UH-PRHLT at SemEval-2016 Task 3:
Combining Lexical and Semantic-based Features for Community Question Answering

Marc Franco-Salvador[1], **Sudipta Kar**[2], **Thamar Solorio**[2], and **Paolo Rosso**[1]

[1] Pattern Recognition and Human Language Technology (PRHLT) research center
Universitat Politècnica de València, Spain
[2] Department of Computer Science
University of Houston, United States
`mfranco@prhlt.upv.es, skar3@uh.edu,`
`tsolorio@uh.edu, prosso@prhlt.upv.es`

Abstract

In this work we describe the system built for the three English subtasks of the SemEval 2016 Task 3 by the Department of Computer Science of the University of Houston (UH) and the Pattern Recognition and Human Language Technology (PRHLT) research center - Universitat Politècnica de València: UH-PRHLT. Our system represents instances by using both lexical and semantic-based similarity measures between text pairs. Our semantic features include the use of distributed representations of words, knowledge graphs generated with the BabelNet multilingual semantic network, and the FrameNet lexical database. Experimental results outperform the random and Google search engine baselines in the three English subtasks. Our approach obtained the highest results of subtask B compared to the other task participants.

1 Introduction

The key role that the Internet plays today for our society benefited the dawn of thousands of new Web social activities. Among those, forums emerged with special relevance following the paradigm of the Community Question Answering (CQA). These type of social networks allow people to post a question to other users of that community. The usage is simple, without much restrictions, and infrequently moderated. The popularity of CQA is a strong indicator that users receive some good and valuable answers.

However, there are several issues related to that type of community. First is the large amount of answers received that makes it difficult and time-consuming for users to search and distinguish the good ones. This is exacerbated with the amount of noise that these questions contain. It is not uncommon to have wrong or misguiding answers that produce more unrelated answers, discussions and sub-threads. Finally, there is a lot of redundancy, questions may be repeated or closely related to previously asked questions.

Details of the SemEval 2016 Task 3 on CQA can be found in the overview paper (Nakov et al., 2016). In this work we evaluate the three English-related Task 3 subtasks on CQA. We first represent each instance to rank — question versus (vs.) comments, question vs. related questions, or question vs. comments of related questions — with a set of similarities computed at two different levels: lexical and semantic. This representation allows us to estimate the relatedness between text pairs in terms of what is explicitly stated and what it means. Our lexical similarities employ representations such as word and character n-grams, and bag-of-words (BOW). The semantic similarities include the use of distributed word bidirectional alignments, distributed representations of text, knowledge graphs, and frames from the FrameNet lexical database (Baker et al., 1998). This type of dual representations have been successfully employed for question answering by the highest performing system in the previous edition of this Se-

mEval task (Tran et al., 2015). Other Natural Language Processing (NLP) tasks such as cross-language document retrieval and categorization also benefited from similar representations (Franco-Salvador et al., 2014). In this task, if the question or comment includes multiple text fields, e.g. body and subject, similarities are estimated using all possible combinations (see Section 3.2). Finally, the ranking of instances is performed using a state-of-the-art machine-learned ranking algorithm: SVM_{rank}.

2 Related Work

Automatic question answering has been a popular interest of research in NLP from the beginning of the Internet to more recently where voice interfaces have been incorporated (Rosso et al., 2012). The use of BOW representations allowed to correctly answer 60% of the questions of the first large-scale question answering evaluation at the TREC-8 Question Answering track (Voorhees, 1999). More complex systems used inference rules to connect expressions between questions and answers (Lin and Pantel, 2001). Similarly, Ravichandran and Hovy (2002) employed bootstrapping to generate surface text patterns in order to successfully answer questions. Other works such as Buscaldi et al. (2010) are based on the redundancy of n-grams in order to find one or more text fragments that include tokens of the original question and the answer. Jeon et al. (2005) studied the semantic relatedness between texts for question answering. They used translation obfuscation to paraphrase the text and to detect which terms are closer in context. Probabilistic topic models have been also useful for detecting the semantics in this task. Celikyilmaz et al. (2010) used Latent Dirichlet Allocation (LDA) (Blei et al., 2003) for representing questions by means of latent topics.

