Support Vector Machines Applied to the Classification of Semantic Relations in Nominalized Noun Phrases

Roxana Girju
Computer Science Department
Baylor University
Waco, Texas
girju@cs.baylor.edu

Ana-Maria Giuglea, Marian Olteanu,
Ovidiu Fortu, Orest Bolohan, and
Dan Moldovan
Department of Computing Science
University of Texas at Dallas
Dallas, Texas
moldovan@utdallas.edu

Abstract

The discovery of semantic relations in text plays an important role in many NLP applications. This paper presents a method for the automatic classification of semantic relations in nominalized noun phrases. Nominalizations represent a subclass of NP constructions in which either the head or the modifier noun is derived from a verb while the other noun is an argument of this verb. Especially designed features are extracted automatically and used in a Support Vector Machine learning model. The paper presents preliminary results for the semantic classification of the most representative NP patterns using four distinct learning models.

1 Introduction

1.1 Problem description

The automatic identification of semantic relations in text has become increasingly important in Information Extraction, Question Answering, Summarization, Text Understanding, and other NLP applications. This paper discusses the automatic labeling of semantic relations in nominalized noun phrases (NPs) using a support vector machines learning algorithm.

Based on the classification provided by the New Webster's Grammar Guide (Semmelmeyer and Bolander 1992) and our observations of noun phrase patterns on large text collections, the most frequently occurring NP level constructions are: (1) *Compound Nominals* consisting of two consecutive nouns (eg *pump drainage* - an INSTRUMENT relation), (2) *Adjective Noun* constructions where the adjectival modifier is derived from a noun (eg *parental refusal* - AGENT), (3) *Genitives* (eg *tone of conversation* - a PROPERTY relation), (4) *Adjective phrases* in which the modifier noun is expressed by a prepositional phrase which functions as an adjective (eg *amusement in the park* - a LOCATION relation), and (5) *Adjective clauses* where the head noun is modified by a relative clause (eg *the man who was driving the car* - an AGENT relation between *man* and *driving*).

1.2 Previous work on the discovery of semantic relations

The development of large semantically annotated corpora, such as Penn Treebank2 and, more recently, PropBank (Kingsbury, et al. 2002), as well as semantic knowledge bases, such as FrameNet (Baker, Fillmore, and Lowe 1998), have stimulated a high interest in the automatic acquisition of semantic relations, and especially of semantic roles. In the last few years, many researchers (Blaheta and Charniak 2000), (Gildea and Jurafsky 2002), (Gildea and Palmer 2002), (Pradhan et al. 2003) have focused on the automatic prediction of semantic roles using statistical techniques. These statistical techniques operate on the output of probabilistic parsers and take advantage of the characteristic features of the semantic roles that are then employed in a learning algorithm.

While these systems focus on verb-argument semantic relations, called semantic roles, in this paper we investigate predicate-argument semantic relations in nominalized noun phrases and present a method for their automatic detection in open-text.

1.3 Approach

We approach the problem top-down, namely identify and study first the characteristics or feature vectors of each noun phrase linguistic pattern and then develop models for their semantic classification. The distribution of the semantic relations is studied across different NP patterns and the similarities and differences among resulting semantic spaces are analyzed. A thorough understanding of the syntactic and semantic characteristics of NPs provides valuable insights into defining the most representative *feature vectors* that ultimately drive the discriminating learning models.

An important characteristic of this work is that it relies heavily on state-of-the-art natural language processing and machine learning methods. Prior to the discovery of semantic relations, the text is syntactically parsed with Charniak's parser (Charniak 2001) and words are semantically disambiguated and mapped into their appropriate WordNet senses. The word sense disambiguation is done manually for training and automatically for testing with a state-of-the-art WSD module, an improved version of a system with which we have participated successfully in Senseval 2 and which has an accuracy of 81% when disambiguating nouns in open-domain. The discovery of semantic relations is based on learning lexical, syntactic, semantic and contextual constraints that effectively identify the most probable relation for each NP construction considered.

2 Semantic Relations in Nominalized Noun Phrases

In this paper we study the behavior of semantic relations at the *noun phrase level* when one of the nouns is nominalized. The following NP level constructions are considered: *complex nominals, genitives, adjective phrases*, and *adjective clauses*.

Complex Nominals
Levi (Levi 1979) defines complex nominals (CNs) as expressions that have a head noun preceded by one or more modifying nouns, or by adjectives derived from nouns (usually called denominal adjectives). Each sequence of nouns, or possibly adjectives and nouns, has a particular meaning as a whole carrying an implicit semantic relation; for example, "*parental refusal*" (AGENT).

