Multi-task Citation Content Analysis for Clinical Research Publications

Anonymous ACL submission

Abstract

Citations are essential building blocks in scientific knowledge production. Citation content analysis using NLP methods has been proposed to benefit tasks such as scientific paper summarization and research impact assessment. In this paper, we propose a new task, *citation subject matter extraction*, and augment an existing citation sentiment corpus with citation context and subject matter annotations to enable a finer-grained study of citation content. We propose a BERT-based multi-task model to jointly address these three classification tasks (i.e., context, subject matter, and sentiment) by enabling knowledge transfer across tasks. Our experimental results show the effectiveness of our joint model over single task models. We also obtain state-of-the-art results for the citation sentiment classification task and demonstrate that isolating the subject matter significantly improves this task. Our error analysis suggests improving annotation consistency and using external knowledge sources could further improve performance. We will make our code, data, and annotation guidelines publicly available upon acceptance.

1 Introduction

Citations play a fundamental role in scholarly communication. It is through citations that scientific claims gain credibility and become beliefs (Greenberg, 2011). Citation-based metrics, such as journal impact factor (Garfield, 1972) and h-index (Hirsch, 2005), are also widely used to measure the scholarly contributions of researchers and journals (Waltman, 2016), although their shortcomings are generally acknowledged (Hicks et al., 2015).

Citation content analysis (Zhang et al., 2013) is concerned with understanding the qualitative nature of the relationship between the citing and the cited papers at finer granularity, including citation context (Abu-Jbara and Radev, 2012; Qazvinian and Radev, 2010), citation sentiment (Athar, 2014; Xu et al., 2015), citation function (Teufel et al., 2006a,b; Jurgens et al., 2018; Lauscher et al., 2021), and citation significance (Zhu et al., 2015; Valenzuela et al., 2015). Citation content analysis can not only augment purely quantitative citation-based metrics, but can also be beneficial for downstream tasks, such as scientific paper summarization (Qazvinian and Radev, 2008) and automatic survey generation (Mohammad et al., 2009).

In this paper, we propose a new fine-grained citation content analysis task, *citation subject matter extraction* and investigate its interaction with citation context and sentiment classification tasks. We define subject matter as "the text span in the citing paper that corresponds to the main topic/argument/claim that is cited from the reference paper." We base our study on a corpus of clinical trial articles (Xu et al., 2015). As the motivation for this task, we argue that current characterizations of citation content may be too simplistic to address the citation tasks that require cross-document linking of citing and reference articles, such as scientific paper summarization (Qazvinian and Radev, 2008; Jaidka et al., 2019; Chandrasekaran et al., 2019, 2020) and citation accuracy assessment (Cohan and Goharian, 2017; Kilicoglu, 2018). First, most related work characterizes citation context as the citation sentence or a fixed number of sentences around the citation (Athar and Teufel, 2012; Jaidka et al., 2019). However, citation context often spans multiple, possibly non-contiguous, sentences (Qazvinian and Radev, 2010) or may correspond to clause-level fragments (Abu-Jbara and Radev, 2012). Second, a citation context often consists of two components (Small, 1978): an objective characterization of the reference paper (i.e., its *subject matter*) and an interpretive component, which indicates a commentary by the authors toward the reference paper, often referred to as *citation sentiment* (Athar, 2014). We hypothesize that distinguishing the subject matter from the authors' interpretation of it would enable a more precise linking of the citing paper to the reference paper and benefit tasks such as citation sentiment classification and citation accuracy assessment. For illustration, consider the example below with two citations (underlined) preceded by their subject matter spans (in bold), taken from a clinical trial article.

(1) CQ was significantly less effective than SP and AQ+AS in treating uncomplicated falciparum malaria, with overall treatment failure of 35.9% within 14 days of follow up. These data show **a higher prevalence of chloroquine resistance than reported in previous studies** [19-21] and **a good effectiveness of SP and AQ** [22, 23].

Both sentences must be included in the context of the citations in the second sentence (due to coreference). Furthermore, two citations refer to different subject matters from cited papers. In cross-document linking, focusing on these specific parts, rather than the full sentence, is likely to be beneficial. Also note that the sentiment of the first citation is negative and that of the second is positive, suggesting that accurate subject matter extraction can lead to better sentiment classification.

