Extractive Evidence Based Medicine Summarisation Based on Sentence-Specific Statistics

Abeed Sarker Diego Mollá
Macquarie University
Department of Computing
Sydney NSW 2109, Australia
Abeed.Sarker@mq.edu.au
Diego.Molla-Aliod@mq.edu.au

Cécile Paris
CSIRO ICT Centre
Crn Vimiera and Pembroke Roads
Marsfield NSW 2122, Australia
Cecile.Paris@csiro.au

Abstract

We present an approach for extracting 3-sentence evidence-based summaries relevant to clinical questions. We approach this task as one of query-focused, extractive, single-document summarisation using sentence-specific statistics for each target sentence. We incorporate simple statistics and domain knowledge and show that such an approach is effective for identifying informative sentences from medical abstracts. Our system is evaluated automatically using ROUGE, and we compare our results with several baselines. The ROUGE-L F-scores of our system outperform all baselines. In addition, our approach is computationally efficient, and, on a percentile rank measure, our system achieves a percentile rank of 97.3%.

1. Introduction

Evidence Based Medicine (EBM) urges the physician to incorporate external evidence at point of care [9], yet the huge amount of medical literature is overwhelming. Resources like PubMed,[1] for example, contain more than 21 million abstracts. There is therefore a strong motivation for systems that can automate or support the process of appraisal and evidence extraction. In this paper we report our work towards the automatic extraction of abstract sentences that are most relevant to the expected evidence-based answer to a clinical query. We frame this task as one of extractive summarisation. Unlike traditional extractive approaches, our approach attempts to produce three-sentence summaries by applying different scoring mechanisms to each target sentence. We show that this approach improves summarisation results in the challenging domain of Evidence Based Medicine.

2. Related Work

While there has been a significant amount of work on automatic text summarisation for some specific domains (e.g., news), research is still very much in its infancy for the medical domain. A research work closely related to ours is that by [6]. The summarisation component of their QA system relies on the classification of information present in medical abstracts into PICO (**P**opulation, **I**ntervention, **C**omparison and **O**utcome) elements [8]. Text segments classified as 'Outcome' are presented as the final summary. Taking a different approach, BioSquash [10] performs question-oriented extractive summarisation of biomedical documents through the use of statistical parsing, named-entity recognition, semantic role labelling and graph generation.

3. Method

3.1. Corpus

We used a corpus that specialises in summarisation for EBM [7]. This corpus is collected from the Clinical Inquiries (CI) section of the Journal of Family Practice[2]. For the research work described in this paper, we use the question, the human-produced summaries, and the abstracts available in the corpus [7]. We choose three as the number of sentences to extract because our preliminary analysis of a sample of the human-produced summaries suggests that it is the appropriate size, and related work in this area suggests likewise [6].

Our final data set consists of 456 clinical questions, over 1,100 distinct answers to these questions generated by domain experts from 2,707 medical publications. We divide the records in the corpus into two parts. We then use the

[1] www.ncbi.nlm.nih.gov/pubmed/

[2] www.jfponline.com

abstracts associated with the first part (henceforth referred to as 'training set') for obtaining statistics and those associated with the second part (evaluation set) for evaluation. Our training set contains a total of 1,388 abstracts, while our evaluation set contains a total of 1,319 abstracts.

3.2. Generation of statistics

We commence our work by generating ideal extractive summaries from the abstracts in the training set. We use the F-score of ROUGE-L as the main evaluation criteria [5], and we define the ideal summary of an abstract as the three-sentence combination that gives the highest ROUGE-L F-score when compared with the human generated summary. These best sentences are found by exhaustive search.

Extractive summarisation approaches generally apply the same statistics for selecting all target sentences from the source text. However, our preliminary analysis, carried out on the best sentences (defined above) of the training set, shows that certain statistics can vary significantly across the three target sentences. We, therefore, consider the selection of three sentences from an abstract as three separate problems and use separate statistics for each sentence selection whenever appropriate.

