Towards Citation-Based Summarization of Biomedical Literature

Arman Cohan, Luca Soldaini, Saket S.R. Mengle, Nazli Goharian

Georgetown University, Information Retrieval Lab, Computer Science Department
arman@ir.cs.georgetown.edu, luca@ir.cs.georgetown.edu
saketmengle@gmail.com, nazli@ir.cs.georgetown.edu

Abstract

Citation-based summarization is a form of technical summarization that uses citations to an article to form its summary. In biomedical literature, citations by themselves are not reliable to be used for summary as they fail to consider the context of the findings in the referenced article. One way to remedy such problem is to link citations to the related text spans in the reference article. The ultimate goal in TAC[1] biomedical summarization track is to generate a citation-based summary, using both the citations and the context information. This paper describes our approach for finding the context information related to each citation and determining their discourse facet (Task 1 of the track). We approach this task as a search task, applying different query reformulation techniques for retrieving the relevant text spans. After finding the relevant spans, we classify each citation to a set of discourse facets to capture the structure of the referenced paper. While our results show 20% improvement over the baseline, the efficiency of the system still leaves much room for improvement.

1 Introduction

A set of citations to an article can be used for its summarization. This summary is a community-generated summary and it is called citation summary of the paper (Elkiss et al., 2008), (Qazvinian et al., 2013). Citation summaries reflect the most important points of the original paper including its different contributions to the scientific community. One benefit of using citations for summary is that they capture the impact of the paper on the community. They may also include comparisons with similar findings from other papers providing further insight into their impact.

However, citations by themselves report findings without considering the context in the original paper. This is specially important in biomedical literature, since circumstances, data and assumptions under which certain findings were obtained are very important in interpreting the results. By finding the related information to each citation in the reference article and using this information alongside the citations, one can alleviate the problem of lack of context in citation summaries. That is the main motivation of task 1a in TAC's Biomedical Summarization track. In this task, the goal is to find text spans in the reference article that best describe the citation text. These text spans are later used to generate the summary of the paper.

We approach this problem as a search task. That is, we index the reference article into different text spans and use the citation text as a query to retrieve the relevant parts. This approach, being search oriented and unsupervised, is highly efficient and scalable in comparison with other text comparison and classification methods. As TAC biomedical summarization track focuses on articles in biomedical literature, we also apply domain targeted query reformulations for finding the reference text spans. After finding the related text spans, we associate each of them with a discourse facet that best describes them. A discourse facet shows the rhetorical function of

[1] Text Analysis Conference

the citation in the reference article describing why it has been cited. The discourse facet can be one of the following: hypothesis, method, results, implications or discussion. The goal of this part (task 1b) is to create a logical ordering of the citations so they can be used in the final summary.

Previous work has studied the citations and the way the can be used for summarization. (Qazvinian and Radev, 2008) analyzed the network of citations to an article to generate its summary. (Elkiss et al., 2008) did a study on the information that exist in the citation texts and concluded that they often include additional information that is absent from the article's abstract. (Abu-Jbara and Radev, 2011) further improved citation-based summaries by focusing on the coherency of the generated summaries. (Teufel et al., 2006) studied the reason why a citation cites a paper by classifying citations into a set of predefined categories.

2 Problem definition

The goal of the system is to identify text segments (text spans) in the reference article that are most relevant to a given citation text. Formally, given a citation text C and a reference text $R = \{s_1, s_2, ...s_n\}$ in which s_i are the semantic units (each can consist of one sentence up to 5 sentences) in the reference text and n is the total number of these units in the reference text, the goal is to find an ordered subset of units $S = \{s'_1, ..., s'_m\}; s'_i \in R$ that is most related to the citation text C.

