# A select and rewrite approach to the generation of related work reports

Ahmed AbuRa'ed, Horacio Saggion

LaSTUS/TALN Group, Universitat Pompeu Fabra, Spain
{name.surname}@upf.edu

Abstract

A related work report is a text which integrates key information from a list of related scientific papers providing context to the work being presented. In this paper we study the automatic generation of related work reports using extractive and abstractive text summarization approaches. Our extractive approach scores the sentences of the scientific papers based on their citations, selecting top scored sentences from each scientific paper to be mentioned in the related work report. The sentences are then organized in the report according to the topic they belong to. In additional experiments we use top scored sentences from our extractive methods and rephrase them using pre-trained abstractive models that generate citation sentences. We discuss automatic and manual evaluation of the generated related work reports showing the viability of the proposed approaches.

Keywords

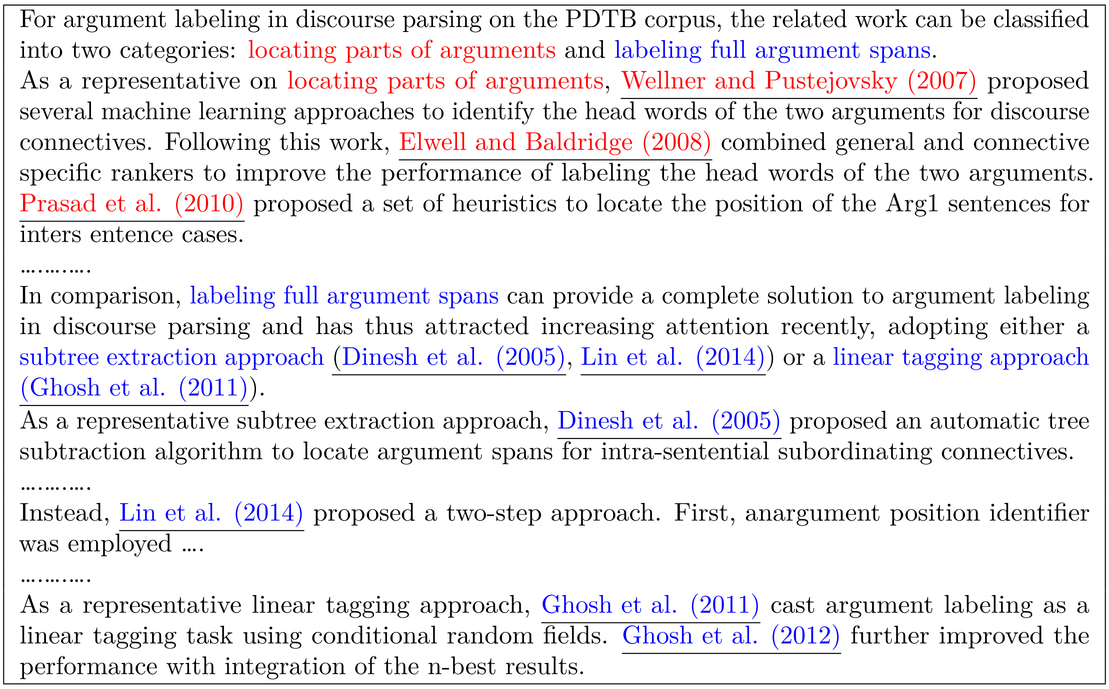
Scientific Summarization, Information Extraction from Scientific Literature, Document Abstracting

## 1. Introduction

Related work reports, or state of the art sections, are an essential part of every scientific paper since they help readers understand the context of a contribution in an area of research, also facilitating any form of comparison between the current paper and previous works. The availability of related work reports is of paramount importance for researchers because they provided condensed information of major contributions in an area of research. However, producing a related work report is a challenging task which requires mastering scientific discourse writing. In this work we are interested in automatically producing state of the art reports by applying text summarization techniques.

A specific example of the type of text we investigate is shown in Figure 1. This related work section introduces previous related works for a paper on Argument Labeling In Discourse Parsing. From Figure 1, we can have a glance at the structure of related work reports. Related work reports usually discuss several different topics: for example in Figure 1 evident topics are "locating parts of arguments" and "labelling full argument spans". Besides facts from previous approaches, comparison statements, differences, and advantages are generally mentioned.

BIR@ECIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval



For argument labeling in discourse parsing on the PDTB corpus, the related work can be classified into two categories: locating parts of arguments and labeling full argument spans.
As a representative on locating parts of arguments, Wellner and Pustejovsky (2007) proposed several machine learning approaches to identify the head words of the two arguments for discourse connectives. Following this work, Elwell and Baldridge (2008) combined general and connective specific rankers to improve the performance of labeling the head words of the two arguments. Prasad et al. (2010) proposed a set of heuristics to locate the position of the Arg1 sentences for inters entence cases.
………..
In comparison, labeling full argument spans can provide a complete solution to argument labeling in discourse parsing and has thus attracted increasing attention recently, adopting either a subtree extraction approach (Dinesh et al. (2005), Lin et al. (2014)) or a linear tagging approach (Ghosh et al. (2011)).
As a representative subtree extraction approach, Dinesh et al. (2005) proposed an automatic tree subtraction algorithm to locate argument spans for intra-sentential subordinating connectives.
………..
Instead, Lin et al. (2014) proposed a two-step approach. First, anargument position identifier was employed ….
………..
As a representative linear tagging approach, Ghosh et al. (2011) cast argument labeling as a linear tagging task using conditional random fields. Ghosh et al. (2012) further improved the performance with integration of the n-best results.

