MinScIE: Citation-centered Open Information Extraction

Anne Lauscher*
University of Mannheim
Mannheim, Germany
anne@informatik.uni-mannheim.de

Yide Song*
University of Mannheim
Mannheim, Germany
yisong@mail.uni-mannheim.de

Kiril Gashteovski*
University of Mannheim
Mannheim, Germany
k.gashteovski@uni-mannheim.com

ABSTRACT

Acknowledging the importance of citations in scientific literature, in this work we present *MinScIE*, an Open Information Extraction system which provides structured knowledge enriched with semantic information about citations. By comparing our system to it's original core, *MinIE*, we show that our approach improves extraction precision by 3 percentage points.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Supervised learning by classification**; *Discourse, dialogue and pragmatics*; • **Applied computing** → **Digital libraries and archives**.

KEYWORDS

open information extraction, citation analysis, citation polarity, citation purpose, suppport vector machines, scitorics

ACM Reference Format:
Anne Lauscher, Yide Song, and Kiril Gashteovski. 2019. MinScIE: Citation-centered Open Information Extraction. In *JCDL '19: ACM/IEEE Joint Conference on Digital Libraries, June 2–6, 2019, Urbana-Champaign, Illinois.* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

With the growing amount of scientific literature published per year, the automatic processing of scientific writing becomes more and more important. For the purpose of generating structured knowledge from large scientific text collections, Open Information Extraction (OIE) can be applied. OIE systems aim to extract information in the form of triples – *(subject, relation, object)* – from natural language sentences in an unsupervised manner. Consider the sentence *"Bell is a telecommunication company, which is based in L.A."* – most of the OIE systems would extract the triples *("Bell"; "is"; "telecommunication company")* and *("Bell"; "is based in"; "L.A.")*. However, Groth et al. [4] recently showed in a systematic evaluation that current off-the-shelf OIE systems perform significantly worse when applied to the science domain, which raises the need for domain adaptation. Acknowledging the importance of citations as argumentative tools in scientific writing [3], in this work we present *MinScIE*,[1] an extension of the OIE system *MinIE* [2], which is specifically designed to handle citations and semantically enrich the obtained triples when applied to scientific content. *MinScIE* provides fixes for most of the issues identified in [4] and further analyzes citation contexts regarding their citation polarity [1] and purpose [7]. We make *MinScIE* publicly available[2] and provide an initial evaluation and analysis of the system's performance.

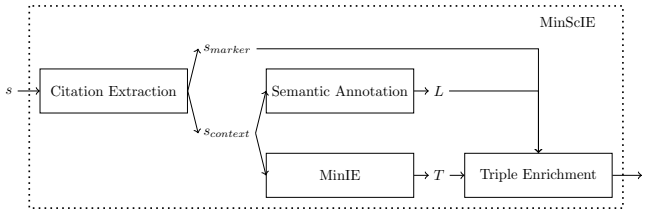

Figure 1: MinScIE's pipeline.

2 SYSTEM DESCRIPTION

MinScIE's pipeline approach is illustrated in Figure 1. Given a sentence $s = [w_1, ..., w_n]$ of length n with individual tokens w_i, the system first extracts the citation marker, a subsequence of tokens $s_{marker} = [w_j, ..., w_k]$, with k, j such that $k - j > 0$ by matching the text to a regular expression. Next, we feed the remaining sequence $s_{context}$, which we consider the citation context, a) through a classifier for each of the respective annotation tasks, i.e., the citation polarity or the citation purpose models, and b) into the original *MinIE* core. Following Lauscher et al. [6], the system currently employs two Support Vector Machines (SVMs) with RBF kernel, for which we represent the context with sentence-averaged 300-dimensional GloVe embeddings trained on computational linguistic publications . Note that the classification models are interchangable and that *MinScIE* can be easily adapted to handle a bigger set of *scitorics* [5]. As a result of the preceding steps, we obtain two labels L reflecting the semantics of the citation based on the context provided by the surrounding sentence as well as a set of extracted triples $T = \{t_1, ..., t_m\}$ of size m. In a last step, we enrich the triples by creating a new *MinIE* triple annotation object consisting of the citation marker and the citation polarity and purpose labels identified, and adding it to each $t_i \in T$. The citation can be uniquely identified with the citation marker, which allows for citation-centered structured knowledge analysis in downstream applications.

*All authors contributed equally to this research.
[1]Pronounciation: [mˈɪnski].

[2]https://github.com/gkiril/MinSCIE

JCDL '19, June 2019, Urbana-Champaign, Illinois

Table 1: Evaluation results on 220 sentences from different scientific domains.

System	# Triples	Correct		Incorrect	
		#	%	#	%
MinIE	846	568	67.14	278	32.86
MinScIE	822	577	70.19	245	29.81

3 EVALUATION

For a detailed evaluation of the SVM model performances, we refer to Lauscher et al. [6]. Here, we focus on the evaluation of the quality of the extracted triples.

Data Set. For evaluating the triple extraction quality of our system, we use the SCI data set created by Groth et al. [4].[3] It consists of 220 sentences which were randomly sampled from 11 scientific articles covering different scientific domains, e.g. medicine, agriculture, and computer science. Before conducting the annotation process, we applied the original *MinIE* distribution in the *safe mode* as well as *MinScIE* to the each of the sentences in the SCI data set.

