Automatic Question Generation from Queries

Chin-Yew Lin
Microsoft Research Asia
5F, Beijing Sigma Center
49 Zhichun Road, Haidian District
Beijing 100190, P.R. China
cyl@microsoft.com

Abstract

With increasing popularity of community-based question answering services such as Yahoo! Answers, huge amounts of user generated questions are available. In this paper, we propose using these data along with search engine query logs to create a question generation shared task that aims to automatically generate questions given a query. At least two benefits could result from such a task. First, it can be used to rank possible candidate questions; therefore, it would enable search of question answer archives. Second, it could be used as a more informative and richer way of query expansion to guide or aid search.

1 Introduction

Generating questions given a query is not a new task. It was used in AskJeeves.com, but had been abandoned in favor of unified search results. Questions at AskJeeves.com were created manually by human editors. This is a laborious process and is difficult to scale up. Community-based question and answering (cQA) services such as Yahoo! Answers, Naver Ji-Sik-In and Baidu Zhidao have become popular ways to elicit user generated content. At of the end of July, 2008, Yahoo! Answers has 40 million English questions and Naver Ji-Sik-In has 70 million Korean questions. Researchers have started asking interesting questions against such collections. For examples, Jeon et al. (2005) proposed methods to search them; Cao et al. (2008) showed how to recommend related questions to help users explore them; Jurczyk and Agichtein (2007) tried to identify experts in a cQA community; Xue et al. (2008) designed new retrieval models for them; Song et al. (2008) examined the utility of a question. The time is ripe for the natural language generation community to also take advantage of the large online cQA archives and start asking interesting questions. Among them, how to automatically generate questions given queries would be a good shared task.

2 Question Generations from Queries as a Shared Task

Mapping queries to existing questions has been tried before with success, e.g. AskJeeves.com. However, the prior work at AskJeeves.com could not scale up due to heavy reliance on human editors. We observe similar efforts in Yahoo! Search Assist and Google Suggest. In this paper, we propose to automate the question generation process given queries.

Our hypothesis is *by automatically generating questions that reflect users' information seeking goals given their queries, we would be able to provide more efficiently and effectively access to information*. If this is proved positive, it would have a huge impact on the information retrieval and web search communities. It would also change the way people interact with information. Therefore, the potential payh off for this task is huge. We summarize the proposed shared task in the following sections.

2.1 Computational Model

Imagining a user issue a query as keywords to a search engine, we might assume that she has a question in mind but it is more convenient and efficient for her to realize her question as one or more queries. Our computational model is to recover the original questions she has in mind given her queries. This process can be expressed as a noisy-channel model as used in statistical machine translation for information retrieval Berger and Lafferty (1999).

2.2 Question Taxonomy

We propose to use the question taxonomy developed by Hovy et al. (2001) that describes a 140 question-type taxonomy[1] created by manually surveying of over 20K online questions. We also look into how to integrate it with more recent results from Liu et al. (2008).

2.3 Question Generation Main Tasks

We propose the main task as generating questions given a query. Queries can be sampled from available search engine query logs to reflect real user needs and ensure coverage of broad topics. An extension to the main task is to generate series of related questions given a user query session.

2.4 Question Generation Sub Tasks

Subtasks to the main task could be: (1) prediction of user goals; (2) learning question generation patterns from cQA archives; (3) question normalization. Just to name a few here.

2.5 Data Creation and Preparation

CQA data and query log access can be negotiated with major search engines and cQA services such as Microsoft Live Search, Ask.com, Yahoo! Answers or Live QnA. Previous TREC QA Track evaluations had experiences in using search engine query logs to create test question sets. Agichtein (Liu et al. 2008) has provided access to data used in his experiments.[2]

2.6 Data Representation and Annotation

Collected data can be stored in a SQL database and dumped into any preferred XML representation formats. We expect to follow previous data annotation schemes in NLG, QA, and summarization communities.

2.7 Evaluations

Evaluations can be addressed in two ways, i.e. *intrinsic* evaluation and *extrinsic* evaluation. In intrinsic evaluation, we estimate how well a method can generate related questions given a query in terms of recall and precision with regarding to question content. Test collection can be created by mapping queries in query logs with question in cQA archives. This would make automatic evaluation possible. We can also measure the quality of generated questions by assessing their grammaticality, conciseness, or other quality metrics that deem important to the NLG community.

For extrinsic evaluation, we can measure how well generated questions help improve users' search efficiency measured in terms of time or their effectiveness by comparing their levels of satisfaction with existing query suggestion services such as Askkids.com, Yahoo! Search Assist, or Google Suggest.

3 Conclusion

In this paper, we propose *automatic generation of questions from queries* as a shared task. With large amount of cQA data available online, together with real world query logs, and interests from both academics and industry, we believe that the time is ripe for such endeavor. The results would change ways that people interact with information and provide new perspectives in natural language generation, information retrieval, and other related fields.

References

Berger, A. and J. Lafferty. Information Retrieval as Statistical Translation. In *Proceedings of SIGIR 1999*.

Cao, Y., C.-Y. Lin, Y. Yu and H.-W. Hon. Recommending Questions Using the MDL-based Tree Cut Model. In *Proceedings of WWW2008*.

Hovy, E., G. Laurie, U. Hermjakob, C.-Y. Lin, and D. Ravichandran. Toward Semantics-Based Answer Pinpointing. In *Proceedings of HLT 2001*.

Jeon, J., W.B. Croft and J. Lee. Finding Similar Questions in Large Question and Answer Archives. In *Proceedings of CIKM 2005*.

Jurczyk, P. and E. Agichtein. Hits on Question Answer Portals: Exploration of Link Analysis for Author Ranking. In *Proceedings of SIGIR 2007*.

Liu, Y., S. Li, Y. Cao, C.-Y. Lin, D. Han and Y. Yu. Understanding and Summarizing Answers in Community-Based Question Answering Services*. In Proceedings of COLING 2008*.

Song, Y.-I., C.-Y. Lin, Y. Cao and H.-J. Rim. Question Utility: A Novel Static Ranking of Question Search. In *Proceedings of AAAI 2008*.

Xue, X., J. Jeon and W.B. Croft. Retrieval Models for Question and Answer Archives. In *Proceedings of SIGIR 2008*.

[1] http://www.isi.edu/natural-language/projects/webclopedia/Taxonomy-data/QuestionAnalysis.html.

[2] http://ir.mathcs.emory.edu/shared/.