An initial Analysis of Topic-based Similarity among Scientific Documents based on their Rhetorical Discourse Parts

Carlos Badenes-Olmedo[1], José Luis Redondo-García[2], and Oscar Corcho[1]

[1] Universidad Politécnica de Madrid, Ontology Engineering Group, Spain
{cbadenes, ocorcho}@fi.upm.es
[2] Amazon Research, Cambridge UK
jluisred@amazon.com

Abstract. Summaries and abstracts of research papers have been traditionally used for many purposes by scientists, research practitioners, editors, programme committee members or reviewers (e.g. to identify relevant papers to read or publish, cite them, explore new fields and disciplines). As a result, many paper repositories only store or expose abstracts, what may limit the capacity of finding the right paper for a specific research purpose.

Given the size limitations and the concise nature of abstracts, they usually omit explicit references to some contributions and impacts of the paper. Therefore for certain information retrieval tasks they cannot be considered as the most appropriate excerpt of the paper to base these operations on. In this paper we have studied other kinds of summaries, built upon textual fragments falling under certain categories of the scientific discourse, such as *outcome*, *background*, *approach*, etc, in order to decide which one is more appropriate in order to substitute the original text. In particular, two novel measures are proposed: (1) *internal-representativeness*, which evaluates how well a summary describes what the full-text is about and (2) *external-representativeness*, which evaluates the potential of a summary to discover related texts.

Results suggest that summaries explaining the method of a scientific article express a more accurate description of the full-content than others. In addition, more relevant related articles are also discovered from summaries describing the method, together with those containing the background knowledge or the outcomes of the research paper.

1 Introduction

In this paper we present our first steps on the analysis of the quality of research article summaries. Our goal is to find the strengths and weaknesses of approaches leveraging exclusively on abstracts against those based on scientific discourse categories such as *approach*, *challenge*, *background*, *outcomes* and *future work*. Since the main contributions and impacts of a research article are not always included explicitly in the abstract, as in the case of [9] describing an architecture, where details about the model architectures are missing, they cannot always be considered as the most adequate scientific summary of a research paper. In order to judge on this accuracy, two novel measures are proposed based on the capability of the summary to substitute the original paper: (1) *internal-representativeness*, which evaluates how well the summary represents the original full-text and (2) *external-representativeness*, which evaluates the summary according to how the summary is able to produce a set of related texts that are similar to what the original full-text has triggered.

The paper is organized as follows: Section 2 highlights recent studies on text mining research articles and presents the steps followed to measure the *representativeness* of abstracts and research article summaries based on rhetorical categories. It describes both the classifier used to identify those categories in papers and the representational model and similarity metric used to compare textual units. Experimental results comparing the different kind of summaries are shown in Section 3. Finally, Section 4 presents conclusions from the experiments.

2 Background and Approach

Recent studies [16][13] have shown that text mining of full research articles give consistently better results than using only their corresponding abstracts. Given the size limitations and concise nature of abstracts, they often omit descriptions or results that are considered to be less relevant but still are important in certain Information Retrieval (IR) tasks. Thus, when other researchers cite a particular paper, 20% of the keywords that they mention are not present in the abstract [6].

In this paper, we show our initial analysis about the *representativeness* of research article summaries, considering those based exclusively on abstracts and those based on their discursive structure (*approach*, *challenge*, *background*, *outcomes* and *future work*)[14]. The *representativeness* of a summary with respect to the original full-text is defined as the degree of relation with the original one (*internal-representativeness*), along with the capacity of mimicking the full text when finding related items (*external-representativeness*). In order to quantify this notions of internal-external representativeness, a probabilistic topic model is trained over the entire set of papers to have a vectorial representation of each text retrieved from a paper: full content-based and summary-based. The vectorial representations of full-papers is used to measure the distance between them and those derived from abstract or summaries (*internal-representativeness*), and also to find similar documents (*external- representativeness*) based on the distance between their vectorial representations. An upper distance threshold is specified to filter less similar pairs and compose a set of related papers for each paper. Then, a comparison in terms of *precision* and *recall* is performed between sets obtained by only using the vectorial representation of full-papers, against sets produced by using other kind of summaries.

