Evaluation of Scientific Elements for Text Similarity in Biomedical Publications

Mariana Neves, Daniel Butzke, Barbara Grune
German Centre for the Protection of Laboratory Animals (Bf3R),
German Federal Institute for Risk Assessment (BfR)
Diedersdorfer Weg 1, 12277, Berlin, Germany
mariana.lara-neves@bfr.bund.de

Abstract

Rhetorical elements from scientific publications provide a more structured view of the document and allow algorithms to focus on particular parts of the text. We surveyed the literature for previously proposed schemes for rhetorical elements and present an overview of its current state of the art. We also searched for available tools using these schemes and applied four tools for our particular task of ranking biomedical abstracts based on text similarity. Comparison of the tools with two strong baselines shows that the predictions provided by the ArguminSci tool can support our use case of mining alternative methods for animal experiments.

1 Introduction

We aim to mine alternative methods to animal experiments from the biomedical literature. These are methods that address any of the so-called 3R principles of replacement (no animals at all or use of invertebrates over vertebrates), reduce (use of less animals), or refinement (cause less harm to animals) (Gruber and Hartung, 2004; Doke and Dhawale, 2015). For such complex natural language processing (NLP) applications, it is necessary to rely on appropriate tools to precisely understand the text and better find the potential relevant documents. The rhetorical elements, such as zones or particular entities, can support NLP algorithms by focusing on the relevant elements of the text (Mann and Thompson, 1987).

Given a certain document that describes an animal experiment for a certain research goal, hereafter called input document, we would like to find potential publications, hereafter called candidate documents, that describe an alternative method for the same research goal. Thus, some of the scientific elements should be similar between input and candidate documents, e.g. research goals and outcomes, while some others should be different, e.g. methods. Finding an alternative method to animal experiment requires two tasks: (a) performing a text similarity task with respect to some aspects of the publication, and (b) precisely understanding the proposed method with respect to the 3R principles. Therefore, the extraction of rhetorical elements has the potential to boost performance for these tasks.

Previous works have proposed many schemes for rhetorical elements in scientific publication, as reviewed in Webber et al. (2012). In a more recent survey, Nasar et al. (2018) present a good overview on both metadata and schemes for scientific articles. On the one hand, many of these schemes are not supported by an annotated corpus for training suitable information extraction tools. On the other hand, some tools based on these schemes are readily available for use.

We surveyed published schemes for rhetorical elements, whether focused on the biomedical domain or not, and we present a short overview on these. For those schemes for which we could find available tools, the latter was used to process a collection of 562 biomedical abstracts. We performed a comparison of the output (rhetorical elements) from the tools in the scope of a text similarity task on a manually annotated dataset. In this work, we limited our evaluation for text similarity but did not address whether the proposed methods comply with the 3R principles.

In summary, the contributions of this work are the following: (a) a short survey on existing schemes and corpora for rhetorical elements in scientific publications; (b) the identification of the schemes for which available tools are readily available for use; and (c) the evaluation of the available tools on a biomedical use case for text similarity. The next section presents a survey on

Proceedings of the 6th Workshop on Argument Mining, pages 124–135
Florence, Italy, August 1, 2019.

the available schemes, followed by the methodology that we propose to compare the tools in the scope of text similarity. We present the results in Section 4 and our discussion in Section 5.

2 Schemes for Rhetorical Elements

We classified the schemes according to the annotation level they address, either on the sentence, entity or relation-level. We present a summary of all schemes that we found, but give a more detailed description for (selected) schemes for which an annotated corpus is available (cf. Table 1).

2.1 Sentence-level Schemes

Many schemes model scientific elements on the level of sentences or phrases, i.e., for document zoning. It consists of splitting the publications (whether abstracts or full texts) on zones according to its scientific content, e.g. introduction, methods, results. Shimbo et al. (2003) proposed five categories and used structured abstracts from Medline while Hirohata et al. (2008) suggested four zoning categories. Further, Mullen et al. (2005) proposed a schema in which labels are grouped in three groups. Agarwal and Yu (2009) defined four categories (IMRAD schema) and manually annotated 148 articles, which was also used by Varga et al. (2012) for the annotation of more than 1,000 biomedical articles. Ruch et al. (2007) also annotated and tried machine learning in biomedical abstracts. However, none of the above data seems to be available for use, but we found many schemes with available corpora:

AZ (Teufel and Moens, 2002). The Argumentative Zoning (AZ) schema was first proposed by Teufel and Moens (2002) and an annotated corpus is freely available for download[1]. The schema is composed of seven rhetorical categories and the corresponding corpus contains 80 articles on computational linguistics. Teufel et al. (2009) extended the schema to 11 categories (the AZ-II schema), applied it to chemistry papers, and later compared it to the CoreSC schema (Liakata et al., 2010).[2] Later, Kovačević et al. (2012) annotated 110 articles in computational linguistics with a modified version of the AZ labels. Mizuta et al. (2006) also adapted the AZ schema to biomedicine by annotating 20 full-text articles. Guo et al. (2010) compared three zoning schemes in abstracts, including a reduced version of the AZ schema composed of seven categories, and annotated 1,000 abstracts with these schemes.[3]

CoreSC (Liakata et al., 2010). This schema consists of three layers of labels and the corresponding ART corpus[4] is composed of 225 full texts. The corpus and schema were used in Guo et al. (2010) (just the first layer) and in Liakata et al. (2012a) for two life sciences applications, while Liakata et al. (2012b) compared it to a schema for biomedical events and developed the the SAPIENTA software[5].