In this paper, we make the following contributions. First, we propose *citation subject matter extraction* as a new citation content analysis task. Second, we present a corpus of clinical trial articles augmented with citation context and subject matter annotations. Third, we propose a multi-task learning approach to recognize citation context, subject matter, and sentiment simultaneously. Fourth, we assess the contribution of each task to the others qualitatively and through ablation, showing that the multi-task setup benefits all tasks.

2 Methods

In this section, we first describe the clinical trial citation corpus used in this study. Next, we provide the details on our multi-task learning model and the experimental setup.

2.1 Clinical trial citation corpus

We used a corpus of the discussion sections of 285 clinical trial articles with 4,182 citations, first reported in Xu et al. (2015). The original corpus consists of citation sentiment annotations only. It was double-annotated with an inter-annotator agreement of 0.504 (Cohen's κ) and adjudicated by a third annotator.

We enriched this corpus with the citation context and subject matter annotations. In line with previous work (e.g., Abu-Jbara and Radev (2012)), we defined citation context as "the text spans that are relevant to understanding the contribution of a particular citation to the article in consideration". Citation context is expected to be interpretable in isolation and can consist of a sentence, a sentence fragment, or a set of, possibly non-contiguous, sentences. For subject matter, we used the definition given in Section 1. The subject matter span can be the same as or be subsumed by the context span. Some citation contexts may not include any explicit subject matter. Annotation guidelines were developed based on a preliminary annotation of 7 articles (of 285). Next, 30 articles were annotated by three annotators to measure inter-annotator agreement and adjudicated. The remaining 248 articles were annotated by a single annotator. F_1 measure was used to calculate inter-annotator agreement on multiply-annotated articles (Hripscak and Rothschild, 2005). Average agreement with partial matches was 0.83 for both citation context and subject matter. With exact match, agreement is lower (0.56 for context and 0.33 for subject matter), indicating that determining the precise boundaries of these elements is challenging.

Table 1 shows several example annotations from the corpus. In the first example, subject matter is in a sentence different from the citation sentence, suggesting that simply using the citation sentence for content analysis is likely to fail. The second example (rows 2-3) illustrates a case in which a sentence contains two citations with overlapping subject matter spans. Accurately identifying these spans could serve the downstream tasks better. In the third example (rows 4-5), the interpretive components of the two citations indicate different sentiment values while their subject matters are the same, similarly illustrating that these tasks are interrelated.

2.2 Model Architecture

While citation analysis tasks are often solved separately (Abu-Jbara and Radev, 2012; Abu-Jbara et al., 2013), some recent work considered two or more tasks together to benefit from multi-task learning (Yousif et al., 2019; Su et al., 2019). Compared with previous work, we make fewer assumptions about the distribution of citation contexts to get as complete a context as possible. We propose a multi-task model to solve the tasks of context

ID	Senti-ment	Citation context (**subject matter**)
1	positive	One possible explanation is that **the combination of aspirin and clopidogrel played an important role in reducing the early risk of stroke**. This conclusion is in accordance with the results of Wong et al. [36].
2	neutral	Many studies that emerged during the past decades described **a benefit of dietary fiber intake**, **such as a decreased risk of colorectal cancer** [10], and lowering of cholesterol and triglycerides levels [11].
3	neutral	Many studies that emerged during the past decades described **a benefit of dietary fiber intake**, **such as** a decreased risk of colorectal cancer [10], and **lowering of cholesterol and triglycerides levels** [11].
4	neutral	Several **phase III randomized studies of cancer vaccines** have been performed [18], but very few of them were successful [19].
5	positive	Several **phase III randomized studies of cancer vaccines** have been performed [18], but very few of them were successful [19].

Table 1: Examples from the corpus. In each row, the relevant citation marker is underlined and the subject matter span corresponding to it in bold.

sentence extraction, subject matter extraction and citation sentiment classification simultaneously to benefit from knowledge transfer across tasks. Annotated citation contexts are sometimes sentence fragments rather than full sentences; however, we perform sentence-level context extraction because we observed that the great majority of context annotations involved full sentences in our corpus (96% intersection-over-union (IoU) between context annotations and context sentences). The overall architecture is shown in Figure 1 and each component is discussed below.

