Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6672–6679
Marseille, 11–16 May 2020

A Multi-level Annotated Corpus of Scientific Papers for Scientific Document Summarization and Cross-document Relation Discovery

Ahmed AbuRa'ed[1], **Horacio Saggion**[1], **Luis Chiruzzo**[2]
[1]Large Scale Text Understanding Systems Lab
TALN Research Group
Universitat Pompeu Fabra, Barcelona, Spain
[2]Universidad de la República, Montevideo, Uruguay
ahmed.aburaed@upf.edu, horacio.saggion@upf.edu, luischir@fing.edu.uy

Abstract

Related work sections or literature reviews are an essential part of every scientific article being crucial for paper reviewing and assessment. The automatic generation of related work sections can be considered an instance of the multi-document summarization problem. In order to allow the study of this specific problem, we have developed a manually annotated, machine readable data-set of related work sections, cited papers (e.g. references) and sentences, together with an additional layer of papers citing the references. We additionally present experiments on the identification of cited sentences, using as input citation contexts. The corpus alongside the gold standard are made available for use by the scientific community.

Keywords: Human Annotated Corpora, Text Summarization, Cross-document Relations, Related Work Sections, Literature Reviews

1. Introduction

Most scientific papers include a related work section providing, in a well organized and condensed form, the key information from a carefully selected list of publications which contextualize and ground the research being presented by an author (Rowley and Slack, 2004). Related work sections are critical for quality assessment since journals pay particular attention to them where evaluation of manuscripts is of concern (Maggio et al., 2016). Past research has shown that related work sections can be produced following cut-and-paste summarization strategies (Jaidka et al., 2013) which are typical of document abstracting (insertion, deletion, substitution, etc.) (Endres-Niggemeyer et al., 1995; Saggion, 2011).

Recent studies have proposed to take advantage of the scientific paper's citation network to approach scientific literature summarization. For that reason we introduce here our corpus which we hope will facilitate the usage of citation networks to boost scientific literature summarization research. The generation of related work sections has been studied from different viewpoints (Hoang and Kan, 2010; Vu, 2010; Hu and Wan, 2014), however no manual annotated data-set, analog to the one we will present here, has been produced until now.

Our corpus expands considerably the data-set of related work sections used in (Hoang and Kan, 2010) 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. While the manually identified cited sentences are useful to support the study of sequence to sequence models in scientific summarization, the new layer of citing papers facilitates the test of citation-based summarization approaches (Qazvinian and Radev, 2008; Jaidka et al., 2014b) which rely on citation networks to assess sentence relevance.

In this corpus we refer to three types of scientific papers: target papers, reference papers, and citing papers which we organize in a two-level network. Level 1 contains target papers with their related work sections we are interested in and, 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.

The rest of this article is organized as follows: The next Section describes related work, then in Section 3 the initial data set is described. Section 4 explains how we extended the initial data set to form our corpus, alongside the data collection process and automatic processing of the data. Section 5 explains the manual annotation process and reports inter-annotator agreement. Then, Section 6 describes several experiments carried out to simulate the retrieval of sentences matching citation contexts and, finally, Section 7 closes the paper with conclusions.

2. Related Work

Good related work sections are difficult to produce since they require the author to select, contrast, and organize key information from several sources. Although there have been a number of studies and guidelines on their functions, types and forms (S. G. Khoo et al., 2011; Jaidka et al., 2013; Pautasso, 2013), our understanding of what is a good related work section is still limited. It is generally agreed that related work sections or literature reviews can either be descriptive or integrative (S. G. Khoo et al., 2011; Jaidka et al., 2013). While a descriptive report will summarize individual papers providing information such as methods and results in citation sentences, integrative reports will focus on key ideas and topics providing in the citation sentences critical views on the presented approaches.

