SHORT TERM SCIENTIFIC MISSION - SCIENTIFIC REPORT

Word Representation Fusion with Global Context Similarities for Semantic Search

Navid Rekabsaz

Correspondence:
rekabsaz@ifs.tuwien.ac.at
Information Management and Preservation Lab,
TU WIEN,
Favoritenstrasse 9 HD0107,
1040 Vienna, Austria
Phone: +43 (1) 58801-188317

STSM details

COST Action	IC1302 - Semantic keyword-based search on structured data sources (KEYSTONE)
STSM Title	Word Representation Fusion with Global Context Similarities for Semantic Search
STSM dates	from 03-Sep-2017 to 16-Sep-2017
Applicant	Navid Rekabsaz
Applicant's institution	Information Management and Preservation Lab (IMP), TU WIEN
Host	Dr. Carsten Eickhoff
Host institution	Data Analytics lab, ETH Zürich

1 Purpose of the STSM

Recent studies show the benefit of exploiting word embedding models such as word2vec [1] for information retrieval [2, 3]. Despite the relative improvements in document retrieval tasks, as shown in our recent work [4], there is still high potential in improving retrieval performance, especially by exploiting other resources of term relatedness.

The purpose of the STSM is to collaboratively study the retrieval effectiveness of a novel word representation which benefits from combining various resources of term-term relatedness. The mainstream word embedding models (i.e. word2vec [1], GloVe [5]) are based on short window context. On the other side, other statistical methods such as Latent Semantic Analysis/Indexing (LSA/LSI) [6], probabilistic Latent Semantic analysis (pLSA) [7], Latent Dirichlet Allocation [8], and Pseudo Relevance Feedback (PRF) capture the relatedness between terms by exploiting wider contexts (paragraphs, documents) of the terms. We refer to the relation between terms using these methods as global context relatedness. We argue that combining window-context with global-context relatedness methods can improve the effectiveness of keyword-based search. During the visit, we aim to study the related methodologies and reach to a solid research framework through brainstorming and eventually develop the initial experiments of the work.

The targeted research questions are two-fold:

- What are the effective methods of capturing global-context similarity in information retrieval?
- How to combine/fuse a window-context word embedding with the term relatedness, achieved from a global-context representation?

Beside working throw these specific research questions, the long-term collaboration of the two labs is also indeed an essential aim of the visit.

2 Work Carried Out During the STSM

The work during the STSM consists of literature review, brainstorming on the research ideas, implementation of a global context similarity method, and finally code analysis of the selected fusion method.

The related work to our study consist of three main areas:

The first area regards the methods to exploit word embedding's term similarities for semantic search. Rekabsaz et al. [3] addresses this issue by introducing the Extended/Generalized Translation Models which expands various probabilistic relevance framework models by using the term similarities. In another recent study, Guo et al. [9] introduce DRMM, a neural network model that uses the histogram of term similarities between a query and documents. We decide to implement both the models to observe the effect of changing word embeddings on document retrieval.

The second area is about the approaches to combine word embedding models (with window-context) with the information of global-context models. A close research area to our work is the studies on enriching word embedding with external resources and in particular WordNet. Several studies in this area suggest different variations of joint learning specially on the word2vec Skipgram model [10, 11]. These methods exploit a list of *connected terms* to a particular term, achieved from the external resource. In the case of WordNet, the connected terms are the ones in the same synset as the term. In contrast, Faruaui et al. [12] introduces a light post-processing method that can be applied on any existing embedding to incorporate the information of an external resource into the embedding. They refer to the method as *embedding retrofitting*. The retrofitting method defines an objective function and optimizes it by making the embeddings of the connected terms closer while maintaining the least possible changes from the original embeddings. In comparison to the other approaches, retrofitting is more efficient since it does not need complete representation learning each time from scratch. We therefore decide to use this method in our experiments.

