Learning Noun-Modifier Semantic Relations with Corpus-based and *WordNet*-based Features

Vivi Nastase
Jelber Sayyad-Shirabad
School of Information Technology and Engineering
University of Ottawa 800 King Edward Avenue
Ottawa, Ontario K1N 6N5, Canada
{vnastase,jsayyad}@site.uottawa.ca

Marina Sokolova*
Département d' informatique et de recherche opérationnelle
Université de Montréal
2920, chemin de la Tour
Montréal, (Québec) H3T 1J4, Canada
sokolovm@iro.umontreal.ca

Stan Szpakowicz
School of Information Technology and Engineering
University of Ottawa 800 King Edward Avenue
Ottawa, Ontario K1N 6N5, Canada
Institute of Computer Science
Polish Academy of Sciences
Ordona 21, 01-237 Warszawa, Poland
szpak@site.uottawa.ca

Abstract

We study the performance of two representations of word meaning in learning noun-modifier semantic relations. One representation is based on lexical resources, in particular *WordNet*, the other – on a corpus. We experimented with decision trees, instance-based learning and Support Vector Machines. All these methods work well in this learning task. We report high precision, recall and F-score, and small variation in performance across several 10-fold cross-validation runs. The corpus-based method has the advantage of working with data without word-sense annotations and performs well over the baseline. The *WordNet*-based method, requiring word-sense annotated data, has higher precision.

Introduction

In understanding a text, it is essential to recognize relations among occurrences[1], entities and their attributes, represented at the surface as verbs, nouns and their modifiers. Semantic relations describe interactions between a noun and its modifiers (noun-modifier relations), a verb and its arguments (case relations/semantic roles), and two clauses. In the past few years we have seen the NLP community's renewed interest in analyzing semantic relations especially between verbs and their arguments (Baker, Fillmore, & Lowe 1998), (Kipper, Dang, & Palmer 2000), nouns and their modifiers (Rosario & Hearst 2001). Semantic role labelling competitions[2] also seem to increase the attractiveness of this topic.

In this paper we consider a specific supervised learning task: assign semantic relations to noun-modifier pairs in base noun phrases (base NPs), composed only of a noun and its modifier. To identify such noun-modifier relations we can rely only on semantic and morphological information about words themselves. For example, in the base NPs *iron gate*, *brick house*, *plastic container*: *iron*, *brick*, *plastic* are substances, and *gate*, *house*, *container* are artifacts; this suggests a MATERIAL relation in these pairs. On the other hand, analyzing case relation or clause-level relations is assisted by prepositions, subordinators, coordinators and maybe more elaborate syntactic structures.

*This work was done at the University of Ottawa.

[1] *Occurrence* encompasses all types of events, actions, activities, processes, states and accomplishments (Allen 1984)

[2] CONLL 2004, 2005 – Semantic Role Labelling shared task http://www.lsi.upc.edu/ srlconll/ .

We experiment with two methods of representing the words in a base NP, to be used in ML experiments for learning semantic relations between nouns and their modifiers. One method is based on features extracted from *WordNet*, which was designed to capture, describe, and relate word senses. In brief, we use hypernyms to describe in more and more general terms the sense of a word in a pair. The other method is based on contextual information extracted from corpora. Contexts are useful for determining word senses, as reflected in research on word-sense disambiguation using a window of words surrounding the target word (starting with (Yarowsky 1995), and more recently (Purandare & Pedersen 2004)). We use grammatical collocations[3] (as opposed to proximity-based co-occurrences) extracted from the British National Corpus, to describe each word in a pair. We compare the learning results produced by using these two types of word sense descriptions with the results obtained by (Turney & Littman 2003) and (Turney 2005), who used a paraphrase method to describe the pair as a whole.

The data we work with consist of noun-modifier pairs labeled with 30 fine-grained semantic relations, grouped into five relation classes. Experiments presented in this paper are based on the five-class coarse-grained grouping.

The corpus-based method gives precision, recall and F-scores well above the baseline, and it works on data without word-sense annotations. The method based on WordNet gives results with higher precision, but requires word-sense annotated data.

The paper consists of an overview of earlier research on learning noun-modifier semantic relations, a description of the data and the representations, a presentation of the experimental setup and the results, and finally a discussion and conclusions.

Related Work

We focus on methods that analyze and learn semantic relations in noun-phrases.

