Exploiting pivot words to classify and summarize discourse facets of scientific papers

Noname manuscript No.
(will be inserted by the editor)

Exploiting pivot words to classify and summarize discourse facets of scientific papers

Moreno La Quatra · Luca Cagliero · Elena Baralis

Received: date / Accepted: date

Abstract The ever-increasing number of published scientific articles has prompted the need for automated, data-driven approaches to summarizing the content of scientific articles. The Computational Linguistics Scientific Document Summarization Shared Task (CL-SciSumm 2019) has recently fostered the study and development of new text mining and machine learning solutions to the summarization problem customized to the academic domain. In CL-SciSumm, a Reference Paper (RP) is associated with a set of Citing Papers (CPs), all containing citations to the RP. In each CP, the text spans (i.e., citances) have been identified that pertain to a particular citation to the RP. The task of identifying the spans of text in the RP that most accurately reflect the citance is addressed using supervised approaches.

This paper proposes a new, more effective solution to the CL-SciSumm discourse facet classification task, which entails identifying for each cited text span what facet of the paper it belongs to from a predefined set of facets. It proposes also to extend the set of traditional CL-SciSumm tasks with a new one, namely the *discourse facet summarization* task. The idea behind is to extract facet-specific descriptions of each RP consisting of a fixed-length collection of RP's text spans. To tackle both the standard and the new tasks, we propose machine learning supported solutions based on the extraction of a selection of discriminating words, called *pivot words*. Predictive features based on pivot words are shown to be of great importance to rate the pertinence and relevance of a text span to a given facet.

The newly proposed facet classification method performs significantly better than the best performing CL-SciSumm 2019 participant (i.e., the classification accuracy has increased by +8%), whereas regression methods achieved promising results for the newly proposed summarization task.

Politecnico di Torino Corso Duca degli Abruzzi, 24 10129 Turin (Italy) E-mail: {name.surname}@polito.it

Keywords Discourse Facet Classification · Faceted Summarization · Classification and Regression · Deep Natural Language Processing

1 Introduction

Thanks to the advances in Information Systems, multimedia processing, and Web-based architectures, in the last decade digital libraries have significantly extended the volume of managed data, the number and kind of provided functionalities, and the user interfaces. They nowadays play a fundamental role in academic research. In fact, researchers can easily access the full-text of the most relevant scientific publications in electronic form. In parallel, scientometric data (e.g., author and co-author relationships, citation and co-citation information) have become easily accessible and usable for various purposes.

Scientists who are interested in enlarging their collaboration network can get in touch with other colleagues through various social platforms (e.g., ResearchGate, Academia.edu [46]), which recommend links and papers based on citation networks. Within the current scenario, however, the in-depth exploration of the content of scientific papers still remains extremely time-consuming, because it mainly relies on manual content retrieval [42].

The Computational Linguistics Scientific Document Summarization Shared Task (CL-SciSumm) [7] is an yearly research challenge focused on text mining and summarization of scientific papers. It fosters the joint analysis of paper full-text and scientometric data in order to gain insights into the analyzed content. More specifically, it aims at bringing together the summarization community to address challenges in scientific communication summarization. The tasks proposed in the fifth edition of the challenge, i.e., CL-SciSumm BIRNDL 2019, address an advanced analysis of a set of topics, where each topic consists of a Reference Paper (RP) and a set of Citing Papers (CPs), all containing citations to the RP. In each CP, the text spans, denoted as *citances*, have been identified that pertain to a particular citation to the RP. The main CL-SciSumm task (1A) is to identify the spans of text in the RP (namely, the cited text spans) that most accurately reflect the citance.

This paper addresses a specific task of CL-SciSumm BIRNDL 2019, namely the discourse facet classification (task 1B). The considered task entails identifying for each cited text span in RP what discourse facet it belongs to. Facets are selected from a predefined set (i.e., *Aim*, *Method*, *Results*, *Hypothesis*, *Implication*). We formulate the task as a classical binary classification problem, i.e., for each cited text span we predict whether a specific discourse facet is pertinent or not. The classifier is trained on a labeled dataset consisting of various cited and citing text spans' descriptors. Among the dataset features, the most discriminating ones for classification purposes rely on the concept of *pivot words*. In a nutshell, the occurrence of these particular words, extracted a priori from the training set, is likely to determine the discourse facet of a text span. The experiments carried out on benchmark data have shown that the proposed classification approach performs significantly better than the winner

of the CL-SciSumm 2019 contest (e.g., classifier accuracy has increased by +8%),

Once the discourse facets have been correctly identified, synthetic descriptions of the RP tailored to each facet would be desirable. Faceted summaries are deemed as useful for gaining insights into specific aspects covered by the paper. In fact, exploring per-facet summaries helps readers to explore particular aspects with limited human effort (i.e., without perusing the entire scientific paper). Hence, we propose to generate a fixed-length summary per facet consisting of the most salient text spans in RP pertaining to that facet. To our best knowledge[1] the discourse facet summarization task is new. We propose a machine learning solution based on regression methods. Specifically, we predict the overlap level between each text span in the RP and the expected per-facet summary. Notably, the results confirm the importance of using the pivot words to drive the summarization process (e.g., Rouge-2 F-measure between 0.30 and 0.40 on most of the analyzed facets). This reinforces the hypothesis that using the pivot words as text span descriptors is particularly effective in this particular scenario.

The innovative contribution of this paper can be summarized as follows:

- **A new machine learning approach**, based on the analysis of the pivot words, **to identifying the most likely discourse facet** of each cited text span in the RP.
- **A new facet summarization task**, which focuses on extracting summaries of the RP's text spans pertaining to a specific facet.
- **A machine learning strategy to solve the newly proposed facet summarization task**, which relies on regression models and also considers the presence of pivot words as predictive feature in the supervised models.

The rest of the paper is organized as follows. Sections 2 and 3 overview the related literature and the official CL-SciSumm 2019 Shared Tasks, respectively. Section 4 formalizes the newly proposed faceted summarization task. Section 5 thoroughly describes the presented method, while Section 6 summarizes the experimental results. Finally, Section 7 draws conclusions and discusses future works.