The previous edition of the SemEval CQA task included two English subtasks (Nakov et al., 2015). The first one was focused on classifying answers as *good*, *bad*, or *potentially relevant* with respect to one question. The second subtask answered a question with *yes*, *no*, or *unsure* based on the list of all answers. In addition, the

first subtask was also available in Arabic. Several teams experimented with complex solutions that included meta-learning, external resources, and linguistic features such as syntactic relations and distributed word representations. Similarly to our work, the highest performing approach employed a combination of lexical and semantic-based similarity measures (Tran et al., 2015). Its semantic features included the use of probabilistic topic models, translation obfuscation-based alignments, and pre-computed distributed representations of words both generated with the word2vec[1] and GloVe[2] toolkits. Their lexical features included BOW, word alignments, and noun matching. They employed a regression model for classification. Another interesting approach, Hou et al. (2015), included textual features — word lengths and punctuation — in addition to syntactical-based features — Part-of-Speech (PoS) tags.

In this work we aim at differentiating from the other approaches by enhancing our ranking model with new similarity measures. These include the use of knowledge graphs obtained using the largest multilingual semantic network — BabelNet — frames from the FrameNet lexical database, and bidirectional distributed word alignments.

3 Lexical and Semantic-based Community Question Answering

In this section we detailed the system that we designed for this CQA task. First in Section 3.1 we described our set of lexical features and semantic-based ones. Next, in Section 3.2 we detail the specific adaptation that we employed for each subtask and the ranking algorithm that we used. We note that all our features are similarity scores obtained with different text similarity measures. More details and examples can be found in their respective papers.

3.1 Feature Description

Our system exploits both the verbatim and the contextual similarities between texts, i.e., questions and comments. In Section 3.1.1 we

---

[1] https://code.google.com/archive/p/word2vec/
[2] http://nlp.stanford.edu/projects/glove/

detailed our lexical and in Section 3.1.2 our semantic-based features.

3.1.1 Lexical Features

The lexical features that we employed are the following:

- **Cosine Similarity.** We used cosine similarity to measure lexical similarity between two text snippets. We calculated cosine similarity based on word n-grams(n=1,2), character 3-grams and tf-idf (Salton and McGill, 1986) scores of words.

- **Word Overlap.** We used the count of common words between two texts. This count was normalized by the length.

- **Noun Overlap.** We used NLTK[3] to part-of-speech tag the text and computed the normalized count of overlapping nouns in two texts as a similarity measure.

- **N-gram Overlap.** We computed the normalized count of common n-grams(n=1,2,3) between two texts.

3.1.2 Semantic Features

The semantic features that we employed are the following:

- **Distributed representations of texts.** We used the continuous Skip-gram model (Mikolov et al., 2013) of the word2vec toolkit to generate distributed representations of the words of the complete English Wikipedia.[4] Next, for each text, e.g. question or comment, we averaged its word vectors in order to have a single representation of its content as this setting has shown good results in other NLP tasks (e.g. for language variety identification (Franco-Salvador et al., 2015a) and discriminating similar languages (Franco-Salvador et al., 2015b)). Finally, the similarity between texts, e.g. question vs. comment, is estimated using the cosine similarity.

- **Distributed word alignments.** The use of word alignment strategies has been employed in the past for textual semantic relatedness (Hassan and Mihalcea, 2011). Tran et al. (2015) employed distributed representations to align the words of the question with the words of the comment. A more recent work introduced the Continuous Word Alignment-based Similarity Analysis (CWASA) (Franco-Salvador et al., 2016a). CWASA uses distributed representations to measure the similarity by double-direction aligning words of texts. In this work we selected as feature the similarity provided by CWASA between questions and comments.

- **Knowledge graphs.** A knowledge graph is a labeled, weighted, and directed graph that expands and relates the concepts belonging to a text. Knowledge Graph Analysis (KGA) (Franco-Salvador et al., 2016b) measures semantic relatedness between texts by means of their knowledge graphs. In this work we used the BabelNet (Navigli and Ponzetto, 2012) multilingual semantic network to generate knowledge graphs from questions and comments, and measured their similarity using KGA.

- **Common frames.** We used Framenet (Baker et al., 1998) to extract the frames associated with the lexical items in the text. For each frame present in the text, we calculated the common lexical items between sentences associated with this frame. The goal is to allow inference of similarity at the level of semantic roles.

As additional feature, for Subtasks A and C we also used the ranking provided by the Google search engine for the questions related to the original questions.