Shared encoder To get the input to our model, we first need to select a text window surrounding the citation which covers the author's discussion about the cited work. This window must be carefully chosen: if the window is too small, it will truncate the citations that span a longer range of text, causing information loss; if the window is too wide, it will introduce too many negative samples for the context sentence extraction, and may include too much irrelevant information from other cited papers that interferes with the model's predictions on this current citation. Adjacent citations often have highly overlapping context (as seen in Table 1) and are meant to be understood together by human readers. Designing a model that benefits from larger context while remaining discriminative enough on adjacent citations is challenging. For each citation mention, a *candidate scope* is selected starting from the citation sentence and going in both directions until it meets the paragraph boundaries or the previous/next citation sentence (inclusive), whichever comes first. More formally, consider a paragraph as a sequence of sentences $[S_1, \ldots, S_n]$, among which the explicit citing sentences are $S_{e_1}, \ldots, S_{e_m}$. Suppose citation q is explicit in sentence $S_{e_i}, 1 \leq i \leq m$. We select a continuous sequence of sentences as the *window* for citation q, which is given by

$$W_q = \begin{cases} [S_1, \ldots, S_{e_2}], & \text{if } i = 1 \\ [S_{e_{m-1}}, \ldots, S_n], & \text{if } i = m \\ [S_{e_{i-1}}, \ldots, S_{e_{i+1}}], & \text{otherwise} \end{cases} \tag{1}$$

Statistics on our dataset show that less than 0.5% context sentences go beyond the window. Following Cohan et al. (2019), we append a special token [SEP] to each sentence in the sequence. Hereafter, we assume that $1 < i < m$ for ease of discussion. For citation q, we get the text string:

$$[S_{e_{i-1}}, \text{[SEP]}, \ldots, S_{e_{i+1}}, \text{[SEP]}]$$

To differentiate citation q from other citations in the window, we replace its span with a special [CLS] token. This gives us the model input for q, denoted as W'_q. We use BERT (Devlin et al., 2019) to encode this text string:

$$\mathbf{W}'_q = \text{BERT}\left(W'_q\right) \tag{2}$$

where $\mathbf{W}'_q = [\mathbf{S}_{e_{i-1}}; [\mathbf{SEP}]_1; \ldots, \mathbf{S}_{e_{i+1}}; [\mathbf{SEP}]_{d_i}]$ is the encoding of the text input, and $d_i = e_{i+1} - e_{i-1} + 1$. The d_i [SEP] tokens are mapped to different embeddings because they are in different context. Intuitively, they are each trained to encode the semantics of the preceding sentence with contextual information from the entire sequence (Cohan et al., 2019).

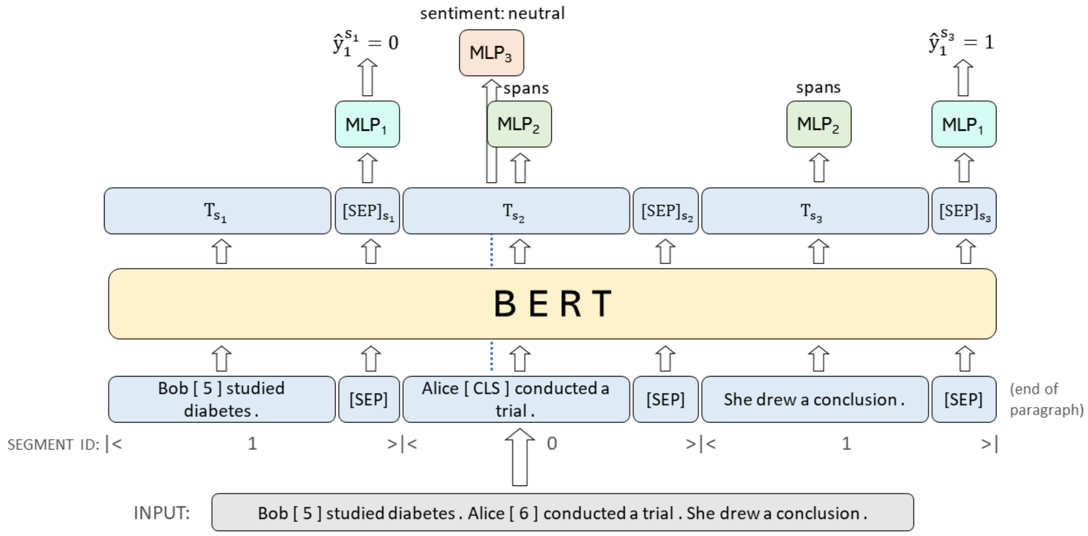

Figure 1: Multi-task citation content analysis model. The window of a citation is bounded by the previous and the next explicit citation sentences as well as paragraph boundaries.