There is a number of corpora related to the work presented here. A large-scale, human-annotated scientific papers corpus is provided by (Yasunaga et al., 2019). It provides over 1,000 papers in the ACL anthology with their citation networks (e.g. citation sentences, citation counts) and their comprehensive, manual summaries. There is also a data-set which has been created for the Computational Linguistics Scientific Document Summarization Shared Task which started in 2014 as a pilot (Jaidka et al., 2014a) and which is now a well developed challenge in its fourth year (Jaidka et al., 2017b; Jaidka et al., 2017a). The shared task provided training data structured in clusters of reference and citing papers together with manual annotations indicating, for each citance, the text span(s) in the reference paper that best represent the citance, as well as their corresponding facets. One of the main problems with the data-set is the lack of agreed manual annotations since only one annotator was in charge of annotating each cluster. Those previously mentioned data-sets are considered the closest to our corpus however they are only equivalent to what we name Level 2 of our corpus and they provide no link between a target paper with a segmented related work section that explicitly mention a set of reference papers.

There are also corpora for the study of scientific text mining and summarization. (Saggion and Lapalme, 2002) have aligned 200 abstracts produced by professional abstractors to their source documents to investigate how to produce non-extractive indicative abstracts. (Fisas et al., 2016) have created a multi-layered annotated corpus from 40 articles in the domain of Computer Graphics. Sentences are annotated with respect to their role in the argumentative structure of the discourse. It specifies the purpose of each citation in the scientific papers and it identifies special features of the scientific discourse such as advantages and disadvantages. In addition, a grade is allocated to each sentence according to its relevance for being included in a summary. (Athar and Teufel, 2012) created a citation context corpus from the ACL Anthology Network (AAN) which consists of 852 papers that are citing 20 papers. The corpus contains 1,034 paper–reference pairs and 203,803 sentences. It is manually annotated by identifying the sentences in the citation context. It also contains a sentiment annotation as well (negative, positive, objective/neutral). (Teufel, 2006) created a corpus based on 80 Argumentative Zoning-annotated conference articles in the computational linguistics domain. The corpus was created to research classifying academic citations in scientific articles according to author claims.

Finally, based on the SAPIENT tool (Liakata et al., 2009) and an annotation guideline (Liakata and Soldatova, 2008) a corpus of 225 papers was created and manually annotated with CISP (Core Information about Scientific Papers) concepts. These papers cover topics in physical chemistry and biochemistry. The Corpus was developed to add value to scientific papers through semantic markup.

3. The Initial Data-Set

The RWSData data-set (Hoang and Kan, 2010) is a publicly available resource that includes twenty articles from sources such as the Special Interest Group on Information Retrieval (SIGIR), the Association for Computational Linguistics (ACL), the North American Chapter of the Association for Computational Linguistics (NAACL), the Empirical Methods for Natural Language Processing (EMNLP) and the International Conference on Computational Linguistics (COLING). (Hoang and Kan, 2010) extracted the related work sections directly from those research articles as well as several references cited in the related work sections. All the scientific papers provided in the RWSData are in PDF format with no further analysis. Moreover, the data-set provides no mapping between the related works section citations and the sentences in the reference papers that are being cited making it challenging to use such data-set for scientific papers summarization.

4. Corpus Extension

We extracted the same twenty target papers considered in (Hoang and Kan, 2010), then for each paper we collected the reference papers mentioned in its related work section. Afterwards, for each reference paper we collected multiple scientific papers citing it. This extra set of citing scientific papers could be used as a citation network which could allow citation network summarization systems to be implemented over the extended corpus.

Figure 1 shows a target paper containing a related work section alongside the reference papers which it cites and, in turn, for each reference paper, a set of scientific papers citing it. The RWSData data-set is the raw data on level 1 while our extension added the citing papers for the reference papers and the (manually identified) links between citing and cited sentences.

