Generating High Quality Questions from Low Quality Questions

Kateryna Ignatova and **Delphine Bernhard** and **Iryna Gurevych**
Ubiquitous Knowledge Processing Lab
Computer Science Department
Technische Universität Darmstadt, Hochschulstraße 10
D-64289 Darmstadt, Germany
{ignatova|delphine|gurevych}@tk.informatik.tu-darmstadt.de

Abstract

We propose an original question generation task consisting in generating high quality questions from low quality questions. Such a system could be used to suggest improvements to questions asked both on social Q&A sites and to automatic QA systems. Low quality question datasets can be easily collected from the Web, based on the questions asked to social Q&A sites, such as WikiAnswers.

1 Introduction

It is well known that asking good questions is a difficult task (Graesser and Person, 1994). Plenty of evidence can be found in the constantly growing social Question and Answer (Q&A) platforms, such as Yahoo! Answers[1] or WikiAnswers[2], where users can ask questions and get answers from other users. The quality of both the contents and the formulation of questions asked on these sites is often low, which has a detrimental effect on the quality of the answers retrieved (Agichtein et al., 2008). There is also a discrepancy between question asking practice, as displayed in social Q&A sites or query logs, and current automatic Question Answering (QA) systems which expect perfectly formulated questions (Rus et al., 2007). We therefore propose to apply Question Generation (QG) to low quality questions in order to automatically improve question quality and increase the users' chance of getting answers to their questions.

[1] http://answers.yahoo.com/
[2] http://wiki.answers.com/

2 Characteristics of Low Quality Questions

While the answerer's lack of knowledge is obviously the most common reason for getting unsatisfactory answers, the askers' inability to formulate grammatically correct or clear questions is yet another major cause for unanswered or badly answered questions. We have identified at least five main factors which negatively influence the answer finding process, both in social Q&A sites and QA systems: misspellings, Internet slang, ill-formed syntax, structurally inappropriate questions such as queries expressed by keywords, and ambiguity. Table 1 lists example questions for these five issues; an exclamation mark signals that the factor negatively influences results in the corresponding environment.

Factor	QA	Q&A	Example
Misspelling	!		***Hou** to cook pasta?*
Internet slang	!		*How **r** plants used **4** medicine?*
Ill-formed syntax	!		*What Alexander Pushkin famous for?*
Keyword search	!	!	*Drug classification, pharmacodynamics*
Ambiguity	!	!	*What is the population of **Washington**?*

Table 1: Question quality factors.

Misspellings, Internet slang and syntactic ill-formedness are common problems which have to be faced when working with spontaneous user input. In order to be able to better quantify these phenomena we manually analyzed 755 questions ex-

tracted from the Yahoo! Answers social Q&A site and found that 18% of them were misspelled, 8% contained Internet slang, and 20% were ill-formed. While misspellings, Internet slang or ill-formedness usually do not prevent human users from understanding and answering questions, noisy text data cannot be correctly processed by most NLP tools, which therefore impedes automatic answer extraction in QA systems.

Keyword queries are the natural way for most people to look for information. In an experiment destined to gather real user questions on a university website, Dale (2007) showed that only about 8% of the queries submitted were questions, despite instructions and examples destined to make users ask full natural language questions. Another study by Spink and Ozmultu (2002) showed that only half of the queries asked to the Ask Jeeves QA Web Search engine were questions. Keyword queries are also commonly found on social Q&A sites. In such cases, the user's information need and the type of the question are unclear, which represents a significant problem for both machines and humans. For example, the following question from Yahoo! Answers *"Drug classification, pharmacodynamics, Drug databases"* gets a low quality answer *"Is there a question here? Ask a question if you want an answer"*. In a similar way, the QuALiM[3] QA system produces the following message in response to a keyword search: *"Hint: Asking proper questions should improve your results"*.

Ambiguity is another important issue which we would like to mention here. The question *"What is the population of **Washington**?"* is ambiguous for both QA systems and humans answering questions on Q&A platforms, and might therefore need special question generation approaches for reformulating the question in order to resolve ambiguity. Since disambiguation usually requires additional contextual information which might not always be available, we will not tackle this issue here.

3 Task Description

The various examples given in the previous section call for the application of question generation techniques to improve low quality questions and subsequently ameliorate retrieval results. The steps which have to be performed to improve questions range from correcting spelling errors to generating full-fledged questions given only a set of keywords.

