Annotating with Pros and Cons of Technologies in Computer Science Papers

Hono Shirai[1], **Naoya Inoue**[1,2], **Jun Suzuki**[1,2], **Kentaro Inui**[1,2]
[1]Tohoku University, [2]RIKEN AIP
{h.shirai, naoya-i, jun.suzuki, inui}@ecei.tohoku.ac.jp

Abstract

This study explores the task of extracting a technological expression and its pros/cons from computer science papers. We report the ongoing efforts on the annotated corpus of pros/cons and the analysis of the nature of the automatic extraction task. Specifically, we show how to adapt the targeted sentiment analysis task for extracting pros/cons from computer science papers and conduct an annotation study. We construct a strong baseline model and conduct an error analysis to identify the challenges of the automatic extraction task. Experimental results show that pros/cons can be consistently annotated by annotators, and that the task is challenging owing to the requirement of domain-specific knowledge. The annotated dataset is made publicly available for research purposes.

1 Introduction

The number of scientific publications has been rapidly increasing. Johnson et al. (2018) showed that over 3 million research articles are published annually. It is increasingly difficult for researchers to have a bird's-eye view of current research trends with such a large number of publications.

This study explores information extraction from computer science papers. The main focus of computer science publications involves problem solving (e.g., optimization algorithm). One typical form of computer science publications is presenting an issue and then discusses solutions for it. Specifically, the pros and cons of previously proposed technologies are discussed and propose new technology. Example (1) discusses the cons of previous technologies for coreference resolution:[1]

(1) *While successful, these approaches require labeled training data, consisting of mention pairs and the correct decisions for them.* (D08-1068)

Therefore, when computer scientists write a paper, it is important to have a bird's-eye view of the pros and cons of previous technologies. As the number of publications rapidly increases, it is desirable to develop an automated tool for mining the pros and cons of technologies.

Previous works have explored automatic extraction of a wide variety of scientific knowledge to assist researchers in collecting relevant publications. This research direction includes domain-independent approaches, such as Citation Network (Kajikawa et al., 2007) and Argumentative Zoning (Teufel et al., 1999), and domain-dependent approaches such as BioNLP (Deléger et al., 2016). These technologies are the foundation of scientific search engines or knowledge discovery tools, such as Semantic Scholar[2] and Dr. Inventor (Ronzano and Saggion, 2015). Nevertheless, less attention has been paid to the mining of the pros and cons of technologies.

This study performs a preliminary investigation on automatically identifying technologies and their pros/cons from computer science papers (henceforth referred to as *pros/cons identification*). We frame pros/cons identification as the well-known NLP task of targeted sentiment analysis (Jiang et al., 2011) and conduct an annotation study. Futhermore, we build a neural baseline model to identify the challenges of pros/cons identification task. The annotation study indicates that the pros/cons identification task can be reasonably framed as the task of targeted sentiment analysis. The experimental results of automatic extraction show that pros/cons identification is difficult mainly owing to the requirement of domain-specific knowledge. The annotated dataset is made

[1]Throughout the paper, an appended 8-character identifier indicates the ACL anthology's paper identifier.

[2]https://www.semanticscholar.org

Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications, pages 37–42
Minneapolis, USA, June 6, 2019.

publicly available.[3]

2 Annotation Scheme

We investigate the task of pros/cons identification task by adopting an existing annotation scheme to our task and conducting an annotation study. Specifically, we apply an annotation scheme from the targeted sentiment analysis task (Jiang et al., 2011), which is mainly developed for mining positive/negative opinion about named entities (e.g. person, products) from twitter.

2.1 TERM

We introduce TERM label to annotate with technological terms. We define TERM as a noun phrase that represents a mechanism, a function, or a method to solve the problem. In Example (2), *recursive neural network* and *AdaRNN* are labeled as TERM because these are types of neural network models.

(2) *We employ a novel adaptive multi-compositionality layer in recursive neural network, which is named as AdaRNN (Dong et al., 2014).* (P14-2009)

Note that we also annotate a general noun phrase (e.g. *our method*) with the TERM label and named entities with the TERM label.

2.2 Sentiment

For each phrase labeled as TERM, we additionally annotate it with a Sentiment attribute, which represents how a technology is evaluated. Following the previous work on targeted sentiment analysis (Jiang et al., 2011, etc.), an evaluation is expressed by three types of attributes: Positive, Negative, and Neutral. These labels represent a local polarity within a sentence and are only judged based on the information obtained from a sentence containing TERM. In Example (3), *the whole-sentence-based classifier* that is labeled TERM is assigned Positive attribute, because it is positively evaluated by the expression "*performs the best*".

(3) *The results indicate that the whole-sentence-based classifier performs the best.* (D09-1019)

Similarly, the negative attribute is assigned to the examples of negative aspects of technologies.