3.2 Data Representation and Ranking

Due to the representation of questions — composed by *subject* and *body* fields — and answers — a *comment* field — we adapted our system for the different English subtasks:

---

[3]http://www.nltk.org/

[4]We used 200-dimensional vectors, context windows of size 10, and 20 negative words for each sample.

- **Subtask A (question-comment similarity ranking):** we used the aforementioned similarity-based features at three levels: question subject vs. comment, question body vs. comment, and full question vs. comment.

- **Subtask B (question-related question similarity ranking):** for this subtask we measured the similarities at body, subject, and full question level.

- **Subtask C (question-external comment similarity ranking):** we employed all the features of Subtasks A and B, plus the similarities of the original question — subject, body, and full levels — with the related question comments.

In order to rank the questions and comments, we selected a variant of Support Vector Machines (SVM) (Hearst et al., 1998) optimized for ranking problems: SVM_{rank} (Joachims, 2002). In our evaluation of Section 4, we call our system as the combination of the acronyms of our member institutions: UH-PRHLT.

Preproscessing steps included stopword removal, lemmatization, and stemming. However, for the distributed representation and knowledge graph-based features we did not employ stemming. These decisions were motivated for performance reasons during our prototyping.

Note that each subtask allows to submit three runs per team: primary, contrastive 1 (contr. 1), and contrastive 2 (contr. 2). We used a linear kernel and optimized the cost factor parameter using Bayesian optimization[5] (Snoek, 2013). Our three runs differ only in the value for that parameter and correspond with the three best — and considerably distant — values. In addition to the ranking, the task requires also to provide with a label for each instance that reflects if the question or comment is relevant to the compared question. For each subtask we optimized a threshold to determine the relevance of each instance that is based on our predicted relevance ranking. In other words, we binarize our ranking.

---

4 Evaluation

This section presents the evaluation of the SemEval 2016 Task 3 on CQA. Details about this task, the datasets, and the three subtasks can be found in the task overview (Nakov et al., 2016). Note that for our system we did not use data from SemEval 2015 CQA as we did not observe gains in performance.

We compared the results of our approach with those provided by the random baseline and the Google search engine when ranking the questions and comments.[6] The official measure of the task is the Mean Average Precision (MAP), but we included also two alternative ranking measures: Average Recall (AvgRec) and Mean Reciprocal Rank (MRR). In addition, we included four classification measures: Accuracy (acc.), Precision (P), Recall (R), and F1-measure (F1).

4.1 Results and Discussion

The best results per partition and subtask are highlighted in bold. In addition, we always refer to the run with the highest performance. Finally, our percentage comparisons use always absolute values. We can see the results of Subtask A (question-comment similarity ranking) in Table 1. In terms of ranking measures, our system outperformed both the random and the search engine baseline. Using the development set, we observed a MAP improvement of 9.4% compared with the results obtained by the search engine. We can see similar differences with respect to the other two ranking measures. Classification results are also superior. We obtain improvements in accuracy and F1 of 24.9% and 5.2% respectively. These results manifest the potential of the selected lexical and semantic-based features for this subtask.

Similar to Subtask A, the performance of our approach has been also superior in Subtask B

		Ranking measures			Classification measures			
	Model	**MAP**	**AvgRec**	**MRR**	**Acc.**	**P**	**R**	**F1**
	Development set results							
(a)	Random baseline	0.456	0.654	0.535	0.433	0.344	**0.764**	0.475
	Search engine	0.538	0.728	0.631	n/a	n/a	n/a	n/a
(b)	UH-PRHLT (primary)	**0.632**	**0.812**	**0.725**	**0.682**	**0.526**	0.500	0.513
	UH-PRHLT (contr. 1)	0.630	0.811	0.722	0.672	0.510	0.545	**0.527**
	UH-PRHLT (contr. 2)	0.630	0.810	0.722	0.674	0.514	0.522	0.518
	Test set results							
(a)	Random baseline	0.528	0.665	0.587	0.525	0.452	0.405	0.428
	Search engine	0.595	0.726	0.678	n/a	n/a	n/a	n/a
(b)	UH-PRHLT (primary)	0.674	0.794	0.770	**0.632**	**0.556**	0.468	0.508
	UH-PRHLT (contr. 1)	**0.676**	**0.795**	**0.771**	0.624	0.541	**0.501**	**0.520**
	UH-PRHLT (contr. 2)	0.673	0.793	0.767	0.630	0.550	0.491	**0.520**

Table 1: Results of **Subtask A: English Question-Comment Similarity**. (a) Baselines; (b) proposed approach.