Levi (1978) analyzes the formation of nominal compounds. She identifies 9 *recoverable deletable predicates* (RDPs): *be*, *cause*, *have*, *make*, *use*, *about*, *for*, *from*, *in*, which, when erased from a more complex expression, generate a noun phrase. Levi writes that relations expressed by the RDPs may be universal, because from a semantic point of view they appear to be quite primitive. Different semantic relations may be associated with each RDP. For example:

[3] Grammatical collocations are collocates that appear with the target word in a grammatical relation, such as subject, object, prepositional complement (Kilgarriff *et al.* 2004).

cause – causes/is caused by; have – possession/possessor; make – physically producing/material.

Berland & Charniak (1999) and Hearst (1992) work with specific relations, *part of* and *type of* respectively. The Automatic Content Extraction project is a research program in information extraction that focuses on detecting specific relations (such as employer-organization, agent-artifact) between seven types of entities (such as person, organization, facility) in texts (Zhao & Grishman 2005). Several types of information – lexical, grammatical and contextual – are combined using kernel methods.

Vanderwende (1994) uses hand-crafted rules and a dictionary built from texts to find clues about the semantic relations in which a word may be involved. This was tested with 97 pairs extracted from the Brown corpus, with an accuracy of 52%.

Several systems use lexical resources (domain-specific like *MeSH* or general like *WordNet* or *Roget's Thesaurus*) to find the appropriate level of generalization for words in a pair, so that words linked by different relations are properly separated.

Rosario & Hearst (2001) learn noun-modifier semantic relations in a medical domain, using neural networks. The list of 13 relations is tailored to the application domain. Rosario, Hearst, & Fillmore (2002), continuing that research, look manually for rules which classify correctly noun compounds in the medical domain, based on the *MeSH* lexical hierarchy. The data are extracted automatically from biomedical journal articles, and sampled for manual analysis. *MeSH* is traversed top-down to find a level at which the noun compounds in different relations are properly separated.

Lauer (1995) maps words in noun compounds onto categories in *Roget's Thesaurus*, in order to find probabilities of occurrence of certain noun compounds and their paraphrases. There is no automatic process to find the best level of generalization. Nastase & Szpakowicz (2003) use the hypernym/hyponym structure of *WordNet*, and *Roget's Thesaurus*, to automatically find the generalization level in these resources that best describe each semantic relation. Several machine learning methods are used in analyzing 30 semantic relations. Girju *et al.* (2005) also use *WordNet* and the generalization/specialization of word senses, in the task of noun compound interpretation. Barker & Szpakowicz (1998) use a memory-based process to assign semantic relations to a new noun phrase, based on previously stored examples. The distance metric employs identity of one or both of the words, and the connective between them (usually a preposition).

Turney & Littman (2003) and Turney (2005) use paraphrases as features to analyze noun-modifier relations. Paraphrases express more overtly the semantic relation between a noun and its modifier. The hypothesis, corroborated by the reported experiments, is that pairs which share the same paraphrases belong to the same semantic relation.

Turney & Littman (2003) use a set of 64 *joining terms* which may appear between the two words in a noun phrase (*in the, at the, because, such that, ...*). For each head noun-modifier (*H-M*) pair in the dataset, and for each joining term J, a query to Alta Vista gave the frequency of the phrases HJM and MJH. The 128 frequency counts were grouped together with the associated semantic relation in a vector that described each noun-modifier pair, and then an ML experiment identified the joining terms that indicate a particular semantic relation, using a 2-nearest-neighbour algorithm.

This has been generalized in Turney (2005) using what he calls Latent Relational Analysis (LRA). For each word in a dataset of pairs, Lin's thesaurus (Lin 1998) gives a set of possible synonyms. All original pairs and pairs generated from synonyms are used to mine a corpus for paraphrases. All paraphrases are gathered and a few thousand of the most frequent ones are selected. The selected paraphrases, the original word pairs and the synonym pairs are used to build an incidence matrix, whose dimensionality is reduced using singular value decomposition (Landauer & Dumais 1997). Similarity between pairs combines scores for similarity between the original word pair and pairs built using synonyms.

Because we use the same data as Turney & Littman (2003) and Turney (2005), we compare the results of learning noun-modifier relations using *WordNet*-based and corpus-based representations, with the results obtained using paraphrase-based information.

