JU: A Supervised Approach to Identify Semantic Relations from Paired Nominals

Santanu Pal **Partha Pakray** **Dipankar Das** **Sivaji Bandyopadhyay**
Department of Computer Science & Engineering, Jadavpur University, Kolkata, India
santanupersonal1@gmail.com, parthapakray@gmail.com,
dipankar.dipnil2005@gmail.com, sivaji_cse_ju@yahoo.com

Abstract

This article presents the experiments carried out at Jadavpur University as part of the participation in Multi-Way Classification of Semantic Relations between Pairs of Nominals in the SemEval 2010 exercise. Separate rules for each type of the relations are identified in the baseline model based on the verbs and prepositions present in the segment between each pair of nominals. Inclusion of WordNet features associated with the paired nominals play an important role in distinguishing the relations from each other. The Conditional Random Field (CRF) based machine-learning framework is adopted for classifying the pair of nominals. Application of dependency relations, Named Entities (NE) and various types of WordNet features along with several combinations of these features help to improve the performance of the system. Error analysis suggests that the performance can be improved by applying suitable strategies to differentiate each paired nominal in an already identified relation. Evaluation result gives an overall macro-averaged F1 score of 52.16%.

1 Introduction

Semantic Relations describe the relations between concepts or meanings that are crucial but hard to identify. The present shared task aims to develop the systems for automatically recognizing semantic relations between pairs of nominals. Nine relations such as Cause-Effect, Instrument-Agency, Product-Producer, Content-Container, Entity-Origin, Entity-Destination, Component-Whole, Member-Collection and Message-Topic are given for SemEval-2010 Task #8 (Hendrix *et al.*, 2010). The relation that does not belong to any of the nine relations is tagged as *Other*. The first five relations also featured in the previous SemEval-2007 Task #4.

The present paper describes the approach of identifying semantic relations between pair of nominals. The baseline system is developed based on the verbs and prepositions present in the sentential segment between the two nominals. Some WordNet (Miller, 1990) features are also used in the baseline for extracting the relation specific attributes (e.g. *Content* type *hypernym* feature used for extracting the relation of *Content-Container*). The performance of the baseline system is limited due to the consideration of only the verb and preposition words in between the two nominals along with a small set of WordNet features. Hence, the Conditional Random Field (CRF) (McCallum *et al.*, 2001) based framework is considered to accomplish the present task. The incorporation of different lexical features (e.g. WordNet *hyponyms, Common-parents, distance*), Named Entities (NE) and syntactic features (direct or transitive dependency relations of parsing) has noticeably improved the performance of the system. It is observed that *nominalization* feature plays an effective role for identifying as well as distinguishing the relations. The test set containing 2717 sentences is evaluated against four different training sets. Some of the relations, e.g. *Cause-Effect*, *Member-Collection* perform well in comparison to other relations in all the four test results. Reviewing of the confusion matrices suggests that the system performance can be improved by reducing the errors that occur in distinguishing the two individual nominals in each relation.

The rest of the paper is organized as follows. The pre-processing of resources and the baseline system are described in Section 2 and Section 3 respectively. Development of CRF-based model is discussed in Section 4. Experimental results along

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 206–209,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

with error analysis are specified in Section 5. Finally Section 6 concludes the paper.

2 Resource Pre-Processing

The annotated training corpus containing 8000 sentences was made available by the respective task organizers. The objective is to evaluate the effectiveness of the system in terms of identifying semantic relations between pair of nominals. The rule-based baseline system is evaluated against the whole training corpus. But, for in-house experiments regarding CRF based framework, the development data is prepared by randomly selecting 500 sentences from the 8000 training sentences. Rest 7500 sentences are used for training of the CRF-model. The format of one example entry in training file is as follows.

"The system as described above has its greatest application in an arrayed <e1>configuration</e1> of antenna <e2>elements</e2>."

Component-Whole (e2, e1)

Comment: Not a collection: there is structure here, organisation.

Each of the training sentences is annotated by the paired nominals tagged as *<e1>* and *<e2>*. The relation of the paired nominals and a comment portion describing the detail of the input type follows the input sentence.

The sentences are filtered and passed through Stanford Dependency Parser (Marneffe *et al.*, 2006) to identify direct as well as transitive dependencies between the nominals. The direct dependency is identified based on the simultaneous presence of both nominals, *<e1>* as well as *<e2>* in the same dependency relation whereas the transitive dependencies are verified if *<e1>* and *<e2>* are connected via one or more intermediate dependency relations.