4.1. Data Collection

The extension of the corpus was done by adding citing papers for each reference paper. The citing papers were collected from Microsoft Academic Graph (MAG) (Tang et al., 2008; Sinha et al., 2015; Wade, 2015; Herrmannova and Knoth, 2016), Semantic Scholar (Xiong et al., 2017; Valenzuela et al., 2015) and the ACL Anthology Network (AAN) (Radev et al., 2013). We queried the APIs of both Semantic scholar and Microsoft Academic Graph in order to obtain detailed information for the scientific papers. Microsoft Academic Graph (MAG) (Tang et al., 2008) is a diverse graph containing scientific publication records, citation relationships between those publications, as well as metadata. Semantic Scholar (Valenzuela et al., 2015) is a publicly available search service with millions of indexed articles. Semantic Scholar identifies citations where the cited publication has a significant impact on the citing publication, making it easier to understand how publications build upon and relate to each other. It also has what is named "influential citations" which are determined by using a machine-learning model analyzing a number of factors including the number of citations to a publication, and the surrounding context for each (Valenzuela et al.,

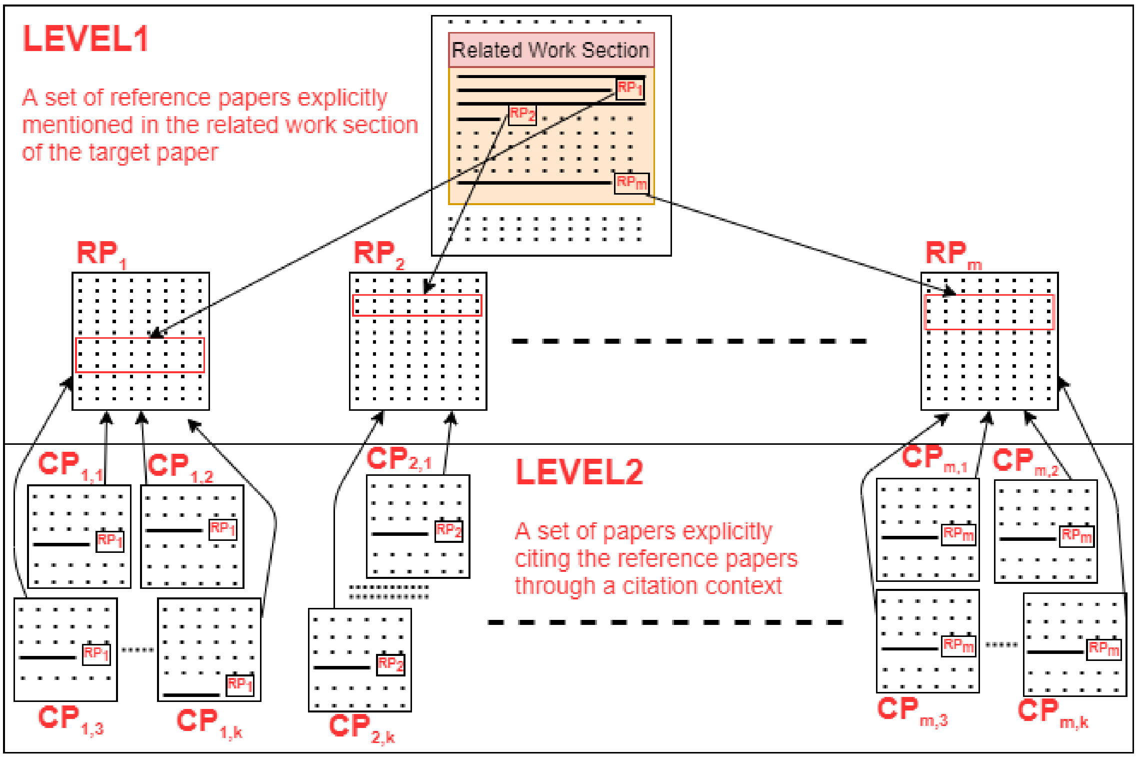
LEVEL1
A set of reference papers explicitly mentioned in the related work section of the target paper
Related Work Section
RP1
RP2
RPm
RP1
RP2
RPm
CP1,1
CP1,2
CP1,3
CP1,k
CP2,1
CP2,k
CPm,1
CPm,2
CPm,3
CPm,k
LEVEL2
A set of papers explicitly citing the reference papers through a citation context

Figure 1: Our corpus outline presenting a target paper, a set of reference papers (Level 1) and for each reference paper a set of citing papers (Level 2)