Figure 1: A sample related work section.

According to previous research [1, 2], related work reports are classified into descriptive or integrative: a descriptive report will summarize individual papers providing information such as methods and results in citation sentences. Instead, integrative reports will focus on key ideas and topics, providing in the citation sentences critical views on the presented approaches.

In this paper we are concerned with the automatic production of descriptive related work sections from a set of selected papers by using extractive and abstractive (sequence-to-sequence) methods. The extractive method to be presented, which achieved state of the art performance in citation-based summarization, learns to score sentences using a Convolutional Neural Network (CNN). We use the top ranked sentences of each paper to generate our extractive report. In order to account for coherence phenomena observed in this type of text (See Figure 1), we propose a topic-based modelling approach for information ordering. Moreover, in order to generate sentences matching the related work report style we rely on an abstractive method based on Bidirectional Recurrent Neural Networks (BRNN) and trained to generate citation sentences. The sentences extracted with the CNN methods are feed into the BRNN to produce an abstractive related work report. As it will be shown, the combination of extractive and abstractive techniques improves the results in terms of automatic evaluation.

The rest of the paper is organized as follows: in Section 2 we report related work in the area, then in Section 3 we explain the methodology and data used. Section 4 presents the experiments and Section 5 discusses automatic and human evaluation results. Finally, Section 6 closes the paper.

## 2. Related Work

In contrast with generic summarization, state of the art generation and summarization has not been extensively explored. Key works in the area are: [3] and [4], making [3] to be the first to generate related work sections from a hierarchical topic-biased tree, and [5] who deal with multi-document scientific article summarization. Other studies investigate mainly single document scientific article summarization. In respect to that, we will cover two main types of related works: Automated related work summarization and Automatic text summarization (ATS) in the domain of scientific texts.

### 2.1. Automatic text summarization (ATS) in the domain of scientific texts

Although research in summarization can be traced back to the 50s [6] and even though a number of important discoveries have been produced in this area, automatic text summarization still faces many challenges given its inherent complexity. Scientific text summarization is of paramount importance and scientific texts were automatic summarization's first application domain [6, 7]. Several methods and techniques have already been reported in the literature to produce text summaries by automatic means [8].[5] tackled the multi-document summarization of scientific articles problem by an original unsupervised method, in which the source document cites a list of papers (also known as a co-citation). From each co-cited article, a topic based clustering of fragments was mined and ranked using a query produced from the context surrounding the co-cited list of papers. [9] proposed a model which uses a clustering approach to summarize a single topic from the article and this summarized topic is further used to summarize the entire topic of the specified article. The main contribution is to use citation summaries and network analysis techniques which yield a summary of a single scientific article as a framework for future research on topic summarization. [10] proposed a summarization approach for scientific articles which takes advantage of citation-context and the document discourse model. They also leverage the inherent scientific article's discourse for producing better summaries. [11] suggested that performing reinforcement ranking on the Semantic Link Network of various representation units within a scientific paper (word, sentence, paragraph and section) can significantly improve extractive summarization of paper. [12] proposed an approach to generate automatic summarization based on 5W1H (who, what, whom, when, where, how) event structure. Sentences in literature are classified and selected for different elements of events by relevance, and then, the importance of each candidate sentence is calculated. Top-k relevant and important sentences are selected to formulate event-based summarization.

### 2.2. Automated related work summarization

[3] and [13] presented the novel problem of automatic related work summarization. A related work summarization system creates a topic-biased summary of related work for a target paper given multiple scientific articles together with a topic hierarchy tree as an input. [3] also stated that three things should be considered to generate a summary. First,

a mandatory input is needed for the summarization process identified as a high-level rhetorical structure in a form of a topic tree. Second, summaries can be seen as transitions along the topic hierarchy tree. Third, sentences either describe generic or specific topics. Generic topics are often characterized by background information. This include definitions or descriptions of a topic's purpose. In contrast, detailed information forms the substance of the summary and often describes key related work that is attributable to specific authors.

[4] investigated on the task of producing a related work section for a target paper, provided a set of Reference Papers along with a target academic paper which has no related work section as input. They developed an Automatic Related Work Generation system (ARWG) that exploits the Probabilistic Latent Semantic Analysis (PLSA) [14] to solve this problem. They used the PLSA model to divide the sentence set of the given papers into different topic-biased parts, and then applies regression models to learn the standing (ranking) of the sentences. Finally, it utilizes an optimization framework to produce the related work section. Their evaluation results on a test set of 150 target papers sideways with their Reference Papers show that ARWG can indeed generate related work sections with improved quality than those of baseline methods MEAD [15] and LexRank [16]. A user study is also carried on to demonstrate that ARWG can achieve improvements over generic multi-document summarization baselines. It is worth noting that in this work they use abstract, introduction, related work and conclusion sections, since other sections corresponding to method and evaluation sections always describe in too much details the specific work.

## 3. Method

We score the sentences of the scientific papers based on their citation network selecting those which score higher, then we generate an organized related work report based on the relation between the scientific papers being summarized. In order to score the sentences we use both supervised and unsupervised approaches. For the unsupervised learning we use two methods: one is based on a modified variant of Jaccard Similarity, and the other is based on the BabelNet Embeddings Distance. As for the supervised approach we use several variations of a Convolutional Neural Network (CNN) [17]. Once we have scored the sentences, we select a unified number of sentences from each scientific paper and add them to the final related work report in topic order. We use Latent Dirichlet Allocation (LDA) [18] to perform topic modeling across the scientific papers detecting any cross document linking between the scientific papers based on their topics.