Data and Representations

Lists of semantic relations in use range from general, as the lists in PropBank (Palmer, Gildea, & Kingsbury 2005) and NomBank (Myers *et al.* 2004), to more and more specific, as in VerbNet (Kipper, Dang, & Palmer 2000) and FrameNet (Baker, Fillmore, & Lowe 1998), to domain-specific (Rosario & Hearst 2001). The data we use consist of 600 noun-modifier pairs, tagged with 30 semantic relations, grouped into 5 classes of relations by general similarity (Barker 1998), (Nastase & Szpakowicz 2003), (Turney & Littman 2003):

1. CAUSAL groups relations enabling or opposing an occurrence. Examples (H denotes the head of a base NP, M denotes the modifier):

   cause - H causes M: *flu virus*;
   effect - H is the effect (was caused by) M: *exam anxiety*;
   purpose - H is for M: *concert hall*;

2. PARTICIPANT groups relations between an occurrence and its participants or circumstances. Examples:

   agent - M performs H: *student protest*;
   object - M is acted upon by H: *metal separator*;
   beneficiary - M benefits from H: *student discount*;

3. SPATIAL groups relations that place an occurrence at an absolute or relative point in space.Examples:

   direction - H is directed towards M: *outgoing mail*;
   location - H is the location of M: *home town*;
   location at - H is located at M: *desert storm*;

4. TEMPORAL groups relations that place an occurrence at an absolute or relative point in time. Examples:

   frequency - H occurs every time M occurs: *weekly game*;
   time at - H occurs when M occurs: *morning coffee*;
   time through - H existed while M existed: *2-hour trip*;

5. QUALITY groups the remaining relations between a verb or noun and its arguments. Examples:

   manner - H occurs as indicated by M: *stylish writing*;
   material - H is made of M: *brick house*;
   measure - M is a measure of H: *heavy rock*;

The words in the pairs from the dataset are also annotated with part of speech and *WordNet 1.6* word senses.

We describe two methods of representing these data. They are evaluated in learning experiments. One representation is based on word hypernym information extracted from *WordNet*. The second representation relies on grammatical collocation information extracted from a corpus.

WordNet-based representation

WordNet was designed to capture and describe word senses, and inter-connect them through a variety of lexical and semantic relations. We make use of the hypernym/hyponym links, to represent each head word and modifier in a pair through their hypernyms (ancestors) in *WordNet*. The data representation we build is used in a different manner by each ML tool we experiment with. Certain ML algorithms, such as decision trees, have a built-in feature selection process. The model of the data is built from a subset of features that is general enough to make good predictions on unseen data but specific enough to classify the training data as correctly as possible. Other ML algorithms, such as memory- or kernel-based, use all the features, possibly with different weights. A representation for the words in our dataset, based on information from *WordNet*, must meet the following constraints: it must be so general that the memory- and kernel-based methods, which do not perform feature selection, will be able to perform well on unseen data, and yet cover a wide enough range of levels to allow the decision tree method to build an accurate model.

WordNet's hypernym/hyponym structure is not uniform. Some domains are presented in greater detail, with a finer distinction in the hierarchy. Below a certain level, however, regardless of the domain represented, the synsets become quite specific and rather technical, and are not helpful in generalization. The maximum depth in *WordNet* reached by words in our data is 14. In an earlier research using this dataset (Nastase & Szpakowicz 2003) we observed that rule-based classifiers pick synsets at levels above 7. We therefore choose level 7 as the cut-off point. This serves as a form of feature selection, which provides more general features to the memory- and kernel-based systems, and enough generalization levels for the decision tree to find the ones that work best for the classes we learn. This choice is supported by high precision, recall and F-measure scores, reported in the "Experimental results" section.

We use a binary feature representation. To represent a word, using the word sense information in the data, we extract all ancestors located at the cut-off level and higher for the corresponding word sense. This produces 959 features to represent the head nouns, and 913 features for the modifiers, to which we add the part of speech. Each noun-modifier pair in the dataset is represented as a vector:

$$< s_{m1}, ..., s_{m913}, pos_m, s_{h1}, ..., s_{h959}, pos_h, relation >$$

s_{hx}, s_{mx} can be 1 or 0: this synset either does or does not appear as an ancestor of the head or the modifier, respectively. This representation will naturally address the problem of multiple inheritance in *WordNet*, since we can represent any number of ancestors of a node, just by setting the corresponding element to 1.

We attempt to connect adjective and adverb modifiers to the noun hierarchy using *pertains to* and *derived from* links. If this is possible, the representation of such a word will consist (mostly) of the representation of the noun synset to which it was linked. If such a connection cannot be made, the representation will be less informative, because the adjective and adverb synsets are not organized in a hierarchy as complex as the nouns'.