Dr. Inventor (Ronzano and Saggion, 2015; Fisas et al., 2015). The Dr. Inventor Framework proposes five categories and annotated 40 Computer Graphics papers, the so-called Dr. Inventor Rhetorically Annotated Corpus. Later, they also annotated another layer for citation purposes (Fisas et al., 2016). An extension of this schema with argumentative components and relations was recently published (Lauscher et al., 2018b), along with a tool for the prediction of the scientific elements (Lauscher et al., 2018a).

MAZEA (Dayrell et al., 2012). This schema considers six categories and the corpus was annotated for 645 abstracts from Physical Sciences and Engineering and Life and Health Sciences.[6] A Web application is available for tagging abstracts.

PIBOSO (Kim et al., 2011). It was designed for the clinical domain and proposes six categories of a modified version of the PICO criteria. It was used for the ALTA-NICTA shared task[7] and recent works using this corpus include Hassanzadeh et al. (2014) and Jin and Szolovits (2018). The latter relies on deep learning methods and the implementation is readily available.

PubMed RCT (Dernoncourt and Lee, 2017). It is a collection that includes two corpora of 20,000 and 200,000 medical abstracts annotated

[1] https://www.cl.cam.ac.uk/~sht25/AZ_corpus.html

[2] However, the AZ-II corpus was not found.

[3] However, the URL informed in a later publication (Guo et al., 2013) no longer exists.

[4] https://www.aber.ac.uk/en/cs/research/cb/projects/art/art-corpus/

[5] http://www.sapientaproject.com/software

[6] http://www.nilc.icmc.usp.br/mazea-web/downloads.php

[7] https://www.kaggle.com/c/alta-nicta-challenge2

	Tools	Categories	Corpora	Topic
Sentence/Phrase	AZ	AIM, TEXTUAL, OWN, BACKGROUND, CONTRAST, BASIC, OTHER	80 (Teufel and Moens, 2002) and 20 (Mizuta et al., 2006)	CL, bio
	CoreSC	[Level 1] Hypothesis, Motivation, Background, Goal, Object, Method, Experiment, Model, Observation, Result, Conclusion	225 (Liakata et al., 2010)	chem
	Dr. Inventor	Approach, Challenge, Background, Outcomes, Future Work	40 (Ronzano and Saggion, 2015)	CG
	MAZEA	background, gap, purpose, method, result, conclusion	645 abstracts (Dayrell et al., 2012)	phy, eng, LS
	PIBOSO	Population, Intervention, Background, Outcome, Study Design, Other	1,000 abstracts (Kim et al., 2011)	bio
	PubMedRCT	background, objective, method, result, conclusion	20,000 and 200,000 abstracts (Dernoncourt and Lee, 2017)	bio
	Wilbur	FOCUS, POLARITY, CERTAINTY, EVIDENCE, DIRECTIONALITY	10,000 sentences (Shatkay et al., 2008)	bio
Ent.	ScienceIE	Task, Process, Material	500 (Augenstein et al., 2017)	CS
Relation	Gábor	USAGE, RESULT, MODEL, PART_WHOLE, TOPIC, COMPARISON	500 abstracts (Gábor et al., 2018)	CL
	SciDTB	[Coarse level] Attribution, Background, Cause-effect, Comparison, Condition, Contrast, Elaboration, Enablement, Evaluation, Explain, Joint, Manner-means, Progression, Same-unit, Summary, Temporal	798 abstracts (Yang and Li, 2018)	CL
Hybrid	Green	[Levels 1-3] 1. Causation, 1.1 One Group, 1.1.1 Agreement Arguments, 1.1.2 Eliminate Candidates, 1.1.3 Explanation-Based, 1.2 Two Group, 1.2.1 Difference, 1.2.2 Analogy (Causal), 1.2.3 Explanation-Based, 2. Other, 2.1 Classification, 2.2 Confirmation	one (Green, 2018)	bio

Table 1: Summary of the selected schemes and corresponding categories, size of the annotated corpora, and topic of the latter. Only the categories from the certain levels were shown for some schemes with various layers. Numbers or the corpora refer to full-text documents, unless otherwise stated. Regarding the topics, "CL" stands for computational linguistics, "bio" for biomedicine, "chem" for chemistry, "CG" for Computer Graphics, "phy" for Physics, "eng" for Engineering, "LS" for Life Sciences, and "CS" for Computer Science.

with five categories. The corpus is freely available[8] as well as at least two tools for its detection, namely the one from Jin and Szolovits (2018) (cf. PIBOSO above) and one based on AllenNLP (Achakulvisut et al., 2018).

