BUAP: Evaluating Features for Multilingual and Cross-Level Semantic Textual Similarity

Darnes Vilariño, David Pinto, Saúl León, Mireya Tovar,Beatriz Beltrán
Benemérita Universidad Autónoma de Puebla
Faculty of Computer Science
14 Sur y Av. San Claudio, CU
Puebla, Puebla, México
{darnes,dpinto,saul.leon,mtovar,bbeltran}@cs.buap.mx

Abstract

In this paper we present the evaluation of different features for multiligual and cross-level semantic textual similarity. Three different types of features were used: lexical, knowledge-based and corpus-based. The results obtained at the Semeval competition rank our approaches above the average of the rest of the teams highlighting the usefulness of the features presented in this paper.

1 Introduction

Semantic textual similarity aims to capture whether the meaning of two texts are similar. This concept is somehow different from the textual similarity definition itself, because in the latter we are only interested in measuring the number of lexical components that the two texts share. Therefore, textual similarity can range from exact semantic equivalence to a complete unrelatedness pair of texts.

Finding the semantic similarity between a pair of texts has become a big challenge for specialists in Natural Language Processing (NLP), because it has applications in some NLP task such as machine translation, automatic construction of summaries, authorship attribution, machine reading comprehension, information retrieval, among others, which usually need a manner to calculate degrees of similarity between two given texts.

Semantic textual similarity can be calculated using texts of different sizes, for example between, a paragraph and a sentence, or a sentence and a phrase, or a phrase and a word, or even a word and a sense. When we consider this difference, we say the task is called "Cross-Level Semantic Similarity", but when this distinction is not considered, then we call the task just as "Semantic Textual Similarity".

In this paper, we evaluate different features for determining those that obtain the best performances for calculating both, cross-level semantic similarity and multilingual semantic textual similarity.

The remaining of this paper is structured as follows. Section 2 presents the features used in both experiments. Section 3 shows the manner we used the features for determining the degree of semantic textual similarity. Section 4, on the other hand, shows the experiments we have carried out for determining cross-level semantic similarity. Finally, in Section 5 the conclusions and findings are given.

2 Description of Features

In this section we describe the different features used for evaluation semantic textual similarity. Basically, we have used three different types of features: lexical, knowledge-based and corpus-based. The first one, counts the frequency of ocurrence of lexical features which include n-grams of characters, $skip$-grams[1], words and some lexical relationships such as synonymy or hypernymy. Additionally, we have used two other features: the Jaccard coefficient between the two text, expanding each term with a set of

[1]They are also known as disperse n-grams because they consider to "skip" a certain number of characters.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 149–153,
Dublin, Ireland, August 23-24, 2014.

synonyms taken from WordReference Carrillo et al. (2012), and the cosine between the two texts represented each by a bag of character n-grams and $skip$-grams. In this case, we did not applied any word sense disambiguation system before expanding with synonyms, a procedure that may be performed in a further work.

The second set of features considers the following six word similarity metrics offered by NLTK: Leacock & Chodorow (Leacock and Chodorow, 1998), Lesk (Lesk, 1986), Wu & Palmer (Wu and Palmer, 1994), Resnik (Resnik, 1995), Lin (Lin, 1998), and Jiang & Conrath[2] (Jiang and Conrath, 1997). In this case, we determine the similarity between two texts as the maximum possible pair of words similarity. The third set of features considers two corpus-based measures, both based on Rada Mihalcea's textual semantic similarity (Mihalcea et al., 2006). The first one uses Pointwise Mutual Information (PMI) (Turney, 2001) for calculating the similarity between pairs of words, whereas the second one uses Latent Semantic Analysis (LSA) (Landauer et al., 1998) (implemented in the R software environment for statistical computing) for that purpose. In particular, the PMI and LSA values were obtained using a corpus built on the basis of Europarl, Project-Gutenberg and Open Office Thesaurus. A summary of these features can be seen in Table 1.

3 Multilingual Semantic Textual Similarity

This task aims to find the semantic textual similarity between two texts written in the same language. Two different languages were considered: English and Spanish. The degree of semantic similarity ranges from 0 to 5; the bigger this value, the best semantic match between the two texts. For the experiments we have used the training datasets provided at 2012, 2013 and 2014 Semeval competitions. These datasets are completely described at the task description papers of these Semeval editions Agirre et al. (2013, 2014).

In order to calculate the semantic textual similarity for the English language, we have used all the features mentioned at Section 2. We have constructed a single vector for each pair of texts of the training corpus, thus resulting 6,627 vectors in total. The resulting set of vectors fed a supervised classifier, in particular, a logistic regression model[3]. This approach has been named as *BUAP-EN-run1*. The most representative results obtained at the competition for the English language can be seen in Table 2. As can be seen, we outperformed the average result in all the cases, except on the case that the *OnWN* corpus was used.