Sentence position related statistics. We compute the relative sentence positions of each of the three best sentences and then generate frequency distributions of the positions.

In particular, we first create histograms of the distributions using 10 bins with bin sizes of 0.1[3]. We then normalise the histograms by dividing each bin by the sum of all bins. Then, for a sentence in a document with relative position p, we assign it a score which is equal to the value of the normalised frequency of the bin for p. Using this approach, we can assign different weights to the same sentence position depending on whether we are attempting to select the first, second or last sentence of the summary.

Sentence length related statistics. Our analysis of the best three-sentence combinations reveals that the lengths of sentences may be good indicators of their importance. This is verified by the fact that the average sentence length for the best sentences is 141.77 characters. The average sentence length for all sentences in the training set is 119.47 characters. During sentence selection, we therefore attempt to reward larger sentences and penalise shorter ones using the following equation:

$$S_{len_i} = \frac{l_s - l_{avg}}{l_d} \tag{1}$$

where l_s is the sentence length in characters, l_{avg} is the average sentence length in characters over the whole training set, and l_d is the length of the document in words. Thus sentences larger than the average length are rewarded by a small amount, while sentences smaller than the average length are penalised. We have experimented with variations of the above equation and observed that this approach works best. This is because, for small documents, long sentences tend to contain a large proportion of important information compared to short sentences, and therefore the rewards/penalties assigned by the equation are also larger in magnitude. Furthermore, using this approach, the magnitude of the score tends to be small for larger documents and sentence selection is primarily influenced by other factors.

Sentence similarity related statistics. Since our intent is to perform *query-focused* extraction of sentences from the abstracts, we attempt to incorporate a technique that rewards sentences similar to the associated queries. At the same time, once a sentence is selected, we try to ensure broad coverage by penalising sentences that are similar to the selected sentence. We perform this through the use of Maximal Marginal Relevance (MMR) [1] and cosine similarity measures.

For each sentence of an abstract and the associated query, we generate vectors as follows. We first lowercase all characters, stem the words and remove stop words. For each remaining stem in a sentence, we then compute the term frequency (tf) in that sentence and the inverse document frequency (idf) over all the sentences in the document. We incorporate domain knowledge into our approach by finding the Unified Medical Language System (UMLS) semantic types of all the terms in each sentence using the MetaMap[4] software package. The semantic types represent broad categories of medical concepts (e.g., disease or syndrome, therapeutic or preventative procedure). The intuition behind the use of semantic types, in addition to words, is that similarity in semantic types between a sentence and the query indicates that the sentence contains the same 'type' of information as the query. For each sentence, we also compute the tf and idf measures for the semantic types. Finally, we generate vectors for each sentence using the $tf \times idf$ values for all pre-processed words and semantic types.

During extraction, the sentence with highest cosine similarity with the query is selected. To score candidate sentences for the following two summary sentences, we use MMR, which is defined as:

$$\begin{aligned} MMR = {} & \lambda(CosSim(S_i, Q)) \\ & - (1-\lambda) max_{S_j \epsilon S}(CosSim(S_i, S_j)) \end{aligned} \tag{2}$$

where $CosSim()$ is the cosine similarity function, S_i is the i-th candidate sentence, S is the set of sentences already

[3] We have experimented with other bin sizes. Using a larger number of bins does not improve performance over the training set as the frequencies tend to get smaller for all bins. Using a smaller number of bins gives comparable performance.

[4] metamap.nlm.nih.gov

selected to be in the summary, and S_j is an already selected sentence. The MMR score is therefore highest for sentences that are similar to the query while at the same time distinct from all other previously selected sentences.

PIBOSO related statistics. We apply the system proposed by [4] to classify all the sentences of the abstracts in our corpus into PIBOSO elements. The PIBOSO elements are a variant of the PICO elements [8] that removes PICO's **C**omparison element and adds **B**ackground, **S**tudy and **O**ther elements.

