Automatic Detection of Answers to Research Questions from Medline Abstracts

Abdulaziz Alamri and Mark Stevenson
Department of Computer Science
The University of Sheffield
Sheffield, UK
adalamri1@sheffield.ac.uk; mark.stevenson@sheffield.ac.uk

Abstract

Given a set of abstracts retrieved from a search engine such as *Pubmed*, we aim to automatically identify the claim zone in each abstract and then select the best sentence(s) from that zone that can serve as an answer to a given query. The system can provide a fast access mechanism to the most informative sentence(s) in abstracts with respect to the given query.

1 Introduction

The large amount of medical literature hinders professionals from analyzing all the relevant knowledge to particular medical questions. Search engines are increasingly used to access such information. However, such systems retrieve documents based on the appearance of the query terms in the text despite the fact that they may describe another problem.

The search engine Pubmed® for example is a well known IR system to access more than 24 million abstracts for the biomedical literature including Medline® (Wheeler et al., 2008). The engine takes a query from user and returns a list of abstracts that can be relevant or partially irrelevant to the query, which requires from the user to go through each abstract for further analysis and evaluation.

Researchers who conduct a systematic review (Gough et al., 2012) tend to use the same approach to collect the studies of interest; however, they are found to spend significant effort identifying the studies that are relevant to the research question. Relevancy is usually measured by scanning the result and conclusion sections to identify authors claim and then comparing the claim with the review question; where a claim can be defined as the summary of the main points presented in a research argument.

Incorporating a middle tier system between the search engine and the user will be useful to minimize the effort required to filter the results. This research presents a system that aids those searching for studies that discuss a particular research question. The system acts as a mediator between the search engine and the user. It interprets the search engine results and returns the most informative sentence(s) from the claim zone of each abstract that are potential answers to the research question. The system reduces the cognitive loads on the user by assisting their identification of relevant claims within abstracts

The system comprises two components. The first component identifies the claim zone in each abstract using the rhetorical moves principle (Teufel and Moens, 2002), and the second component uses the sentences in the claim zone to predict the most informative sentence(s) from each abstract to the given query.

This paper makes three contributions: presenting a new set of features to build a classifier to identify the structure role of sentences in an abstract that is at least shows similar performance to the current systems; building a classifier to detect the best sentence(s) (lexically) that can be an answer to a given query; and introducing a new feature (*Z-score*) for this task.

2 Related Work

We are not aware of any work that has explicitly discussed the detection of claim sentence most related to a predefined question, however, studies have discussed related research.

Ruch et al. (2007) for example used the rhetorical moves approach to identify the conclusion sentences in abstracts. Their system was based on a Bayesian classifier, and normalized n-grams and relative position features. The main objective of that research was to identify sentences that belong to the conclusion sections of abstracts; they re-

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 141–146,
Beijing, China, July 30, 2015.

garded such information as *key* information to determine the research topic. Our research is similar to that work since we use the conclusion section to identify the key information in an abstract with respect to a query, but we also include the result sections.

Hirohata et al. (2008) showed a similar system using CRFs to classify the abstract sentences into four categories: objective, methods, results, and conclusions. That classifier takes into account the neighbouring features in sentence S_n such as the n-grams of the previous sentence S_{n-1} and the next sentence S_{n+1}.

Agarwal et al. (2009) described a system that automatically classifies sentences appear in full biomedical articles into one of four rhetorical categories: introduction, methods, results and discussions. The best system was achieved using Multinominal Naive Bayes. They reported that their system outperformed their baseline system which was a rule-based.

Recently, Yepes et al. (2013) described a system to index Gene Reference Into Function (GeneRIF) sentences that show novel functionality of genes mentioned in Medline. The goal of that work was to choose the most likely sentences to be selected for GeneRIF indexing. The best system was achieved using Naive Bayes classifier and various features including the discourse annotations (the NLM category labels) for the abstracts sentences.

Our research is close to Hirohata et al. (2008) system since we use the same algorithm, but use a different set of features to build the model. Moreover, it similar to Yepes et al.(2013) system since we use the value of the *nlmCategory* attribute rather than the labels provided by the authors to learn the role of sentences.

3 Method

3.1 Claim Zoning Component

This component is based on the hypothesis that the contribution of a research paper tend to be found within the result or conclusion sections of its abstract (Lin et al., 2009). Identifying these sections manually especially in unstructured abstracts is a tedious task. Medical abstracts tend to have logical structure (Orasan, 2001) in which each section represent a different role.

Unfortunately, about 70% of Medline abstracts are unstructured (have no section labels). Structured abstracts use a variety of these labels. The National Library of Medicine (NLM) have reported that 2,779 headings have been used to label abstracts sections in Medline (Ripple et al., 2012).

