Australasian Medical Journal [AMJ 2012, 5, 9, 478-481]

Extractive summarisation of medical documents using domain knowledge and corpus statistics

Abeed Sarker[1], Diego Mollá[1], Cecile Paris[2]

1. Centre for Language Technology, Department of Computing, Macquarie University, Sydney, Australia

2. CSIRO --- ICT Centre, Locked Bag 17, North Ryde, Sydney, Australia

RESEARCH

Please cite this paper as: Sarker A, Mollá D, Paris C. Extractive summarisation of medical documents using domain knowledge and corpus statistics. AMJ 2012, 5, 9, 478-481. http://doi.org/10.21767/AMJ.2012.1361

Corresponding Author:
Abeed Sarker
Centre for Language Technology, Department of Computing, Macquarie University, Sydney, NSW 2109, Australia.
Email: abeed.sarker@mq.edu.au

Abstract

Background

Evidence Based Medicine (EBM) practice requires practitioners to extract evidence from published medical research when answering clinical queries. Due to the time-consuming nature of this practice, there is a strong motivation for systems that can automatically summarise medical documents and help practitioners find relevant information.

Aim

The aim of this work is to propose an automatic query-focused, extractive summarisation approach that selects informative sentences from medical documents.

Method

We use a corpus that is specifically designed for summarisation in the EBM domain. We use approximately half the corpus for deriving important statistics associated with the *best* possible extractive summaries. We take into account factors such as sentence position, length, sentence content, and the *type* of the query posed. Using the statistics from the first set, we evaluate our approach on a separate set. Evaluation of the qualities of the generated summaries is performed automatically using ROUGE, which is a popular tool for evaluating automatic summaries.

Results

Our summarisation approach outperforms all baselines (best baseline score: 0.1594; our score 0.1653). Further improvements are achieved when query types are taken into account.

Conclusion

The quality of extractive summarisation in the medical domain can be significantly improved by incorporating domain knowledge and statistics derived from a specialised corpus. Such techniques can therefore be applied for content selection in end-to-end summarisation systems.

Key Words

Automatic summarisation, extractive summarisation evidence based medicine, medical document summarisation

What this study adds:

1. An approach to automatically summarise medical text – a topic on which existing research is quite limited.
2. An investigation of the effect of incorporating domain knowledge and corpus statistics on the quality of extractive summarisation.
3. A possible way of helping EBM practitioners in the future by automatically identifying informative text in medical documents.

Background

Evidence Based Medicine (EBM) practice requires practitioners to extract evidence from published medical research when answering clinical queries. Research has shown that practitioners often fail to follow EBM guidelines during practice, particularly at point-of-care, primarily due to its time-consuming nature. Thus, there is a strong motivation for systems that can automatically summarise information present in medical documents for practitioners and reduce the time required for EBM practice. Despite the strong motivation, research in this area is still very much in its infancy, due to the domain-specific nature of the text.

We propose an extractive, query-focused, single document summarisation system for the medical domain. Our approach utilises domain knowledge and statistical

information derived from a specialised corpus. We further show that the qualities of the extracted summaries can be improved by customising the sentence extraction technique to the type of query posed.

Related work

The earliest works on automatic text summarisation were extractive in nature, where features such as word frequencies, sentence positions, key words and other lexical features were utilised.[1,2,3] Edmundson[4] defined the framework for much of the work on extractive summarisation in what is known as the *Edmundsonian Paradigm.* The author used a linear function to rank sentences for extraction, and we adopt this technique in our approach. More recently, numerous statistical approaches have been proposed that utilise noun phrases, named entities, discourse structures, rhetorical statuses etc.

Summarisation for the medical domain

Research on automatic text summarisation for the medical domain is still very much in its infancy, primarily because of the vast amount of domain knowledge required for this task. Early summarisation systems in this domain were mostly extractive as well and only addressed definitional or factoid questions. Some work in this domain has been carried out under the broader research area of Question Answering (QA). Lin and Demner-Fushman[5] present a summarisation system where text segments classified as the outcomes of the study are presented as the final summary. The BioSquash[6] system performs question-oriented text summarisation of biomedical documents through the use of statistical parsing, named-entity recognition, semantic role labeling and graph generation. The summarisation component of the EpoCare[7] system performs sentence level polarity classification of sentences in medical abstracts, and applies this information for summarisation. None of these systems, however, are fully functional and only have prototypes available.