3 Methodology

In this section we describe our main methodology for the task. First we index the text spans s_i in the reference article $R = \{s_1, s_2, ...s_n\}$. We consider the smallest semantic unit as a set of consecutive sentence from length 1 up to 5. This selection is based on the annotation guidelines which state that a reference text span can include 1 to 5 sentences. Our methodology consists of the following steps:

1. Create a sentence level index from the reference article in which each semantic unit s_i is indexed.
2. Find the most relevant text spans using the citation text C as the query.
3. Rerank and merge the retrieved spans to form the final subset S of R that correctly provides context for the citation text C.
4. Classify each citation to a discourse facet that best describes it's function within the paper.

3.1 Model for identification of the relevant spans (Task 1a)

We use the vector space retrieval model for retrieving the related reference spans. Specifically, we use this model to measure the cosine similarity of a given citation with each text span in the reference article.

After retrieving the initial spans, we combine and merge these spans to form the final result set. This is based on the fact that indexed spans can overlap each other. The number of such spans that overlap indicates the importance of that part of the article. That is, if in top results we have many spans that have some overlap with each other, we rank them higher than another span with no overlap with other results. Therefore, we rerank the retrieved results based on the number of overlapping spans. We also merge the overlapping spans to a single span, which is the union of these spans. Finally, we choose a cut-off point for our ranked list of spans and return the spans that are above that cut-off point. Our cut-off point is set to 3, following the specifications of the TAC's annotation guidelines in which the retrieved spans can be up to 3 different segments of the text.

3.2 Query reformulations for identification of the relevant spans

We applied several query reformulation techniques on top of our retrieval model for finding the relevant text spans to citations. The citation text by itself as the query is often very large and includes terms that are not informative (do not represent the content of the query). Therefore, we reduce the query to limit it to only informative terms. On the other hand, the author of the citing article and reference article might use different terminology to refer to same concepts. To address this, we also expand the query to include the related biomedical concepts. Our query reformulation approaches are described below:

3.2.1 Unmodified query - baseline

We consider the citation text as the query after preprocessing and removing the citation marker (i.e., the actual indicator of the citation), we use this method as our baseline.

3.2.2 Biomedical concepts

We reduce the query to contain only the biomedical concepts in the citation. To do so, we take advantage of two thesauri. First, we use the MeSH terms thesaurus; in this approach we reduce the query to only contain the terms that match one of terms in the MeSH thesaurus. MeSH (Medical Subject Headings)[1] is a thesaurus that contains biomedicine and health related terminology; it is maintained by NLM[2]. We call this method *MeSH_terms* throughout the rest of the paper. Second, we use the comprehensive biomedical thesaurus, UMLS[3]. This approach works similar to *MeSH_terms* by only keeping the terms that match a UMLS concept. We use MetaMap[4] to map text to UMLS medical concepts. We refer to this method as UMLS*_concepts*.

3.2.3 Noun phrases

We observed that most of the important terms and medical concepts in a query are in form of noun phrases. Hence, we extract noun phrases from the query and remove all other terms. Our chunks are up to 3 terms, since long noun phrases will be too specific and highly unlikely to match any phrase in the target textual content.

3.2.4 Keyword extraction

Informative keywords are more likely to help us in identifying the correct textual spans. We use a statistical measure to find term informativeness. Specifically, we use *idf* (inverse document frequency) of the terms as an indicator of their importance. We leveraged Wikipedia to calculate the *idf* of the terms in the citation text and then filter out the terms that do not meet a minimum *idf* threshold. We chose the threshold empirically based on the resource it was drawn from. We refer to this method as *idf-wiki* throughout the rest of the paper.

3.2.5 Wikipedia health terms

Inspired by (Parker et al., 2013) and (Soldaini et al., 2015), we use Wikipedia to filter non health-related terms. Specifically, we estimate for each term its likelihood of being associated with a health-related page on Wikipedia by evaluating the odds ratio between the probability of that term appearing in a health-related Wikipedia page over its probability of appearing in a non-health related Wikipedia page. For each term t, we calculate its likelihood of being associated with a health-related Wikipedia entry:

$$\text{OR}(t) = \frac{Pr\{\text{P is health related } |t \in \text{P}\}}{Pr\{\text{ P is not health related } |t \in \text{P}\}} \quad (1)$$

In which OR(t) is the odds ratio of term t belonging to a health related wikipedia page P over the probability of t appearing in a non-health related Wikipedia page P. We consider the term t as health-related if it's odds ratio is above some threshold δ. We empirically set δ to 5. We refer to this method as *wiki-health-terms*.

