and take only AGENT semantic relations. For example, *the departure of the girl*.

d. Inherently passive nouns such as *the capture of the soldier*. These nouns, like the verbs they are derived from, assume a default AGENT (subject) and being transitive, associate to their internal argument (introduced by "of" in the example above) the THEME relation.

B. Romance Features

F8, F9, F10, F11 and **F12.** *Prepositional cues* that link the two nouns are extracted from each translation of the English instance: F8 (Es.), F9 (Fr.), F10 (It.), F11 (Port.), and F12 (Ro.). These can be either simple or complex prepositions (e.g., *de*, *in materia de* (Es.)) in all five Romance languages, or the Romanian genitival article *a/ai/ale*. In Romanian the genitive case is assigned by the definite article of the first noun to the second noun, case realized as a suffix if the second noun is preceded by the definite article or as one of the genitival articles *a/ai/ale*. For example, the noun phrase *the beauty of the girl* is translated as *frumuseţea fetei* (*beauty*-the *girl*-gen), and *the beauty of a girl* as *frumuseţea unei fete* (*beauty*-the gen *girl*). For N N instances, this feature is "–".

F13. *Noun inflection* is defined only for Romanian and shows if the modifier noun is inflected (indicates the genitive case). This feature is used to help differentiate between instances encoding IS-A and other semantic relations in N N compounds in Romanian. It also helps in features F6 and F7, case a) when the choice of syntactic construction reflects different semantic content. For example, *iubirea pentru copii* (N P N) (*the love for children*) and not *iubirea copiilor* (N N) (*love expressed by the children*).

4.2 Learning Models

We have experimented with the support vector machines (SVM) model[6] and compared the results against two state-of-the-art models: a supervised model, Semantic Scattering (SS), (Moldovan and Badulescu, 2005), and a web-based unsupervised model (Lapata and Keller, 2004). The SVM and SS models were trained and tested on the Europarl and CLUVI corpora using a 8:2 ratio. The test dataset was randomly selected from each corpus and the test nouns (only for English) were tagged with the corresponding sense in context using a state of the art WSD tool (Mihalcea and Faruque, 2004).

After the initial NP instances in the training and test corpora were expanded with the corresponding features, we had to prepare them for SVM and SS. The method consists of a set of automatic iterative procedures of specialization of the English nouns on the WordNet IS-A hierarchy. Thus, after a set of necessary specialization iterations, the method produces specialized examples which through supervised machine learning are transformed into sets of semantic rules. This specialization procedure improves the system's performance since it efficiently separates the positive and negative noun-noun pairs in the WordNet hierarchy.

Initially, the training corpus consists of examples in the format exemplified by the feature space. Note that for the English NP instances, each noun constituent was expanded with the corresponding WordNet top semantic class. At this point, the generalized training corpus contains two types of examples: unambiguous and ambiguous. The second situation occurs when the training corpus classifies the same noun – noun pair into more than one semantic category. For example, both relationships "*chocolate cake*"-PART-WHOLE and "*chocolate article*"-TOPIC are mapped into the more general type <*entity#1, entity#1,* PART-WHOLE/TOPIC>[7]. We recursively specialize these examples to eliminate the ambiguity. By specialization, the semantic class is replaced with the corresponding hyponym for that particular sense, i.e. the concept immediately below in the hierarchy. These steps are repeated until there are no

[6]We used the package LIBSVM with a radial-based kernel *http://www.csie.ntu.edu.tw/∼cjlin/libsvm/*

[7]The specialization procedure applies only to features 1, 2.

more ambiguous examples. For the example above, the specialization stops at the first hyponym of *entity*: *physical entity* (for *cake*) and *abstract entity* (for *article*). For the unambiguous examples in the generalized training corpus (those that are classified with a single semantic relation), constraints are determined using cross validation on SVM.

A. Semantic Scattering uses a training data set to establish a boundary G^* on WordNet noun hierarchies such that each feature pair of noun – noun senses f_{ij} on this boundary maps uniquely into one of a predefined list of semantic relations, and any feature pair above the boundary maps into more than one semantic relation. For any new pair of noun–noun senses, the model finds the closest WordNet boundary pair.