2.1 Annotation of Rhetorical Discourse Parts

First of all, we need to identify the rhetorical parts of a research paper. Some approaches have been proposed to summarize scientific articles [4] taking advantage of the citation context and the document discourse model. We have used the scientific discourse annotator proposed by [11] to automatically create summaries from scientific articles by classifying each sentence as belonging to one of the following scientific discourse categories: *approach*, *challenge*, *background*, *outcomes* and *future work*. These categories were identified from the schemata proposed by [15] with the original purpose of characterizing the content of Computer Graphics papers. The annotator is based on a Support Vector Machine classifier that combines both lexical and syntactic features to model each sentence in a paper. This tool[3] was integrated in the *librAIry* [1] *Rhetoric Module*[4] to automatically annotate research papers with their rhetorical content.

2.2 Representational Model

A representational model is required not only to measure distances between text fragments but, more importantly, to help to understand the differences in their content. Topic models are widely used to uncover the latent semantic structure from text corpora. In particular, Probabilistic Topic Models represent documents as a mixture of topics, where topics are probability distributions over words. Latent Dirichlet Allocation (LDA)[2] is the simplest *generative* topic model that adds Dirichlet priors for the document-specific topic mixtures, making it possible to characterize documents not previously used during the training task. This is a key feature for our evaluations because, although the model used for the experiments will be trained from the full-content of papers, it will be also used to describe the texts summaries.

Thus, we have used a LDA model to describe the inherent topic distribution of papers in the corpus. Some hyper-parameters need to be estimated: the *number of topics (k)*, the concentration parameter (α) for the prior placed on documents' distributions over topics and the concentration parameter (β) for the prior placed on topics distributions over terms. Since the target of this experiment is not to evaluate the quality of the representational model, but to compare their topic distributions, we accepted as valid values those widely used in the literature: $\alpha = 0.1$, $\beta = 0.1$, and $k = 2 * \sqrt{n/2} = 44$ where n is the size of the corpus.

Similarity Measure Feature vectors in Topic Models are topic distributions expressed as vectors of probabilities. Hence we opt for *Jensen-Shannon divergence* (JSD)[8] instead of the commonly used *Kullback-Liebler divergence* (KLD). The reason for this is that KLD (1) is not defined when a topic distribution is zero and (2) is not symmetric, what does not fit well with semantic similarity

[3] http://backingdata.org/dri/library/

[4] https://github.com/librairy/annotator-rhetoric

measures which in general are symmetric [12]. JSD considers the average of the distributions as follows :

$$JSD(p,q) = \sum_{i=1}^{T} p_i * \log \frac{2 * p_i}{p_i + q_i} + \sum_{i=1}^{T} q_i * \log \frac{2 * q_i}{q_i + p_i} \tag{1}$$

where T is the number of topics and p, q are the topics distributions.

And the *similarity measure* used in our analysis is based on the JSD transformed into a similarity measure as follows [5] :

$$similarity(D_i, D_j) = 10^{-JSD(p,q)} \tag{2}$$

where D_i, D_j are the documents and p, q the topics distributions of each of them.

3 Experiments

The corpus used in the experiments was created by combining journals in different scientific domains such as *Advances in Space Research*, *Procedia Chemistry*, *Journal of Pharmaceutical Analysis* and *Journal of Web Semantics*. In total 1,000 papers were added, 250 from each journal. Both the abstract and the *full-content* of these documents were directly retrieved from the Elsevier API [5] by using the *librAIry* [1] *Harvester module* [6]. The code used to perform the analysis along with the results obtained are available in GitHub[7].

