Does *Similarity* Matter? The Case of Answer Extraction from Technical Discussion Forums

Rose Catherine[1] *Amit Singh*[1]
Rashmi Gangadharaiah[1] *Dinesh Raghu*[1] *Karthik Visweswariah*[1]
(1) IBM Research - India, Bangalore
{rosecatherinek, amitkumarsingh, rashgang, dinraghu, v-karthik}@in.ibm.com

ABSTRACT
Extracting question–answer pairs from social media discussions has garnered much attention in recent times. Several methods have been proposed in the past that pose this task as a post or sentence classification problem, which label each entry as an answer or not. This paper makes the first attempt at the following two–fold objectives: (a) In all classification based approaches towards this direction, one of the foremost signals used to identify answers is their *similarity* to the question. We study the contribution of content similarity specifically in the context of technical problem–solving domain. (b) We introduce hitherto unexplored features that aid in high–precision extraction of answers, and present a thorough study of the contribution of all features to this task. Our results show that, it is possible to extract answers using these features with high accuracy, when their similarity to the question is unreliable.

KEYWORDS: Question Answering, Information & Content Extraction, Text Mining.

Proceedings of COLING 2012: Posters, pages 175–184,
COLING 2012, Mumbai, December 2012.

1 Introduction

Online discussion forums are internet sites that provide a channel for users to discuss and share their views on various topics ranging from troubleshooting products to choosing holiday resorts. Over a period of time, they have accumulated huge amounts of data, thus making them excellent sources of information for future reference. Mining question–answer knowledge from these online forums, and social media discussions in general, has garnered much research and commercial interest of late. Such mined data can be used to provide enhanced access to the forum content, augment chatbot knowledge (Huang et al., 2007), supplement the data in Community Question Answering (CQA) sites (Cong et al., 2008) etc.

All answer extraction methods suggested in the past use a multitude of features that include similarity based and lexical features, structural features constructed from the organization of the discussion etc. Of these, similarity of the answer candidate to the question post has been a de facto standard feature, whose contribution to the accuracy of extraction have so far only been assumed, but never really measured.

The goals and contributions of this paper are as below:

- Study the characteristics of technical discussion forums and their points of difference from other domains, thus motivating the rest of the contributions of this paper.
- Analyze the effectiveness of similarity of candidates to the question, as a feature towards the task of identifying answers, specifically in the case of technical discussion forums. Unlike other domains, here, the answers have minimal lexical overlap with the question.
- Propose new features and study the contribution of all features to the overall goal of answer extraction. Particularly, we aim to test if similarity–independent features can act as an understudy to question similarity for this task, when the latter is unavailable/unreliable.

To the best of the authors' knowledge, this is the first paper which attempts the above objectives.

2 Related Work

Classification–based approaches proposed in the past for detecting answers in online discussion forums like (Ding et al., 2008), (Hong and Davison, 2009), (Yang et al., 2009), (Kim et al., 2010), and in email discussions like (Shrestha and McKeown, 2004) use similarity of the sentence or post to the question as one of the main features for identifying answers. Other approaches like graph based methods (Cong et al., 2008) and (Otterbacher et al., 2005) rely on similarity to construct the graph. However, none of these approaches test systematically, the inadequacy or indispensability, which ever is the case, of similarity to the task.

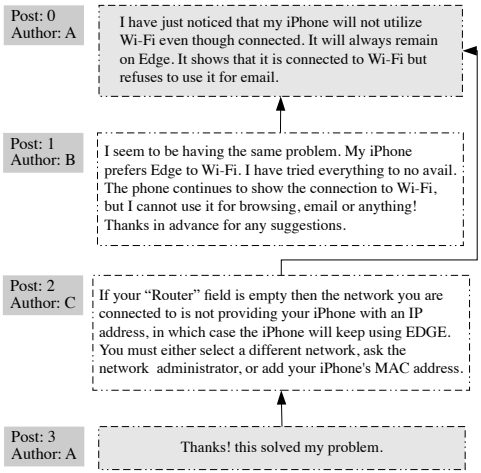

Figure 1: Technical Discussion Thread - Example

The low similarity between questions and answers is due to the lexical chasm between them, which some prior works had

incidentally observed (Cong et al., 2008), (Ding et al., 2008), (Hong and Davison, 2009), and used external data like Yahoo! Answers[1] to either expand the content or learn a translation model. For learning such models, it should be noted that, such data may not always be available and is required in good amounts to train a decent model. Also, (Hong and Davison, 2009), while experimenting on technical discussions, reported that a combination of two non–similarity based features gave better accuracy than a language model. In Section 4.1, we show that, in addition to these features, with the aid of other non–similarity based features, the accuracy of the task can be greatly improved.