We also perform experiments using information from *WordNet* when word-sense information is ignored. In this case, a word's representation contains the ancetors at and above the cut-off level for all its possible senses. The purpose of these experiments is to measure the impact of knowing the sense of words in a pair for determining the semantic relation between them. The representation is similar to the one described above. The length of the vector representing a pair increases: there are now 1918 (hypernym) features to represent the head nouns, and 1741 for the modifier.

Corpus-based representation using grammatical collocations

Contexts provide strong and consistent clues to the sense of a word (Yarowsky 1993). If a corpus captures a large sample of language use, it allows us to describe the senses of a word through collocated words. Suppose that noun N, denoting entity E, is the subject of a sentence. Verbs that co-occur with N characterize occurrences in which E can participate, for example a *child* can *grow, eat, sleep, play, cry, laugh ....* Adjectives that modify N tell us about E's attributes, so for example a *child* can be *good, happy, sad, small, tall, chubby, playful, ...,* and so on.

We test such a context-based word-sense description for the task of learning noun-modifier relations. The Word Sketch Engine (WSE) (Kilgarriff *et al.* 2004) gives us collocation information organized by grammatical relations. It runs on a corpus – in our case, the British National Corpus – and extracts, for a given word, collocation information based and organized on grammatical categories. Thus, for a noun the engine builds a list containing: verbs with which the noun appears as a subject, verbs with which it appears as an object, the prepositional phrases attached to it (grouped by prepositions), the head nouns it modifies, its adjectival and nominal modifiers, and so on. The advantage of having such a resource is that it eliminates most, if not all, of the noise that we would encounter had we used a simple proximity-based process to gather co-occurrences – n-grams that are not proper phrases, are not connected to the words we consider, or do not span the entire phrase. Figure 1 shows a partial word sketch for the noun *cloud*[4].

cloud-n

and/or	*object_of*	*subject_of*	*a_modifier*	*pp_of-p*	...
mist-n	watch-v	scud-v	dark-j	smoke-n	
rain-n	swirl-v	drift-v	scudding-j	dust-n	
sky-n	billow-v	gather-v	black-j	steam-n	
cloud-n	form-v	hang-v	grey-j	ash-n	
...	...	...	...	...	

Figure 1: Sample (partial) word sketch for the noun *cloud* produced by the Word Sketch Engine

We produce a word sketch for each word in each noun-modifier pair in the data. From each word sketch we obtain a list of strings by concatenating each grammatical relation G_i with each word in this relation. For example, for the noun $cloud$, we will generate the list { *and/or_mist, and/or_rain, ..., object_of_watch, object_of_swirl, ...*} .

From the strings generated for all words that appear in our data, we keep the most frequent ones to obtain a binary feature set to represent each word. The corresponding value for a feature $G_i\_w_k$ will be 1 for word w if w appears in grammatical relation G_i with w_k in the corpus.

This feature construction and selection process produces a vector of 4969 *grammatical relation_word* strings. The final

[4]To simplify, we have omitted frequency and other statistical information that the WSE produces.

Rel. class	Examples	P_{TL}	R_{TL}	F_{TL}	P_{LRA}	R_{LRA}	F_{LRA}	P_{WN}	R_{WN}	F_{WN}	P_{WNas}	R_{WNas}	F_{WNas}	P_{WS}	R_{WS}	F_{WS}
causal	86 (14.3%)	21.2	24.4	22.68	38.8	38.4	**38.59**	**69.56**	18.6	29.35	52.63	11.76	19.23	17.37	**67.05**	27.60
participant	260 (43.3%)	55.3	51.9	53.54	**66**	67.3	**66.64**	52.16	88.41	65.61	47.16	**92.46**	62.46	59.01	28.57	38.5
quality	146 (24.3%)	45.4	47.3	46.33	54.2	**57.5**	**55.80**	**54.94**	34.48	42.37	50	20.42	29	46.42	27.46	34.51
spatial	56 (9.3%)	29.1	28.6	28.84	43.1	**39.3**	**41.11**	**85.71**	10.71	19.04	42.85	5.55	9.83	21.42	5.55	8.82
temporal	52 (8.7%)	66	63.5	64.72	77.3	**65.4**	70.85	**89.47**	**65.38**	**75.55**	80	8	14.54	88.57	62	72.94

Table 1: Precision, recall and f-score for memory-based learning experiments (2 nearest neighbour)

set of features that represents the noun-modifier pair has 2 * 4969 + 1 features; the added feature is the class (semantic relation) to which the pair belongs. We have chosen a binary representation rather than a representation which includes frequency information. We have two reasons: (i) the fact that two words appear together, connected by a grammatical relation, indicates that they are related; (ii) the number of co-occurrences is corpus-specific, so including frequency information may skew results. The results will show that this representation gives good learning results. Frequency information can be used to filter out noise (with the potential of deleting important, but infrequent, collocations) or, with a higher threshold, for feature selection.