Wilbur (Wilbur et al., 2006). It consists of a schema developed for biomedical articles on five dimensions. Later, the authors annotated 10,000 sentences from full-text publications (Shatkay et al., 2008), which was made available after a detailed analysis (Rzhetsky et al., 2009).[9] The annotation are on the level of fragments, which usually correspond to either the sentences or phrases.

2.2 Entity-level Schemes

Entity-level schemes aim at annotating the elements on the level of entities. Gupta and Manning (2011) proposed a simple schema based on three concepts and labeled 474 abstracts of computational linguistics. More recently, Jung (2017) defined five entity types and annotated 1,000 articles about information and communication technology (ICT) and chemical engineering. Blake (2010) also proposed a schema based on various levels of evidence (implicit and explicit claims) and annotated 29 full-text biomedical articles. However, none of the above data seems to be available but we found one schema with annotated corpus:

ScienceIE (Augenstein et al., 2017). This schema proposes three elements on the entity level as well as the annotation of keyphrases. The corpus contains 500 articles about Computer Science, Material Sciences and Physics, which were split into training, development and test datasets and used for the a SemEval task in 2017. We found the implementation from two of the participants on the shared task, namely (Prasad and Kan, 2017) and (Eger et al., 2017).

2.3 Relation-level Schemes

Previous work also considered schemes that consider relations between scientific elements. Prasad et al. (2011) defined eight discourse relations in the Biomedical Discourse Relation Bank (BioDRB) and annotated 24 articles from the GENIA corpus, which was later used in a couple of works (Ramesh and Yu, 2010; Polepalli Ramesh et al., 2012). Tateisi et al. (2013) defined 16 relations and annotated 30 articles, while Meyers et al. (2014) proposed five relations and sub-relations with which they annotated 200 biomedical articles. However, none of the data above seems to be available, but we found corpora for the following two schemes:

Gábor (Gábor et al., 2016) It is a schema in the form of an ontology of 18 relations for the scientific literature, besides three more general relations. Six of these relations were recently addressed in the SemEval'18 Task 7, for which annotated data is available (Gábor et al., 2018). For sub-task 2 in SemEval'18 Task 7, the code from the team that obtained the best scores in this task is available (Luan et al., 2018).

SciDTB (Yang and Li, 2018). It is a discourse treebank for scientific articles that includes 17 coarse-grained and 26 fine-grained relation types. They annotated 798 abstracts from the ACL Anthology that are available for download.[10]

2.4 Hybrid Schemes

Hybrid schemes contain labels which cover more than one of the levels above. Tateisi et al. (2016) created an ontology of entities and relations and annotated 400 abstracts about computational linguistic. However, we found only one hybrid schema for which annotated data is available:

Green (Green, 2018). It is schema of 15 arguments annotated for one single article from the biomedical domain. The schema includes both entities and relations that are organized in a short taxonomy. Both schema and the annotated article are available.[11]

3 Methods

We evaluated tools that consider some of the schemes that we found (cf. Section 2) for the task of text similarity in the scope of our use case of mining alternative methods for animal experiments. In this section we described the data and the tools that we used as well as the evaluation methodology.

3.1 Data

We evaluated the selected schemes and tools for the task of text similarity. For this purpose, we

[8] https://github.com/Franck-Dernoncourt/pubmed-rct

[9] https://doi.org/10.1371/journal.pcbi.1000391.s002

[10] https://github.com/PKU-TANGENT/SciDTB

[11] https://github.com/greennl/BIO-Arg

model our problem as the following: given an input document that describes an animal experiment, we would like to mine similar candidate documents that are potential alternatives to animal testing. Our definition of similarity requires that both input and candidate documents should have similar research goal and comparable outcomes. However, the methods in the input document should be substantial different from those in the candidate documents. Therefore, we aim to compare input and candidate documents based on certain rhetorical elements as opposed to using the whole text.

Our evaluation datasets consist of seven input documents from Medline whose identifiers (PMIDs) are 11489449, 11932745, 16192371, 16850029, 19735549, 21494637 and 24204323. For each input document, we collected the top 200 documents (titles and abstracts) retrieved from PubMed's "similar articles" functionality. On one hand, the candidate documents are already very similar to the input document. On the other hand, the list of candidates returned by PubMed does not consider our definition of similarity.

In order to build a suitable test set for our use case, a biomedical researcher manually validated at least the top 100 documents with regards to three degrees of similarity: very similar, similar and not similar. These three labels only consider the similarity of the research goals of each pair of abstracts (input vs. candidate documents) but do not address the 3R principles. Some documents were ignored because either they were only partially similar or because no decision could be made only based on the title and the abstract.

After manual validation by the expert, our seven datasets encompass a total of 562 publications (titles and abstracts). Figure 1 illustrates the distribution of the labels for each input document. Only four from the seven input documents had very similar publications (from only 2 to 8 of them), while similar ones (from only 4 to 19) could be found for all of them. However, the non similar publications are still the largest part (from 56 to 76) of the list. The annotated data is available for download [12].

