BUAP: A First Approximation to Relational Similarity Measuring

Mireya Tovar, J. Alejandro Reyes, Azucena Montes
CENIDET, Department of Computer Science
Int. Internado Palmira S/N, Col. Palmira Cuernavaca, Morelos, México
{mtovar, alexreyes06c, amr}@cenidet.edu.mx

Darnes Vilariño, David Pinto, Saul León
B. Universidad Autónoma de Puebla, Faculty of Computer Science
14 Sur y Av. San Claudio, CU Puebla, Puebla, México
{darnes, dpinto}@cs.buap.mx saul.ls@live.com

Abstract

We describe a system proposed for measuring the degree of relational similarity between a pair of words at the Task #2 of Semeval 2012. The approach presented is based on a vectorial representation using the following features: i) the context surrounding the words with a window size = 3, ii) knowledge extracted from WordNet to discover several semantic relationships, such as meronymy, hyponymy, hypernymy, and part-whole between pair of words, iii) the description of the pairs with their POS tag, morphological information (gender, person), and iv) the average number of words separating the two words in text.

1 Introduction

The Task # 2 of Semeval 2012 focuses on measuring the degree of relational similarity between the reference words pairs (training) and the test pairs for a given class (Jurgens et al., 2012). The training data set consists of 10 classes and the testing data set consists of the 69 classes. These datasets as well as the particularities of the task are better described at overview paper (Jurgens et al., 2012). In this paper we report the approach submitted to the competition, which is based on a vector space model representation for each pair (Salton et al., 1975). With respect to the type of features used, we have observed that Fabio Celli (Celli, 2010) considers that contextual information is useful, as well the lexical and semantic information are in the extraction of semantic relationships task. Additionally, in (Chen et al., 2010) and (Negri and Kouylekov, 2010) are proposed WordNet based features with the same purpose.

In the experiments carried out in this paper, we use a set of lexical, semantic, WordNet-based and contextual features which allows to construct the vectors. Actually, we have tested a subset of the 20 contextual features proposed by Celli (Celli, 2010) and some of those proposed by Chen (Chen et al., 2010) and Negri (Negri and Kouylekov, 2010).

The cosine similarity measure is used for determining the degree of relational similarity (Frakes and Baeza-Yates, 1992) among the vectors.

The rest of this paper is structured as follows. Section 2 describes the system employed. Section 3 show the obtained results. Finally, in Section 4 the final conclusions are given.

2 System description

The approach reported in this paper measures the relational similarity of a set of word pairs that belong to the same semantic relationship. Those word pairs are represented by means of the vector space model (Salton et al., 1975). Each value of the vector represents the average value of the corresponding feature. This average is calculated using 100 samples obtained from Internet by employing the Google search engine. The search process is carried out assuming that those words co-occurring in the same context contain some kind of semantic relationship.

Let \((w_1, w_2)\) be a word pair, then the vectorial representation of this pair \((\vec{x})\) using semantic, contextual, lexical, and WordNet-based features may be expressed as it can be seen in Eq. (1).
The previous example is only illustrative, since we have gathered 100 sentence per word pair. In total, we collected a corpus containing 2,054,687 tokens, with an average class terms of 26,684 and with an average class vocabulary of 4,006.

The features extracted are described as follows:

2.1 Lexical features

The lexical features describe morphologically and syntactically the word pair \((w_1, w_2)\). The lexical features extracted are the following:

- Average number of words separating the two words \((w_1, w_2)\) in the text.
- The position of \(w_1\) with respect to \(w_2\) in the text. If \(w_1\) appears before \(w_2\) then the feature value is 1, otherwise, the value is 2.
- The Part of Speech Tag for each word in the pair (two features). We use the FreeLing PoS-tagger (Padró et al., 2010) for obtaining the grammatical category. The possible values are the following: adjective=1; adverb=2; article=3; noun=4; verb=5; pronoun=6; conjunction=7; preposition=8

2.2 Semantic features

The following four semantic features are boolean values (true or false) indicating:

- If \(w_1\) and \(w_2\) are named entities (two features)\(^1\).
- If \(w_1\) and \(w_2\) are entities defined (two features)\(^2\).

The following two semantic features indicate:

- The type of prepositional phrase in case of existing for \(w_1\) and \(w_2\). The feature values are nominal: about=1; after=2; at=3; behind=4; between=5; by=6; except=7; from=8; into=9; near=10; of=11; over=12; through=13; until=14; under=15; upon=16; without=17; above=18; among=19; before=20; below=21; beside=22; but=23; down=24; for=25; in=26; on=27; since=28; to=29; with=30.

2.3 WordNet-based features

The semantic features are boolean values (true or false) indicating whether or not \(w_2\) is contained in:

- the synonym set of \(w_1\)
- the antonym set of \(w_1\)
- the meronymy set of \(w_1\)
- the hyponymy set of \(w_1\)
- the hypernymy set of \(w_1\)
- the part-whole set of \(w_1\)
- the gloss set of \(w_1\)

We used WordNet (Fellbaum, 1998) in order to determine the relationship set for word \(w_1\).

\(^1\)A named entity is defined by a Proper Noun Phrase, which was detected using the module NER-Named Entity Recognition of the FreeLing 2.1 tool.

