# Finding Good Answers in Online Forums: Community Question Answering for Bulgarian

**Tsvetomila Mihaylova, Ivan Koychev**
FMI, Sofia University “St. Kliment Ohridski”, Sofia, Bulgaria

**Preslav Nakov**
ALT Research Group, Qatar Computing Research Institute, HBKU, Doha, Qatar

**Ivelina Nikolova**
IICT, Bulgarian Academy of Sciences, Sofia, Bulgaria

## Abstract

Community Question Answering (CQA) is a form of question answering that is getting increasingly popular as a research direction recently. Given a question posted in an online community forum and the thread of answers to it, a common formulation of the task is to rank automatically the answers, so that the good ones are ranked higher than the bad ones. Despite the vast research in CQA for English, very little attention has been paid to other languages. To bridge this gap, here we present our method for Community Question Answering in Bulgarian. We create annotated training and testing datasets for Bulgarian, and we further explore the applicability of machine translation for reusing English CQA data for building a Bulgarian system. The evaluation results show improvement over the baseline and can serve as a basis for further research.

## 1. Introduction

With the ever growing user-generated content, it is becoming increasingly harder and time-consuming for users to find valuable information. This is especially true for web forums, where a question can generate a thread of hundreds of answers. Thus, there is a need to filter the answer thread and to present to the user the most relevant answers first, i.e., to rerank the answers in a forum not chronologically as they naturally occur, but based on how well they answer the original forum question.

Community Question Answering (CQA) is a special case of the more general problem of Question Answering (QA), which has been an active research area for years (Webber and Webb, 2010). The TREC conference has had QA tasks since 1999 (Voorhees, 1999), focusing on various aspects of the problem.

CQA is a topic with growing research interest. The specifics of CQA include user-generated content in free text, without necessarily following strict rules. Important difference between the traditional content and the user-generated content is that the latter shows higher variance in quality (Agichtein et al., 2008; Ahn et al., 2013; Baltadzhieva and Chrupała, 2015). This problem is well-studied for English, e.g., there has been a shared tasks for CQA at SemEval-2015 (Nakov et al., 2015) and SemEval-2016 (Nakov et al., 2016b).

However, the field is not explored for Bulgarian yet. To bridge this gap, in this paper, we experiment with CQA data from the biggest online forum in Bulgaria - BGMamma.[1] We create annotated training and testing datasets for Bulgarian. While annotating data for testing is not that hard, annotating a lot of data for training is a rather time-consuming task. Therefore, we annotate small sets for training and testing, and we translate them from Bulgarian to English, and we train a system that works for English. In order to make a larger training set, we use additional publicly available annotated data for English. Then we apply domain adaptation to combine the small translated in-domain data with the large out-of-domain data.

[1] BG Mamma: `http://www.bg-mamma.com/`

The remainder of the paper is organized as follows: Section 2. introduces the research in the field related to our task. Section 3. describes the features and the method used for classification. Section 4. contains the result of our experiments, where we compare the results from using different training sets and different feature groups. In section 5., we conclude and we point to possible directions for future work.

## 2. Related Work

Community Question Answering is a topic of great research interest. For example, there has been a shared task for CQA in SemEval-2015 (Nakov et al., 2015) and SemEval-2016 (Nakov et al., 2016b) editions. Various approaches for CQA have been explored by the systems in those competitions. For example, Belinkov et al. (2015) used vectors of the question and of the comment, metadata features, and text-based similarities. Nicosia et al. (2015) used similarity measures, URLs in the comment text and statistics about the user profile: number of good, bad, and potentially useful comments. In our system, we use similar features to those systems, such as the number of posts by the same user in the thread, topic model-based feature, special words, URLs, word embeddings of the question and comment, metadata features, and text similarities.

Other approaches for CQA, used in the top systems in SemEval-2016 Task 3 on CQA (Nakov et al., 2016b) include troll user features by (Mihaylov et al., 2015a; Mihaylov et al., 2015b; Mihaylov and Nakov, 2016a), fine-tuned word embeddings as in the SemanticZ system (Mihaylov and Nakov, 2016b), and PMI-based goodness polarity lexicons as in the PMI-cool system (Balchev et al., 2016), as well as sentiment polarity features (Nicosia et al., 2015). Other systems are based on a deep learning architecture, e.g., as in the MTE-NN system (Guzmán et al., 2016a; Guzmán et al., 2016b; Nakov et al., 2016a), which borrowed an entire neural network framework and architecture from previous work on machine translation evaluation (Guzmán et al., 2015). We do not currently use such kinds of features in our system.