3.2.6 Combination of reduction and expansion approaches

By using the UMLS ontology, we find related medical concepts to the terms that exist in the citation text and expand the original citation with the relevant biomedical concepts. Specifically, we first reduce the citation text using one of the described methods above to limit it to contain potentially informative terms. Then we use the UMLS terminology for expanding the concepts by adding other biomedical terms that are related to them. We do not expand concepts for the following semantic types: "functional concepts", "qualitative concepts", "quantitative concept" and "intellectual product"[5]. These types are not related to a specific biomedical con-

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

[2] National Library of Medicine

[3] Unified Medical Language System

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

[5] Functional concept: A functional concept pertains to the carrying out of a process or activity*. Qualitative concepts: Concepts which are assessment of some quality, rather than a direct measurement*. Quantitative concepts: A concept which involves the dimensions, quantity or capacity of something using some unit of measure, or which involves the quantitative comparison of entities*. Intellectual product: A conceptual entity resulting from human endeavor. Concepts assigned to this type generally refer to information created by humans for some purpose*.

* http://semanticnetwork.nlm.nih.gov/Download/RelationalFiles/SRDEF

cept and therefore expanding them would introduce many general terms and cause query drift.

3.3 Identifying the citation facet (Task 1b)

After identifying the related text spans for each citation, we associate each with a specific discourse facet. Discourse facets are to be selected from the following predefined values: hypothesis, method, results, implication and discussion. We use supervised algorithms to predict the discourse facet for each citation. Discourse facets could later be used in generating a coherent and comprehensive summary of the referenced article. We use both the citation and reference text spans as training data for our classifier. We use tf-idf features for training the classifier after stopword removal and stemming. We train five classifiers for this task: Support Vector Machine (SVM), Supervised Latent Dirichlet Allocation (SLDA), Decision Tree, Boosting and Random Forests, as well as the ensemble of these classifiers. For training and testing, ten fold cross validation is used.

4 Dataset

The TAC Biomedical Summarization training dataset consists of 20 topics, each of which having a set of citing articles and one reference article. For each topic, four annotators have annotated the citation texts, the corresponding reference spans in the designated reference article, and the discourse facet. To have a better understanding of the TAC's dataset, we performed some statistical analysis on data, which we present in Table 1 and Table 2.

In Table 1, *Full overlap* means that the offsets for correct reference spans identified by the different annotators should fully overlap with each other. *Partial overlap* means that the intersection between identified spans should not be empty (e.g. the following text spans: offsets: [200-400] and [350-700]). *Majority* of annotators indicates three out of four and *minority* indicates that two out of four annotators agree on a span (partially or fully). *Number of combinations* refers to different combinations of annotators. For example, partial agreement with 2 combinations means that there are two sets of annotators that agree with each other at least partially. (e.g. There is overlap between correct offsets identified by annotator "A" and annotator "B", and overlap between annotator "C" and annotator "D"). As it is shown in the table, there is not a single citation whose reference span is agreed upon by all annotators. The number of citations whose reference spans are agreed partially by majority of annotators is also limited. Overall low agreement among annotators, corroborates the fact that this task is highly non-trivial even for the domain expert.

For task 1b, the training data consists of the discourse facet for each citation in topics determined by each annotators. Our analysis of the data shows that the agreement on the annotation of discourse facets among annotators is similarly low (Table 2). The Fleiss' Kappa agreement among annotators in annotating the correct discourse facet is 0.187. The dataset is also unbalanced for different discourse facets (Table 3).