For additional experiments we rephrase the selected sentences using pretrained sequence-to-sequence models trained with citation-sentences to paraphrase the extracted sentences in a citation style.

### 3.1. Data

In order to generate related work reports through extractive summarization of scientific papers we utilize a Multi-level Annotated Corpus of Scientific Papers. The corpus is proposed by [19] which expands considerably the data-set of related work sections used in [3] by providing: (i) related work sections, (ii) a manually annotated layer of cited papers and sentences, (iii) citing papers referring to the cited papers in the related work section, and (iv) a layer of rich linguistic, rhetorical, and semantic annotations computed automatically. The corpus contains three types of scientific papers: target papers, reference papers, and citing papers which are organized in a two-level network. Level 1 contains target papers with their related work sections which cite a set of reference papers. Level 2 extends the corpus by adding a layer representing a set of scientific papers explicitly citing the reference papers in Level 1. This data-set is ideal for our research: we use Level 2 of the corpus to score the reference papers' sentences through their connection with the scientific papers citing them in the citation network. Next, we select the top sentences of the reference papers and generate an organized related work report using Level 1 of the corpus, in which we try to re-create the related work section of the target paper by summarizing the reference papers mentioned in it. Finally, for evaluation we use the gold related work sections of the target papers provided by the corpus.

### 3.2. Scoring sentences of the Reference Papers

Researchers tend to cite the major contributions of a scientific paper. Therefore, utilizing the citation network between the scientific paper and the papers that are citing it will provide an insight of what those researchers consider an important context in the scientific paper.

We performed experiments using two unsupervised methods using two similarity metrics: Modified Jaccard similarity and BabelNet (a multilingual lexicalized semantic network and ontology) Embeddings Distance [20]. We used a metric similar to the Jaccard similarity coefficient [21] for comparing two sentences (i.e., the citation sentence of a citing paper with every sentence within a reference paper). This metric considers the union and intersection of words (like the Jaccard coefficient) but uses the inverted frequency information to give more weight to words in the intersection that are less common. Our modification assigns greater weight to matching words that are infrequent in the corpus, based on the idea that two text spans that share infrequent words are more likely to be semantically related. The modified Jaccard similarity between two text spans $s_1$ and $s_2$ is defined in Equation 1.

$$MJ(s_1, s_2) = \frac{\sum_{t \in s_1 \cap s_2} 2^{idf(t)}}{|s_1 \cup s_2|} \tag{1}$$

As for the second method we obtained the BabelNet synsets for both sentences and transformed them into synset embeddings [22]. We then take the cosine similarity between the centroids of the synset embeddings for both the reference and citation sentences. Once we have collected the similarities based on these two metrics between each citation context in a citing paper and each sentence in the reference paper they cite, we formed the final score for each sentence in the reference paper. The final score takes into account the similarity of the citation context in the citing paper with the sentence in the reference paper alongside the assigned weight of that citing paper. The weight for each scientific paper is based on the number of scientific papers citing it.

On the other hand, for our supervised approach, we needed a source of data to train our CNN models. To do so we make use of a data set provided by the CL-Scisumm Shared challenge [23] which addresses the problem of summarizing a scientific paper taking advantage of its citation network. The challenge organizers provided a cluster of $n$ documents where one is a reference paper (RP) and the $n-1$ remaining documents are papers (i.e., citing papers (CPs)) citing the reference paper, they also provide three gold summaries for each reference paper alongside manual annotations stating which sentences in the reference paper have been cited by the citation context of the citing papers. The three types of summaries for each Reference Paper are:

- the abstract, written by the authors of the research paper.
- the community summary, collated from the majority of the reference spans of its citances.
- a human-written summary, written by the annotators of the CL-SciSumm annotation effort.

The CNN scores the sentences of the scientific paper using linguistic and semantic features from the paper itself alongside the papers that are citing it. The aim of our CNN is to learn the relation between a sentence and a scoring value indicating its relevance. Since the CL-Scisumm Challenge Dataset provides several types of gold standard summaries, we use them to train several CNN models. The CNN learns the relation between features and a score, that is regression [24]. The scoring functions are defined below:

- Cosine Distance: we calculated the maximum cosine similarity between each sentence vector in the Reference Paper with each vector in the gold standard summaries. This method produced three scoring functions (based on SUMMA [25] word vectors, ACL embeddings, and Google embeddings) for each summary type.
- ROUGE-2 Similarity: we also calculated similarities based on the overlap of bigrams between sentences in the Reference Paper and gold standard summaries. In this regard, each sentence in the Reference Paper is compared with each gold standard summary using ROUGE-2 [26].
- Scoring Functions Average: Moreover, we computed the average between all cosine scoring functions (SUMMA, ACL, Google and ROUGE-2; referred to as SGAR in the tables) for each summary type. In addition, we also calculated a simplified average with vectors that are not based on word-frequencies (ACL, Google and ROUGE-2; referred to as GAR in the tables).

We have used the same set of features that [27] used at their participation in the CL-SciSumm challenge which achieved state of the art performance. After training the CNN models, we pass the reference papers of the Multi-level Annotated Corpus (Section 3.1) as testing data to score their sentences. Once we have scored the sentences of the reference papers using the supervised and unsupervised methods, we sort the sentences in descending order before moving to selecting and organizing the final output.

