

Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites

Delphine Bernhard and Iryna Gurevych

Ubiquitous Knowledge Processing Lab

Computer Science Department

Technische Universität Darmstadt, Hochschulstraße 10

D-64289 Darmstadt, Germany

{delphine|gurevych}@tk.informatik.tu-darmstadt.de

Abstract

Information overload is a well-known problem which can be particularly detrimental to learners. In this paper, we propose a method to support learners in the information seeking process which consists in answering their questions by retrieving question paraphrases and their corresponding answers from social Q&A sites. Given the novelty of this kind of data, it is crucial to get a better understanding of how questions in social Q&A sites can be automatically analysed and retrieved. We discuss and evaluate several pre-processing strategies and question similarity metrics, using a new question paraphrase corpus collected from the WikiAnswers Q&A site. The results show that viable performance levels of more than 80% accuracy can be obtained for the task of question paraphrase retrieval.

1 Introduction

Question asking is an important component of efficient learning. However, instructors are often overwhelmed with students' questions and are therefore unable to provide timely answers (Feng et al., 2006). Information seeking is also rendered difficult by the sheer amount of learning material available, especially online. The use of advanced information retrieval and natural language processing techniques to answer learners' questions and reduce the difficulty of information seeking is henceforth particularly promising. Question Answering (QA) systems seem well suited for this task since they aim at generating precise answers to natural language questions instead of merely returning documents con-

taining answers. However, QA systems have to be adapted to meet learners' needs. Indeed, learners do not merely ask concrete or factoid questions, but rather open-ended, explanatory or methodological questions which cannot be answered by a single sentence (Baram-Tsabari et al., 2006). Despite a recent trend to render the tasks more complex at large scale QA evaluation campaigns such as TREC or CLEF, current QA systems are still ill-suited to meet these requirements.

A first alternative to full-fledged QA consists in making use of already available question and answer pairs extracted from archived discussions. For instance, Feng et al. (2006) describe an intelligent discussion bot for answering student questions in forums which relies on answers retrieved from an annotated corpus of discussions. This renders the task of QA easier since answers do not have to be generated from heterogeneous documents by the system. The scope of such a discussion bot is however inherently limited since it relies on manually annotated data, taken from forums within a specific domain.

We propose a different solution which consists in tapping into the wisdom of crowds to answer learners' questions. This approach provides the compelling advantage that it utilises the wealth of already answered questions available in online social Q&A sites. The task of Question Answering can then be boiled down to the problem of finding question paraphrases in a database of answered questions. Question paraphrases are questions which have identical meanings and expect the same answer while presenting alternate wordings. Several methods have already been proposed to identify question

paraphrases mostly in FAQs (Tomuro and Lytinen, 2004) or search engine logs (Zhao et al., 2007).

In this paper, we focus on the problem of question paraphrase identification in social Q&A sites within a realistic information seeking scenario: given a user question, we want to retrieve the best matching question paraphrase from a database of previously answered questions in order to display the corresponding answer. The use of social Q&A sites for educational applications brings about new challenges linked to the variable quality of social media content. As opposed to questions in FAQs, which are subject to editorial control, questions in social Q&A sites are often ill-formed or contain spelling errors. It is therefore crucial to get a better understanding of how they can be automatically analysed and retrieved. In this work, we focus on several pre-processing strategies and question similarity measures applied to the task of identifying question paraphrases in a social Q&A site. We chose WikiAnswers which has been ranked by comScore as the first fastest growing domain of the top 1,500 in the U.S. in 2007.

The remainder of the paper is organised as follows. Section 2 first discusses related work on paraphrase identification and question paraphrasing. Section 3 then presents question and answer repositories with special emphasis on social Q&A sites. Our methods to identify question paraphrases are detailed in section 4. Finally, we present and analyse the experimental results obtained in section 5 and conclude in section 6.

2 Related Work

The identification of question paraphrases in question and answer repositories is related to research focusing on sentence paraphrase identification (section 2.1) and query paraphrasing (section 2.2). The specific features of question paraphrasing have also already been investigated (section 2.3).

