SimBow at SemEval-2017 Task 3: Soft-Cosine Semantic Similarity between Questions for Community Question Answering

Delphine Charlet and Géraldine Damnati
Orange Labs
Lannion, France
delphine.charlet,geraldine.damnati@orange.com

Abstract

This paper describes the SimBow system submitted at SemEval2017-Task3, for the question-question similarity subtask B. The proposed approach is a supervised combination of different unsupervised textual similarities. These textual similarities rely on the introduction of a relation matrix in the classical cosine similarity between bag-of-words, so as to get a soft-cosine that takes into account relations between words. According to the type of relation matrix embedded in the soft-cosine, semantic or lexical relations can be considered. Our system ranked first among the official submissions of subtask B.

1 Introduction

Social networks enable people to post questions, and to interact with other people to obtain relevant answers. The popularity of forums show that they are able to propose reliable answers. Due to this tremendous popularity, forums are growing fast, and the first reflex for an internet user is to check with his favorite search engine if a similar question has already been posted.Community Question Answering at SemEval focuses on this task, with 3 different subtasks. SubtaskA (resp. subtaskC) aims at re-ranking the comments of one original question (resp. the comments of a set of 10 related questions), regarding the relevancy to the original questions. SubtaskB aims at re-ranking 10 related questions proposed by a search engine, regarding the relevancy to the original question. Subtasks A and C are question-answering tasks. SubtaskB can be viewed as a pure semantic textual similarity task applied on community questions, with noisy user-generated texts, making it different from SemEval-Task1 (Agirre et al., 2016), which focuses on semantic similarity between short well-formed sentences.

In this paper, we only focus on subtaskB, with the purpose of developing semantic textual similarity measures for such noisy texts. Question-question similarity appeared in SemEval2016 (Nakov et al., 2016), and is pursued in SemEval2017 (Nakov et al., 2017). The approaches explored last year were mostly supervised fusion of different similarity measures, some being unsupervised, others supervised. Among the unsupervised measures, many were based on overlap count between components (from n-grams of words or characters to knowledge-based components such as named entities, frame representations, knowledge graphs, e.g. (Franco-Salvador et al., 2016)...). Much attention was also paid for the use of word embeddings (e.g. (Mihaylov and Nakov, 2016)), with question-level averaged vectors used directly with a cosine similarity or as input of a neural classifier. Finally, fusion was often performed with SVMs (Filice et al., 2016)

Our motivation in this work was slightly different: we considered that forum data were too noisy to get reliable outputs from linguistic analysis and we wanted to focus on core textual semantic similarity. Hence, we avoided using any metadata analysis (such as user profile...) to get results that could easily generalize to other similarity tasks.Thus, we explore unsupervised similarity measures, with no external resources, hardly any linguistic processing (except a list of stopwords), relying only on the availability of sufficient unannotated corpora representative of the data. And we fuse them in a robust and simple supervised framework (logistic regression).

The rest of the paper is organized as follows: in section 2, the core unsupervised similarity measure is presented, the submitted systems are described in section 3, and section 4 presents results.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 315–319,
Vancouver, Canada, August 3 - 4, 2017.

2 Soft-Cosine Similarity Measure

In a classical bag-of-words approach, texts are represented by a vector of TF-IDF coefficients of size N, N being the number of different words occurring in the texts. Computing a cosine similarity between 2 vectors is directly related to the amount of words which are in common in both texts.

$$cos(X,Y) = \frac{X^t.Y}{\sqrt{X^t.X}\sqrt{Y^t.Y}} \text{whith } X^t.Y = \sum_{i=1}^{n} x_i y_i \quad (1)$$

When there are no words in common between texts X and Y (*i.e.* no index i for which both x_i and y_i are not equal to zero), cosine similarity is null. However, even with no words in common, texts can be semantically related when the words are themselves semantically related. Hence we propose to take into account word-level relations by introducing in the cosine similarity formula a relation matrix M, as suggested in equation 2.