The authors define with $SC^m = \{f_i^m\}$ and $SC^h = \{f_j^h\}$ the sets of semantic class features for modifier noun and, respectively head noun. A pair of <modifier – head> nouns maps uniquely into a semantic class feature pair $< f_i^m, f_j^h >$, denoted as f_{ij}. The probability of a semantic relation *r* given feature pair f_{ij}, $P(r|f_{ij}) = \frac{n(r,f_{ij})}{n(f_{ij})}$, is defined as the ratio between the number of occurrences of a relation *r* in the presence of feature pair f_{ij} over the number of occurrences of feature pair f_{ij} in the corpus. The most probable semantic relation r̂ is $\arg\max_{r \in R} P(r|f_{ij}) = \arg\max_{r \in R} P(f_{ij}|r)P(r)$.

B. (Lapata and Keller, 2004)'s web-based unsupervised model classifies noun - noun instances based on Lauer's list of 8 prepositions and uses the web as training corpus. They show that the best performance is obtained with the trigram model $f(n_1, p, n_2)$. The count used for a given trigram is the number of pages returned by Altavista on the trigram corresponding queries. For example, for the test instance *war stories*, the best number of hits was obtained with the query *stories about war*.

For the Europarl and CLUVI test sets, we replicated Lapata & Keller's experiments using Google[8]. We formed inflected queries with the patterns they proposed and searched the web.

[8] As Google limits the number of queries to 1,000 per day, we repeated the experiment for a number of days. Although (Lapata and Keller, 2004) used Altavista in their experiments, they showed there is almost no difference between the correlations achieved using Google and Altavista counts.

5 Experimental results

Table 2 shows the results obtained against SS and Lapata & Keller's model on both corpora and the contribution the features exemplified in one baseline and six versions of the SVM model. The baseline is defined only for the English part of the NP feature set and measures the the contribution of the WordNet IS-A lexical hierarchy specialization. The baseline does not differentiate between unambiguous and ambiguous training examples (after just one level specialization) and thus, does not specialize the ambiguous ones. Moreover, here we wanted to see what is the difference between SS and SVM, and what is the contribution of the other English features, such as preposition and nominalization (F1–F7).

The table shows that, overall the performance is better for the Europarl corpus than for CLUVI. For the Baseline and SVM_1, SS [F1 + F2] gives better results than SVM. The inclusion of other English features (SVM [F1–F7]) adds more than 15% (with a higher increase in Europarl) for SVM_1.

The contribution of Romance linguistic features. Since our intuition is that the more translations are provided for an English noun phrase instance, the better the results, we wanted to see what is the impact of each Romance language on the overall performance. Thus, SVM_2 shows the results obtained for English and the Romance language that contributed the least to the performance (F1–F12). Here we computed the performance on all five English – Romance language combinations and chose the Romance language that provided the best result. Thus, SVM #2, #3, #4, #5, and #6 add Spanish, French, Italian, Portuguese, and Romanian in this order and show the contribution of each Romance preposition and all features for English.

The language ranking in Table 2 shows that Romance languages considered here have a different contribution to the overall performance. While the addition of Italian in Europarl decreases the performance, Portuguese doesn't add anything. However, a closer analysis of the data shows that this is mostly due to the distribution of the corpus instances. For example, French, Italian, Spanish, and Portuguese are most of the time consistent in the choice of preposition (e.g. most of the time, if the preposition 'de' ('of') is used in French, then the