2.1 Sentence Paraphrase Identification

Paraphrases are alternative ways to convey the same information (Barzilay and McKeown, 2001). Paraphrases can be found at different levels of linguistic structure: words, phrases and whole sentences. While word and phrasal paraphrases can be assimilated to the well-studied notion of syn-

onymy, sentence level paraphrasing is more difficult to grasp and cannot be equated with word-for-word or phrase-by-phrase substitution since it might entail changes in the structure of the sentence (Barzilay and Lee, 2003). In practice, sentence paraphrases are identified using various string and semantic similarity measures which aim at capturing the semantic equivalence of the sentences being compared. String similarity metrics, when applied to sentences, consist in comparing the words contained in the sentences. There exist many different string similarity measures: word overlap (Tomuro and Lytinen, 2004), longest common subsequence (Islam and Inkpen, 2007), Levenshtein edit distance (Dolan et al., 2004), word n-gram overlap (Barzilay and Lee, 2003) etc. Semantic similarity measures are obtained by first computing the semantic similarity of the words contained in the sentences being compared. Mihalcea et al. (2006) use both corpus-based and knowledge-based measures of the semantic similarity between words. Both string similarity and semantic similarity might be combined: for instance, Islam and Inkpen (2007) combine semantic similarity with longest common subsequence string similarity, while Li et al. (2006) make additional use of word order similarity.

2.2 Query Paraphrasing

In Information Retrieval, research on paraphrasing is dedicated to query paraphrasing which consists in identifying semantically similar queries. The overall objective is to discover frequently asked questions and popular topics (Wen et al., 2002) or suggest related queries to users (Sahami and Heilman, 2006). Traditional string similarity metrics are usually deemed inefficient for such short text snippets and alternative similarity metrics have therefore been proposed. For instance, Wen et al. (2002) rely on user click logs, based on the idea that queries and questions which result in identical document clicks are bound to be similar.

2.3 Question Paraphrasing

Following previous research in this domain, we define question paraphrases as questions which have all the following properties: (a) they have identical meanings, (b) they have the same answers, and (c) they present alternate wordings. Question para-

phrases differ from sentence paraphrases by the additional condition (b). This definition encompasses the following questions, taken from the WikiAnswers web site: *How many ounces are there in a pound?*, *What's the number of ounces per pound?*, *How many oz. in a lb.?*

Question paraphrases share some properties both with declarative sentence paraphrases and query paraphrases. On the one hand, questions are complete sentences which differ from declarative sentences by their specific word order and the presence of question words and a question focus. On the other hand, questions are usually associated with answers, which makes them similar to queries associated with documents. Accordingly, research on the identification of question paraphrases in Q&A repositories builds upon both sentence and query paraphrasing.

Zhao et al. (2007) propose to utilise user click logs from the Encarta web site to identify question paraphrases. Jeon et al. (2005) employ a related method, in that they identify similar answers in the Naver Question and Answer database to retrieve semantically similar questions, while Jijkoun and de Rijke (2005) include the answer in the retrieval process to return a ranked list of QA pairs in response to a user's question. Lytinen and Tomuro (2002) suggest yet another feature to identify question paraphrases, namely question type similarity, which consists in determining a question's category in order to match questions only if they belong to the same category.

Our focus is on question paraphrase identification in social Q&A sites. Previous research was mostly based on question paraphrase identification in FAQs (Lytinen and Tomuro, 2002; Tomuro and Lytinen, 2004; Jijkoun and de Rijke, 2005). In FAQs, questions and answers are edited by expert information suppliers, which guarantees stricter conformance to conventional writing rules. In social Q&A sites, questions and answers are written by users and may hence be error-prone. Question paraphrase identification in social Q&A sites has been little investigated. To our knowledge, only Jeon et al. (2005) have used data from a Q&A site, namely the Korean Naver portal, to find semantically similar questions. Our work is related to the latter since it employs a similar dataset, yet in English and from a different social Q&A site.

3 Question and Answer Repositories

3.1 Properties of Q&A Repositories

Question and answer repositories have existed for a long time on the Internet. Their form has evolved from Frequently Asked Questions (FAQs) to Ask-an-expert services (Baram-Tsabari et al., 2006) and, even more recently, social Q&A sites. The latest, which include web sites such as Yahoo! Answers and AnswerBag, provide portals where users can ask their own questions as well as answer questions from other users. Social Q&A sites are increasingly popular. For instance, in December 2006 Yahoo! Answers was the second-most visited education/reference site on the Internet after Wikipedia according to the Hitwise company (Prescott, 2006). Even more strikingly, the Q&A portal Naver is the leader of Internet search in South Korea, well ahead of Google (Sang-Hun, 2007).

