Automated Detection of Subject Area for Question Triage in Digital Reference

Keisuke Inoue
School of Information Studies
Syracuse University, Syracuse NY

Abstract

This poster presents an on-going study, which attempts to implement an automated detection of the subject area of digital reference questions for the purpose of question triage. The study attempts to show that automated question triage is an achievable task, by incorporating the technological developments in Information Retrieval and conceptual developments in Library Science. The preliminary experiment using a machine learning classifier produced promising results.

1 Introduction

There has been a rapid increase in the use and technological advancement of digital communication media worldwide in recent years. Among the implications of this phenomenon to digital reference services are the corresponding growth in the use of the services and diversification of users and questions, which challenge the maintenance of the quality of digital reference services. Moreover, digital reference services are facing another challenge, posed by a new generation, who grew up in the digital communication environment. According to Radford and Connaway (2007b), people in this new generation "want not just speedy answers, but full gratification of their information requests on the spot" (p.5). Automated question triage in digital reference was first proposed by Pomerantz and Lankes (2003), in order to overcome such challenges. Despite the fact that the process of question triage is often incorporated in a fully-automated experimental IR system, to the best knowledge of the author, the technology is yet to be utilized in digital reference services.

2 Current Study

2.1 Automated Detection of Subject Area

According to Radford and Connaway (2007a), "subject search" is the most common type of reference question asked in online chat reference services. This study investigates the automated detection of subject area of reference questions using the initial input from users. The assumption is that identifying the subject area of reference questions will enable a digital reference service to forward questions to the appropriate subject expert (or general reference librarian), which, in turn, will help provide a more efficient service.

In order to identify the subject area of reference questions, this study plans to employ two major tools of information organization in library science: classification schemes, such as the Dewey Decimal Classification or the Library of Congress Classification, and thesauri, such as the ERIC Thesaurus or the Library of Congress Subject Headings. The concept was already suggested by Pomerantz and Lankes (2003), so the contribution of this study will be to investigate how to operationalize the idea on actual data and implement it. The author believes that the study will also contribute to the research of question answering systems and information retrieval systems by providing an approach to understand an aspect of questions that are asked through online services. Furthermore, it will help the evaluation of the digital reference services by enabling annotating a large amount of reference questions with question types and subject categories.

2.2 Approach

The current approach of this study is to conduct a two-stage experiment. In each stage, the experiment uses a machine learning (ML) classifier system (Witten and Frank, 2005) that receives manually annotated reference questions as an input and learns to automatically classify the questions. In the first stage, the system classifies the questions into the types of questions. In the second stage, the system classifies the questions into the subject areas. As some question types (e.g. a "procedural/policy" question, such as *"Are you able to renew my library card?"*) do not belong to a particular subject category by its nature, the second stage experiment should use questions of a certain types (e.g. "subject search", such as "I need info on Attention Deficit Disorder (ADD)") As a baseline, simple word counts are used as the attributes for the ML. The study attempts to improve the performance of the baseline system by utilizing additional attributes created by using a library classification scheme and thesaurus (Figure 1). The current implementation of the module that generates the additional attributes (+LCCO) utilizes Library of Congress Classification (LCC) Outline. It counts the number of occurrences of terms that are listed under each "class", the top category of the classification scheme, in the contents of reference questions and use the number as an attribute. Thus it generates as many attributes as the classes in the scheme. While the full LCC is most likely too large for the purpose of the experiment, the outline version include around 20000 words, which is probably still too large, but manageable.