The main tasks are the *recognition*, and the *interpretation* of complex nominals. The recognition task deals with the identification of CN constructions in text, while the interpretation of CNs focuses on the detection and classification of a comprehensive set of semantic relations between the noun constituents.

Genitives
In English there are two kinds of genitives; in one, the modifier is morphologically linked to the possessive clitic *'s* and precedes the head noun (s-genitive e.g. "*John's conclusion*"), and in the second one the modifier is syntactically marked by the preposition *of* and follows the head noun (of-genitive, e.g. "*declaration of independence*").

Adjective Phrases are prepositional phrases attached to nouns and act as adjectives (cf. (Semmelmeyer and Bolander 1992)). Prepositions play an important role both syntactically and semantically ((Dorr 1997). Prepositional constructions can encode various semantic relations, their interpretations being provided most of the time by the underlying context. For instance, the preposition "*with*" can encode different semantic relations: (1) It was the girl *with* blue eyes (MERONYMY), (2) The baby *with* the red ribbon is cute (POSSESSION), (3) The woman *with* triplets received a lot of attention (KINSHIP).
The conclusion for us is that in addition to the nouns semantic classes, the preposition and the context play important roles here.

Adjective Clauses are subordinate clauses attached to nouns (cf. (Semmelmeyer and Bolander 1992)). Often they are introduced by a relative pronoun/adverb (ie *that, which, who, whom, whose, where*) as in the following examples: (1) Here is *the book which I am reading* (*book* is the THEME of *reading*) (2) *The man who was driving the car was a spy* (*man* is the AGENT of *driving*). Adjective clauses are inherently verb-argument structures, thus their interpretation consists of detecting the semantic role between the head noun and the main verb in the relative clause. This is addressed below.

3 Nominalizations and Mapping of NPs into Grammatical Role Structures

3.1 Nominalizations

A further analysis of various examples of noun - noun pairs encoded by the first three major types of NP-level constructions shows the need for a different taxonomy based on the syntactic and grammatical roles the constituents have in relation to each other. The criterion in this classification splits the noun - noun examples (respectively, adjective - noun examples in complex nominals) into **nominalizations** and **non-nominalizations**. Nominalizations represent a particular subclass of NP constructions that in general have "a systematic correspondence with a clause structure" (Quirk et al.1985). The head or modifier noun is derived from a verb while the other noun (the modifier, or respectively, the head) is interpreted as an argument of this verb. For example, the noun phrase "*car owner*" corresponds to "*he owns a car*". The head noun *owner* is morphologically related to the verb *own*. Otherwise said, the interpretation of this class of NPs is reduced to the automatic detection and interpretation of semantic roles mapped on the corresponding verb-argument structure.

As in (Hull and Gomez 1996), in this paper we use the term *nominalization* to refer only to those **senses of the nominalized nouns** which are derived from verbs. For example, the noun "*decoration*" has three senses in WordNet 2.0: *an ornament* (#1), *a medal* (#2), and *the act of decorating* (#3). Only the last sense is a nominalization. However, there are more complex situations when the underlying verb has more than one sense that refers to an action/event. This is the case of "*examination*" which has five senses of which four are action-related. In this case, the selection of the correct sense is provided by the context.

We are interested in answering the following questions: *(1) What is the best set of features that can capture the meaning of noun - noun nominalization pairs for each NP-level construction?* and *(2) What is the semantic behavior of nominalization constructions across NP levels?*

3.2 Taxonomy of nominalizations

Deverbal vs verbal noun.

(Quirk et al.1985) generally classify nominalizations based on the morphological formation of the nominalized noun. They distinguish between *deverbal* nouns, i.e. those derived from the underlying verb through word formation; e.g., "*student examination*", and *verbal* nouns, i.e. those derived from the verb by adding the gerund suffix "-ing"; e.g.: "*cleaning woman*". Most of the time, verbal nouns are derived from verbs which don't have a deverbal correspondent.

Table 1 shows the mapping of the first three major syntactic NP constructions to the grammatical role level. By analyzing a large corpus, we have observed that Quirk's grammatical roles shown in Table 1 are not uniformly distributed over the types of NP-constructions. For example, the "$N_{Obj} - N_{AgentialNoun}$" pattern cannot be encoded by s-genitives (e.g., "*language teacher*", "*teacher of language*").