5 Evaluation

Evaluation of task 1a is based on the weighted overlaps between the retrieved spans and the correct spans identified by annotators. Character level precision and recall is used for the evaluations which are calculated based on agreement between annotators. Specifically, weighted precision and weighted recall for a system returning a span S with respect to a set of annotations from m assessors, consisting ground truth spans G_1, ..., G_m are defined as follows:

$$\text{WeightedRecall} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{m} |S \cap G_i|}{\sum_{i=1}^{m} |G_i|} \qquad (2)$$

$$\text{WeightedPrecision} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{m} |S \cap G_i|}{m \times |S|} \qquad (3)$$

The overall performance is measured by Weighted F-1, i.e the harmonic mean of weighted average of precision and recall.

Task 1b is evaluated on the weighed accuracy of the correct citation facets. Specifically, the weighted accuracy $A_w(f)$ for a returned discourse facet f is defined as:

$$A_w(f) \stackrel{\text{def}}{=} \frac{|(F_i : F_i = f)|}{m} \qquad (4)$$

In which F_i is the facet identified by annotator i for i=$\{1, ..., m\}$; m is the total number of annotators and $(.)$ denotes a list of items. Therefore a 100% accuracy is only obtainable if all annotators agree on the correct discourse facet.

Type of agreement, subset of annotators, [comments]	number of annotations	average overlap	standard deviation of overlaps
total	313	-	-
full, all	0	-	-
partial, all	66	21.77%	±15.44%
full, majority	1	-	-
partial, majority, (1 combination)	104	19.08%	±11.13%
partial, majority, (2 combination)	17	21.27%	±14.26%
full, minority	1	-	-
partial, minority, (1 combination)	45	31.76%	±17.79%
partial, minority, (2 combinations)	46	26.26%	±16.55%
partial, minority, (3 combinations)	19	21.10%	±12.56%
partial, minority, (4 combinations)	3	14.45%	±5.27%
no overlap	11	-	-

Table 1: Our analysis of the dataset for task 1a. Full agreement: complete overlap between identified offsets; Partial: There exists some overlap between identified offsets; Majority: three annotators; Minority: two annotators; Combinations: sets of annotators that agree with each other; the overlap percentage and standard deviations are undefined when there is no agreement or full agreement between annotators.

Type of agreement	number of annotations
Full agreement,	45
Majority agreement	123
Minority agreement	97
Tie	45
No agreement	4

Table 2: Our analysis of the dataset for task 1b. Agreement between annotators in identifying discourse facets. Majority means 3 out of 4 annotators agree on a facet, minority means 2 out of 4 agree on a facet and tie means two annotators agree on one facet and two others on another facet.

	M	H	I	D	R
number of facets	155	21	140	446	490

Table 3: Facet category distribution in the dataset, facets are abbreviated by following letters: M: Method, H: Hypothesis, I: Implication, D: Discussion and R: Results.

Method	recall (% increase)	precision (% increase)	F-1 (% increase)
random	0.0421 (-74.64%)	0.0449 (-71.81%)	0.0401 (-75.47%)
baseline	0.166 (0.00%)	0.1593 (0.00%)	0.1635 (0.00%)
MeSH_terms	0.105 (-36.75%)	0.1075 (-32.52%)	0.1038 (-36.51%)
UMLS*_concepts*	0.1887 (+13.67%)	0.173 (+8.60%)	0.1782 (+8.99%)
noun_phrases	0.209 (+25.90%)	0.1689 (+6.03%)	0.1846 (+12.91%)
idf-wiki	0.1285 (-22.59%)	0.1051 (-34.02%)	0.1143 (-30.09%)
wiki-health-terms	0.0793 (-52.23%)	0.0753 (-52.73%)	0.0755 (-53.82%)
comb_1	0.2139 (+28.86%)	0.1842 (+15.63%)	0.1957 (+19.69%)
comb_2	**0.2147 (+29.34%)**	**0.1855 (+16.45%)**	**0.1967 (+20.31%)**

Table 4: Results of identification of correct reference spans for all the methods (task 1a). % increase indicates relative increase to the baseline. Comb_1 is the combination of UMLS_concepts reduction with query expansion. Comb_2 is the combination of UMLS_concepts and noun_phrases reductions along with query expansion. *random* shows the performance of random retrieval.