2 Related Works

This work is an extended version of the paper presented by [23]. The preliminary version describes a submission to the 2019 CL-SciSumm Shared tasks [7], which is based on an ensemble of traditional classification and regression models. The submission achieved very good results on summarization task 2 (i.e., 1st out of 104 runs against the community summary), whereas got halfway ranks on tasks 1A and 1B (i.e., 36th out of 98 submitted runs for task 1A,

[1] The formulated task has not been included among the official tasks of the CL-SciSumm challenges

57th over 98 submitted runs for task 1B). The achieved results have pushed us, on one hand, to deepen the search of effective solutions for a specific task (i.e., task 1B: discourse facet classification) and, on the other hand, to explore further extensions of the original problem related to the area of scientific paper summarization. Compared to our previous submission, this extended version (i) presents a new methodology to identify the facets of each cited text span of the RP, (ii) proposes a new task, called *discourse facet summarization*, (iii) describes a method to solve the newly proposed task.

A huge body of work has been devoted to proposing new summarization algorithms. Summarization focuses on identifying the most significant content from the body of a document [34]. Extractive methods pick existing parts of the original document, such as words, phrases, or sentences, whereas abstractive methods produce summaries including also newly generated content. Summarization techniques have already been applied to documents related to a variety of contexts, such as news articles [6,15,16], tweets [1,31], learning documents [5,11], and scientific papers [9,35]. The latter application context (scientific paper summarization) is the main target of this work. Notice that, unlike news articles, scientific papers are usually enriched with bibliographic references. In our context, a *citing* paper contains a reference to a given *cited* one. A subset of sentences of the citing paper pertaining to a given citation will be denoted as *citing sentences*, whereas the sentences in the cited paper that are most likely related to the citation will be denoted as *cited sentences*. Scientific papers can be summarized in many different ways: (i) by the abstract and title that the author provides [35,30,28], (ii) by the text snippets in the citing papers that pertain to it [43,32,40], (iii) by the text snippets including the most representative content in a semantic link network [47], (iv) by the presentation slides [47], or (v) by the highlight statements submitted along with the manuscript [10]. The contribution of this paper provides one step further towards the use of text spans in the cited papers surrounding a citation.

The task of summarizing a scientific paper using its corresponding set of citation sentences is called citation-based summarization [39]. It entails aggregating all the citation sentences that cite a paper and ranking them based on various criteria (e.g., coherence, readability) [2]. Unlike [2,39], in the CL-SciSumm Shared tasks and in the newly extension proposed in the present work the paper summary does not include citing sentence but only the sentences of the cited paper.

A pilot study of the Biomedical Summarization Track of the Text Analysis Conference 2014[2] indicates that most citations clearly refer to one or more specific aspects of the cited paper (usually Aim, Method, Result/Data, or Conclusion) [41]. Considering this insight could help analysts to create more coherent citation-based summaries. They recommend to identify the cited text span first, then identify the facet, and finally create a unique summary covering all the aspects. The newly proposed *facet summarization* task differs from

[2] `https://tac.nist.gov/2014/BiomedSumm/index.html`

those presented in [41] because it focuses on creating facet-specific, fixed-length summaries (a separate summary for each facet). Hence, each summary covers a separate aspect of the paper and provides deep knowledge about that aspect.

Citation-Sensitive In-Browser Summariser (CSIBS) systems (e.g.,[51, 52, 50]) generate facet-weighted previews, which vary according to the section of the text that the user is reading. This approach is complementary to the method proposed for the faceted summarization task, as the facet-specific summaries can be provided as detailed overview.

The participants to the previous editions of the Computational Linguistic Scientific Document Summarization Shared Tasks have proposed different solutions to automatically identify the discourse facets in the reference papers. For example, the winners of the 2018 challenge edition [19] have proposed to predict the syntactic distance between the candidate and actual cited text spans using regression models. Parallel attempts have been devoted to applying binary classifiers that combine various text embedding features [12, 37, 44]. For example, the approach presented in [37] has integrated contextualized word vectors trained on GoogleNews and on the ACL Antology Network collections. Alternative solutions adopted advanced word- or sentence-based distance metrics such as the Word Mover's Distance and the Earth Mover's Distance [24], the IDF-weighted Average Embedding based similarity, and the Smooth Inverse Frequency based similarity distances [14]. This paper proposes to use the concept of pivot words, proposed, with different formalizations, by [49, 13]. As shown in Section 6, the solution to the facet classification task (1B) presented in this paper performs significantly better than the winner of the 2019 CL-SciSumm Shared task 1B.

3 Computational Linguistics Scientific Document Summarization Shared Tasks

This section introduces the official tasks of the fifth edition of the Computational Linguistics Scientific Document Summarization Shared Task (i.e., CL-SciSumm 2019) [7]. The proposed challenges focus on text mining and summarization of scientific papers and follow up on the tasks proposed in the previous edition (CL-SciSumm 2018 [21]). The CL-SciSumm 2019 tasks' descriptions, the approaches proposed by the participants, and the achieved results have been presented at the joint workshop on Bibliometric-enhanced IR and NLP for Digital Libraries (BIRNDL@SIGIR 2019).

3.1 Description of the Data

The corpus of scientific papers released by the organizers of the CL-SciSumm 2019 Shared task consists of the full-text of 40 academic papers ranging over various topics (mostly from scientific research areas). The corpus is available online at `https://github.com/WING-NUS/scisumm-corpus`.

To support computational linguistics analyses, the text of each paper is partitioned into sentences (according to the presence of in-text punctuation marks) and sections (according to the original structure of the paper) [18]. Furthermore, the information about the citation network is known. Specifically, for each citation the citing and the cited (reference) papers are enriched with the following information: (i) The text spans in the citing paper that pertain to the citation (i.e., hereafter denoted as *citances*). (ii) The text span in the cited paper that are referenced by the citation. (iii) The discourse facets of the link between citing and cited text spans, which belong to the following set of predefined categories: *Aim, Hypothesis, Implication, Results, Method*. The text spans (i) and (ii) may include one or more consecutive sentences and may belong to many facets. Each paper contains, on average, 20 citations.

3.2 Official CL-SciSumm 2019 Tasks

The goal of the CL-SciSumm 2019 Shared Task is threefold. First, given a *Reference Paper*, it aims at identifying the text spans that are referenced by each of the corresponding citances (hereafter denoted as *task 1A*). It entails automatically linking the text spans belonging to the reference and citing papers, respectively. Secondly, for each cited text span it identifies the discourse facet it belongs to (*task 1B*). It entails addressing a multi-label text classification problem [48]. Lastly, it generates a unique summary per reference paper, consisting of its most salient text spans. It is a supervised text summarization problem, hereafter denoted as *task 2*. A more formal description of the proposed tasks is given below.