Some of the tools that we compared require some linguistic information not originally included in our documents, such as sentences and tokens. We utilized syntok[13] for both sentence splitting and tokenization to build input data for one of the tools, namely, Prasad and Kan (2017).

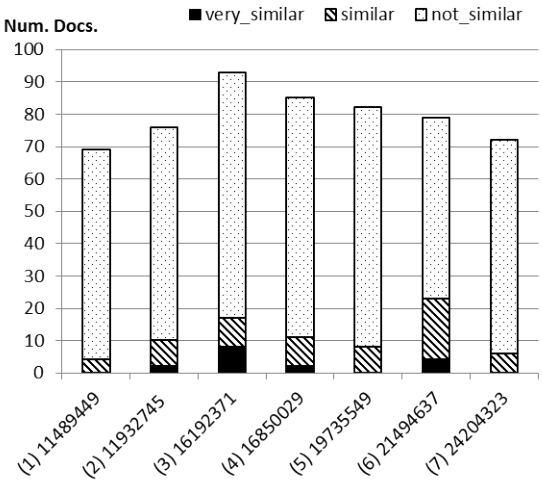

Figure 1: Number of documents according to the degree of similarity to the input document. The number of the dataset (1-7) is shown before the PMID.

3.2 Tools

We found a few available tools that address some of schemes discussed in Section 2. However, we had dismiss some of them due to various problems.

We experienced many problems with the TensorFlow library while trying the tool[14] developed by (Eger et al., 2017) for the ScienceIE schema. The tool seems to require a version of the library that it is no longer available and we could not resolve this issue not even after contacting the tool's developers. We also dismissed the tool[15] from Jin and Szolovits (2018) for the PIBOSO and PubMedRCT schemes. The installation worked but we were not able to train it due to memory problems. Finally, we did not try the tool[16] from Luan et al. (2018) since it addresses a relation-based schema (Gábor) that requires pre-tagged entities. Using named entities provided by other tools would probably add too much noise to the experiment. Finally, we had to dismiss the SAPIENTA tool (Liakata et al., 2012b) because it only allows uploading documents one by one to the Web application and we could not overcome this problem. We describe below the four tools that we

[12] https://github.com/mariananeves/scientific-elements-text-similarity

[13] https://github.com/fnl/syntok

[14] https://github.com/UKPLab/semeval2017-scienceie

[15] https://github.com/jind11/HSLN-Joint-Sentence-Classification

[16] https://bitbucket.org/luanyi/semeval2018/src/master/

tried for the extraction of rhetorical elements. Examples for the sentence-based (zones) and entity-based annotations are shown in Figure 2. We released in the GitHub repository the annotations extracted by the tools in the JSON format supported by the TextAE tool[17].

Achakulvisut et al.[18] (Achakulvisut et al., 2018) (PubMedRCT schema). It addresses the PubMed RCT schema, thus provides predictions for five zoning labels, namely, "Background", "Objective", "Method", "Results" and "Conclusions". We utilized the pre-trained models for Conditional Random Fields (CRF) as provided by the tool. Given that there is no publication, it is not clear what methods are behind the available models, but probably CRF.

ArguminSci[19] (Lauscher et al., 2018a) (Dr. Inventor schema extended). ArguminSci is available both for download as well as on-line (Web application). It provides predictions for five schemes but we considered only the "Discourse Role Classification (DRC)" whose labels are "Background", "Challenge", "Approach", "Outcome" and "Future Work". ArguminSci's models are based on bidirectional recurrent networks with long short-term memory cells (Bi-LSTMs) and we utilized the command line version of the tool.

MAZEA tool[20] and schema (Dayrell et al., 2012). The tool addresses six categories, namely, "Background", "Gap", "Purpose", "Method", "Result" and "Conclusion". It is currently not available for download but only as a Web tool that requires to manually upload each document individually. However, the developers kindly processed our documents locally and sent the predictions back to us. The tool utilizes machine learning algorithms, such as Support Vector Machines (SVM) and Decision Trees.

Prasad and Kan[21] (Prasad and Kan, 2017) (ScienceIE schema). It addresses the three labels for entities from the ScienceIE schema, namely, "Task", "Process" and "Material". From the repository, we utilized the scripts for feature processing and the template to train the model with CRF++[22]. We had to correct the provided template in order to successfully train the system. The entity recognition approach is based on various features and uses the CRF algorithm.

3.3 Evaluation

We evaluated the tools for the task of text similarity. Therefore, we calculated the similarity between the input and candidate documents, either based on the whole text or on selected rhetorical elements as provided by the tools. When utilizing the output from the various tools, we built a pseudo-document based either on the sentences or entities that we obtained. For the zoning tools, we concatenated the sentences to form a single text, while we printed the entities (one per line) for the entity-based predictions. Similarly, when evaluating combination of various labels, we concatenated the text from various labels into a single file.