Annotation Process. In the annotation process, we evaluated the extracted triples from each of the two systems regarding their correctness. To this end, we hired three annotators and instructed them to follow the annotation guidelines created by Gashteovski et al. [2].[4] We split the data for both systems evenly between the annotators. Gashteovski et al. [2] report a moderate inter-annotator agreement of 50 % − 53 % Cohen's κ for untrained labelers.

Results. Table 1 shows the results of our evaluation. Overall, for 26 sentences (11.89 %) the processing between *MinIE* and *MinScIE* differed. As a result, *MinScIE* extracted 24 triples less than the original system and improved the performance by around 3 % precision. Removing the citation markers before extracting the triples avoids confusing the core of our system, *MinIE*, in many cases and leads to better extractions. As an example, consider the following sentence:

> *"Some use strong extractants which dissolved strongly bound P and hence does not necessarily represent the actual labile pool of P in soils and others use weak extractants like water or weak acids which might underestimate available soil P (Neyroud and Lischer, 2003)."*

The original version of *MinIE* outputs the following triples:

t_1: *(water; is; weak extractant)*
t_2: *(weak acids; is; weak extractant)*
t_3: *(QUANT_S_1; use; strong extractants); QUANT_S_1 = some*
t_4: *(strong extractants; dissolved; strongly bound P)*
t_5: *(strong extractants; does represent; actual labile pool of P in soils)*
t_6: *(others; use weak extractants like; water)*
t_7: *(others; use weak extractants like; weak acids)*
t_8: *(others; use; weak extractants)*
t_9: *(water; underestimate; available soil P)*
t_{10}: *(weak acids; underestimate; available soil P)*
t_{11}: *(available soil P; is; Neyroud)*
t_{12}: *(available soil P; is; Lischer)*

[3]https://data.mendeley.com/datasets/6m5dyx4b58/2/files/f187c790-5770-408a-b04e-ba849d7d0261
[4]https://dws.informatik.uni-mannheim.de/fileadmin/lehrstuehle/pi1/pi1/minie/minie-labeling-guide.pdf

In this example, the triples t_1 to t_{10} are correctly extracted, but the last two triples, t_{11} and t_{12} are incorrectly extracted. The reason for this issue is that *MinIE* (which is based on dependency parsing output) cannot deal with citation markers, because the dependency parser confuses the words from within brackets from the citation markers as having appositional relationship with the head noun of the preceding noun phrase. In contrast, *MinScIE* extracts the same triples t_1 to t_{10}, but omits outputting the wrong triples t_{11} and t_{12}. In addition, *MinScIE* extracts and preserves the citation marker, semantically analyzes the citation context, and attaches the obtained information to each of the triples:

Citation Marker: (Neyroud and Lischer, 2003)
Citation Polarity: Neutral
Citation Function: Criticizing

However, some issues identified in [4] are not resolved by applying this rather simple extension, e.g., the correct extraction of triples from sentences including complex mathematical formula.

4 CONCLUSION

In this work we presented *MinScIE*, an OIE system adapted to the scientific domain. By accounting for the occurance of citation markers in the text, the system offers a higher precision than it's non-adapted core, *MinIE*, and additionally provides a semantic analysis of the context of a citation in terms of citation polarity and purpose. We strongly believe that citation-centered OIE triples are useful representations of the knowledge inherent to scientific literature as they enable the user to connect factual knowledge with references to the scientific discourse. We also think that our technique for avoiding confusion in the triple extraction can be easily extended to other sources of confusion, e.g. intra-paper references to figures and tables. In the future, we plan to extend *MinScIE* for handling these cases and further enrich the semantic analysis component with more *scitorics* [5], e.g., argumentative information.

REFERENCES

[1] Awais Athar. 2011. Sentiment Analysis of Citations using Sentence Structure-Based Features. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Student Session*. 81–87. http://aclweb.org/anthology/P11-3015

[2] Kiril Gashteovski, Rainer Gemulla, and Luciano Del Corro. 2017. MinIE: Minimizing Facts in Open Information Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 2630–2640. http://aclweb.org/anthology/D17-1278

[3] G Nigel Gilbert. 1977. Referencing as persuasion. *Social Studies of Science* 7, 1 (1977), 113–122.

[4] Paul T. Groth, Michael Lauruhn, Antony Scerri, and Ron Daniel. 2018. Open Information Extraction on Scientific Text: An Evaluation. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. 3414–3423. http://aclweb.org/anthology/C18-1289

[5] Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018. ArguminSci: A Tool for Analyzing Argumentation and Rhetorical Aspects in Scientific Writing. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*. 22–28. http://aclweb.org/anthology/W18-5203

[6] Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. 2017. Investigating Convolutional Networks and Domain-Specific Embeddings for Semantic Classification of Citations. In *Proceedings of the 6th International Workshop on Mining Scientific Publications*. Association for Computing Machinery, Toronto, ON, Canada, 24–28.

[7] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic Classification of Citation Function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sydney, Australia, 103–110.