Segment embedding Suppose citation q is explicit in sentence S_{e_i}. In the context sentence extraction task, for each of the sentences in W_q except S_{e_i}, we predict whether it is also relevant to q. From this perspective, extracting context sentences of a citation is akin to classifying a set of sentence pairs $\{(S_{e_i}, S_k); S_k \in W_q, S_k \neq S_{e_i}\}$. In BERT-based sentence pair classification, segment IDs 0 and 1 with pretrained embeddings are often used to differentiate two sentences that are concatenated as a text input. For this task, we leverage the pretrained segment embeddings to mark the position of the explicit citing sentence, which differs in each window. Specifically, we used segment ID 0 for the explicit citing sentence, and 1 for all other sentences. Experiments show that this design is crucial for the successful training of our model.

Task classifiers We use different parts of the text encoding as input to multi-layer perceptron (MLP) classifiers for respective tasks. To identify citation context, we perform binary classification on each sentence in the window except S_{e_i}, using the representation of the [SEP] token, to predict whether it belongs to the citation context

$$\hat{\mathbf{y}}_1 = \mathrm{MLP}_1\left(\{[\mathbf{SEP}]_p; 1 \leq p \leq d_i, \\ p \neq e_i - e_{i-1} + 1\}\right) \quad (3)$$

The positive sentences together with S_{e_i} constitute the context of citation q, denoted as C_q, from which we extract subject matter spans. A sentence can be written as a sequence of words $S_j = [w_j^1, \ldots, w_j^{l_j}]$, where l_j is the number of words in S_j. Likewise, we write its encoding $\mathbf{S}_j$ in terms of contextualized word embeddings, $\mathbf{S}_j = [\mathbf{w}_j^1; \ldots; \mathbf{w}_j^{l_j}], e_{i-1} \leq j \leq e_{i+1}$.

We perform binary classification on each token in the citation context to predict whether it is contained in a subject matter span:

$$\hat{\mathbf{y}}_2 = \mathrm{MLP}_2\left(\{\mathbf{w}_j^k; S_j \in C_q, 1 \leq k \leq l_j\}\right) \quad (4)$$

We use the representation of the [CLS] token to predict a sentiment label for this citation: positive, negative, or neutral.

$$\hat{\mathbf{y}}_3 = \mathrm{MLP}_3\left([\mathbf{CLS}]\right) \quad (5)$$

Loss function We use the Gradient Harmonizing Mechanism (GHM) loss (Li et al., 2019) to compute the loss value of each task. The GHM loss makes statistics of the Gradient Norm density to reweight training samples, which has shown to improve performance on noisy and imbalanced data. This loss function can be written as follows:

$$\mathcal{L}_t = \mathrm{GHM}\left(\{\hat{\mathbf{y}}_t\}, \{\mathbf{y}_t\}\right), \quad t = 1, 2, 3 \quad (6)$$

where $\{\hat{\mathbf{y}}_t\}$ is the set of predictions for a task t on all citations in the training data, and $\{\mathbf{y}_t\}$ is the set of corresponding labels. We sum up the task losses to optimize them jointly. Following Cipolla et al. (2018), we use learned parameters $\{\sigma_t\}_{t=1}^3$ to dynamically adjust the loss weights

$$\mathcal{L} = \sum_{t=1}^{3} \frac{1}{\sigma_t^2} \mathcal{L}_t + log(\sigma_t) \quad (7)$$