### 3.3. Generating the Related Work Report

Since authors of related work reports usually starts with a certain related topic and move onward stating each and every reference paper related to that specific topic before moving to the next topic, we group the sentences of reference papers that share the same topic together. To find the topics across the reference papers we used Latent Dirichlet Allocation (LDA) [18] and modeled each reference paper based on its Title and Abstract. In order to find the optimal number of topics to train the LDA model on, we build many LDA models based on different number of topics (*numT*) and pick the one that gives the highest coherence value or till the coherence value converges, choosing a '*numT*' that marks the end of a rapid growth of topic coherence usually offers meaningful and interpretive topics, while picking an even higher value can sometimes provide more granular sub-topics.

Once we identify the LDA model with the ideal number of topics to train on (*numT*), we use it to identify the topics that each reference paper belongs to. We choose the topic with the highest probability as the representative of a reference paper's topic, assigning each reference paper to only one topic.

For ordering the sentences, we start with the topic of the paper with most citations, adding the sentences from the papers that belongs to the same topic. Afterwards, we repeat the process till all the papers have been included in the report.

Tables 5 and 4 in the appendix show examples of generated related work reports with and without topic modeling applied.

### 3.4. Rewriting Models

In our recent work [28], we have used pointer–generator neural networks with two different architectures; Bidirectional Recurrent Neural Networks (BRNN) [29] and Transformers [30] to train a system able to generate citation-sentences from the title and abstract of a research paper. The pointer–generator networks can copy words from the source text via pointing, which aids accurate reproduction of information while retaining the ability to produce novel words through the generator.

In order to improve our extractive summaries, we use the selected sentences and rephrase them using a set of pretrained abstractive models [28]. These models were trained using over 16K pairs of Title and Abstract of scientific papers as a source sequence (input) with a citation context as a target sequence (output), making them ideal for our experiments. We feed the pretrained models with a scientific context extracted from our extractive methods to generate a rephrased citation context that can represent a scientific paper. Such additional experiments will avoid the use of copy-paste techniques adopted by the extractive models. Once we have a citation context for each reference paper we concatenate them into the final related work report.

## 4. Experiments

We use the Multi-level Annotated Corpus of Scientific Papers as our main data set (testing data) aiming at recreating the related work section (report) of the target paper for each cluster provided. We compared our systems against a set of off-the-shelf baselines. We performed both automatic and human evaluations by comparing the systems to the gold related work sections of the target paper in the Multi-level Annotated Corpus of Scientific Papers.

For scoring the sentences of the reference papers, we ran the unsupervised approaches: Modified Jaccard and BabelNet Embeddings Distance directly over each cluster in the test data. We modeled a pair of vectors from Level 2 of the corpus: the citation context mentioned in the citing papers alongside each and every sentence in the corresponding cited reference paper. Then we scored the sentences, that is, the cosine similarity score in case of the BabelNet Embeddings and the score of the Modified Jaccard similarity. We also trained several CNN models based on the similarity between the reference papers sentences and the abstract, community and human written summaries provided by the CL-Scisumm Challenge Dataset as a training dataset. We trained 6 models for each summary of the reference paper each representing one of the scores: cosine distance: based on ACL, Google and SUMMA, ROUGE-2 similarity and two averages of the four scores: including and excluding SUMMA. After training our CNN models we fed them each and every sentence from the reference paper in the testing data set (the multi-level corpus). Once we scored the sentences of the reference papers using all our methods we sort the sentences in each reference paper in descending order based on the score.

After that, we select the top $m$ sentences from each reference paper in a unified way based on the length of the gold standard. Moreover, we organize the sentences of each reference papers based on their topics. All reference papers that belongs to the same topic were grouped together, the position of the sentences inside each reference paper were respected. We have also generated related work reports without applying topic modeling. Finally, we always generate related work reports that has the same number of sentences as the gold related work report ($N\,\text{system} = N\,\text{gold}$).

We perform additional experiments by selecting the top 1, 3, and 6 sentences from each reference paper, passing them through a set of pretrained abstractive models to produce citation contexts. Finally, we concatenate the citation contexts of the reference papers to form the final related work report. Since [28] produced several pretrained models based on BRNN and Transformer architectures, we have used all of the pretrained models they provided only reporting here the best models.

### 4.1. Baselines

For our experiments we implemented several extractive summarization baselines alongside a set of simple baselines based on the observations arising from the analysis of citation sentences and scientific abstracts on the use of titles and abstracts [2, 31]. The title baseline is to use the title of each cited article as citation sentences. The abstract first baseline uses as citation sentences the first sentence of the abstract of the cited articles while the abstract last baseline uses the last sentence.

The second set of baselines is composed of available systems that use well-established extractive techniques. We have made sure that all the baselines have the same conditions as our systems. That is we fed each and every scientific paper to the baseline and guaranteed that at least the system will select one sentence from each. We also instructed the system to generate the same number of sentences as the gold related work sections ($N$ system = $N$ gold). We describe the systems as follows:

- MEAD [15] is a well-known extractive document summarizer which generates summaries using centroids alongside other features such as the position of the sentence and the length.
- TextRank [32] and LexRank [16] are both extractive and unsupervised graph-based text summarization systems which create sentence graphs in order to compute centrality values for each sentence. Both algorithms have similar underlying methods to compute centrality which are based on the PageRank ranking algorithm. They differ in how links are weighted in the document graph.
- SUMMA [25] is a Java implementation of several sentence scoring functions. We use the implementation of the centroid scoring functionality to select the most central sentence in a document.

