Towards Two-step Multi-document Summarisation for Evidence Based Medicine: A Quantitative Analysis

Abeed Sarker , Diego Mollá-Aliod
Centre for Language Technology
Macquarie University
Sydney, NSW 2109
{abeed.sarker, diego.molla-aliod}@mq.edu.au

Cécile Paris
CSIRO – ICT Centre
Sydney, NSW 2122
cecile.paris@csiro.au

Abstract

We perform a quantitative analysis of data in a corpus that specialises on summarisation for Evidence Based Medicine (EBM). The intent of the analysis is to discover possible directions for performing automatic evidence-based summarisation. Our analysis attempts to ascertain the extent to which good, evidence-based, multi-document summaries can be obtained from individual single-document summaries of the source texts. We define a set of scores, which we call *coverage scores*, to estimate the degree of information overlap between the multi-document summaries and source texts of various granularities. Based on our analysis, using several variants of the *coverage scores*, and the results of a simple task oriented evaluation, we argue that approaches for the automatic generation of evidence-based, bottom-line, multi-document summaries may benefit by utilising a two-step approach: in the first step, content-rich, single-document, query-focused summaries are generated; followed by a step to synthesise the information from the individual summaries.

1 Introduction

Automatic summarisation is the process of presenting the important information contained in a source text in a compressed format. Such approaches have important applications in domains where lexical resources are abundant and users face the problem of information overload. One such domain is the medical domain, with the largest online database (PubMed[1]) containing over 21 million published medical articles. Thus, a standard clinical query on this database returns numerous results, which are extremely time-consuming to read and analyse manually. This is a major obstacle to the practice of Evidence Based Medicine (EBM), which requires practitioners to refer to relevant published medical research when attempting to answer clinical queries. Research has shown that practitioners require botom-line evidence-based answers at point of care, but often fail to obtain them because of time constraints (Ely et al., 1999).

1.1 Motivation

Due to the problems associated with the practise of EBM, there is a strong motivation for automatic summarisation/question-answering (QA) systems that can aid practitioners. While automatic text summarisation research in other domains (e.g., news) has made significant advances, research in the medical domain is still at an early stage. This can be attributed to various factors: (i) the process of answer generation for EBM requires practitioners to combine their own expertise with medical evidence, and automatic systems are only capable of summarising content present in the source texts; (ii) the medical domain is very complex with a large number of domain specific terminologies and relationships between the terms that systems need to take into account when performing summarisation; and (iii) while there is an abundance of medical documents available electronically, specialised corpora for performing summarisation research in this domain are scarce.

1.2 Contribution

We study a corpus that specialises on the task of summarisation for EBM and quantitatively analyse the contents of human generated evidence-based summaries. We compare bottom-line evidence-based summaries to source texts and human-generated, query-focused, single-document summaries of the source texts. This enables us to determine if good single-document summaries contain sufficient content, from source texts, to be used for the generation of multi-document, bottom-line summaries. We also study single-document extractive summaries

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

Abeed Sarker, Diego Mollá-Aliod and Cecile Paris. 2012. Towards Two-step Multi-document Summarisation for Evidence Based Medicine: A Quantitative Analysis. In *Proceedings of Australasian Language Technology Association Workshop*, pages 79−87.

generated by various summarisation systems and compare their performance relative to source texts and human generated summaries. In terms of content, our experiments reveal that there is no statistically significant difference between the source texts and the human-generated, single-document summaries, relative to the bottom-line summaries. This suggests that that the generation of bottom-line summaries *may be* considered to be a two step summarisation process in which the first step is single-document summarisation, and the second step involves information synthesis from the summaries, as illustrated in Figure 1. In the figure, d represents a source document, s represents a summary of a source document, and b represents a bottom-line summary generated from multiple single-document summaries.

In addition to this analysis, we attempt to make *estimations* about the extent to which the core contents of the bottom-line summaries come from the source texts. Such an analysis is of paramount importance in this domain because, if only a small proportion of the summaries contain information from the source articles, we can assume that the summaries are almost entirely generated from specialised human knowledge, making it impossible to perform text-to-text summarisation automatically in this domain without intensive use of domain-specific knowledge. We conclude that there is sufficient overlap between the source texts and evidence-based summaries for the process to be automated. Our analysis is purely numerical and is based on various statistics computed from the available corpus.

