Unlocking the Power of Word2Vec for Identifying Implicit Links

2017 IEEE 17th International Conference on Advanced Learning Technologies

Unlocking the Power of Word2Vec for Identifying Implicit Links

Gabriel Gutu, Mihai Dascalu, Stefan Ruseti, Traian Rebedea, Stefan Trausan-Matu
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania
{gabriel.gutu, mihai.dascalu, stefan.ruseti, traian.rebedea, stefan.trausan}@cs.pub.ro

***Abstract*—This paper presents a research on using Word2Vec for determining implicit links in multi-participant Computer-Supported Collaborative Learning chat conversations. Word2Vec is a powerful and one of the newest Natural Language Processing semantic models used for computing text cohesion and similarity between documents. This research considers cohesion scores in terms of the strength of the semantic relations established between two utterances; the higher the score, the stronger the similarity between two utterances. An implicit link is established based on cohesion to the most similar previous utterance, within an imposed window. Three similarity formulas were used to compute the cohesion score: an unnormalized score, a normalized score with distance and Mihalcea's formula. Our corpus of conversations incorporated explicit references provided by authors, which were used for validation. A window of 5 utterances and a 1-minute time frame provided the highest detection rate both for exact matching and matching of a block of continuous utterances belonging to the same speaker. Moreover, the unnormalized score correctly identified the largest number of implicit links.**

Keywords—implicit links; CSCL; Word2Vec; semantic models; text cohesion

I. INTRODUCTION

Computer-Supported Collaborative Learning (CSCL) was imposed in recent years due to the advance of communication and collaborative technologies on the social web [1]. Moreover, CSCL emerged as a well-suited method for learning through a knowledge building process according to the socio-cultural paradigm [1]. One of the most popular technologies used in CSCL is instant messenger (chat). Chat environments, when also integrating explicit referencing facilities, enable small groups of students to generate complex parallel threads of discussions, inter-animating in a polyphonic framework [2].

The polyphonic weaving of knowledge construction in CSCL chats involves threads composed of explicit and implicit links. The latter are pairs of utterances part of a discussion thread, which are logically connected in a discursive structure. Implicit links may be detected using Natural Language Processing (NLP) techniques: repetitions, lexical chains, adjacency pairs of speech acts [2], and semantic models or other means for measuring the similarity between two utterances [3].

In this paper, we investigate the use of Word2Vec [4] for identifying implicit links in chat conversations. Word2Vec, a new NLP technique developed for assessing the semantics of a text, proved in recent years to provide better performance for several NLP tasks involving semantic analysis than previous approaches [4].

II. THE READERBENCH FRAMEWORK

ReaderBench is an open-source multi-lingual natural language analysis framework that supports various NLP-related scenarios [3]. Of these scenarios, text cohesion analysis is the one of high interest for our study as it can detect relations between units of texts of different granularity. This study considers text cohesion in terms of the strength of the semantic relations established between two utterances. A semantic relation is expressed as a cohesion score – a higher score shows a stronger similarity between two utterances. *ReaderBench* encapsulates some of the latest and the most utilized semantic models such as Latent Semantic Analysis, Latent Dirichlet Allocation [3] and the more recently Word2Vec [4]. Regarding ontologies, WordNet (http://wordnet.princeton.edu) for different languages, together with corresponding semantic distances, is also integrated within our framework [3].

Word2Vec [4] is one of the most recent methods used for representing words and phrases in a vector-space model within a limited number of dimensions, called word embeddings, which are computed using a neural network model. The resulted embedded space can be used afterwards to compute a semantic similarity between words and phrases. In the case of Word2Vec, each embedding is computed using the context before and after each word occurrence in the training dataset. This way, words co-occurring in similar contexts are represented closer in the embedded space, while words that do not share similar context are represented in different regions of this space (are farther apart).

Our previous experiment [5] conducted using the WordNet ontology and semantic models, but not Word2Vec, emphasized that the path-length similarity score based on WordNet was the most suitable for determining implicit links. This study introduces Word2Vec as a new semantic model for detection of implicit links.

2161-377X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICALT.2017.120

III. RESULTS

For this research, a corpus of 55 CSCL chat conversations given in a specific XML format was used. The conversations were annotated by participants with references to previous utterances, i.e., explicit links. The collection summed a total of more than 17,500 utterances and was refined by eliminating conversations lasting less than 30 minutes or having less than 50 utterances or 10 explicit links [5].