3 Does Question Similarity Matter for Answer Extraction?

Discussion forums provide an online medium for users to collaboratively solve a problem or answer a query. Figure 1 shows a typical discussion in an online forum – it starts with the first post, which we refer to as the `question` post. The directed edges show the `reply-to` relation, where the start node of the edge – child post, was posted in reply to the end node of the edge – parent post. In this paper, we use the term 'thread' interchangeably with 'discussion' to refer to a single multi–user conversation of the above form.

Discussions frequently have digressions, where new questions are posted and discussed within the same thread. We do not attempt to find these questions; question detection is a well researched area (Cong et al., 2008), and is outside the scope of this paper. We treat the first post as the main question and find answers to only this question. Answers to other questions within the same thread are not considered.

3.1 Characteristics of Technical Discussion Forums

Technical discussion forums differ from other forums like travel and shopping in that, they are characterized by low lexical overlap between the problem statement and the answer.

3.1.1 Lexical Overlap

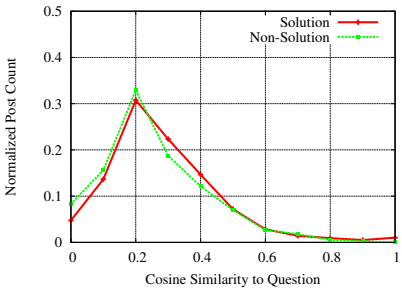

Figure 2: Similarity Histogram

To verify the above hypothesis about the overlap between answer posts and question posts, we studied about 450 threads which together have about 2000 reply posts, from Apple Discussions[2] (details about the dataset is in Section 4). Figure 2 plots a histogram (counts normalized) of cosine similarity (bucketized) of answers and non–answers to their respective question posts. The cosine similarity was computed on the *tf-idf* representation of the posts, after removing common English stopwords and stemming the words using a Porter Stemmer (Porter, 1980). It is clear from the figure that, a large fraction of answers have very minimal overlap with the question, and the fraction of answers with high overlap is very minimal. It is interesting to note that, the same trend is exhibited by non-answers too, thereby making it difficult to separate out the two using question similarity alone. In-depth inspection showed that, a large fraction of posts whose overlap with the question post is high, are in fact, other users complaining about facing the same or a similar problem, while the actual answer

[1] http://answers.yahoo.com
[2] https://discussions.apple.com/community/iphone

Forum	Avg. Spam %	Avg. Digression %
Apple Discussions (discussions.apple.com)	0	10.9
Ubuntu (ubuntuforums.org)	0	5.9
Photography (photography-on-the.net)	0	8.9
Avg. for Technical	0	8.5
Trip Advisor (www.tripadvisor.com)	4.1	25.2
Lonely Planet (www.lonelyplanet.com)	0	33.5
Vogue (forums.vogue.com.au)	0	21.3
Avg. for Non Technical	1.3	26.6

Table 1: Avg. Spam and Digressions per Thread

Statistics	Training	Test
No. of Threads	451	150
No. of Posts	2003	702
Avg. Replies	3.5	3.7
Avg. Answers per Thread	1.6	1.8

Table 2: Statistics of the Training and Test datasets

uses a different set of words, thus resulting in a low lexical overlap. This is also noticeable in the sample discussion of Figure 1. Here, similarity with the question post is actually misleading.

3.1.2 Spam and Digressions

When the similarity of answers to questions is low or unreliable, are there other properties of the post, the thread or its structure that we can rely on for accurate extraction? To explore such options, we conducted a small study to compare the amount of spam and digressions in technical forums versus other forums.