In order to calculate the semantic textual similarity for the Spanish language, we have submitted two runs, the first one is a supervised approach which constructs a regression model, similar that the one constructed for the English language, but considering only the following features: character n-grams, character $skip$-grams, and the cosine similarity of bag of character n-grams and $skip$-grams. This approach was named *BUAP-run1*. Given that the number of Spanish samples was so small, we decided to investigate the behaviour of training with English and testing with Spanish language. It is quite interesting that this approach obtained a relevant ranking (17 from 22 runs), even if the type of features used were naïve.

The second approach submitted for determining the semantic textual similarity for the Spanish language is an unsupervised one. It uses the same features of the supervised approach for Spanish, but these features were used to create a representation vector for each text (independently), so that we may be able to calculate the similarity by means of the cosine measure between the two vectors. The approach was named *BUAP-run2*.

The most representative results obtained at the competition for the Spanish language can be seen in Table 3. There we can see that our unsupervised approach slightly outperformed the overall average, but the supervised approach was below the overall average, a fact that is expected since we have trained using the English corpus and testing with the Spanish language. Despite this, it is quite interesting that the result obtained with this supervised approach is not so bad.

Due to space constraints, we did not reported the complete set of results of the competition, however, these results can be seen at the task 10 description

[2]Natural Language Toolkit of Python; http://www.nltk.org/

[3]We used the version of the logistic classifier implemented in the the Weka toolkit

Table 1: Features used for calculating semantic textual similarity

Feature	Type
n-grams of characters ($n = 2, \cdots, 5$)	Lexical
$skip$-grams of characters ($skip = 2, \cdots, 5$)	Lexical
Number of words shared	Lexical
Number of synonyms shared	Lexical
Number of hypernyms shared	Lexical
Jaccard coefficient with synonyms expansion	Lexical
Cosine of bag of character n-grams and $skip$-grams	Lexical
Leacock & Chodorow's word similarity	Knowledge-based
Lesk's word similarity	Knowledge-based
Wu & Palmer's word similarity	Knowledge-based
Resnik's word similarity	Knowledge-based
Lin's word similarity	Knowledge-based
Jiang & Conrath's word similarity	Knowledge-based
Rada Mihalcea's metric using PMI	Corpus-based
Rada Mihalcea's metric using LSA	Corpus-based

Table 2: Results obtained at the Task 10 of the Semeval competition for the English language

Team Name	deft-forum	deft-news	headlines	images	OnWN	tweet-news	Weighted mean	Rank
DLS@CU-run2	0.4828	0.7657	0.7646	0.8214	0.8589	0.7639	0.7610	1
Meerkat_Mafia-pairingWords	0.4711	0.7628	0.7597	0.8013	0.8745	0.7793	0.7605	2
NTNU-run3	0.5305	0.7813	0.7837	0.8343	0.8502	0.6755	0.7549	3
BUAP-EN-run1	**0.4557**	**0.6855**	**0.6888**	**0.6966**	**0.6539**	**0.7706**	**0.6715**	**19**
Overall average	0.3607	0.6198	0.5885	0.6760	0.6786	0.6001	0.6015	27-28
Bielefeld_SC-run2	0.2108	0.4307	0.3112	0.3558	0.3607	0.4087	0.3470	36
UNED-run22_p_np	0.1043	0.3148	0.0374	0.3243	0.5086	0.4898	0.3097	37
LIPN-run2	0.0843	-	-	-	-	-	0.0101	38
Our difference against the average	9%	7%	10%	2%	-2%	17%	7%	-

Table 3: Results obtained at the Task 10 of the Semeval competition for the Spanish language (NOTE: The * symbol denotes a system that used Wikipedia to build its model for the Wikipedia test dataset)

Team Name	System type	Wikipedia	News	Weighted correlation	Rank
UMCC_DLSI-run2	supervised	0.7802	0.8254	0.8072	1
Meerkat_Mafia-run2	unsupervised	0.7431	0.8454	0.8042	2
UNAL-NLP-run1	weakly supervised	0.7804	0.8154	0.8013	3
BUAP-run2	**unsupervised**	**0.6396**	**0.7637**	**0.7137**	**14**
Overall average	-	0.6193	0.7504	0.6976	14-15
BUAP-run1	**supervised**	**0.5504**	**0.6785**	**0.6269**	**17**
RTM-DCU-run2	supervised	0.3689	0.6253	0.5219	20
Bielefeld_SC-run2	unsupervised*	0.2646	0.5546	0.4377	21
Bielefeld_SC-run1	unsupervised*	0.2632	0.5545	0.4371	22
Difference between our run1 and the overall average	-	-7%	-7%	-7%	-
Difference between our run2 and the overall average	-	2%	1%	2%	-

paper (Agirre et al., 2014) of Semeval 2014.

4 Cross-Level Semantic Similarity

This task aims to find semantic similarity between a pair of texts of different length written in English language, actually each text belong to a different level of representation of language (paragraph, sentence, phrase, word, and sense). Thus, the pair of levels that were required to be compared in order to determine their semantic similarity were: paragraph-to-sentence, sentence-to-phrase, phrase-to-word, and word-to-sense.

