A Corpus for Evidence Based Medicine Summarisation

Diego Molla
Macquarie University
Sydney, Australia
diego.molla-aliod@mq.edu.au

Abstract

In this paper we motivate the need for a corpus for the development and testing of summarisation systems for evidence-based medicine. We describe the corpus which we are currently creating, and show its applicability by evaluating several simple query-based summarisation techniques using a small fragment of the corpus.

1 Introduction

Current clinical guidelines urge medical practitioners to practise Evidence Based Medicine (EBM) when providing care for their patients (Sackett et al., 2000). EBM has been defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al., 1996). To find and appraise the evidence the medical practitioner has access to systematic reviews available through search tools such as the Cochrane Library[1] and UpToDate[2]. However, there is not always a systematic review that addresses the specific topic at hand (Sackett et al., 2000) and then a search on the primary literature becomes necessary.

The amount of documents that exist in the primary literature is overwhelming. The US National Library of Medicine, for example, offers PubMed,[3] a database of medical publications that comprises more than 20 million citations. A search in PubMed often returns thousands of documents. With such amount of text, summarising the information becomes crucial. The tools available to the medical practitioner — see e.g. the Survey by Berkowitz (2002) — typically focus on finding and ranking the relevant papers, often with easy access to the abstracts and type of study, and sometimes with highlight of matching terms. But surprisingly little effort has been placed on summarising the information for easy perusal and appraisal by the user.

In this paper we stress the lack of corpora to help research in evidence-based summarisation of clinical articles (Section 2). We present the characteristics of the corpus we are developing (Section 3), and we show the use of a small fragment of the corpus for the evaluation of simple summarisation techniques (Section 4).

2 Where is the Corpus for Summarisation?

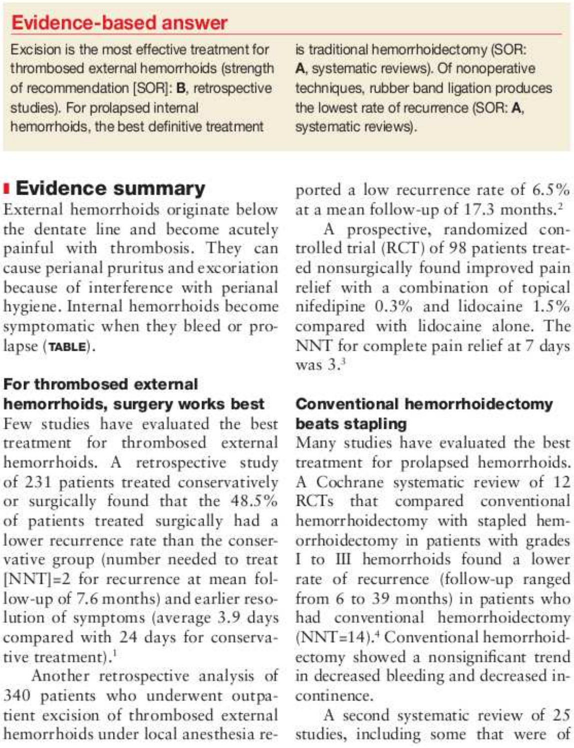
Current summarisation systems have been developed and tested by using corpora built ad-hoc and there is no common corpus readily available specifically for the task. Afantenos et al. (2005) surveys research in summarisation from medical documents. One such summariser is CENTRIFUSER/PERSIVAL (Elhadad et al., 2005), which builds structured query-based representations of the documents as source for the summaries. The system was built using an iterative design that accommodates the feedback of a cohort of users. However, their developers acknowledge the lack of appropriate corpora, and to our knowledge neither CENTRIFUSER nor PERSIVAL were tested on a specific corpus for comparison with other systems.

SemRep (Fiszman et al., 2004) provides abstractive summarisation by producing a semantic representation based on the UMLS concepts and their relations (Bodenreider, 2004) as found in the text. The evaluation was based on human judgement and therefore its results are not readily comparable.

The system by Demner-Fushman and Lin (2006) produces multi-document summaries based on clusters of the main intervention found.

