Adapting naturally occurring test suites for evaluation of clinical question answering

Dina Demner-Fushman
Lister Hill National Center for Biomedical Communications,
National Library of Medicine, NIH, Bethesda, MD 20894, USA
ddemner@mail.nih.gov

Abstract

This paper describes the structure of a test suite for evaluation of clinical question answering systems; presents several manually compiled resources found useful for test suite generation; and describes the adaptation of these resources for evaluation of a clinical question answering system.

1 Introduction

The community-wide interest in rapid development in many areas of natural language processing and information retrieval resulted in creation of reusable test collections in large-scale evaluations such as the Text REtrieval Conference (TREC)[1]. Researchers in more specific areas, for which no TREC or other collections are available, have to create or find suitable test collections to evaluate their systems.

For example, Cramer et al. (2006) recruited volunteers and quickly gathered a sizeable corpus of question-answer pairs for evaluation of German open-domain question answering systems. This was achieved through a Web-based tool that allowed marking up "interesting" passages in Wikipedia articles and then asking questions about the content of those passages. This appealing approach can not easily be applied in the domain of clinical question answering because the quality of the questions and answers as well as the answer completeness are paramount. A test suite for evaluation of clinical question answering systems should contain a set of real-life questions asked by clinicians and high-quality answers compiled by experts. The answers should be presented in the form deemed useful by clinicians.

One of the benefits of focusing on a specific domain, such as clinical question answering, is that the user-needs and desirable results are well-studied and their descriptions are readily-available. In the case of clinical question answering, clinicians' desiderata are: to see a "bottom-line advice" first, have on-demand access to the context that was used in generation of the advice, and finally have access to the original sources of information (Ely et al., 2005). A fair number of high-quality manually created collections present answers to clinical questions in this form and could be obtained online. Three partially freely-available sources: Family Practitioner Inquiry Network (FPIN)[2], Parkhurst Exchange Forum (PE)[3], and BMJ Clinical Evidence (BMJ-CE)[4] were used to design and develop the presented test suites and evaluation methods.

Although there seems to be a distinction between `test collections` and `test suites` (Cohen et al., 2004) (the former defined as "pieces of text" and associated with corpora, the latter, as lists of specially constructed sentences, or sentence sequences, or sentence fragments (Balkan et al., 1994)), evaluation of answers to clinical questions crosses this boundary and requires the availability of carefully generated sentence fragments as well as suitable document collections.

[1] http://trec.nist.gov/

[2] http://www.primeanswers.org/primeanswers/

[3] http://www.parkhurstexchange.com/qa/index.php

[4] http://www.clinicalevidence.com/ceweb/conditions/index.jsp

Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pages 21–22,
Columbus, Ohio, USA, June 2008.

2 Test suite structure

The multi-tiered answer model of the FPIN and BMJ-CE resources is adapted in this work. The top tier contains the "bottom-line advice". FPIN provides the key-points of the advice in the form of a short sentence sequence, whereas BMJ-CE provides a list of sentence fragments (see Figure 1). Both sources employ experts in question areas to carefully construct the answers. The second tier elaborates each of the key-points in 2-3 paragraph-long summaries generated by the same experts. The third tier provides references to the original sources used in answer compilation.

Likely to be beneficial:

- Angiotensin converting enzyme inhibitors
- Aspirin
- β Blockers . . .

Trade-off between benefits and harms:

- Nitrates (in the absence of thrombolysis)

Likely to be ineffective or harmful: . . .

Figure 1: The top tier of a multi-tiered answer to the clinical question *How to improve outcomes in acute myocardial infarction?* contains key-points generated by a panel of cardiologists.

3 Using the test suite in an evaluation

The answer presented in Figure 1 can be used to evaluate a system's answer to this question by extracting the reference list from the FPIN or BMJ-CE answer. Similarly, the second-tier summaries can be used to evaluate the context for the key-points generated by a system. The references can be used to evaluate the quality of the original sources retrieved by a system if the documents in both lists are represented using their unique identifiers: DOI or a PubMed[5] identifier. Availability of these test suites provides for the following evaluation forms:

- diagnostic, in which developers could evaluate how a tier is affected by changes in its own module(s) or in the underlying tiers;
- task-oriented, in which the system is evaluated as a whole on its ability to answer clinical questions.

It is conceivable to evaluate a system as a whole by evaluating its performance in each tier and then combining the results. In a task-oriented evaluation, it seems reasonable to evaluate the quality of the first-tier answer and verify the adequacy of the second-tier context.

3.1 Caveats

Even the simplest case of the top-tier evaluation, checking the list of fragments generated by a system against the reference list, ideally should be conducted manually by a person with biomedical background. For example, *Acetylsalicylic acid* in a system's answer needs to be matched to *Aspirin* in the reference list. Automation of this step is possible through mapping of both lists to an ontology, e.g., UMLS[6], but such evaluation will be significantly less accurate and potentially biased (if a system uses the same mapping algorithm to find the answer).

A manual evaluation based on 30 of 54 BMJ-CE question-answer pairs in the presented test suite is described in (Demner-Fushman and Lin, 2006). Another 50 question-answer pairs originated in FPIN and PE.

References

Cramer I., Leidner J.L. and Klakow D. 2006. Building an Evaluation Corpus for German Question Answering by Harvesting Wikipedia. *LREC-2006*, Genoa, Italy.

Cohen K.B., Tanabe L., Kinoshita S., and Hunter L. 2004. A resource for constructing customized test suites for molecular biology entity identification systems. *HLT-NAACL 2004 Workshop: Biolink 2004*, Boston, Massachusetts

Balkan L., Netter K., Arnold D. and Meijer S. 1994. TSNLP. Test Suites for Natural Language Processing. *Language Engineering Convention*, Paris, France.

Ely J.W., Osheroff J.A., Chambliss M.L., Ebell M.H. and Rosenbaum M.E. 2005. Answering Physicians' Clinical Questions: Obstacles and Potential Solutions. *JAMIA*, 12(2):217–224.

Demner-Fushman D. and Lin J. 2006. Answer Extraction, Semantic Clustering, and Extractive Summarization for Clinical Question Answering. *ACL 2006*, Sydney, Australia

[5]http://www.ncbi.nlm.nih.gov/sites/entrez

[6]http://www.nlm.nih.gov/research/umls/