FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2016

Interpretable Semantic Textual Similarity with Distributional Semantics for Chunks

Ondřej Pražák[1], David Steinberger[2], Miloslav Konopík[3], Tomáš Brychcín[4]

Introduction

The goal of the *Interpretable Semantic Textual Similarity* task is to go deeper with the assessment of semantic textual similarity of sentence pairs. It is requested to add an explanatory layer that offers a deeper insight into the sentence similarities. The sentences are split into chunks and the first goal is to find corresponding chunks (with respect to their meanings) among the compared sentences. When the corresponding chunks are known, the chunks are annotated with their similarity scores and their relation types (e.g. equivalent, more specific, etc).

Machine Learning Approach

The main effort of our team was focused on the machine learning approach to the task. We divided the task into to three classification / regression tasks:
Alignment binary classification – we decide whether two given chunks should be aligned with each other.
Score classification / regression – we experiment with both classification and regression of the chunks similarity score.
Type classification – we classify all aligned pairs of chunks into a predefined set of types.

Classifiers

We experiment with the following classifiers: *Maximum Entropy Classifier*, *Support Vector Machines Classifier Multilayer perceptron* and *Voted perceptron neural networks* and with *Decision / regression tree learning*.

Features

We use following features for classifiers: word base form overlap, word lemma overlap, chunk length difference, word sentence positions difference, parse tree path and position difference features, number of part of speech differences and *WordNet* puth differences. The main feature is chunk semantic similarity estimate. Our attempts to estimate the similarity score between chunks are based upon estimating semantic similarity of individual words and compiling them into one number for a given chunk pair. We experiment with Word2Vec [2] and GloVe [3] for estimating similarity of words. We compile all the word similarities in one number that

[1] student of master study programme Computer Science and Engineering, software engeneering, e-mail: ondfa@students.zcu.cz
[2] student of master study programme Computer Science and Engineering, Intelligent Computer Systems e-mail: fenic@students.zcu.cz
[3] Faculty of Applied Sciences, e-mail: konopik@kiv.zcu.cz
[4] Faculty of Applied Sciences, e-mail: brychcín@kiv.zcu.cz

reflects semantic similarity of whole chunks via *lexical semantic vectors*.

Modified lexical semantic vectors method is based upon [1]. At first we create a combined vocabulary of all unique words from chunks $\mathbf{CH}_k^a$ and $\mathbf{CH}_l^b$: $\mathbf{L} = \mathtt{unique}(\mathbf{CH}_k^a \cup \mathbf{CH}_l^b)$. Then we take all words from vocabulary $\mathbf{L}$: $w_i \in \mathbf{L}$ and look for maximal similarities with words from chunks a and b, respectively. This way we get vectors $\vec{m}$ and $\vec{n}$ containing maximal similarities of chunk words and words from the combined vocabulary:

$$\begin{aligned} m_i &= \max_{j:1 \leq j \leq |\mathbf{CH}_k^a|} \mathtt{sim}(w_i, w_j) : \quad \forall w_i \in \mathbf{L} \\ n_i &= \max_{j:1 \leq j \leq |\mathbf{CH}_l^b|} \mathtt{sim}(w_i, w_j) : \quad \forall w_i \in \mathbf{L} \end{aligned} \tag{1}$$

where m_i and n_i are elements of vectors $\vec{m}$ and $\vec{n}$. The similarity of words is given by the Word2Vec semantic space. We compare the vectors $\vec{m}$ and $\vec{n}$ with cosine distance

Results

Run	Ali	Type	Score	T+S	Rank
1	0.6672	0.6212	0.6248	0.6377	1
3	**0.6708**	0.6296	0.6114	0.6373	2
2	0.6206	0.6013	0.4748	0.5656	12

Table 1: Official system evaluation. of out system with 3 different configurations

Conclusion

In the overall comparison of the 'Gold standard chunk scenario', our supervised system took the first and second places among other 19 systems from 10 international teams (see Table 1). Our unsupervised system placed in the middle. SemEval is a highly respected workshop in the NLP field.

Acknowledgement

This publication was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports and by Grant No. SGS-2016-018 Data and Software Engineering for Advanced Applications.

References

Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. on Knowl. and Data Eng.*, 18(8):1138–1150, August 2006.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.