Learning models		**Results** [%]			
		CLUVI		**Europarl**	
		8-PP	**22-SR**	**8-PP**	**22-SR**
Baseline (En.) (no specializ.)	SS (F1+F2)	44.11	48.03	38.7	38
	SVM (F1+F2)	36.37	40.67	31.18	34.81
	SVM (F1-F7)	–	52.15	–	47.37
SVM$_1$ (En.)	SS (F1+F2)	56.22	61.33	53.1	56.81
	SVM (F1+F2)	45.08	46.1	40.23	42.2
	SVM (F1-F7)	–	62.54	–	74.19
SVM$_2$ (En. + Es.)	SVM (F1-F8)	–	64.18	–	75.74
SVM$_3$ (En.+Es.+Fr.)	SVM (F1-F9)	–	67.8	–	76.52
SVM$_4$ (En.+Es.+Fr.+It.)	SVM (F1-F10)	–	66.31	–	75.74
SVM$_5$ (En.+Es.+Fr.+It+Port.)	SVM (F1-F11)	–	67.12	–	75.74
SVM$_6$ (En.+Romance: F1–F13)		–	**74.31**	–	**77.9**
Lapata & Keller's unsupervised model (En.)		44.15	–	45.31	–

Table 2: The performance of the cross-linguistic SVM models compared against one baseline, SS model and Lapata & Keller's unsupervised model. *Accuracy* (number of correctly labeled instances over the number of instances in the test set).

corresponding preposition is used in the other four language translations). A notable exception here is Romanian which provides two possible constructions: the N P N and the genitive-marked N N. The table shows (in the increase in performance between SVM_5 and SVM_6) that this choice is not random, but influenced by the meaning of the instances (features F12, F13). This observation is also supported by the contribution of each feature to the overall performance. For example, in Europarl, the WordNet verb and nominalization features of the head noun (F3, F6) have a contribution of 4.08%, while for the modifier nouns it decreases by about 2%. The preposition (F5) contributes 4.41% (Europarl) and 5.24% (CLUVI) to the overall performance.

A closer analysis of the data shows that in Europarl most of the N N instances were naming noun compounds such as *framework law* (TYPE) and, most of the time, are encoded by N N patterns in the target languages (e.g., *legge quadro* (It.)). In the CLUVI corpus, on the other hand, the N N Romance translations represented only 1% of the data. A notable exception here is Romanian where most NPs are represented as genitive–marked noun compounds. However, there are instances that are encoded mostly or only as N P N constructions and this choice correlates with the meaning of the instance. For example, *the milk glass* (PURPOSE) translates as *paharul de lapte* (*glass*-the *of milk*) and not as *paharul laptelui* (*glass*-the *milk*-gen), *the olive oil* (SOURCE) translates as *uleiul de măsline* (*oil*-the *of olive*) and not as *uleiul măslinei* (*oil*-the *olive*-gen). Other examples include CAUSE and TOPIC.

Lauer's set of 8 prepositions represents 94.5% (Europarl) and 97% (CLUVI) of the N P N instances. From these, the most frequent preposition is "of" with a coverage of 70.31% (Europarl) and 85.08% (CLUVI). Moreover, in the Europarl corpus, 26.39% of the instances are synthetic phrases (where one of the nouns is a nominalization) encoding AGENT, EXPERIENCER, THEME, BENEFICIARY. Out of these instances, 74.81% use the preposition *of*. In CLUVI, 11.71% of the examples were verbal, from which the preposition *of* has a coverage of 82.20%. The many-to-many mappings of the prepositions (especially *of/de*) to the semantic classes adds to the complexity of the interpretation task. Thus, for the interpretation of these constructions a system must rely on the semantic information of the preposition and two constituent nouns in particular, and on context in general.

In Europarl, the most frequently occurring relations are PURPOSE, TYPE, and THEME that together represent about 57% of the data followed by PART-WHOLE, PROPERTY, TOPIC, AGENT, and LOCATION with an average coverage of about 6.23%. Moreover, other relations such as KINSHIP, DEPICTION, MANNER, MEANS did not occur in this corpus and 5.08% represented OTHER-SR relations. This semantic distribution contrasts with the one in CLUVI, which uses a more descriptive language. Here, the most frequent relation by far is PART-WHOLE (32.14%), followed by LOCATION (12.40%), THEME (9.23%) and OTHER-SR (7.74%). It is interesting to note here that only 5.70% of the TYPE relation instances in Europarl were unique. This is in contrast with the other relations in both corpora, where instances were mostly unique.