Some of the non-nominalization NP constructions can also capture the arguments of a particular verb that is missing (e.g., *subject - object*, *subject - complement*). The "*General*" subclass refers to all other types of noun - noun constructions that cannot be mapped on verb-argument relations (e.g., "*hundreds of dollars*"). Adjective clauses are not part of Table 1 as they describe by default verb-argument relations (semantic roles). Thus they cannot be classified as nominalizations or non-nominalizations.

Two other useful classifications for nominalizations are: *paraphrased vs. non-paraphrased*, and the classification according to the nominalized noun's verb-argument underlying structures as provided by the *NomLex dictionary* of English nominalizations (Macleod et al. 1998) discussed more later.

Paraphrased vs non-Paraphrased.

In most cases, the relation between the nominalized noun and the other noun argument can be captured from the subcategorization properties of the underlying verb. Otherwise said, most of the time, there is a systematic correspondence between the nominalized NP construction and the predicate-argument structure of the corresponding verb in a clausal paraphrase (*paraphrased nominalization*). The predicate-argument structure can be captured by three grammatical roles: *verb-subject*, *verb-object*, and *verb-complement*. We call the arguments of the verb that appear more frequently or are obligatory - *frame arguments*. From this point of view the non-nominalized noun can be mapped on the verb-argument frame or not. Thus we can classify paraphrased nominalizations in *framed* and *non-framed* according to the presence or absence of the non-nominalized noun in the frame of the verb. The semantic classification of nominalizations involves first the detection of a nominalization, the selection of the correct sense of the root verb, and finally the detection of the semantic relationship with the other noun.

Besides the *paraphrase nominalization*, there is another type which occurs less frequently. We call this type *non-paraphrased nominalization* as its meaning is different from its most related paraphrase clause. Examples: *research budget*, *design contract*, *preparation booklet*, *publishing sub-industry* and *editing error*. An important observation is that the nominalized noun occurs most of the time on the first position in an NP construction.

The criteria presented here consider also nominalizations with adjectival modifiers such as "*parental refusal*". These adjectives are derived from nouns, so the construction is just a special case of nominalization between nouns.

NomLex classification

The *NomLex dictionary of nominalizations* (Macleod et al. 1998) contains 1025 lexical entries and lists the verbs from which the nouns are derived. This dictionary specifies the complements allowed for a nominalization. The mapping is done at a syntactic level only. NomLex is used in the first phase of our algorithm in order to detect a possible nominalization and the corresponding root verb. The criterion of NomLex classification is based on the verb-argument correspondence:

a. *Verb-nom*: The nominalized noun represents the action/state of the verb (e.g., "*acquisition challenge*", "*depositary receipt*)",

b. *Subj-nom*: The nominalized noun refers to the subject of the verb (e.g., "*auto maker*", "*math teacher*"). This type is also called *agential nominalization* (Quirk et al.1985) as the nominalized noun captures information about both the subject and verb.

c. *Obj-nom*: The nominalized noun refers to the object of the verb (e.g., "*court order*", "*company employee*"),

d. *Verb-part*: the nominalized noun is derived from a compositional verb (e.g., "*takeover target*").

3.3 Corpus Analysis at NP level

The data

We have assembled a corpus from the Wall Street Journal articles from TREC-9. Table 2 shows for each syntactic

Syntactic Patterns		Grammatical Roles	CNs		Genitives		Adjective Phrase	Example
			N N	Adj-N	’s	of		
Nominalization	Deverbal noun	$N_{Obj} - N_{DeverbalNoun}$	∠	∠	∠	∠		*“heart massage”*
		$N_{Obj} - N_{AgentialNoun}$	∠	∠		∠		*“language teacher”*
		$N_{AdvNoun} - N_{DeverbalNoun}$	∠	∠		∠	∠	*“smallpox vaccination”*
		$N_{Subj} - N_{DeverbalNoun}$	∠	∠	∠	∠		*“boy’s application”*
	Verbal noun	$N_{VerbalNoun-ing} - N_{Subj}$	∠			∠		*“cleaning woman”*
		$N_{VerbalNoun-ing} - N_{Obj}$	∠			∠		*“spending money”*
		$N_{VerbalNoun-ing} - N_{AdvNoun}$	∠				∠	*“carving knife”*
		$N_{AdvNoun} - N_{VerbalNoun-ing}$	∠	∠			∠	*“horse riding”*
Non-nominalization		$N_{Subj} - N_{Obj}$	∠	∠			∠	*“wind mill”*
		$N_{Subj} - N_{Complement}$	∠	∠			∠	*“pine tree”*
		General	∠	∠	∠	∠	∠	*“hundreds of dollars”*

Table 1: Classification of noun phrase constructions on the types of grammatical roles (cf. (Quirk et al.1985)) needed for semantic roles detection. The classification is the result of our observations of nominalization patterns at noun phrase level.

category the number of randomly selected sentences, the number of instances found in these sentences, and finally the number of nominalized instances our group managed to annotate by hand. The annotation of each example consisted of specifying its feature vector and the most appropriate semantic relation as defined in (Moldovan et al. 2004).