A Word2Vec semantic model was used to compute semantic similarity scores between utterances. The model was trained on a custom-built corpus that contained the TASA corpus (Touchstone Applied Science Associates, Inc., http://lsa.colorado.edu/spaces.html) and a collection of more than 500 CSCL-related scientific papers. The Word2Vec model was trained using the Skip-gram model with negative sampling as described by Mikolov, et al. [4]. As hyper-parameters, we used a vector size of 300, a windows size of 5 words, 3 epochs and 3 iterations. Words with less than 5 occurrences were ignored during training.

Three similarities formulas were used: 1) *W2V_SIM* – the basic Word2Vec semantic similarity score computed using cosine similarity; 2) *W2V_NSIM* – the normalized score by the inverse of the distance between an utterance and its referee; 3) *W2V_MSIM* – Mihalcea's formula [6] which increases the similarity score of a pair of utterances by the highest similarity score between one word belonging to an utterance, and another belonging to the other utterance.

Two criteria were considered in detecting implicit links: *exact matching* and *in-turn matching*. The second one detected implicit links within the block of continuous utterances belonging to the same participant. The pairs with the highest similarity scores within an imposed window of utterances were considered implicit links. Distance (in terms of number of utterances) and time frame (in terms of interval between posting times) were considered as windows. Significant differences in coverage of explicit links were observed for the distances of 20, 10 and 5 utterances and for the time frames of 5, 3, 2, 1 and 0.5 minutes [5]. Significant differences in coverage of explicit links were observed for the 5 and 10 utterances distance, and for the 1 and 2-minute time frames. All the other potential scenarios were discarded.

Experiments were performed considering combinations of distance and time frame pairs to determine the optimal window for identifying implicit links (see Table I). The highest detection rates were obtained using the W2V_SIM formula for both the exact matching and in-turn matching criteria. However, there was no improvement when adjusting the time frame – for both distances, the percentages remained unchanged for the 1 and 2-minute time frames. Regarding distance, the smaller window size of 5 utterances provided better results. Overall, a window of 5 utterances and a 1-minute time frame provided the best detection rate.

TABLE I. IMPLICIT LINKS DETECTION RATE *

Window Size, Time Frame	*Method*	*Exact matching*	*In-turn matching*
5 utt., 1 min	W2V_SIM	**30.56%**	**39.36%**
	W2V_NSIM	25.32%	34.64%
	W2V_MSIM	28.16%	38.18%
10 utt., 1min	W2V_SIM	**29.46%**	**37.64%**
	W2V_NSIM	25.37%	34.71%
	W2V_MSIM	26.81%	36.16%
5 utt., 2 mins	W2V_SIM	**30.56%**	**39.36%**
	W2V_NSIM	25.32%	34.64%
	W2V_MSIM	28.16%	38.18%
10 utt., 2 mins	W2V_SIM	**29.46%**	**37.64%**
	W2V_NSIM	25.37%	34.71%
	W2V_MSIM	26.81%	36.16%

* Highest percentages are emphasized in bold

IV. CONCLUSIONS

This paper extends previous experiments focused on determining implicit links in chat conversations [5] by integrating the Word2Vec semantic model. Implicit links were detected by determining the preceding utterance having the highest semantic similarity score. Our experiments showed that a window of 5 utterances and a time frame of 1-minute provided the highest detection rate when employing the standard Word2Vec similarity measure.

Future work covers the separation of explicit references into two categories, i.e., simple and ambiguous references. This will enable a new analysis centered on determining whether semantic similarity techniques could provide better results for one category over the other.

ACKNOWLEDGEMENT

This research was partially supported by the 644187 EC H2020 RAGE project, as well as by University Politehnica of Bucharest through the "Excellence Research Grants" Program UPB–GEX 12/26.09.2016.

REFERENCES

[1] G. Stahl, Group cognition. Computer support for building collaborative knowledge. Cambridge, MA: MIT Press, 2006.

[2] S. Trausan-Matu and T. Rebedea, "A polyphonic model and system for inter-animation analysis in chat conversations with multiple participants," in 11th Int. Conf. Computational Linguistics and Intelligent Text Processing (CICLing 2010), New York, 2010, pp. 354–363.

[3] M. Dascalu, S. Trausan-Matu, D. S. McNamara, and P. Dessus, "ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism," International Journal of Computer-Supported Collaborative Learning, vol. 10, pp. 395–423, 2015.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representation in Vector Space," in Workshop at ICLR, Scottsdale, AZ, 2013.

[5] G. Gutu, M. Dascalu, T. Rebedea, and S. Trausan-Matu, "Which Semantic Similarity Measure Best Captures Implicit Links in CSCL Conversations?," in 12th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2017), Philadelphia, PA, in press.

[6] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," presented at the 21st Int. Conf. AAAI, Boston, Massachusetts, 2006.