2015). The ACL Anthology Network (AAN) (Radev et al., 2013) is a wide-range manually curated networked database of citations, collaborations, and summaries in the field of Computational Linguistics. AAN provides citation and collaboration networks of the articles included in the ACL Anthology (Bird et al., 2008) (excluding book reviews). The order of querying the data sources was first Semantic Scholar, then MAG and, finally, ACL. The citing papers were collected from the same source where the reference paper was found. We kept the most cited or most influential papers depending on the source from where the papers were collected. Overall, we collected up to 15 citing papers for each reference paper (with an average of 12 per reference paper).

4.2. Corpus Basic Data Processing

We converted the PDF documents for the entire corpus into GATE (Maynard et al., 2002) documents using three converters; Grobid (Lopez, 2009), PDF Digest (Ferrés et al., 2018) and PDFX (Constantin et al., 2013). The three converters provide basic information about each scientific paper contents including: title, authors, affiliations, abstract and paper sections. Finally, we also identified the sentences of each scientific paper by annotating a sentence ID for the GATE documents, this ID was used to help map the sentences during the annotation process.
Table 1 provides information about the different paper types: Target Papers (TP), Reference Papers (RP) and Citing Papers (CP). It shows the number of papers, sentences and tokens alongside their averages.

5. Annotation Process

The corpus is designed to be used by scientific papers summarization systems including those based on citation networks. We manually annotated the relationship between the target papers and the reference papers (See the upper half of Figure 1). These annotations provide a mapping between the related work section and the texts fragments which are considered semantically close to the citation sentence in the reference papers. In order to facilitate the annotation process, we also provided the citation context computed using the state-of-the-art approach described in (AbuRa'ed et al., 2018b).

Three annotators with expertise in computational linguistics carried out the annotation, one of the annotators is the first author of the paper, the other two were hired for the project. Annotators were asked to identify which parts of the reference papers (one or more sentences) have been cited by the citing target papers by means of the citation sentence. We used the open-source, web-based text mining tool WARP-Text (Kovatchev et al., 2018) for the manual annotations process since it allows for annotating relationships between pairs of texts. We customized the tool to perform annotations at a sentence level.

We organized the annotation process in screens showing a citation context and a set of sentences representing the cited reference paper to choose from, see Figure 2. The annotator will then select which of the reference paper sentences best reflect the citation context of the citing

Paper Type	#	#Sentences	avg#Sentences	#Tokens	avg#Tokens
TP	20	8,151	407.55	148,732	7,436.60
RP	222	73,225	329.84	1,285,168	5,789.04
CP	2,216	829,003	374.10	15,073,031	6,810.90

Table 1: Corpus statistics presenting information about the different paper types: Target Papers (TP), Reference Papers (RP) and Citing Papers (CP). It presents the number of papers, sentences and tokens alongside their averages.

target paper. See Figure 3 for a citation/cited sentence pair.

The annotation process was straight forward, the web page allowed multiple sentences selection and once a sentence was selected by an annotator it was highlighted. After an annotator selected all the sentences from a reference paper that he or she believed were reflecting the citation context, he or she could then record the annotations. In cases where a screen presents more than a citation marker in the citing paper side only the target citation would be capitalized to avoid confusion. Finally, the scientific papers names and titles were also visible on the screens. The annotators had also access to all PDF articles which they regularly used to ground their decisions.

We divided the corpus into 5 batches: first batch was aimed to get an initial feedback from the annotators, it contained only one target paper's related work with the references mentioned in it. Second batch has 4 target papers with their references and the last 3 batches each contained 5 target papers with their references.

The annotation process was iterative: once a batch was finished we got feedback from the annotators, computed agreement and we improved annotation recommendations and display accordingly. For example, after the first batch, we realized that furnishing all of the sentences of a reference paper at once over one screen was inconvenient for annotation. Therefore, we decided to filter out non-relevant sentences and to divide the rest of the sentences of the reference paper over more than one screen where each screen contains maximum 15 sentences. We used the work done by (Abura'ed et al., 2018a) to filter out unrelated sentences and keep the ones that are most similar to the citation context. All sentences were also available by consulting the original paper in PDF in case no suitable match was found.