The current study is based on our previous work for Task 3 of SemEval-2016[2] (Mihaylova et al., 2016). We rank a set of comments according to their relevance to a question. The task is solved with a classification approach where each question-comment pair is tagged as *Good* or *Bad* and the rank of a comment is a function of the probability that the pair is *Good*. The following feature groups are considered during classification: metadata, semantic, lexical, credibility and user features. In the current research for Bulgarian, we only apply metadata, semantic and lexical features. The user features are still applicable and will be included in future experiments.

Using machine translation for solving tasks in languages different from English is used in various domains. For example, Mohammad et al. (2016) translated text from Arabic to English for sentiment analysis. Balahur and Turchi (2014) used machine translation for sentiment analysis of tweets. In cross-lingual and multilingual information retrieval, machine translation is often applied to the query, to the results or to both (PothulaSujatha and Dhavachelvan, 2011). The experiments show that using machine translation in such settings yields meaningful results and could be applied for translating the target documents to languages rich in resources such as English as we do in the current study. Translation is used for CQA by Zhou et al. (2011; 2012) who experiment with machine translation for question retrieval.

The problem with no sufficient data available for classification in a target domain is solved by using domain adaptation (Daume III, 2007). For the current study, we do not have sufficient data for Bulgarian, so we are using domain adaptation for including publicily available annotated data to expand the training set.

In our research of CQA for Bulgarian, we use machine translation to translate the collected training and testing target data from Bulgarian to English. We further use domain adaptation to expand the insufficient training set. After that, we use an existing pipeline developed for CQA for English, and we run it on the translated data together with the additional training data.

[2] SemEval-2016, Task 3: `http://alt.qcri.org/semeval2016/task3/`

## 3. Method

The purpose of our experiments is to see whether a system that performed well for CQA in English (ranked 1st on the main Substask C of Task 3 on SemEval 2016), would deliver strong results for Bulgarian too.

The experimental environment is built on top of the framework developed for CQA in English (Mihaylova et al., 2016). It solves the task of ranking comments with respect to their relevance to a given question. The pipeline includes variety of features, some of which are extracted from external data sources, i.e., the Qatar Living (QL) forum.

In this work, we use all features of the system for English, excluding user statistics from the QL forum, pointwise mutual information (PMI) and credibility features. Those features are still relevant for our study in Bulgarian, and we plan to add them in future experiments.

### 3.1. Features

In our experiments with data in Bulgarian, we use lexical, semantic and metadata features in the way they are described in (Mihaylova et al., 2016).

**Metadata Features**

These features present observations for the thread and for the comment structure and properties.

- Whether the comment is written by the author of the question.
- Comment's rank in the thread.
- Ratio of the comment length to the question length (in terms of number of tokens).
- Number and order of comments from the same user in a particular thread.
- Presence and the number of URLs in the question and in the comment.

**Lexical Features**

For obtaining the lexical features, the question and the comment texts are annotated with GATE (Cunningham et al., 2002; Cunningham et al., 2011).

- Number of each question word (*where*, *who*, *what*, etc.) in the question and in the comment text.
- Whether the comment contains an answer to a wh-question (*where*, *who*, *what*, etc.). For example, if the question contains *where* and the comment contains an address or location, this is considered as a response to such a question.
- Number of verbs, nouns, pronouns, and adjectives in the question and in the comment.
- Number of question marks and question words in the question and in the comment.
- Comment contains smileys, currency units, e-mails, phone numbers, only laughter, "thank you" phrases, personal opinions, disagreement.
- Number of misspelled words and offensive words from a dictionary.
- Dictionary of unigram and bigram occurrences across the classes.
- Lexical similarity between a question and a comment using *SimHash* (Sadowski and Levin, 2007).
- Level of readability and complexity of the text (Aluisio et al., 2010). The standard readability measures include Automated Readability Index, Coleman-Liau Index, Flesch Reading Ease, Gunning Fog Index, Flesch-Kincaid Grade Level, LIX, SMOG grade. We also use statistics about the average number of words per sentence in the comment or in the question, and type-to-token ratios.
- Average number of words per sentence in the comment or in the question.
- Type-to-token ratios in the question and in the comment.