Each of the sentences is passed through a Stanford Named Entity Recognizer (NER)[1] for identifying the named entities. The named entities are the useful hints to separately identify the relations like *Entity-Origin* and *Entity-Destination* from other relations as the *Origin* and *Destination* entities are tagged by the NER frequently than other entities.

Different seed lists are prepared for different types of verbs. For example, the lists for *causal* and *motion* verbs are developed by processing the XML files of English VerbNet (Kipper-Schuler, 2005). The list of the *causal* and *motion* verbs are prepared by collecting the member verbs if their corresponding class contain the semantic type "*CAUSE*" or "*MOTION*". The other verb lists are prepared manually by reviewing the frequency of verbs in the training corpus. The WordNet stemmer is used to identify the root forms of the verbs.

3 Baseline Model

The baseline model is developed based on the similarity clues present in the phrasal pattern containing verbs and prepositions. Different rules are identified separately for the nine different relations. A few WordNet features such as *hypernym*, *meronym*, *distance* and *Common-Parents* are added into the rule-based baseline model. Some of the relation specific rules are mentioned below.

For example, if any of the nominals contain their *meronym* property as "*whole*" and if the *hypernym* tree for one of the nominals contains the word "*whole*", the relation is identified as a *Component-Whole* relation. But, the ordering of the nominals *<e1>* and *<e2>* is done based on the combination of "*has*", "*with*" and "*of*" with other word level components.

The relations *Cause-Effect*, *Entity-Destination* are identified based on the *causal verbs* (cause, lead etc.) and *motion verbs* (go, run etc.) respectively. One of the main criteria for extracting these relations is to verify the presence of *causal* and *motion* verbs in between the text segment of *<e1>* and *<e2>*. Different types of specific *relaters* (as, because etc.) are identified from the text segment as well. It is observed that such specific causal *relaters* help in distinguishing other relations from *Cause-Effect*.

If one of the nominals is described as *instrument* type in its *hypernym* tree, the corresponding relation is identified as *Instrument-Agency* but the base level filtering criterion is applied if both the nominals belong to *instrument* type. On the other hand, if any of the nominals belong to the *hypernym* tree as *content* or *container* or *hold* type, it returns the relation *Content-Container* as a probable answer. Similarly, if both of them belong to the same type, the condition is fixed as false criterion for that particular category. The nominals identified as the part of *collective nouns* and associated with

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

phrases like "*of*", "*in*", "*from*" between *<e1>* and *<e2>* contain the relation of *Member-Collection*. The relations e.g. *Message-Topic* uses seed list of verbs that satisfy the *communication* type in the *hypernym* tree and *Product-Producer* relation concerns the *hypernym* feature as *Product* type.

But, the identification of the proper ordering of the entities in the relation, i.e., whether the relation is valid between *<e1, e2>* or *<e2, e1>* is done by considering the passive sense of the sentence with the help of the keyword "*by*" as well as by some passive dependency relations.

The evaluation of the rule-based baseline system on the 8000 training data gives an average F1-score of 22.45%. The error analysis has shown that use of lexical features only is not sufficient to analyze the semantic relation between two nominals and the performance can be improved by adopting strategies for differentiating the nominals of a particular pair.

4 CRF-based Model

To improve the baseline system performance, CRF-based machine learning framework (McCallum *et al.*, 2001) is considered for classifying the semantic relations that exist among the ordered pair of nominals. Identification of appropriate features plays a crucial role in any machine-learning framework. The following features are identified heuristically by manually reviewing the corpus and based on the frequency of different verbs in different relations.

- 11 WordNet features (*Synset, Synonym, Gloss, Hyponym, Nominalization, Holonym, Common-parents, WordNet distance, Sense ID, Sense count, Meronym*)
- Named Entities (NE)
- Direct Dependency
- Transitive Dependency
- 9 separate verb list containing relation specific verbs, each for 9 different semantic relations

Different singleton features and their combinations are generated from the training corpus. Instead of considering the whole sentence as an input to the CRF-based system, only the pairs of nominals are passed for classification. The previous and next token of the current token with respect to each of the relations are added in the template to identify their co-occurrence nature that in turn help in the classification process. Synsets containing synonymous verbs of the same and different senses are considered as individual features.

4.1 Feature Analysis

The importance of different features varies according to the genre of the relations. For example, the *Common-parents* WordNet feature plays an effective role in identifying the *Content-Container* and *Product-Producer* relations. If the nominals in a pair share a common *Sense ID* and *Sense Count* then this is considered as a feature. The combination of multiple features in comparison with a single feature generally shows a reasonable performance enhancement of the present classification system. Evaluation on the development data for the various feature combinations has shown that the *nominalization* feature effectively performs for all the relations. WordNet *distance* feature is used for capturing the relations like *Content-Container* and *Component-Whole*. The direct and transitive dependency syntactic features contribute in identifying the relation as well as identify the ordering of the entities *<e1>* and *<e2>* in the relation.