The rest of the paper is organised as follows: Section 2 presents a brief overview of research work related to ours; Section 3 provides a description of the corpus we study; Section 4 details our analytical techniques; Section 5 presents the results we obtain, along with a discussion; and Section 6 concludes the paper and provides a brief discussion of our planned future work.

2 Related Work

2.1 Evidence Based Medicine

There is a good amount of published work on EBM practice, which is defined by Sackett et al. (1996) as *"the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients"*. The goal of EBM is to improve the quality of patient care in the long run through the identification of practices that work, and the elimination of ineffective or harmful ones (Selvaraj et al., 2010). The necessity of searching for, appraising, and synthesising evidence makes EBM practice time-consuming. Research has shown that practitioners generally spend about 2 minutes to search for evidence (Ely et al., 2000). Consequently, practitioners often fail to provide evidence-based answers to clinical queries, particularly at point of care (Ely et al., 1999; Ely et al., 2002). The research findings strongly motivate the need for end-to-end medical text summarisation systems.

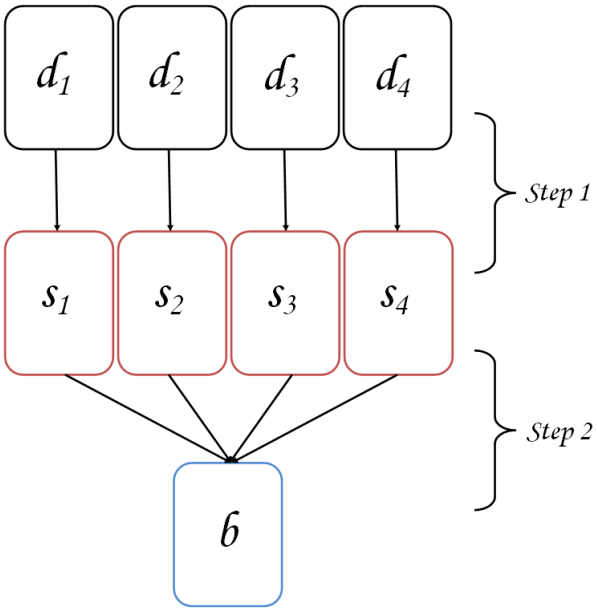

Figure 1: The two-step summarisation process.

2.2 Summarisation for EBM

As already mentioned, the task of automatic text summarisation is particularly challenging for the medical domain because of the vast amount of domain-specific knowledge required (Lin and Demner-Fushman, 2007) and the highly complex domain-specific terminologies and semantic relationships (Athenikos and Han, 2009). Text processing systems in this domain generally use the Unified Medical Language System (UMLS)[2], which is a repository of biomedical vocabularies developed by the US National Library of Medicine. It covers over 1 million biomedical concepts and terms from various vocabularies, semantic categories for the concepts and both hier-

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

archical and non-hierarchical relationships among the concepts (Aronson, 2001). In the UMLS vocabulary, each medical concept is represented using a *Concept Unique Identifier (CUI)*. Related concepts are grouped into generic categories called *semantic types*. Our analysis heavily relies on the CUIs and semantic types of medical terms.

There has been some progress in research for EBM text summarisation (i.e., query-focused summarisation of published medical texts) in recent years. Lin and Demner-Fushman (2007) showed the use of knowledge-based and statistical techniques in summary generation. Their summarisation system relies on the classification of text nuggets into various categories, including *Outcome*, and presents the sentences categorised as *outcomes* as the final summary. Niu et al. (2005, 2006) presented the EPoCare[3] system. The summarisation component of this system performs sentence-level polarity classification to determine if a sentence presents a positive, negative or neutral outcome. Polarised sentences are extracted to be part of the final summary. Shi et al. (2007) presented the BioSquash system that performs query-focused, extractive summarisation through the generation of text graphs and the identification of important groups of concepts from the graphs to be included in the final summaries. More recently, Cao et al. (2011) proposed the AskHermes[4] system that performs multi-document summarisation via key-word identification and clustering of information. The generated summaries are extracted, paragraph-like text segments. Sarker et al. (2012) showed the use of a specialised corpus to perform evidence-based summarisation. In their recent approach, the authors introduce target-sentence-specific, extractive, single-document summarisation, and use various statistics derived from the corpus to rank sentences for extraction. All these systems, however, have limitations. Inspired by this fact, our analyses attempt to test if automatic summairsation is in fact possible for EBM. We also attempt to identify possible summarisation approaches that are likely to be effective in this domain.