We performed text similarity using the TextFlow tool (Mrabet et al., 2017) and utilized these similarity scores to rank the candidate documents. Subsequently, we evaluated the ranked list with regard the metrics of precision, recall and f-score at rank 10, i.e. P@10, R@10 and F@10. P@10 is the rate of correct positive candidate documents in the top 10 highest ranked documents, i.e. $P@10 = \frac{TP@10}{10}$. The R@10 corresponds to the rate of positives candidate documents in the top 10 over the total of all positive instances, i.e. $R@10 = \frac{TP@10}{Num.Positive}$. Finally, the F@10 is the harmonic average of the P@10 and R@10 above, i.e. $F@10 = \frac{2*P@10*R@10}{P@10+R@10}$.

We considered as positive examples all those publications manually classified by our expert as "very similar" or "similar". Given the few of these instances in our datasets, we decided to make no distinction between both categories. As a result, the number of positive examples for the input documents in Figure 1 are 4, 10, 16, 11, 8, 23 and 6, respectively. We evaluated at rank 10 due to the reason that only two datasets have more than 20 positive instances, while only two of them over 10 positive instances. For datasets which contain more than 10 positive examples, we considered the number of positive instances to be equal to 10 in the equation of R@10. For the final comparison between the various tools and baselines, we per-

[17] http://textae.pubannotation.org/

[18] https://github.com/titipata/detecting-scientific-claim

[19] https://github.com/anlausch/ArguminSci

[20] http://www.nilc.icmc.usp.br/mazea-web/

[21] https://github.com/animeshprasad/science_ie

[22] https://taku910.github.io/crfpp/

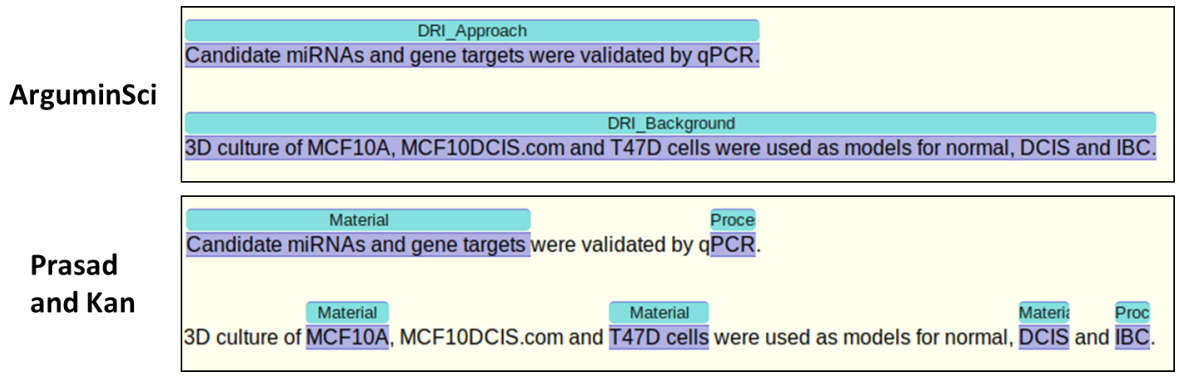

Figure 2: Visualization in the TextAE tool of the annotations provided by two of the tools that we used.

formed an average of the metrics over the seven datasets.

We defined two baselines for comparison: (i) the original order of the candidate documents as returned by PubMed's "similar articles" functionality; and (ii) string similarity based on the whole text (title and abstract) without any pre-processing on the text. For the first baseline, we searched in PubMed for each of the seven PMIDs and downloaded the list of the top 100 similar articles (stand of March 13th, 2019). Given that the current list of similar articles might include citations not present at the time when our corpus was annotated, we dismissed any document not included in our dataset when calculating the above metrics, i.e., we did not consider them as false positives.

Tools	P@10	R@10	F@10
PubMed	0.30	0.33	0.31
Title+Abstract	0.43	0.51	0.45
Achakulvisut et al	0.44	0.52	0.47
ArguminSci	0.47	0.56	0.50
MAZEA	0.4	0.47	0.42
Prasad and Kan	0.44	0.54	0.47
Min score	0.14	0.16	0.15
Max score	0.83	1.0	0.90

Table 2: Summary of the results from the two baselines (two first rows) and when using the selected tools. The maximum scores represent the maximum value of P@10, R@10 and F@10 that could have been obtained by any of the approaches. The minimum scores are the ones obtained when randomly selecting 10 candidates in each dataset, averaged over 1,000 experiments.

4 Results

We compared the tools based on the metrics of P@10, R@10 and F@10 that assess the performance of the various tools for the ranking task. We performed a total of 38 experiments which includes the four tools and baselines, as well as some combinations of selected labels from the tools. The combination of labels were decided based on the performance of the single labels and on our understanding of which labels are more relevant for our use case. Table 2 presents the results for our two baselines and the best results for each tool. In the following we specify the labels that obtained the best results:

- Achakulvisut et al: the combination of all labels, i.e. "Background-Conclusions-Methods-Objective-Results"
- ArguminSci: two combinations of labels were equally good: "Background-Challenge-Outcome" and "Background-Challenge-Outcome-FutureWork".
- MAZEA: the combination "Method-Result".
- Prasad and Kan: the combination "Process-Material".