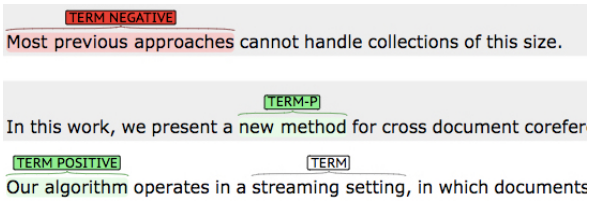

Figure 1: The brat annotation interface used for conducting our annotation study.

Neutral attribute is given to TERM if only the neutral features and properties of technology are described in the sentence. In Example (2), *recursive neural network* and *AdaRNN* are assigned to Neutral attributes.

3 Annotation Study

In this section, we describe our annotation study used for creating a dataset for the automatic extraction of pros/cons.

3.1 Dataset

We retrieved 92 computational linguistics papers that contained the keyword "*coreference resolution*" in the title or body texts using Google Custom Search in ACL Anthology.[4] Various methods have been proposed for coreference resolution because it has been a subject of research for numerous years. This is suitable for our trial annotation. These papers we considered were published from 1999 to 2017.

In a publication, the pros and cons of the proposed/existing methods are generally discussed in the introduction section. Therefore, we focus on annotating only the introduction section to reduce the cost of annotation.

3.2 Settings

We employed three fluent-English speakers who specialize in NLP. We assigned two annotators per paper to investigate the inter-annotator agreement. Figure 1 illustrates the annotation interface *brat* (Stenetorp et al., 2012), which is used for conducting our annotation.

3.3 Results and Discussion

We measured the inter-annotator agreement after the annotation was completed.

TERM The percentage of the exact match of TERM spans between annotators was 24.0 %. We observed multiple of cases where one annotator labeled a phrase as TERM, but the other annotator

[3] https://github.com/cl-tohoku/scientific-paper-pros-cons

[4] http://www.aclweb.org/anthology/

did not. Such examples included *joint inference* and *a learned cluster ranker*. We speculate that this is because these noun phrases indirectly indicate whether a phrase is a mechanism, function, or method.

The percentage of *partial* match between annotators was 38.2 %. We observed that the interpretation of span was sometimes different across annotators in certain cases. For example, one annotator included a modifier such as *a simplified semantic role labeling (SRL) framework*, but the other did not (i.e., *semantic role labeling (SRL) framework*).

Sentiment We calculated the inter-annotator agreement of the Sentiment attributes for 390 instances whose TERM span annotation matched exactly matched between annotators. We obtained a Fleiss's Kappa of 0.65, which indicated substantial agreement (Fleiss, 1971).

Even though the inter-annotator agreement was generally high, there are a few disagreements. The primary cause of disagreements is that one annotator assigned the Neutral attribute, and the other assigned the non-Neutral attributes (i.e., Positive or Negative). Among the disagreements, we found numerous cases where domain-specific knowledge was required. In Example (4), one annotator labeled *ranking models* as Positive and the other labeled them as Neutral. To judge the sentiment attributes correctly, one required the domain knowledge of coreference resolution that *directly capturing the competition among potential antecedent candidates* is appropriate.

(4) *In essence, ranking models directly capture during training the competition among potential antecedent candidates, instead of considering them independently.* (D08-1069)

We found a large number of cases where sentences took the form of concession. In Example (5), one annotator labeled *the pairwise approach* as Negative and the other Neutral. We speculate that annotators were confused because *the pairwise approach* is evaluated positively by the phrase *high precision* in the subordinate clause, but negatively by the phrase *neither realistic nor scalable* in the main clause.

(5) *While the pairwise approach has high precision, it is neither realistic nor scalable to explicitly enumerate all pairs of compatible word pairs.* (N10-1061)

	# TERM spans		
# sentences	Positive	Neutral	Negative
2,058	255	1,100	116

Table 1: Statistics of annotated corpus.

4 Experiments

To identify the challenges of the automatic extraction task, we ran a strong baseline model to conduct an error analysis.

4.1 Dataset

To obtain high-recall annotations, we aggregated all annotations from each annotator pair. We solved the conflicts between Sentiment attributes by employing the following rules: (i) if both labels are Positive and Negative, Neutral label is applied, and (ii) if one label being Positive or Negative and the other Neutral, the non-Neutral attribute is applied. Furthermore, we manually cleaned the data by resolving the conflicts between the spans assigned by two annotators (e.g., *a model* v.s. *model*). The statistics of the final corpus are shown in Table 1.