[3]http://www.cs.toronto.edu/km/epocare/index.html
[4]http://www.askhermes.org/

2.3 Evaluation and Analysis of Summarisation Systems

The most important research related to automatic evaluation of summarisation systems is perhaps that by Lin and Hovy (2003) . The authors propose a set of metrics called Recall-Oriented Understudy for Gisting Evaluation (ROUGE) that have become very much the standard for automatic summary evaluation. The intent of the ROUGE measures is to find the similarity between automatically generated summaries and reference summaries and it has been shown that ROUGE scores of summaries have a high correlation with human evaluations. We incorporate some ROUGE statistics in our analysis.

ROUGE has also been used for analysis tasks in automatic text summarisation, such as the analysis of extractive summarisation provided by Ceylan et al. (2011) . The authors use ROUGE to show that the combination of all possible extractive summaries follow a long-tailed gaussian distribution, causing most summarisation systems to achieve scores that are generally close to the mean and making it difficult for systems to achieve very high scores. This analysis of extractive summaries has opened a new direction for relative comparison of summarisation systems and the approach has been used in recent work (Sarker et al., 2012). Another recent analysis work on text summarisation, similar to the one we present here, is that by Louis and Nenkova (2011), who show that human-generated summaries generally contain a large proportion of generic content along with specific content. From the perspective of our research, this means that some of the disagreement between different summarisers, in terms of content, may be attributed to dissimilar generic (stylistic) content that are not contained in the source documents, rather than dissimilar query-specific content.

3 Data

3.1 Corpus

The corpus we study (Mollá-Aliod and Santiago-Martinez, 2011) was created from the Journal of Family Practice[5] (JFP). The 'Clinical Inquiries' section of the JFP contains clinical queries and evidence-based answers from real-life EBM practice, and the corpus was built from the information in this section. The corpus consists of a set

[5]www.jfponline.com

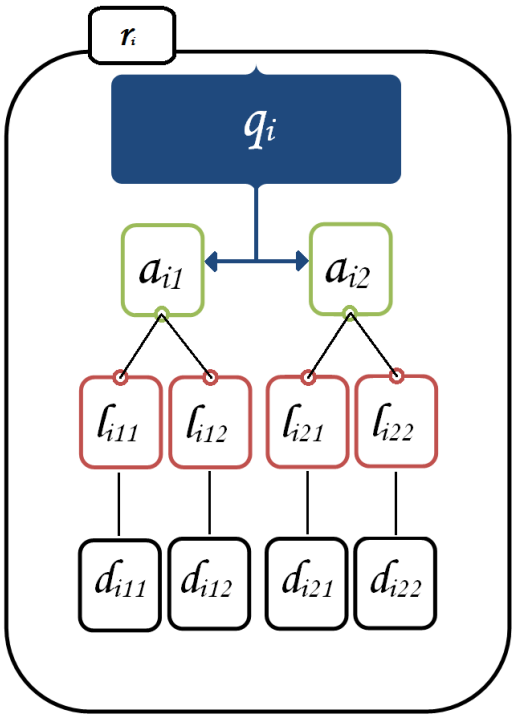

Figure 2: Structure of a sample record the corpus.

of records, $R = \{r_1 \ldots r_m\}$. Each record, r_i, contains one clinical query, q_i, so that we have a set of questions $Q = \{q_1 \ldots q_m\}$. Each r_i has associated with it a set of one or more bottom-line answers to the query, $A_i = \{a_{i1} \ldots a_{in}\}$. For each bottom-line answer of r_i, a_{ij}, there exists a set of human-authored detailed justifications (single-document summaries) $L_{ij} = \{l_{ij1} \ldots l_{ijo}\}$. Each detailed justification in turn l_{ijk} is associated with at least one source document d_{ijk}. Thus, the corpus has a set of source documents, which we denote as $D = \{d_{ij1} \ldots d_{ijo}\}$.