[1] http://www.thecochranelibrary.com/
[2] http://www.uptodateonline.com
[3] http://www.ncbi.nlm.nih.gov/pubmed

Diego Molla. 2010. A Corpus for Evidence Based Medicine Summarisation. In *Proceedings of Australasian Language Technology Association Workshop*, pages 76−80



Evidence-based answer

Excision is the most effective treatment for thrombosed external hemorrhoids (strength of recommendation [SOR]: **B**, retrospective studies). For prolapsed internal hemorrhoids, the best definitive treatment is traditional hemorrhoidectomy (SOR: **A**, systematic reviews). Of nonoperative techniques, rubber band ligation produces the lowest rate of recurrence (SOR: **A**, systematic reviews).

Evidence summary

External hemorrhoids originate below the dentate line and become acutely painful with thrombosis. They can cause perianal pruritus and excoriation because of interference with perianal hygiene. Internal hemorrhoids become symptomatic when they bleed or prolapse (**TABLE**).

For thrombosed external hemorrhoids, surgery works best

Few studies have evaluated the best treatment for thrombosed external hemorrhoids. 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).[1]

Another 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.[2]

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.[3]

Conventional hemorrhoidectomy beats stapling

Many studies have evaluated the best treatment for prolapsed hemorrhoids. A Cochrane systematic review of 12 RCTs that compared conventional hemorrhoidectomy with stapled hemorrhoidectomy in patients with grades I to III hemorrhoids found a lower rate of recurrence (follow-up ranged from 6 to 39 months) in patients who had conventional hemorrhoidectomy (NNT=14).[4] Conventional hemorrhoidectomy showed a nonsignificant trend in decreased bleeding and decreased incontinence.

A second systematic review of 25 studies, including some that were of

Figure 1: Extract of a clinical inquiry from the Journal of Family Practice for the question "Which treatments work best for hemorrhoids?".

The authors present a fine review of possible evaluation methods and they finally settled for a combination of a factoid-based evaluation method, together with the automatic tool for summary evaluation ROUGE (Lin, 2004). The model summaries used for the automatic evaluation were the original paper abstracts. However, by evaluating on a set of abstracts the evaluation was not able to measure the system's ability to perform query-based summarisation, since the abstracts were written prior to any query.

The system by Fiszman et al. (2009) uses factoid-based evaluation that tests the summary ability to find good interventions. This kind of evaluation is not suitable for assessing the summary's ability to indicate the quality of the clinical evidence or other aspects of the summaries that could be important to the medical doctor.

There are collections of clinical questions with their answers that could be used as development and evaluation corpora, such as the Parkhurst Exchange collection,[4] but to our knowledge none of the answers in these collections contain explicit pointers to primary literature. Therefore, as they stand these collections could be used for question-answering tasks but not for query-based summarisation.

3 A Corpus for Summarisation

We are currently developing a corpus of questions and evidence-based information sourced from the Journal of Family Practice (JFP)[5]. We are using all the 496 publicly available documents of the "Clinical Inquiries" section (JFPCI henceforth).[6] Each clinical inquiry from JFPCI contains a clinical question, a short evidence-based answer that includes the strength of recommendation as specified by the Strength of Recommendation Taxonomy (SORT) (Ebell et al., 2004), and a justification of the answer that includes specific references. An extract of a clinical inquiry is shown in Figure 1.

There are two main advantages of using JFPCI rather than direct systematic reviews such as the Cochrane Reviews[7] as a source for our corpus.

1. The format of each inquiry is relatively uniform across all inquiries and therefore it enables a semi-automatic method to convert the data to a corpus that can be used by a machine.

2. The text in each inquiry is much more compact than in a Cochrane review. This results on target text that is closer to what a busy medical practitioner would want to read.

There are other sources of evidence-based text that could be used, such as the project ATTRACT by Public Health Wales (Brassey, 2001).[8] We prefer JFPCI because their procedure to find the answers is more methodical than ATTRACT's and JFPCI includes a short evidence-based answers followed by longer explanations, thus allowing for the use of the corpus for multiple-document and single-document summarisation.

The corpus we are developing is being encoded in XML and each item has the following information (see Figure 2 for a fragment of the encoding of the information from Figure 1):

[4] http://www.parkhurstexchange.com/searchQA

[5] http://jfponline.com/

[6] As of 6 September 2010.

