# Question Similarity in Community Question Answering: A Systematic Exploration of Preprocessing Methods and Models

Florian Kunneman[1,3], Thiago Castro Ferreira[2], Antal van den Bosch[3,4], and Emiel Krahmer[2]

[1] Vrije Universiteit Amsterdam, The Netherlands
[2] Tilburg University, The Netherlands
[3] Radboud University, The Netherlands
[4] KNAW Meertens Institute, The Netherlands

Community Question Answering forums are popular among Internet users, and a basic problem they encounter is trying to find out if their question has already been posed before. To address this issue, Natural Language Processing researchers have developed methods to automatically detect question-similarity, which was one of the shared tasks in SemEval[3]. The best performing systems for this task made use of Syntactic Tree Kernels (SPTK) [2] or the SoftCosine metric [1]. However, it remains unclear why these methods seem to work, whether their performance can be improved by better preprocessing methods and similarity metrics and what kinds of errors they (and other methods) make. In this study, we therefore systematically combine and compare these two approaches with the more traditional BM25 [4] and translation-based models (TRLM) [5]. Moreover, we analyze the impact of preprocessing steps (lowercasing, suppression of punctuation and stop words removal) and word meaning similarity based on different distributions (word translation probability, Word2Vec, fastText and ELMo) on the performance of the task. We conduct an error analysis to gain insight into the differences in performance between the system set-ups.[1][2]

We applied the aforementioned alternated set-ups to two benchmark datasets: Qatar Living[3] and Quora[4]. We added two ensemble settings to test whether a combination of approaches can lead to an improved performance. In Table 1 we

[1] The implementation is made publicly available: https://github.com/fkunneman/DiscoSumo/tree/master/ranlp

[2] Original paper presented at RANLP 2019: Kunneman, F., Castro Ferreira, T., Krahmer, E. & van den Bosch, A. (2019). Question Similarity in Community Question Answering: A Systematic Exploration of Preprocessing Methods and Models. In Proceedings of the International Conference Recent Advances in Natural Language Processing (pp. 593-601).

[3] http://alt.qcri.org/semeval2017/task3/index.php?id=data-and-tools

[4] http://qim.fs.quoracdn.net/quora_duplicate_questions.tsv

| **Preproc.** | **BM25** | **TRLM** | **SoftCosine** | **SPTK** | **Ensemble** | **EnsSPTK** |
|---|---|---|---|---|---|---|
| L.S.P. | 68.80 | 68.43 | **72.75** | - | **71.62** | **72.40** |
| L.S. | 67.31 | 63.25 | 69.15 | - | 69.50 | 71.29 |
| L.P. | **69.95** | 68.42 | 65.33 | - | 68.70 | 69.16 |
| S.P. | 66.03 | **68.65** | 68.56 | - | 68.67 | 70.37 |
| L. | 67.07 | 66.42 | 63.68 | 54.34 | 67.04 | 67.41 |
| S. | 63.77 | 64.53 | 67.01 | - | 67.85 | 68.36 |
| P. | 65.05 | 64.38 | 60.04 | - | 65.31 | 66.66 |
| - | 63.52 | 64.95 | 60.66 | **54.44** | 63.08 | 64.31 |

| **Metric** | **BM25** | **TRLM** | **SoftCosine** | **SPTK** | **Ensemble** | **EnsSPTK** |
|---|---|---|---|---|---|---|
| Translation | - | 68.43 | 70.75 | 48.10 | 70.80 | 70.80 |
| Word2Vec | - | **72.90** | 72.75 | 54.44 | 71.40 | 72.64 |
| fastText | - | 70.93 | 71.07 | 53.49 | 71.92 | 71.92 |
| Word2Vec+ELMo | - | 71.41 | **73.89** | **54.78** | **73.90** | **74.63** |
| fastText+ELMo | - | 70.56 | 73.43 | 54.77 | 73.73 | 73.73 |

**Table 1.** MAP results on the different preprocessing and word-relation metric conditions in the development set. *L.*, *S.* and *P.* denote lowercase, stop words removal and punctuation suppression methods respectively.

highlight the outcomes on the development set of the Qatar Living dataset. The task here is to rerank a pre-selection of ten questions that are either similar to a given target or not, where the most similar questions should be ranked highest. This is evaluated by the Mean Average Precision (MAP): the average precision when measuring the precision at each rank.

Our findings show that lowercasing the input and removing both punctuation and stopwords yields the most robust outcomes, especially for the SoftCosine metric. In addition, representing the meaning of words by means of Word2Vec combined with the top layer of ELMo is the most beneficial word similarity implementation. The error analysis showed that the BM25 model is most stable across different preprocessing metrics, while the SoftCosine model mostly profits from preprocessing. Given the semantic matching that is done as part of SoftCosine and is absent in BM25, we can infer that preprocessing is an important prerequisite for effectively ranking question pairs based on semantic links.

Most of our experimentation was conducted on the SemEval dataset, in which similarity between questions is labeled. We also showed that adjusting preprocessing and word similarity settings led to better results in the task of identifying question duplicates, in the Quora dataset. More research is needed to see whether the patterns that we find are dataset-independent.

## References

1. Charlet, D., Damnati, G.: Simbow at semeval-2017 task 3: Soft-cosine semantic similarity between questions for community question answering. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 315–319 (2017)

2. Filice, S., Croce, D., Moschitti, A., Basili, R.: Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 1116–1123 (2016)
3. Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, a.A., Glass, J., Randeree, B.: Semeval-2016 task 3: Community question answering. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 525–545. Association for Computational Linguistics, San Diego, California (June 2016), `http://www.aclweb.org/anthology/S16-1083`
4. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2009)
5. Xue, X., Jeon, J., Croft, W.B.: Retrieval models for question and answer archives. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 475–482. ACM (2008)