Experiments with SemScat2 in the Task of Semantic Relation Extraction

Alla Rozovskaya and **Brandon Beamer** and **Roxana Girju**
Department of Linguistics and
Beckman Institute
University of Illinois at Urbana-Champaign
{rozovska, bbeamer, girju}@uiuc.edu

Abstract

This paper addresses the task of automatic classification of semantic relations between nouns. An evaluation of an improved WordNet-based learning model is presented. The experiments are performed using data from the 2007 SemEval-1 Task 4 competition and other sources. We report substantial improvements over previous WordNet-based methods on the SemEval-1 data. The experiments also show that WordNet's IS-A hierarchy is better suited for some semantic relations compared with others. Finally, learning curves show that the task difficulty varies across relations and that adding more training data from other sources does not increase the performance.

Introduction

The identification of semantic relations is at the core of Natural Language Processing (NLP) and many of its applications. Detecting semantic relations between various text segments, such as phrases, sentences, and discourse spans is important for automatic text understanding. Furthermore, semantic relations represent the core elements in the organization of lexical semantic knowledge bases intended for inference purposes. In the past few years at many workshops, tutorials, and competitions this research topic has received considerable interest from the NLP community.

Semantic relation identification is the problem of recognizing, for example, the CAUSE-EFFECT (cycling, happiness) relation in the sentence *He derives great joy and happiness from cycling*. This task requires several local and global decisions needed for relation identification. This involves the meaning of the two noun entities along with the meaning of other words in context.

In this paper we present an evaluation of efficient WordNet-based learning model that identifies and extracts noun features from the WordNet IS-A backbone, which was designed to capture and relate noun senses. The basic idea is that noun - noun pairs which have the same or similar sense collocation tend to encode the same semantic relation. We perform various experiments on the SemEval-1 dataset and compare the results against another state-of-the-art WordNet-based algorithm (Moldovan & Badulescu 2005) and the top-ranked systems in SemEval-1(Girju *et al.* 2007).

The results show that our WordNet-based semantic relation model places 5th with respect to the top-ranked systems in the SemEval-1 competition. We believe this is a significant result considering that our model uses only the WordNet noun IS-A hierarchy. Moreover, we also compute the learning curves for each relation and show that this model does not need a lot of training data to learn the classification function.

The paper is organized as follows. In the next section we present previous work, followed by the description of the SemEval task and datasets. We then give a brief overview of our model. Finally, we present various experiments and discuss the results.

Previous Work

Most of the attempts in the area of noun - noun semantic interpretation have studied the problem in different limited syntactic contexts, such as noun–noun compounds and other noun phrases (e.g., "N preposition N", "N, such as N"), and "N verb N". Recent work in this area follows roughly two main approaches: interpretation based on semantic similarity with previous seen examples (Rosario & Hearst 2001), (Moldovan *et al.* 2004), (Nastase *et al.* 2006), and semantic disambiguation relative to an underlying predicate or semantically-unambiguous paraphrase (Lapata 2002), (Kim & Baldwin 2006).

Most methods employ rich ontologies and disregard the sentence context in which the nouns occur, partly due to the lack of annotated contextual data on which they are trained and tested, and partly due to the claim that axioms and ontological distinctions are more important than the information derived from the context in which the nouns occur. In this paper, our experimental results also support this claim. However, we show that some semantic relations are better suited for WordNet-based models than others, and that contextual data are important in the performance of a noun - noun semantic parser.

Furthermore, most approaches use supervised learning employing fairly large feature sets on fairly small datasets. For example, (Nastase *et al.* 2006) train their system on 600 noun–modifier pairs classified into six high-level semantic relations (Cause, Participant, Spatial, Temporal, Quality). The problem is that these relations are not uniformly distributed in the dataset. Moreover, most of the SemEval

participant systems[1] were trained on an average of 140 examples per relation. This raises questions about the effectiveness of the algorithms and their capability to generalize over seen examples in order to efficiently classify unseen instances. Additionally, it is difficult to talk about the learning curve of each semantic relation, and thus, impossible to draw valid conclusions about the difficulty of the task across different relations.

In this paper we train our interpretation model on various large annotated datasets and present observations on the learning curves generated for each relation.

Classification of Semantic Relations between Nominals

The SemEval-1 task on Semantic Relations between Nominals is to identify the underlying semantic relation between two nouns in the context of a sentence. The SemEval effort focuses on seven separate semantic relations: *Cause-Effect, Instrument-Agency, Product-Producer, Origin-Entity, Theme-Tool*, *Part-Whole*, and *Content-Container*. The dataset provided consists of 140 training and about 70 test sentences for each of the seven relations considered.