For the work described in this paper, we use the sets A, L and D from the corpus. Figure 2 visually illustrates a sample record from the corpus with two bottom-line summaries associated with the query. We analyse a total of 1,279 bottom-line summaries, associated with 456 queries, along with source texts and human summaries.

4 Methods

4.1 Coverage Analysis

Our first analytical experiments attempt to estimate the extent to which information in the set of bottom-line summaries, A, are contained in the source documents, D_a, associated with each summary (a). This gives us a measure of the extent to which extra information are added to the final summaries by the authors of the JFP articles from which the corpus has been built. For this, we define a set of scores, which we call *coverage scores*. The greater the score, the better is the bottom-line summary coverage. Consider a bottom-line summary a, which contains a set of m terms, and the associated source documents, D_a. The first variant of the coverage scores that we use is a term-based measure and is given by the following equation:

$$Coverage(a, D_a) = \frac{|a \cap D_a|}{m} \quad (1)$$

where $|a \cap D_a|$ represents the number of terms common to a summary and the associated source texts. We first preprocess the text by removing stop words and punctuations, lowercasing all terms and stemming the terms using the Porter stemmer (Porter, 1980). Term tokenisation is performed using the *word tokeniser* of the nltk[6] toolbox. Such a term-level coverage measurement, however, often fails to identify matches in the case of medical concepts that may be represented by multiple distinct terms. An example of this is the term *high blood pressure*. In our corpus, this term has various other representations including *hypertension* and *hbp*.

4.1.1 Incorporation of CUIs and Semantic Types

To address the problem of distinct lexical representations of the same concepts, we identify the semantic types and CUIs of all the terms in the corpus and incorporate this information in our coverage computation. Using CUIs in the computation reduces the dependence on direct string matching because distinct terms representing the same medical concept have the same CUI. For example, all the different variants of the term *high blood pressure* have the same CUI (C0020538). However, it is also possible for terms with different CUIs to have the same underlying meaning in our corpus. For example, the terms [African] women (CUI: C0043210) and African Americans (CUI:C008575) have different CUIs but have been used to represent the same population group. The two terms have the same UMLS semantic type: popg (population group) and this information may be used to match the two terms in our experiments.

[6] nltk.org

We use the MetaMap[7] tool to automatically identify the CUIs and semantic types for all the text in our corpus.

We introduce two more variants of the coverage scores. In our first variation, we use individual terms and CUIs; and in the second variation we use terms, CUIs and semantic types. We apply a sequence of functions that, given a and D_a, along with the CUIs and semantic types of the terms in a and D_a, compute $a \cap D_a$ utilising all the available information (i.e., terms, CUIs, semantic types). Term-based matching is first performed and the terms in a that are exactly matched by terms in D_a are collected. Next, for the unmatched terms in a, CUI matching is performed with the CUIs of D_a. This ensures that different lexical versions of the same concept are detected correctly. All the matched terms are added to the covered terms collection. In our first variant, this value is used for $|a \cap D_a|$ in equation 1. For the second variant, for terms that are still uncovered after CUI matching, semantic type matching is performed and the terms in a with matching semantic types are added to the covered terms collection before computing the coverage score.

A problem with the use of semantic types in coverage score computation is that they are too generic and often produce incorrect matches. For example, the terms *pneumonia* and *heart failure* are two completely distinct concepts but have the same semantic type (*dsyn*). The use of semantic types, therefore, leads to incorrect matches, resulting in high coverage scores. We still use semantic types along with terms and CUIs in our experiments because their coverage scores give an idea of the coverage upper limits.