For our datasets, all approaches using rhetorical tools obtained a better performance than the baseline from PubMed. Further, three tools scored higher than our strong baseline that uses TextFlow over the whole text (titles and abstracts). Two of the tools (Achakulvisut et al and ArguminSci) address zoning elements while one of them (Prasad and Kan) returns entity-level annotations. However, none of the tools scored close the maximum possible scores. Given that we do not have at least 10 positive instances ("very similar" or "similar") for some of our input documents, our maximum P@10 is of 0.83 instead of 1.0.

The three zoning tools rely on labels that can

Tools	Labels	P@10	R@10	F@10
Achakulvisut	Background	0.28	0.32	0.30
	Objective	0.33	0.41	0.35
	Methods	0.31	0.40	0.34
	Results	0.20	0.25	0.22
	Conclusions	0.23	0.26	0.24
ArguminSci	Background	0.23	0.25	0.24
	Challenge	0.23	0.26	0.24
	Approach	0.26	0.32	0.28
	Outcome	0.41	0.50	0.44
	Future Work	0.33	0.41	0.35
MAZEA	Background	0.24	0.28	0.25
	Purpose	0.24	0.25	0.25
	Method	0.30	0.37	0.32
	Result	0.28	0.32	0.30
	Conclusion	0.23	0.30	0.25
Prasad	Process	0.37	0.48	0.40
	Material	0.31	0.35	0.33
	Task	0.28	0.36	0.31

Table 3: Performance of the single labels in the re-ranking task.

be mapped to one another, as shown by the order of their labels in Table 3. When examining the performance of single labels, only the "Outcome" label from ArguminSci tool could perform close our strong baseline.

The labels that we expected to be more relevant, i.e. the ones more related to the background and outcome sections and less with the methods section, did not always perform better in the ranking task. For instance, the F@10 obtained by the label "Approach" from ArguminSci performed slightly better (0.28) than the "Background" (0.24) and "Challenge" (0.24) labels. Similarly, the label "Method" from MAZEA performed better (0.32) than "Background" (0.25) and "Purpose" (0.25) sections. We wonder whether the good performance of methods-related labels were actually due to mistakes in the classification performed by the tools.

Our experiments showed that a combination of labels always performed better than the single ones, while some combinations of labels performed better than others (cf. Figure 3). We could not find any difference in the text similarity scores (as computed by TextFlow) when considering different order of the same labels in the concatenation of the text.

5 Discussion

We carried out a total of 38 experiments that involved diverse tools, single labels and combination of various labels. We ran an error analysis to learn more about the false negatives and false positives that we obtained.

At least one positive document was missed by any of the tools, i.e. was not placed among the top 10 positions. Many of the documents that we missed are certainly due to the limitation of considering only the top 10 highest ranked positions. However, none of the experiments obtained a recall of 1.0. The highest recall that we obtained was 0.9 for the dataset 3 (16192371) using the ArguminSci tool and either the single label "Outcome" or the combination of labels "Challenge-Outcome-FutureWork".

On one hand, five documents were missed by all experiments (38 times), namely, candidate documents "19155551", "29133591", "21362567", "19667187" and "26047474" from datasets 3, 5, 6, 6, and 7, respectively. On the other hand, the candidate document "25174890" from dataset 6 was the least missed one: only by three experiments. A total of 333 documents were wrongly classified as positive, i.e. were placed among the top 10 ones, by any of the 38 experiments. No candidate document was mistakenly classified by all approaches, but the more frequent ones were: "21501651" (27 times) and "23571276" (25 times), both from dataset 4, and "11494364" (25 times) from dataset 7. Our expert checked again the labels assigned to the top FPs and FNs above described and confirmed that their labels are correct and that the documents have been wrongly classified by the corresponding approaches.

Our experiments have shown that many of the tools can indeed support our use case, specially when compared to the original list provided by PubMed. Regarding the integration of these tools into a workflow, one of the tools is currently not available (MAZEA), while all the others need some adaptations to be used in real-life applications. With respect to the methods behind the tools, ArguminSci, which is based on LSTM, performed slightly better than the ones based on CRF (Achakulvisut et al, Prasad and Kan) and superior than the machine learning algorithms in MAZEA. However, we did not evaluate the predictions made by the tools, but only their impact in a specific text similarity task.