$$cos_M(X,Y) = \frac{X^t.M.Y}{\sqrt{X^t.M.X}\sqrt{Y^t.M.Y}} \quad (2)$$

$$X^t.M.Y = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i m_{i,j} y_j \quad (3)$$

where M is a matrix whose element $m_{i,j}$ expresses some relation between word i and word j. With such a metric, the similarity between two texts is non null as soon as the texts share related words, even if they have no words in common. Introducing the relation matrix in the denominator normalization factors ensures that the reflexive similarity is 1. If the words are only related with themselves ($m_{i,i} = 1$ and $m_{i,j} = 0\ \forall i,j$ with $i \neq j$), M is the identity matrix and the *soft*-cosine turns out to be the cosine.

We first investigated this modified cosine similarity in the context of topic segmentation of TV Broadcast News (Bouchekif et al., 2016), using semantic relations between words to improve the computation of semantic cohesion between consecutive snippets. Other researchers have also proposed this measure (e.g (Sidorov et al., 2014)) along with the *soft-cosine* denomination, where the matrix was based for instance on Levenshtein distance between n-grams. In this work, we investigate different kinds of word relations that can be used for computing M.

2.1 Semantic relations

Distributed representations of words, such as the `word2vec` approach proposed by (Mikolov et al., 2013) have known a tremendous success recently. They enable to obtain relevant semantic relations between words, based on a simple similarity measure (e.g. cosine) between the vector representations of these words.

In this work, 2 distributed representations of words are computed, using the `word2vec` toolkit, in the cbow configuration: one is estimated on English Wikipedia, and the other is estimated using the unannotated corpus of questions and comments on Qatar-Living forum, distributed in the campaign, which contains 100 millions of words. The vectors dimension is 300 (experiments with various vector dimensions didn't provide any significant difference), and only the words with a minimal frequency of 50 are taken into account.

Once the `word2vec` representations of words are available, M can be computed in different ways. We have explored different variants, and the best results were obtained with the following framework, where v_i stands for the `word2vec` reprsentation of word w_i:

$$m_{i,j} = max(0, cosine(v_i, v_j))^2 \quad (4)$$

Grounding to 0 is motivated by the observation that negative *cosine* between words are hard to interpret, and often irrelevant. Squaring is applied to emphasize the dynamics of the semantic relations: insisting more on strong semantic relations, and flattening weak semantic relations. Actually we have observed in several applicative domains that high semantic similarities derived from word embedding are more significant than low similarities.

2.2 Edit-distance based relations

Using a Levenshtein distance between words, an edit relation between words can be computed: it enables to cope, for instance, with little typographic errors which are frequent in social user-generated corpora such as Qatar Living forum. It is defined as $m_{i,i} = 1$ and for $i \neq j$:

$$m_{i,j} = \alpha * \left(1 - \frac{Levenshtein(w_i, w_j)}{max(||w_i||, ||w_j||)}\right)^{\beta} \quad (5)$$

$||w||$ is the number of characters of the word, α is a weighting factor relatively to diagonal elements, and β is a factor that enables to emphasize the score dynamics. Experiments on train and dev led to set $\alpha = 1.8$ and $\beta = 5$.

3 System Description

3.1 Data pre-processing

Some basic preprocessing steps are applied on the text: lowercase, suppression of punctuation marks and stopwords, replacing urls and images with the generic terms "_url_" and "_img_". As for the bag of word representation, TF-IDF coefficients are computed in a specific way: TF coefficients are computed in the text under consideration as usual but IDF coefficients are computed from the large unannotated Qatar Living forum corpus.

3.2 Supervised combination of unsupervised similarities

For a given pair of texts to compare, 3 textual similarity measures are considered: cos_{Mrel} (soft-cosine with semantic relations), cos_{Mlev} (soft-cosine with Levenshtein distance), $wavg\_word2vec$ (cosine between weighted averaged `word2vec`). Soft-cosine measures are computed as explained in section 2. For the latter, weights are given by the TF-IDF coefficients computed as described in section3.1.