In each training and test example sentence, the nouns are identified and manually labeled with their corresponding WordNet 3.0 senses. Moreover, each example is accompanied by the heuristic pattern (query) the annotators used to extract the sentence from the web and the position of the arguments in the relation. Positive and negative examples of the Cause–Effect relation are listed in (1) and (2) below. Cause–Effect relations are semantically similar to other relations such as Temporal, Source, Origin–Entity, and Product–Producer. Instances encoding these relations are called near-miss examples, as shown in (2).

(1) "He derives great joy and <e$_1$>happiness</e_1> from <e_2>cycling</e_2>." WordNet(e_1) = "happiness%1:12:00::" [2], WordNet(e_2) = "cycling%1:04: 00::", Cause-Effect(e_2,e_1) = "true", Query = "happiness from *"

(2) "Women may experience <e$_1$>anxiety</e_1> from the <e_2>time</e_2> they first learn about the breast abnormality." WordNet(e1) = "anxiety%1:12:00::", WordNet(e2) = "time%1:11:00::", Cause-Eff ect(e2,e1) = "false"; Query = "anxiety from *"

The task is defined as a binary classification problem. Thus, given a pair of nouns and their sentential context, a semantic interpretation system decides whether the nouns are linked by the target semantic relation. Based on the information employed, systems can be classified in four types of classes: (A) systems that use neither the given WordNet synsets nor the queries, (B) systems that use only WordNet senses, (C) systems that use only the queries, and (D) systems that use both.

In this paper we present an evaluation of a knowledge-intensive semantic interpretation system of type-B (Beamer, Rozovskaya, & Girju 2008). The system relies on WordNet semantic features employed in a supervised learning model.

[1]The exception here is UIUC's system which was trained on external data sets.

[2]The numbers after a noun refer to WordNet synset sense keys.

The Model

The learning model we use is described in detail in (Beamer, Rozovskaya, & Girju 2008) and is a significant improvement over Semantic Scattering (Moldovan & Badulescu 2005). The main idea of the Semantic Scattering model is to find the best set of noun semantic classes that would separate the positive and the negative examples and that would accurately classify unseen instances. This is done by finding a boundary (a division in the WordNet noun hierarchy) that would best generalize over the training examples. The distinguishing feature of the present model is a much improved boundary detection algorithm. In particular, we introduce a more efficient probability function for better boundary specialization, and thus for better semantic interpretation of unseen examples.

Experiments

We describe four experiments. The first three experiments are performed on the SemEval-1 dataset and focus on the behavior of our model from several perspectives. Specifically, we show how our model compares against Semantic Scattering and the top-ranked systems in the SemEval-1 competition. Furthermore, we distinguish between two types of data instances, depending on the type of information required for the identification of a relation between two nouns. We compare the performance on both types of data and discuss the effectiveness of our model in this context. Finally, the experiment IV shows for each relation the learning curves as generated by our model and Semantic Scattering. In what follows, we refer to Semantic Scattering as SemScat1 and to our algorithm as SemScat2.

Experiment I

Experiment I evaluates SemScat2 models with respect to the SemEval-1 test data. Table 1 presents accuracy results of SemScat1 and SemScat2, where a model is trained for each relation on the training data from SemEval-1 and tested on the corresponding SemEval-1 test set. SemScat2 outperforms SemScat1 on all relations, except Theme-Tool, with an absolute increase of 6% on average.

Table 2 shows that SemScat2 places fifth among the 14 systems in the SemEval-1 competition. This is despite the fact that it does not make use of sentence context, making a prediction using the noun–noun pair only.

Experiment II

As in Experiment I, only the SemEval-1 data are used. However, for each relation, the test and training sets are lumped together and 10-fold cross-validation is performed, yielding a prediction for each example.

Table 3 compares the performance of SemScat1[3] and

[3]It should be noted that since the choice of development set for SemScat1 is crucial for the performance, the performance of SemScat1 can change dramatically on the same test set with the same training set due to a choice of the development set.