A spam is a completely off–topic post, while a digression is a post that is related to, but not discussing the same exact problem stated in the first post. Spam posts are usually advertisements generated automatically by spambots[3] and can be safely ignored without affecting the rest of the discussion. A digression, however, is still related to the overall discussion; at times, the result of this seemingly different problem might be useful in solving the main problem, and hence cannot be ignored completely. Nevertheless, for the purposes of this paper, we do not attempt to collate any of the sub–problems or their answers discussed within the thread.

Table 1 summarizes our findings on three technical and non-technical forums each. The numbers give the fraction (percentage) of the number of replies per thread, averaged over 15 randomly chosen threads from each of the forums. In the table, we note that the former has fewer spam and digressions, which suggests that it might be possible to find answers to the main question without regard to the question post or similarity to it, in a technical domain.

3.2 Features for Answer Extraction

The features that we study in this paper for the answer extraction task are detailed in Table 3. All Part-Of-Speech tags were generated using the Open NLP POS Tagger[4]. The column **Type** groups the features and **Availability** gives the fraction of forums in which each feature is publicly available, from 12 technical forums that we inspected. For example, the `Reply-to` structure of the thread may not always be displayed (Seo et al., 2009), and is usually flattened to their chronological order. Where the entry is `Always`, the data is always available, usually because it is computed from the text of the post.

Out of these, `Has_Link`, `Has_Navigation`, `Post_Belongs_to_First_N_Posts`, `In_Reply_to_Question_Author` and `Is_Replied_by_Question_Author` have not been proposed before, to the best of our knowledge.

4 Experiments

We crawled about 147,000 threads from Apple Discussions[5] of which we discarded those that had only 2 or fewer number of reply posts (88, 565 threads) and those that had more than 30

[3] http://en.wikipedia.org/wiki/Forum_spam
[4] http://opennlp.apache.org
[5] https://discussions.apple.com/community/iphone

	Feature	Description	Type	Availability
1	Has Noun	True or False depending on whether this post has nouns.	Lexical	Always
2	Has Proper Noun	True or False depending on whether this post has proper nouns.		
3	Has Verb	True or False depending on whether this post has verbs.		
4	No. of Non-Stopwords	The number of words in the post after discarding common English stopwords.		
5	Has Link	True if the post has a hyper–link, for example, to another thread or an online manual; else False.	Content	Always
6	Has Navigation	True if the post gives a navigational instruction like 'Settings→Sounds→Ringtone'; else False.		
7	Author Authority	A forum specific value - numerical (e.g. 1000 points) or categorical (e.g. Beginner) - assigned to the author, and indicative of their level of expertise in the context of the forum.	Forum Specific	100%
8	Post Rating	Numerical (e.g. 5 votes) or categorical (e.g. Helpful) value assigned by the question author or other users, indicating the usefulness of the post in answering the question.		36.3%
9	Relative Post Position in Thread	Computed from the ordinal position of the post in the thread, which is usually chronological. This value is grouped into 3 buckets - Beginning, Middle and End.	Structural	Always
10	Post Belongs to First N Posts	True if the ordinal number of the post is less than N, which was set to 5 in our experiments. Else, False.		
11	Post Author is Not Question Author	True if the two authors are different; else False.		
12	Time Difference to Question Post	Difference between the time of posting of the question post and the reply post, bucketized into hour, day and more.		100%
13	In Reply to Question Author	True or False depending on whether this post was in reply to the Question Author.	Reply-to	75%
14	Is Replied by Question Author	True or False depending on whether this post was replied by the Question Author.		
15	In Reply to Question Post	True or False depending on whether this post was in reply to the first post.		
16	No. of Replies to this Post	Number of replies to this post, as a fraction of the total number of replies in the thread.		
17	No. of Replies to Parent Post	Number of replies to the parent post, as a fraction of the total number of replies in the thread.		

Table 3: Features generated for a post, their types and availability

reply posts (845 threads), which gave us 58, 356 threads. From this, about 600 threads were randomly chosen for manual tagging. Posts in these threads were tagged as 'Answer' if they proposed an answer to the question post, and as 'Other', otherwise. If there were more than one answer post, ALL were marked as 'Answer's. Answers to other questions within the thread (digressions) were marked as 'Other'. Table 2 gives statistics of the training and test datasets.