Each original question contains a *subject* and a *body*. Each related question additionally contains a thread of *comments*. We have considered several variants for text selection: (*subject*, *body*, *subject+body*, or *comments*). The 3 textual similarities are then computed between every possible 12 text pairings (3 possible texts for original questions $\times$ 4 possible texts for relative questions), constituting the set of 36 potential features. We also include in this set the IR system reciprocal rank rrk. Logistic regressions, combining these features, is then trained on the "train-part1" set of 1999 paired texts of SemEval2016. Thus, we evaluate all possible subsets of features among the set of 37 potential features, and we keep for our primary submission the one that gave the best result on average on dev and test2016. *contrastive1* was chosen as the best candidate including only soft-cosine metrics and rrk and *contrastive2* was chosen as the best candidate system with the lowest amount of features.

4 Evaluation

In this section, we present detailed evaluations of Task3/subtaskB. Given a new question (aka original question), the task consists in reranking the 10 questions (aka related questions) proposed by a search engine. A precise description of the corpus and metrics can be found in Task3 description paper (Nakov et al., 2017). Results are presented with the MAP evaluation measure, on 3 corpora: dev (50 original questions $\times$ 10 related questions), test2016 (70 original questions $\times$ 10 related questions) and test2017 (88 original questions $\times$ 10 related questions).

It is worth noticing that the MAP scorer used in this campaign is sensitive to the amount of original questions which don't have any relevant related questions in the gold labels. In fact, these questions always account for a precision of 0 in the MAP scoring. Hence, an Oracle evaluation, giving a score of 1 to all related questions labeled as "true", and a score of 0 to all related questions labeled as "false" in the gold labels, doesn't provide a 100% MAP but an Oracle MAP which corresponds to the proportion of original questions that have at least 1 relevant related question. Hence the upper bound of MAP performances is 86.00% for dev, 88.57% for test2016, and only 67.05% for test2017 (29 original questions without any relevant related question out of 88). Another difference between test2016 and test2017 is the average number of "true" labels for questions that have at least one relevant associated question (3.7 for test2016 and 2.7 for test2017). On the overall test2017 is more difficult for the Task.

4.1 Unsupervised textual similarity measures

Table 1 presents the MAP results obtained for different unsupervised textual similarities. Here, the focus is made on unsupervised textual similarity measure, and we only present results for the *subject+body* configuration for both the original and related questions. Performances of the Information Retrieval system (IR), and of the best system submitted at SemEval2016 (Franco-Salvador et al., 2016) are reported for comparison purpose.

similarity	**dev**	**test** 2016	**test** 2017
IR	71.35	74.75	41.85
best SemEval2016	-	77.33	-
baseline_token_cos	62.22	68.54	40.88
baseline_pp_cos	67.49	71.05	42.80
baseline_pp_cos_tfidf	69.41	75.53	44.37
cos_{Mrel} relations WP	72.25	77.11	45.38
cos_{Mrel} relations QL	**75.24**	**77.96**	45.27
cos_{Mlev} Levenshtein	70.02	76.34	46.10
wavg-word2vec on QL	73.31	75.77	**46.99**

Table 1: MAP results for unsupervised textual similarity measures

As a baseline, we use the *baseline_token_cos* defined in SemEval2015-Task2 (Agirre et al., 2015), for semantic textual similarity between sentences. It is a simple cosine similarity between bag-of-tokens binary vectors (a token is a non-white-space sequence between 2 white spaces, and weights are 1 or 0). Performances of *baseline_pp_cos*, which is also a cosine of binary vectors but obtained after the pre-processing step show the importance of suitable pre-processing. *baseline_pp_cos_tfidf* show the influence of appropriate term weighting over simple binary coefficients. Next results reveal significant improvements when introducing a relation matrix M in the soft-cosine metric (cos_M). When M contains semantic relations, a significant difference is observed on dev, between relations estimated on a general corpus WP (Wikipedia, 2.7 Bwords) and on a specialized corpus QL (Qatar Living, 100 Mwords). The difference is much lower for test2016, and even negative for test2017. On the contrary, the Levenshtein-based M matrix performs best on test2017, whereas its gain is only marginal for dev and test2016. In all cases, introducing a carefully chosen relation matrix M in the cosine-based similarity measure improves performances. Finally, the cosine between TF-IDF weighted average `word2vec` is less effective on dev and test2016, but performs well on test2017.