**Semantic Features**

This group of features aims to find the similarity of the question and of the comment meaning.

- Topic Modeling with Mallet (McCallum, 2002) is used for training of 100 topics from questions and comments from the QL training data.
- Word Embeddings trained with Word2Vec(Mikolov et al., 2013) on the QL forum data.
- Cosine distances between the text of the question and of the comment: between vectors of all words, between different parts of speech (nouns, verbs, adjectives). The cosine distance was calculated between the sum of the embeddings of all words in the question and in the comment text.

### 3.2. Domain Adaptation

Since the training data we prepared for Bulgarian is relatively small, we expanded the training set by using domain adaptation (Daume III, 2007). The idea of domain adaptation is, when insufficient training data exists for a target domain, to use available data from another domain as an additional training set, called the source set. Suppose we have a set of features we can extract from the target and from the source data. We can perform domain adaptation using equation 1.

$$\Phi^s = \langle x, x, 0\rangle, \Phi^t = \langle x, 0, x\rangle \tag{1}$$

In this equation, $x$ is the vector of features and $0 = \langle 0, 0, ...\rangle$ is the zero vector. In order to put the features extracted from the target and from the source domain into one classifier, we expand the feature space and we use three parts for the feature vector. The first part contains features extracted from both the target and the source sets. The second part contains features extracted from the source set only. The third part contains features extracted from the target set only.

For the domain adaptation, we construct a training set that contains two subsets: the features from Qatar Living as a source set formatted as $\Phi^s$, and the features from the BG-Mamma training set as a target set ($\Phi^t$). The test set is the test set from BG-Mamma, again formatted as a target set ($\Phi^t$).

### 3.3. Classifier

The task of ordering the comments with respect to the question is a ranking problem. It aims to order the comments according to their relevance as a response to the given question. It is important that the *Good* comments are ranked higher than the *Bad* ones. We approach the problem as a classification task. Each example is a question-answer pair, and the following feature vector is formed for the examples:

$$v_{q_1}, ..., v_{q_k}, v_{c_1}, ..., v_{c_k}, f_1, ..., f_m \tag{2}$$

where $v_q$ and $v_c$ are the $k$-sized vectors of the word embeddings of the question and of the comment, and $f$ is a vector of the non-embedding features.

We used LibSVM (Chang and Lin, 2011; Hsu et al., 2003) for the classification. The results from the classification give probability for each class. The probability for the *Good* class is used as a ranking score for the question-comment pair. We experimented with different kernels, but the best results are achieved with an RBF kernel. Thus, we only report results when using an RBF kernel.

## 4. Experiments

### 4.1. Data

The data for the current study is collected from the largest online forum in Bulgaria - BGMamma. The forum has topics in various categories, each topic is a thread with comments from different users. In order to prepare the data in a format suitable for our task, we first selected topics with titles containing 'въпрос' (the Bulgarian word for 'question'). The first comment in the topic is considered as a question. The next five comments in the topic are considered as answers to this question. We annotated manually 80 questions with the first 5 answers from the thread for each of them, i.e., 400 question-comment pairs. Each answer is annotated as either *Good* (it gives a direct answer to the given question) or *Bad* (it does not give a direct answer to the question).

We split the annotated questions into training and test set. The training set has 50 questions with 5 answers each, i.e., 250 question-answer pairs. The test set has 30 questions with 5 answers each, i.e., 150 question-answer pairs. Table 1 shows more detailed statistics for the training and test sets.

After the data was annotated, the topic categories, question texts, question subjects and comment texts were translated from Bulgarian to English with the *Microsoft Translation API*.[3] As an additional training data we use the Train-1 set from SemEval-2016 Task 3. From them, we took only the comments on positions from 1 to 5 in the forum thread. The difference of the SemEval labeling of the comments is that they also include *Potentially Useful* labels. We consider those labels *Bad*.

[3] https://www.microsoft.com/en-us/translator/translatorapi.aspx

| | Questions Count | Comments Count | Good Comments | Bad Comments |
|---|---|---|---|---|
| Test Set from BG-Mamma | 30 | 150 | 49 | 110 |
| Train Set from BG-Mamma | 50 | 250 | 84 | 166 |
| Additional Train Set from QL | 1411 | 7055 | 3021 | 4034 |

Table 1: Statistics about the data sets.