## 5. Results, Evaluation and Discussion

In this section we compare our systems against the baseline systems for the task of automatic generation of related work reports. We have performed both automatic and human evaluation. For automatic evaluation we have used 4 ROUGE metrics: ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-SU4. We present all the systems except for the CNN approach for which we only present the top five systems. ROUGE measures combine precision and recall in a harmonic F-measure which is generally used to assess the systems' performance. The results of ROUGE-1 and ROUGE-2 metrics can be found in Table 1, while for ROUGE-L and ROUGE-SU4 the results are presented in Table 6 in the appendix.

The non-informed extractive baselines which do not perform any analysis of the input (e.g. use of titles or sentences from abstracts) tend to have a high precision but low recall, especially precise is the title. Except for LexRank, the off-the-shelf baselines have low performance, which was expected since they are based on poor word-based representations of the document. A CNN approach which uses word embedding, rich summarization features, and scores sentences based on similarity to abstract outperforms (in terms of ROUGE-1) all the other systems.

Finally, we perform automatic evaluation using ROUGE of the rephrasing experiments (only the top five best models are shown, all of them based on BRNN). We report the results of selecting the top one, three and six sentences from the reference papers and

| | ROUGE-1 | | | ROUGE-2 | | |
|---|---|---|---|---|---|---|
| SYSTEM | R | P | F | R | P | F |
| Titles | 0.074 | 0.375* | 0.119 | 0.013 | 0.072 | 0.022 |
| AbsFS | 0.126 | 0.272 | 0.155 | 0.019 | 0.041 | 0.023 |
| AbsLS | 0.114 | 0.263 | 0.150 | 0.013 | 0.035 | 0.018 |
| SUMMA | 0.293 | 0.103 | 0.154 | 0.102 | 0.027 | 0.047 |
| MEAD | 0.361 | 0.137 | 0.205 | 0.118 | 0.025 | 0.049 |
| LexRank | 0.312 | 0.228 | 0.259 | 0.107 | 0.060 | 0.076 |
| TexRank | 0.367 | 0.116 | 0.186 | 0.117 | 0.017 | 0.040 |
| Babelnet | 0.409* | 0.242 | 0.299 | 0.158* | 0.087 | 0.110 |
| MJ | 0.350 | 0.270 | 0.299 | 0.154 | 0.112 | 0.127 |
| $CNN_{ROUGE-2-abstract}$ | 0.322 | 0.291 | 0.302 | 0.142 | 0.120* | 0.128 |
| $CNN_{AvgGAR-abstract}$ | 0.360 | 0.273 | 0.307 | 0.148 | 0.108 | 0.124 |
| $CNN_{AvgSGAR-abstract}$ | 0.359 | 0.278 | 0.310* | 0.149 | 0.110 | 0.125 |
| $CNN_{ROUGE-2-community}$ | 0.353 | 0.275 | 0.305 | 0.152 | 0.111 | 0.126 |
| $CNN_{ROUGE-2-human}$ | 0.319 | 0.285 | 0.298 | 0.145 | 0.120* | 0.130* |

Table 1
Automatic evaluation results of our systems against the baselines for ROUGE-1 and ROUGE-2 metrics. Only the top 5 systems of the CNN approach are shown.

| | | ROUGE-1 | | | ROUGE-2 | | |
|---|---|---|---|---|---|---|---|
| #SEN | SYSTEM | R | P | F | R | P | F |
| 1 | $CNN_{ACL-community}$ | 0.301 | 0.365* | 0.319* | 0.191* | 0.196* | 0.190* |
| | $CNN_{ACL-abstract}$ | 0.296 | 0.363 | 0.315 | 0.184 | 0.180 | 0.179 |
| | $CNN_{AvgSGAR-abstract}$ | 0.297 | 0.358 | 0.314 | 0.184 | 0.180 | 0.179 |
| | $CNN_{AvgGAR-community}$ | 0.296 | 0.358 | 0.313 | 0.185 | 0.178 | 0.178 |
| | $CNN_{AvgGAR-human}$ | 0.299 | 0.351 | 0.313 | 0.172 | 0.178 | 0.172 |
| 3 | $CNN_{AvgGAR-community}$ | 0.287 | 0.358 | 0.306 | 0.178 | 0.174 | 0.173 |
| | $CNN_{ROUGE-2-abstract}$ | 0.299 | 0.326 | 0.303 | 0.189 | 0.178 | 0.180 |
| | $CNN_{AvgSGAR-community}$ | 0.286 | 0.346 | 0.302 | 0.175 | 0.168 | 0.169 |
| | $CNN_{AvgGAR-human}$ | 0.286 | 0.345 | 0.302 | 0.180 | 0.174 | 0.174 |
| | $CNN_{AvgGAR-abstract}$ | 0.282 | 0.350 | 0.301 | 0.177 | 0.168 | 0.170 |
| 6 | $CNN_{AvgGAR-abstract}$ | 0.291 | 0.360 | 0.311 | 0.170 | 0.171 | 0.168 |
| | MJ | 0.303* | 0.332 | 0.307 | 0.186 | 0.176 | 0.178 |
| | $CNN_{ACL-human}$ | 0.294 | 0.338 | 0.305 | 0.178 | 0.173 | 0.172 |
| | Babelnet | 0.298 | 0.331 | 0.304 | 0.171 | 0.166 | 0.166 |
| | $CNN_{ROUGE-2-human}$ | 0.297 | 0.330 | 0.304 | 0.170 | 0.165 | 0.165 |

Table 2
Automatic evaluation results of paraphrasing the top sentences of reference papers. Only the top five abstractive models for ROUGE-1 and ROUGE-2 metrics are reported.

generating a citation context using the pretrained abstractive models trained by [28]. The results are encouraging, the rewriting system improves the results of the extractive methods in terms of ROUGE and also when comparing this with the pure abstractive approach presented in [28], confirming that the selecting and rewriting approach to summarization is a viable alternative to pure extractive or abstractive approaches.