5.1. Inter-Annotator Agreement

We used Cohen's kappa coefficient (Cohen, 1960) in order to measure the inter-annotators agreement for each target paper with the reference papers mentioned in it. During the annotation process throughout the 5 batches there were some conflicts between the annotators. We held meetings to address the conflicts in which we presented the annotators with a list of pairs presented by the tool sorted by agreement from worst to best. Going through the list of less agreed to most agreed papers and discussing the reasons that could lead to such disagreement improved the annotation process. One of the annotators was more likely to select sentences which included definitions or background information not reflected in the citations but which she considered important for her understanding of the paper. Situations like these made higher agreement levels difficult to achieve, that is why the meetings helped to better clarify what information to search for.

Table 2 reports the pair-wise agreement as well as the average of Cohen's kappa results over the entire corpus. The agreement level $\kappa > 0.5$ indicates moderate agreement between the annotators. The final corpus contains the cited sentences which were selected by majority agreement.

A_1 **&** A_2	A_1 **&** A_3	A_2 **&** A_3	**Average**
0.64	0.57	0.35	0.52

Table 2: Pairwise and Average Inter-annotator Agreement

6. Experiments

In order to identify relevant sentences for writing a related work section, it is first important to know which sentences in a citing paper contain relevant information. We have implemented several automatic systems to simulate the annotators' task casting the problem as one of retrieving sentences which better reflect the citation and its context. For this purpose, we have also enriched the corpus with annotations relevant for scientific text processing in the hope to make it easier for additional related tasks (these annotations are being made available).

6.1. Corpus Enrichment

Each GATE document was annotated using processing resources from the GATE system (Maynard et al., 2002; Cunningham et al., 2002), the SUMMA library (Saggion, 2008), and the freely available Dr Inventor library (DRI Framework) (Ronzano and Saggion, 2015). The tools semantically enrich the corpus by providing rhetorical annotation, causality identification, coreference, and BabelNet synsets. The SUMMA library was used to produce different normalized term vectors for each document. Vector of terms and BabelNet synsets are created using tf*idf weighting computed from a corpus of 4K ACL scientific papers. Using 58 gazetteer lists created from the lexicons proposed by (Teufel and Moens, 2002) we identified scientific concepts and actions useful for text summarization.

The corpus is available for research and development purposes in two versions[1], one version contains the manual

[1] http://taln.upf.edu/sciencecorpus

Citing Paper: C08-1031: Mining Opinions in Comparative Sentences	
Cited Paper: Fiszman-et-al-2007: Interpreting Comparative Constructions in Biomedical Text	
Citation	FISZMAN ET AL (2007) studied the problem of identifying which entity has more of certain features in comparative sentences.
PAGE 4 of 4	
Cited Paper	*55: In our sample, expressions interpreted as empty heads include those referring to drug dosage and formulations, such as extended release (the latter often abbreviated as XR).* *56: Examples of missed interpretations are in sentences (28) and (29), where the empty heads are in bold.* *57: These mechanisms are being incorporated into the processing for comparative structures.* *58: 6 CONCLUSION* *59: We expanded a symbolic semantic interpreter to identify comparative constructions in biomedical text.* *60: The method relies on underspecified syntactic analysis and domain knowledge from the UMLS.* *61: We identify two compared terms and scalar comparative structures in MEDLINE citations.*

Figure 2: Schematic View of the Data during the Annotation Process (on top a citation sentence in a related work section, in the bottom, sentences from the cited paper i.e. reference paper)

Citing Paper: C08-1031: Mining Opinions in Comparative Sentences	
Cited Paper: Fiszman-et-al-2007: Interpreting Comparative Constructions in Biomedical Text	
Citation	FISZMAN ET AL (2007) studied the problem of identifying which entity has more of certain features in comparative sentences.
PAGE 4 of 4	
Cited paper sentences	We expanded a symbolic semantic interpreter to identify comparative constructions in biomedical text.