Friendly Adversarial Training Adversarial training has been shown to improve the generalization of NLP models (Miyato et al., 2017). Zhang et al. (2020) proposed the Friendly Adversarial Training (FAT) method, which reaches a good balance between the generalizability and robustness of neural models. Instead of finding the most adversarial example under constraints maximizing the loss, they find the least adversarial example minimizing the loss as long as it is confidently misclassified by the model. It can be written as:

$$\tilde{x}_i = \arg\min_{\tilde{x} \in B_\epsilon(x_i)} l(f(\tilde{x}), y_i)$$

$$\text{s.t. } l(f(\tilde{x}), y_i) - \min_{y \in \mathcal{Y}} l(f(\tilde{x}), y) \geq \rho \quad (8)$$

where $B_\epsilon(x_i)$ is a closed ball of radius ϵ centered at x_i, and ρ is a margin representing the confidence of the adversarial example being misclassified. To prevent over-fitting and improve performance, we fine-tuned our model with FAT, which was implemented as an early stopped version of the Projected Gradient Descent (PGD) method (Madry et al., 2019).

2.3 Experimental Setup

We used PubmedBERT (Gu et al., 2021) as the pre-trained language model (containing 110M parameters), and implemented our method with Hugging Face Transformers (Wolf et al., 2020). We first conducted cross validation to find the best batch size among {8, 16, 32}, learning rate among {1e-5, 2e-5, 5e-5} and number of training epochs between 4 to 10 by random search on different combinations. We then chose the batch size of 8, learning rate of 2e-5, and 5 training epochs. The training of the our joint model (base) took about one hour on Google Colab with a P100 GPU, and 4 hours with adversarial training. We evaluated our model using a 80-20 training/test split and averaged our results over 5 random seeds. In addition to evaluating the performance of our joint citation content analysis model, we also assessed the effect of removing one or two tasks on the remaining task(s). As baseline for each task, we consider the single task model based on the same BERT architecture.

3 Results

Table 2 shows descriptive statistics of the corpus. We observe that implicit context sentences (those without the citation marker) constitute 7.2% of all candidate sentences and that more than 75% of the sentiment labels are neutral, indicating the data for these tasks are imbalanced. Citations indicating disagreement (negative sentiment) are rare (7.4%), as has been observed in similar work (Athar, 2014). On average, there are 0.24 implicit sentences per citation. While not very high, when they occur, implicit sentences often include informative context for the citation (as shown in Example 1). Subject matter spans are typically long and, on average, correspond to about half of the context window. Each citation context window contains about 1.7 disjoint subject matter spans, suggesting that discussion of points from the reference paper can be diffuse within the context window (Table 1 row 3).

General characteristics	
Number of articles	285
Number of sentences	11,845
Number of words	338,750
Number of citations	4,182
Context sentences	
Number of implicit context sentences per citation	0.24±0.59
Number of candidate context sentences per citation	3.39±2.05
Ratio of implicit context sentences	7.2%
Subject matter spans	
Number of subject matter words per citation	20±15
Number of words in each citation context	40±21
Number of words in each subject matter span	12±9
Ratio of positive words (words inside a subject matter span)	49.2%
Sentiment	
Neutral	3,172 (75.8%)
Positive	702 (16.8%)
Negative	308 (7.4%)

Table 2: Descriptive statistics of the corpus

The evaluation results for our joint model are shown in Table 3. We use F_1 score as the evaluation metric for context sentence classification. Because subject matter spans are typically much longer than typical named entities, we consider partial match better than exact match and use the average IoU score for subject matter extraction. We use macro-F_1 and accuracy to evaluate citation sentiment classification, in line with previous work on this corpus (Xu et al., 2015; Kilicoglu et al., 2019).

The results show that joint model improves performance broadly by enabling effective knowledge

Model	Context		Subject matter		Sentiment	
	F_1	Δ	IoU	Δ	F_1	Δ
Joint (base)	61.18	-	**74.54**	-	76.05	-
Joint (FAT)	**62.14**	+0.96	73.90	-0.66	**76.88**	+0.83
Ablating tasks	-	-	73.34	-1.20	75.59	-0.46
	61.04	-0.14	-	-	75.06	-0.99
	59.93	-1.25	73.80	-0.74	-	-
	59.82	-1.36	-	-	-	-
	-	-	73.89	-0.65	-	-
	-	-	-	-	74.22	-1.83

Table 3: Performance of our joint citation content analysis model on the test split and effects of ablating different tasks. "Δ" corresponds to the difference from the Joint (base) model. In the ablating task rows, if a cell is empty, it corresponds to training the multi-task model without the data corresponding to the task of that column.