Notation. Let rp be a reference scientific paper and let CP be the set of scientific papers citing rp (hereafter denoted as *citing papers*). Given an arbitrary citing paper $cp \in CP$, let $c_{CP}=\{c_1, c_2, \ldots, c_n\}$ be the citances in cp that pertain to any citation to rp (i.e., the text spans where citations to rp are placed).

The CL-SciSumm summarization challenge is comprised of the following tasks

- **Task 1A: Cited text span identification**: For each citance c_j ($1 \leq j \leq n$), identify the spans of text $rp(c_j)$ in the reference paper (hereafter denoted as *cited text spans*) that are most likely to be pertinent to c_j. The cited text spans can be either a single sentence or a sequence of sentences.
- **Task 1B: discourse facet classification**: For each cited text span, identify the discourse facets it belongs to from a predefined set of facets.
- **Task 2: Reference paper summarization**: Produce a short summary of the reference paper (no longer than 250 words), consisting of a selection of the most salient text spans[3]. Although the text spans included in the

[3] This task was optional in the BIRNDL CL-SciSumm 2019 challenge.

summary are not necessarily referenced by any citance, the summarization process should be driven by the knowledge extracted at the previous steps (i.e., the citation links generated as output of task 1A and the facets discovered for task 1B).

The paper summaries produced as output of task 2 are compared with the following three different types of summaries: (i) the abstract summary of the reference paper, which was written by the authors of the research paper. (ii) the community summary, which collects all the referenced text spans in the reference paper. (iii) a human-written summary, written by the annotators of the CL-SciSumm annotation effort.

4 The Newly Proposed Task: Discourse Facet Summarization

We propose to extend the set of the Computational Linguistics Scientific Document Summarization Shared Tasks with a new one called *Discourse Facet Summarization*. In coherence with the notation used by the task proponents [7], we will also denote it as *task 2B*.

The goal is to produce a summary of the reference paper tailored to a specific discourse facet. Unlike in task 2, whose aim is to extract general-purpose summaries, in the newly proposed task each summary will consist of a selection of RP's text spans containing the key information related to a given discourse facet.

Similar to other summarization tasks, the idea behind is to provide a synthetic overview on the RP content. The peculiarity of the proposed task is the use of facet labels to drive the selection of the summary content. As mentioned in [34], facet-specific summaries can be explored to quickly grasp all aspects of a scientific paper with limited human effort.

Let $\mathcal{F}$ be the set of predefined facets (i.e., Aim, Hypothesis, Implication, Results, Method) and let rp be the reference paper. The discourse facet summarization task entails generating facet-specific, fixed-length summaries $\mathcal{S}(F_k)$ of rp, one for each facet in $\mathcal{F}$. $\mathcal{S}(F_k)$ consists of a selection of text spans in rp. Notice that the correct labeling produced by the facet classification is assumed to be unknown.

5 Presented methods

This section thoroughly describes the methods proposed to tackle both the official task 1B (cited text span classification) and the newly proposed task 2B (discourse facet summarization).

The section is organized as follows. Sections 5.1 and 5.2 describes the data preparation and feature engineering steps. Section 5.3 presents the new methodology, based on pivot word extraction, to address task 1B. Finally, Section 5.4 presents the regression-based approach proposed to address the new task 2B.

5.1 Text parsing and preprocessing

We process the text of the scientific papers and the related citation network to tailor the input data to the subsequent analyses. First, we perform a parsing of the text, provided in xml format, and store the structure of the text (e.g., the organization of the text into sentences and sections). Then, the input text is tokenized into separate words and the less relevant or non-informative words (e.g., conjunctions, prepositions) are removed. Word tokenization and stop-word removal were based on English vocabulary provided by the Natural Language Toolkit [4]. Such data are then used in the subsequent stages of the system both for the classification and the summarization task.

5.2 Feature engineering

The feature engineering step aims at generating a rich description of the cited and citing text spans. The selected features will be used to discriminate a sentence as pertinent and relevant for a given facet.

For each pair of cited and citing sentences the features for the classifier belongs to three different classes.

S_1: Relative position of the cited text span in the RP

S_2: Relative position of the citing text span in the CP

S_3: Cited text span length (expressed in number of words)

S_4: Presence of non-alphabetic characters in the cited text span

S_5: Section where the cited text span is placed (encoded using the IMRaD standard [45])

P_1-P_n: Number of pivot words peculiar to facet F_k [$1 \leq$ k $\leq$ n] that occur in the cited text span

P_{n+1}-P_{2n}: Number of pivot words peculiar to facet F_k [$1 \leq$ k $\leq$ n] that occur in the citing text span

P_1^{W2V}-P_n^{W2V}: Distance between the cited text span and the pivot words peculiar to facet F_k [$1 \leq$ k $\leq$ n] in the Word2Vec latent space [29] (computed using the Word Mover Distance (WMD) [22])

P_{n+1}^{W2V}-P_{2n}^{W2V}: Distance between the citing text span and the pivot words peculiar to facet F_k [$1 \leq$ k $\leq$ n] in the Word2Vec latent space [29] (computed using the Word Mover Distance (WMD) [22])

PR: PageRank [36] of the cited text span in the SciBERT vector similarity graph (vector similarity is computed using the cosine similarity).

Features denoted as S_* describe structural and syntactical features related to the text spans. They are considered as they provide basic textual properties.

Features denoted as P_* and P_*^{W2V} indicate the syntactic and semantic coherence of the text span with a set of pivot words generated using the state-of-art approach presented by [13]. Pivot words are collections of words from all sentences peculiar to a specific facet. Hence, their probability of occurrence within a text is likely to be related to the membership of the text to that facet. The idea behind using pivot words in facet classification and summarization is to identify the units of text that are most discriminating while deciding the membership of a text span to a given facet. We extract a separate set of pivot words for each facet. Features P_1-P_{2n} indicate the number of occurrences of the pivot word set in each text span, while features P_1^{W2V}-P_{2n}^{W2V} denote the semantic similarity between the pivot word set and the text span in the Word2Vec latent space. Using two complementary text representations (i.e., frequency-based and Deep NLP-based) allows us to capture both syntactic and semantic pertinence of the text span to the facet under analysis. To tailor word embeddings to the context of scientific papers, the embedding vector representation of text are trained on the large collection of scientific papers introduced in [10].