The *Named-Entity* (NE) relation plays an important role in distinguishing the relations, e.g., *Entity-Origin* and *Entity-Destination* from other relations. The *person* tagged NEs have been excluded from the present task as such NEs are not present in the *Entity-Origin* and *Entity-Destination* relations. It has been observed that the relation specific verbs supply useful clues to the training phrase for differentiating relations among nominals.

The system is trained on 7500 sentences and the evaluation is carried out on 500 development sentences achieving an F1-Score of 57.56% F1-Score. The tuning on the development set has been carried out based on the performance produced by the individual features that effectively contains *WordNet* relations. In addition to that, the combination of dependency features with verb feature plays an contributory role on the system evaluation results.

Relations	TD1			TD2			TD3			TD4		
	Prec.	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Cause-Effect	76.33	65.85	70.70	78.55	65.85	71.64	79.86	68.90	73.98	79.26	72.26	75.60
Component-Whole	49.25	31.41	38.36	48.76	37.82	42.60	50.77	42.31	46.15	58.40	49.04	53.31
Content-Container	31.35	30.21	30.77	37.93	34.38	36.07	40.65	32.81	36.31	51.15	34.90	41.49
Entity-Destination	37.58	62.67	46.98	43.43	63.36	51.53	43.09	63.01	51.18	47.07	60.62	52.99
Entity-Origin	62.50	46.51	53.33	61.95	49.22	54.86	60.18	52.71	56.20	64.02	53.10	58.05
Instrument-Agency	19.46	23.08	21.11	21.18	27.56	23.96	26.43	23.72	25.00	32.48	24.36	27.84
Member-Collection	50.97	67.81	58.20	54.82	70.82	61.80	59.93	72.53	65.63	66.80	71.67	69.15
Message-Topic	41.70	41.38	41.54	50.23	42.15	45.83	52.81	46.74	49.59	57.78	49.81	53.50
Product-Producer	52.94	7.79	13.58	48.94	9.96	16.55	59.09	16.88	26.26	53.17	29.00	37.54
Other	21.10	27.09	23.72	24.48	33.70	28.36	26.28	37.44	30.88	26.64	42.07	32.62
Average F1 score	**42.62**			**44.98**			**47.81**			**52.16**		

Table 1: Precision, Recall and F1-scores (in %) of semantic relations in (9+1) way directionality-based evaluation

5 Experimental Results

The active feature list is prepared after achieving the best possible F1-score of 61.82% on the development set of 500 sentences. The final training of the CRF-based model is carried out on four different sets containing 1000, 2000, 4000 and 8000 sentences. These four training sets are prepared by extracting sentences from the beginning of the training corpus and the final evaluation is carried out on 2717 test sentences as provided by the organizers. The results on the four test sets termed as TD1, TD2, TD3 and TD4 are shown in Table 1. The error analysis is done based on the information present in the confusion matrices. The fewer occurrence of *Entity-Destination (e2, e1)* instance in the training corpus plays the negative role in identifying the relation. Mainly, the strategy used for assigning the order among the entities, i.e., either *<e1, e2>* or *<e2, e1>* in the already identified relations is the main cause of errors of the system. The *Entity-Origin, Product-Producer* and *Message-Topic* relations suffer from overlapping problem with other relations. Each of the tested nominal pairs is tagged with more than one relation. But, selecting the first output tag produced by CRF is considered as the final relational tag for each of the nominal pairs. Hence, a distinguishing strategy needs to be adopted for fine-grained selection.

6 Conclusion and Future Task

In our approach to automatic classification of semantic relations between nominals, the system achieves its best performance using the lexical feature such as *nominalization* of WordNet and syntactic information such as dependency relations. These facts lead us to conclude that semantic features from WordNet, in general, play a key role in the classification task. The present system aims for assigning class labels to discrete word level entities but the context feature is not taken into consideration. The future task is to evaluate the performance of the system by capturing the context present between the pair of nominals.

References

Andrew McCallum, Fernando Pereira and John Lafferty. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and labeling Sequence Data. ICML-01, 282 – 289.

George A. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4): 235–312.

Karin Kipper-Schuler. 2005. VerbNet. A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. *(LREC 2006).*

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid ´O S´eaghdha, Sebastian Padok , Marco Pennacchiotti, Lorenza Romano, Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals. *5th SIGLEX Workshop.*