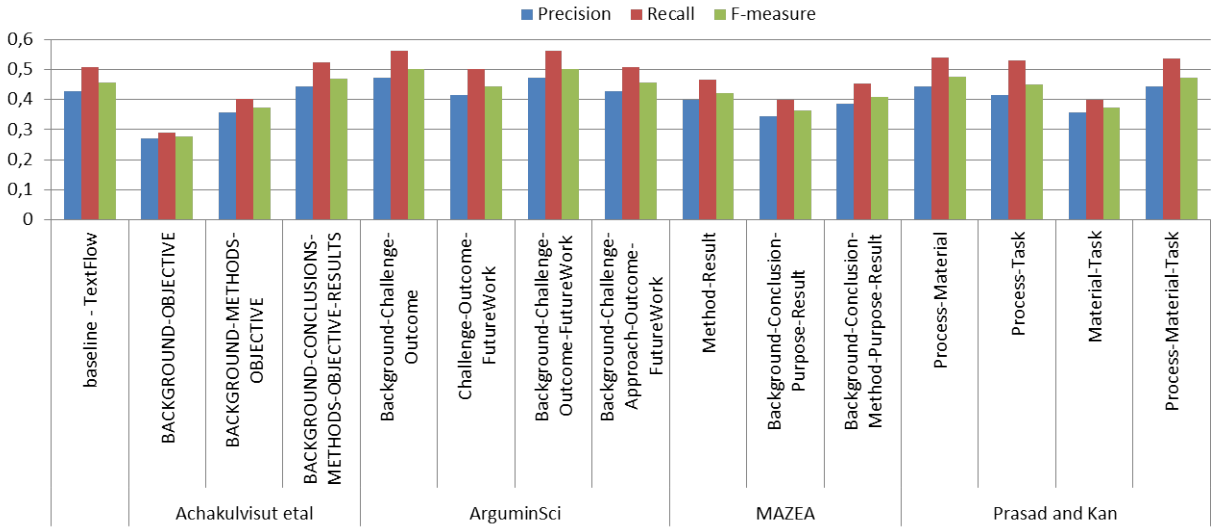

Figure 3: Comparison to the baselines of various combinations of labels as provided by the tools.

We expected that the best performing tools would be the ones that utilized corpora specifically built for the biomedical domain. From the tools that we evaluated, only Achakulvisut et al and MAZEA were specifically trained on documents from the biomedical or health domains. Nevertheless, ArguminSci, the best performing one, was trained on documents from computer graphics while and Prasad and Kan utilizes documents about computational linguistics.

We also investigated whether there was any impact of the document type in the corpora, i.e. either full texts or only abstracts, on the performance of the corresponding tools. However, we did not observe any clear association between these two aspects. While the best performing tool (ArguminSci) was trained on full texts, Achakulvisut et al utilizes only Medline abstracts. Similar to ArguminSci, the tool from Prasad and Kan is also based on full text documents.

We carried out experiments with various tools but limited to a very specific use case. Even though our datasets contains a reasonable number of documents (562), the similarity of the candidate documents was computed with respect to only seven input documents, and datasets were annotated by only one annotator. Further, we only considered titles and abstracts in our evaluation, while some tools were trained on full-text documents. Previous work has already shown the differences of information and performance of NLP tools in biomedical abstracts and full texts (Verspoor et al., 2012; Mons et al., 2004). Our future work will address many aspects: (i) use of full texts; (ii) improvement of the datasets with additional annotators; (iii) estimation of the compliance with the 3R principles by a candidate document, in addition to the calculation of similarity; (iv) evaluation of the relation-based tool (Luan et al., 2018) and the one for which we experienced memory problems (Jin and Szolovits, 2018); and (v) evaluation of other schemes (e.g. Wilbur et al. (2006)) for which an implementation is currently not available.

6 Conclusions

We surveyed schemes that model scientific elements in publications and selected four schemes for which we could find an available tool. We utilized the predictions from these tools for assessing the text similarity between documents and further ranking them in the scope of mining alternative methods to animal testing. Our experiments show that a considerable improvement can be obtained when using ArguminSci, with respect to the original ranking returned by PubMed and to the strong baseline that we considered. However, there is still much room for improvement given that the obtained scores are still far below the possible maximum values.

Acknowledgments

We would like to thank Arnaldo Candido Junior and Sandra Maria Aluísio from the MAZEA tool for kindly processing our documents. We also would like to thank Animesh Prasad and Min-Yen Kan for their support when using their tool.

References

Titipat Achakulvisut, Chandra Bhagavatula, Daniel E Acuna, and Konrad P Kording. 2018. Claim extraction for scientific publications. https://github.com/titipata/detecting-scientific-claim.

Shashank Agarwal and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Summit on Translat Bioinforma*, 2009:6–10. Amias2009-6[PII].

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2):173 – 189.

Carmen Dayrell, Arnaldo Candido Jr., Gabriel Lima, Danilo Machado Jr., Ann Copestake, Valria Feltrim, Stella Tagnin, and Sandra Aluisio. 2012. Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313. Asian Federation of Natural Language Processing.

Sonali K. Doke and Shashikant C. Dhawale. 2015. Alternatives to animal testing: A review. *Saudi Pharmaceutical Journal*, 23(3):223 – 229.

Steffen Eger, Erik-Lân Do Dinh, Ilia Kuznetsov, Masoud Kiaeeha, and Iryna Gurevych. 2017. Eelection at semeval-2017 task 10: Ensemble of neural learners for keyphrase classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 942–946. Association for Computational Linguistics.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 42–51, Denver, Colorado, USA. Association for Computational Linguistics.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 Task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Kata Gábor, Haifa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the acl anthology corpus for the automatic analysis of scientific literature. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Nancy Green. 2018. Proposed method for annotation of scientific arguments in terms of semantic relations and argument schemes. In *Proceedings of the 5th Workshop on Argument Mining*, pages 105–110. Association for Computational Linguistics.