4.1 Classification Experiments

We trained LibSVM classifiers[6] (Chang and Lin, 2011) on different sets of features as listed below, to obtain classifiers that mark each post as an answer or not, the precision–recall plot[7] of which is given in Figure 3:

- Question Similarity: uses the cosine similarity of the answer candidate and its respective question post, after discarding English stopwords and Porter stemming. As expected, it fails to give good accuracy for the task.
- Word: the features of this classifier are the words of the post after stopword removal and stemming. This is to test if answer posts use similar terminology which can be leveraged,

[6]With default settings (svm-type: C-SVC, kernel-type: RBF) and no tuning of hyperparameters

[7]Precision–Recall Plot: To obtain this plot, for each post in the test set, the trained classifier was used to get the probability of it being an answer. Let t be a threshold where all posts whose predicted probability is greater than t are labeled as answers. Then, t was varied from 0 to 1 in steps of 0.05 to get the different precision–recall values.

but gives an unimpressive performance.

- `Hong and Davison` (Hong and Davison, 2009): this classifier uses `Relative Post Position in Thread` and `Author Authority` alone, as reported in their paper. As can be seen from the figure, it gives better performance than the above two classifiers.
- `Forum Features`: this is the classifier that uses all features listed in Table 3 and is able to show a significant improvement over `Hong and Davison`.
- `Forum Features and Question Similarity`: it uses question similarity in addition to `Forum Features`, but overlaps almost completely with it indicating that similarity does not give any value addition.

4.2 Feature Selection Experiments

To study the relative importance of features for the answer extraction task, we performed two sets of feature selection experiments – a permutation test (Section 4.2.1) and a feature ablation study (Section 4.2.2), discussed in the below sections. The latter technique gauges the importance based on the performance of a classifier, while the former uses a statistical measure and does not depend on an external classifier.

4.2.1 Permutation Test

Permutation test (Good, 2000) is a popular non-parametric technique for statistical analysis of data and provides an empirical estimate for the distribution of the statistic under the null hypothesis ($\mathcal{H}_0$). Let l, m be the number of class 0, 1 samples respectively. For each feature, a test statistic θ (like information gain, mutual information) indicating similarity between the two class conditional densities, is calculated. Next, the data for the feature is randomly permuted and partitioned into sets of size l and m, on which the test statistic θ_p is calculated. This procedure is repeated over all possible such partitions of the feature into sets of size l and m. p-value is then estimated as the fraction of times $\theta_p > \theta$ and is an indicator of feature importance. Table 4 shows the Mutual Information scores for all features along with p-value. As a standard practice, any feature with p-value $<$ 0.05 (marked with *) is deemed important and the ones with p-value $\gg$ 0.05 (marked with **) are suggestively weak. From Table 4, it can be seen that `Question_Similarity` ranks very low on mutual Information. More detailed analysis is in Section 4.2.3.

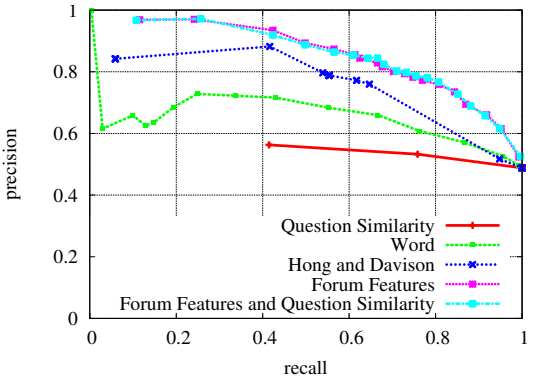

Figure 3: Precision Recall plots for answer classification

4.2.2 Feature Ablation Study

The goal of this study is to find the most reliable features that a classifier can use for the answer extraction task. In the feature ablation analysis (Arguello et al., 2009), at each step, each feature is individually omitted and the classifier is trained on the rest of the features. The importance of the feature is then measured as the classifier's percentage decrease in F–measure; higher the decrease, higher is the contribution. This process is repeated with the best feature of each step