Table 2 shows an example of a question thread and its comments from the forum (the translation in English is presented in Table 3). It also illustrates the difference between the relevant vs. non-relevant answers. The comments marked as *Good* give a direct answer to the asked question. The answers marked as *Bad* can be for example a 'Thank you' statement, an irrelevant comment, could be a new question or a reply to some question in the comment thread rather than to the original question.

<table>
<tr><th colspan="2">Question Subject</th><th>Question Text</th></tr>
<tr><td colspan="2">Въпроси относно камина Ерато</td><td>Моля тези от вас, които имат такава камина да се включат с отговор на няколко въпроса: 1. Запалихте ли вече камините. Първоначално само вечер ли? 2. Какви настройки сте направили? 3. Имате ли някакво ръководство? 4. Какви пелети ползвате? Предварително благодаря на всички :lol:</td></tr>
<tr><th colspan="3">Comments</th></tr>
<tr><th>Position</th><th>Relevance</th><th>Comment Text</th></tr>
<tr><td>1</td><td>Good</td><td>Имам Пони9 на Ерато. Днес я запалих за 2 часа. Пелетите са български от Разлог, но имаме още 2-3 торби от тях. За тази зима сме поръчали етрополски пелети 2,5 тона. Засега не сме настройвали нищо. Миналата година бяхме настроили да се включва сутрин в 5:30, после по някое време се изключваше, пак се включваше и т.н., но не помня подробности. Имам книжка с инструкции.</td></tr>
<tr><td>2</td><td>Bad</td><td>Благодаря за отговора. Използвали ли сте някакъв екорецим?</td></tr>
<tr><td>3</td><td>Bad</td><td>Нямам идея какво е това. :shock:</td></tr>
<tr><td>4</td><td>Bad</td><td>Бухахаха :D ей такива смешки стават, когато пишеш през телефона. Имах предвид ЕКО РЕЖИМ :hug: Междудругото вашата камина, когато достигне определена темп спира ли работа?</td></tr>
<tr><td>5</td><td>Bad</td><td>Първата зима спираше, но после от фирмата, откъдето я купихме, ѝ промениха настройките и сега не спира. Проблемът със самоизключването бе, че трябва температурата да падне с 2 градуса под зададената, за да се включи. По този начин се получаваха големи температурни амплитуди.</td></tr>
</table>

Table 2: Example of question and comments from the forum in Bulgarian.

The feature extraction pipeline includes word embeddings trained on the Qatar Living[4] forum with Word2Vec (Mikolov et al., 2013). This data was provided as an unannotated data for SemEval-2016 Task 3 and it includes 200,000 questions and 2 million comments. The vectors were trained with Gensim (Řehůřek and Sojka, 2010).

## 4.2. Experiment Setup

We train our models on several different training sets and we measure which one achieves best results when testing on the test set. The first one is the training set from the Bulgarian forum, translated to English. The second one is the training set from the QL forum - questions and comments originally written in English.

[4] Qatar Living: `http://www.qatarliving.com/forum`

| **Question Subject** | | **Question Text** |
|---|---|---|
| Questions about fireplace Erato | | Please those of you who have that fireplace to get involved with the answer to a few questions: 1. You lit the fireplaces. Initially only night? 2. What settings have you done? 3. Do you have any guidance? 4. What pellets you use? thanks in advance to all : lol: |
| **Comments** | | |
| **Position** | **Relevance** | **Comment Text** |
| 1 | Good | I have Poni9 on Erato. Today I lit it for 2 a.m. pellets are Bulgarian from Razlog, but we still have 2-3 bags of them. For this winter we ordered etropole pellets 2.5 tons. so far, we haven't set up any thing. Last year we were set up to turn on at 5:30 in the morning, then at some time is excluded, it still included, etc, but I don't remember the details. Have book with instructions. |
| 2 | Bad | Thanks for the reply. Have you used any ekorecim? |
| 3 | Bad | I have no idea what that is. :shock: |
| 4 | Bad | Buhahaha :D These jokes become, when you write in the phone. I meant the ECO MODE : hug: by the way your fireplace when it reaches a certain temp stops work? |
| 5 | Bad | The first winter, but then stopped by the company where we bought it, I changed the settings and now I can't stop. problem with turning itself off, you need the temperature to drop to 2 degrees below the set to be turned on. Thus received large temperature amplitudes. |

Table 3: Example of question and comments from the previous table, translated to English.

