Creation of a corpus for evidence based medicine summarisation

Diego Mollá, María Elena Santiago-Martínez

Department of Computing, Macquarie University, Australia

RESEARCH

Please cite this paper as: Mollá, D, Santiago-Martínez, ME. Creation of a corpus for evidence based medicine summarisation. AMJ 2012, 5, 9, 503-506. http://doi.org/10.21767/AMJ.2012.1375

Corresponding Author:
Diego Mollá
Department of Computing, Macquarie University, Sydney, NSW 2109. Australia Email: diego.molla-aliod@mq.edu.au

Abstract

Background

Automated text summarisers that find the best clinical evidence reported in collections of medical literature are of potential benefit for the practice of Evidence Based Medicine (EBM). Research and development of text summarisers for EBM, however, is impeded by the lack of corpora to train and test such systems.

Aims

To produce a corpus for research in EBM summarisation.

Method

We sourced the "Clinical Inquiries" section of the Journal of Family Practice (JFP) and obtained a sizeable sample of questions and evidence based summaries. We further processed the summaries by combining automated techniques, human annotations, and crowdsourcing techniques to identify the PubMed IDs of the references.

Results

The corpus has 456 questions, 1,396 answer components, 3,036 answer justifications, and 2,908 references.

Conclusion

The corpus is now available for the research community at http://sourceforge.net/projects/ebmsumcorpus.

Key Words

Evidence Based Medicine, corpora, text summarisation, natural language processing.

What this study adds:

1. A corpus with questions and summaries that can be used to assist in the research, development and test of natural language processing for evidence based medicine.
2. A description of how the corpus was built.
3. An indication of the kind of research that has been done with this corpus and what else could be done.

Background

Evidence Based Medicine (EBM) recommends physicians to incorporate published evidence when providing care for their patients[1]. Systematic reviews and specialised journals summarise the major findings on those topics that are of highest interest to the physician. However, when the physician is confronted with a specific condition that is not covered by a review, the physician needs to perform a time-consuming sequence of steps to search through the available literature, appraise the quality of the information found, and decide whether the information is applicable to the patient. Resources such as PubMed, a database of more than 20 million abstracts of medical publications, and specialised search engines, help the physician find the relevant literature; but very little has been done to appraise the research findings and extract the specific information that the physician needs.

Method

To help the physician, we propose the development of query-based multi-document summarisation systems that, given a clinical question, find the relevant documents, appraise their medical quality, and summarise them within the context of the question. The expected output of such systems would be synthesised summaries that highlight the key answers to the clinical question as given by the medical literature. Several summarisation systems have been proposed, such as those reviewed by Afantenos et al.[2] However, there is no corpus available to compare the performance of those systems, and there are no means to tell what is the upper limit of achievement of summarisation systems. In this paper we introduce a corpus that we have developed for this purpose. For further details see our past work[3].

Figure 1: Extract of the corpus, edited and reformatted to enhance readability and as an example of the ideal output of an automatic summariser. The underlined text represents links to the source documents. The text below answer 2 has been deleted.

Question: Which treatments work best for hemorrhoids?
Answer 1: Excision is the most effective treatment for thrombosed external haemorrhoids.
Strength of recommendation: B, retrospective studies.

1. A retrospective study of 231 patients treated conservatively or surgically found that the 48.5% of patients treated surgically had a lower recurrence rate than the conservative group (number needed to treat [NNT]=2 for recurrence at mean follow-up of 7.6 months) and earlier resolution of symptoms (average 3.9 days compared with 24 days for conservative treatment).
   Ref PMID= 15486746, Greenspon J, Williams SB, Young HA ,et al. Thrombosed external hemorrhoids: outcome after conservative or surgical management. Dis Colon Rectum. 2004; 47: 1493-1498.
2. A retrospective analysis of 340 patients who underwent outpatient excision of thrombosed external hemorrhoids under local anesthesia reported a low recurrence rate of 6.5% at a mean follow-up of 17.3 months.
   Ref PMID=12972967, Jongen J, Bach S, Stubinger SH ,et al. Excision of thrombosed external hemorrhoids under local anesthesia: a retrospective evaluation of 340 patients. Dis Colon Rectum. 2003; 46: 1226-1231.
3. A prospective, randomized controlled trial (RCT) of 98 patients treated nonsurgically found improved pain relief with a combination of topical nifedipine 0.3% and lidocaine 1.5% compared with lidocaine alone. The NNT for complete pain relief at 7 days was 3.
   Ref PMID=11289288, Perrotti P, Antropoli C, Molino D ,et al. Conservative treatment of acute thrombosed external hemorrhoids with topical nifedipine. Dis Colon Rectum. 2001; 44: 405-409.

Answer 2 For prolapsed internal haemorrhoids, the best definitive treatment is traditional hemorrhoidectomy.
Strength of recommendation: A, systematic reviews.

1. … *(text deleted)* …

The corpus has been sourced from the Clinical Inquiries section of the JFP. These Clinical Inquiries are short reviews of about two pages each. Each Clinical Inquiry addresses key clinical questions for family practice. We have downloaded a total of 456 publicly available clinical inquiries with the kind permission of the publishers. The corpus is formatted in XML to facilitate its processing by a computer. An extract of the corpus is shown in Figure 1, reformatted to ease readability and to illustrate the ideal output that a summariser should produce.