Feature PR indicates the centrality of a cited text span in the reference paper. It indicates the relative importance of a text span compared to all the other ones in the RP. The similarity graph consists of undirected weighted graph, where each node is a sentence in the RP while edges indicate that the two vector representations are similar in terms of SciBERT embedding vectors [3]. SciBERT is an established sentence embedding model trained on a large corpus of scientific documents. For each sentence, the vector representation is extracted considering last hidden layer of the SciBERT architecture. Pairwise vector similarity is computed using the cosine similarity [48] and indicates the semantic relatedness between the two text snippets. The similarity scores are normalized per paper to range in the interval $[0, 1]$. Figure 1 shows the distribution of the similarity scores between sentence pairs across the analyzed data collection, which fits, to a good approximation, a Gaussian distribution, with mean 0.75 and standard deviation 0.1. Notice that in the graph-based representation the edges representing least similar sentences are kept to guarantee PageRank algorithm convergence.

5.3 Discourse facet classification (Task 1B)

For each cited text span we decide whether to assign it to a given facet using a classification approach. Specifically, the data model described in Section 5.2 is used to train a multi-class classifier. Let rp be an arbitrary reference paper

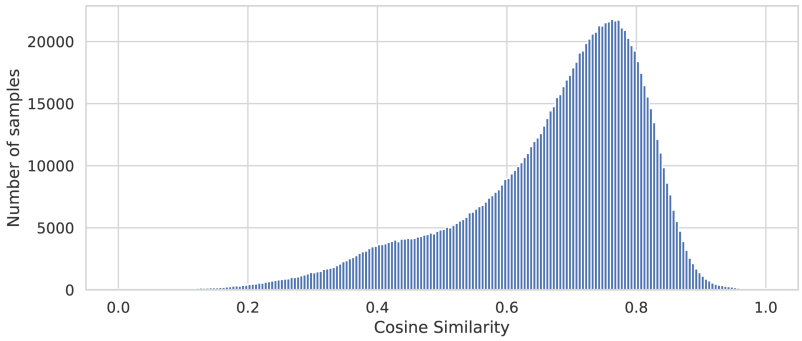


Fig. 1: Distribution of similarity scores between sentence pairs.

and let $\mathbf{C}$ be a target variable taking values in $\mathcal{F}$. The classifier is an arbitrary function $\mathcal{G}$: $rp \rightarrow \mathbf{C}$ indicating for each cited text span in rp the assigned facet.

To accomplish task 1B, we tested various classification methods (i.e., Gradient Boosting, Multi-Layer Perceptron, AdaBoost [48]). The considered classifiers are able to capture both linear and non-linear trends in the analyzed data.

5.4 Discourse facet summarization (Task 2B)

Given a facet F_k and a collection of reference papers, the facet summarization task entails learning a model able to extract a fixed-length, facet-specific summary of a new reference paper rp. The summary consists of a subset of rp's text spans pertaining to facet F_k.

Hence, we formulate the aforesaid problem as a regression task.

Regression target: For each text span of the reference paper we first compute the full set of features defined in Section 5.2. Then, we choose as target variable the level of overlap of the text span with the expected summary content using the Rouge-L precision [27].

Rouge is an established evaluation tool for summarization algorithms. It indicates the unit overlap (in terms of N-grams) between the generated and the expected summary. Rouge-L indicates the overlap between the longest common sub-sequence. Its use in evaluating supervised summarization techniques is established [10].

Objective function: The objective is to maximize the overlap between the text span and the expected summary.

Model application: We iteratively pick, in a greedy way, the top-ranked sentences (i.e., the one maximizing the predicted overlap score). To generate summaries of the same paper tailored to different facet, a separate regression model is trained for each facet.

Tested models: To accomplish task 2B, we used Linear Regression, Decision Tree Regressor, Random Forest Regressor, AdaBoost Regressor, Multi-Layer Perceptron Regressor, and Gradient Boosting Regressor [48].

6 Experimental results

We empirically analyzed the performance of the proposed approach on the benchmark dataset released for the CL-SciSumm 2019 Shared Task [17]. All the experiments were run on a machine equipped with Intel® Xeon® X5650, 32 GB of RAM and running Ubuntu 18.04.1 LTS.

6.1 Experimental design

We tested the performance of the proposed approach on benchmark data provided by the organizers of the CL-SciSumm 2019 Shared Task. Specifically, we considered the Training-Set-2018 collection (`https://github.com/WING-NUS/scisumm-corpus`), which consists of 40 annotated papers. We applied an 75%-25% hold-out validation strategy. Specifically, we divided it into two parts: 30 papers (along with the corresponding citation network) was used for training, whereas the remaining 10 for test. To make the results fully reproducible, we made the train-test dataset splits publicly available to the research community (see Section 6.6).

Algorithms. To accomplish both tasks we used the implementations available in the Scikit-Learn library[4] [38] and we tested various parameter settings to suit them to the underlying data distributions.

To extract the pivot words we replicated the methodology described in [13] to the best of our understanding starting from the implementation available in the official repository of the project[5].

Evaluation metrics. We evaluated the system performance on task 1B using three established information retrieval measures, i.e., precision, recall, and F-measure [25]. The *precision* is the fraction of cited text spans pertinent to the given facet among all the retrieved pertinent text spans, whereas the *recall*, is the fraction of pertinent text spans that have been retrieved over the total number of pertinent text spans. Finally, the *F-measure* is the harmonic mean of precision and recall. Since task 1B is a multi-class classification problem, we separately report (i) macro-average (i.e., the average over all the class), (ii) micro-average (i.e., the average computed considering both true positives, false negatives and false positives), and (iii) weighted average (i.e., the weighted average per class).

[4] `http://www.scikit-learn.org`

[5] `https://github.com/FranxYao/pivot_analysis`

Facet	Occurrences	Percentage
Method	563	69.6%
Results	105	12.9%
Aim	62	7.7%
Implication	60	7.4%
Hypothesis	18	2.2%

Table 1: Facet distribution in CL-SciSumm data collection.