Inter-annotator Agreement

The annotators, four PhD students in Computational Semantics worked in groups of two, each group focusing on one half of the corpus to annotate. Besides the type of relation, the annotators were asked to provide information about the order of the modifier and the head nouns in the syntactic constructions if applicable. For example, “*owner of the car*” and “*car of the owner*”.

The annotators were also asked to indicate if the instance was a nominalization and if yes, which of the noun constituents was derived from a verb (e.g. the head noun nominalization “*student protest*”, or the modifier noun nominalization “*working woman’*” cf. (Quirk et al.1985)).

The annotators’ agreement was measured using the Kappa statistics (Siegel and Castellan 1988), one of the most frequently used measure of inter-annotator agreement for classification tasks: $K = \frac{Pr(A)-Pr(E)}{1-Pr(E)}$, where $Pr(A)$ is the proportion of times the raters agree and $Pr(E)$ is the probability of agreement by chance. The K coefficient is 1 if there is a total agreement among the annotators, and 0 if there is no agreement other than that expected to occur by chance.

For each construction, the corpus was split after agreement with an 80/20 training/testing ratio. For each pattern, we computed the K coefficient only for those instances tagged with one of the 35 semantic relations (K value for: NN (0.64), AdjN (0.70), s-genitive (0.69), of-genitive (0.73), adjective phrases (0.67), and adjective clauses (0.71)). For each pattern, we also calculated the number of pairs that were tagged with OTHERS by both annotators, over the number of examples classified in this category by at least one of the judges, averaged by the number of patterns considered (agreement for OTHERS: 75%).

The K coefficient shows a good level of agreement for the training and testing data on the set of 35 relations, taking into consideration the task difficulty. This can be explained by the instructions the annotators received prior to annotation and by their expertise in lexical semantics.

3.4 Distribution of Semantic Relations

Even noun phrase constructions are very productive allowing for a large number of possible interpretations, Table 3 shows that a relatively small set of 35 semantic relations covers a significant part of the semantic distribution of these constructions on a large open-domain corpus. Moreover, the distribution of these relations is dependent on the type of NP construction, each type encoding a particular subset. For example, in the case of s-genitives, there were 13 relations found from the total of 35 relations considered. The most frequently occurring relations were AGENT, TEMPORAL, LOCATION, and THEME. By comparing the subsets of semantic relations in each column we can notice that these semantic spaces (the set of semantic relations an NP construction can encode) are not identical, proving our initial intuition that the NP constructions cannot be alternative ways of packing the same information. Table 3 also shows that there is a subset of semantic relations that can be fully encoded by all types of NP constructions. The statistics about the annotated *nominalized* examples are as follows (lines 3 and 4 in Table 2): N-N (32.30%), Adj-N (30.80%), s-genitive (21.09%), of-genitive (21.8%), adjective phrase (40.5%). 80% of the examples in adjective phrases (respectively in 94% in s-genitives) had the nominalized noun on the head position.

This simple analysis leads to the important conclusion that the NP constructions must be treated separately as their semantic content is different. We can draw from here the following conclusions:

1. Not all semantic relations can be encoded by all NP

	Wall Street Journal					
	CNs		Genitives		Adjective Phrases	Adjective Clauses
	NN	AdjN	’s	of		
No. of sentences	7067	5381	50291	27067	14582	31568
No. of instances	5557	500	2990	4185	3502	6520 2
No. of annotated instances	2315	383	1816	3404	1341	563
No. of annotated nominalized instances	747	118	383	742	543	563
No. of annotated nominalized instances used in the learning task	312	118	383	344	297	563

Table 2: Corpus statistics.

syntactic constructions.
2. There are semantic relations that have preferences over particular syntactic constructions.