[7] http://www.cochrane.org/cochrane-reviews

[8] http://www.attract.wales.nhs.uk/

<question>**Which treatments work best for hemorrhoids?**</question>
<answer> <snip ID="1">*Excision is the most effective treatment for thrombosed external hemorrhoids. <SOR type="B">retrospective studies</SOR>*

> <long>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 ID="15486746"/ ></long>
>
> <long>Another retrospective analysis of 340 patients ... <ref ID="12972967"/ ></long></snip>

</answer>

Figure 2: Information extracted from a clinical inquiry (formatted to enhance readability)

- A question, which corresponds to the title of the clinical inquiry.
- The answer, which is split into "snips" each one delimited by its evidence level in the original clinical inquiry (e.g. there are 3 answer snips in Figure 2).
- The evidence level of each answer snip (A, B, C) as marked by the source.
- Additional "long" text that justifies the answer by providing a summary of the explicit evidence. This text is manually extracted from the main text body.
- References used in the additional text. Manual lookup in PubMed is being done to locate the PubMed ID.

Not all of the text from the original source is mapped to the XML data (e.g. the sentence "Few studies ..." has been removed in Figure 2), and sometimes minor rephrasing is required to avoid incoherent text.

4 Summarisation Experiments

At the time of writing we had 12 clinical inquiries available for a pilot study. With this fragment we have evaluated several simple query-based single-document summarisation methods. Given a question and an abstract, the summarisers attempt to find those sentences that best satisfy the question information needs. The evaluation system uses ROUGE[9] taking the corresponding <long> element as the model text. For example, in Figure 2, given the abstract with PubMed ID 15486746, the model text is the first <long> element. The 12 clinical inquiries produce a total of 73 text-reference pairs that were used for our evaluation.

We used two baselines:

System	n	*Avg F*	*Confidence*
Last	3	0.183	[0.159–0.206]
Outcomes	3	0.181	[0.158–0.205]

Table 1: Baseline results

1. (*Last*): Return the last n sentences of the abstract for $n = 1, 3, 7$. We obtained the best values for $n = 3$ with no statistically significant difference between $n = 3$ and $n = 7$.[10]
2. (*Outcomes*): Return the sentences extracted by the US National Library of Medicine (NLM)'s outcome extractor (Demner-Fushman et al., 2006). We chose this system because it reports very good results in the task of finding the outcome information and it is the closest that we have found to the aims of our summarisers. The system returns 3 sentences ($n = 3$).

The results of the evaluation of the baselines are summarised in Table 1.

4.1 Finding the most similar sentences

The two baselines introduced in the previous section return summaries that do not incorporate information from the question. We tested the following summarisers that reward sentences with higher similarity with the question:

1. (*Simple*): Return the n sentences that share any words (except stop words)[11] with the question, for $n = 1, 3, 7$. We found the best results for $n = 3$.

[9] We used the default settings of ROUGE.

[10] All tests of statistical significance in this paper are based on the 95% confidence intervals returned by ROUGE.

[11] The stop words used are: '[', ']', 'of', 'a', 'the', 'in', 'to', 'and', 'or', 'should', 'than', 'both', 'for', 'with',' through', 'is', 'as', 'that', '.', ',', ';', ':', '(', ')', 'who', 'are', 'this', 'those', 'at', 'has', 'have', 'had', 'been', 'be', 'it', 'were', 'was'.

System	*n*	*Avg F*	*Confidence*
Simple	3	0.180	[0.157–0.203]
UMLS Concepts	3	0.185	[0.161–0.209]
UMLS Graph	3	0.172	[0.149–0.194]

Table 2: Results with query similarity methods

System	*n*	*Avg F*	*Confidence*
No Overlap	3	0.184	[0.161–0.206]
Word	3	0.178	[0.154–0.199]
UMLS	3	0.185	[0.160–0.209]

Table 3: Results with abstract structure

2. (*UMLS Concepts*): Attempt to account for the existence of synonyms by incorporating the information from UMLS. In particular, return the last n sentences that share any UMLS concepts with the question. UMLS concepts are extracted via NLM's MetaMap (Aronson, 2001).

