An Empirical Assessment of Citation Information in Scientific Summarization

Francesco Ronzano and Horacio Saggion

Natural Language Processing Group
Universitat Pompeu Fabra, Barcelona, Spain
{francesco.ronzano,horacio.saggion}@upf.edu

Abstract. Considering the recent substantial growth of the publication rate of scientific results, nowadays the availability of effective and automated techniques to summarize scientific articles is of utmost importance. In this paper we investigate if and how we can exploit the citations of an article in order to better identify its relevant excerpts. By relying on the BioSumm2014 dataset, we evaluate the variation in performance of extractive summarization approaches when we consider the citations to extend or select the contents of an article to summarize. We compute the maximum ROUGE-2 scores that can be obtained when we summarize a paper by considering its contents together with its citations. We show that the inclusion of citation-related information brings to the generation of better summaries.

Keywords: citation-based summarization, scientific text mining, summary evaluation

1 Introduction: citation-based summarization

A new scientific article is published every 20 seconds and such publication rate is going to increase during then next decade [1]. Considering this trend, everyone can agree that any person who has to deal with scientific publications, including researchers, editors, reviewers, is flooded by a huge amount of information. Keeping track of recent advances has become an extremely difficult and time consuming activity.

New and improved approaches to select and organize contents of interest from the huge amount of scientific papers that are accessible on-line could help to solve or at least reduce this problem. In this context, automated summarization techniques play a central role. By exploiting these techniques, it is possible to identify the set of excerpts of a text that better recap its contents, thus substantially reducing the effort needed to organize and evaluate scientific information. Current

This work is (partly) supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and by the European Project Dr. Inventor (FP7-ICT-2013.8.1 - Grant no: 611383).

approaches to summarize scientific literature usually extend existing document summarization techniques [2] by taking advantage of distinctive traits of scientific publications. Citations represent by far the feature of scientific articles that has been more often exploited in order to improve the quality of automatically generated summaries.

When a paper is cited, the part of the citing article that explains the reason of the citation is usually referred to as *citation context* or *citance* [3]. Each citance includes the sentence where the citation occurs (citing sentence) or part of it and possibly one or more surrounding sentences. To automatically identify the *citation context* different approaches have been proposed relying on the dependency tree of the citing sentence [6], on classifiers like Support Vector Machines or on sequence taggers, including Conditional Random Fields [7, 8].

Several analyses aim at characterizing what type of information about a paper is provided by its citances. [5] analyze a corpus of open access articles from PubMed Central¹ and notice that there is a small but quantifiable difference between the informative content of the abstract of a paper and the information provided by its set of citances. [9] prove the utility of citances to improve the quality of multi-document summaries: graph-based summarization methodologies are evaluated over two topic-homogeneous collections of papers from the ACL Anthology².

Distinct approaches have been proposed to exploit citances so as to select the set of clues or sentences that better summarize a paper. Starting from an article, [10] build a similarity graph to cluster sentences concerning the same topic and explore different selection strategies to choose from each cluster the ones to include in the summary [11] exploit the citances of a paper to build a language model that, in turn, is used to select the most influential sentences. [15] show that the citation information improves sensibly the quality of a summary by relying on different summarization approaches and exploiting the BioSumm2014 dataset. In [4] the contents of a paper are summarized by extracting relevant keyphrases (n-grams) from the collection of citances of the considered paper. A background language model is built by gathering several papers of the same domain of the one to summarize. The summary includes the sentences with the lowest keyphrase overlap with the background language model. [6] rely on the citances of an article to improve the diversity and readability of its automatically generated summary.

We investigate to what extent the citances of a paper are useful to create an improved summary of its contents. In particular we analyze how the contents of different parts of a paper, including abstract, body and citances, contribute to maximize a widespread summary evaluation metric, ROUGE-2. To this purpose we exploit the Biomedical Summarization Dataset (BioSumm2014), released in the context of the Biomedical Summarization Track of the Text Analytics Conference 2014³. The BioSumm2014 dataset consists of 20 collections of annotated

¹ <http://www.ncbi.nlm.nih.gov/pubmed>

² <https://aclweb.org/anthology/>

³ <http://www.nist.gov/tac/2014/BiomedSumm/>

papers, each one including a reference article and 10 citing articles. Section 2 provides a detailed description of the BioSumm2014 dataset and the citation-related manually annotated information that it includes. In Section 2 we also briefly describe the pre-processing steps we perform over the contents of the BioSumm2014 dataset to support further data analyses. In Section 3 we present and discuss how the maximum ROUGE-2 of a summary of a paper varies if we consider its citances to automatically build its summary. In Section 4 we provide our conclusions, outlining future venues of research.