3.5 Model

3.6 Support Vector Machines

Support Vector Machines (SVM) have a strong mathematical foundation (Vapnik 1982) and have been applied successfully to text classification (Tong and Koller 2001), speech recognition, and other applications. We applied SVM to the semantic classification problem and obtained encouraging results.

SVM algorithms are a special class of hyperplane classifiers that use the information encoded in the dot-products of the transformed feature vectors as a similarity measure. The similarity between two instances $\mathbf{x}$ and $\mathbf{x'}$ is given as a function $K : X \times X \rightarrow I\!R$, $K(\mathbf{x}, \mathbf{x'}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x'})$. The *Kernel function* K is the inner product of the non-linear function $\Phi : X \rightarrow F$ that maps the original feature vectors into real feature space.

The function that provides the best classification is of the form: $f(x) = sgn \sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}, \mathbf{x_i})$. The vectors $\mathbf{x_i}$ for which the Lagrange multipliers $\alpha_i \neq 0$ are called *support vectors*. Intuitively, they are the closest to the separating hyperplane. SVM provide good classifiers with few, well chosen training examples.

In order to achieve classification in n classes, $n > 2$, a binary classifier is built for each pair of classes (a total of C_n^2 classifiers). A voting procedure is then used to establish the class of a new example. For the experiments with semantic relations, the simplest voting scheme has been chosen; each binary classifier has one vote which is assigned to the class it chooses when it is run. Then the class with the largest number of votes is considered to be the answer. Using the specific nature of the semantic relation detection problem, new voting schemes can be designed, with good perspectives of improving the overall precision.

The software used in these experiments is the package LIBSVM, *http://www.csie.ntu.edu.tw/∼cjlin/libsvm/* which implements the SVM algorithm described above.

The choice of the kernel is the most difficult part of applying SVM algorithms as the performance of the classifier might be enhanced with a judicious choice of the kernel. We used in our experiments 4 types of general kernels (linear, polynomial, radial-based and sigmoid), with good results. All of them had nearly the same performance, with slight deviations between 2% and 4% on a reduced testing set. However, remarkable is the fact that all classifiers, regardless of the kernel used, made the same mistakes (misclassified the same examples - eg, a classifier with 58% precision makes the same mistakes as one with 62% precision, plus some of its own, and this situation occurred even when the two classifiers had different kernels), while the overall precision seems to be around to the same value during the coefficient tuning. This shows that the limitation is rather imposed by the classification task than by the kernel type.

3.7 Feature space

The key to a successful semantic classification of NP constructions is the identification of their most specific lexical, syntactic, semantic and contextual features. We developed algorithms for finding their values automatically. The values of these features are determined with the help of some important **resources** mentioned below.

ComLex (Grishman et al.1994) is a computational lexicon providing syntactic information for more than 38,000 English headwords. It contains detailed syntactic information about the attributes of each lexical item and the subcategorization frames when words have arguments. This last feature is the most useful for our task as the senses of verbs are clustered by the syntactic frames. We will use ComLex in combination with VerbLeX to map the syntactic behaviors to verb semantic classes.

VerbLeX is an in-house verb lexicon built by enriching VerbNet (Kipper et al. 2000) with verb synsets from WordNet and verbs extracted from the semantic frames of FrameNet. It contains information about the semantic roles that can appear within a class of verbs together with the selectional restrictions for their lexical realizations, syntactic subcategorization and WordNet verb senses. The syntactic information is less detailed than in ComLex, but a mapping between these two resources will provide both the semantic and syntactic information needed for the task. From the total of 13,213 verbs in the extended VerbNet, 6,077 were distinct. It also provides a mapping from the FrameNet deep semantic roles to general thematic roles (list defined in (Moldovan et al. 2004)), and use cases for VerbNet.

