Matching Citation Text and Cited Spans in Biomedical Literature: a Search-Oriented Approach

Arman Cohan, Luca Soldaini, Nazli Goharian
Georgetown University, Information Retrieval Lab, Computer Science Department
{arman, luca, nazli}@ir.cs.georgetown.edu

Abstract

Citation sentences (citances) to a reference article have been extensively studied for summarization tasks. However, citances might not accurately represent the content of the cited article, as they often fail to capture the context of the reported findings and can be affected by epistemic value drift. Following the intuition behind the TAC (Text Analysis Conference) 2014 Biomedical Summarization track, we propose a system that identifies text spans in the reference article that are related to a given citance. We refer to this problem as citance-reference spans matching. We approach the problem as a retrieval task; in this paper, we detail a comparison of different citance reformulation methods and their combinations. While our results show improvement over the baseline (up to 25.9%), their absolute magnitude implies that there is ample room for future improvement.

1 Introduction

The size of scientific literature has increased dramatically during recent decades. In biomedical domain for example, PubMed – the largest repository of biomedical literature – contains more than 24 million articles. Thus, there is a need for concise presentation of important findings in the scientific articles being published. Text summarization of scientific articles is a method for such presentation. One obvious form of scientific summaries, is the abstract of the articles. Another type of scientific summaries relates to citance-based summaries which are summaries created using the set of citations to a reference article. This kind of summary covers some aspects of the reference article which might not be present in its abstract (Elkiss et al., 2008).

Citances often cover important and novel insights about findings or aspects of a paper that others have found interesting; thus, they capture contributions that had an impact on the research community (Elkiss et al., 2008; Qazvinian and Radev, 2008).

Reference Article

(Voorhoeve et al., 2006): *"These miRNAs neutralize p53-mediated CDK inhibition, **possibly** through direct inhibition of the expression of the tumor suppressor LATS2."*

Citing Article

(Okada et al., 2011): *"Two oncogenic miRNAs, miR-372 and miR-373, **directly inhibit** the expression of Lats2, **thereby** allowing tumorigenic growth in the presence of p53 (Voorhoeve et al., 2006)."*

Figure 1: Example of epistemic value drift from (De Waard and Maat, 2012). The claim in (Voorhoeve et al., 2006) becomes fact in (Okada et al., 2011).

In the past, many have focused on citance extraction and citance-based summarization. Example of citance extraction include (Siddharthan and Teufel, 2007), who used a machine learning approach with linguistic, lexical, statistical and positional features, and (Kaplan et al., 2009), who studied a coreference resolution based approach. Citance extraction has been also studied in the context of automatic summarization. For example, (Qazvinian and Radev, 2010) proposed a framework based on probabilistic inference to identify citances, while (Abu-Jbara and Radev, 2011) approached the problem as a classification task. In the biomedical domain, the use of citances was first studied by (Nakov et al., 2004).

While useful, citances by themselves lack the appropriate evidence to capture the exact content of the original paper, such as circumstances, data and assumptions under which certain findings were obtained. Citance-based summaries might also modify the epistemic value of a claim presented in the cited work (De Waard and Maat, 2012); that is, they might report a preliminary result or a claim as a definite fact (example in figure 1).

Recently, a new track at TAC has been introduced to explore ways to generate better citance-based

summaries[1]. One way to achieve this, is to link citances to text spans in the reference article to obtain a more informative collection of sentences representing the reference article (figure 2). A framework designed to solve such problem requires two components: (*i*) a method to identify the most relevant spans of text in the reference text and (*ii*) a system to automatically generate a summary given a set of citances and reference spans.

In this paper, we propose an information retrieval approach designed to address the first task. We explore the impact of several query reformulation techniques – some domain independent, others tailored to biomedical literature – on the performance of the system. Furthermore, we apply combined reformulations, which yields an additional improvement over any single method (25% over the baseline).

As a related area, passage retrieval in biomedical articles has been studied in the context of the genomics track (Hersh et al., 2006; Hersh et al., 2007) and in following efforts (Urbain et al., 2008; Urbain et al., 2009; Chen et al., 2011). In these works, the goal is to find passages that relate to a given term or keyword (e.g. GeneRIF). In contrast, our system considers citances as queries, which are substantially longer than keyword-based queries and have a syntactical structure.

In summary, our contributions are: (*i*) A search-based, unsupervised (thus easily scalable to other domains) approach to citance-reference spans matching and (*ii*) adaptation of various query reformulation techniques for the citatnce-refrence span matching.