Several factors might explain the success of social Q&A sites:

- they provide answers to questions which are difficult to answer with a traditional Web search or using static reference sites like Wikipedia, for instance opinions or advice about a specific family situation or a relationship problem;
- questions can be asked anonymously;
- users do not have to browse a list of documents but rather obtain a complete answer;
- the answers are almost instantaneous and numerous, due to the large number of users.

Social Q&A sites record the questions and their answers online, and thus constitute a formidable repository of collective intelligence, including answers to complex questions. Moreover, they make it possible for learners to reach other people worldwide. The relevance of social Q&A sites for learning has been little investigated. To our knowledge, there has been only one study which has shown that Korean users of the Naver Question and Answer platform consider that social Q&A sites can satisfactorily and reliably support learning (Lee, 2006).

3.2 WikiAnswers

For our experiments we collected a dataset of questions and their paraphrases from the WikiAnswers

web site. WikiAnswers¹ is a social Q&A site similar to Yahoo! Answers and AnswerBag. As of February 2008, it contained 1,807,600 questions, sorted in 2,404 categories (Answers Corporation, 2008).

Compared with its competitors, the main originality of WikiAnswers is that it relies on the *wiki* technology used in Wikipedia, which means that answers can be edited and improved over time by all contributors. Moreover, the Answers Corporation, which owns the WikiAnswers site, explicitly targets educational uses and even provides an educator toolkit.² Another interesting property of WikiAnswers is that users might manually tag question reformulations in order to prevent the duplication of questions asking the same thing in a different way. When a user enters a question which is not already part of the question repository, the web site displays a list of questions already existing on the site and similar to the one just asked by the user. The user may then freely select the question which paraphrases her question, if available, or choose to view one of the proposed alternatives without labelling it as a paraphrase. The user-labelled question reformulations are stored in order to retrieve the same answer when the question rephrasing is asked again. The wiki principle holds for the stored reformulations too, since they can subsequently be edited by other users if they consider that they correspond to another existing question or actually ask an entirely new question. It should be noted that contributors get not reward in terms of trust points for providing or editing alternate wordings for questions.

We use the wealth of question paraphrases available on the WikiAnswers website as the so called user generated gold standard in our question paraphrasing experiments. User generated gold standards have been increasingly used in recent years for research evaluation purposes, since they can be easily created from user annotated content. For instance, Mihalcea and Csomai (2007) use manually annotated keywords (links to other articles) in Wikipedia articles to evaluate their automatic keyword extraction and word sense disambiguation algorithms. Similarly, quality assessments provided by users in social media have been used as gold

standards for the automatic assessment of post quality in forum discussions (Weimer et al., 2007). It should however be kept in mind that user generated gold standards are not perfect, as already noticed by (Mihalcea and Csomai, 2007), and thus constitute a trade-off solution.

For the experiments described hereafter, we randomly extracted a collection of 1,000 questions along with their paraphrases (totalling 7,434 question paraphrases) from 100 randomly selected FAQ files in the Education category of the WikiAnswers web site. In what follows, the corpus of 1,000 questions is called the *target questions* collection, while the 7,434 question paraphrases constitute the *input questions* collection. The objective of the task is to retrieve the corresponding target question for each input question. The target question selected is the one which maximises the question similarity value (see section 4.2).

4 Method

In order to rate the similarity of input and target questions, we have first pre-processed both the input and target questions and then experimented with several question similarity measures.

4.1 Pre-processing

We employ the following steps in pre-processing the questions:

Stop words elimination however, we keep question words such as *how*, *why*, *what*, etc. since these make it possible to implicitly identify the question type (Lytinen and Tomuro, 2002; Jijkoun and de Rijke, 2005)

Stemming using the Porter Stemmer³

Lemmatisation using the TreeTagger⁴

Spelling correction using a statistical system based on language modelling (Norvig, 2007).⁵

³<http://snowball.tartarus.org/>

⁴<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

⁵We used a Java implementation of the system, jSpellCorrect available at <http://developer.gauger.org/jspellcorrect/>, trained with the default English training data, to which we appended the myspell English dictionaries.