Figure 3: Sentences Selected by Annotator Matching a Citation in the Related Work Section

annotations (agreed cited sentences) and the other the full machine readable corpus with the automatic analysis just described.

6.2. Automatic Systems

We implemented several automatic systems in which we provide them with a citation context from a related work section in a citing target paper and retrieve the reference sentences sorted by the most similar to the citation context. The systems are as follows:

- Google News: Using a collection of 300 dimensional `word2vec` embeddings trained over a corpus of 100 billion words from Google News[2], this heuristic calculates the centroid of each sentence in the reference and compares it to the centroid of the citing sentence, and returns the most similar ones according to cosine similarity.

- ACL: Similar to the previous case, the heuristic calculates the centroids using 100 dimensional vectors from the ACL Anthology Reference Corpus embeddings (Liu, 2017) trained over a corpus of ACL papers (Bird et al., 2008).

- Google + ACL: The same as before, but using the concatenation of Google News and ACL vectors, creating 400 dimensional vectors. When a word was not present in either of the embeddings collections, it was replaced by a null vector of equivalent size.

- BabelNet: This heuristic first obtains the BabelNet synsets present in each sentence using the Babelfy API[3] and creates an embedding for the sentence by averaging the embeddings of each synset from a BabelNet 300 dimensional embeddings collection trained over a corpus of 300 million words tagged with BabelNet synsets (Mancini et al., 2016), then returns the sentences sorted by cosine similarity.

- SUMMA normalized vectors: In this case we model each sentence as the vector of normalized tf-idf values for each of the terms in the sentence, calculating the frequencies in an ACL reference corpus of around 4,000 papers. We compare the citing sentence to all sentences in the reference and return the results sorted by cosine similarity.

- Modified Jaccard: This heuristic uses a metric similar to the Jaccard similarity coefficient for comparing the citing sentence to each sentence of the reference paper. This version of the metric (AbuRa'ed et al., 2017) considers the union and intersection of words (like the Jaccard coefficient) but also includes information about the inverted frequency to give more weight to words in the intersection that are less common.

6.3. Results

We used Precision at k ($P@k$) (Sujatha and Dhavachelvan, 2011) to evaluate the task of selecting the sentences in each reference paper that best reflects the content expressed in the citation context from the related work section of the target scientific paper. See Table 3 to see Precision at positions 1 to 5. This is a hard task due to the large number of sentences a scientific paper has and the natural difference between citing/cited papers because of the rephrasing characteristics of cited sentences.

[2] `https://code.google.com/archive/p/word2vec/`

[3] `http://babelfy.org/guide`

In this sense results are not surprising, it can be noticed that the BabelNet system is the best one which may indicate that comparing sentences by semantic similarity (instead of lexical) is a good option for achieving good result. Worst results are achieved by systems which use more superficial representations based on words or lemmas. Word embedding perform better than superficial representations, still worst than semantics based on lexical resources, and embedding combinations shows positive improvements.

System	P@1	P@2	P@3	P@4	P@5
ACL	0.1213	0.1416	0.1388	0.1362	0.1369
Babelnet	0.1934	0.1844	0.1852	0.1789	0.1776
Google	0.1593	0.1361	0.1422	0.1344	0.1228
G+ACL	0.1653	0.1428	0.1470	0.1421	0.1321
MJ	0.0988	0.0957	0.0887	0.0878	0.0794
SUMMA	0.0609	0.0590	0.0498	0.0473	0.0478

Table 3: Average Precision for the automatic systems at position 1 to 5

7. Conclusion

In this paper, we have presented a corpus in the field of scientific text mining and summarization to allow the study of automatic related work text generation. The corpus provides related work sections of scientific papers, a manually annotated layer of referenced cited papers, a level of citing papers referring to the cited papers in the related work section, and a layer of rich linguistic, rhetorical, and semantic annotations computed automatically.