Relying on the labels provided by the abstracts authors to identify the roles of the sentences could be useful for research purpose; but in practice this means all Medline abstracts need to be reannotated even the structured abstracts to guarantee that they are labelled with the same set of annotations to understand their roles. This is not efficient especially when we consider the huge volume of the Medline repository.

To accommodate that problem, we use the NLM category value assigned to each section in the XML abstract (*nlmCategory* attribute). The NLM assigns five possible values (categories): *Objective, Background, Methods, Results and Conclusions*. This research uses these categories as an alternative way to learn the roles of abstracts sentences. This resolves two problems: first, the roles of sentences in structured abstracts can be automatically learned from the the value of the *nlmCategory* attribute without any further processing, consequently, the roles of sentences in 30% of the Medline abstracts can be accurately identified; second, those labels can be used to build a machine learning classifier to predict the role sentences of the unstructured abstracts in Medline.

The claim zoning component regards identifying the roles of sentences as a sequence labelling problem. This requires an algorithm that takes into account the neighbouring observations rather than only current observation as in other ordinary classifiers e.g. SVM and Naive bayes. Conditional Random Fields (CRF) algorithm have been used successfully for such task (Hirohata et al., 2008; Lin et al., 2009). Therefore, we use the CRF algorithm along with lexical, structural and sequential features to build a classifier model to identify the claim zones in abstracts. The classifier is implemented using the CRFsuite library (Okazaki, 2007) using L-BFGS method. Note that we modify the NLM five categories to become four where the *Background* and *Objective* categories are merged into a new category called *Introduction*. That is because the background and objectives sections in Medline tend to overlap with each other (Lin et al., 2009). Moreover, these sections usually appear sequentially and merging them together is sensible to avoid the overlapping problem. Therefore, this component identifies the

sentences roles in abstracts by labelling them with one of the four possible categories: *Introduction, Methods, Results and Conclusions*.

3.1.1 Data

The claim zoning component is built using a dataset consisting of 10,000 structured abstracts collected from Medline using the query "*cardiovascular disease*".

3.1.2 Features

The claim zoning component employs various features:

N-grams: N-grams are lexical features that have been reported as useful to capture the general context of text (Turney, 2002; Yu and Hatzivassiloglou, 2003). For every sentence, uni-grams and bi-grams are extracted from the abstract's title, the current sentence S_n, the previous sentence S_{n-1}, and the next sentence S_{n+1}.

Sentence-Title similarity (*st-sim*): This feature is the cosine similarity score $sim(s, T)$ between each sentence in an abstract and its title. This feature has been previously found useful for summarization tasks (Teufel and Moens, 2002). Achieving an accurate similarity score between the sentences and the title in an abstract is not a straightforward task. Many abstracts in the medical domain use multiple forms (i.e abbreviation and its expansions) to describe the same medical concept e.g. ACE and angiotensin-converting enzyme.

Such variation may cause inaccurate scores particularly when computing the similarity between an abbreviation and its expansion. Fortunately, the pattern of using abbreviations and their expansions in medical research can be predicted using an algorithm developed by Schwartz and Hearst (2003). We automatically replace all long-forms concepts with their abbreviations to unified their appearance within an abstract. Similarity scores are binned into 11 values starting from 0 to 10.

Relative Sentence location: The relative location of a sentence is important to identify its role within the abstract. The introduction sentences for instance tend to occur at the beginning of an abstract and the conclusion sentences occur at the end. Rather than using the original position of the sentence, we adjusted the all sentences positions to have the same scale from 1 to 10.

Tense feature: The tense of verbs used in sentences often correlates with its rhetorical moves (Teufel and Moens, 2002). For example, some authors use the present perfect tense in the introduction section and past simple in the conclusion section. For each sentence in an abstract, the main verb tense (ROOT-0, verb) is extracted using the dependency tree generated from the Stanford parser (de Marneffe and Manning, 2008).

3.2 Answers Detection Component

This component uses the sentences that belong to the result or conclusion sections of abstracts (claim zone) to identify the most informative sentence(s) to a given query. It relies on three assumptions, two from the literature (Lin et al., 2008; Ruch et al., 2007; Lin et al., 2009; Otani and Tomiura, 2014) and the last one that is conventional: the first assumption is that any sentence in abstract that shares many words with the title tends to express important information about the topic. The second is that any sentence that applies the first assumption within certain threshold and exist in the result and conclusion sections is considered as a key sentence concerning the research topic. The third assumption is that any sentence that applies the previous two assumptions and has a high lexical similarity score with the query is considered an informative sentence with respect to the query.

The component classifier is built using a decision tree algorithm (Quinlan, 1993). The decision tree algorithm builds a tree-like model that can be converted into rules which can be easily interpreted and analysed by human. We use the open source implementation of decision tree (J48) in Weka (Hall et al., 2009) to build the model.