Relation	SemScat1 [% Acc.]	SemScat2 [% Acc.]
Cause-Effect	60.8	73.0
Instrument-Agency	50.7	70.0
Product-Producer	65.5	67.9
Origin-Entity	59.7	63.6
Theme-Tool	63.6	54.5
Part-Whole	63.4	70.4
Content-Container	60.8	67.6
Average	**60.7**	**66.8**

Table 1: Experiment I results: Models are trained on SemEval-1 training data and tested on SemEval-1 test data. Acc. means "Accuracy".

System	F [%]	Acc [%]
UIUC	72.4	76.3
FBK-IRST	71.8	72.9
ILK	71.5	73.2
UCD-S1	66.8	71.4
SemScat2	**65.8**	**66.8**

Table 2: Experiment I results: Comparison of SemScat2 with top-ranked B-systems of the SemEval-1 competition. F and Acc mean "F1" and "Accuracy" respectively.

SemScat2 with respect to the accuracy of each algorithm on each relation. The results show that SemScat2 significantly outperforms SemScat1 on Cause-Effect, Instrument-Agency, Origin-Entity, Content-Container. On the remaining three relations, the two algorithms exhibit comparable results. Overall, SemScat2 outperforms SemScat1 by 6%.

Moreover, for each relation, we split the lumped training and test examples into *regular* and *context-sensitive*. *Regular* examples are those where the relation between the two given nouns can be determined out of context. *Context-sensitive* examples are those in which sentence context is required for their correct interpretation. This split was performed manually by an annotator based on his judgment. Consider, for example, the following sentences with respect to the "Cause-Effect" relation. In (3) and (4) we can say with high confidence that "Cause-Effect" relation is "True". By contrast, in (5) and (6) it is the sentence context that determines the answer. We consider example (3) regular and (4) context-sensitive.

(3) <e_1>*tumor shrinkage*</e_1> <e_2>*radiation therapy*</> Cause-Effect(e_2,e_1)="true" "The period of <e_1>tumor shrinkage</e_1> after <e_2>radiation therapy</e_2> is often long and varied".

(4) <e_1>*activation*</e_1> <e_2> *summer*</e_2> Cause-Effect(e_2,e_1)="false" "One newly formed Iraqi battalion is on duty, with scheduled for <e_1>activation</e_1> by <e_2>summer</e_2> 2004."

(5) <e_1>*nausea*</e_1> <e_2>*abnormal sensations*</e_2> Cause-Effect(e_2,e_1)="true" "It is likely that the <e_1>nausea</e_1> comes from <e_2>abnormal sensations</e_2> originating from areas of the brain that are sensing the lack of oxygen."

(6) "<e_1>*anxiety*</e_1> <e_2>*exam*</e_2> Cause-Effect(e_2,e_1)="false" "The following are very basic tips which may help you manage your <e_1>anxiety</e_1> in the <e_2>exam</e_2>." Comment: Time; the context does not imply that the exam is the cause for the anxiety.

Each relation contains between 26 and 60 context-sensitive examples. Table 4 compares the performance in accuracy of SemScat2 in 10-fold cross-validation on regular (column 2) and context-sensitive (column 3) examples. In parentheses, we list the performance on positive and negative examples within each group[4].

Relation	SemScat1 [% Acc.]	SemScat2 [% Acc.]
Cause-Effect	57	69
Instrument-Agency	61	70
Product-Producer	64	63
Origin-Entity	61	68
Theme-Tool	62	63
Part-Whole	72	71
Content-Container	58	75
Average	**62**	**68**

Table 3: Experiment II results: 10-fold cross-validation on SemEval-1 data. Columns 2 and 3 show accuracy for each relation of SemScat1 and SemScat2 respectively. Acc. means "Accuracy".

Relation	Examples (pos.; neg.)	
	Regular [% Acc.]	Context-sensitive [% Acc.]
Cause-Effect	71 (79; 63)	63 (71; 50)
Instrument-Agency	78 (82; 74)	37 (42; 32)
Product-Producer	61 (80; 34)	71 (78; 33)
Origin-Entity	70 (71; 69)	58 (56; 67)
Theme-Tool	66 (54; 72)	48 (42; 58)
Part-Whole	75 (82; 70)	42 (54; 31)
Content-Container	77 (85; 70)	63 (57; 69)
Average	**71** (75; 65)	**55** (57; 49)

Table 4: Experiment II results on regular and context-sensitive SemEval-1 examples. Columns 2 and 3 show accuracy for each relation of SemScat2 on regular and context-sensitive examples, respectively.