		Ranking measures			Classification measures			
	Model	**MAP**	**AvgRec**	**MRR**	**Acc.**	**P**	**R**	**F1**
	Development set results							
(a)	Random baseline	0.559	0.732	0.622	0.488	0.443	**0.766**	0.562
	Search engine	0.713	0.861	0.766	n/a	n/a	n/a	n/a
(b)	UH-PRHLT (primary)	**0.759**	**0.911**	**0.830**	**0.762**	**0.721**	0.724	**0.723**
	UH-PRHLT (contr. 1)	0.757	**0.911**	**0.830**	0.758	0.712	0.729	0.721
	UH-PRHLT (contr. 2)	0.755	0.910	0.817	0.758	0.714	0.724	0.719
	Test set results							
(a)	Random baseline	0.470	0.679	0.510	0.452	0.404	0.326	0.361
	Search engine	0.747	0.883	0.838	n/a	n/a	n/a	n/a
(b)	UH-PRHLT (primary)	0.767	0.903	0.830	0.766	0.635	0.695	0.664
	UH-PRHLT (contr. 1)	0.766	0.902	0.830	0.763	0.627	**0.708**	0.665
	UH-PRHLT (contr. 2)	**0.773**	**0.908**	**0.840**	**0.767**	**0.636**	0.704	**0.668**

Table 2: Results of **Subtask B: English Question-Question Similarity**. (a) Baselines; (b) proposed approach.

		Ranking measures			Classification measures			
	Model	**MAP**	**AvgRec**	**MRR**	**Acc.**	**P**	**R**	**F1**
	Development set results							
(a)	Random baseline	0.138	0.096	0.160	0.284	0.070	**0.759**	0.128
	Search engine	0.306	0.346	0.360	n/a	n/a	n/a	n/a
(b)	UH-PRHLT (primary)	**0.383**	0.413	0.420	0.894	0.242	0.252	0.247
	UH-PRHLT (contr. 1)	**0.383**	**0.421**	0.425	0.897	**0.252**	0.249	**0.250**
	UH-PRHLT (contr. 2)	**0.383**	0.419	**0.435**	0.899	0.251	0.232	0.241
	Test set results							
(a)	Random baseline	0.150	0.114	0.152	0.167	0.296	0.094	0.143
	Search engine	0.404	0.460	0.459	n/a	n/a	n/a	n/a
(b)	UH-PRHLT (primary)	0.432	**0.480**	0.478	0.886	0.376	**0.342**	**0.359**
	UH-PRHLT (contr. 1)	**0.434**	**0.480**	**0.484**	**0.888**	**0.386**	0.327	0.354
	UH-PRHLT (contr. 2)	0.433	**0.480**	**0.484**	**0.888**	0.382	0.327	0.353

Table 3: Results of **Subtask C: English Question-External Comment Similarity**. (a) Baselines; (b) proposed approach.

(question-related question similarity ranking). As we can see in Table 2, using the development set, the improvement of MAP, AvgRec, and MRR has been of 4.6%, 5%, and 6.4% respectively compared to the search engine baseline. In this case, the similarity between questions was easier to estimate — also for the baselines — and the improvements in performance were slightly reduced. With respect to the classification measures, we outperformed the random baseline with 27.4% and 16.1 % of accuracy and F1-measure respectively.

In Table 3 we can see the results of the Subtask C (question-external comment similarity ranking). In this case, we are ranking 100 comments (10 times more compared to the other subtasks). Therefore, this has been the most difficult subtask. However, we obtained improvements in line with those reported for the other subtasks. Compared to the search engine baseline, the MAP, AvgRec, and MRR improved 8.7%, 8.5%, and 7.5% respectively when using the development partition. The accuracy and F1-measure improved 61.5% and 12.2% respectively. The largest number of comments to rank, and the use of top 10 results when measuring results, benefited our approach with this especially high difference in accuracy.