One advantage of using grammatical collocations extracted from a corpus is that we do not need data annotated with word senses. On the other hand, the representation obtained will group together contextual information for all possible senses of a word. The empirical results show that, despite this, we can still find common characteristics among words involved in the same semantic relation. Having word-sense disambiguated associations may, however, lead to better results. We will test this hypothesis in future work.

Experiments

As we write in the "Related work" section, Turney & Littman (2003) and Turney (2005) applied the nearest neighbour method to the task of learning semantic relations on the same dataset that we use. They used the leave-one-out method to measure the performance of their predictions, and the class (semantic relation) of a test example is predicted based on its two nearest neighbours. Table 1 shows the reported results when using 64 joining terms (P_{TL}, R_{TL}, F_{TL}) and when using LRA ($P_{LRA}, R_{LRA}, F_{LRA}$). Here, *P, R* and *F-score* stand for precision, recall and F(1) measure (which gives the same weight to recall and precision). To compare these paraphrase-based representations with the corpus-based and *WordNet*-based ones, Table 1 includes the results obtained using *WordNet* with word sense information (P_{WN}, R_{WN}, F_{WN}) and without ($P_{WNas}, R_{WNas}, F_{WNas}$ – "WN all senses"), and Word Sketches (P_{WS}, R_{WS}, F_{WS}) in methodologically similar experiments – using an instance-based learner (TiMBL v. 5.1.0 (Daelemans *et al.* 2004)), with 2 nearest neighbour and leave-one-out testing.

The paraphrase-based representation which uses latent relational analysis performs better. LRA performs more poorly in terms of precision in 4 of 5 relation classes, with large differences in 3 of those 4 cases (30.76% for CAUSAL, 42.61% for SPATIAL, and 12.17% for TEMPORAL) compared to *WordNet* with word-sense information. On the other hand, its recall and F-score are higher than all the other representations.

The experiments that follow use a different methodology. Several 10-fold cross-validation runs verify that the learners have a consistent performance on different random data splits. The results of these experiments are not directly comparable with the ones in Table 1, because they are produced with different training-testing methods.

We apply memory-based learning (TiMBL v. 5.1.0 (Daelemans *et al.* 2004)), decision tree induction (C5.0 v.1.16 (Quinlan)) and Support Vector Machine (SVMlight v. 6.01 (Joachims)) to compare word representation methods, discussed above, for the task of learning noun-modifier semantic relations. To give more reliable results, we perform five runs. For each run we split the data into 10 random splits which preserve the class distribution of the original data set. We perform 10-fold cross-validation experiments on these splits with the three machine learning algorithms, adjusting the formatting of the files to fit each tool. These are binary experiments, in which examples of each relation class in turn become the positive instances, and the rest of the examples become the negative instances.

Preliminary runs found a configuration for each classifier. The results presented in this section were obtained with the following configurations: TiMBL uses the IGTREE classification algorithm (decision-tree-based optimization) and the χ^2 feature-weighing scheme; C5.0 runs with the default configuration; SVMlight uses the linear kernel.

Empirical Results and Discussion

Table 2 shows the results of learning the assignment of five classes of semantic relations in binary classification experiments for each relation class. The F-score baseline for each binary classification experiment combines (with equal weights) the precision when all examples are classified as positive – which is equal to the percentage of examples in the positive class – and the corresponding 100%recall. This baseline is independent of the learning method used, and is also higher for most classes than other baselines tried (based on a representation consisting of only the words in the pair, in both bag of words binary representation, and 2 multi-valued attribute representation).

The precision, recall and equally-weighted F-score results for each representation are averages over the five runs of 10-fold cross-validation experiments, plus-minus the standard deviation for each average. Because of class imbalance (averaging 1:5 for the five-class problem), accuracy is not as informative as precision, recall and F-score, and is not reported.

The data representation is very high-dimensional, as it often happens in NLP problems. Not all features have the same effect on the learning of noun-modifier relations. Using feature weighting schemes in TiMBL and C5.0's built-in feature selection gives better learning results than SVMlight in terms of F-score.