It is worth noticing that the mere cos_{Mrel} soft-cosine on QL would have won the 2016 challenge.

4.2 Evaluation of supervised combination

Table 2 presents the MAP obtained for different supervised combinations of similarity measures.

First, for a given unsupervised textual similarity measure, all the possible combinations of paired texts are evaluated, and we give the result of the subset which gives the best performance on average on dev, test2016, and test2017. Interestingly, it is the same combination of paired texts which performs best for the 3 textual similarity measure: similarity between *subject+body* for both questions and *subject+body* for the original question and *comments* for the related question. This last pairing performs poorly alone but is interesting in combination with the first one.

Then we report the results of the submitted systems to the official evaluation. As can be seen in Table 2, *contrastive2* was more robust to the more difficult conditions of test2017. Additionally, as the IR performs really worse in test2017, we re-trained the systems excluding rrk from features. Actually, if rrk was helpful for both dev and test2016 corpora, we can see that removing rrk provides better results on test2017, yielding a maximum MAP score of 48.38. This performance is obtained with the following set of similarities: cos_{Mrel} between $subject+body$, cos_{Mlev} between $subject + body$ and $subject$ of the original question and $body$ of the relative question, and $wavg - w2v$ between $subject + body$ and between $subject + body$ of the original question and $comments$ of the relative question.

similarity	**dev**	**test** 2016	**test** 2017
IR	71.35	74.75	41.85
best SemEval2016	-	77.33	-
text combination			
cos_{Mrel} relations QL	75.76	78.76	46.67
cos_{Mlev} Levenshtein	72.26	78.19	47.48
wavg-word2vec on QL	75.91	76.70	47.40
submissions			
primary	**77.30**	**79.77**	**47.22**
contrastive1	77.04	79.12	46.84
contrastive2	77.30	79.43	**47.87**
removing rrk			
primary$-rrk$	76.71	78.61	47.68
contrastive1$-rrk$	76.09	78.90	46.96
contrastive2$-rrk$	76.73	78.97	**48.38**

Table 2: MAP results for supervised combination of textual similarity measures

5 Conclusion

In this work, we have explored a modified version of the cosine similarity between bag-of-words representation of texts. In this so-called *soft-cosine* similarity, a relation matrix M is embedded, allowing relations between words to be taken into account. The computation of M is unsupervised, and can be derived from distributed representations of words. *soft-cosine* performed well at SemEval-Taks3 question-question similarity subtaskB. A simple supervised logistic regression combination of different unsupervised similarity measures over different text selection strategies ranked first at the official evaluation. In the future, we plan to pursue the work on *soft-cosine* in two directions: including other relations between words, for instance using semantic role labeling, and studying how this matrix M, efficiently initialized in an unsupervised way, could be further trained for specific tasks.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, pages 252–263. http://www.aclweb.org/anthology/S15-2045.

Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*. pages 497–511.

Abdessalam Bouchekif, Géraldine Damnati, Delphine Charlet, Nathalie Camelin, and Yannick Estève. 2016. Title assignment for automatic topic segments in TV broadcast news. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. pages 6100–6104.

Simone Filice, Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2016. Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers. In *SemEval@NAACL-HLT*.

Marc Franco-Salvador, Sudipta Kar, Thamar Solorio, and Paolo Rosso. 2016. UH-PRHLT at semeval-2016 task 3: Combining lexical and semantic-based features for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*. pages 814–821.

Todor Mihaylov and Preslav Nakov. 2016. Semanticz at semeval-2016 task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. *Proceedings of SemEval* pages 879–886.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Vancouver, Canada, SemEval '17.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. Semeval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 525–545. http://www.aclweb.org/anthology/S16-1083.

Grigori Sidorov, Alexander F. Gelbukh, Helena Gómez-Adorno, and David Pinto. 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas* 18(3). http://cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2043.