To evaluate the performance of the summarization approach on the newly proposed task 2B, we exploited F-measure of the Rouge-2 which denote the overlap in terms of bi-grams [27]. For each facet we consider as reference summary the sentences of the community summary that have been classified as pertinent to the given facet.

6.2 Preliminary Data Characterization

We first analyzed the frequency of occurrence of the facet labels in the analyzed data. The distribution of the facets in the collection is reported in Figure 1. It shows that using predictive models to predict *Hypothesis* is practically unfeasible, because the number of training data is too limited. Hence, hereafter we will report the results achieved only on the other facets.

We analyzed also the importance of the input features in the classification process using a standard feature weighting method[6].

Figure 2 summarizes the importance of different features in the classification task. Notice that, since we address a multi-class problem, each feature has multiple relevance scores, one for each candidate facet. Hence, due to the inherent characteristics of the addressed task, the scores reported in Figure 2 are aggregated over all the facets.

The features related to the similarity with the pivot words in the Word2Vec space appear to be the most discriminating ones. This confirms the hypothesis that pivot words are potentially useful for classifying the cited text spans. However, notice that the simple count of the number of pivot words is weakly correlated with the target class as it is sensitive to the presence of noise in the input data.

6.3 Results on Discourse Facet Classification

We compare the performance of the proposed approach, namely *Pivot-Based Classification*, the with that of (i) the best performing approach presented in the latest edition (2019) of the CL-SciSumm Shared Task 1B, i.e., CIST [26][7]

[6] We exploit the *feature importance* function of Scikit-Learn to measure the relevance of each input feature to the classification phase.

[7] The approach presented by [26] has been re-implemented at the best of the authors' understanding.

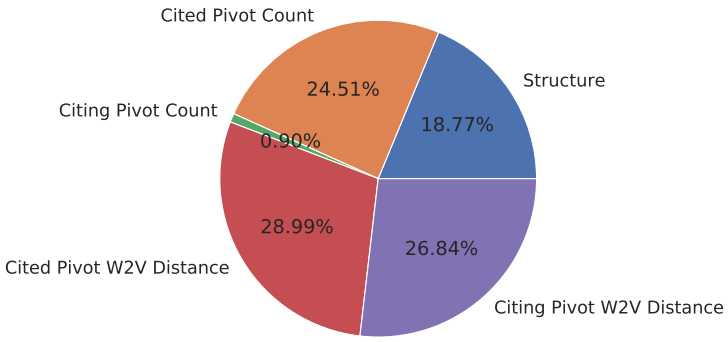


Fig. 2: Aggregated feature importance in facet classification.

and (ii) the previous version of the proposed method (Poli2Sum), presented in [23].

In Table 2 we report the results achieved on task 1B. The results were obtained by applying the model described in Section 5.3. The Weighted Average F-measure is around 8% higher than those of CIST and 14% higher than Poli2Sum [23]. This indicates that the classifiers based on pivot words are able to identify the given facets more accurately than previous approaches.

System	Type	Precision	Recall	F-Measure
Poli2Sum	Micro Average	0.59	0.59	0.59
	Macro Average	0.24	0.22	0.23
	Weighted Average	0.77	0.59	0.67
CIST	Micro Average	0.67	0.67	0.67
	Macro Average	0.24	0.24	0.23
	Weighted Average	0.81	0.67	0.73
Pivot-Based Classification	Micro Average	**0.77**	**0.77**	**0.77**
	Macro Average	**0.36**	**0.55**	**0.41**
	Weighted Average	**0.88**	**0.77**	**0.81**

Table 2: Comparison of different classification systems for the Task 1B.

6.4 Results on Discourse Facet Summarization

We have compared the performance of multiple summarization models, each one trained using a different regression method. Each model is trained to predict, as target value, the Rouge-L precision score with respect to the reference summary. Figure 3 reports the F-Measure Rouge-2 scores. The bars coloured in blue indicate the performance of the regression models trained on the full feature set (see Section 5.2). Conversely, the orange bars indicate the per-

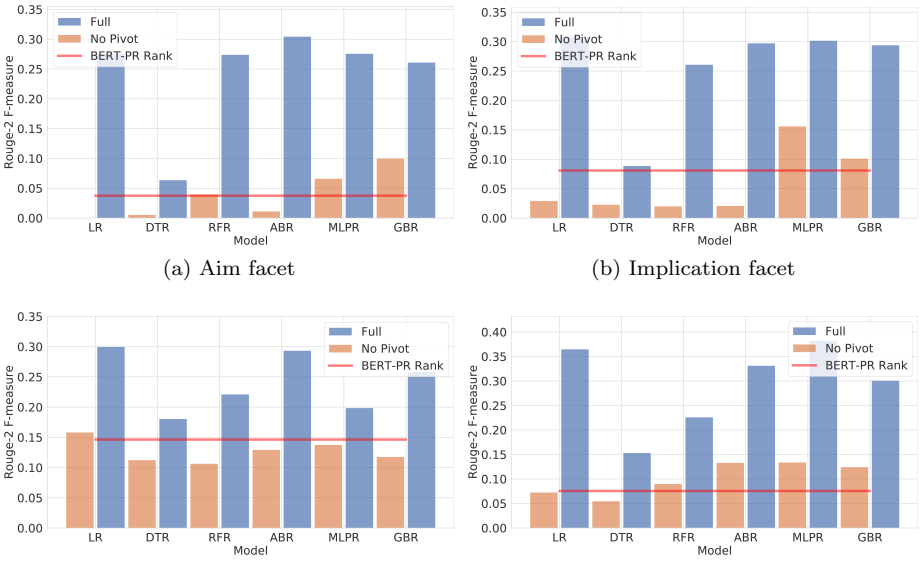
formance of the models trained by excluding the pivot word features. The comparison between blue and orange bars allows us to highlight the impact of pivot words on the summarizer performance. Excluding the pivot words results in a significant performance drop for all the considered facets and algorithms. This indicates that pivot words are of great importance even for tackling the summarization task.

In red we plotted also the performance of a baseline ranking method, which picks the top ranked sentences in order of decreasing PageRank score until the constraint on the maximum summary length is met. For the sake of clarity, in the plots we have reported only the Rouge score achieved by the best ranking function (i.e., the PageRank computed on the SciBERT vector similarity graph).