The task cross level similarity judgments are based on five rating levels which goes from 0 to

4. The first (0) implies that the two items do not mean the same thing and are not on the same topic, whereas the last one (4) implies that the two items have very similar meanings and the most important ideas, concepts, or actions in the larger text are represented in the smaller text. The remaining rating levels imply something in the middle.

For word-to-sense comparison, a sense is paired with a word and the perceived meaning of the word is modulated by virtue of the comparison with the paired sense's definition. For the experiments presented at the competition, a corpus of 2,000 pairs of texts were provided for training and other 2,000 pairs for testing. This dataset considered 500 pairs for each type of level of semantic similarity. The complete description of this task together with the dataset employed is given in the task description paper Jurgens et al. (2014).

We submitted two supervised approaches, to this task employing all the features presented at Section 2. The first approach simply constructs a single vector for each pair of training texts using the aforementioned features. These vectors are introduced in Weka for constructing a classification model based on logistic regression. This approach was named *BUAP-run1*.

We have observed that when comparing texts of different length, there may be a high discrepancy between those texts because a very small length in the texts may difficult the process of determining the semantic similarity. Therefore, we have proposed to expand small text with the aim of having more term useful in the process of calculating the degree of semantic similarity. In particular, we have expanded words for the phrase-to-word and word-to-sense cases. The expansion has been done as follows. When we calculated the similarity between phrases and words, we expanded the word component with those related terms obtained by means of the Related-Tags Service of Flickr. When we calculated the semantic similarity between words and senses, we expanded the word component with their WordNet Synsets (none word sense disambiguation method was employed). This second approach was named *BUAP-run2*.

The most representative results for the cross-level semantic similarity task (which include our results) are shown in Table 4. There we can see that the features obtained a good performance when we computed the semantic similarity between paragraphs and sentences, and when we calculated the similiraty between sentences to phrases. Actually, both runs obtained exactly the same result, because the main difference between these two runs is that the second one expands the word/sense using the Related Tags of Flickr. However, the set of expansion words did not work properly, in particular when calculating the semantic similarity between phrases and words. We consider that this behaviour is due to the domain of the expansion set do not match with the domain of the dataset to be evaluated. In the case of expanding words for calculating the similarity between words and senses, we obtained a slightly better performance, but again, this values are not sufficient to highly outperform the overall average. As future work we consider to implement a self-expansion technique for obtaining a set of related terms by means of the same training corpus. This technique has proved to be useful when the expansion process is needed in restricted domains Pinto et al. (2011).

5 Conclusions

This paper presents the results obtained by the BUAP team at the Task 3 and 10 of SemEval 2014. In both task we have used a set of similar features, due to the aim of these two task are quite similar: determining semantic similarity. Some special modifications has been done according to each task in order to tackle some issues like the language or the text length.

In general, the features evaluated performed well over the two approaches, however, some issues arise that let us know that we need to tune the approaches presented here. For example, a better expansion set is required in the case of the Task 3, and a great number of samples for the spanish samples of Task 10 will be required.

References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *sem 2013 shared task: Semantic textual similarity. In *2nd Joint Conference on Lexical and Computational*

Table 4: Results obtained at Task 3 of Semeval 2014

Team	System	Paragraph-to-Sentence	Sentence-to-Phrase	Phrase-to-Word	Word-to-Sense	Rank
SimCompass	run1	0.811	0.742	0.415	0.356	1
ECNU	run1	0.834	0.771	0.315	0.269	2
UNAL-NLP	run2	0.837	0.738	0.274	0.256	3
BUAP	**run1**	**0.805**	**0.714**	**0.162**	**0.201**	**9**
BUAP	**run2**	**0.805**	**0.714**	**0.142**	**0.194**	**10**
Overall average	-	0.728	0.651	0.198	0.192	11-12
Our run1 - Overall average		8%	6%	-4%	1%	-
Our run2 - Overall average		8%	6%	-6%	0%	-

*Semantics (*SEM)*, pages 32–43, Atlanta, Georgia, USA, 2013.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland, 2014.

Maya Carrillo, Darnes Vilariño, David Pinto, Mireya Tovar, Saul León, and Esteban Castillo. Fcc: Three approaches for semantic textual similarity. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (SemEval 2012)*, pages 631–634, Montréal, Canada, 2012.

Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc of 10th International Conference on Research in Computational Linguistics, ROCLING'97*, pages 19–33, 1997.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland, 2014.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, (25):259–284, 1998.

Claudia Leacock and Martin Chodorow. Combining local context and wordnet similarity for word sense identification. In Christiane Fellbaum, editor, *MIT Press*, pages 265–283, 1998.

Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26. ACM, 1986.

Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 775–780, 2006.

David Pinto, Paolo Rosso, and Héctor Jiménez-Salazar. A self-enriching methodology for clustering narrow domain short texts. *Computer Journal*, 54(7):1148–1165, 2011.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995.

Peter D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, pages 491–502. Springer-Verlag, 2001.

Zhibiao Wu and Martha Stone Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, 1994.