We observe that consistently across all relations, accuracy on both positive and negative examples is better in the regular group than in the corresponding context-sensitive group. Overall, performance on regular examples is considerably higher for all relations, with an average accuracy of 71% for regular examples and 55% for context-sensitive examples. And while for Product-Producer the numbers for the context-sensitive group are higher than for the regular group,

[4]Positive and negative examples are those labeled as "True" and as "False", respectively in Gold Standard.

the results for Product-Producer are compatible with the results for the other relations. This is because the proportion of negative examples within the context-sensitive group for this relation is much lower than the proportion of negative examples within the regular group, and the accuracy on negative context-sensitive examples is much lower than on positive examples from that group.

Separating regular examples allows us also to see which relations are best captured with the WordNet hierarchy. In particular, the results on regular examples in Table 4 demonstrate that the best-processed relations are Instrument-Agency, Part-Whole, and Content-Container, while the poorest is Product-Producer.

Experiment III

Experiment III is concerned with the role of context-sensitive examples in training. First, we wish to determine how a model trained on regular examples performs on context-sensitive examples. Table 5, column 2 shows the performance of SemScat2 trained on regular examples and tested on context-sensitive examples (as in Experiment II). This is compared with the performance obtained when the model is trained and tested on both regular and context-sensitive examples with 10-fold cross-validation (mixed model)in column 3. We note that models in experiment IV do not perform as well as the models from experiment II. Since the training sizes are roughly the same in both experiments, the main difference between the models is the absence of context-sensitive examples in the models of experiment IV. It can be conjectured that the presence of context-sensitive examples in training is beneficial, when the test set is composed of such examples.

Furthermore, we wish to determine whether the presence of context-sensitive examples in training is detrimental, when the test set consists only of regular examples. Table 5 compares the results of experiment II (10-fold cross-validation on all the data) (column 5) with 10-fold cross-validation on regular examples only (column 4) [5]. The test sets contain only regular examples in both cases. We observe that there is no significant difference in performance. When context-sensitive examples are removed from training, the performance improves for relations Instrument-Agency, Product-Producer, and Theme-Tool, and decreases for Cause-Effect and Content-Container.

Experiment IV

In this experiment, we determine the learning curve for each relation. Because the SemEval-1 training data contain only 140 examples per relation, which is not sufficient to obtain an accurate learning curve, we use additional datasets [6]. The models are tested on SemEval-1 test data and trained on all other data available. Since the number of examples varies considerably from one relation to another, we group the relations into classes based on the size of their data sets. Thus, class I (approx. 130 examples per relation) contains {Content-Container, Instrument-Agency, Theme-Tool} and class II (approx. 1,000 examples) has {Origin-Entity, Cause-Effect, Product-Producer, Part-Whole}.

Figures 1 and 2 show the learning curves for each class. Each figure displays the results of both SemScatt1 and SemScatt2. The models are tested on SemEval-1 test data and trained on all other data available. SemScat2's performance is plotted with thick lines.

There are several observations to be made. First, at each level of training, our system either outperforms Semantic Scattering by a significant margin, or performs similarly to it. Second, we note the stability of our system when compared to Semantic Scattering. More specifically, our system's performance reaches the saturation point much faster than Semantic Scattering. This unpredictable behavior is a result of the randomly selected development set, which Semantic Scattering uses to judge the performance of its boundary during training. Finally, it should be noted that the learning curves look quite flat. The low saturation point may be an indication that our system needs a relatively small amount of training data to achieve maximum performance.

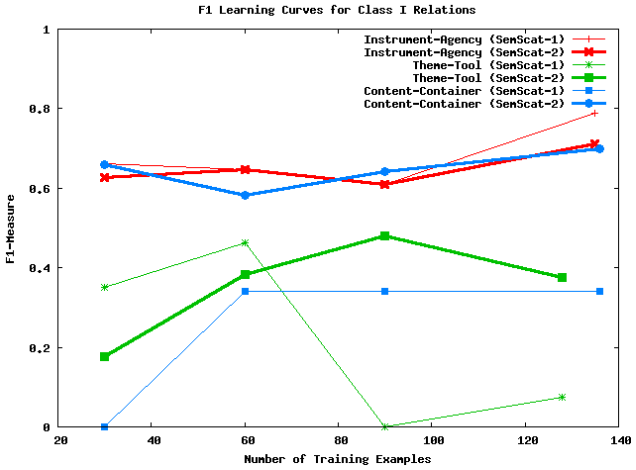

Figure 1: The learning curves for Class I relations.