No.	Semantic Relations	Frequency (%)						Examples
		CNs		Genitives		Adjective Phrases	Adjective Clauses	
		NN	AdjN	’s	of			
1	POSSESSION	0.46	1.16	1.06	0.37	0	0.23	*“stock holders”*
2	KINSHIP	0	0	0	0	0	0	
3	ATTRIBUTE-HOLDER	1.37	6.97	1.41	8.24	1.82	0	*“intensity of intervention”*
4	AGENT	15.13	23.25	33.68	1.87	11.41	25.81	*“trading companies”*
5	TEMPORAL	0.92	0	26.24	1.50	11.87	6.27	*“date of purchase”*
6	DEPICTION-DEPICTED	0	0	0	0.75	0	0	*“evidence of cheating”*
7	PART-WHOLE	0	8.13	1.41	3.37	1.82	0	*“world consumer”*
8	IS-A (HYPERNYMY)	0	0	0	0	0	0	
9	ENTAIL	0	0	0	0	0	0	
10	CAUSE	0.46	0	0.35	1.12	0.91	0	*“fire destruction”*
11	MAKE/PRODUCE	0	3.48	0.35	3.00	0.91	0	*“computer’s maker”*
12	INSTRUMENT	0	0	0	0.75	0.45	0.46	*“oven cooking”*
13	LOCATION/SPACE	2.75	16.27	3.19	0.75	8.67	4.65	*“meeting in Philadelphia”*
14	PURPOSE	10.09	3.48	0	1.50	5.93	0	*“research budget”*
15	SOURCE	0	13.95	0	0.37	3.19	0.46	*“Japanese buyer”*
16	TOPIC	24.77	3.48	0	9.73	5.02	6.51	*“price discussion”*
17	MANNER	4.13	0	0	0	1.37	2.32	*“shock reaction”*
18	MEANS	0	0	0	0	0	0	
19	ACCOMPANIMENT	0	0	0	0	0	0	
20	EXPERIENCER	0	1.16	0.35	0	1.37	2.79	*“risk for the investor”*
21	RECIPIENT	0	0	0.35	0.37	0.45	0.23	*“ovations for the champions”*
22	FREQUENCY	0	8.13	0	0	0	0	*“daily jogging”*
23	INFLUENCE	0	0	0	0	0	0	
24	ASSOCIATED WITH	0	1.16	0.71	1.12	1.82	0	*“designer’s attorney”*
25	MEASURE	0	0	0	5.24	0.45	0	*“5-mile running”*
26	SYNONYMY	0	0	0	0	0	0	
27	ANTONYMY	0	0	0	0	0	0	
28	PROBABILITY	0	0	0	0	0	0.46	*“chance that he is single”*
29	POSSIBILITY	0	0	0	0	0	0	
30	CERTAINTY	0	0	0	0	0	0	
31	THEME	25.22	4.65	22.34	43.82	32.87	38.14	*“use of cards”*
32	RESULT	5.04	0	0.35	3.00	0.45	1.62	*“ship construction”*
33	STIMULUS	0	0	0	0	0	2.32	*“the beautiful painting he saw”*
34	EXTENT	0	0	0	0	0	0	
35	PREDICATE	0.46	0	0	0	0.45	0.70	*“sun king, as Louis XVI is called”*
OTHERS		4.58	4.65	8.15	13.11	8.67	6.98	*“editing error”*
Total no. of examples		100% (218)	100% (86)	100% (282)	100% (267)	100% (219)	100% (430)	

Table 3: The distribution of the semantic relations on the annotated corpus after agreement. The list of 35 semantic relations was presented in (Moldovan et al. 2004). The percentages represent the number of examples that encode a semantic relation for a particular pattern. The last row shows the number of examples covered by each pattern in the entire annotated corpus (1502 pairs).

An essential aspect of our approach below is the word sense disambiguation (**WSD**) of the content words (nouns, verbs, adjectives and adverbs). Using a state-of-the-art open-text WSD system, each word is mapped into its corresponding **WordNet 2.0** sense. When disambiguating each word, the WSD algorithm takes into account the surrounding words, and this is one important way through which *context* gets to play a role in the semantic classification of NPs.

So far, we have identified and experimented with the following NP **features**:

1. Semantic class of the non-nominalized noun. The non-nominalized noun is classified into one of the 39 EuroWordNet noun semantic classes. VerbNet classes extended in VerbLeX contain selectional restrictions for different semantic roles inside the verb frame. These restrictions are provided based on the EuroWordNet noun semantic classes. Example: “*computer maker*”, where “*computer*” is mapped to the ABSTRACT noun category in EuroWordNet. We intend to map the EuroWordNet top noun semantic classes into their WordNet correspondents.

2. Verb class for nominalized noun, or verb in adjective clauses maps the nominalizing verb into its VerbLeX class. The intuition behind this feature is that semantic relations cluster around specific VerbLeX verb classes.

3. Type of nominalization indicates the NomLex nominalization class. For this experiment we considered only examples that could be found in NomLex. By specifying *subj-nom*, *obj-nom*, and *verb-nom* types of nominalization, we reduce the list of possible semantic relations the verb can have with the non-nominalized noun. Example: “*computer maker*”, where “*maker*” is an agential deverbal noun that captures both the subject (respectively, AGENT) and the verb. Thus, the noun “*computer*” can only map to object (respectively, THEME).