Model	Overall		Per Category			
	Accu.	Macro-F_1	Cat	Pr.	Rec.	F_1
Joint model (this paper)	**87.4**	**76.1**	Neutral	**93.4**	91.7	92.5
			Positive	77.0	**79.2**	**78.1**
			Negative	55.3	**61.8**	**58.3**
Single model (this paper)	86.5	74.2	Neutral	87.6	96.3	91.7
			Positive	77.1	73.4	75.2
			Negative	58.5	54.9	56.6
(Kilicoglu et al., 2019)	88.2	72.1	Neutral	89.5	**98.2**	**93.7**
			Positive	78.3	68.1	72.8
			Negative	**93.0**	34.1	49.7
(Xu et al., 2015)	87.0	71.9	Neutral	88.6	96.6	92.4
			Positive	**82.3**	64.4	72.3
			Negative	71.1	39.9	51.1

Table 4: Comparison of our models with previously reported results on sentiment classification. Best results are shown in bold.

sharing across tasks. We observe that removing one task or two tasks from multi-task learning consistently decreases the performance of the remaining task(s). Compared to the baseline (single-task), we observe a 1.36% increase in absolute points for context classification (row 4 in Table 3), 0.65% increase for subject matter extraction (row 5), and 1.83% increase for sentiment classification (row 6). It is not surprising that the subject matter extraction is improved less by the multi-task setting, since the baseline BERT model already takes advantage of the balanced dataset for this token prediction task. Using FAT (Zhang et al., 2020) further improves the performance for context sentence and sentiment tasks by 0.96% and 0.83% respectively, despite a slight drop in the subject matter performance.

Table 4 compares the per-class sentiment classification performance to previous work. We observe that, with the joint model, macro-F_1 score is improved by 4% absolute points over the previous best result, while the accuracy is slightly lower (by 0.8%). On the other hand, recognition of positive and negative sentiment labels is significantly improved with this model (5.3% and 7.2% points, respectively). While the baseline single-task BERT model is not as successful as the joint model, it still outperforms the previously reported models, when it comes to positive and negative sentiment labels.

4 Discussion

Our hypothesis was that better resolution of citation context and subject matter would benefit citation sentiment classification. Ablation results in Table 3 show that both citation context and subject matter extraction tasks do indeed benefit sentiment classification, with their joint effect being the best. The benefit from the context classification task is expected. Since our input window often contains multiple citations, the supervision from the citation context task helps the model better focus on the context of the current citation to predict its sentiment. Moreover, we find that the subject matter extraction task plays a more important role in improving sentiment classification. To better understand the benefit brought by the subject matter extraction task, we observed examples of citations that would have been classified incorrectly without this task. Table 5 shows a selection of examples. We find that the subject matter task is helpful because it provides: (a) fine-grained localization of the content of the cited work to distinguish it from other citations or clauses comparing it to the current work within the same context (Table 5 row 1); (b) important linguistic clues showing the authors' interpretive commentary toward the the cited work (Table 5 row 2).

Citation context classification errors Some citation context classification errors were due to missing the coreference between one entity in the implicit citation sentence with another in the explicit citing sentence. Synonymy of biomedical terms had a similar effect (e.g. *ADHD* and *hyperactivity*), suggesting that infusing knowledge into the models beyond what is included in pretrained language models (e.g., explicit knowledge from UMLS (Bodenreider, 2004)) could further enhance the model performance. We also observed annotation inconsistencies, which potentially misled the model.

Subject matter span extraction errors Table 6 shows some typical subject matter extraction errors. We find three main error types: (a) the prediction omits a few words from the annotated span (row 1), possibly because the subject matter spans are too long; (b) subject matter span can be somewhat ambiguous (row 2); (c) several citations form a complex case of coordination ellipsis (row 3).

Casting the problem as span prediction (Lee et al., 2017) rather than sequence labeling could alleviate the first problem, although long spans may also lead to an explosion of candidate spans. More specific annotation guidelines could help with consistency and improve the second problem, while enhancing representations with AMR graphs (Banarescu et al., 2013) or dependency trees could help with the third problem.