2 The BioSumm2014 dataset

The BioSumm2014 dataset has been released in the context of the Biomedical Summarization Trak (Text Analytics Conference 2014) to provide a manually annotated corpus of scientific publications useful to investigate several aspects of citations. All the experiments presented in the following Sections of this paper rely on this dataset.

The BioSumm2014 dataset includes 20 collections of scientific publications dealing with the bio-molecular, pharmaceutical and medical domains, indexed by PubMed Central. Each collection consists of one cited paper (also referred to as Reference Paper) and 10 Citing Papers. Every Citing Paper of each collection cites one or more times the corresponding Reference Paper (see Figure 1). Each citation of each Citing Paper has been manually annotated by four annotators who were asked to identify:

- the *context of the citation* (i.e. citance), consisting of one to three text spans and including the related in-line citation marker ([1], (Rossi et al., 2010), etc.);
- the *citing spans*, that are one to three text spans in the Reference Paper that most accurately reflect the citance, thus representing the part of the Reference Paper that contains the reason why such paper has been cited by the Citing Paper.

In each Reference Paper we can identify the following parts (see Figure 1): the *abstract*, the *body* and, for each citation of a Citing Paper from the same collection, *four citing spans*, each one identified by a different annotator. On average, one third of the sentences the body of each Reference Paper intersects one or more citing spans. In addition, for each collection of papers, each annotator provided a summary of the Reference Paper having an approximate length of 250 words and including the opinions expressed by the related citations. In the BioSumm2014 dataset, for each citation of a Reference Paper, different citances have been identified by distinct annotators in the related Citing Paper. We associate to each citation, a *global citance span*. The global citance span represents the union of the citance text spans identified by the four human annotators for that particular citation. All the experiment we discuss in the next Section are based on the manual annotations of the citation contexts and the citing spans. The automatic identification of these text spans is out of the scope of the experiments presented in this paper.

Fig. 1. Structure of the BioSumm2014 dataset.

2.1 Pre-processing BioSumm2014 papers

To enable the execution of the experiments described in this paper, we imported the articles of the BioSumm2014 dataset together with their stand-off annotations into the text engineering framework GATE⁴. Then, we performed the following set of pre-processing steps on the textual contents of each article (both Reference and Citing Papers):

- **Custom sentence splitting:** we customized GATE regular-expression sentence-splitter in order to properly identify sentence boundaries in the scientific texts included in the BioSumm2014 dataset. We added rules to correctly deal with abbreviations that are characteristic of scientific literature (Fig., Sect., etc.);
- **Tokenization and POS-tagging:** we exploited GATE English Tokenizer and POS-tagger to perform these actions on the texts.
- **Sentence sanitization:** we filtered out incorrectly annotated sentences, relying on a set of rules and heuristics. For instance, we excluded sentences with no verbal tokens or short sentences, with less than 4 tokens (excluding punctuation). These sentences often represent headers or captions;
- **Sentence TF-IDF vector calculation:** we determined the TF-IDF vector associated to each sentence, where the IDF values are computed over all the papers of the related collection (10 Citing Papers and one Reference Paper). We removed stop-words to compute TF-IDF vectors.

⁴ <https://gate.ac.uk/>

3 ROUGE-2 maximization analysis

The ROUGE metrics [14] represent one of the most adopted sets of measures that aim at evaluating the quality of an automatically generated summary (also referred to as candidate summary) with respect to one or more reference summaries, usually created by humans. Summary evaluation results based on ROUGE metrics are proven to have a high correlation with human judgements. In our experiments we rely on ROUGE-2, one of the measures of the ROUGE family of metrics. ROUGE-2 quantifies the bi-gram recall between a candidate summary and a set of reference summaries: higher is the ROUGE-2 value, higher is the similarity of the candidate summary with the reference ones in terms of shared bi-grams.