After the analysis of results in the three English subtasks, we highlight that the combination of lexical and semantic-based features that we employ in this work offers a competitive performance for the CQA task. This is true also when comparing results with other task participants. Our approach obtained the highest results — with considerable margin (1.04%) — for subtask B. It is worth mentioning that we designed our system for the subtask B and adapted it later for the other tasks. However, for the other two subtasks, we obtained a low ranking position. At this point we have not discovered any coding error that could explain this difference. In addition, we analysed the information gain ratio of the features for the three subtasks. That results showed an average decrease of ~66% for subtasks A and C. Therefore, we conclude that our approach is more adequate for tasks of similarity rather than

question answering. That analysis also manifests that the most relevant features are the word n-gram ones followed by the CWASA, distributed representation-based, and knowledge-graph-based ones. The comparison of results of all the submitted systems and task participants can be found in the task overview (Nakov et al., 2016).

5 Conclusions

In this work we evaluated the three English subtasks of the SemEval 2016 Task 3 on CQA. In order to measure similarities, our proposed approach combined lexical and semantic-based features. We included simple — and effective — representations based on BOW, character and word n-grams. We also employed semantic features which used distributed representations of words to represent documents or to directly measure similarity by means of distributed word bidirectional alignments. The use of knowledge graphs generated with the BabelNet multilingual semantic network has been exploited too. Experimental results showed that our system was able to outperform — with considerably differences — the random and Google search engine baselines in all the evaluated subtasks. In addition, our approach obtained the highest results in subtask B compared to the other task participants. This fact manifests the potential of our combination of lexical and semantic features for the CQA subtask.

As future work we will continue studying how to approach CQA with knowledge graphs and distributed representations. In addition, we will further explore how to employ this type of lexical and semantic-based representations for other NLP tasks such as plagiarism detection.

Acknowledgments

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Davide Buscaldi, Paolo Rosso, José Manuel Gómez-Soriano, and Emilio Sanchis. 2010. Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems*, 34(2):113–134.

Asli Celikyilmaz, Dilek Hakkani-Tur, and Gokhan Tur. 2010. Lda based similarity modeling for question answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, pages 1–9. Association for Computational Linguistics.

Marc Franco-Salvador, Paolo Rosso, and Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 414–423. Association for Computational Linguistics.

Marc Franco-Salvador, Francisco Rangel, Paolo Rosso, Mariona Taulé, and M. Antònia Martí. 2015a. Language variety identification using distributed representations of words and documents. In *Proceeding of the 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF 2015)*, volume LNCS(9283), page n/a. Springer-Verlag.

Marc Franco-Salvador, Paolo Rosso, and Francisco Rangel. 2015b. Distributed representations of words and documents for discriminating similar languages. In *Proceeding of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial), RANLP*.

Marc Franco-Salvador, Parth Gupta, Paolo Rosso, and Rafael E. Banchs. 2016a. Cross-language plagiarism detection over continuous-space representations of language. *Pre-print submitted to journal*.

Marc Franco-Salvador, Paolo Rosso, and Manuel Montes y Gómez. 2016b. A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management*.

Samer Hassan and Rada Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *AAAI*.

Marti A. Hearst, Susan T Dumais, Edgar Osman, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28.

Yongshuai Hou, Cong Tan, Xiaolong Wang, Yaoyun Zhang, Jun Xu, and Qingcai Chen. 2015. Hitszicrc: Exploiting classification approach for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*, volume 15 of *SemEval '15*, pages 196–202. Association for Computational Linguistics.

Jiwoon Jeon, W Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 84–90. ACM.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(04):343–360.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Preslav Nakov, Lluís Màrquez, Walid Magdy, and Alessandro Moschitti. 2015. Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 269–281. Association for Computational Linguistics.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 41–47. Association for Computational Linguistics.

Paolo Rosso, Lluís-F Hurtado, Encarna Segarra, and Emilio Sanchis. 2012. On the voice-activated question answering. *IEEE Transactions on Systems, Man, and Cybernetics–Part C*, 42(1):75–85.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Jasper Snoek. 2013. *Bayesian Optimization and Semiparametric Models with Applications to Assistive Technology*. Ph.D. thesis, University of Toronto.

Quan Hung Tran, Vu Tran, Tu Vu, Minh Nguyen, and Son Bao Pham. 2015. Jaist: Combining multiple features for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, volume 15 of *SemEval '15*, pages 215–219. Association for Computational Linguistics.

Ellen M. Voorhees. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.