2 Methodology

The goal of the proposed system is to retrieve text spans from the reference paper that match the finding(s) each citance is referring to. We approach this problem as a search task. That is we consider the citance as a query and the reference text spans as documents. Then, using a retrieval model along with query reformulation, we find the most relevant text spans to a given citance. Our methodology consist of the following steps:

1. Create sentence level index from the reference article.

[1]http://www.nist.gov/tac/2014/BiomedSumm/

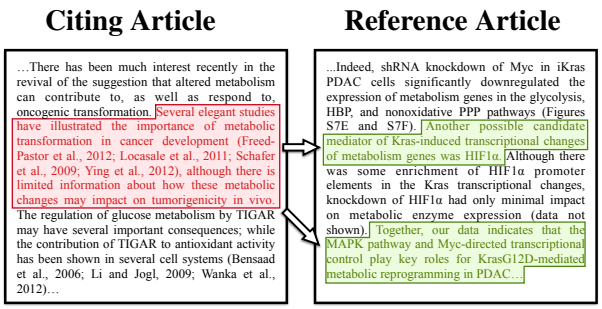

Figure 2: Example of a citance/reference article pair from the TAC training set[1]. The text in the red box on the left is referred to as the citance text, while the text in the green boxes on the right is referred to as the reference text.

2. Apply query reformulation to the given citance and retrieve the most relevant spans.
3. Rerank and merge the retrieved spans that correctly describe the citance.

We will describe each step in the following sections.

2.1 Creating the index

To create an index of spans, each reference article is tokenized at a sentence level using the Punkt tokenize (Kiss and Strunk, 2006). Because each relevant reference span in the reference text can be formed by several consecutive sentences (according to the annotation guidelines, each span can consist of one up to five consecutive sentences), we index text spans comprised of one up to five sentences.

2.2 Retrieval model

We evaluated the performance of several retrieval models during experimentation, i.e. vector space model (Salton et al., 1975), probabilistic BM25 (Robertson and Zaragoza, 2009), divergence from randomness (DFR) (Amati and Van Rijsbergen, 2002), and language models (Ponte and Croft, 1998) with Dirichlet priors. All models showed very similar performances (with only DFR constantly underperforming all other models) and we did not observe any statistically significant differences between each set of runs. Therefore, we opted for the vector space model as our retrieval model.

2.3 Query reformulation

We apply several query reformulation techniques to the citance to better retrieve the related text spans. We leverage both general and domain specific query reformulations for this purpose. Specifically,

we use biomedical concepts, ontology information, keyphrases and the syntactic structure of the citance.

2.3.1. Unmodified query (*baseline*): The citance after removing stop words, numeric values and citation markers (i.e. the actual indicator of the citation) serves as our baseline.

2.3.2. Biomedical concepts (UMLS-*reduce*): We remove from the query those terms that do not map to any medical concept in the UMLS[1] metathesaurus. We use MetaMap (Aronson, 2001) to map biomedical expressions in the citances to UMLS concepts. More specifically, our heuristic greedily matches the longest expressions in the citance to concepts in the UMLS metathesaurus; such strategy was deemed the most appropriate after experimenting with various matching approaches. We limited the scope of UMLS-reduce to SNOMED Clinical Terms (Bos et al., 2006) collection of UMLS and the "preferred concepts" (i.e., concepts that are determined by the National Library of Medicine to provide the best representation for a concept); terms that are not mapped to any UMLS concept were removed.

2.3.3. Noun phrases (*NP*): Citances include many important biological concepts, often appearing as noun phrases. For this reason, we reformulate citance by only keeping noun phrases and filtering out other parts of speech. We retain noun phrases that consist of up to 3 terms, as longer phrases were empirically determined to be too specific. Stopwords are removed from noun phrases.

2.3.4. Keyword based (*KW*): We consider a statistical measure for identifying key terms in the citance. Specifically, we computed the *idf*[2] of the terms in the citance in a domain-specific corpus to evaluate their importance. Given the domain of our dataset, we used the Open Access Subset of PubMed Central[3]. We filter out the terms whose *idf* value is less than a fixed threshold (after empirical evaluation, this threshold was set to 2.5).