Method

We use a corpus that is specifically designed for summarisation for EBM[8]. The corpus consists of real-life clinical queries, human generated summaries for each query and abstracts of articles referenced to generate the summaries. We divide the abstracts of the corpus into two sets – one for deriving statistics associated with good quality summaries (*training set*: 1388 abstracts) and another for evaluation of our approach (*test set:* 1319 abstracts). The goal of the task is therefore to use the statistics derived from the first set to select the three most informative sentences from each abstract in the second set. For evaluation of the extracted summaries, we use ROUGE[9], which is a popular tool for evaluating the performance of summarisation systems.

We incorporate domain knowledge into our system by using a sentence classifier tailored for the EBM domain.[10] The classifier classifies each sentence in a medical abstract as one of: Population, Intervention, Background, Outcome, Study and Other (PIBOSO). The classification of sentences into these categories enables us to analyse the type of content that is generally present in medical summaries. We also identify the medical concepts or *semantic types* that are present in the text of our corpus. For this we use the Unified Medical Language System (UMLS) and identify the concepts using the publicly available MetaMap[11] tool. Similar to the PIBOSO information, this information enables us to identify important medical concepts that are generally present in summaries.

We commence our work by generating *ideal* extractive summaries from the abstracts in our training set using the popular summary evaluation tool ROUGE. We do this by generating all three-sentence combinations from each abstract and then calculating the ROUGE-L[1] f-score score of each combination to identify the best three-sentence combination for that abstract. The ROUGE-L score gives a measure of the similarity of an extract with the associated human generated summary in our corpus. Thus, the highest scoring three-sentence combination can be considered as the best extractive summary. We then use these best combinations to derive various statistics based on which our system performs the summarisation task.

We consider the problem of selecting the three sentences for a summary from a source text as three separate problems and derive statistics for each sentence position using the best three sentences in our training set. The score for each source text sentence, therefore, varies across the three target sentences and it can have a different score depending on whether the first, second or third sentence of the summary is being extracted. The statistics are based on factors such as relative sentence position (rps), sentence length (sl), PIBOSO classification of sentence (spib). The following is a brief discussion about each of these factors and how statistics related to each are generated and used.

Relative sentence position: From the best sentence combinations of our training set, we create approximate probability distributions of relative sentence positions for each of the three target sentences. Thus, during summarisation, each sentence is given a score based on the

[1] ROUGE-L is a similarity score based on the Longest Common Subsequence (LCS) between two sequences of text.

probability of its relative position and the target sentence number.

Sentence length: Our analyses show that longer sentences tend to be more informative and therefore are generally more likely to be present in the final summary. Therefore, our summarisation approach rewards longer sentences and penalises shorter ones by assigning positive or negative scores.

PIBOSO: From our training set, we derive the probabilities for each of the six PIBOSO types of sentences of being in the final summary. While existing research suggests that summaries of medical documents consist of *Outcome* sentences, there has not been any concrete analysis of this assumption. We therefore use our training set to obtain probability estimates of each type of sentence. The probability for a specific type of sentence is estimated by dividing the proportion of that type of sentence among the best sentence combinations by its proportion among all the sentences in the training set. The probability distributions for each of the three target sentences show that while it is highly probable for the last target sentence to be an *Outcome* sentence, the two other target sentences tend to include some *Background*, *Population* or generic (*Other*) information. Thus, incorporating this measure enables our summariser to include a number of different topics in our final extracted summaries based on probability, similar to the human generated summaries.

For each sentence, each of these factors contributes a score, which indicates the likeliness of the sentence of being in the final summary based on that factor. These scores are combined using the following *Edmundsonian*[4] equation to generate the final score for a sentence:

$$score = (\alpha \times rps) + (\beta \times sl) + (\gamma \times spib) \qquad (1)$$

To calculate optimal values for the weights α, β and γ we perform an exhaustive search through values from 0 to 1 (with step sizes of 0.2) and choose the values that give the best results over the training set.

Query *type* information in summarisation

To investigate if and how the contents of summaries vary depending on the *types* of queries, we manually identify all *treatment* and *diagnosis* questions in our corpus. We then identify the important semantic types among answers to both these types of questions from the training set summaries. We perform this by generating semantic type frequency distributions of the human generated summaries belonging to treatment and diagnosis questions and comparing them with the semantic type distributions of answers to all other types of questions. A semantic type that has a high frequency for a specific query type relative to all other queries is considered to be an important semantic type for that query type. For each of the two query types mentioned, we compute the top four semantic types. The four top-ranked *treatment* semantic types are: Pharmacologic Substance (phsu), Therapeutic or Preventative Procedure (topp), Organic Chemical (orch) and Disease or Syndrome (dsyn). The four common *diagnosis* semantic types are: Diagnostic Procedure (diap), Disease or Syndrome (dsyn), Laboratory Procedure (lbpr) and Finding (fndg).