The last direction is the method to capture the term-term similarity on the global level. As mentioned before, some immediate approaches are LSA/LSI, pLSA, and LDA. Another established method in the IR community is the Pseudo Relevance Feedback (PRF). The PRF method is mainly used for many-to-many term relatedness such that the similarity of multi terms (a query) to a set of terms. However, PRF can simply be used for capturing one-to-many term relatedness by passing queries with only one word. In our research, we decide to evaluate the effectiveness of the LSI and PRF (with one-to-many term relatedness) methods.

Considering the mentioned ideas, I created word2vec and LSI models on a recent corpus of Wikipedia. I then implemented the one-to-many term relatedness PRF method and executed it for every word in Wikipedia (with collection frequency of higher than 100) to find the top-200 most related terms. Some observation of the results are discussed in the next section. Finally, I studied the code of the retrofitting method[1], and start to change it for our purpose.

[1] `https://github.com/mfaruqui/retrofitting`

Table 1: Top 10 related terms to the *asthma* and *Argentina* using the PRF method and the Cosine similarity between the vectors of a word2vec model.

asthma		Argentina	
PRF	word2vec	PRF	wor2vec
ISAAC	bronchitis	Riquelme	Uruguay
symptoms	allergies	Argentine	Paraguay
breathing	bronchial	cup	Chile
pneumoniae	rheumatic	Boca	Argentine
airway	allergy	team	Argentinian
pneumonia	COPD	Chile	Buenos_Aires
atopic	arthritis	Pumas	Brazil
disease	chronic	match	Bolivia
respiratory	rheumatoid	Maradona	Colombia
exposure	diabetes	bondholders	Salta

3 Main Results

The main results of the visit are the analysis of the research questions, brainstorming on the possible directions, achieving a framework for the experiments design, and finally initial experiments and comparison on term relatedness methods. While the research framework is discussed in the previous section, in the following we report the results of the initial experiments.

Table 1 shows the top 10 related terms to the *asthma* and *Argentina* using the PRF approach (global context relatedness) and the Cosine similarity of the vector representations of the term from the word2vec model (window context relatedness). The results show interesting difference between the related terms, achieved from the methods. The related terms using the word2vec model contains several cases of topic shifting, e.g. asthma is related to bronchitis, rheumatic, and diabetes; and Argentina is related to Uruguay, Paraguay, Chile, and Brazil. On the other hand, the PRF method retrieves terms which are in a similar topic but has a much broader perspective, e.g. asthma is related to breathing, airway, and disease; and Argentina to Riquelme, match, and Maradona. However when by considering the terms that both the approaches see as related terms, we could provide more relevant semantically related terms for retrieval tasks.

4 Future Collaborations

Encouraged by the results discussed in the previous section, we will continue the experiments by focusing on the retrofitting of word embeddings with the term similarities of the global context methods. We continue the collaboration between the IMP and Data Analytics labs on this research project, by targeting a submission for the SIGIR 2018 conference.

Besides, we pursue collaborative research on the areas of interest between the labs such as neural information retrieval models, word and document representation learning, and health information retrieval.

References

1. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
2. Rekabsaz, N., Bierig, R., Lupu, M., Hanbury, A.: Toward optimized multimodal concept indexing. In: Proceedings of the 1st International KEYSTONE Conference (IKC), pp. 141–152 (2015). Springer
3. Rekabsaz, N., Lupu, M., Hanbury, A., Zuccon, G.: Generalizing translation models in the probabilistic relevance framework. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 711–720 (2016). ACM

4. Rekabsaz, N., Lupu, M., Hanbury, A., Zamani, H.: Word embedding causes topic shifting; exploit global context! In: Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR) (2017)
5. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
6. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391 (1990)
7. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57 (1999). ACM
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
9. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A Deep Relevance Matching Model for Ad-hoc Retrieval. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16 (2016)
10. Fried, D., Duh, K.: Incorporating both distributional and relational semantics in word representations. arXiv preprint arXiv:1412.4369 (2014)
11. Kiela, D., Hill, F., Clark, S.: Specializing word embeddings for similarity or relatedness. In: EMNLP, pp. 2044–2048 (2015)
12. Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A.: Retrofitting word vectors to semantic lexicons. (2015). Association for Computational Linguistics