4. Verbal nominalization is a binary feature indicating whether the nominalized noun is gerundive or

not. Chomsky (Chomsky 1970) showed that gerundive nominalizations have different behavior than derived nominalizations. Example: ”*woman worker*” vs “*working woman*”; here “*working*” is a verbal nominal.

5. Semantic class of the coordinating word. This is a contextual feature and can be either a noun (if the phrase that contains the nominalization is attached to a noun) or a verb (if the phrase is an argument of the verb in the sentence). The feature value is either the VerbLeX class of the verb or the root of the noun in the WordNet hierarchy. The coordinating word captures some properties present in the noun phrase, properties that help to discriminate between various competing semantic relations. Example: “*Foreigners complain that they have limited access to [government procurement] in Japan.*” - the coordinating word is “access” which is a psychological feature.

6. Position of the nominalized noun depicts the position of the nominalizing verb in the compound; ie, either head or modifier. Example: “*working woman*”, where the nominalized noun is the modifier, and “*computer maker*” where the nominalized noun is the head noun.

7. In frame is a three-value feature indicating whether the compound has a paraphrase or if the peer in the compound is framed or not. If the peer in the NP noun-noun pair is in the corresponding VerbLeX predicate-argument frame, than the relation is captured in the predicate-argument structure. If it is not in the VerbLeX frame, but is an external argument (eg, LOCATION, TEMPORAL, MANNER, etc.), then it is no-frame. Otherwise, there is no paraphrase that keeps the meaning, so the relation is not defined by the predicate-argument frame. Example: “*computer maker*” is framed where as “*backyard composting*” is non-framed, and “*editing error*” is no-paraphrase (has no paraphrase of type verb-argument).

8. Relative pronoun/adverb applies only to adjective clauses and embeds information about the grammatical and/or semantic role of the head noun in the subordinate clause. Example: “*the room where the meeting took place*” - the word *where* implies location.

9. Grammatical role of relative pronoun/adverb applies only to adjective clauses and specifies the grammatical role of the relative pronoun/adverb, if one exists. This feature depicts better the grammatical role played in the sentence by the head noun. We used for this purpose an in-house rule-based grammatical role detection module, which annotates the following roles (cf. (Quirk et al.1985): *subject, direct object, indirect object, subject complement* (argument for copular verbs), *object complement* (second argument for complex transitive verbs), *object oblique, free predicative*, and approximates *extent* and *temporal semantic roles*. Example: “*the man who gave Mary the book*” - *Mary* and *the book* are indirect object and, respectively direct object, so *man* cannot be THEME or RECIPIENT.

10. Voice. This feature applies only to adjective clauses and indicates the voice of the verb in the relative clause. The voice plays an important role in the correlation between grammatical roles and semantic roles in a sentence. Example: “*the child that was taken to the zoo*” - passive voice, so the child is in a THEME relation with the verb *take*.

Let’s consider an example of nominalization with its features.

“Several candidates have withdrawn their names from consideration after administration officials asked them for their views on abortion and *fetal-tissue transplants*.”

The noun compound “*fetal-tissue#1 transplant#1*” is detected as a nominalization as the noun “*transplant*” is derived from the verb “*to transplant#3*”. The features and their values are: Feature 1: semantic class for fetal-tissue: *body-part*; Feature 2: verb class for *transplant*: *fill-9.8*; Feature 3: type of nominalization: *verb-nom*; Feature 4: gerundive: *no* (0); Feature 5: semantic class for coordinating word (“view”) = *psychological_feature#1*; Feature 6: position of the nominalized noun = *second*; Feature 7: in frame = *yes*.

The in-house extended verb lexicon *VerbLeX* shows the following semantic frame for the verb class *fill-9.8*: *Agent[+animate] Destination[+location -region] Theme[+concrete] Body-part is a subcategory of concrete*. Thus, for this example the semantic relation is THEME.

4 Overview of Results

The f-measure results obtained so far are summarized in Table 4. They are divided in two categories, nominalizations, and adjective clauses since the feature vectors differ from one category to another. We have compared the performance of SVM with three other learning algorithms: (1) semantic scattering (Moldovan et al. 2004), (2) decision trees (a C4.5 implementation), and (3) Naive Bayes. We considered as baseline *semantic scattering* which is a new learning model (Moldovan et al. 2004) developed in-house for the semantic classification of noun-noun pairs in NP constructions. The semantic relation derives from the WordNet semantic classes of the two nouns participating in those constructions, as well as the surrounding context provided by the WSD module.