Conclusions

We have presented an evaluation of a learning model that identifies and extracts noun features efficiently from WordNet's IS-A backbone. Our experiments provide insight into the problem of the identification of semantic relations between nominals. More specifically, we have shown that our model is superior to Semantic Scattering (Moldovan & Badulescu 2005) both in terms of performance accuracy and system stability. Moreover, our system places 5th with respect to the top-ranked B-systems in the SemEval-1 competition. We believe this is an important result, given the information that the model requires. We have also shown that WordNet structure is capable of capturing some relations better than others. Additionally, we have made a distinction between regular examples and those that require sentence context for the relation identification. Our system performs much better

[5] Note that the training sets are slightly smaller when context-sensitive examples are removed from the data

[6] The datasets are described in (Beamer *et al.* 2007)

Relation	Regular models#1 [% Acc.]	Mixed models#1 [% Acc.]	Regular models#2 [% Acc.]	Mixed models#2 [% Acc.]
Cause-Effect	56	63	68	71
Instrument-Agency	28	37	82	78
Product-Producer	52	71	70	61
Origin-Entity	36	58	71	70
Theme-Tool	45	48	67	66
Part-Whole	38	42	79	75
Content-Container	56	63	74	77
Average	**44**	**55**	**73**	**71**

Table 5: Comparison of models trained on regular examples and mixed data, when tested on context-sensitive examples (columns 2 and 3) and regular examples (columns 4 and 5).

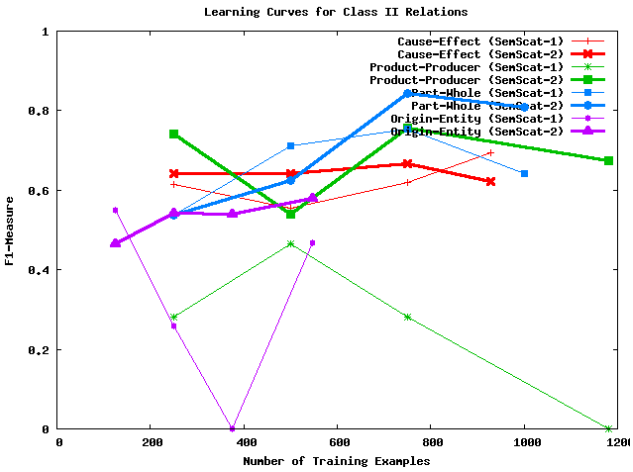

Figure 2: The learning curves for Class I relations.

on regular examples, as expected. Finally, learning curves show that the task difficulty varies across relations and that adding more training data from other sources does not increase the performance.

References

Beamer, B.; Bhat, S.; Chee, B.; Fister, A.; Rozovskaya, A.; and Girju, R. 2007. UIUC: A Knowledge-rich Approach to Identifying Semantic Relations between Nominals. In *The 4th ACL Workshop on Semantic Evaluations*.

Beamer, B.; Rozovskaya, A.; and Girju, R. 2008. Automatic Semantic Relation Extraction with Multiple Boundary Generation. In *AAAI*.

Girju, R.; Nakov, P.; Nastase, V.; Szpakowicz, S.; Turney, P.; and Yuret, D. 2007. Semeval-2007 Task 04: Classification of Semantic Relations between Nominals. In *The 4th ACL International Workshop on Semantic Evaluations*.

Kim, S. N., and Baldwin, T. 2006. Interpreting Semantic Relations in Noun Compounds via Verb Semantics. In *International Computational Linguistics Coference (COLING)*.

Lapata, M. 2002. The Disambiguation of Nominalizations. *Computational Linguistics* 28(3):357–388.

Moldovan, D., and Badulescu, A. 2005. A Semantic Scattering Model for the Automatic Interpretation of Genitives. In *The Human Language Technology Conference (HLT)*.

Moldovan, D.; Badulescu, A.; Tatu, M.; Antohe, D.; and Girju, R. 2004. Models for the Semantic Classification of Noun Phrases. In *The HLT/NAACL workshop on Computational Lexical Semantics*.

Nastase, V.; Shirabad, J. S.; Sokolova, M.; and Szpakowicz, S. 2006. Learning Noun-Modifier Semantic Relations with Corpus-based and WordNet-based Features. In *The 21st National Conference on Artificial Intelligence (AAAI)*.

Rosario, B., and Hearst, M. 2001. Classifying the semantic relations in noun compounds. In *Conference on Empirical Methods in Natural Language Processing*, 82–90.