2.3.5. Biomedical expansion (UMLS-*expand*): The terminology used by the citing author and the referenced author is not necessarily identical. Multiple terms or multi-word expressions can be mapped to the same concepts and each author might use their own choice of terms for describing a concept. In this approach, we add related terminology to the important concepts in the citance to solve this issue. Since our dataset consists of articles from biomedical literature, we took advantage of the UMLS metathesaurus to expand terms or multi-word expressions with their synonyms. We did not enforce any threshold for the number of terms added by UMLS-*expand*. However, in order to prevent query drift, we expanded citances using only UMLS's "preferred concepts" and concepts from the "SNOMED Clinical Terms" (SNOMED CT) terminology.

2.3.6. Combined reformulation: Due to the narrative structure of citances and their relative long length, using all citance terms for expansion is likely to cause query drift. Therefore, we first reduce the citance using one of previously described reduction approaches and then apply query expansion. In detail, we evaluated the combination of noun phrases and UMLS expansion, as well as UMLS reduction and expansion.

2.4 Combining retrieved spans

Due to our indexing strategy described in section 2.1, some text spans retrieved by the search engine could overlap with each other. Intuitively, if a span containing multiple contiguous sentences $\{s_1, \ldots, s_l\}$ is retrieved alongside any of its constituent sentences s_i, its relevance score should be increased to account for the relevance of s_i.

We exploited such intuition by adding the score of each span with the score of any of the constituent sentences or sub-spans retrieved alongside it. After the score is updated, the constituent sentences or sub-spans are removed from the list of retrieved results. Finally, because the number of reference spans indicated by the annotators in our data set is at most three, the system returns the top three results.

It is worth mentioning that we also looked at some other query reformulation approaches such as pseudo relevance feedback (Buckley et al., 1995) and Wikipedia based biomedical term filtering (Cohan et al., 2014); however, our experimentations should that these methods performed substantially worse than the baseline, consequently, we do not report those results nor their relevant discussions.

[1] http://www.nlm.nih.gov/research/umls/
[2] Inverted Document Frequency
[3] http://www.ncbi.nlm.nih.gov/pmc/

Type of agreement	Count	Average overlap
Full agreement	2	100%
Partial agreement between all annotators	66	$21.7 \pm 15.4\%$
Partial agreement between a majority of annotators	121	$19.2 \pm 11.4\%$
Partial agreement between a minority of annotators	113	$27.0 \pm 15.9\%$
No agreement at all	11	0%

Table 1: Levels of agreement between annotators. The 4 annotators fully agree on just 2 of the 313 annotations. In most cases, a majority (3 annotators) or a minority (2 annotators) agrees on a portion of reference spans, indicating that the task is not trivial even for domain experts.

3 Evaluation and Dataset

The system was evaluated on TAC 2014 Biomedical Summarization track training dataset. It consists of 20 topics, each of which contains between 10 to 20 citing articles and 1 reference article. For each topic, four domain experts were asked to identify the appropriate reference spans for each citance in the reference text. To better understand the dataset, we analyzed the agreement between annotators (table 1). This table shows that the overall agreement is relatively low.

We used two sets of metrics for evaluation of the task. The first one is based on the weighted overlaps between the retrieved spans and the correct spans designated by annotators and is meant to reward spans overlapping with the ground truth. Weighted recall and precision for a system returning span S with respect to a set of M annotators, consisting of gold spans G_1, ..., G_M are defined as follows:

$$\text{Recall} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{M} |S \cap G_i|}{\sum_{i=1}^{M} |G_i|} \quad \text{Prec} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{M} |S \cap G_i|}{M \times |S|} \quad (1)$$

The overall score of the system is the mean F-1 (harmonic mean of the weighted precision and recall) over all the topics.

Based on the weighted F-1 score, a method could be penalized for retrieving any spans that are not indicated as gold spans by the annotators. Even if those spans are semantically similar to the gold spans, they will not receive any score. This is not ideal because, as the high disagreement shown in table 1 implies, gold spans by offset locations are highly controversial. For this reason, we also considered ROUGE-L (Lin, 2004) as another evaluation metric, as it rewards a method for retrieving spans that are similar to the gold spans. Specifically, ROUGE-L, takes into account the sentence similarity by considering the longest in sequence n-grams between the retrieved spans and gold spans.

4 Results and discussion

The problem of matching citations with cited spans in scientific articles is a new task and to the best of our knowledge, there is no prior work on this task. Thus to evaluate the effectiveness of our different methods, we compared the performance of our proposed approaches against the unmodified query baseline. The results are shown in Table 2.