In this Section we evaluate the contribution of different portions of BioSumm2014 papers (abstracts, body, citing spans, etc.) with respect to the variations of ROUGE-2. Specifically, we compute the maximum ROUGE-2 that can be obtained when we produce a summary of 250 words by picking sentences that belong to distinct portions of the papers. We truncate the contents of the summaries to evaluate to their 250th word.

Relying on the sentence detection output (see pre-processing subsection 2.1), we are able to divide the sentences of the papers of each collection into the following groups:

- **ABSTRACT**: sentences that belong to the abstract of the Reference Paper;
- **BODY**: sentences of the body of the Reference Paper;
- **CITING SPANS**: sentences of the body of the Reference Paper that intersect totally or partially a citing span (abstract sentences intersecting a citing span are not included in this group);
- **CITATION CONTEXT**: sentences that intersect at least one global citation span in a Citing Paper (i.e. sentences that belong to the context of the citations of the Reference Paper).

In each collection of papers, for each combination of the previous groups of sentences, we determine the subset of sentences (up to 250 words) that maximizes the ROUGE-2, by considering two types of reference summaries:

- **ABST_SUMM**: the abstract of the Reference Paper;
- **GS_SUMM**: the four 250-words summaries written by BioSumm2014 annotators.

In Tables 1 and 2 we show the average value (across the 20 collections) of the maximized ROUGE-2, when we generate a 250-words summary by choosing sentences from different groups.

3.1 ROUGE-2 maximization discussion

In both Tables 1 and 2, as expected, the best ROUGE-2 result is obtained when we consider all the sentences available (from the abstract, body and citation context in Table 1 and from the body and citation context in Table 2).

ABST.	ABST. BODY	ABST. CIT. SP.	ABST. CIT. CTX	ABST. BODY CIT. CTX	ABST. CIT. SP. CIT. CTX
0,126486	0,229957	0,228127	0,224505	0,245432	0,242601
BODY	BODY CIT. CTX	CIT. SP.	CIT. SP. CIT. CTX	CIT. CTX	
0,211558	0,234545	0,230904	0,204874	0,167024	

Table 1. Average max ROUGE-2 across collections with respect to GS-SUMM when the sentences to include in the 250-words generated summary are chosen by combining the following groups: abstract (ABST.), body (BODY), citing spans (CIT. SP.) and citation context (CIT. CTX). The best ROUGE-2 is highlighted in bold.

BODY	BODY CIT. CTX	CIT. SP.	CIT. SP. CIT. CTX	CIT. CTX
0,305349	0,317036	0,290046	0,275215	0,141528

Table 2. Average max ROUGE-2 across collections with respect to ABST-SUMM (abstract as reference summary) when the sentences to include in the 250-words generated summary are chosen by combining the following groups: body (BODY), citing spans (CIT. SP.) and citation context (CIT. CTX). The best ROUGE-2 is highlighted in bold.

When we refer to human generated reference summaries (Table 1), we can notice that if we consider only the sentences from the abstract or the sentences from the citation context, lower quality extractive summaries are generated with respect to other combinations of groups of sentences. As a consequence the contents of the abstract or the contents of the citation context alone are not enough for a good quality summarization.

If we add the sentences of the citation contexts to the sentences of the Reference Paper (abstract and body), we can observe a sensible improvement of the average maximum ROUGE-2, from 0,229957 to 0,245432. We can observe the same trend if we add the sentences of the citation context to the sentences of the body (from 0,211558 to 0,234545). Both trends show that by considering the sentences of the citation context of a paper, we can potentially build a summary with a higher ROUGE-2 score.

The sentences of the body of each Reference Paper that belong to a citing span represent about one third of the whole set of sentences of each Reference Paper body. The summary generated thanks to these citing span sentences has a quality comparable to the summary generated by considering all the sentences of the body. As a consequence, we can state that without a considerable loss of summarization quality, we can select from the sentences of the body of a Reference Paper, the ones that most accurately reflect the contents of its citation contexts and then generate an extractive summary from this subset of sentences. This selection would act as a filter, thus considerably reducing the number of

candidate sentence to evaluate when generating a summary of the Reference Paper.