To produce the corpus we have processed each clinical inquiry and the following information has been extracted:

The clinical inquiries, e.g. ``What is the most effective treatment for tinea pedis athlete's foot?''. This was obtained straight from the title of the clinical inquiry.

The evidence-based answers: The answer to each clinical inquiry is composed of several parts addressing different topics related to the question. Each part was identified automatically by using the formatting conventions of the source text. In particular, we took advantage of the fact that each part was followed by an evidence grade that was easy to identify (see below).

The evidence grades of the answer parts: The evidence grade of each answer part follows the Strength of Recommendation (SOR) taxonomy that is used by JFP. It was extracted from the source text by exploiting the text formatting conventions, in particular by looking at the presence of the keyword "SOR", followed by a letter indicating the strength of the recommendation (A, B, C, or D).

The answer justifications: The main text of each clinical inquiry was inspected manually and fragments of it were allocated to the relevant answer components. This was a major annotation undertaking. The source text was distributed to three annotators (members of the research team), with some overlap to check consistency. During the annotation process several checks were made until a final consensus was reached. The whole annotation process took place between December 2010 and February 2011. During the annotation process the annotators also double-checked the automatically extracted components (clinical inquiry, answer, and evidence grade) and corrected them when necessary.

The references: During the annotation process, the citation text was automatically extracted and then manually allocated to the corresponding answer justifications. For each reference, the PubMed ID was identified by running a crowdsourcing annotation task using Amazon's Mechanical Turk (AMT). The references were grouped in sets of 10 references per group (called "hit" in AMT's framework), and each individual hit was assigned to 5 Turkers from the pool of Turkers provided by AMT. After passing a test where they were asked to simulate the annotation task given references with known IDs, the Turkers could choose what hits to annotate. After the annotation was complete, the following automatic checks were made to detect the quality of the annotations: (i) include references with known IDs and check them against the IDs found by the Turkers; (ii) check any errors reported after searching PubMed with the IDs; (iii) compute the percentage of overlapping text between the reference text and the title of the PubMed article retrieved using the ID; and (iv) check the agreement with the other Turkers. These tests highlighted potentially incorrect IDs returned by the Turkers, which were then reviewed manually and corrected if necessary. A final test after the crowdsourcing task was completed and double-checked as described above revealed 100% correct annotations from a random sample of 100 annotations.

Results

In total, the corpus has 456 questions, 1,396 answer components, 3,036 answer justifications, and 2,908 references. There is an average of 3.06 answer components per question and 2.17 answer justifications per answer component. There are 1.22 references per answer justification, which is more than 2,908/3,036 due to the common presence of shared references across answer justifications in a question. There is an average of 6.57 references per question, which is different from 2,908/456 due to the presence of shared references across distinct answer components in a question and the occasional shared references across different questions. The distribution of evidence-based grades was: 345 for A, 535 for B, 330 for C, 15 for D, and 171 without grade.

Discussion

The most immediate application of this corpus is for single-document summarisation. For example, in related work[3] we used the question and reference abstracts as the source, and the answer justifications as the target summary. Our summarisers scored each sentence of the abstract according to several measures including sentence position, similarity with the question, and section information. We evaluated the results using the ROUGE evaluation tool[4], which returns an automatic measure of the similarity between the result summary and the answer justifications.

An additional use of the corpus is for the development of automatic grading systems that determine the strength of the evidence reported. For example, we have trained[5] a supervised classifier using several combinations of features and discovered that publication type alone (such as meta-analysis, systematic review, randomised controlled trial, etc.) gives an accuracy close to 70%.

Currently we are working towards the more ambitious goal of combining the information of multiple documents to provide summaries that are closer to human summaries.

Conclusion

We have completed a corpus of clinical questions and answers. The corpus aims to help the development and testing of text-processing technology to assist the physician in the practice of evidence based medicine. We envisage the use of the corpus for: (i) single-document summarisation and query-focused multi-document summarisation; (ii) appraisal of the answers; and (iii) clustering of references according to the answer components.

References

1. Sackett DL, Rosenberg WM, Gray J, Haynes RB, Richardson WS. Evidence Based Medicine: What it is and what it isn’t. BMJ. 1996; 312(7023): 71-72.
2. Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical documents: a survey. Artif Intell Med. 2005 Feb; 33(2):157-77.
3. Mollá D, Santiago-Martínez ME. Development of a corpus for evidence based medicine summarisation. Proc ALTA 2011, Canberra, Australia; 2011.
4. Lin C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. ACL Workshop on Tech Summarisation Branches Out.
5. Sarker A, Mollá D, Paris C. Towards automatic grading of evidence. Proc LOUHI 2011, Bled, Slovenia; 2011.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their helpful comments. Thanks to Macquarie University for funding this research project.

PEER REVIEW

Not commissioned. Externally peer reviewed.

CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

FUNDING

This research is funded by an internal Macquarie University Research Development Grant.

ETHICS COMMITTEE APPROVAL

Macquarie University Human Research Ethics Committee, reference 5201000828.