	Random	**Probability Voting**	**Logit Boost**	**SLDA**	**Random Forests**	**SVM**	**Tree**	**Ensemble Voting**	**Oracle**
Weighted Accuracy	0.1094	0.4224	0.4489	0.3842	0.4864	**0.5256**	0.4065	**0.5**	0.6665

Table 5: Mean weighted accuracy for different methods for identification of the citation facets (task 1b); Oracle shows the maximum possible weighted accuracy; Random is the performance of a random classifier.

6 Results and discussion

The results for task 1a are shown in Table 4. *Random* refers to the performance of a random retrieval system that randomly returns text spans from the indexed document. The *baseline* method is the unmodified query which achieves F-1 score of 0.164. We compared the performance of all approaches against the *baseline*.

We observe that the performance of *MeSH_terms* is poor with F-1 score of 0.104; we attribute this to the focused vocabulary that exist in MeSH. In particular, using MeSH to reduce the query leaves us only with highly focused concepts many of which might not appear in the target paper with the same form. More importantly, many less specific words will not be selected. UMLS*_concepts* is essentially the same approach, but uses UMLS thesaurus for query reduction. This approach works better than the baseline (+8.99% higher F-1) since UMLS thesaurus consists of a broader range of biomedical and biomedicine related concepts and in comparison with *MeSH_terms*, captures a higher number of important concepts in the citation.

Using noun phrases for query reduction also shows improvement over the baseline (+12.91% higher F-1). This is due to the fact that many informative terms that help in identifying the correct spans are noun phrases in the citation sentence. The statistical keyword extraction method (*idf-wiki*) performs poorly with F-1 score of 0.114. We observed that many terminology used in the biomedical articles (e.g. names of specific proteins and genes or their codes) are not mentioned in any Wikipedia entry. That is why Wikipedia index fails to capture keywords in this domain. In order for this approach to work, one needs to opt for a better knowledge base that is suited for this domain for extracting *idf* values.

The reduction approaches that outperform the baseline, are UMLS*_concepts* and *noun_phrases*. As the wordings between the referenced authors and the citing authors differ, we expect to further improve the performance by using query expansion. In fact, our results show that the overall best performing methods are these combination approaches. Our expansion method adds the related biomedical terminology from UMLS to the selected terms from the query. In the first approach (*comb_1*), we use UMLS_concepts to reduce the query and then only use those concepts to expand the query. With *comb_1*, we could achieve 0.196 F-1 score. In second combination approach (*comb_2*), we use both noun phrases and UMLS concepts for reduction and biomedical terminology from the UMLS thesaurus for expansion. This approach, yielded The highest overall F-1 score among all methods (0.197). We did not observe any significant differences between these two methods.

The overall low performance of all methods in terms of weighted precision and recall is expected because of the difficulty of the task in finding exact related text spans and also the fact that the performance measures are computed at character level. The latter aspect makes it difficult for any system to achieve high levels of F-1, as it needs to exactly match the same spans as the annotators. As it was previously mentioned, this fact is also reflected in the low agreement among domain expert annotators.

Table 5 shows the results of classification of citations into different discourse facets. We calculated the performance of each of the runs that we have submitted using the validation data. The training and test was done using 10 fold cross validation. As it is shown in Table 5, we observe that SVM algorithm yields the best accuracy (0.526). The ensemble of SVM and random forest algorithms also shows high performance. We experimented with two methodologies for ensemble classifiers. The first approach used the probabilities generated by both the classifiers to weigh their prediction, while