As expected, all the supervised methods outperform the baseline strategy. The MultiLayer Perceptron and the ensemble methods (ABR, RFR, GBR) achieved similar performance (Rouge-2 F-measure around 0.30).

6.4.1 Applicability of Neural Network models

In the previous evaluation we have considered just a simple fully-connected Neural Network Model, i.e., MultiLayer Perceptron (MLP). This kind of models require a relatively large number of training data to avoid data overfitting. To gain insights into the reliability of the Neural Network model, we have tested the performance of the simple MLP model on paper collections of var-



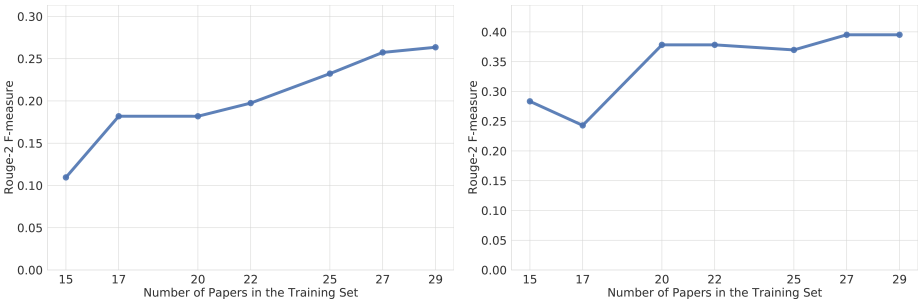
(a) Aim facet (b) Implication facet

(c) Method facet (d) Results facet

Fig. 3: Regression model comparison on task 2B. Rouge-2 F-measure.

ious size. The results, for the most frequent facets, are shown in Figure 4. By increasing the number of papers, the performance increases roughly linearly. Thus the models seem to be not affected by data overfitting.

Notice that as soon as a sufficient amount of training data will become available, various state-of-the-art deep summarization models (e.g., [8,33,20]) will become eligible for tackling the newly proposed summarization task.



(a) Method facet (b) Results facet

Fig. 4: Performance of MultiLayer Perceptron by varying the training set size. Task 2B. Rouge-2 F-measure.

6.5 Qualitative evaluation

We validated the outcomes of the facet classification and summarization processes with the help of a domain expert. Table 4 reports some examples of facet annotations produced by the classification strategy. For example, the *Method* annotation was assigned to a cited text span describing the features used in the proposed methodology. Notice that the citing text also refers to the feature model. The *Aim* annotation was correctly assigned to a sentence where the authors clarify the objectives of their research work.

Table 3 compares the automatically generated and expected summaries. The automatically selected content reflects, to a large extent, the meaning of the expected summary.

6.6 Contribution to the research community

To foster further contributions on the newly proposed facet summarization task, at `https://git.io/JvOe7` we have released (i) the list of pivot words extracted from the the Scisumm collection, (ii) the facet summaries generated from Scisumm, (iii) the automatic facet assignments from ScisummNet [53,54] (i.e., a larger collection of 1,000 papers, for which the manual annotations are missing).

Facet	Type	Summary
Method	Proposed Summary (AdaBoost)	Discovering Corpus-Specific Word Senses Following the method in (Widdows and Dorow, 2002), we build a graph in which each node represents a noun and two nodes have an edge between them if they co-occur in lists more than a given number of times 1. (Kilgarriff, 1992)). 7. Here we only mention a few direct results of our work. This gives rise to an automatic, unsupervised word sense disambiguation algorithm which is trained on the data to be disambiguated. section 2), and we plan to evaluate the uses of our clustering algorithm for unsupervised disambiguation more thoroughly. Let
	Reference summary	The algorithm is based on a graph model representing words and relationships between them. Sense clusters are iteratively computed by clustering the local graph of similar words around an ambiguous word. Following the method in (Widdows and Dorow, 2002), we build a graph in which each node represents a noun and two nodes have an edge between them if they co-occur in lists more than a given number of times 1. To detect the different areas of meaning in our local graphs, we use a cluster algorithm for graphs (Markov clustering, MCL) developed by van Dongen (2000).
Results	Proposed Summary (AdaBoost)	To investigate the influence of these factors, we analyze Modern Standard Arabic (henceforth MSA, or simply "Arabic") because of the unusual opportunity it presents for comparison to English parsing results. Particles are uninflected. of Arabic. pre-processing. It
	Reference summary	Better Arabic Parsing: Baselines, Evaluations, and Analysis To investigate the influence of these factors, we analyze Modern Standard Arabic (henceforth MSA, or simply "Arabic") because of the unusual opportunity it presents for comparison to English parsing results.

Table 3: Comparison between automatically generated and reference summaries for Method and Results facets. AdaBoost regression model.

7 Conclusions and future works

This paper addresses the problems of assigning facets to cited texts spans and summarizing the cited paper with text spans pertaining to a specific facet. The classification problem is part of the CL-Scisumm Shared tasks. We propose a new machine learning approach, based on the concept of pivot words, that has achieved performance superior to the best performing participant to the 2019 CL-Scisumm Shared Task 1B. The facet summarization problem addressed in this paper is new (to our best knowledge). We deem the newly proposed task as relevant to provide a facet-specific overview of the reference paper.

The results of the 2019 CL-Scisumm have highlighted the great potential of deep learning methods in tackling summarization problems tailored to the academic domain [7]. In light of these results, as a future work we plan to perform a semi-automatic, human-driven validation method that allows us to use the larger ScisummNet collection (1,000 papers [53, 54]). This would enable the use of various deep summarization methods to address task 2B.

Facet	Type	Text
Method	Citing	We used the base feature model defined in (Nivre et al, 2006) for all the languages but Arabic, Chinese, Czech, and Turkish.
	Cited	Features of the type DEPREL have a special status in that they are extracted during parsing from the partially built dependency graph and may therefore contain errors, whereas all the other features have gold standard values during both training and parsing.2 Based on previous research, we defined a base model to be used as a starting point for language specific feature selection.
Results	Citing	Gildea and Palmer (2002) achieve a recall of 0.50, a precision of 0.58, and an F-measure of 0.54 when using the full parser of Collins (1999).
	Cited	In fact, this system achieves 27.6% precision and 22.0% recall.
Implication	Citing	Unfortunately, parallel corpora are not readily available in large quantities, except for a small subset of the world's languages (see Resnik and Smith (2003) for discussion), therefore limiting the potential use of current SMT systems.
	Cited	Unfortunately, they are not readily available in the necessary quantities.
Aim	Citing	Second, we use the Twitter data set created by Ritter et al (2010)
	Cited	An initial conversation model can be created by simply applying the content modeling framework to conversation data.