We also report here our observations on Lapata & Keller's unsupervised model. An analysis of these results showed that the order of the constituent nouns in the N P N paraphrase plays an important role. For example, a search for *blood vessels* generated similar frequency counts for *vessels of blood* and *blood in vessels*. About 30% noun - noun paraphrasable pairs preserved the order in the corresponding N P N paraphrases. We also manually checked the first five entries generated by Google for each most frequent prepositional paraphrase for 50 instances and noticed that about 35% of them were wrong due to syntactic and/or semantic ambiguities. Thus, since we wanted to measure the impact of these ambiguities of noun compounds on the interpretation performance, we further tested the probabilistic web-based model on four distinct test sets selected from Europarl, each containing 30 noun - noun pairs encoding different types of ambiguity: in set#1 the noun constituents had only one part of speech and one WordNet sense; in set#2 the nouns had at least two possible parts of speech and were semantically unambiguous, in set#3 the nouns were ambiguous only semantically, and in set#4 they were ambiguous both syntactically and semantically. For unambiguous noun-noun pairs (set#1), the model obtained an accuracy of 35.01%, while for more semantically ambiguous compounds it obtained an accuracy of about 48.8%. This shows that for more semantically ambiguous noun - noun pairs, the web-based probabilistic model introduces a significant number of false positives. Thus, the more abstract the categories, the more noun compounds are covered, but also the more room for variation as to which category a compound should be assigned.

6 Discussion and Conclusions

In this paper we presented a supervised, knowledge-intensive interpretation model which takes advantage of new linguistic information from English and a list of five Romance languages. Our approach to NP interpretation is novel in several ways. We defined the problem in a cross-linguistic framework and provided empirical observations on the distribution of the syntax and meaning of noun phrases on two different corpora based on two state-of-the-art classification tag sets.

As future work we consider the inclusion of other features such as the semantic classes of Romance nouns from aligned EuroWordNets, and other sentence features. Since the results obtained can be seen as an upper bound on NP interpretation due to perfect English - Romance NP alignment, we will experiment with automatic translations generated for the test data. Moreover, we like to extend the analysis to other set of languages whose structures are very different from English and Romance.

References

T. W. Finin. 1980. *The Semantic Interpretation of Compound Nominals*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

A. Giorgi and G. Longobardi. 1991. The syntax of noun phrases. Cambridge University Press.

R. Girju, D. Moldovan, M. Tatu, and D. Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, 19(4):479–496.

R. Girju. 2007. Experiments with an annotation scheme for a knowledge-rich noun phrase interpretation system. The Linguistic Annotation Workshop at ACL, Prague.

Su Nam Kim and T. Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. *COLING-ACL*.

K. Kipper, H. Dong, and M. Palmer. 2000. Class-based construction of a verb lexicon. *AAAI* Conference, Austin.

M. Lapata and F. Keller. 2004. The Web as a baseline: Evaluating the performance of unsupervised Web-based models for a range of NLP tasks. *HLT-NAACL*.

M. Lauer. 1995. Corpus statistics meet the noun compound: Some empirical results. *ACL*, Cambridge, Mass.

A. Meyers, R. Reeves, C. Macleod, R. Szekeley V. Zielinska, and B. Young. 2004. The cross-breeding of dictionaries. *LREC-2004*, Lisbon, Portugal.

R. Mihalcea and E. Faruque. 2004. Senselearner: Minimally supervised word sense disambiguation for all words in open text. *ACL/SIGLEX Senseval-3*, Barcelona, Spain.

D. Moldovan and A. Badulescu. 2005. A semantic scattering model for the automatic interpretation of genitives. *HLT/EMNLP Conference*, Vancouver, Canada.

P. Pantel and M. Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. *COLING/ACL*, Sydney, Australia.

B. Rosario and M. Hearst. 2001. Classifying the semantic relations in noun compounds. *EMNLP Conference*.

R. Snow, D. Jurafsky, and A. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. *COLING-ACL*.

P. Turney. 2006. Expressing implicit semantic relations without supervision. *COLING/ACL*, Sydney, Australia.