To incorporate this information in our summarisation technique, we add another score to equation (1) based on the presence of these semantic types in a sentence. Thus the same sentence can have different scores when the type of query posed is different. We find the optimal weights for this combination of scores in the same way as explained earlier.

Results and Discussion

We compare the ROUGE-L f-scores obtained using our technique against several baselines (one of which includes the summarisation system proposed by Lin and Demner-Fushman[5] that uses sentences classified as *Outcome* for the final summary). Domain independent summarisation techniques such as Naïve Bayes and SumBasic are also used. The first *n* sentences baseline that is invariably used in summarisation for specific domains (e.g., news) is also included. The comparison of scores is shown in Table 1 along with the 95% confidence intervals computed using ROUGE. It can be seen that our system outperforms even the best baseline system.

Table 1: Comparison of ROUGE scores between our system and several baselines along with 95% confidence intervals.

System	ROUGE-L f-score	95% CI
Last 3 sentences	0.1548	0.151-0.158
Last 3 *Outcome* sentences	0.1592	0.155-0.163
First 3 Sentences	0.1399	0.136-0.143
Random	0.1516	0.147-0.154
All *Outcome* sentences	0.1594	0.155-0.164
Naïve Bayes	0.1555	0.152-0.159
SumBasic	0.1582	0.155-0.162
Our System	**0.1653**	0.161-0.169

For question specific summarisation, incorporating the score based on medical semantic types provides additional

improvements over our generic approach. For treatment questions, the ROUGE-L f-score of our summarisation system increases from **0.1619** (95% CI: 0.159 – 0.164) to **0.1644** (95% CI: 0.162 – 0.167) once this new information is incorporated. Similarly, for diagnosis questions, the ROUGE-L f-score increases from **0.1343** (95% CI: 0.132 – 0.136) to **0.1362** (95% CI: 0.134 – 0.137).

The results clearly indicate that incorporation of domain knowledge and statistics carefully derived from a specialised corpus can improve automatic summarisation techniques in this domain. Furthermore, improvements can be obtained by customising the summarisation approach to the type of question. Medical questions can be categorised into various types and the content of the generated summaries can vary depending on the type of the question. In our work, we identify the types of questions manually and only incorporate two types of questions.

Future work will focus on the use automated techniques for identifying question types and customising the information extraction technique for each type of question, taking into account the similarity of each query and the candidate summary. We will also investigate the effect of customising summarisation techniques for different medical publication types based on their differing structure and content. Future work will also focus on analysing more target sentence specific features, such as the distribution of PIBOSO elements and medical semantic types for each of the three target sentences.

Conclusion

EBM practice is time-consuming in nature, as it requires practitioners to search through and extract information from medical research papers. Automatic summarisation techniques can benefit practitioners by extracting relevant information associated with their queries. We show that the use of specialised corpora and domain knowledge can help identify useful information in medical text. The approach can be further improved by customising the extraction technique for different types of questions with differing information needs. Such techniques can therefore be applied for content selection in end-to-end summarisation systems that can present practitioners with bottom-line answers at point of care.

References

1. Luhn HP. The automatic creation of literature abstracts. IBM Journal of Research Development. 1958; 2: 159–65.
2. Baxendale PB. Machine-made index for technical literature – an experiment. IBM Journal of Research Development. 1958; 2(4): 354–61.
3. Earl LL. Experiments in automatic extracting and indexing. Information Storage and Retrieval. 1970; 6: 313–34.
4. Edmundson HP. New methods in automatic extracting. JACM. 1969; 16(2):264–85.
5. Lin J, Demner-Fushman D. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics. 2007; 33(1), 63–103.
6. Shi Z, Melli G, Wang Y, Liu Y, Gu B, Kashani MM et al. Question answering summarisation of multiple biomedical documents. In Proceedings of the 20th Canadian Conference on Artificial Intelligence. 2007.
7. Niu Y, Zhu X, Hirst G. Using outcome polarity in sentence extraction for medical question-answering. In Proceedings of the AMIA annual symposium. 2005; 570–574.
8. Mollá D, Santiago-Martinez ME. Development of a Corpus for Evidence Based Medicine Summarisation. In Proceedings of the Australasian Language Technology Association Workshop. 2011; 86-94.
9. Lin C. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of NAACL-HLT. 2004; 74-81.
10. Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support Evidence Based Medicine. BMC bioinformatics. 2011; 12 (Suppl 2):S5.
11. Aronson AR. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. In Proceedings of AMIA Symposium. 2001; 17–21.

ACKNOWLEDGEMENTS

The authors would like thank the anonymous reviewers for their helpful comments. Thanks to CSIRO and 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 jointly funded by Macquarie University and CSIRO.

ETHICS COMMITTEE APPROVAL

Not Applicable.