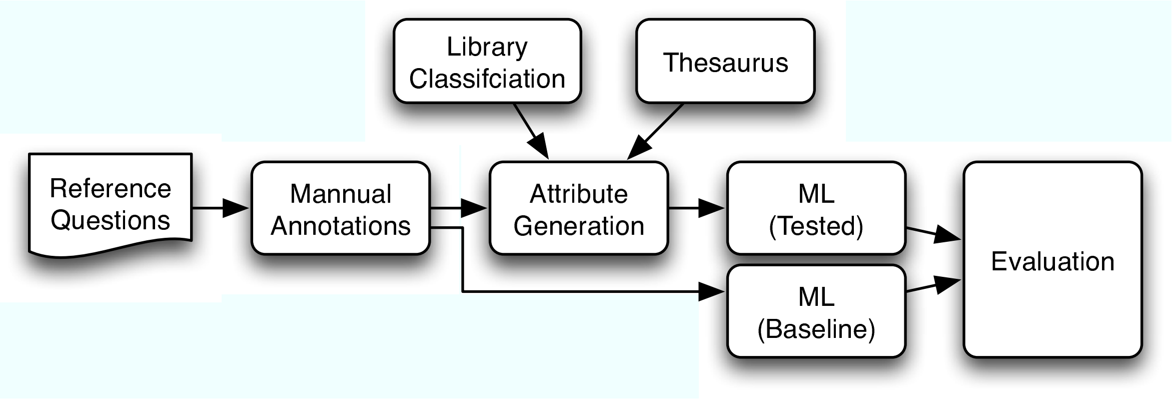

Figure 1: Experiment in each stage

2.3 Data

The data of this study is a collection of initial user inputs from randomly selected transaction logs of the OCLC QuestionPoint service. The data originally contained 500 online chat reference sessions, dated from December 2005 to December 2006[1]. Out of the 500 sessions, 60 sessions were excluded from the analysis because the sessions were for testing or training purposes or for other reasons, such as the sessions were not initiated by the user. Thus 440 sessions remained, providing 440 first messages from the users.

2.4 Annotation

Each initial input is annotated with a question type and a subject category. (Table 1) The current study followed Radford and Connaway (2007a) for the the question types and the subject category were extracted from the LCC.

Table 1: List of Question Types and Subject Categories

Question Types	Subject Categories
Subject Search	Philosophy. Psychology. Religion
Ready Reference	History
Procedural/Policy	Geography. Anthropology. Recreation
Holding	Social Sciences
Research	Political Science
Directional	Law
Inappropriate	Education
Unknown	Music
	Fine Arts
	Language and Literature
	Science
	Medicine
	Agriculture
	Technology
	Military Science
	Naval Science
	Bibliography. Library Science.

2.5 Preliminary Results

At the time of submission, preliminary results from the first stage experiment are available. 168 questions have been annotated by the author. Three ML algorithms were chosen for comparison: Decision Tree, Naive Bayes, and Support Vector Machine (SVM). As in Table 2, the evaluation of the baseline systems shows 40.12% at the lowest to 46.11% at the highest in detecting the type of question types correctly. This is a promising start, given the small number of data (168) and the relatively high number of types of questions to be classified (10). In the current implementation, using the new attributes worsened the performance of the system. Following are some of the possible reasons:

[1] The data was originally prepared for an on-going research project by Radford and Connaway (2005) and became available to the author by courtesy of Dr. Radford, Dr. Connaway, and the OCLC.

1. The number of words in the LCCO is too high and do not contribute to differentiate reference questions.
2. Some words in the LCCO are not representative of the subject category.
3. The number of baseline attributes is too large.
4. The current data size is too small to learn from the +LCCO attributes.

In order to solve the problems above, the author is planning to work on the followings:

1. Produce more annotated data.
2. Manually edit the term list to eliminate unrelated terms. by using a thesaurus.
3. Create a stoplist for the baseline attributes.
4. Use a parser to recognize phrases in a text and the syntactic structure of the text.
5. Use a thesaurus in order to refine the term lists.

In addition, the +LCCO attributes are more likely to contribute to the second stage of the experiment, which attempts to detect the subject categories. Thus the second-stage experiment may prove the utility of the +LCCO attributes.

Table 2: Preliminary Results

	Classifier	**Correctly Identified**	**Incorrectly Identified**
Baseline	Decision Tree	67 (40.12%)	100 (59.88%)
	Naive Bayes	77 (46.11%)	90 (53.89%)
	SVM	74 (44.31%)	93 (55.69%)
+LCCO	Decision Tree	65 (38.92%)	102 (61.08%)
	Naive Bayes	75 (43.71%)	94 (56.29%)
	SVM	71 (42.51%)	96 (57.49%)

3 Conclusion

The preliminary results indicate a good potential. With the further developments above listed, the author hopes that the study will achieve its goal in the near future.

References

Pomerantz, J. and Lankes, R. D. (2003). Taxonomies for automated question triage in digital reference. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 119–121, Washington, DC, USA. IEEE Computer Society.

Radford, M. L. and Connaway, L. S. (2005). Seeking synchronicity: Evaluating virtual reference services from user, non-user, and librarian perspectives. Proposal for a research project, submitted February 1, 2005, to the National Leadership Grants for Libraries program of the Institute of Museum and Library Services (IMLS). Available online at: http://www.oclc.org/research/projects/synchronicity/proposal.pdf.

Radford, M. L. and Connaway, L. S. (2007a). Not dead yet! ready reference in live chat reference. In *13th RUSA New Reference Research Forum ALA Annual Conference*.

Radford, M. L. and Connaway, L. S. (2007b). "screenagers" and live chat reference: Living up to the promise. *Scan*, 26(1):31–39.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA, USA, 2nd edition edition.