Citation sentiment classification errors We observe that the main confusion in sentiment classification comes from misclassifying positive and negative citations as neutral. This is in line with previous studies, which indicate that positive and negative sentiment in scientific articles is often implicit (negative sentiment more so) (Athar, 2011). We present two types of errors in Table 7, the first involving positive polarity and the second negative, both misclassified as neutral. Note that important clues are somewhat implicit. The second example also illustrates that domain knowledge could help the model better capture the implicit sentiment (*no randomization* indicating a less rigorous study).

Limitations Our study has limitations. We find that the annotations have some consistency issues. Annotating citation context and subject matter boundaries precisely are both challenging tasks, as shown by relatively low inter-annotator agreement score for exact matches. Improving corpus quality through additional annotation and adjudication would improve model performance and utility.

We cast citation context extraction as sentence classification. Although clause level contexts occur (Abu-Jbara and Radev, 2012), they were uncommon in our data (96% IoU of context spans and sentences). We also did not consider contexts beyond adjacent citations (0.5% of the cases).

5 Related Work

Most NLP research in citation analysis has focused on the computational linguistics literature, owing to the availability of the ACL Anthology Corpus (Radev et al., 2013), which has been used to study citation significance (Athar, 2014), sentiment (Athar, 2011; Athar and Teufel, 2012), and context (Qazvinian and Radev, 2010; Abu-Jbara and Radev, 2012). The effect of multi-sentence context identification on citation sentiment has also been investigated, with contradictory results (Athar and Teufel, 2012; Abu-Jbara et al., 2013). Multi-task learning for citation content analysis has focused on citation function/provenance (Su et al., 2019) and sentiment/purpose classification (Yousif et al., 2019). In the biomedical domain, citation content analysis is relatively understudied, existing work focusing primarily on citation function (Agar-

Correct prediction	Wrong prediction	Citation context
neutral	negative	Given that pulp therapy in the hands of specialists can often have a failure rate of over 10% [47-50], and that it is quite an invasive treatment for a child to be expected to cope with, it would seem prudent to revise the recommendation made by Duggal [44].
negative	neutral	Lobo et al. did not report the ASA classification, but their exclusion criteria very likely prohibited inclusion of patients classified as ASA 3 [17]. Thus, our patients were at a higher perioperative risk due to the higher prevalence of co-morbidity and therefore, they may have benefited from a more conventional fluid intake.

Table 5: Examples of the subject matter extraction task correcting sentiment predictions. The true positive, false positive and false negative words for the subject matter task are marked in green, blue, and red respectively.

Citation Context
Endothelial dysfunction is often seen in patients with metabolic syndrome, and it is recognized as a primary pathogenic factor of atherosclerosis [4, 18].
The technique of using the consumption of morphine during PCA treatment of postoperative pain, as a measure of the effect of the analgesic regime under study, has been used in several other studies of this kind [5, 6].
CU has been widely studied throughout literature for its anti-inflammatory [13, 14], anti-oxidant [15], antibacterial [16] and wound healing [17] properties.

Table 6: Examples of subject matter span extraction errors. True positive, false positive and false negative words are marked in green, blue, and red respectively.

Citation Context
Also, Ashley [10] found a decrease in the intake of saturated fat and cholesterol by the inclusion of PMR. **As expected**, PMR + I and INU groups significantly increased total fiber intake from 13.9 to 17.5, and 13.6 to 20.8 g/d per day, respectively. An increase in dietary fiber intake is highly recommended in obese subjects [39].
An observational study of 398 ICU patients with suspected VAP reported that the mortality rate was significantly (P=0.001) lower in patients with DE (17%) than in those with no change in therapy (23.7%) or escalation (42.6%) [5]. That study, however, was observational, **with no randomization**, and other factors, such as baseline disease severity, may have influenced treatment outcomes, rather than the DE itself.

Table 7: Examples of citation sentiment wrongly predicted as neutral. Important clues are marked in bold.

wal et al., 2010) and sentiment (Xu et al., 2015; Kilicoglu et al., 2019).