We also present initial experiments to assess several text representation mechanisms (e.g. lemmas, embeddings, synsets) for the retrieval of sentences likely to be cited by scientific papers comparing system results to the gold standard annotations. The manually annotated corpus with its automatically enriched documents is being made available for the community.

We hope this corpus would provide the research community means for fair comparisons of various summarization approaches. Considering recent work in citation-based summarization, our future work will consider the use of this corpus for the generation of automatic related work sections given the reference papers and their citation networks.

8. Acknowledgments

This work is (partly) supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502). We thank the reviewers for their remarks. We are grateful to Alicia Burga and Beatriz Fisas for their work.

9. Bibliographical References

AbuRa'ed, A., Chiruzzo, L., and Saggion, H. (2017). What sentence are you referring to and why? identifying cited sentences in scientific literature. In *RANLP 2017. International Conference Recent Advances in Natural Language Processing; 2017 Sep 2-8; Varna, Bulgaria.[Stroudsburg (PA)]: ACL; 2017. p. 9-17.* ACL (Association for Computational Linguistics).

Abura'ed, A., Bravo, A., Chiruzzo, L., and Saggion, H. (2018a). Lastus/taln+ inco@ cl-scisumm 2018-using regression and convolutions for cross-document semantic linking and summarization of scholarly literature. In *Proceedings of the 3nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2018). Ann Arbor, Michigan (July 2018).*

AbuRa'ed, A., Chiruzzo, L., and Saggion, H. (2018b). Experiments in detection of implicit citations. In *WOSP 2018. 7th International Workshop on Mining Scientific Publications; 2018 May 7; Miyazaki, Japan.[Paris (Francce)]: European Language Resources Association; 2018. 7 p.* ELRA (European Language Resources Association).

Athar, A. and Teufel, S. (2012). Detection of implicit citations for sentiment detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 18–26, Jeju Island, Korea, July. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Constantin, A., Pettifer, S., and Voronkov, A. (2013). Pdfx: fully-automated pdf-to-xml conversion of scientific literature. In *Proceedings of the 2013 ACM symposium on Document engineering*, pages 177–180. ACM.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

Endres-Niggemeyer, B., Elisabeth Maier, E., and Alexander Sigel, A. (1995). How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. *Information Processing and Management*, 31(5):631 – 674.

Ferrés, D., Saggion, H., Ronzano, F., and Bravo, À. (2018). Pdfdigest: an adaptable layout-aware pdf-to-xml textual content extractor for scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Herrmannova, D. and Knoth, P. (2016). An analysis of the microsoft academic graph. *D-Lib Magazine*, 22(9/10).

Hoang, C. D. V. and Kan, M.-Y. (2010). Towards automated related work summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 427–435. Association for Computational Linguistics.

Hu, Y. and Wan, X. (2014). Automatic generation of related work sections in scientific papers: An optimization approach. In *EMNLP*, pages 1624–1633.

Jaidka, K., Khoo, C., and Na, J.-C. (2013). Deconstructing human literature reviews–a framework for multi-document summarization. In *Proceedings of the 14th*

European Workshop on Natural Language Generation, pages 125–135.

Jaidka, K., Chandrasekaran, M. K., Elizalde, B. F., Jha, R., Jones, C., Kan, M.-Y., Khanna, A., Molla-Aliod, D., Radev, D. R., Ronzano, F., and Saggion, H. (2014a). The computational linguistics summarization pilot task. In *Proceedings of TAC 2014*.

Jaidka, K., Chandrasekaran, M. K., Jha, R., Jones, C., Kan, M.-Y., Khanna, A., Mollá-Aliod, D., Radev, D. R., Ronzano, F., Saggion, H., and Wee, W. K. (2014b). The computational linguistics summarization pilot task. In *Proceedings of TAC 2014*.

Jaidka, K., Chandrasekaran, M. K., Jain, D., and Kan, M.-Y. (2017a). Overview of the CL-SciSumm 2017 shared task. *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries*, August.