¹<http://wiki.answers.com/>

²<http://educator.answers.com/>

Stop words were eliminated in all the experimental settings, while stemming and lemmatisation were optionally performed to evaluate the effects of these pre-processing steps on the identification of question paraphrases. We added spelling correction to the conventional pre-processing steps, since we target paraphrasing of questions which often contain spelling errors, such as *When was indoor plumbing invented?* or *What is the largest country in the western Hemipher?* Other related endeavours at retrieving question paraphrases have identified spelling mistakes in questions as a significant source of errors in the retrieval process, but have not attempted to solve this problem (Jijkoun and de Rijke, 2005; Zhao et al., 2007).

4.2 Question Similarity Measures

We have experimented with several kinds of question similarity measures, belonging to two different families of measures: string similarity measures and vector space measures.

4.3 String Similarity Measures

Basic string similarity measures compare the words contained in the questions without taking word frequency into account.

Matching coefficient The matching coefficient of two questions q_1 and q_2 represented by the set of distinct words Q_1 and Q_2 they contain is computed as follows (Manning and Schütze, 1999):

$$\text{matching coefficient} = \frac{|Q_1 \cap Q_2|}{|Q_1 \cup Q_2|}$$

Overlap coefficient The overlap coefficient is computed according to the following formula (Manning and Schütze, 1999):

$$\text{overlap coefficient} = \frac{|Q_1 \cap Q_2|}{\min(|Q_1|, |Q_2|)}$$

Normalised Edit Distance The edit distance of two questions is the number of words that need to be substituted, inserted, or deleted, to transform q_1 into q_2 . In order to be able to compare the edit distance with the other metrics, we have used the following formula (Wen et al., 2002) which normalises the minimum edit distance by the length of the longest question and transforms it into a similarity metric:

$$\text{normalised edit distance} = 1 - \frac{\text{edit_dist}(q_1, q_2)}{\max(|q_1|, |q_2|)}$$

Word Ngram Overlap This metric compares the word n -grams in both questions:

$$\text{ngram overlap} = \frac{1}{N} \sum_{n=1}^N \frac{|G_n(q_1) \cap G_n(q_2)|}{\min(|G_n(q_1)|, |G_n(q_2)|)}$$

where $G_n(q)$ is the set of n -grams of length n in question q and N usually equals 4 (Barzilay and Lee, 2003; Cordeiro et al., 2007).

4.4 Vector Space Based Measures

Vector space measures represent questions as real-valued vectors by taking word frequency into account.

Term Vector Similarity Questions are represented as term vectors V_1 and V_2 . The feature values of the vectors are the *tf.idf* scores of the corresponding terms:

$$\text{tf.idf} = (1 + \log(\text{tf})) * \log \frac{N + 1}{df}$$

where tf is equal to the frequency of the term in the question, N is the number of target questions and df is the number of target questions in which the term occurs, computed by considering the input question as part of the target questions collection (Lytinen and Tomuro, 2002).

The similarity of an input question vector and a target question vector is determined by the cosine coefficient:

$$\text{cosine coefficient} = \frac{V_1 \cdot V_2}{|V_1| \cdot |V_2|}$$

Lucene's Extended Boolean Model The problem of question paraphrase identification can be cast as an Information Retrieval problem, since in real-world applications the user posts a question and the system returns the best matching questions from its database. We have therefore tested the results obtained using an Information Retrieval system, namely Lucene⁶, which combines the Vector Space Model and the Boolean model. Lucene has already been successfully used by Jijkoun and de Rijke (2005) to retrieve answers from FAQ web pages by combining several fields: question text, answer text and the whole FAQ page. The target questions are indexed as documents and retrieved by transforming the input questions into queries.