4.2 Model

We formulate the automatic extraction task as a BIO sequence tagging task. Specifically, given a sentence, the model tags each word as one of {O, B-POS, I-POS, B-NEG, I-NEG, B-NEU, I-NEU}, where a combination of BI tags represents a Positive (POS), Negative (NEG), and Neutral (NEU) technical term span.

We use the BiLSTM-CRF model proposed by Lample et al. (2016) which was originally designed for the task of named entity recognition.[5] Regarding word embedding, we use word2vec (Mikolov et al., 2013) embeddings trained on ACL Anthology Corpus (Aizawa et al., 2018) (henceforth, CL), and ELMo (Peters et al., 2018) embeddings trained on 1 Billion Word Benchmark (henceforth, EL).

4.3 Configurations

For the detection, TERM and Sentiment are judged as correct only if they exactly match with gold-standard spans. We report F1 scores as an evaluation measure. We evaluate our models in two configurations.

[5]We use the implementation provided at https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf

ID	Sentence	Gold	Prediction
(i)	*Several studies report successful applications of concept maps in this direction...* (I17-1081)	Positive	N/M
(ii)	*Second, they have limitations in their expressiveness.* (D09-1101)	Negative	N/M
(iii)	*While successful, these approaches require labeled training data, consisting of mention pairs and...* (D08-1068)	Negative	Neutral
(iv)	*We compare the prediction accuracy of memory network with an existing state-of-the-art coreference resolution...* (W17-2605)	Neutral	N/M

Table 2: Examples of the model predictions. Underlined words indicate a TERM phrase span. N/M indicates that the model does not label it as TERM.

Setting	Emb.	dev F1	test F1 / Prec. / Rec.
10-FCV	CL	50.70	49.79 / 50.0 / 49.7
	CL+EL	54.23	52.35 / 54.4 / 50.8
NEWYEAR	CL+EL	53.29	42.69 / 51.8 / 36.3

Table 3: Performance of pros/cons identification.

10-FCV We employ 10-fold cross validation in this configuration. When data are split, we ensure that the paper IDs in the training set do not have an overlap with the paper IDs in the test set. For model selection, we reserve 10% of the training dataset as the development set. We report F1 scores averaged across all folds.

NEWYEAR In this configuration, to evaluate the models in real-life situations, we verify whether the models are able to extract the pros and cons of new papers after being trained on older papers. We utilize the papers from 2017 (i.e. the latest papers) and data from other years as the test and training sets, respectively.

4.4 Results and Discussion

The results are shown in Table 3. ELMo embeddings improve the prediction performance on the test and dev sets.[6] This indicates that contextual information is important for pros/cons identification.

The results also highlight the difficulty of our task. We analyzed the results given by the best model (CL+EL model) to investigate how challenging the task is. Model predictions along with their gold labels are shown in Table 2.

First, we observe that when an input does not include a word that directly indicates a method, then we are likely to obtain a false negative error (i.e., the recognition of TERM fails). In sentence (i), the model is unable to predict a label for the term *concept maps* because it does not include a word that indicates a TERM. Sentence (ii) is another case in which the model cannot predict whether *they* is TERM. Although *they* refers to a model, our model cannot recognize it because it does not resolve coreference.

We also discovered that it is difficult to predict Sentiment attributes when the phrase implicitly expresses sentiment. In sentence (iii), the gold label for *these approaches* is Negative. However, the model predicts Neutral because *successful* is a positive expression for *these approaches* and *require labeled training data, ...* is negative.

The performance of the models in the NEWYEAR configuration is poorer than that in the 10-FCV configuration. We observed that prediction fails for sentences that contain unknown words. For example, in sentence (iv), *memory network* is not observed in training data.

5 Use Case

To show the use cases of our study, we parsed 60 ACL papers published in 2017 with our best performing model. One use case is to employ our system with a search engine-style interface. We implemented a prototype pros/cons identification system. We consider a situation in which we want to obtain an overview of the evaluation measures of dialogue responses and we already have several keywords such as *ADEM* and *BLEU*. Given a search query *ADEM*, our system lists pros/cons of *ADEM*, as illustrated in Figure 2. Analyzing the results, the cons of *ADEM* are provided such as "*ADEM tends to be too conservative when predicting response scores*". We believe that this search interface will provide useful information for re-

[6] The improvement is statistically significant (Wilcoxon's signed-rank test, $p < 0.05$).