When our reference summary is the abstract of the Reference Paper (Table 2), we can note that the sentences of the citation contexts alone do not constitute the best choice to summarize the paper (citation context column, avg. max ROUGE-2 equal to 0,141528). Moreover, as previously observed with human generated reference summaries, if we add to the sentences of the body of the paper, the sentences of the citation context, the average maximum ROUGE-2 improves (from 0,305349 - only body - to 0,317036 - body and citation context). These fact confirms the usefulness of citation context sentences to generate better summaries of a paper.

4 Conclusions and future work

Nowadays, the growing publication rate of scientific results is more and more stressing the importance of effective approaches to easily structure and summarize the contents of articles. We have investigated how the citations of an article can contribute to generate a better summary. To this purpose we relied on the BioSumm2014 dataset that includes 20 manually annotated collections, each one made of one cited paper and 10 other papers that cite the first one. We evaluated how different parts of each cited paper (abstract, body, citing spans) together with its citation context contribute to the generation of a summary. We adopted the following approach: by considering any possible combination of parts of the cited paper and its citation context, we computed the maximum ROUGE-2 score achievable. As a result, we noted that the maximum ROUGE-2 score achievable sensibly increases when we exploit the sentences from the citation context to generate a summary by considering as reference summaries both human generated summaries and the abstract.

In general, we have to notice that when we want to generate a citation-based summary of a paper, we need to gather and process the set of articles that cite the paper. Currently this process requires a considerable effort and the contents of all the citing articles are not always available. As a consequence, besides the improvement of the quality of the summary, when we consider citation-based summarization we have also to take into account the efforts needed in order to collect the citation related information concerning a paper.

As venues for future research we would like to validate the set of experiments presented in this paper with scientific publications from different domains. It would be interesting to perform a detailed evaluation of the consistency and diversity of the information included in automatically generated summaries, since the contents of a paper are often repeated or paraphrased by its citation contexts. Moreover, we would like to investigate into details to what extent different summarization approaches in different domains generate better summaries when we consider citation contexts as well as other kinds of information from citing papers. We would also perform more extensive analyses by contemplating, besides the ROUGE-2 score, alternative summary evaluation metrics.

References

1. AAAS Science Magazine: Communication in Science Pressures and Predators, volume 1. Science, 4 October 2013. Vol. 342 no. 6154 pp. 58-59 DOI: 10.1126/science.342.6154.58.
2. Gupta, V., & Lehal, G. S.: A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence, 2(3), 258-268 (2010)
3. Nakov, P. I., Schwartz, A. S., & Hearst, M.: Citances: Citation sentences for semantic analysis of bioscience text. In Proceedings of the SIGIR04 workshop on Search and Discovery in Bioinformatics, pp. 81-88 (2004)
4. Qazvinian, V., Radev, D. R., & zgr, A.: Citation summarization through keyphrase extraction. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 895-903). Association for Computational Linguistics (2010)
5. El Kiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D.: 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 (2008)
6. Abu-Jbara, A., & Radev, D.: 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 (pp. 500-509). Association for Computational Linguistics (2011)
7. Abu-Jbara, A., & Radev, D.: Reference scope identification in citing sentences. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 80-90). Association for Computational Linguistics (2012)
8. Qazvinian, V., & Radev, D. R.: Identifying non-explicit citing sentences for citation-based summarization. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 555-564). Association for Computational Linguistics (2010)
9. Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishnan, P., Qazvinian, V., Radev D., & Zajic, D.: Using citations to generate surveys of scientific paradigms. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 584-592). Association for Computational Linguistics (2009)
10. Qazvinian, V., Radev, D. R., Mohammad, S., Dorr, B. J., Zajic, D. M., Whidby, M., & Moon, T.: Generating Extractive Summaries of Scientific Paradigms. J. Artif. Intell. Res.(JAIR), 46, 165-201 (2013)
11. Mei, Q., & Zhai, C.: Generating Impact-Based Summaries for Scientific Literature. In ACL, Vol. 8, pp. 816-824 (2008)
12. Erkan, G., & Radev, D. R.: LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 457-479 (2004)
13. Pennington, J., Socher, R., & Manning, C. D.: Glove: Global vectors for word representation. Proceedings of the Empirical Methods in Natural Language Processing, 12, 1532-1543 (2014)
14. Lin, C. Y.: Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, Vol. 8 (2004)
15. Cohan, A., & Goharian, N.: Scientific Article Summarization Using Citation-Context and Articles Discourse Structure. Empirical Methods in Natural Language Processing (2015).