We first generate five frequency distributions of PIBOSO elements: (1) For all sentences in the training set, (2) for all best sentences of the training set, (3) for all 'first' sentences from the best sentences, (4) for all 'second' sentences from the best sentences, and (5) for all the 'last' sentences from the best sentences. We normalise all the frequency distributions by dividing each bin value by the sum of all the bin values for that distribution.

Taking into account the PIBOSO category of each candidate sentence, we derive two scores from these frequency distributions — one that does not take into account whether the first, second or last sentence is being selected and one that does take this position information into account.

The first score, which we call the Position Independent PIBOSO Score (PIPS) is computed as follows:

$$S_{PIPS_i} = \frac{P_{best}}{P_{all}} \tag{3}$$

where P_{best} is the proportion for that PIBOSO element among the best sentences, and P_{all} is the proportion of that PIBOSO element among all sentences. Thus, this score is higher for sentences belonging to PIBOSO categories that have a higher proportion among the best sentences compared to all sentences; and the larger the difference between the two proportions, the higher the magnitude of this score.

The second score is computed as follows:

$$S_{PDPS_i} = \frac{P_{pos}}{P_{best}} \tag{4}$$

where P_{best} is as before, and P_{pos} is the proportion for that PIBOSO element depending on the sentence number being selected. We call this the Position Dependent PIBOSO Score (PDPS). Thus, when selecting the first sentence, a sentence classified as Background is given a much higher score compared to a sentence classified as Outcome. Similarly, when selecting the last sentence, Outcome sentences receive a much higher score compared to sentences belonging to other categories.

3.3. Combining statistics for sentence extraction

Following [3] we combine all the scores in a single Edmundsonian equation to give the overall sentence score:

$$\begin{aligned} S_{S_i} = \alpha S_{rpos_i} + \beta S_{len_i} + \gamma S_{PIPS_i} \\ + \delta S_{PDPS_i} + \epsilon S_{MMR_i} \end{aligned} \tag{5}$$

where S_{S_i} is the score for a candidate sentence S_i calculated as the weighted sum of the score due to its relative position (S_{rpos_i}), length (S_{len_i}), PIPS (S_{PIPS_i}), PDPS (S_{PDPS_i}), and MMR (S_{MMR_i}). Note that when extracting the first sentence, we replace the MMR score with the cosine similarity score in the equation. In the case of ties, the sentence with greater length is chosen.

Our algorithm is fast, since it does not perform computationally expensive NLP (e.g., parsing). It takes a few seconds to summarise all the documents in our evaluation set on a standard personal computer.

To find good approximations for optimal values of the five weights (α, β, γ, δ and ϵ) and the λ parameter in MMR, we use the training set to perform an exhaustive search through all values from 0.0 to 1.0 using step sizes of 0.1. The obtained values for the weights are: $\alpha = 1.0$, $\beta = 0.8$, $\gamma = 0.1$, $\delta = 0.8$, $\epsilon = 0.1$ and $\lambda = 0.1$.

4. Evaluation

We use a percentile-based approach for evaluating the performance our system versus other approaches using the technique proposed by [2]. The ROUGE-L F-scores of all possible three-sentence combinations of each abstract are binned into 1,000 bins. The resulting histograms are normalised by dividing each bin by the sum of all bins, and the normalised distribution is used as an approximation for the probability density function (*pdf*) for the ROUGE-L F-scores of the sentence combinations for the abstract. The *pdf*s for all abstracts in the evaluation set are then convolved together to generate a *pdf* for the whole set. The resulting distribution is long-tailed, meaning that the F-scores for most of the extracts in the summary space are clustered around the mean. According to the distribution, the minimum F-score that a summarisation system in this domain can have is 0.055 and the maximum is 0.264. However, 95% of the F-scores will lie within a very small range — between the values 0.148 (approximate percentile rank of 2.5%) and 0.167 (approximate percentile rank of 97.5%).