### 5.1. Human Evaluation

In order to assess the quality of the automatically generated related work reports, we selected 10 clusters that discusses different varieties of topics, each cluster was manually evaluated by three subjects with the age group between 25-34 with an expert level of English language, they have a range between good and very good in their expertise in Natural Language Processing (the topic of the analyzed summaries).

The objective of this evaluation is to assess the appropriateness of four different related work sections for a given target paper in the test data set i.e. Multi-level corpus. The four related work sections represent: the best system of the baselines i.e. LexRank, the gold related work section and the best system we have with and without topic modeling applied i.e. $CNN_{AvgSGAR-abstract}$.

To carry out the evaluation we prepared each reference paper's Title, Abstract and Introduction in PDF format. Alongside the scientific paper we provided in a random order the related work sections in text format to be evaluated. We also added a folder with the references that are mentioned in the related work section, we also provided the bibliographic information about each of the references which will be cited in the related work section. Given the target scientific paper Title, Abstract and Introduction section alongside a related work section, we asked them for their opinion on three fronts:

- Responsiveness: How good do you consider the related work section given that it must include information on the list of reference papers and must fit in the target paper.

- Linguistic quality: How do you rate the readability and grammaticality of the related work section? That is: is it understandable? is it grammatically correct (are the sentences correct)? Are there any spelling mistakes? Is punctuation appropriate?

- Text organization: How well organized and coherent the related work section is? That is: does the discourse (topics) flows from sentences to sentence? Are the sentences organized in a coherent way? Is the text not redundant?

We instructed them to read the target scientific paper's Title, Abstract and Introduction (the pdf file), and then to read each related work section (the text file). Once they had finished reading the related work section we informed them to fill the evaluation form indicating the scores for each metric, all the scores were on the scale of 1 to 5. Finally, we requested that they should not check the web for a related work section or the target paper to avoid influence from external variables and use the references folder if they felt they had to.

Table 3 present the average of all the metrics across the 10 clusters for our system with and without topic modeling applied, LexRank: the best baseline in the automatic evaluation and finally the gold related work report. What can be noticed is that our system with topic modeling super-passes the baseline in all metrics and it is considered an improvement over not implementing topic modeling for our system.

| System | Responsiveness | Linguistic quality | Text organization |
|---|---|---|---|
| Gold | 4.5** | 4.4** | 4.3** |
| LexRank | 2.4 | 2.5 | 2.0 |
| WithoutTM | 2.9 | 3.3 | 2.3 |
| TopicModeling | 3.1* | 3.6* | 2.5* |

Table 3
The results of the Human Evaluation over our system with and without applying topic modeling against the LexRank baseline.

## 6. Conclusion

In this paper, we have presented a number of computational approaches for the automatic generation of related work reports. These approaches utilize extractive summarization to describe each scientific paper to be mentioned in the related work report. We utilize a corpus of articles connected through their citation network to test our approach. We perform automatic and human evaluation over of our extractive methods showing their viability. Moreover, we perform additional experiments to rephrase the sentences selected by the extractive methods finding a viable alternative to pure extractive or abstractive approaches. We have found that our regression learning extractive approach obtains competitive results in terms of ROUGE scores when compared with well known baselines. A human evaluation also confirms that the reports generated by the extractive approach are also preferred to the best baseline. Moreover, automatically rewriting the sentences which were automatically extracted produces abstracts which are better in terms of ROUGE. We believe that ”select and rewrite” is a valid strategy to the generation of summaries specially in the case of very long documents such as scientific articles. There are many limitations of our work which indicate further research. In future work different sentence ordering strategies should be investigated. Moreover, the generation of citation sentences for multiple scientific papers is an interesting topic to take on.

## Acknowledgments

This work was (partly) supported by the Spanish Government under the María de Maeztu Units of Excellence Programme (MDM-2015-0502). We also acknowledge support from the project Context-aware Multilingual Text Simplification (ConMuTeS) PID2019-109066GB-I00/AEI/10.13039/501100011033 awarded by Ministerio de Ciencia, Innovación y Universidades (MCIU) and by Agencia Estatal de Investigación (AEI) of Spain.

## References

[1] C. S. G. Khoo, J.-C. Na, K. Jaidka, Analysis of the macro-level discourse structure of literature reviews., Online Information Review 35 (2011).

[2] K. Jaidka, C. Khoo, J.-C. Na, Deconstructing human literature reviews–a framework

for multi-document summarization, in: Proceedings of the 14th European Workshop on Natural Language Generation, 2013, pp. 125–135.

[3] C. D. V. Hoang, M.-Y. Kan, Towards automated related work summarization, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 427–435.

[4] Y. Hu, X. Wan, Automatic generation of related work sections in scientific papers: An optimization approach., in: EMNLP, 2014, pp. 1624–1633.

[5] N. Agarwal, K. Gvr, R. S. Reddy, C. P. Rosé, Towards multi-document summarization of scientific articles: making interesting comparisons with scisumm, in: Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, Association for Computational Linguistics, 2011, pp. 8–15.

[6] H. P. Luhn, The automatic creation of literature abstracts, IBM J. Res. Dev. 2 (1958) 159–165.

[7] H. P. Edmundson, New methods in automatic extracting, J. ACM 16 (1969) 264–285.