A low standard deviation indicates that the performance in the real world will be close to the estimated average. A combination of high precision and recall values and low standard deviation shows classes of relations that are learned well in a particular experimental configuration. This is the case for the TEMPORAL class learned using C5.0 or TiMBL with

Rel.class	Baseline	TiMBL			C5.0			SVM light		
	F-score	P	R	F-score	P	R	F-score	P	R	F-score
WordNet-based representation, with word sense information										
CAUSAL	25.43	52.96±3.98	**23.8**±2.62	**31.14**±2.96	56.82±10	14.33±2.13	21.8±3.25	**68.66**±9.68	17.79±0.85	27.48±1.30
PARTICIPANT	**60.35**	68.86±1.78	48.22±1.58	56.22±1.65	71.65±1.75	31.56±1.40	43.05±1.51	**69.06**±1.79	**50.31**±1.56	57.68±1.44
QUALITY	39.16	64.71±1.58	**30.24**±1.19	**40**±1.23	**68.14**±5.11	30.02±1.48	39.72±1.98	66.28±2.49	24.14±2.23	34.57±2.19
SPATIAL	16.95	**74.73**±6.18	**37.86**±2.39	**47.86**±2.44	67.43±17.03	28.73±5.28	37.72±6.77	66±8.13	25.52±1.51	35.73±1.71
TEMPORAL	15.78	92.63±1.33	76.4±3.44	82.47±2.55	**94**±0.45	**79.2**±1.6	**84.73**±0.92	77.46±8.49	52.15±1.31	60±1.44
WordNet-based representation, without word sense information										
CAUSAL	25.43	36.63±10.82	7.77±1.88	12.32±3.00	50.73±5.75	**16.24**±1.98	23.68±2.62	**62.40**±13.29	15.46±1.37	**24.07**±1.87
PARTICIPANT	**60.35**	67.61±1.79	23.02±0.97	33.78±1.21	**73.04**±1.92	27.73±1.37	39.74±1.55	63.72±2.60	**39.22**±1.24	47.88±1.75
QUALITY	**39.16**	**65.08**±4.45	**23.16**±1.16	33.16±1.56	53.79±3.95	18.39±0.95	26.74±1.17	62.83±6.55	15±1.55	23.75±2.27
SPATIAL	16.95	46.67±13.82	16.4±3.06	23.18±4.75	**54.93**±5.09	**18.73**±1.58	**26.59**±2.50	34.66±10.5	11.37±1.82	16.91±3.30
TEMPORAL	15.78	**97.10**±0.55	**57.6**±4.07	**70.38**±3.72	83.03±2.81	34±2.52	45.57±2.29	61.66±4.11	27.15±2.77	37.14±3.00
Word Sketch-based representation										
CAUSAL	**25.43**	**32.07**±3.48	**19.79**±3.12	23.58±3.32	27.96±3.74	18.5±1.34	21.12±1.30	**32.06**±12.56	7.83±3.41	12.43±5.32
PARTICIPANT	**60.35**	59.48±1.25	51.16±1.47	54.43±0.98	54.02±1.19	**53.01**±1.95	53.18±1.30	**71.12**±1.56	42.23±0.64	52.47±0.65
QUALITY	39.16	53.52±3.69	39.28±1.28	**44.67**±1.67	43.43±1.05	**44.33**±2.54	43.08±1.31	**56.53**±4.05	19.81±0.97	28.67±1.42
SPATIAL	16.95	**43**±5.33	26.08±5.23	**30.57**±5.07	32.49±6.10	**30.46**±3.66	29.8±4.29	37±8.05	13.68±1.32	19.7±2.57
TEMPORAL	15.78	**81.9**±3.17	**73.5**±5.89	**75.62**±3.31	77.72±4.28	69.6±1.49	70.2±2.16	80.03±6.33	36.81±2.83	48.3±2.44

Table 2: Learning results for TiMBL, C5.0, SVMlight for *WordNet*-based and *Word Sketch*-based representations