Feature	Mutual Information
Relative Post Position in Thread	0.1256**
Post Author is Not Question Author	0.0977*
Has Link	0.0336**
Post Belongs to First N Posts	0.0299**
Time Difference to Question Post	0.0265**
Author Authority	0.0250**
Is Replied by Question Author	0.0223
No. of Replies to Parent Post	0.0201**
Has Proper Noun	0.0080*
Has Verb	0.0046*
No. of Non-Stopwords	0.0041*
Post Rating	0.0033
Has Noun	0.0027
Has Navigation	0.0019
In Reply to Question Post	0.0013
Question Similarity	0.0013
No. of Replies to Parent Post	0.0010*

Table 4: Permutation test results

Feature	Precision	Recall	F measure
Post Author is Not Question Author	0.82	0.69	0.75
Author Authority	0.81	0.65	0.72
Has Link	0.79	0.64	0.71
In Reply to Question Post	0.77	0.64	0.70
In Reply to Question Author	0.83	0.59	0.69
Relative Post Position in Thread	0.84	0.51	0.63
Is Replied by Question Author	0.70	0.46	0.55
Post Belongs to First N Posts	0.69	0.43	0.53
No. of Non-Stopwords	0.71	0.39	0.50
Has Verb	0.69	0.36	0.48
Has Navigation	0.70	0.35	0.46
Has Proper Noun	0.71	0.35	0.47
Post Rating	0.67	0.37	0.47
Has Noun	0.63	0.34	0.44
No. of Replies to this Post	0.63	0.34	0.44
Time Difference to Question Post	0.63	0.34	0.44
Question Similarity	0.63	0.34	0.44
No. of Replies to Parent Post	0.58	0.34	0.43

Table 5: Feature Ablation Study

progressively removed until all features are exhausted, to give them in their decreasing order of importance. For the experiment, we used a LibSVM classifier (Chang and Lin, 2011), and the results are in Table 5. The table lists the most helpful to the least helpful of features; the Precision, Recall and F–measure values (Chakrabarti, 2002) shown against each feature gives the accuracy numbers obtained when that feature and all those below it in the table were used to train the classifier. Detailed analysis is in Section 4.2.3.

4.2.3 Feature Selection Results Discussion

The results of permutation test in Table 4 and that of feature ablation study in Table 5 differ slightly because the latter is dependent on the performance of a classification algorithm while the former uses a statistical measure. However, it can be noted that, the following features show up as the best in both the tests:

- Post Author is Not Question Author
- Author Authority
- Has Link*
- Relative Post Position in Thread
- Is Replied by Question Author*
- Post Belongs to First N Posts*

Note that, out of the best 6 features, 3 were newly proposed in this paper (marked with *). Also note that, `Question_Similarity` ranks among the lowest in both the tests, thus showing its insignificance to this task. Another rather surprising observation is that `Post_Rating`, which gives the usefulness of the post, also does not contribute highly, which could be because, the number of posts that can be marked as `Helpful` is limited in the Apple discussions forum, thus missing out on useful suggestions that exceed the limit.

4.3 Feature Correlation Study

Correlation[8] refers to any of the broad class of statistical relationships between two random variables. In this paper, we use Pearson Product–Moment Correlation Coefficient[9] (Pearson's r), a widely used measure of correlation, defined as $\frac{cov(X,Y)}{\sigma_X \sigma_Y}$ for two variables X and Y, where cov and σ are the covariance and the standard deviation respectively.

[8] http://en.wikipedia.org/wiki/Correlation_and_dependence
[9] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

Feature	Correlation to Answer
Time Difference to Question Post	58.92
Author Authority	57.41
Is Replied by Question Author	40.92
Has Link	39.81
Post Rating	32.17
Relative Post Position in Thread	30.89
In Reply to Question Post	19.08
No. of Non-Stopwords	18.16
In Reply to Question Author	12.22
Post Belongs to First N Posts	12.06
Has Navigation	10.47
Has Proper Noun	8.73
No. of Replies to Parent Post	8.21
Post Author is Not Question Author	5.47
Has Noun	4.18
Has Verb	3.80
No. of Replies to this Post	1.44