Since the annotation process to automatically discovers the rhetorical parts of a research paper (Section 2.1) is sensitive to the structure of the phrases that are used when writing the text, only 20% of papers in the corpus could be fully annotated with all the fragments considered. In fact, these categories are not present in the same proportion in the corpus: *approach* (90%), *background* (78%), *outcome* (73%), *challenge* (57%) and *future work* (21%)

3.1 Internal Representativeness

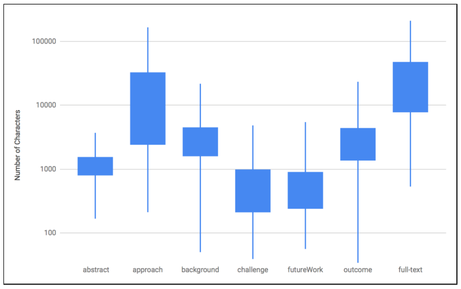
The *internal-representativeness* of a summary measures the similarity of this summary against the original full-text research paper. This similarity is based on the JSD between the topic distribution of each of them.

Since LDA considers documents as *bag-of-words*, the text length (e.g. full-content or summaries) affects the accuracy of the topic distributions inferred by the topic model described in Section 2.2. The occurrences of words in short texts are less discriminative than in long texts where the model has more word counts to know how words are related [7]. In view of the above, the *approach*, the *background* and the *outcome* content of a paper generate more accurate topic distributions than those created from other approaches such as the abstract.

[5] https://dev.elsevier.com

[6] https://github.com/librairy/harvester-elsevier

[7] https://github.com/librairy/study-semantic-similarity



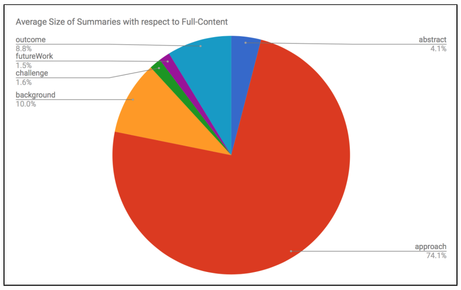


Fig. 1. length of summaries

Fig. 2. relative size of parts of an article

Also, the relative presence of each of them in a paper (figure 2) shows an unexpected result when compared to the IMRaD format [10]. This style proposes to distribute the content of an abstract, and by extension the full-paper, as follows: *Introduction*(25%), *Methods*(25%), *Results*(35%) and *Discussion*(15%). However, the results (figure 2) show that *Method* section (*approach* content) is more extensive than *Results* section (*outcome* content) in our corpus.

All pairwise similarities between full-papers, abstracts and rhetorical-based summaries are calculated to measure the ***internal-representativeness*** of a summary with respect to the original text, i.e. the topic-based similarity value (equation 2) between the probability distributions of the full-text and each of the summaries. Results (table 1) suggest than summaries created from the *approach* content are more representative than others, i.e. the distribution of topics describing the text created from the *approach* content is the most similar to the one corresponding to the full-content of the paper.

	Min	Lower Quartile	Upper Quartile	Max	Dev	Median
abstract	0.0489	0.9109	0.9840	1.0000	0.1443	0.9741
approach	**0.0499**	**0.9969**	**1.0000**	**1.0000**	**0.0872**	**0.9998**
background	0.0463	0.8967	0.9937	0.9988	0.2037	0.9822
challenge	0.0426	0.7503	0.9517	0.9940	0.2224	0.8829
futureWork	0.0000	0.6003	0.9435	0.9948	0.2842	0.8814
outcome	0.0485	0.9267	0.9925	0.9990	0.1721	0.9835

Table 1. Internal-Representativeness

3.2 External-Representativeness

The *external-representativeness* metric tries to measure how different is the set of related documents obtained from summaries with respect to those derived from the original full-text. In terms of *precision*, *recall* and *f-measure*, a comparison has been performed to analyze the behavior of the summaries when trying to discover related content compared to use the full-text of the article.

By using the same topic model previously created, similarities among all pairs of documents were also calculated according to equation 2. Then, a minimum score or similarity threshold is required to define when a pair of papers are related. Each threshold is used to create a gold-standard which relates articles to others based on their similarity values. In order to discover that lower bound of similarity, a study about trends in the similarity scores (fig 3) as well as distributions of topics in the corpus (fig 4) was performed. We can see that topics are not equally balanced across papers. This fact generates separated groups of strongly related papers. We think this phenomena is due to our usage of a corpus created from journals where different domains are equally balanced. Then, we considered a similarity score equals to 0.99 (fig 3) as the threshold from which strong relations appear. However, to cover different interpretations of similarity, from those based on sharing general ideas or themes to those that imply to share a more specific content, the following list of thresholds was considered in the experiments: 0.5, 0.6, 0.7, 0.8, 0.9, 0.95 and 0.99.