4.1.2 Concept Coverage

In an attempt to reduce the number of non-medical terms in our coverage score computation, we introduce a fourth varaint to our coverage scores which we call *Concept Coverage* (CC). We noticed that often non-medical terms such as entities, numbers etc. are the primary causes of mismatch among different terms. This coverage score only takes into account the concepts (CUIs) in a and D_a. Referring to equation 1, m in this case represents the number of unique CUIs in a, while $|a \cap D_a|$ is computed as a combination of direct CUI matches and similarity measures among unmatched CUIs. That is, besides considering direct matches between CUIs, we also consider *similarities* among concepts when performing this calculation. This is important because often bottom-line summaries contain generic terms representing the more specific concepts in the source texts (e.g., the generic term *anti-depressant* in the bottom-line summary to represent *paroxetine*, *amitriptyline* and so on). The concept similarity between two concepts gives a measure of their *semantic relatedness* or how *close* two concepts are within a specific domain or ontology (Budanitsky and Hirst, 2006).

In our concept coverage computation, each CUI in a receives a score of 1.0 if it has an exact match with the CUIs in D_a. For each unmatched CUI in a, its concept similarity value with each unmatched concept in D_a is computed and the *maximum similarity* value is chosen as the score for that concept. To compute the similarity between two concepts, we use the similarity measure proposed by Jiang and Conrath (1997). The authors apply a corpus-based method that works in conjunction with lexical taxonomies to calculate semantic similarities between terms, and the approach has been shown to agree well with human judgements. We use the implementation provided by McInnes et al. (2009), and scale the scores so that they fall within the range [0.0,1.0), with 0.0 indicating no match and 0.99 representing near perfect match (theoretically). The direct match score or *maximum similarity* score of each CUI in a are added and divided by m to give the final concept coverage score.

4.1.3 Comparison of Coverage Scores

Our intent is to determine the extent to which the contents of the bottom-line summaries in the corpus are contained in source texts of different granularities. This gives us an estimate of the information loss that occurs when source text is compressed by various compression factors. More specifically, in our experiments, a (in equation 1) is always the bottom-line summary, while for D_a, we use:

i all the text from all the article abstracts associated with a (FullAbs),

ii all the text from all the human-generated, single-document summaries (from L) (HS),

iii all the text from all the *ideal* three-sentence

[7] http://metamap.nlm.nih.gov/

extractive summaries associated with a (IdealSum),

iv all the text from all the single document, three-sentence extractive summaries, produced by a state of the art summarisation system (Sarker et al., 2012), associated with a, and

v all the text from random three sentence extractive single document summaries associated with a (Random).

The IdealSum summaries are three-sentence, single-document, extractive summaries that have the highest ROUGE-L *f-scores* (Lin and Hovy, 2003; Lin, 2004) when compared with the human generated single document summaries (l)[8]. Using these five different sets enables us to estimate the degradation, if any, in coverage scores as the source text is compressed. Table 1 presents the coverage scores for these five data sets along with the concept coverage scores for (i) and (ii)[9].

For each data set, we also compute their ROUGE-L recall scores (after stemming and stop word removal) with the bottom-line summaries, and compare these scores. This enables us to compare the voerage of these data sets using another metric. Table 2 shows the recall scores along with the 95% confidence intervals.

4.2 Task Oriented Evaluation

To establish some estimates about the performances of these variants of the source texts, we performed simple task oriented evaluations. The evaluations required annotation of the data, which is extremely time-consuming. Therefore, we used a subset of the corpus for this task. We manually identified 33 questions from the corpus that focus on 'drug treatments for diseases/syndromes'. All the questions are of the generic form: *'What is the best drug treatment for disease X?'*. Given a question, the task for the system is to identify drug candidates for the disease from the source texts.

From the bottom-line summaries for each of the 33 questions, we manually collected the list of all mentioned drug interventions. Using these, we measured a system's performance by computing its recall for the drug interventions. Our intent, in fact, is not to evaluate the performances of different systems. Instead, it is to evaluate the performances of different source texts on the same task. To extract drug candidates from text, the system relies fully on the MetaMap system's annotation of the data set. All terms identified as *drugs* or *chemicals*[10] are extracted by the system and returned as a list. Recall and precision for each type of source text is computed from this list of drug names.