Table 6: Feature – Answer Correlation

Feature A	Feature B	Correlation
In Reply to Question Post	In Reply to Question Author	86.72
Post Belongs to First N Posts	Relative Post Position in Thread	63.68
In Reply to Question Author	Post Author is Not Question Author	58.14
Time Difference to Question Post	Author Authority	55.41
In Reply to Question Post	Post Author is Not Question Author	51.94
Time Difference to Question Post	In Reply to Question Post	43.91
No. of Replies to Parent Post	In Reply to Question Post	41.68
Post Belongs to First N Posts	In Reply to Question Author	40.75
Is Replied by Question Author	Author Authority	38.89
Time Difference to Question Post	In Reply to Question Author	37.61
In Reply to Question Author	No. of Replies to Parent Post	37.29
Time Difference to Question Post	Is Replied by Question Author	35.40

Table 7: Feature – Feature Correlation

Table 6 gives the correlation of all the features to the answer label of the post. Higher the score, higher is the influence of the feature on the label. However, a higher score alone does not imply that the feature is important. If the feature is also highly correlated to many other features, it introduces redundancy, thus reducing its significance. The top 12 inter–feature correlation are listed in Table 7. Though `Time_Difference_to_Question_Post` shows the highest correlation to the answer label (Table 6), Table 7 shows that it is also highly correlated to many other features. Another contradicting result is that, in Section 4.2.3, `Post_Rating` was not ranked high. But Tables 6 and 7 show that it is highly correlated to the answer label and at the same time, not correlated to other features, suggesting that it might still prove to be useful.

Some of the features chosen in Section 4.2.3 from the feature selection experiments show correlation amongst themselves, as shown in Table 7. However, `Has_Link` proves to be a high ranking feature according to both (a) Feature Selection, as well as, (b) Feature Correlation, since it highly correlates to the answer, but does not overlap with other features.

5 Conclusion

In this paper, we studied the contribution and importance of similarity to question in extracting answers from technical discussions, and showed that this feature does not contribute significantly towards the task of answer extraction, contrary to its perceived significance. We also presented the characteristics of technical discussion forums that distinguish them from other domains thus suggesting that it is possible to extract answers with high accuracy using other non-similarity based features when question similarity is unreliable, which was then demonstrated through experiments. We also presented a careful study of all features to determine which ones contributed highly to this task. The results of one set of experiments – Feature Selection – showed that out of the 6 best features, 3 were the ones newly proposed in this paper. Further analysis using Feature Correlation tests showed that all but one of the 6 best features from the former experiments were in fact highly correlated amongst themselves. The one feature that proved to be highly important in all the tests is `Has_Link`, proposed for the first time in this paper.

As part of future work, we aim to test the importance of the features proposed in this paper in other domains, and the marginal improvement in accuracy that they can provide even in the presence of high similarity of answers to question posts.

References

Arguello, J., Callan, J., and Diaz, F. (2009). Classification-based resource selection. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1277–1286, New York, NY, USA. ACM.

Chakrabarti, S. (2002). *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman.

Chang, C. and Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Cong, G., Wang, L., Lin, C.-Y., Song, Y.-I., and Sun, Y. (2008). Finding question-answer pairs from online forums. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 467–474, New York, NY, USA. ACM.

Ding, S., Cong, G., Lin, C.-Y., and Zhu, X. (2008). Using conditional random fields to extract contexts and answers of questions from online forums. In *ACL*, pages 710–718.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, 2nd edition.

Hong, L. and Davison, B. D. (2009). A classification-based approach to question answering in discussion boards. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 171–178, New York, NY, USA. ACM.

Huang, J., Zhou, M., and Yang, D. (2007). Extracting chatbot knowledge from online discussion forums. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 423–428, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kim, S. N., Wang, L., and Baldwin, T. (2010). Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 192–202, Stroudsburg, PA, USA. Association for Computational Linguistics.

Otterbacher, J., Erkan, G., and Radev, D. R. (2005). Using random walks for question-focused sentence retrieval. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 915–922, Stroudsburg, PA, USA. Association for Computational Linguistics.

Porter, M. (1980). An algorithm for suffix stripping. In *Program*.

Seo, J., Croft, W. B., and Smith, D. A. (2009). Online community search using thread structure. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1907–1910, New York, NY, USA. ACM.

Shrestha, L. and McKeown, K. (2004). Detection of question-answer pairs in email conversations. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yang, W.-Y., Cao, Y., and Lin, C.-Y. (2009). A structural support vector method for extracting contexts and answers of questions from online forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 514–523, Stroudsburg, PA, USA. Association for Computational Linguistics.