⁶<http://lucene.apache.org/java/docs/>

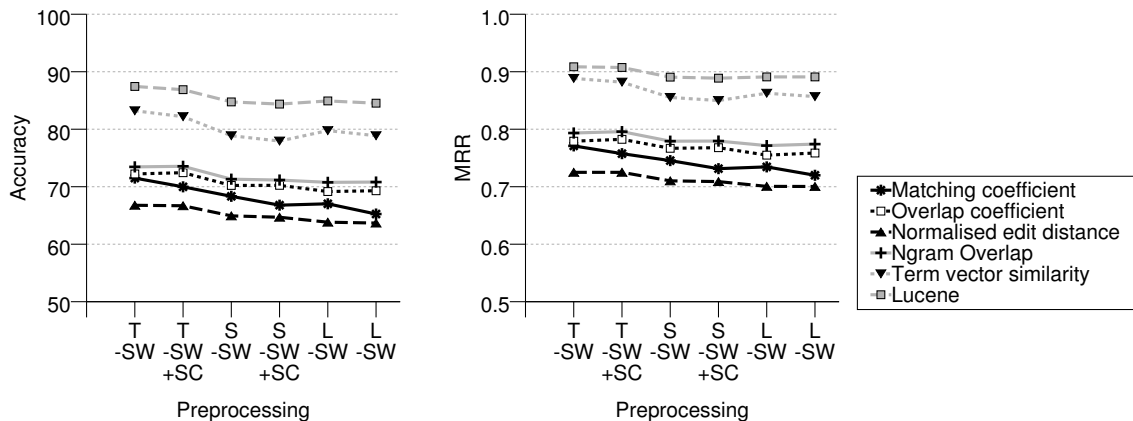


Figure 1: Accuracy (%) and Mean Reciprocal Rank obtained for different question similarity measures and pre-processing strategies: tokens (T), stemming (S), lemmatisation (L), stop words removal (-SW), spelling correction (+SC).

5 Evaluation and Experimental Results

5.1 Evaluation Measures

We use the following evaluation measures for evaluating the results:

Mean Reciprocal Rank For a question, the reciprocal rank RR is $\frac{1}{r}$ where r is the rank of the correct target question, or zero if the target question was not found. The Mean Reciprocal Rank (MRR) is the mean of the reciprocal ranks over all the input questions.

Accuracy We define accuracy as $\text{Success}@1$, which is the percentage of input questions for which the correct target question has been retrieved at rank 1.

5.2 Experimental Results

Figure 1 displays the accuracy and the mean reciprocal ranks obtained with the different question similarity measures and pre-processing strategies. As could be expected, vector space based similarity measures are consistently more accurate than simple string similarity measures. Moreover, both the accuracy and the MRR are rather high for vector space metrics (accuracy around 80-85% and MRR around 0.85-0.9), which shows that good results can be obtained with these retrieval mechanisms. Additional pre-processing, i.e. stemming, lemmatisation

and spelling correction, does not ameliorate the tokens minus stop words (T -SW) baseline.

5.3 Detailed Error Analysis

Stemming and lemmatisation Morphological pre-processing brings about mitigated improvements over the tokens-only baseline. On the one hand, it improves paraphrase retrieval for questions containing morphological variants of the same words such as *What are analogies for mitochondria?* and *What is an analogy for mitochondrion?* On the other hand, it also leads to false positives, such as *How was calculus started?*, stemmed as *How was calculus start?* and lemmatised as *How be calculus start?*, which is mapped by Lucene to the question *How could you start your MA English studies?* instead of *Who developed calculus?*. The negative effect of stemming has already been identified by (Jijkoun and de Rijke, 2005) and our results are consistent with this previous finding.

Spelling correction We expected that spelling correction would have a positive impact on the results. There are indeed cases when spelling correction helps. For instance, given the question *How do you become an anesthesiologist?*, it is impossible to retrieve the target question *How many years of medical school do you need to be an anesthesiologist?* without spelling correction since *anesthesiologist* is ill-spelled both in the paraphrase and the target question.

error patterns, as suggested by our detailed results analysis. Figure 2 confirms that the pre-processing techniques as well as the methods employed result in dissimilar error patterns. We therefore combined several methods and pre-processing techniques in order to verify if we could improve accuracy.

We obtained the best results by performing a majority vote combination of the following methods and pre-processing strategies: Lucene, Term Vector Similarity with stemming and Ngram Overlap with spelling correction. The combination yielded an accuracy of 88.3%, that is 0.9% over the best Lucene results with an accuracy of 87.4%.

6 Conclusion and Outlook

In this paper, we have shown that it is feasible to answer learners' questions by retrieving question paraphrases from social Q&A sites. As a first step towards this objective, we investigated several question similarity metrics and pre-processing strategies, using WikiAnswers as input data and user generated gold standard. The approach is however not limited to this dataset and can be easily applied to retrieve question paraphrases from other social Q&A sites.

We also performed an extended failure analysis which provided useful insights on how results could be further improved by performing named entity analysis and using semantic similarity metrics.