Interestingly, we observe that UMLS-*reduce* performs worse than the baseline in terms of F-1. This can be attributed to the fact that multiple expressions in the biomedical literature can be used to refer to the same concept. Such diversity is not captured by UMLS-*reduce*, as it only performs query reduction. Moreover, a citance often contains expressions that, while not mapping to any biomedical concepts, provide useful context and therefore are fundamental in conveying the meaning of the citance (we will refer to such expressions as *supporting expressions* in the reminder of the paper). These supporting expressions are not captured by UMLS-*reduce*.

NP outperforms the baseline (+18.8% F-1). This outcome is expected, as most important biomedical concepts in the citance are noun phrases. Moreover, supporting expressions are also captured, as most of them are noun phrases.

KW also shows promising results (+11.5% F-1 and +15.2% ROUGE-L F-1 improvement), proving that the *idf* of the terms in citance over a large biomedical corpus is a valid measure of their informativeness for this task.

When comparing *KW* and *NP*, we notice that the former obtains higher precision values than the latter; this outcome is reversed with respect to recall (i.e., *NP*'s recall is higher than *KW*'s). Such behavior can be motivated by the fact that *NP*, as it extracts noun phrases that are likely to appear in the gold reference span, has a higher chance of retrieving relevant sections of the reference text. However, *NP* is more likely to retrieve non-relevant spans, as the extracted noun phrases, which are often describing the main findings of the cited paper, are preva-

	Recall	Precision	F-1	ROUGE-L Recall	ROUGE-L Prec	ROUGE-L F-1
baseline	0.169	0.152	0.156	0.496	0.200	0.280
UMLS-*reduce*	0.132 (-22.0%)	0.146 (-4.08%)	0.136 (-12.5%)	0.496 (0.0%)	0.224* (12.0%)	0.293 (4.8%)
KW	0.173* (3.0%)	0.193** (27.6%)	0.174** (11.5%)	0.491 (-0.1%)	0.273** (36.3%)	0.323** (15.2%)
NP	0.199** (18.3%)	0.178** (17.6%)	0.185** (18.8%)	0.550** (11.1 %)	0.211* (5.5 %)	0.280 (0.0%)
UMLS-*expand*	0.182** (8.1%)	0.148 (-2.1%)	0.160* (3.2 %)	0.498 (0.5%)	0.245** (22.2%)	0.315** (12.3%)
UMLS-*reduce* + UMLS-*expand*	**0.201** (19.6%)**	0.179** (18.0%)	0.187** (20.0%)	**0.558** (12.6 %)**	0.209** (4.4 %)	0.293* (4.4%)
NP + UMLS-*expand*	0.180* (7.1%)	**0.224** (47.8%)**	**0.196** (25.9%)**	0.501 (1.13%)	**0.280** (39.9%)**	**0.333** (18.8%)**

Table 2: Results for reference span matching; *KW*: reduction using KeyWords; *NP*: reduction using Noun Phrases; UMLS-*expand*: expansion using UMLS; UMLS-*reduce*: reduction using UMLS; * (**) indicates statistical significance at $p < 0.05$ ($p < 0.01$) using student's t-test over the baseline.

lent throughout the reference article. On the other hand, *KW* selects highly discriminative terms which are highly effective in retrieving some relevant reference spans, but might not appear in others.

We observe that UMLS-*expand*, by adding related concepts to the query, achieves significant improvement over the baseline in terms of recall (+8.1%). Such improvement is expected, as UMLS-*expand* augments the citance with all possible formulations of the detected biomedical concepts. However, its precision is only comparable with the baseline, as it does not remove any noisy terms from the citance. Interestingly, we notice that its ROUGE-L precision greatly outperforms the baseline (+22.2%). This behavior is motivated by the fact that UMLS-*expand*, even when not retrieving all the correct reference spans, extracts certain parts of the reference articles that share many biomedical concepts with the gold spans, thus achieving high structural similarity.

The two combined methods (*NP* + UMLS-*expand* and UMLS-*reduce* + UMLS-*expand*) obtain the best overall performance compared to the baseline. UMLS-*reduce* + UMLS-*expand* obtains the highest recall among all methods. This outcome directly depends on the fact that all the synonyms of a certain biomedical concept are captured using UMLS-*expand*. However, unlike UMLS-*expand*, this combined method also achieves statistically significant improvement in terms of precision, as UMLS-*reduce* removes terms that can cause query drift.

NP + UMLS-*expand* has the highest overall performance, achieving a 25.9% increase over the baseline in terms of F-1, and an 18.8% increase in terms of ROUGE-L F-1. As previously mentioned, noun phrases are highly effective in identifying relevant biomedical concepts, as well as supporting expressions. Given the addition of UMLS-*expand*, synonyms of the extracted noun phrases are also considered, further increasing the chance of retrieving relevant reference spans.