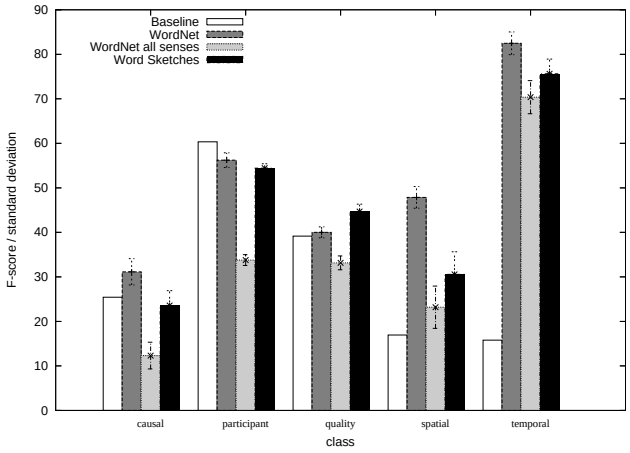

Figure 2: F-scores for learning with TiMBL

both the *WordNet*-based and the *Word Sketch*-based representations. In situations where the standard deviation is high, we cannot make confident predictions about future performance.

Figure 2 plots the F-scores for the word representation explored, when TiMBL was the learning method. We plot TiMBL's results, because – according to the standard deviation – it was the most stable learner. We observe that the performance of the representation based on Word Sketches, which does not distinguish word senses, performs better than *WordNet* without word sense information. It is also close to *WordNet* with sense information.

The advantage of using corpora is that no knowledge-intensive preprocessing is necessary, and the method does not rely on other lexical resources. The process may therefore be ported to other languages. In order to use paraphrases effectively, a larger corpus is needed so sufficiently many paraphrases can be found. The same is true of building descriptions of word meaning based on grammatical collocations in a corpus: the larger the corpus, the higher the chances that we find the most informative collocations. Here are some collocation features picked by C5.0 during the learning phase: *happen Modifier, occur during Modifier, wait until Modifier* [5] indicate a TEMPORAL relation; *predict Head-noun, Head-noun and/or fear*[6] indicate a CAUSAL relation.

We observe the impact of having word-sense information when we compare the results of learning experiments with the *WordNet*-based representation with and without word-sense annotation. The difference in results is quite dramatic. The F-scores drop for all relation classes and all ML methods used. Moreover, the difference in results when using Word Sketches and when using non-annotated data – in favour of Word Sketches – indicate that when no word-sense information is available, corpus-based word descriptions are more informative and useful for the task of learning semantic relations. The interesting exceptions are the recall for the PARTICIPANT class in 2-nearest neighbour experiments – 92.46% compared to the next best one of 88.41% – and the precision for the TEMPORAL class in cross-validation runs with TiMBL – 97.1%, compared to the 92.63% precision when word senses are used. The fact that an increase in precision is accompanied by a sharp drop in recall (from 76.4 to 57.6) means that the learner reduces the number of examples incorrectly assigned to the temporal class, but at the same time more temporal examples are assigned to an incorrect class. The effect of including all word hypernyms is that it introduces ambiguity between previously well separated words (when sense information was used) through shared hypernyms that do not pertain to the word sense in the pair. This causes more of the pairs to become ambiguous from the semantic relation point of view, and these will be misclassified. The pairs with stronger commonalities or non-ambiguous hypernyms will be fewer, but will be classified better. A reverse effect explains the increase in recall for PARTICIPANT , accompanied by a drop in precision (from 52.16% to 47.16%) – when more examples of the class are caught, but are classified less correctly. PARTICIPANT con-

[5] *object_of happen-v, pp_obj_during-p occur-v, pp_obj_until-p wait-v*

[6] *object_of predict-v, and/or fear-n*

tains the most instances, 43.22% of the dataset. Previously discriminating hypernyms will now cover a more heterogeneous mixture of instances.

Using *WordNet* with word-sense information gives very high results – 82.47% F-score – especially in terms of precision – 92.63%. This shows that indeed there are inherited characteristics of word senses which determine the semantic relations in which these words are involved. Here are some features chosen by the decision tree method: {*clock time, time*}, {*measure, quantity, amount, quantum*} for the modifier indicate a TEMPORAL relation; {*ill health, unhealthiness, health problem*} for the modifier indicate a CAUSAL relation; {*causal agent, cause, causal agency*} for the head indicate a PARTICIPANT relation. The fact that recall is lower may suggest that some word senses could not be connected, probably because what they share cannot be captured by the hypernym/hyponym relation. The word representation can be extended to make use of other relations in *WordNet*, such as meronym/holonym.

Conclusions and Future Work

We have compared different methods of representing data for learning to identify semantic relations between nouns and modifiers in base noun phrases.