Using this technique we evaluate the performance of the five previously mentioned source texts. The recall for the FullAbs set acts as the upper limit and this evaluation enables us to determine how much information is lost when the source texts are summarised either manually or automatically. The performance of the Random data set indicates the lower limit. The results of this experiment are presented in the next section.

System	T	T & C	T, C & ST	CC
FullAbs	0.596	0.643	0.782	0.659
HS	0.595	0.630	0.737	0.644
IdealSum	0.468	0.511	0.654	..
Sarker et al.	0.502	0.546	0.683	..
Random	0.403	0.451	0.594	..

Table 1: Coverage scores for the five data sets with the bottom-line summaries. T = Terms, C = CUIs, ST = Semantic Types, and CC = Concept Coverage.

System	Recall	95% CI
FullAbs	0.418	0.40 - 0.44
HS	0.405	0.39 - 0.42
IdealSum	0.284	0.27 - 0.30
Sarker et al.	0.318	0.30 - 0.34
Random	0.229	0.21 - 0.24

Table 2: ROUGE-1 recall scoresand 95% confidence intervals for the five data sets with the bottom-line summaries.

5 Results and Discussion

Table 1 shows that, when terms and CUIs are used, the source texts cover approximately 65% of the summary texts, and incorporation of se-

[8]These summaries were produced by generating all three-sentence combinations for each source text, and then computing the ROUGE-L f-score for each combination.

[9]We only compute the concept coverage scores for these two sets because of the extremely long running time of our similarity measurement algorithm.

[10]The semantic types included in these two categories are: aapp, antb, hops, horm, nnon, orch, phsu, strd, vita, bacs, carb, eico, elii, enzy, imft, inch, lipd nsba, opco. A list of the current UMLS semantic types can be found at: www.nlm.nih.gov/research/umls/

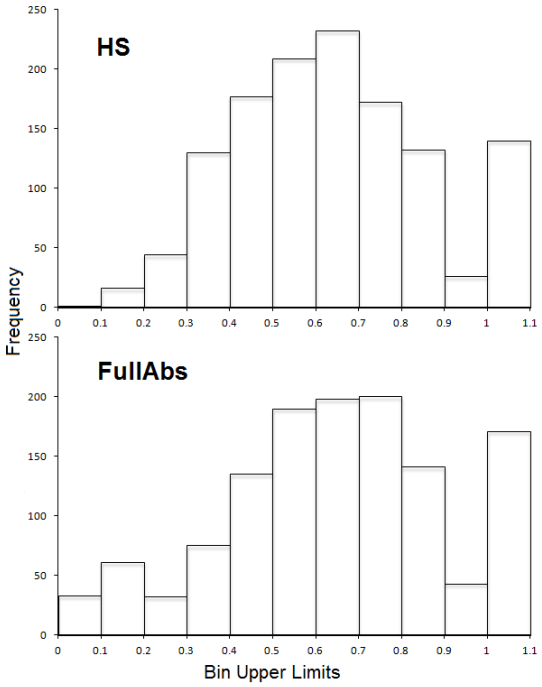

Figure 3: Distributions for concept coverage scores.

mantic types takes the coverage score to close to 80%. The concept coverage scores are similar to the term and CUI overlap scores. Analysis of the *uncovered* components reveal a number of reasons behind coverage mismatches. First of all, as already metioned earlier in this paper, authors often prefer using generalised medical terms in the bottom-line summaries while the source texts contain more specific terms (e.g., *antibiotics vs penicillin*). Incorporating semantic types ensures coverage in such cases, but also leads to false matches. Secondly, MetaMap has a relatively low word sense disambiguation accuracy (Plaza et al., 2011) and often fails to disambiguate terms correctly, causing variants of the same term to have different CUIs, and often different semantic types. Thirdly, a large portion of the uncovered components consists of text that improves the qualitative aspects of the summaries and do not represent important content. Considering the analysis presented by Louis and Nenkova (2011), it is no surprise that the texts of all granularities contain a significant amount of generic information, which may be added or lost during summarisation.

Interestingly, Table 1 reveals that the human generated single-document summaries have almost identical coverage scores to full source articles. Figure 3 shows the distributions of the concept coverage scores for the two sets, and it can be seen that the distributions are very similar. The coverage scores obtained by the two summarisation systems (IdealSum and Sarker et al.) also have high coverage scores compared to the Random summaries.