[8] H. Saggion, T. Poibeau, Automatic text summarization: Past, present andfuture, in: T. Poibeau, H. Saggion, J. Piskorski, R. Yangarber (Eds.), Multi-source, Multilingual Information Extraction and Summarization, Springer Verlag, Berlin, 2013.

[9] V. Qazvinian, D. R. Radev, Scientific paper summarization using citation summary networks, in: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics, 2008, pp. 689–696.

[10] A. Cohan, N. Goharian, Scientific article summarization using citation-context and article's discourse structure, arXiv preprint arXiv:1704.06619 (2017).

[11] X. Sun, H. Zhuge, Summarization of scientific paper through reinforcement ranking on semantic link network, IEEE Access 6 (2018) 40611–40625.

[12] J. Zhang, K. Li, C. Yao, Y. Sun, Event-based summarization method for scientific literature, Personal and Ubiquitous Computing (2019) 1–10.

[13] H. C. D. Vu, Towards automated related work summarization, Ph.D. thesis, 2010.

[14] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1999, pp. 50–57.

[15] D. R. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, et al., Mead-a platform for multidocument multilingual text summarization., in: LREC, 2004.

[16] G. Erkan, D. R. Radev, Lexrank: Graph-based lexical centrality as salience in text summarization, Journal of Artificial Intelligence Research 22 (2004) 457–479.

[17] W. Zhang, K. Itoh, J. Tanida, Y. Ichioka, Parallel distributed processing model with local space-invariant interconnections and its optical architecture, Applied optics 29 (1990) 4790–4797.

[18] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[19] A. AbuRa'ed, H. Saggion, L. Chiruzzo, A multi-level annotated corpus of scientific papers for scientific document summarization and cross-document relation discovery, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6672–6679. URL:

https://www.aclweb.org/anthology/2020.lrec-1.824.

[20] R. Navigli, S. P. Ponzetto, BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence 193 (2012) 217–250.

[21] S. Niwattanakul, J. Singthongchai, E. Naenudorn, S. Wanapu, Using of jaccard coefficient for keywords similarity, in: Proceedings of the international multiconference of engineers and computer scientists, volume 1, 2013, pp. 380–384.

[22] M. Mancini, J. Camacho-Collados, I. Iacobacci, R. Navigli, Embedding words and senses together via joint knowledge-enhanced training, arXiv preprint arXiv:1612.02703 (2016).

[23] M. K. Chandrasekaran, M. Yasunaga, D. Radev, D. Freitag, M.-Y. Kan, Overview and results: Cl-scisumm shared task 2019, arXiv preprint arXiv:1907.09854 (2019).

[24] H. Saggion, A. AbuRa'ed, F. Ronzano, Trainable citation-enhanced summarization of scientific articles, in: G. Cabanac, M. K. Chandrasekaran, I. Frommholz, K. Jaidka, M. Kan, P. Mayr, D. Wolfram (Eds.), Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) co-located with the Joint Conference on Digital Libraries 2016 (JCDL 2016), Newark, NJ, USA, June 23, 2016, volume 1610 of CEUR Workshop Proceedings, CEUR-WS.org, 2016, pp. 175–186.

[25] H. Saggion, SUMMA. A Robust and Adaptable Summarization Tool, TAL 49 (2008) 103–125.

[26] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text summarization branches out: Proceedings of the ACL-04 workshop, volume 8, Barcelona, Spain, 2004.

[27] A. AbuRa'ed, À. Bravo Serrano, L. Chiruzzo, H. Saggion, Lastus/taln+ inco@ cl-scisumm 2018-using regression and convolutions for cross-document semantic linking and summarization of scholarly literature, in: Mayr P, Chandrasekaran MK, Jaidka K, editors. BIRNDL 2018. 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries; 2018 Jul 21; Ann Arbor, MI.[place unknown]: CEUR; 2018. p. 150-63., CEUR Workshop Proceedings, 2018.

[28] A. AbuRa'ed, H. Saggion, A. Shvets, À. Bravo, Automatic related work section generation: experiments in scientific document abstracting, Scientometrics 125 (2020) 3159–3185. URL: https://doi.org/10.1007/s11192-020-03630-2. doi:10.1007/s11192-020-03630-2.

[29] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (1997) 2673–2681.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[31] H. Saggion, Using linguistic knowledge in automatic abstracting, in: 27th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, USA, 20-26 June 1999., 1999.

[32] R. Mihalcea, P. Tarau, Textrank: Bringing order into texts, Association for

Computational Linguistics, 2004.

# A. Appendix

|  |
|---|
| (Blunsom et al. 2007) Statistical machine translation (SMT) has seen a resurgence in popularity in recent years ... (Kumar and Byrne 2004) We also show how MBR decoding can be used to incorporate syntactic structure into a statistical MT system ... template model for statistical machine translation. (Matsusaki et al. 2005) This paper defines a generative probabilistic model of parse trees, which we call PCFG-LA. This paper defines a generative model of parse trees that we call PCFG with latent annotations (PCFG-LA). (May and Knight 2006) We also demonstrate our algorithm's effectiveness ... to deal with grammars that produce trees. (Petrov et al. 2006) In this paper, we investigate the learning of a grammar consistent with a treebank at ... likelihood of the training trees. We present a method that combines the strengths of both manual and automatic approaches while addressing some of their common shortcomings. (Tromble et al. 2008) In this paper we explore a different strategy to perform MBR decoding over Translation Lattices ... that compactly encode a huge number of translation ... We begin with a review of MBR decoding for Statistical Machine Translation (SMT). |

Table 4
An example of a summary generated by the our system without topic modeling applied.

