Using Word Semantics on Entity Names for Correspondence Set Generation

Rafael Vieira[1] and Kate Revoredo[2]

[1,2]Federal University of the State of Rio de Janeiro (UNIRIO), Brazil,
[1]katerevoredo@uniriotec.br, [2]rvieira.research@gmail.com.br

1 Introduction

On ontology Matching, many works make use of word semantics to align the ontologies. One commonly used resource is WordNet[4][5], which groups words that share the same meaning together. Thesaurus and lexicons like WordNet indeed provide rich semantic information but require large amounts of human effort to be created and maintained.

Vector space representations of word semantics are a family of language models that associate words with vectors in a semantic space, where each dimension represents a component of the meaning of words[2][1][3]. The semantic similarity of words is exploited by these methods, providing vectors close in space when their related words are close in meaning. These vectors are usually calculated by a learning algorithm on large corpora like Wikipedia and then used to evaluate the similarity between two words.

In this work, we exploit the word-word similarities in the GloVe model as external resources for Ontology Matching. The hypothesis is that two entities can be matched based on the words in their names using the word-word similarity provided by the model. We built a prototype and evaluated its performance against the baselines from OAEI.

2 Prototype

To build the simplest prototype, we used pre-trained vectors[1] from GloVe and two ontologies O_1 and O_2. Then, each entity e defined in O_1 or O_2 is associated with one vector $\overrightarrow{v_e} = (a_1, \ldots, a_n)$, based on its name, where each component a_i represents the semantic dimension of words that have related meaning. In case entity e has a compound name, we average the vectors of each word in its name, and set the resulting vector as $\overrightarrow{v_e}$.

To generate a correspondence between two entities e_1 and e_2, from O_1 and O_2 respectively, we calculate the cosine similarity on vectors $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$, associated with e_1 and e_2, respectively. If the value of cosine similarity is above a lower bound, we continue with this correspondence, otherwise, it is discarded. This lower bound was empirically set to 0.7 as this value showed the better results.

[1] Obtained at http://nlp.stanford.edu/data/glove.6B.zip

After doing this procedure for all entity pairs, we have the complete alignment. Finally, we compare this alignment with the baseline alignments edna(edit distance based) and StringEquiv(string equivalence based) from OAEI 2016 on the conference and benchmark data sets. The results are presented in table 1.

Dataset (method)	Precision	Recall	F_1-measure
Conference (edna)	0.74	0.45	0.56
Conference (StringEquiv)	0.76	0.41	0.53
Conference (Prototype)	0.71	0.45	0.54
Benchmark (edna)	0.35	0.51	0.41
Benchmark (Prototype)	0.72	0.26	0.34

Table 1. Comparison between the prototype and baselines of each data set

The prototype obtained low recall on both data sets. The majority of errors on the benchmark data set were on tests with random entity names, resulting in the low recall. This is expected since our method uses only this source of information to gather the entity semantics and then generate correspondences.

On the conference data set, the prototype performed between the two baselines. Many words from entity names were not in the vocabulary of the vectors, and were assigned the vector $\vec{0}$, which contributes to the average recall.

3 Conclusion

These results are not ground-breaking, but also promising. Furthermore, given the simplicity of the prototype, there are many places where it can be improved. For example, in a future experiment, we should train our own vectors and fine tune the hyperparameters of the model. We believe that these improvements may provide increased performance and lead to further research in the area.

References

1. Pennington, J., Socher, R. Manning, C. D.: GloVe: Global Vectors for Word Representation. Empirical Methods in Natural Language Processing (EMNLP), 1532–1543 (2014)
2. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space Computing Research Repository (CoRR), abs-1301-3781 (2013)
3. Gabrilovich, E., Markovitch, S.: Wikipedia-based Semantic Interpretation for Natural Language Processing J. Artif. Intell. Res., 34, 443–498 (2009)
4. He, W., Yang, X., Huang, D.: A Hybrid Approach for Measuring Semantic Similarity between Ontologies Based on WordNet Knowledge Science, Engineering and Management - 5th International Conference, 68–78 (2011)
5. Lin, F., Sandkuhl, K.: A Survey of Exploiting WordNet in Ontology Matching. Artificial Intelligence in Theory and Practice II, 43, 341–350 (2008)