Franz P Gruber and Thomas Hartung. 2004. Alternatives to animal experimentation in basic research. *ALTEX*, 21 Suppl 1:331.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, pages 99–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yufan Guo, Ilona Silins, Ulla Stenius, and Anna Korhonen. 2013. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. *Bioinformatics*, 29(11):1440–1447.

Sonal Gupta and Christopher D. Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *In Proceedings of IJCNLP*.

Hamed Hassanzadeh, Tudor Groza, and Jane Hunter. 2014. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of Biomedical Informatics*, 49:159 – 170.

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

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3100–3109. Association for Computational Linguistics.

Yuchul Jung. 2017. A semantic annotation framework for scientific publications. *Quality & Quantity*, 51(3):1009–1025.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*, 12(2):S5.

Aleksandar Kovačević, Zora Konjović, Branko Milosavljević, and Goran Nenadic. 2012. Mining methodologies from nlp publications: A case study in automatic terminology recognition. *Computer Speech & Language*, 26(2):105 – 126.

Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018a. Arguminsci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 22–28. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018b. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46. Association for Computational Linguistics.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012a. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Maria Liakata, Paul Thompson, Anita de Waard, Raheel Nawaz, Henk Pander Maat, and Sophia Ananiadou. 2012b. A three-way perspective on scientific discourse annotation for knowledge extraction. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, ACL '12, pages 37–46, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. The uwnlp system at semeval-2018 task 7: Neural relation extraction model with selectively incorporated concept embeddings. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 788–792. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1987. *Rhetorical Structure Theory: Description and Construction of Text Structures*. Springer Netherlands, Dordrecht.

Adam Meyers, Giancarlo Lee, Angus Grieve-Smith, Yifan He, and Harriet Taber. 2014. Annotating relations in scientific articles. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75(6):468 – 487. Recent Advances in Natural Language Processing for Biomedical Applications Special Issue.

B. Mons, B. J. A. Schijvenaars, C. C. van der Eijk, E. M. van Mulligen, J. A. Kors, M. Weeber, M. J. Schuemie, and R. Jelier. 2004. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–2604.

Yassine Mrabet, Halil Kilicoglu, and Dina Demner-Fushman. 2017. Textflow: A text similarity measure based on continuous sequences. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 763–772. Association for Computational Linguistics.

Tony Mullen, Yoko Mizuta, and Nigel Collier. 2005. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *SIGKDD Explor. Newsl.*, 7(1):52–58.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics*.

Balaji Polepalli Ramesh, Rashmi Prasad, Tim Miller, Brian Harrington, and Hong Yu. 2012. Automatic discourse connective detection in biomedical text. *Journal of the American Medical Informatics Association*, 19(5):800–808.

Animesh Prasad and Min-Yen Kan. 2017. Wing-nus at semeval-2017 task 10: Keyphrase extraction and classification as joint sequence labeling. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 972–976, Vancouver, Canada. Association for Computational Linguistics.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC Bioinformatics*, 12(1):188.

Balaji Polepalli Ramesh and Hong Yu. 2010. Identifying discourse connectives in biomedical text. *AMIA Annu Symp Proc*, 2010:657–661. Amia-2010_sympproc_0657[PII].

Francesco Ronzano and Horacio Saggion. 2015. Dr. inventor framework: Extracting structured information from scientific publications. In *Discovery Science*, pages 209–220, Cham. Springer International Publishing.

Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbhler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, and Anne-Lise Veuthey. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76(2):195 – 200. Connecting Medical Informatics and Bio-Informatics - MIE 2005.

Andrey Rzhetsky, Hagit Shatkay, and W. John Wilbur. 2009. How to get the most out of your curation effort. *PLOS Computational Biology*, 5(5):1–13.

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

Masashi Shimbo, Takahiro Yamasaki, and Yuji Matsumoto. 2003. Using sectioning information for text retrieval: a case study with the medline abstracts. In *Proceedings of Second International Workshop on Active Mining (AM'03)*, pages 32–41.

Yuka Tateisi, Tomoko Ohta, Sampo Pyysalo, Yusuke Miyao, and Akiko Aizawa. 2016. Typed entity and relation annotation on computer science papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Yuka Tateisi, Yo Shidahara, Yusuke Miyao, and Akiko Aizawa. 2013. Relation annotation for understanding research papers. In *LAW@ACL*.

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

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1493–1502, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrea Varga, Daniel Preotiuc-Pietro, and Fabio Ciravegna. 2012. Unsupervised document zone identification using probabilistic graphical models. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Karin Verspoor, Kevin Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L. Johnson, Christophe Roeder, Jinho D. Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, Nianwen Xue, William A. Baumgartner, Michael Bada, Martha Palmer, and Lawrence E. Hunter. 2012. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13(1):207.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437490.

W. John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7(1):356.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency treebank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449. Association for Computational Linguistics.