To construct the third training set, we use domain adaptation as described in (Daume III, 2007). The details were described in Section 3.2. above. As the source set, we use the training set from QL. The training set in Bulgarian is included as target in the training data and the test set on the Bulgarian forum is also processed as target. Comparison of the results is shown in Section 4.4..

The baseline is calculated by ranking the comment with respect to their chronological position in the question-comment thread. The first posted comment in the thread has position 1, the second one has position 2 etc. For the baseline, *1/comment position* is used as the ranking score for the comment in the thread.

### 4.3. Evaluation

In Section 4., we present the results of our experiments. We first compare the test results when the classifier was trained on different training sets with all features. After that, we compare different feature groups to find the most important ones for our task.

As a main evaluation measure, we use Mean Average Precision at 5 (MAP@5), as we are interested in the most useful answers appearing at the top of the result. As an additional measure, we use accuracy. When a ranked result is given, *MAP* (formula 3) calculates the mean of the average precision for each query (question) $q$. Average precision *AveP(q)* takes the precision at each position for the given question (i.e., for the first 1 result, for the first 2 results) and then takes the average of those values (precision *P(k)* measures the ratio of the positively classified - *Good* examples to all given examples up to position $k$). Finally, *accuracy* measures the ratio of the number of correctly classified examples to the total number of examples.

$$MAP@5 = \sum_{q=1}^{Q} AveP(q)/Q, AveP@5 \sum_{k=1}^{5} P(k)/5 \tag{3}$$

### 4.4. Results

Table 4 shows the results when training the classifier with different training sets: from BG-Mamma, from Qatar Living (including all comments and only the first 5 comments), and using domain adaptation. For this comparison, all features are used. The results show that only using the data from Qatar Living as a training set does not yield very good results. The best results are achieved when the training set from BG-Mamma is used, as well as when domain adaptation is applied. For further experiments, we use only the training set from BG-Mamma, as is yields comparable results to domain adaptation, but training the classifier is faster because of the smaller feature space and the smaller set size.

| Training Set | MAP | Accuracy (%) |
|---|---|---|
| Baseline | 70.76 | – |
| Training data from BG-Mamma | 90.39 | 78.67 |
| Training data from Qatar Living - all data | 83.67 | 73.33 |
| Training data from Qatar Living - only answers up to 5 | 87.06 | 74.67 |
| Domain adaptation - data from Qatar living and BG-Mamma | 90.39 | 79.33 |

Table 4: Comparison of different training sets. The shown results are trained on the corresponding training set with all features.

For our next experiments, we wanted to determine which groups of features are significant for the results and which ones are not. Tables 5 and 6 show experiments with different features groups. The classifier for those experiments was trained on the training set from BG-Mamma, translated to English. The compared feature groups contain logically related features, described in Section 3.1.. The results show that the most significant features are the word embeddings and the metadata of question and comment. Those feature groups improve the baseline when used on their own and the result is lower when they are excluded from the feature set. In our previous work (Mihaylova et al., 2016), the word embeddings and the metadata also turned out to be among the most significant features.

The described experiments show that the approach of using machine translation and a pipeline prepared for English works well. The achieved results significantly improve the baseline.

| | MAP | Accuracy (%) |
|---|---|---|
| All Features | 90.39 | 78.67 |
| only Semantic features / Word embeddings | 81.06 | 67.33 |
| only Metadata features / Thread structure | 76.89 | 72.00 |
| only Metadata features / Comment structure | 74.28 | 67.33 |
| only Semantic features / Cosine distances | 66.42 | 67.33 |
| only Metadata features / URLs | 68.15 | 67.33 |
| only Lexical features / Question words | 59.13 | 67.33 |
| only Lexical features / Parts of speech | 69.67 | 66.00 |

Table 5: Experiments with different feature groups. The results are obtained when only the features from the given group are used for classification.

| | MAP | Accuracy (%) |
|---|---|---|
| All Features | 90.39 | 78.67 |
| All − Semantic features / Word embeddings | 86.22 | 72.67 |
| All − Metadata features / Thread structure | 85.83 | 78.00 |
| All − Metadata features / Comment structure | 90.11 | 77.33 |
| All − Semantic features / Cosine distances | 88.72 | 76.67 |
| All − Metadata features / URLs | 90.39 | 78.67 |
| All − Lexical features / Question words | 90.39 | 74.00 |
| All − Lexical features / Parts of speech | 90.94 | 76.67 |

Table 6: Experiments with different feature groups. The results are obtained when all the features are used, excluding the features in the given group.