3.2.1 Data

This component uses two subsets (corpus-2 and corpus-3) of a corpus that was originally developed to recognize contradictory claims in medical abstracts. That corpus consists of abstracts that were collected from the studies used in systematic reviews that discuss various problems about cardiovascular diseases. Note that each systematic review attempts to answer one question. Two independent annotators were asked to identify the best claim sentence from each abstract that answers the review question e.g. (1). In this research the most informative information with respect to the research question is considered to be the claim.

1. In patients with dilated cardiomyopathy, are HLA genes associated with development of Dilated Cardiomyopathy? **[Question]**

2. In the IDC group, the frequency of human leukocyte antigen DR4 was similar to that reported in the normal population. **[PMID#9220309][ANSWER]**

The classifier of answer detection is trained and evaluated using corpus-2 (structured abstracts). That corpus consists of 183 sentences annotated as *answers* and 987 sentences annotated as *non-answers* to 24 review questions. Note that it is possible for more than one sentence to answer a review question, however, only the most informative sentence was annotated as answer.

Corpus-3 (unstructured abstracts) consists of 69 abstracts (69 *answer* sentences and 357 *non-answer* sentences) which answer 15 review questions. It is used to evaluate the system resulted from the integration of the claim zoning component and the answer detection component.

3.2.2 Features

This component uses four features which are extracted from the result and conclusions sentences:

Sentence Structure Role (*role-label*): This feature comes from annotating the abstract sentences using the claim zoning component if the abstract is unstructured, otherwise the value of *nlmCategory* is extracted and used as a feature.

Sentence-Title Similarity (*st-sim*): This feature is similar to *st-sim* feature used in the claim zoning component. The scores are normalized to a scale of 0 to 50 since this was shown to improve performance.

Sentence-Query Similarity (*sq-sim*): This feature captures the relationship between the research question and sentences in the abstract. Those with a high lexical similarity to the question are more likely to be answers to it than others. Similar to *st-sim* feature, the cosine similarity score between sentences and their related questions are computed and the scores are normalized to a scale of 0 to 50.

Z-score Value: This feature is used to exploit assumption (2) described in section 3.2. This feature identifies the position of the similarity score of a sentence with respect the distribution of the similarity scores of the other sentences within an abstract. It assumes that the similarities of the sentences in the result and conclusion sections are normally distributed. The goal of using this feature is to enable the classifier to learn a similarity threshold score that can be used to identify the potential answer sentences.

The *Z-score* value is a standard score that shows the number of standard deviations (σ) above the mean (μ) (Wonnacott and Wonnacott, 1990). This value is identified for each sentence by exploring all possible Z values using equation (1) that makes the similarity *st-sim* of that sentence is just equal or above the score X.

$$X = \mu + Z\sigma \quad (1)$$

4 Result and Discussion

Table (1) describes the performance of the claim zoning component using corpus-1. Table (2) describes the performance of Hirohata et al. (2008) system using the same corpus. Although, the difference was not significant, our system showed an alternative set of features that can achieve at least similar performance to the state of the art systems.

	Precision	Recall	F1-score
Introduction	0.96	**0.95**	**0.96**
Method	**0.83**	0.82	0.83
Results	0.87	**0.89**	**0.88**
Conclusions	**0.93**	**0.92**	**0.92**
Overall	**0.89**	**0.89**	**0.89**

Table 1: Claim zoning performance

	Precision	Recall	F1-score
Introduction	0.96	0.94	0.95
Method	0.81	**0.84**	0.83
Results	**0.88**	0.86	0.87
Conclusions	0.91	0.91	0.91
Overall	0.88	0.88	0.88

Table 2: Hirohata et al. (2008) system performance.

The output of the first component, particularly the sentences in the results and conclusions sections were then used as input in the answer detection component. That component was trained and evaluated on *corpus-2* using 10-folds cross validation. Table (3) shows the component's performance using five different combinations of features as follows:

- feature-set 1: *st-sim, sq-sim*
- feature-set 2: *Z-score, sq-sim*
- feature-set 3: *st-sim, role-label*

- feature-set 4: *Z-score, role-label*
- feature-set 5: *st-sim, sq-sim, role-label*
- feature-set 6: *Z-score, sq-sim, role-label*

The goal of trying different features combinations was to measure the effect of the *Z-score* feature on enhancing the overall performance of the component. The component achieved F1-score of 45% using set 1 compared to 56% using set 2. At this stage it was clear that the *Z-score* feature outperformed the *st-sim* feature.

Next, the *sq-sim* feature was replaced with the *role-label* as in set 3 and 4; however the results showed that using set 3 enhanced the F1-score by 22% compared to using set 1; and 19% using set 4 compared to set 2. This suggested that combining the *st-sim* feature with *sq-sim* was better than combining the *Z-score* and *sq-sim*.