As expected, the results vary from pattern to pattern. SVM and Naive Bayes seem to perform better than other models for the nominalizations and adjective clauses. Overall, these results are very encouraging given the complexity of the problem. By comparison with the baseline, the feature vector presented here gives better results.

Syntactic Pattern		Semantic Scattering (Baseline)	Decision Tree	Naive Bayes	Support Vector Machines
Complex Nominals	NN	48.88%	68.00%	70.00%	72.00%
	AdjN	51.08%	NA	NA	NA
Genitives	'S	57.24%	72%	70%	67%
	Of	33.74%	66.7%	68%	61%
Adjective Phrases		37.84%	58%	62%	64%
Adjective Clauses		NA	60%	62%	74%

Table 4: F-measure results for the semantic classification of NP patterns obtained with four learning models on a corpora with an 80/20 training/testing ratio. "NA" means not available.

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
Nominalized	H	M	M	M	L	H	H			
Adjective Clauses	M	H						H	L	H

Table 5: The impact of each feature f_i on the overall performance; H-high (over 8%), M-medium(between 2% and 8%), and L-low (below 2%). Empty boxes indicate the absence of features.

This explains in part our initial intuition that nominalization constructions at NP level have a different semantic behavior than the NP non-nominalization patterns.

We studied the influence of each feature on the performance, and since there are too many cases to discuss we only show in Table 5 the average impact as High, Medium, or Low. This table also shows the features used in each case.

References

C. Baker, C. Fillmore, and J. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLLING/ACL*.

D. Blaheta and E. Charniak. 2000. Assigning function tags to parsed text. In *Proceedings of the NAACL*.

E. Charniak. 2001. Immediate-head parsing for language models. In *Proceedings of ACL*, Toulouse, France.

N. Chomsky. 1970. Remarks on Nominalization. In *Readings in English Transformational Grammar*. Jacobs, R. A., and Rosenbaum, P.S. (eds), Ginn and Company.

B. Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. In *Machine Translation*, 12(4).

D. Gildea and D. Jurafsky. 2002. Automatic Labeling of Semantic Roles. In *Computational Linguistics*, 28(3).

D. Gildea and M. Palmer. 2002. The Necessity of Parsing for Predicate Argument Recognition. In *ACL*.

R. Grishman, C. Macleod, and A. Meyers. 1994. Comlex syntax: Building a computational lexicon. In *Proceedings of COLING*, Kyoto, Japan.

R. Hull and F. Gomez. 1996. Semantic interpretation of nominalizations. In *AAAI* conference, Oregon.

K. Kipper, H. T. Dang, and M. Palmer. 2000. Class-based construction of a verb lexicon. In *AAAI*, Austin, Texas.

P. Kingsbury, M. Palmer, and M. Marcus. 2002. Adding Semantic Annotation to the Penn TreeBank. In *Proceedings of HLT*, California.

P. Kingsbury and M. Palmer. 2002. From Treebank to Propbank. In *Third International Conference on Language Resources and Evaluation*, LREC-02, Las Palmas, Canary Islands.

Judith Levi. 1979. The Syntax and Semantics of Complex Nominals. New York: Academic Press.

C. Macleod, R. Grishman, A. Meyers, L. Barrett, and R. Reeves. 1998. Nomlex: A lexicon of nominalizations. In *Proceedings of the 8th International Congress of the European Association for Lexicography*, Belgium.

D. Moldovan, A. Badulescu, M. Tatu, D. Antohe, and R. Girju. 2004. Semantic Classification of Non-nominalized Noun Phrases. In *Proceedings of HLT/NAACL 2004 - Computational Lexical Semantics workshop*, Boston, MA.

S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J. Martin, and D. Jurafsky. 2003. Semantic role parsing: Adding semantic structure to unstructured text. In *ICDM*, Florida.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. A comprehensive grammar of English language, Longman, Harlow.

S. Siegel and N.J. Castellan. 1988. *Non Parametric Statistics for the behavioral science*. New York: McGraw-Hill.

M. Semmelmeyer and D. Bolander. 1992. *The New Webster's Grammar Guide*. Lexicon Publications, Inc.

S. Tong and D. Koller. 2001. Support Vector Machine Active Learning with Applications to Text Classification. In *Journal of Machine Learning Research*.

V. Vapnik. 1982. Estimation of Dependences Based on Empirical Data. In *Springer Verlag*.