|  |
|---|
| (Blunsom et al. 2007) Statistical machine translation (SMT) has seen a resurgence in popularity in recent years ... (Kumar and Byrne 2004) We also show how MBR decoding can be used to incorporate syntactic structure into a statistical MT system ... template model for statistical machine translation. (Tromble et al. 2008) In this paper we explore a different strategy to perform MBR decoding over Translation Lattices ... that compactly encode a huge number of translation ... We begin with a review of MBR decoding for Statistical Machine Translation (SMT). (Matsusaki et al. 2005) This paper defines a generative probabilistic model of parse trees, which we call PCFG-LA. This paper defines a generative model of parse trees that we call PCFG with latent annotations (PCFG-LA). (May and Knight 2006) We also demonstrate our algorithm's effectiveness ... to deal with grammars that produce trees. (Petrov et al. 2006) In this paper, we investigate the learning of a grammar consistent with a treebank at ... likelihood of the training trees. We present a method that combines the strengths of both manual and automatic approaches while addressing some of their common shortcomings. |

Table 5
An example of a summary generated by the our system with topic modeling applied.

| | ROUGE-L | | | ROUGE-SU4 | | |
|---|---|---|---|---|---|---|
| SYSTEM | R | P | F | R | P | F |
| Titles | 0.087 | 0.363* | 0.134 | 0.029 | 0.147* | 0.046 |
| AbsFS | 0.149 | 0.260 | 0.174 | 0.051 | 0.082 | 0.056 |
| AbsLS | 0.127 | 0.221 | 0.151 | 0.045 | 0.079 | 0.054 |
| SUMMA | 0.250 | 0.091 | 0.135 | 0.156 | 0.046 | 0.075 |
| MEAD | 0.269 | 0.117 | 0.168 | 0.198 | 0.034 | 0.071 |
| LexRank | 0.243 | 0.197 | 0.215 | 0.169 | 0.074 | 0.105 |
| TexRank | 0.282 | 0.117 | 0.172 | 0.196 | 0.025 | 0.060 |
| Babelnet | 0.345* | 0.203 | 0.252 | 0.228* | 0.112 | 0.147 |
| MJ | 0.273 | 0.219 | 0.240 | 0.215 | 0.135 | 0.162* |
| $CNN_{ROUGE-2-abstract}$ | 0.273 | 0.255 | 0.262 | 0.184 | 0.137 | 0.154 |
| $CNN_{AvgGAR-abstract}$ | 0.309 | 0.230 | 0.262 | 0.198 | 0.129 | 0.154 |
| $CNN_{AvgSGAR-abstract}$ | 0.302 | 0.229 | 0.258 | 0.199 | 0.132 | 0.156 |
| $CNN_{ROUGE-2-community}$ | 0.288 | 0.228 | 0.251 | 0.204 | 0.130 | 0.156 |
| $CNN_{ROUGE-2-human}$ | 0.273 | 0.246 | 0.256* | 0.185 | 0.136 | 0.154 |

Table 6
Automatic evaluation results of our systems against the baselines for ROUGE-L and ROUGE-SU4 metrics. Only the top 5 systems of the CNN approach are shown.

| | | ROUGE-L | | | ROUGE-SU4 | | |
|---|---|---|---|---|---|---|---|
| #SEN | SYSTEM | R | P | F | R | P | F |
| 1 | $CNN_{ACL-community}$ | 0.261 | 0.344 | 0.286 | 0.220* | 0.195 | 0.202 |
| | $CNN_{ACL-abstract}$ | 0.260 | 0.347* | 0.287 | 0.207 | 0.204* | 0.202 |
| | $CNN_{AvgSGAR-abstract}$ | 0.270* | 0.339 | 0.290* | 0.204 | 0.204* | 0.200 |
| | $CNN_{AvgGAR-community}$ | 0.60 | 0.344 | 0.286 | 0.204 | 0.200 | 0.199 |
| | $CNN_{AvgGAR-human}$ | 0.265 | 0.328 | 0.284 | 0.212 | 0.204* | 0.204* |
| 3 | $CNN_{AvgGAR-community}$ | 0.261 | 0.314 | 0.276 | 0.211 | 0.189 | 0.195 |
| | $CNN_{ROUGE-2-abstract}$ | 0.265 | 0.324 | 0.281 | 0.196 | 0.190 | 0.189 |
| | $CNN_{AvgSGAR-community}$ | 0.255 | 0.308 | 0.270 | 0.201 | 0.190 | 0.191 |
| | $CNN_{AvgGAR-human}$ | 0.258 | 0.304 | 0.271 | 0.192 | 0.185 | 0.186 |
| | $CNN_{AvgGAR-abstract}$ | 0.263 | 0.326 | 0.281 | 0.191 | 0.188 | 0.186 |
| 6 | $CNN_{AvgGAR-abstract}$ | 0.257 | 0.318 | 0.274 | 0.206 | 0.188 | 0.193 |
| | MJ | 0.253 | 0.305 | 0.269 | 0.217 | 0.189 | 0.197 |
| | $CNN_{ACL-human}$ | 0.245 | 0.300 | 0.262 | 0.207 | 0.185 | 0.191 |
| | Babelnet | 0.261 | 0.312 | 0.276 | 0.202 | 0.181 | 0.187 |
| | $CNN_{ROUGE-2-human}$ | 0.267 | 0.311 | 0.278 | 0.205 | 0.183 | 0.189 |

Table 7
Automatic evaluation results of paraphrasing the top sentences of reference papers. Only the top five abstractive models for ROUGE-L and ROUGE-SU4 metrics are reported.