5. Results and discussion

The baselines we use for comparison are as follows:

System	F-Score	95% CI	Percentile (%)
L3	0.159	0.155–0.163	60.3
O3	0.161	0.158–0.165	77.5
R	0.158	0.154–0.161	50.3
O	0.159	0.155–0.164	60.3
PI	0.160	0.157–0.164	69.4
Our	**0.166**	**0.162–0.170**	**97.3**

Table 1. Comparison of ROUGE-L F-scores.

L3 Last three sentences. The last sentences in a medical abstract usually present conclusions, and this has been used as a baseline for summarisation tasks in this domain before [6].

O3 Last three PIBOSO outcome sentences. This is comparable to the summarisation component used by [6]. In our approach, there can be more than three conclusion sentences. Hence, we use the last three[5]. If there are less than three outcome sentences, all outcome sentences are chosen along with the last occurring non-outcome sentences.

R Random. Three sentences are randomly selected from each abstract.

O All Outcomes. All PIBOSO outcome sentences are chosen irrespective of the number of sentences.

PI Sentence position independent. Using the same approach as our system but applying the same statistics for all target sentences. The relative position is used as S_{rpos_i} (i.e., higher score for later sentences) and there is no PDPS.

Table 1 presents the results. Our system outperforms all systems with a percentile rank of 97.3%, and the difference is statistically significant as evidenced by the confidence intervals (CI).

6. Summary and future work

In this paper, we have presented an approach for extractive summarisation for Evidence Based Medicine (EBM). We formulate the task as one of query-focused, automatic extractive summarisation using sentence-specific statistics based on best-sentence extracts. We show that, using separate scoring measures for each target sentence, the performance of the system improves. We derive all statistics from a corpus that specialises in summarisation for EBM and evaluate our approach automatically using ROUGE-L F-scores that are generated by comparing extracted sentences against summaries generated by domain experts. We compare the ROUGE-L F-scores of our system with baseline approaches using a percentile rank-based approach. The best ROUGE-L F-score obtained by our system has a percentile rank of 97.3% and is a statistically significant improvement over the best performing baseline.

Our summarisation approach is extractive and we do not take into account factors such as summary coherence. Instead, our focus is to select informative sentences that can be used to generate bottom-line answers. The primary goal of future work will therefore be to combine informative sentences selected in this manner to perform multi-document summarisation, taking into account factors such as coherence and redundancy, and the query type and publication type.

[5] For purely empirical reasons: we have compared this baseline against one that randomly chooses among outcome sentences. There is no significant difference in scores.

References

[1] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR'98*, pages 335–336, 1998.

[2] H. Ceylan, R. Mihalcea, U. Özertem, E. Lloret, and M. Palomar. Quantifying the limits and success of extractive summarization systems across domains. In *Proceedings of NAACL 2010*, pages 903–911, Los Angeles, California, June 2010. Association for Computational Linguistics.

[3] H. P. Edmundson. New methods in automatic extracting. *J. ACM*, 16(2):264–285, 1969.

[4] S. N. N. Kim, D. Martinez, L. Cavedon, and L. Yencken. Automatic classification of sentences to support Evidence Based Medicine. *BMC bioinformatics*, 12 Suppl 2, 2011.

[5] C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the 2003 Conference of the North American Chapter of the Association fo Computational Linguistics on Human Language Technology*, 2004.

[6] J. J. Lin and D. Demner-Fushman. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

[7] D. Mollá-Aliod and M. E. Santiago-Martinez. Development of a Corpus for Evidence Based Medicine Summarisation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, December 2011.

[8] S. W. Richardson, M. C. Wilson, J. Nishikawa, and R. S. Hayward. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12–A13, 1995.

[9] D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.

[10] Z. Shi, G. Melli, Y. Wang, Y. Liu, B. Gu, M. M. Kashani, A. Sarkar, and F. Popowich. Question answering summarization of multiple biomedical documents. In *Proceedings of the 20th Canadian Conference on Aritificial Intelligence (CanAI '07)*, 2007.