The limited performance of all methods in terms of the overall weighted F-1 and ROUGE-L scores is expected due to the difficulty of the task, as further corroborated by the low agreement between annotators. As previously stated, this makes the task particularly challenging for any system, as identifying the most appropriate reference spans is highly nontrivial even for domain experts. Nevertheless, while full agreement between domain experts is not present, as it is shown in table 1, more than 60% of the time, annotators agree – at least partially – on the position of the reference spans. This makes the task worth exploring.

5 Conclusion

In this paper, we propose an information retrieval approach for the problem of matching reference text spans with citances. Our approach takes advantage of several general and domain specific query reformulation techniques. Our best performing method obtains a significant increase over the baseline (25.9% F-1). However, as the absolute performance of the system indicates, the task of identifying matching reference spans to a given citance is highly non trivial. This fact is also reflected by the high disagreement between domain experts annotations and suggests that further exploration of the task is needed.

Acknowledgments

This work was partially supported by NSF (grant CNS-1204347).

References

Abu-Jbara, A. and Radev, D. (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.

Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Bos, L. et al. (2006). Snomed-ct: The advanced terminology and coding system for ehealth. *Stud Health Technol Inform*, 121:279–290.

Buckley, C., Singhal, A., Mitra, M., and Salton, G. (1995). New retrieval approaches using smart: Trec 4. In *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 25–48.

Chen, R., Lin, H., and Yang, Z. (2011). Passage retrieval based hidden knowledge discovery from biomedical literature. *Expert Systems with Applications*, 38(8):9958–9964.

Cheung, E. C., Athineos, D., Lee, P., Ridgway, R. A., Lambie, W., Nixon, C., Strathdee, D., Blyth, K., Sansom, O. J., and Vousden, K. H. (2013). Tigar is required for efficient intestinal regeneration and tumorigenesis. *Developmental cell*, 25(5):463–477.

Cohan, A., Soldaini, L., and Goharian, N. (2014). Towards citation-based summarization of biomedical literature. Proceedings of the Text Analysis Conference (TAC '14).

De Waard, A. and Maat, H. P. (2012). Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 47–55. Association for Computational Linguistics.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., and Radev, D. (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.

Hersh, W. R., Cohen, A. M., Roberts, P. M., and Rekapalli, H. K. (2006). Text retrieval conference 2006 genomics track overview. In *TREC*.

Hersh, W. R., Cohen, A. M., Ruslen, L., and Roberts, P. M. (2007). Text retrieval conference 2007 genomics track overview. In *TREC*.

Kaplan, D., Iida, R., and Tokunaga, T. (2009). Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 88–95. Association for Computational Linguistics.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Nakov, P. I., Schwartz, A. S., and Hearst, M. (2004). Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*, pages 81–88.

Okada, N., Yabuta, N., Suzuki, H., Aylon, Y., Oren, M., and Nojima, H. (2011). A novel chk1/2–lats2–14-3-3 signaling pathway regulates p-body formation in response to uv damage. *Journal of cell science*, 124(1):57–67.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM.

Qazvinian, V. and Radev, D. R. (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.

Qazvinian, V. and Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 555–564. Association for Computational Linguistics.

Robertson, S. and Zaragoza, H. (2009). *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Siddharthan, A. and Teufel, S. (2007). Whose idea was this, and why does it matter? attributing scientific work to citations. In *HLT-NAACL*, pages 316–323. Citeseer.

Urbain, J., Frieder, O., and Goharian, N. (2009). Passage relevance models for genomics search. *BMC bioinformatics*, 10(Suppl 3):S3.

Urbain, J., Goharian, N., and Frieder, O. (2008). Probabilistic passage models for semantic search of genomics literature. *Journal of the American Society for Information Science and Technology*, 59(12).

Voorhoeve, P. M., Le Sage, C., Schrier, M., Gillis, A. J., Stoop, H., Nagel, R., Liu, Y.-P., Van Duijse, J., Drost, J., Griekspoor, A., et al. (2006). A genetic screen implicates mirna-372 and mirna-373 as oncogenes in testicular germ cell tumors. *Cell*, 124(6):1169–1181.

Ying, H., Kimmelman, A. C., Lyssiotis, C. A., Hua, S., Chu, G. C., Fletcher-Sananikone, E., Locasale, J. W., Son, J., Zhang, H., Coloff, J. L., et al. (2012). Oncogenic kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. *Cell*, 149(3):656–670.