## 5. Conclusion and Future Work

We have presented our research on Community Question Answering for Bulgarian using machine translation. First, we translate the text of the questions and answers from Bulgarian to English and then run a pipeline tested for English with the translated texts. The experiments show that this approach works very well and the improvement over the baseline is comparable to the one used in the original system tested in English. The results show that this approach can be used for further work in CQA for Bulgarian.

In future work, we plan to try ideas from the top systems that participated in SemEval-2016 Task 3 on CQA (Nakov et al., 2016b). In particular, we want to incorporate several rich knowledge sources, e.g., as in the SUper Team system (Mihaylova et al., 2016), including troll user features as inspired by (Mihaylov et al., 2015a; Mihaylov et al., 2015b; Mihaylov and Nakov, 2016a), fine-tuned word embeddings as in the SemanticZ system (Mihaylov and Nakov, 2016b), and PMI-based goodness polarity lexicons as in the PMI-cool system (Balchev et al., 2016), as well as sentiment polarity features (Nicosia et al., 2015).

We further want to use our features in a deep learning architecture, e.g., as in the MTE-NN system (Guzmán et al., 2016a; Guzmán et al., 2016b; Nakov et al., 2016a), which borrowed an entire neural network framework and architecture from previous work on machine translation evaluation (Guzmán et al., 2015).

Moreover, we plan to use information from entire threads as well as from other question-answer threads to make better predictions, as using thread-level information for answer classification has already been shown useful for SemEval-2015 Task 3, subtask A, e.g., by using features modeling the thread structure and dialogue (Nicosia et al., 2015; Barrón-Cedeño et al., 2015), or by applying thread-level inference using the predictions of local classifiers (Joty et al., 2015; Joty et al., 2016). How to use such models efficiently in our ranking evaluation setup is an interesting research question.

Finally, we plan to experiment with different CQA tasks, such as ranking similar questions to a given question and finding useful answer to a new question entered by a user of the forum as in SemEval-2016 Task 3. We could run a pipeline for Bulgarian using the same features and we will can compare the results to the current approach. This can include translation of the English resources to Bulgarian.

**Acknowledgments.** This research was performed by Tsvetomila Mihaylova, a M.Sc. student in Computer Science in the Sofia University “St Kliment Ohridski”. It is also part of the Interactive sYstems for Answer Search (Iyas) project, which is developed by the Arabic Language Technologies (ALT) group at the Qatar Computing Research Institute (QCRI), HBKU, part of Qatar Foundation in collaboration with MIT-CSAIL.

## References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding High-quality Content in Social Media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM ’08, pages 183–194, Palo Alto, California, USA.

Ahn, J., Butler, B. S., Weng, C., and Webster, S. (2013). Learning to Be a Better Q'Er in Social Q&A Sites: Social Norms and Information Artifacts. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, ASIST '13, pages 4:1–4:10, Montreal, Quebec, Canada.

Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability Assessment for Text Simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, IUNLPBEA '10, pages 1–9, Los Angeles, California, USA.

Balahur, A. and Turchi, M. (2014). Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis. *Comput. Speech Lang.*, 28(1):56–75.

Balchev, D., Kiprov, Y., Koychev, I., and Nakov, P. (2016). PMI-cool at SemEval-2016 Task 3: Experiments with PMI and Goodness Polarity Lexicons for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Baltadzhieva, A. and Chrupała, G. (2015). Question Quality in Community Question Answering Forums: A Survey. *SIGKDD Explor. Newsl.*, 17(1):8–13.

Barrón-Cedeño, A., Filice, S., Da San Martino, G., Joty, S., Màrquez, L., Nakov, P., and Moschitti, A. (2015). Thread-Level Information for Comment Classification in Community Question Answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '15, pages 687–693, Beijing, China.

Belinkov, Y., Mohtarami, M., Cyphers, S., and Glass, J. (2015). VectorSLU: A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 282–287, Denver, Colorado, USA.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: an Architecture for Development of Robust HLT applications. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, ACL '12, pages 168–175, Philadelphia, Pennsylvania, USA.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*.