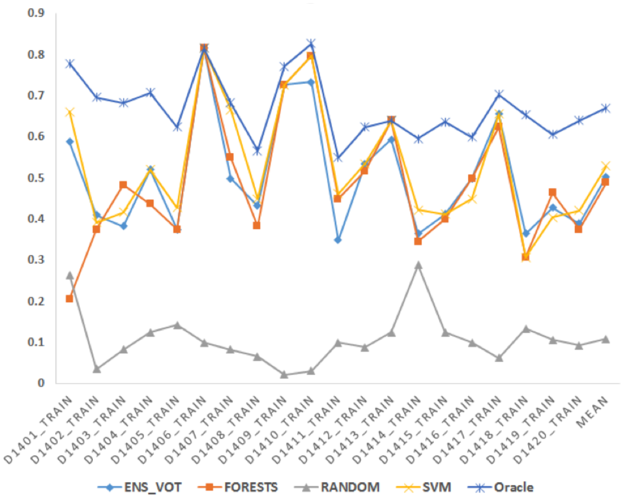

Figure 1: Mean weighted accuracy for each topic. The oracle is indicated with dark blue line (the topmost line) and shows the maximum possible achievable accuracy.

the second approach used the actual ranks of predictions. Both approaches yielded similar results. Random forests algorithm uses bootstrap aggregations of decision trees and shows significantly better performance than decision tree. We also observed significantly lower accuracy for SLDA and Boosting and decision tree approaches.

On this classification task, an oracle would get the maximum score of 0.667 as indicated in the table (highest possible score). Such system always returns the discourse facet identified by majority of annotators. Due to the low agreement between annotators, the oracle score is also relatively low. Comparison of our best method with the oracle shows reasonable performance for task 1b.

The results of classifications per each topic are also shown in figure 1. This figure shows the performance of our top 3 methods as well as the highest possible accuracy achievable by the oracle for each topic. The performance of a random classifier is included for reference. As it is illustrated, we achieved the highest results for topics 6, 9 and 10. The per topic performance chart shows that low accuracy is for topics with lower agreement among the annotators as reflected in the oracle score. We can see that our top methods' performance is low on the topics that the oracle is also performing low.

7 Submitted runs

Based on our experiments on the training data, we chose two of our best approaches from task 1a (combination approaches) and two of our best approaches from task 1b (SVM and Ensemble voting) and we submitted 4 different combinations of them for the track (run #1 to #4). In the analysis of dataset, we observed that some annotators had identified reference spans in parts that are not in the main body of the text (e.g figure captions, tables, etc). Since the documents were parsed from PDF, contents of the tables and figures are also present in the text files. These sections include keywords that cause performance loss and in the preprocessing step these usually need to be removed. But based on training data, sine some annotations included reference spans from these sections, we had to also include them in our index. By the intuition that usually the spans belong to main body of the article and not to figure captions and tables, our last run consists of our best methods for task 1a and 1b, ran on the filtered documents in which figures, tables, acknowledgments and other non-pertinent sections were removed from the index (run # 5).

8 Conclusion

In this paper we described our system for the first task of TAC's biomedical summarization track. We approached the problem, from an information retrieval perspective and used different indexing and query reformulation methods for retrieving the correct results. While we could obtain up to 20% improvement over the baseline, the low overall weighted F-1 score, proves the difficulty of this task in comparison with regular text retrieval tasks. This fact is further confirmed by observing high disagreement between annotators in identification of correct reference spans. This proves that the task is non-trivial and demands further exploration.

9 Acknowledgments

This work was partially supported by the US National Science Foundation through grant CNS-1204347.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 500–509. Association for Computational Linguistics.

Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir Radev. 2008. Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1):51–62.

Jon Parker, Yifang Wei, Andrew Yates, Ophir Frieder, and Nazli Goharian. 2013. A framework for detecting public health trends with twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 556–563, New York, NY, USA. ACM.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 689–696, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vahed Qazvinian, DR Radev, and SM Mohammad. 2013. Generating Extractive Summaries of Scientific Paradigms. *J. Artif. Intell.*, 46:165–201.

Luca Soldaini, Arman Cohan, Andrew Yates, Nazli Goharian, and Ophir Frieder. 2015. Retrieving medical literature for clinical decision support. In *37th European Conference on Information Retrieval*, ECIR '15.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110. Association for Computational Linguistics.