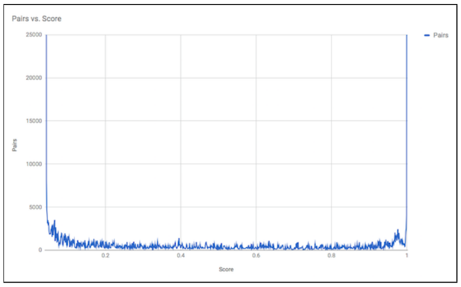


Fig. 3. number of pairwises by similarity score (rounded up to two decimals)

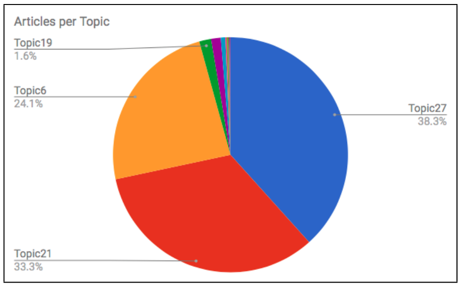
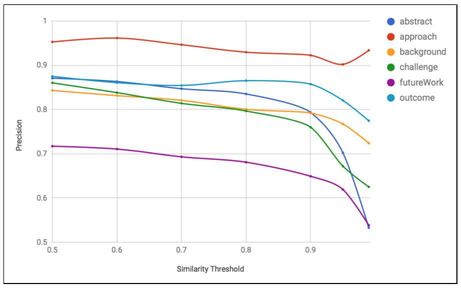


Fig. 4. topics per article with value above 0.5

For each similarity threshold, a gold-standard was created based on considering as related those papers with a similarity value upper than the selected threshold. Results (figure 5) comparing the related papers inferred from the full-content with those inferred from the partial-content representation (i.e. abstract or rhetorical parts) suggest that strongly related papers are mainly discovered by using the summary created from the *approach* section. The reason for this may be based on the average size of this type of summaries or the particular content included in this part of a paper. While other summaries include more general-domain words, the *approach* content includes more specific words that describe the method or the final objective of the paper. So, for higher similarity thresholds, i.e. for strongly related papers, the recommendations discovered by using the *approach* are more precise than those discovered by using the abstract.

In terms of *recall* (figure 6), the upward trend followed by the *approach*, the *outcome* and the *background* content remarks the assumption of summaries containing key words allow to discover more similar papers than others. Moreover,



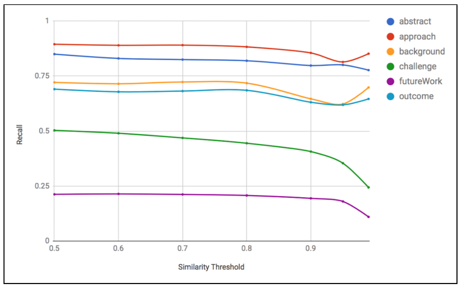
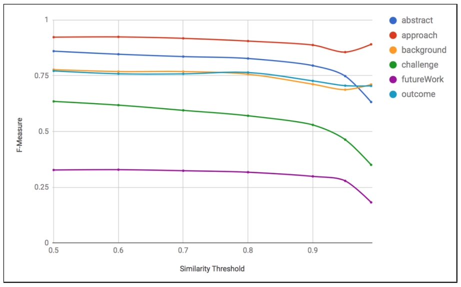


Fig. 5. P at different similarity thresholds **Fig. 6.** R at different similarity thresholds

since *recall* overlooks false-negatives classifications, it suggests that these parts of a research paper share more words than others with strongly related papers but they may also present commonalities with highly related papers, except in case of *approach* which still exhibits higher *precision*.

As expected, only summaries created from the *approach*, the *outcome* and the *background* content maintain high accuracy values (fig 7) even for high similarity thresholds. Along with the results showed in figure 8, where the same three rhetorical classes present the lowest standard deviation over the *f-measure*, they can be considered as the most robust summaries containing the ideas that better characterize the paper compared to others.