The experiment was repeated using set 5 and set 6 which included the *sq-sim* feature in set 3 and 4; and the results were consistent with the results of using set 3 and 4. The component using set 5 outperformed set 6 due to the recall score (85%) in set 5. However, the precision score using set 6 was higher than using set 5 (73% vs 70%). This result was consistent with the component performance using set 3 and 4.

The above experiments showed a comparison between the *st-sim* and the *Z-score* features. The results suggest that using the *Z-score* feature contributes more than the *st-sim* feature with respect to the precision score, but less with respect to the recall score.

	Precision	Recall	F1-score
features-set(1)	0.68	0.34	0.45
features-set(2)	0.67	0.48	0.56
features-set(3)	0.70	**0.85**	**0.77**
features-set(4)	**0.73**	0.78	0.75
features-set(5)	0.70	0.83	0.76
features-set(6)	**0.73**	0.75	0.74

Table 3: The performance of the answer detection component using different combinations of features

Table (4) shows the performance of integrating the two components (the claim zoning and answer detection) using *corpus-3*. Note that the corpus only consists of unstructured abstracts (see *section (3.2.1)*. The integrated system was able to achieve precision of 56%, recall of 57% and F1-score of 56%. The main reason for the reduction in the performance score was due to the number of the answers examples used in the corpus being relatively small (69 answers). Another reason was the errors generated from the claim zoning component, which may have influenced the decisions made by the answer detection component.

	Precision	Recall	F1-score
Answer	0.56	0.57	0.56
Non-answer	0.92	0.92	0.92
Overall	0.86	0.86	0.86

Table 4: Answer detection performance using both components

5 Conclusion

This paper explored the problem of identifying the sentence(s) in an abstract that are the most informative information for a given query. It described a system for automatically identifying these sentences that consisted of two components: claim zone detection and answers detection. The system used the attribute value of *nlmCategory* to learn the sentences roles, which was found useful. Moreover, the component used different set of features that achieved at least similar performance to other systems for similar task. Finally, the research examined a new feature (*Z-score*) that was extracted from the same information used in *(st-sim)* feature. The *Z-score* feature was found more useful to enhance the precision score of the system compared with the *st-sim*.

References

Shashank Agarwal and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180, Dec.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Gough, Sandy Oliver, and James Thomas. 2012. *An introduction to systematic reviews*. Sage Publications.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *In Proc. of the IJCNLP 2008*.

Antonio J. Jimeno-Yepes, J Caitlin Sticco, James G. Mork, and Alan R. Aronson. 2013. Generif indexing: sentence selection based on machine learning. *BMC Bioinformatics*, 14:171.

Ryan T.K. Lin, Hong-Jei Dai, Yue-Yang Bow, Min-Yuh Day, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2008. Result identification for biomedical abstracts using conditional random fields. In *Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on*, pages 122–126.

Ryan T. K. Lin, Hong-Jie Dai, Yue-Yang Bow, Justing Lian-Te Chiu, and Richardg Tzon-Han Tsai. 2009. Using conditional random fields for result identification in biomedical abstracts. *Integr. Comput.-Aided Eng.*, 16(4):339–352, December.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs). http://www.chokkan.org/software/crfsuite/.

Constantin Orasan. 2001. Patterns in scientific abstracts. In *In Proceedings of Corpus Linguistics 2001 Conference*, pages 433–443. Lancaster University.

S. Otani and Y. Tomiura. 2014. Extraction of key expressions indicating the important sentence from article abstracts. In *Advanced Applied Informatics (IIAIAAI), 2014 IIAI 3rd International Conference on*, pages 216–219.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Anna M. Ripple, James G. Mork, John M. Rozier, and Lou S. Knecht. 2012. Structured abstracts in medline: Twenty-five years later.

Patrick Ruch, Clia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbhler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, and Anne-Lise Veuthey. 2007. Using argumentation to extract key sentences from biomedical abstracts. *I. J. Medical Informatics*, 76(2-3):195–200.

A.S. Schwartz and M.A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *In Proceedings of Pacic Symposium on Biocomputing*, volume 4, pages 451–462, November.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Comput. Linguist.*, 28(4):409–445.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

David L. Wheeler, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael Dicuccio, Ron Edgar, Scott Federhen, Lewis Y. Geer, Yuri Kapustin, Oleg Khovayko, David L, David J. Lipman, Thomas L. Madden, Donna R. Maglott, James Ostell, Vadim Miller, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Re Souvorov, Grigory Starchenko, Roman L. Tatusov, Tatiana A. Tatusova, Lukas Wagner, and Eugene Yaschenko. 2008. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, pages 13–21.

Thomas H. Wonnacott and Ronald J. Wonnacott. 1990. *Introductory Statistics*. John Wiley and Sons, fifth edition edition.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.