Jaidka, K., Chandrasekaran, M. K., Rustagi, S., and Kan, M.-Y. (2017b). Insights from CL-SciSumm 2016: the faceted scientific document summarization shared task. *International Journal on Digital Libraries*, Jun.

Kovatchev, V., Martí, M. A., and Salamó, M. (2018). WARP-text: a web-based tool for annotating relationships between pairs of texts. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 132–136, Santa Fe, New Mexico, August. Association for Computational Linguistics.

Liakata, M. and Soldatova, L. (2008). Guidelines for the annotation of general scientific concepts. *Aberystwyth University, JISC Project Report http://ie-repository. jisc. ac. uk/88*.

Liakata, M., Soldatova, L. N., et al. (2009). Semantic annotation of papers: Interface & enrichment tool (sapient). In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 193–200. Association for Computational Linguistics.

Liu, H. (2017). Sentiment analysis of citations using word2vec. *arXiv preprint arXiv:1704.00177*.

Lopez, P. (2009). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*, pages 473–474. Springer.

Maggio, L., Sewell, J., and Artino, A. (2016). The literature review: A foundation for high-quality medical education research. *Journal of Graduate Medical Education*, 8:297–303, 07.

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

Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., and Wilks, Y. (2002). Architectural elements of language engineering robustness. *Natural Language Engineering*, 8(2-3):257–274.

Pautasso, M. (2013). Ten simple rules for writing a literature review. *PLoS computational biology*, 9:e1003149, 07.

Qazvinian, V. and Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 689–696. Association for Computational Linguistics.

Ronzano, F. and Saggion, H. (2015). Dr. Inventor Framework: Extracting structured information from scientific publications. In *International Conference on Discovery Science*, pages 209–220. Springer.

Rowley, J. and Slack, F. (2004). Conducting a literature review. *Management Research News*, 27, 06.

S. G. Khoo, C., Na, J.-C., and Jaidka, K. (2011). Analysis of the macro-level discourse structure of literature reviews. *Online Information Review*, 35, 04.

Saggion, H. and Lapalme, G. (2002). Generating indicative-informative summaries with sumum. *Computational Linguistics*, 28(4):497–526.

Saggion, H. (2008). SUMMA. A Robust and Adaptable Summarization Tool. *TAL*, 49(2):103–125.

Saggion, H. (2011). Learning predicate insertion rules for document abstracting. In *Computational Linguistics and Intelligent Text Processing - 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II*, pages 301–312.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-j. P., and Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM.

Sujatha, P. and Dhavachelvan, P. (2011). Precision at k in multilingual information retrieval.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Teufel, S. (2006). Argumentative zoning for improved citation indexing. In *Computing attitude and affect in text: Theory and Applications*, pages 159–169. Springer.

Valenzuela, M., Ha, V. A., and Etzioni, O. (2015). Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data*.

Vu, H. C. D. (2010). *Towards automated related work summarization*. Ph.D. thesis.

Wade, A. D. (2015). Overview of microsoft academic graph. *Alonso et al.[2]*, page 8.

Xiong, C., Power, R., and Callan, J. (2017). Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279. International World Wide Web Conferences Steering Committee.

10. Language Resource References

Bird, Steven and Dale, Robert and Dorr, Bonnie J and Gibson, Bryan and Joseph, Mark Thomas and Kan, Min-

Yen and Lee, Dongwon and Powley, Brett and Radev, Dragomir R and Tan, Yee Fan. (2008). *The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics*. EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA, ISLRN 150-170-243-077-5.

Fisas, Beatriz and Ronzano, Francesco and Saggion, Horacio. (2016). *A multi-layered annotated corpus of scientific papers*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).

Radev, Dragomir R. and Muthukrishnan, Pradeep and Qazvinian, Vahed and Abu-Jbara, Amjad. (2013). *The ACL anthology network corpus*. Language Resources and Evaluation.

Michihiro Yasunaga and Jungo Kasai and Rui Zhang and Alexander Fabbri and Irene Li and Dan Friedman and Dragomir Radev. (2019). *ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks*. Proceedings of AAAI 2019.