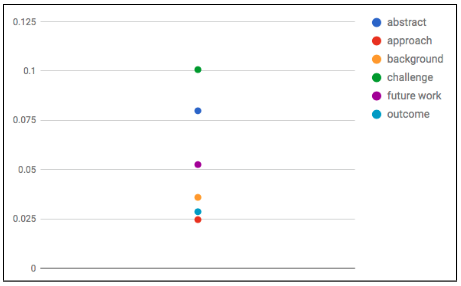


Fig. 7. f-measure VS similarity thresholds **Fig. 8.** σ of the f-measure

4 Conclusions and Future Work

We have studied the Topic-based similarities among scientific documents based on their abstract sections with respect to summaries corresponding to their scientific discourse categories. For this purpose, two novel measures have been proposed: (1) *internal-representativeness* and (2) *external-representativeness*.

Results show that summaries created from the *approach*, *outcome* or *background* content of a paper describe more accurately its full-content in terms of overall ideas and related documents than abstracts. Although those summaries are more extensive in number of characters than other with similar *precision* such as the abstract content, they have proven to be particularly helpful discovering strongly related papers, i.e. papers with a similarity value close to 1.0.

In order to avoid an influence of the size of the summaries on the accuracy of the results, in future work we plan to use probabilistic topic model algorithms oriented to handle short-texts such as BTM [3] to describe texts.

References

1. Badenes-Olmedo, C., Redondo-Garcia, J.L., Corcho, O.: Distributing Text Mining tasks with librAIry. In: Proceedings of the 17th ACM Symposium on Document Engineering (DocEng) (2017)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3(4-5), 993–1022 (2003)
3. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: Topic Modeling over Short Texts. Knowledge and Data Engineering, IEEE Transactions on PP(99), 1 (2014)
4. Cohan, A., Goharian, N.: Scientific Article Summarization Using Citation-Context and Article's Discourse Structure. In: Conference on Empirical Methods in Natural Language Processing. pp. 390–400 (2015)
5. Dagan, I., Lee, L., Pereira, F.C.N.: Similarity-Based Models of Word Cooccurrence Probabilities. Machine Learning 34(1-3), 43–69 (1999)
6. Divoli, A., Nakov, P., Hearst, M.A.: Do peers see more in a paper than its authors? Advances in Bioinformatics (2012)
7. Hong, L., Davison, B.D.: Empirical study of topic modeling in Twitter. Proceedings of the First Workshop on Social Media Analytics - SOMA '10 pp. 80–88 (2010)
8. Lin, J.: Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory 37(1), 145–151 (1991)
9. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR 2013) pp. 1–12 (2013)
10. Nair, P.R., Nair, V.D.: Organization of a Research Paper: The IMRAD Format. In: Scientific Writing and Communication in Agriculture and Natural Resources, p. 150 (2014)
11. Ronzano, F., Saggion, H.: Dr. Inventor Framework: Extracting Structured Information from Scientific Publications. In: Discovery Science: 18th International Conference. pp. 209–220 (2015)
12. Rus, V., Niraula, N., Banjade, R.: Similarity Measures Based on Latent Dirichlet Allocation. In: Computational Linguistics and Intelligent Text Processing, pp. 459–470 (2013)
13. Sciences, E.R.S.f.P., life: Harnessing the power of content - Extracting value from scientific literature: the power of mining full-text articles for pathway analysis Harnessing the Power of content (2016)
14. Simone Teufel: The Structure of Scientific Articles - Applications to Citation Indexing and Summarization. In: CSLI Studies in Computational Linguistics (2010)
15. Teufel, S., Siddharthan, A., Batchelor, C.: Towards discipline-independent Argumentative Zoning: Evidence from chemistry and computational linguistics. Conference on Empirical Methods in Natural Language Processing pp. 1493–1502 (2009)
16. Westergaard, D., Stærfeldt, H.h., Tønsberg, C., Jensen, L.J., Brunak, S.: Text mining of 15 million full-text scientific articles. bioRxiv (2017)