Looking at the results obtained with the different representation methods, we can conclude that we can detect successfully the TEMPORAL relation between words by looking at either of the following: individual word senses as described by *WordNet*, word meaning as described by its contexts, or the prepositions or paraphrases that connect the words in the pair. For the other four relation classes, describing a word using sense specific *WordNet* information allows for high precision in identifying the correct relation class, but in order to increase the number of relation instances recognized, using corpus-based features helps. When no word-sense information is available, corpora-based features will lead to better results than using all word senses in *WordNet*.

As we said previously, using the word meaning representation methods described generates very high dimensional data. While we do obtain results well above the baseline, it is quite likely that the ML tools are overwhelmed by the large number of attributes. We will experiment with different feature selection methods to find a small set of word meaning descriptors that may produce even better results.

Because we use sets of features from different sources, which achieve high precision on different classes, we could use co-training to bootstrap the automatic tagging of a new set of pairs (Balcan, Blum, & Yang 2005). This would allow us to incrementally increase a starting (small) dataset with examples classified at high precision. Obtaining a larger dataset would help address the problem of data sparseness.

References

Allen, J. F. 1984. Towards a general theory of action and time. *Artifi cial Intelligence* 23(2):123–154.

Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The Berkeley FrameNet project. In *COLING-ACL 1998*, 86–90.

Balcan, M.-F.; Blum, A.; and Yang, K. 2005. Co-training and expansion: Towards bridging theory and practice. In Saul, L. K.; Weiss, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press. 89–96.

Barker, K., and Szpakowicz, S. 1998. Semi-automatic recognition of noun-modifier relationships. In *Proc. of COLING-ACL ' 98*, 96–102.

Barker, K. 1998. *Semi-Automatic Recognition of Semantic Relationships in English Technical Texts*. Ph.D. Dissertation, University of Ottawa, Department of Computer Science. http://www.cs.utexas.edu/users/kbarker/thesis.

Berland, M., and Charniak, E. 1999. Finding parts in very large corpora. In *Proc. of ACL 1999*, 57–67.

Daelemans, W.; Zavrel, J.; van der Sloot, K.; and van den Bosch, A. 2004. TiMBL: Tilburg Memory Based Learner, version 5.1. Reference Guide. ILK Technical Report 04-02, Available from http://ilk.uvt.nl/downloads/pub/papers/ilk0402.pdf.

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

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of CoLing 1992*, 539–545.

Joachims, T. SVM light. http://svmlight.joachims.org.

Kilgarriff, A.; Rychly, P.; Smrz, P.; and Tugwell, D. 2004. The sketch engine. In *Proc. of EURALEX 2004*, 105–116.

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

Landauer, T. K., and Dumais, S. 1997. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction and representation of knowledge. *Psychological Review* (104):211–240.

Lauer, M. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. Dissertation, Macquarie University, Australia.

Levi, J. 1978. *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proc. of COLING-ACL ' 98*, 768–774.

Myers, A.; Reeves, R.; Macleod, C.; Szekely, R.; Zielinska, V.; Young, B.; and Grishman, R. 2004. The NomBank project: An interim report. In *HLT-EACL Workshop: Frontiers in corpus annotation*, 24–31.

Nastase, V., and Szpakowicz, S. 2003. Exploring noun-modifier semantic relations. In *Proc. of IWCS 2003*, 281–301.

Palmer, M.; Gildea, D.; and Kingsbury, P. 2005. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–106.

Purandare, A., and Pedersen, T. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proc. of CONLL 2004*, 41–48.

Quinlan, R. C5.0. http://www.rulequest.com.

Rosario, B., and Hearst, M. 2001. Classifying the semantic relations in noun-compounds via a domain specific hierarchy. In *Proc. of EMNLP 2001*, 82–90.

Rosario, B.; Hearst, M.; and Fillmore, C. 2002. The descent of hierarchy, and selection in relational semantics. In *Proc. of ACL 2002*, 417–424.

Turney, P., and Littman, M. 2003. Learning analogies and semantic relations. Technical Report Technical Report ERB-1103. (NRC #46488), National Research Council, Institute for Information Technology.

Turney, P. 2005. Measuring semantic similarity by latent relational analysis. In *Proc. of IJCAI 2005*, 1136–1141.

Vanderwende, L. 1994. Algorithm for automatic interpretation of noun sequences. In *Proc. of ACL 1994*, 782–788.

Yarowsky, D. 1993. One sense per collocation. In *ARPA Human Language Technology Workshop*.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. ACL 1995*, 189–196.

Zhao, S., and Grishman, R. 2005. Extracting relations with integrated information using kernel methods. In *Proc. of ACL 2005*, 419–426.