3.1 Main Subtasks

Spelling and Grammatical Error Correction
The performance of spelling correction depends on the training lexicon, and therefore unrecognized words, which frequently occur in user-generated content, lead to wrong corrections. For instance, as we have shown in (Bernhard and Gurevych, 2008), the question *"What are the **GRE** score required to get into top 100 US universities?"*, where GRE stands for Graduate Record Examination, is badly corrected as *"What are the **are** score required to get into top 100 US universities?"* by a general purpose dictionary-based spelling correction system (Norvig, 2007). Spelling correctors on social Q&A sites fare no better. For example, in response to the question *"Wat r Wi-Fi systems?"*, both Yahoo! Answers and WikiAnswers suggest to correct the word *'Wi-Fi'*, but do not complain about the Internet slang words *'wat'* and *'r'*. Internet slang should therefore be tackled in parallel to spelling correction. Grammar checking and correction is yet another complex issue. A thorough study of the kind of grammatical errors found in questions would be needed in order to correctly handle them.

Question generation from a set of keywords
The task of question generation from keywords is very challenging and, to our knowledge, has not been addressed yet. The converse task, extracting keywords from user queries, has already been largely investigated both for Question Answering and Information Retrieval (Kumaran and Allan, 2006). Moreover, it is important to generate not only grammatical but also *important* (Vanderwende, 2008) and *useful* questions (Song et al., 2008). For this reason, we suggest to reuse the questions previously asked on Q&A platforms to learn the preferred question types and patterns given specific keywords. Additional information available on Q&A sites, such as user profiles, could further help to generate questions by taking into consideration the preferences of a user. The analysis of user profiles has been already proposed by Jeon et al. (2006)

[3] http://demos.inf.ed.ac.uk:8080/qualim/

and Liu et al. (2008) for predicting the quality of answers and the satisfaction of information seekers.

3.2 Datasets

Evaluation datasets for the proposed task can be easily obtained online. For instance, the WikiAnswers social Q&A site repertories questions as well as manually tagged paraphrases for the questions which include low quality question reformulations. For example, the question *"What is the height of the Eiffel Tower?"* is mapped to the following paraphrases: *"Height eiffel tower?", "Whats the height of effil tower?", "What is the eiffel towers height?", "The height of the eiffel tower in metres?"*, etc.[4]

3.3 Evaluation

Beside an intrinsic evaluation of the quality of generated questions, we propose an extrinsic evaluation of the generated questions which would consist in measuring the impact on automatic QA of (i) low quality questions vs. (ii) automatically generated high quality questions.

Acknowledgments

This work has been supported by the Emmy Noether Program of the German Research Foundation (DFG) under the grant No. GU 798/3-1, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under the grant No. I/82806.

References

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining (WSDM '08)*, pages 183–194.

Delphine Bernhard and Iryna Gurevych. 2008. Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications, ACL 2008*, pages 44–52, Columbus, Ohio, USA, June 19.

Robert Dale. 2007. Industry Watch: Is 2007 the Year of Question-Answering? *Natural Language Engineering*, 13(1):91–94.

Arthur C. Graesser and Natalie K. Person. 1994. Question Asking During Tutoring. *American Educational Research Journal*, 31:104–137.

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 *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06)*, pages 228–235.

Giridhar Kumaran and James Allan. 2006. A Case for Shorter Queries, and Helping User Create Them. In *Proceedings of the Human Language Technology Conference and the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 220–227, Rochester, New York, April. Association for Computational Linguistics.

Yandong Liu, Jiang Bian, and Eugene Agichtein. 2008. Predicting Information Seeker Satisfaction in Community Question Answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08)*, pages 483–490, New York, NY, USA. ACM.

Peter Norvig. 2007. How to Write a Spelling Corrector. [Online; visited February 22, 2008]. `http://norvig.com/spell-correct.html`.

Vasile Rus, Zhiqiang Cai, and Arthur C. Graesser. 2007. Experiments on Generating Questions About Facts. In Alexander F. Gelbukh, editor, *Proceedings of CICLing*, volume 4394 of *Lecture Notes in Computer Science*, pages 444–455. Springer.

Young-In Song, Chin-Yew Lin, Yunbo Cao, and Hae-Chang Rim. 2008. Question Utility: A Novel Static Ranking of Question Search. In *Proceedings of AAAI 2008 (AI and the Web Special Track)*, pages 1231–1236, Chicago, Illinois, USA, July.

Amanda Spink and H. Cenk Ozmultu. 2002. Characteristics of question format web queries: an exploratory study. *Information Processing & Management*, 38(4):453–471, July.

Lucy Vanderwende. 2008. The Importance of Being Important: Question Generation. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, Arlington, VA, September.

[4] For the full list of paraphrases, see `http://wiki.answers.com/Q/What_is_the_height_of_the_Eiffel_Tower`