\(^2\)A defined sentence is one that begins with a definite article.
2.4 Contextual features

Contextual features considers values for the words that occur in the context of $w_1$ and $w_2$ (in a window size of 3). The description of those features follows.

- Nominal values indicating the Part of Speech Tag (adjective=1; adverb=2; article=3; noun=4; verb=5; pronoun=6; conjunction=7; preposition=8) for the three words at:
  - the left context of $w_1$ (three features).
  - the right context of $w_1$ (three features).
  - the left context of $w_2$ (three features).
  - the right context of $w_2$ (three features).

- A Nominal value indicating number of the following grammatical categories between $w_1$ and $w_2$: verbs, adjectives and nouns (three features).

- Nominal values indicating the frequencies of the verbs: be, do, have, locate, know, make, use, become, include, take between $w_1$ and $w_2$ (ten features).

2.5 Feature selection

We carried out a feature selection process with the aim of discarding irrelevant features. In this step, we apply the attribute selection filter reported in (Hall, 1999), that evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them and an exhaustive search method.

The following features were obtained as relevant: the average number of words between $w_1$ and $w_2$; Named Entity of $w_1$ and $w_2$; phrase defined of $w_1$ and $w_2$; prepositional phrase type $w_1$ and $w_2$; part of speech tag $w_1$ and $w_2$; part of speech tag of right context of $w_1$ with a windows size of 3; occurrences of verbs between $w_1$ and $w_2$; frequency of verbs be, do, make, locate, take; synonym, antonym, meronymy, hyponymy, hypernymy, part-whole and gloss relationships between $w_1$ and $w_2$.

After applying the aforementioned feature selection method, we removed 17 features, and the vectorial representation of each word pair will be done with only 25 values (features).

2.6 Determining the degree of similarity

We have used the features mentioned before for constructing a prototype vector representing a given semantic class. In order to do so, we have employed the training corpus for gathering samples from Internet and, thereafter, we average the feature values in order to construct such prototype vector.

For each word pair in the test dataset, we obtained a vector using the same process explained before. We determined the similarity for each test feature vector with respect to the prototype of the given class by using the cosine similarity coefficient (Frakes and Baeza-Yates, 1992), i.e., measuring the cosine of the angle between the two vectors.

In this way, we obtain a similarity measure of each test word pair with respect to its corresponding class. Finally, we may output a ranking of all the word pairs at the test dataset by sorting these similarity values obtained.

3 Experimental results

The approach submitted to the Task #2 of SemEval 2012 obtained very poor results. The Spearman correlation coefficient, which measured the correlation of the approach with respect to the gold standard, it is quite low (see Table 1).

| Team-Algorithm | Spearman | MaxDiff |
|----------------|----------|---------|
| UTD-NB         | 0.23     | 39.4    |
| UTD-SVM        | 0.12     | 34.7    |
| DULUTH-V0      | 0.05     | 32.4    |
| DULUTH-V1      | 0.04     | 31.5    |
| DULUTH-V2      | 0.04     | 31.1    |
| BUAP           | 0.01     | 31.7    |
| Random         | 0.02     | 31.2    |

Table 1: Spearman and MaxDiff scores obtained at the Task #2 of SemEval 2012

Actually, it shows that the run submitted does not correlate with the gold standard. We consider that this behavior is derived from the nature of the support corpus used for obtaining the features set. The number of sentences (100) used for representing the word pairs was not enough for constructing a real prototype of both, the semantic class and the word pairs. A further analysis will confirm this issue.
Despite this limitation we note that the MaxDiff score was 31.7% slightly above the baseline (31.2%) and not far from the best score of the task (39.4%). That is, we achieved an average of 31.7% of questions answered correctly.

4 Discussion and conclusion

In this paper we report the set of features used in the approach submitted for measuring the degrees of relational similarity between a given reference word pair and a variety of other pairs. The results obtained are not encouraging with a Spearman correlation coefficient close to zero, which mean that there are not correlation between the run submitted and the gold standard. A deeper analysis of the approach is needed in order to determine if the limitation of the system falls in the features used, the similarity measure, or the support corpus used for extracting the features.

Acknowledgments

This project has been partially supported by projects CONACYT #106625, VIEP #PIAD-ING11-II and #VIAD-ING11-II.

References

Fabio Celli. 2010. Unitn: Part-of-speech counting in relation extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval ’10, pages 198–201, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yuan Chen, Man Lan, Jian Su, Zhi Min Zhou, and Yu Xu. 2010. Ecnu: Effective semantic relations classification without complicated features or multiple external corpora. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval ’10, pages 226–229, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press.

William B. Frakes and Ricardo A. Baeza-Yates, editors. 1992. *Information Retrieval: Data Structures & Algorithms*. Prentice-Hall.

Mark A. Hall. 1999. *Correlation-based Feature Subset Selection for Machine Learning*. Ph.D. thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

David A. Jurgens, Saif M. Mohammad, Peter D. Turney, and Keith J. Holyoak. 2012. *SemEval-2012 task 2: Measuring degrees of relational similarity*. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montreal, Canada.

Matteo Negri and Milen Kouylekov. 2010. Fbk_nk: A wordnet-based system for multi-way classification of semantic relations. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval ’10, pages 202–205, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lluis Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November.