	T	**T & C**	**CC**
z	-1.5	-1.27	-1.33
p-value (2-tail)	0.13	0.20	0.16

Table 3: z and p-values for Wilcoxon rank sum tests.

Table 2 shows that the ROUGE-L recall scores are also very similar for the HS and FullAbs data sets and lie within each other's 95% confidence intervals, indicating that there is no statistically significant difference between the contents of the HS and FullAbs sets.

To verify if the difference in the coverage scores between the HS and FullAbs sets are statistically significant, we perform statistical significance tests for the two pairs of coverage scores. Due to the paired nature of the data, we perform the Wilcoxon rank sum test with the null hypothesis that the coverage scores for the two sets are the same ($\mu_0 = 0$). Table 3 shows the z and p-values for the tests performed for the term, term and CUI and concept coverage scores for the HS and FullAbs sets. In all cases $p > 0.05$, meaning that we cannot reject the null hypothesis. Therefore, the difference in the two sets of coverage scores are not statistically significant. This adds further evidence to the hypothesis that single document summaries may contain sufficient content for bottom-line summary generation. This, in turn, strengthens our claim that the generation of bottom-line summaries by humans *may be* considered to be a two step process, in which the first step involves summarising individual documents, based on the information needs of queries, and the second step synthesises information from the individual summaries.

The compression factors (CF) in Table 4 show the relative compression rates required for the various source texts to generate the bottom-line summaries. It can be seen that generating bottom-line summaries from original source texts require approximately 5 times more compression compared to the generation from single document summaries, suggesting that the single document

System	Recall (%)	Precision (%)	CF
FullAbs	77	41	0.05
HS	75	68	0.26
IdealSum	66	48	0.20
Sarker et al.	68	45	0.15
Random	52	35	0.21

Table 4: Task oriented evaluation results and summary compression factors (CF) for the five sets of source texts.

summaries contain important information from the source texts in a much compressed manner. Thus, for a summarisation system that focuses on generating bottom-line summaries, it is perhaps better to use single document summaries as input rather than whole source texts, as the information in the source texts are generally very noisy. Considering the balance between coverage scores and compression factors of IdealSum and Sarker et al., such content-rich automatic summaries may prove to be better inputs for the generation of bottom-line summaries than original texts.

Table 4 also presents the drug name recall and precision values for the five source text sets from the task-oriented evaluation. The relative recall-based performances of the different source text sets closely resemble their coverage scores. The performance of the HS summaries is almost identical to the FullAbs system, and the systems IdealSum and Sarker et al. are close behind. Primary reasons for drops in recall are the use of generic terms in bottom-line summaries, as already discussed, and errors made by MetaMap. For the former problem, automatic summarisation systems such as IdealSum and Sarker et al. suffer the most, as the articles in the FullAbs set generally contain the generic terms (e.g., antibiotic) and also the specific terms (e.g., penicillin). However, the compressed versions of the source texts, in the IdealSum and Sarker et al. sets, only the specific terms tend to occur. Importantly, the low precision score for the FullAbs set illustrates the high amount of noise present. The precision scores for the HS set and the two summarisation systems are higher than the FullAbs set, indicating that selective compression of the source text may help to efficiently remove noise.

6 Conclusions and Future Work

We performed analyses on a corpus that is specialised for automatic evidence-based summarisation. Our intent was to analyse the extent to which: (i) information in the bottom-line summaries are directly contained in the source texts; and, (ii) good, evidence-based, multi-document summaries can be obtained from individual single-document summaries of the source texts. We applied various statistics from the corpus to ascertain the difference in content among source texts and summaries of the source texts.