Table 4: Examples of facet assignments. ScisummNet collection [53, 54].

Acknowledgements The research leading to these results has been partly funded by the Smart-Data@PoliTO center for Big Data and Machine Learning technologies. Computational resources were provided by HPC@POLITO, a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino (http://www.hpc.polito.it).

References

1. A. P. "Naik" and S. "Bojewar": Tweet analytics and tweet summarization using graph mining. In: 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), vol. 1, pp. 17–21 (2017). DOI {10.1109/ICECA.2017.8203674}
2. Abu-Jbara, A., Radev, D.: Coherent citation-based summarization of scientific papers. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, p. 500–509. Association for Computational Linguistics, USA (2011)
3. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3606–3611 (2019)
4. Bird, Steven and Klein, Ewan and Loper, Edward: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009)
5. Cagliero, L., Farinetti, L., Baralis, E.: Recommending personalized summaries of teaching materials. IEEE Access **7**, 22729–22739 (2019). DOI 10.1109/ACCESS.2019.2899655. URL `https://doi.org/10.1109/ACCESS.2019.2899655`
6. Cagliero, Luca and Garza, Paolo and Baralis, Elena: ELSA: A Multilingual Document Summarization Algorithm Based on Frequent Itemsets and Latent Semantic Analysis. ACM Trans. Inf. Syst. **37**(2), 21:1–21:33 (2019). DOI {10.1145/3298987}. URL `{http://doi.acm.org/10.1145/3298987}`
7. Chandrasekaran, M.K. and Yasunaga, M. and Radev, D. and Freitag, D. and Kan, M.-Y.: Overview and Results: CL-SciSumm SharedTask 2019. In: In Proceedings of the 4th

Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) @ SIGIR 2019, Paris, France. (2019)
8. Cheng, J., Lapata, M.: Neural summarization by extracting sentences and words. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 484–494. Association for Computational Linguistics, Berlin, Germany (2016). DOI 10.18653/v1/P16-1046. URL https://www.aclweb.org/anthology/P16-1046
9. Collins, E., Augenstein, I., Riedel, S.: A supervised approach to extractive summarisation of scientific papers. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 195–205. Association for Computational Linguistics, Vancouver, Canada (2017). DOI 10.18653/v1/K17-1021. URL https://www.aclweb.org/anthology/K17-1021
10. Collins, Ed and Augenstein, Isabelle and Riedel, Sebastian: A Supervised Approach to Extractive Summarisation of Scientific Papers. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 195–205 (2017)
11. Elena Baralis and Luca Cagliero: Highlighter: Automatic Highlighting of Electronic Learning Documents. "IEEE" Trans. Emerging Topics Comput. **6**(1), 7–19 (2018). DOI {10.1109/TETC.2017.2681655}. URL {https://doi.org/10.1109/TETC.2017.2681655}
12. Elnaz Davoodi and Kanika Madan and Jia Gu: CLSciSumm Shared Task: On the Contribution of Similarity measure and Natural Language Processing Features for Citing Problem. In: BIRNDL@SIGIR, *"CEUR" Workshop Proceedings*, vol. 2132, pp. 96–101. CEUR-WS.org (2018)
13. Fu, Y., Zhou, H., Chen, J., Li, L.: Rethinking text attribute transfer: A lexical analysis. In: K. van Deemter, C. Lin, H. Takamura (eds.) Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019, pp. 24–33. Association for Computational Linguistics (2019). URL https://aclweb.org/anthology/papers/W/W19/W19-8604/
14. Gaurav Baruah and Maheedhar Kolla: Klick Labs at CL-SciSumm 2018. In: BIRNDL@SIGIR, *"CEUR" Workshop Proceedings*, vol. 2132, pp. 134–141. CEUR-WS.org (2018)
15. George Giannakopoulos and Jeff Kubina and John M. Conroy and Josef Steinberger and Beno""̂ı""t Favre and Mijail A. Kabadjov and Udo Kruschwitz and Massimo Poesio: MultiLing 2015: Multilingual Summarization of Single and Multi-Documents, On-line Fora, and Call-center Conversations. In: Proceedings of the "SIGDIAL" 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic, pp. 270–274 (2015). URL {http://aclweb.org/anthology/W/W15/W15-4638.pdf}
16. Giannakopoulos, G.: Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop. In: Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pp. 20–28. Association for Computational Linguistics (2013). URL http://www.aclweb.org/anthology/W13-3103
17. Jaidka, K., Yasunaga, M., Chandrasekaran, M.K., Radev, D., Kan, M.Y.: The cl-scisumm shared task 2018: Results and key insights. arXiv preprint arXiv:1909.00764 (2019)
18. Jaidka, Kokil and Chandrasekaran, Muthu Kumar and Rustagi, Sajal and Kan, Min-Yen: Overview of the CL-SciSumm 2016 Shared Task. In: In Proceedings of Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries (2016)
19. Jaidka, Kokil and Yasunga, Michihiro and Chandrasekaran, Muthu and Radev, Dragomir and Kan, Min-Yen: The CL-SciSumm Shared Task 2018: Results and Key Insights. pp. 1–10 (2018)
20. Kedzie, C., McKeown, K., Daumé III, H.: Content selection in deep learning models of summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1818–1828 (2018)
21. Kumar Chandrasekaran, Muthu and Jaidka, Kokil and Mayr, Philipp: Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018). In: The 41st International ACM SIGIR Conference