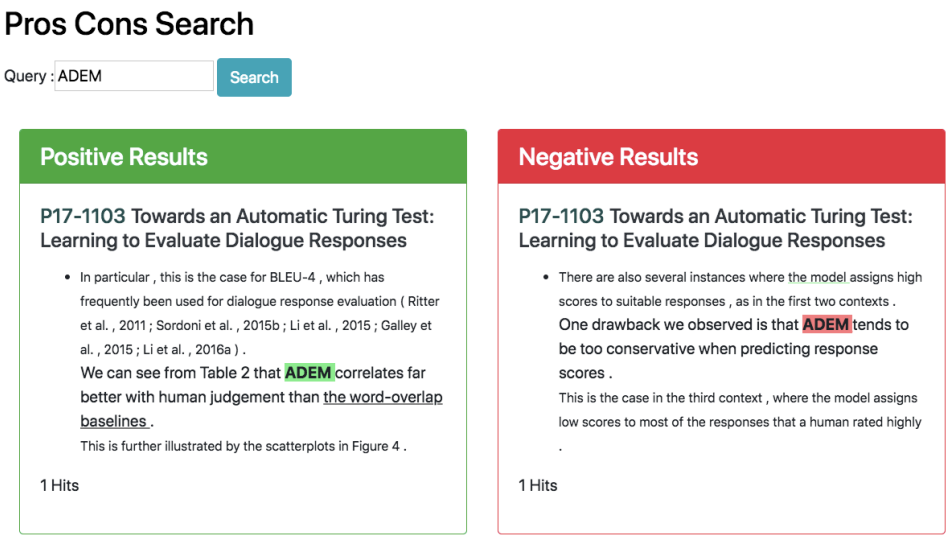

Figure 2: Search results obtained from our pros/cons identification prototype system.

searchers who are starting work in a new field.

Another possible interface is an "add-on" for a PDF viewer. For each important keyword in a PDF, a pop-up window can appear and inform the user about the pros/cons of the keyword.

6 Related Work

There are several types of attempts on extracting useful information from scientific papers. Citation Network (Kajikawa et al., 2007) analyzes the trends of important technology in papers. Argumentative Zoning (Teufel et al., 1999) classifies the sentences in papers into an argumentative type such as BACKGROUND and RELATEDWORK, etc.

A few studies annotate scientific papers with relations between entities such as "APPLY-TO(CRF, POS tagger)". Tateisi et al. (2016) propose an annotation scheme for describing the semantic structures of research articles. SemEval, which is one of the shared tasks workshop in NLP, proposes some information extraction tasks in the scientific paper domain. ScienceIE (Augenstein et al., 2017) is the task of extracting phrases and relationships from papers in multiple domains. SemEval-2018 Task 7 (Gábor et al., 2018) proposes a classification task that classifies the relations between entities in the ACL Anthology. BioNLP (Deléger et al., 2016) aims to extract technical terms, such as proteins, relations between proteins, and substances and their side effects, in the biological and medical domains.

In the field of sentiment analysis, Aspect-Based Sentiment Analysis is performed in the domain of review documents is performed. SemEval-2015 Task 12 (Pontiki et al., 2015) is the task of performing sentiment analysis based on the defined viewpoints such as the prices, cooking or quality of service in hotels and restaurants. Targeted sentiment analysis (Jiang et al., 2011) is the task of classifying a sentiment towards a certain target entity in given sentences. The target entity is the name of persons, companies, and products. In the sentiment analysis in the scientific paper domain, Citation Sentiment Analysis (Yousif et al., 2017) has been performed to analyze the sentiment polarity of an author against documents cited in a paper. However, targeted sentiment analysis of the paper content itself has not been explored.

7 Conclusion

We have proposed the task of pros/cons identification. We have designed a scheme for annotating technological terms and its pros/cons. An annotation study shows that annotators can consistently annotate sentiment attributes. Experiments performed on automatic extraction show that the task is still challenging because domain-specific knowledge and inference are required.

In our future work, we plan to expand our annotation to other domains such as computer vision. We also plan to develop a mechanism of recognizing sentiment attributes using domain-specific knowledge.

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR1513.

References

Akiko Aizawa, Takeshi Sagara, Kenichi Iwatsuki, and Goran Topic. 2018. Construction of a new acl anthology corpus for deeper analysis of scientific papers. In *Third International Workshop on SCIentific DOCument Analysis (SCIDOCA-2018)*.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières, and Claire Nédellec. 2016. Overview of the bacteria biotope task at bionlp shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160, Portland, Oregon, USA. Association for Computational Linguistics.

Rob Johnson, Anthony Watkinson, and M Wabe. 2018. The stm report. *An overview of scientific and scholarly publishing. 5th edition October*.

Yuya Kajikawa, Junko Ohno, Yoshiyuki Takeda, Katsumori Matsushima, and Hiroshi Komiyama. 2007. Creating an academic landscape of sustainability science: an analysis of the citation network. *Sustainability Science*, 2(2):221.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

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

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

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

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Abdallah Yousif, Zhendong Niu, John K. Tarus, and Arshad Ahmad. 2017. A survey on sentiment analysis of scientific citations. *Artificial Intelligence Review*.