Another important challenge in using social Q&A sites for educational purposes lies in the quality of the answers retrieved from such sites. Previous research on the identification of high quality content in social Q&A sites has defined answer quality in terms of correctness, well-formedness, readability, objectivity, relevance, utility and interestingness (Jeon et al., 2006; Agichtein et al., 2008). It is obvious that all these elements play an important role in the acceptance of the answers by learners. We therefore plan to integrate quality measures in the retrieval process and to perform evaluations in a real educational setting.

Acknowledgments

This work was supported by the Emmy Noether Programme of the German Research Foundation (DFG) under grant No. GU 798/3-1.

References

- Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 183–194.
- Answers Corporation. 2008. WikiAnswers Journalist Quick Guide. [Online; visited March 4, 2008]. http://site.wikianswers.com/resources/WikiAnswers_1-pager.pdf.
- Ayelet Baram-Tsabari, Ricky J. Sethi, Lynn Bry, and Anat Yarden. 2006. Using questions sent to an Ask-A-Scientist site to identify children's interests in science. *Science Education*, 90(6):1050–1072.
- Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of NAACL-HLT 2003*, pages 16–23. Association for Computational Linguistics.
- Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *ACL '01: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57. Association for Computational Linguistics.
- João Cordeiro, Gaël Dias, and Pavel Brazdil. 2007. Learning Paraphrases from WNS Corpora. In David Wilson and Geoff Sutcliffe, editors, *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 193–198, Key West, Florida, USA, May 7-9. AAAI Press.
- Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, pages 350–356. Association for Computational Linguistics.
- Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. An Intelligent Discussion-Bot for Answering Student Queries in Threaded Discussions. In *Proceedings of the 11th international conference on Intelligent user interfaces (IUI'06)*, pages 171–177.
- Aminul Islam and Diana Inkpen. 2007. Semantic Similarity of Short Texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria, September.
- Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 84–90.

- Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. 2006. A framework to predict the quality of answers with non-textual features. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 228–235.
- Valentin Jijkoun and Maarten de Rijke. 2005. Retrieving answers from frequently asked questions pages on the web. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 76–83.
- Yu Sun Lee. 2006. Toward a New Knowledge Sharing Community: Collective Intelligence and Learning through Web-Portal-Based Question-Answer Services. Masters of arts in communication, culture & technology, Faculty of the Graduate School of Arts and Sciences of Georgetown University, May. [Online; visited February 15, 2008], <http://hdl.handle.net/1961/3701>.
- Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. 2006. Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150.
- Steven L. Lytinen and Noriko Tomuro. 2002. The Use of Question Types to Match Questions in FAQFinder. In *Proceedings of the 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pages 46–53.
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.
- Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on information and knowledge management*, pages 233–242.
- Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of AAAI 2006*, Boston, July.
- Peter Norvig. 2007. How to Write a Spelling Corrector. [Online; visited February 22, 2008]. <http://norvig.com/spell-correct.html>.
- Lee Ann Prescott. 2006. Yahoo! Answers Captures 96% of Q and A Market Share. Hitwise Intelligence [Online; visited February 26, 2008]. http://weblogs.hitwise.com/leeann-prescott/2006/12/yahoo_answers_captures_96_of_q.html.
- Mehran Sahami and Timothy D. Heilman. 2006. A web-based kernel function for measuring the similarity of short text snippets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 377–386.
- Choe Sang-Hun. 2007. To outdo Google, Naver taps into Korea's collective wisdom. International Herald Tribune, July 4. <http://www.iht.com/articles/2007/07/04/technology/naver.php>.
- Noriko Tomuro and Steven Lytinen. 2004. Retrieval Models and Q&A Learning with FAQ Files. In Mark T. Maybury, editor, *New Directions in Question Answering*, pages 183–194. AAAI Press.
- Markus Weimer, Iryna Gurevych, and Max Mühlhäuser. 2007. Automatically Assessing the Post Quality in Online Discussions on Software. In *Proceedings of the Demo and Poster Sessions of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 125–128, Prague, Czech Republic, June. Association for Computational Linguistics.
- Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. 2002. Query clustering using user logs. *ACM Trans. Inf. Syst.*, 20(1):59–81.
- Shiqi Zhao, Ming Zhou, and Ting Liu. 2007. Learning Question Paraphrases for QA from Encarta Logs. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1795–1801, Hyderabad, India, January 6-12.