3. (*UMLS Graph*): Incorporate word relations other than synonymy. We do this by incorporating a word similarity measure that is based on random walks through the graph formed by UMLS relations (Agirre et al., 2009). The summarisers of this group return the n sentences that have the greatest similarity score with the question.

We found the best results for $n = 3$ as reported in Table 2. None of the approaches have statistically significant differences on the value of the average F against each other nor against the baselines.

4.2 Using the structure of the abstracts

Many of the source abstracts contain labelled sections. In the next group of summarisers we have attempted to use such structured abstracts to help the summarisers focus on specific sections of the abstracts. We have mapped each abstract section labels into one of "background", "setting", "design", "results", "conclusion", "evidence" and "appendix".[12] Then we have used this information to build summarisers that extract n sentences using this sequence of steps:

1. Extract the first n sentences of the "conclusion" sections.[12]

2. If we have less than n sentences, fill from the first sentences of the last "results" section. If there are still less than n sentences, fill from the first sentences of the second last "results" section, and so on until we have n sentences or we have exhausted all "results" sections.

3. If we still have less than n sentences, fill form the "design" sections using the same method as with the "results" sections described in step 2.

If the abstract did not have structure, the summariser returns the last n sentences as in Section 4.1. We are also studying methods to automatically structure the unstructured abstracts.

We tried a variation that did not use information from the question (*No Overlap*), another one that selected only sentences with word overlap with the question (*Word*), and another one that selected sentences with UMLS overlap with the question (*UMLS*). The results are shown in Table 3. The results are not statistically different among each other or against the results of the previous section.

5 Summary and Conclusions

We have argued for the creation of a corpus for evidence-based medical summarisation. The corpus is currently under construction, and here we have presented a pilot study of the use of a fragment of the corpus to test simple evidence-based summarisers.

We have seen no statistically different results between the approaches presented. We expect to complete the corpus by end 2010. Then we will repeat the experiments and confirm whether there is no real difference in the results. More importantly, we will release the corpus and test its use with more data-intensive approaches including machine learning methods.

The corpus is designed to facilitate the development of multi-document summarisation techniques and this will be one of the of the main research paths that we plan to follow.

Acknowledgments

Parts of this work have been funded by a travel grant by ORISE and by a Macquarie University Research Development Grant (MQRDG).

[12] Note that an abstract may have several sections that result mapped to the same target label.

References

Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. 2005. Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177, February. PMID: 15811783.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Morristown, NJ, USA. Association for Computational Linguistics.

A. R. Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proc AMIA Symp*, pages 17–21.

Lyle Berkowitz. 2002. Review and evaluation of internet-based clinical reference tools for physicians. Technical report, UpToDate.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270. PMID: 14681409.

J. Brassey. 2001. Just in time information for clinicians: a questionnaire evaluation of the ATTRACT project. *BMJ*, 322(7285):529–530.

Dina Demner-Fushman and Jimmy Lin. 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings ACL*. The Association for Computer Linguistics.

Dina Demner-Fushman, Barbara Few, Susan E Hauser, and George Thoma. 2006. Automatically identifying health outcome information in medline records. *J Am Med Inform Assoc*, 13(1):52–60.

Mark H Ebell, Jay Siwek, Barry D Weiss, Steven H Woolf, Jeffrey Susman, Bernard Ewigman, and Marjorie Bowman. 2004. Strength of recommendation taxonomy (sort): a patient-centered approach to grading evidence in the medical literature. *Am Fam Physician*, 69(3):548–556, Feb.

N Elhadad, M-Y Kan, J L Klavans, and K R McKeown. 2005. Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine*, 33(2):179–198, February. PMID: 15811784.

Marcelo Fiszman, Thomas C. Rindflesch, and Halil Kilicoglu. 2004. Abstraction summarization for managing the biomedical research literature. In *Procs. HLT-NAACL Workshop on Computational Lexical Semantics*, pages 76–83.

Marcelo Fiszman, Dina Demner-Fushman, Halil Kilicoglu, and Thomas C. Rindflesch. 2009. Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of Biomedical Informatics*, 42:801–813.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.

David L. Sackett, William M. Rosenberg, Jamuir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72.

David L. Sackett, Sharon E. Straus, W. Scott Richardson, William Rosenberg, and R. Brian Haynes. 2000. *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, 2 edition.