Our analyses show that human summarisers rely significantly on information from published research when generating bottom-line evidence-based summaries. This is demonstrated by the coverage scores presented in the previous section and the manual analysis following it. This indicates that, content-wise, it is possible to generate summaries for EBM automatically in a text-to-text manner. Our experiments also show that human-generated single-document summaries contain approximately the same relevant information as the source texts but in a much more compressed format. This suggests that, for generating bottom-line summaries, it might be a good idea to apply a two-step summarisation. The first step involves single-document, query-focused summarisation. The second step, which is dependent on the output of the first step, performs further summarisation of the already compressed source texts to generate bottom-line answers. For such an approach, it is essential that the first step produces content-rich, high precision summaries. With the advent of new, efficient, single-document summarisation systems in this domain, a multi-step summarisation system has the potential of producing very good results.

Future work will focus on performing more comprehensive task-oriented experiments using these different datasets to evaluate their usefulness in the summarisation task. We will also attempt to develop a two-step summarisation system and compare its performance with other state of the art summarisation systems in this domain.

Acknowledgements

This research is jointly funded by Macquarie University and CSIRO. The authors would like to thank the anonymous reviewers for their valuable comments.

References

Alan R. Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. In *Proceedings of the AMIA Annual Symposium*, pages 17–21.

Sofia J. Athenikos and Hyoil Han. 2009. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, pages 1–24.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

John W. Ely, Jerome A. Osheroff, Mark H. Ebell, George R. Bergus, Barcey T. Levy, Lee M. Chambliss, and Eric R. Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319(7206):358–361.

John W. Ely, Jerome A. Osheroff, Paul Gorman, Mark H. Ebell, Lee M. Chambliss, Eric Pifer, and Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *BMJ*, 321:429–432.

John W. Ely, Jerome A. Osheroff, Mark H. Ebell, Lee M. Chambliss, DC Vinson, James J. Stevermer, and Eric A. Pifer. 2002. Obstacles to answering doctors' questions about patient care with evidence: Qualitative study. *BMJ*, 324(7339):710–716.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, pages pp. 19–33.

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

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of HLT-NAACL 2003*, pages 71–78.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of NAACL-HLT 2004*, pages 74–81.

Annie Louis and Ani Nenkova. 2011. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 34–42.

Bridget T. McInnes, Ted Pedersen, and Serguei V. S. Pakhomov. 2009. UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. In *Proceedings of the AMIA Annual Symposium 2009*, pages 431–435.

Diego Mollá-Aliod and Maria Elena Santiago-Martinez. 2011. Development of a Corpus for Evidence Based Medicine Summarisation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 86–94.

Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. 2005. Analysis of polarity information in medical text. In *Proceedings of the AMIA Annual Symposium*, pages 570–574.

Yun Niu, Xiaodan Zhu, and Graeme Hirst. 2006. Using outcome polarity in sentence extraction for medical question-answering. In *Proceedings of the AMIA Annual Symposium*, pages 599–603.

Laura Plaza, Antonio Jimeno-Yepes, Alberto Diaz, and Alan Aronson. 2011. Studying the correlation between different word sense disambiguation methods and summarization effectiveness in biomedical texts. *BMC Bioinformatics*, 12(1):355–368.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

David L Sackett, William M C Rosenberg, J A Muir 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.

Abeed Sarker, Diego Mollá, and Cecile Paris. 2012. Extractive Evidence Based Medicine Summarisation Based on Sentence-Specific Statistics. In *Proceedings of the 25th IEEE International Symposium on CBMS*, pages 1–4.

Sanchaya Selvaraj, Yeshwant Kumar, Elakiya, Prarthana Saraswathi, Balaji, Nagamani, and SuraPaneni Krishna Mohan. 2010. Evidence-based medicine - a new approach to teach medicine: a basic review for beginners. *Biology and Medicine*, 2(1):1–5.

Zhongmin Shi, Gabor Melli, Yang Wang, Yudong Liu, Baohua Gu, Mehdi M. Kashani, Anoop Sarkar, and Fred Popowich. 2007. Question answering summarization of multiple biomedical documents. In *Proceedings of the 20th Canadian Conference on Aritificial Intelligence (CanAI '07)*.

Yonggang Cao and Feifan Liu and Pippa Simpson and Lamont D. Antieau and Andrew Bennett and James J. Cimino and John W. Ely and Hong Yu. 2011. AskHermes: An Online Question Answering System for Complex Clinical Questions. *Journal of Biomedical Informatics*, 44(2):277 – 288.