on Research & Development in Information Retrieval, SIGIR '18, pp. 1415–1418. ACM, New York, NY, USA (2018). DOI {10.1145/3209978.3210194}. URL {http://doi.acm.org/10.1145/3209978.3210194}
22. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, p. 957–966. JMLR.org (2015)
23. La Quatra, M., Cagliero, L., Baralis, E.: Poli2sum@cl-scisumm-19: Identify, classify, and summarize cited text spans by means of ensembles of supervised models. pp. 233–246 (2019). URL https://www2.scopus.com/inward/record.uri?eid=2-s2.0-85071194418&partnerID=40&md5=e8f54672c3477c87a07010397cc60d28
24. Lei Li and Junqi Chi and Moye Chen and Zuying Huang and Yingqi Zhu and Xiangling Fu: CIST@CLSciSumm-18: Methods for Computational Linguistics Scientific Citation Linkage, Facet Classification and Summarization. In: BIRNDL@SIGIR, *"CEUR" Workshop Proceedings*, vol. 2132, pp. 84–95. CEUR-WS.org (2018)
25. Leskovec, Jure and Rajaraman, Anand and Ullman, Jeffrey David: Mining of Massive Datasets, 2nd edn. Cambridge University Press, New York, NY, USA (2014)
26. Li, L., Zhu, Y., Xie, Y., Huang, Z., Liu, W., Li, X., Liu, Y.: Cist@ clscisumm-19: Automatic scientific paper summarization with citances and facets. In: BIRNDL@SIGIR (2019)
27. Lin, Chin-Yew and Hovy, Eduard: Automatic evaluation of summaries using N-gram co-occurrence statistics. In: Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, pp. 71–78 (2003)
28. Lloret, E., Romá-Ferri, M.T., Palomar, M.: Compendium: A text summarization system for generating abstracts of research papers. Data & Knowledge Engineering **88**, 164 – 175 (2013). DOI https://doi.org/10.1016/j.datak.2013.08.005
29. Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119 (2013)
30. Minsoo Kim and Dennis Singh Moirangthem and Minho Lee: Towards Abstraction from Extraction: Multiple Timescale Gated Recurrent Unit for Summarization. In: Rep4NLP@ACL, pp. 70–77. Association for Computational Linguistics (2016)
31. Naik, S., Lade, S., Mamidipelli, S., Save, A.: Tweet summarization: A new approch. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 1022–1025 (2018). DOI 10.1109/ICICCT.2018.8473327
32. Nakov, P.I., Schwartz, A.S., Hearst, M.A.: Citances: Citation sentences for semantic analysis of bioscience text. In: In Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics (2004)
33. Nallapati, R., Zhai, F., Zhou, B.: Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, p. 3075–3081. AAAI Press (2017)
34. Nenkova, Ani and McKeown, Kathleen: A survey of text summarization techniques. In: Mining text data, pp. 43–76. Springer (2012)
35. Nikolov, Nikola I and Pfeiffer, Michael and Hahnloser, Richard HR: Data-driven Summarization of Scientific Articles. In: Proc. of the 7th International Workshop on Mining Scientific Publications, LREC 2018 (2018)
36. Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry: The PageRank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
37. Pancheng Wang and Shasha Li and Ting Wang and Haifang Zhou and Jintao Tang: "NUDT" @ CLSciSumm-18. In: Proceedings of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries "(BIRNDL" 2018) co-located with the 41st International "ACM" "SIGIR" Conference on Research and Development in Information Retrieval "(SIGIR" 2018), Ann Arbor, USA, July 12, 2018., pp. 102–113 (2018)
38. Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and

Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E.: Scikit-learn: Machine Learning in "P"ython. Journal of Machine Learning Research **12**, 2825–2830 (2011)
39. Qazvinian, V., Radev, D.R.: Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 689–696. Coling 2008 Organizing Committee, Manchester, UK (2008). URL https://www.aclweb.org/anthology/C08-1087
40. Qazvinian, V., Radev, D.R.: Identifying non-explicit citing sentences for citation-based summarization. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, p. 555–564. Association for Computational Linguistics, USA (2010)
41. Ronzano, F., Saggion, H.: An empirical assessment of citation information in scientific summarization. In: E. Métais, F. Meziane, M. Saraee, V. Sugumaran, S. Vadera (eds.) Natural Language Processing and Information Systems, pp. 318–325. Springer International Publishing, Cham (2016)
42. Saggion, H., Ronzano, F.: Scholarly data mining: Making sense of scientific literature. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1–2 (2017). DOI 10.1109/JCDL.2017.7991622
43. Schwartz, A.S., Hearst, M.: Summarizing key concepts using citation sentences. In: Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology, LNLBioNLP '06, p. 134–135. Association for Computational Linguistics, USA (2006)
44. Shutian Ma and Jin Xu and Chengzhi Zhang: Automatic identification of cited text spans: a multi-classifier approach over imbalanced dataset. Scientometrics **116**(2), 1303–1330 (2018)
45. Sollaci, L.B., Pereira, M.G.: The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey. Journal of the medical library association **92**(3), 364 (2004)
46. Steven Ovadia: ResearchGate and Academia.edu: Academic Social Networks. Behavioral & Social Sciences Librarian **33**(3), 165–169 (2014). DOI {10.1080/01639269.2014.934093}
47. Sun, X., Zhuge, H.: Summarization of scientific paper through reinforcement ranking on semantic link network. IEEE Access **6**, 40611–40625 (2018). DOI 10.1109/ACCESS.2018.2856530
48. Tan, P.N., Steinbach, M., Karpatne, A., Kumar, V.: Introduction to Data Mining (2nd Edition), 2nd edn. Pearson (2018)
49. Wan, S., Dale, R., Dras, M., Paris, C.: Seed and grow: Augmenting statistically generated summary sentences using schematic word patterns. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 543–552 (2008)
50. Wan, S., Paris, C., Dale, R.: Whetting the appetite of scientists: producing summaries tailored to the citation context. In: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, pp. 59–68. ACM (2009)
51. Wan, S., Paris, C., Dale, R.: Invited paper: Supporting browsing-specific information needs: Introducing the citation-sensitive in-browser summariser. Web Semant. **8**(2–3), 196–202 (2010). DOI 10.1016/j.websem.2010.03.002. URL https://doi.org/10.1016/j.websem.2010.03.002
52. Wan, S., Paris, C., Muthukrishna, M., Dale, R.: Designing a citation-sensitive research tool: An initial study of browsing-specific information needs. In: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL), pp. 45–53. Association for Computational Linguistics, Suntec City, Singapore (2009). URL https://www.aclweb.org/anthology/W09-3606
53. Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A., Li, I., Friedman, D., Radev, D.: ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In: Proceedings of AAAI 2019 (2019)
54. Yasunaga, M., Zhang, R., Meelu, K., Pareek, A., Srinivasan, K., Radev, D.R.: Graph-based neural multi-document summarization. In: Proceedings of CoNLL 2017 (2017)