Daume III, H. (2007). Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, pages 256–263, Prague, Czech Republic.

Guzmán, F., Joty, S., Màrquez, L., and Nakov, P. (2015). Pairwise Neural Machine Translation Evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL-IJCNLP '15, pages 805–814, Beijing, China.

Guzmán, F., Màrquez, L., and Nakov, P. (2016a). Machine Translation Evaluation Meets Community Question Answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, Berlin, Germany.

Guzmán, F., Màrquez, L., and Nakov, P. (2016b). MTE-NN at SemEval-2016 Task 3: Can Machine Translation Evaluation Help Community Question Answering? In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). *A Practical Guide to Support Vector Classification*. Technical report, Department of Computer Science, National Taiwan University.

Joty, S., Barrón-Cedeño, A., Da San Martino, G., Filice, S., Màrquez, L., Moschitti, A., and Nakov, P. (2015). Global Thread-level Inference for Comment Classification in Community Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 573–578, Lisbon, Portugal.

Joty, S., Màrquez, L., and Nakov, P. (2016). Joint Learning with Global Inference for Comment Classification in Community Question Answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '16, San Diego, California, USA.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Mihaylov, T. and Nakov, P. (2016a). Hunting for Troll Comments in News Community Forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, Berlin, Germany.

Mihaylov, T. and Nakov, P. (2016b). SemanticZ at SemEval-2016 Task 3: Ranking Relevant Answers in Community Question Answering Using Semantic Similarity Based on Fine-tuned Word Embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Mihaylov, T., Georgiev, G., and Nakov, P. (2015a). Finding Opinion Manipulation Trolls in News Community Forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, CoNLL '15, pages 310–314, Beijing, China.

Mihaylov, T., Koychev, I., Georgiev, G., and Nakov, P. (2015b). Exposing Paid Opinion Manipulation Trolls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '15, pages 443–450, Hissar, Bulgaria.

Mihaylova, T., Gencheva, P., Boyanov, M., Yovcheva, I., Mihaylov, T., Hardalov, M., Kiprov, Y., Balchev, D., Koychev, I., Nakov, P., Nikolova, I., and Angelova, G. (2016). SUper Team at SemEval-2016 Task 3: Building a Feature-Rich System for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 746–751, Atlanta, Georgia, USA.

Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016). How Translation Alters Sentiment. In *Journal of Artificial Intelligence Research*, volume 55, pages 95–130.

Nakov, P., Màrquez, L., Magdy, W., Moschitti, A., Glass, J., and Randeree, B. (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, Denver, Colorado, USA.

Nakov, P., Guzmán, F., and Màrquez, L. (2016a). It Takes Three to Tango: Triangulation Approach to Answer Ranking in Community Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, Austin, Texas, USA.

Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A. A., Glass, J., and Randeree, B. (2016b). SemEval-2016 Task 3: Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Nicosia, M., Filice, S., Barrón-Cedeño, A., Saleh, I., Mubarak, H., Gao, W., Nakov, P., Da San Martino, G., Moschitti, A., Darwish, K., Màrquez, L., Joty, S., and Magdy, W. (2015). QCRI: Answer Selection for Community Question Answering - Experiments for Arabic and English. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 203–209, Denver, Colorado, USA.

PothulaSujatha and Dhavachelvan, P. (2011). A Review on the Cross and Multilingual Information Retrieval. *International Journal of Web & Semantic Technology (IJWesT)*, 2(4):115–124.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta.

Sadowski, C. and Levin, G. (2007). *SimiHash: Hash-based similarity detection*. Technical Report UCSC-SOE-11-07, University of California, Santa Cruz, USA.

Voorhees, E. M. (1999). The TREC-8 Question Answering Track Report. In *In Proceedings of TREC-8*, pages 77–82.

Webber, B. and Webb, N. (2010). Question Answering. In Alexander Clark, Chris Fox, S. L., Ed., *The Handbook of Computational Linguistics and Natural Language Processing*, chapter 22, pages 630–654.

Zhou, G., Cai, L., Zhao, J., and Liu, K. (2011). Phrase-based Translation Model for Question Retrieval in Community Question Answer Archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 653–662, Portland, Oregon.

Zhou, G., Liu, K., and Zhao, J. (2012). Exploiting Bilingual Translation for Question Retrieval in Community-Based Question Answering. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 3153–3170.