6 Conclusions and Future Work

In this paper, we proposed a multi-task model to jointly address three citation content analysis tasks: citation context classification, subject matter extraction, and sentiment classification. Our experimental results show that all tasks benefit from multi-task learning. Our citation sentiment model outperformed previous best model. We also illustrated how subject matter extraction benefits sentiment classification. Finally, we observed error cases to gain insights into the remaining challenges in our models and data. These models can serve as a step toward better models of linking citation in citing papers to relevant reference paper spans and can ultimately support challenging tasks, such as citation accuracy assessment (Kilicoglu, 2018).

In future work, we plan to address data quality and consistency issues in the dataset. We will also explore methods to evaluate the contribution of external knowledge to enhance our model (e.g., UMLS embeddings (Maldonado et al., 2019)). Finally, we are interested in exploring how citation context and subject matter analysis could interact with other citation content analysis tasks, such as citation function (Jurgens et al., 2018).

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 596–606, Atlanta, Georgia. Association for Computational Linguistics.

Amjad Abu-Jbara and Dragomir Radev. 2012. Reference scope identification in citing sentences. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 80–90, Montréal, Canada. Association for Computational Linguistics.

Shashank Agarwal, Lisha Choubey, and Hong Yu. 2010. Automatically classifying the role of citations in biomedical articles. In *AMIA Annual Symposium Proceedings*, volume 2010, page 11. American Medical Informatics Association.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87.

Awais Athar. 2014. Sentiment analysis of scientific citations. Technical Report UCAM-CL-TR-856, University of Cambridge, Computer Laboratory.

Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 597–601.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):267–270.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224.

Muthu Kumar Chandrasekaran, Michihiro Yasunaga, Dragomir R Radev, Dayne Freitag, and Min-Yen Kan. 2019. Overview and Results: CL-SciSumm Shared Task 2019. In *BIRNDL@SIGIR*.

Roberto Cipolla, Yarin Gal, and Alex Kendall. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. 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)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2017. Contextualizing citations for scientific summarization using word embeddings and domain knowledge. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1133–1136.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eugene Garfield. 1972. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479.

Steven A Greenberg. 2011. Understanding belief using citation networks. *Journal of evaluation in clinical practice*, 17(2):389–393.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1).

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. 2015. Bibliometrics: the Leiden Manifesto for research metrics. *Nature*, 520(7548):429–431.

J. E. Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *JAMIA*, 12(3):296–298.

Kokil Jaidka, Michihiro Yasunaga, Muthu Kumar Chandrasekaran, Dragomir Radev, and Min-Yen Kan. 2019. The CL-SciSumm shared task 2018: Results and key insights. *arXiv preprint arXiv:1909.00764*.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Halil Kilicoglu. 2018. Biomedical text mining for research rigor and integrity: tasks, challenges, directions. *Briefings in Bioinformatics*, 19(6):1400–1414.

Halil Kilicoglu, Zeshan Peng, Shabnam Tafreshi, Tung Tran, Graciela Rosemblat, and Jodi Schneider. 2019. Confirm or refute? a comparative study on citation sentiment classification in clinical research publications. *Journal of Biomedical Informatics*, 91:103123.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, David Jurgens, Arman Cohan, and Kyle Lo. 2021. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. *arXiv preprint arXiv:2107.00414*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Buyu Li, Yu Liu, and Xiaogang Wang. 2019. Gradient harmonized single-stage detector. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8577–8584.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards deep learning models resistant to adversarial attacks.

Ramon Maldonado, Meliha Yetisgen, and Sanda M Harabagiu. 2019. Adversarial learning of knowledge embeddings for the Unified Medical Language System. *AMIA Summits on Translational Science Proceedings*, 2019:543.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 584.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 689–696.

Vahed Qazvinian and Dragomir R. Radev. 2010. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 555–564.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

Henry G Small. 1978. Cited documents as concept symbols. *Social studies of science*, 8(3):327–340.

Xuan Su, Animesh Prasad, Min-Yen Kan, and Kazunari Sugiyama. 2019. Neural multi-task learning for citation function and provenance. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 394–395.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006a. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL'06, pages 80–87.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006b. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP'06, pages 103–110.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop*, pages 21–26.

Ludo Waltman. 2016. A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2):365–391.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jun Xu, Yaoyun Zhang, Yonghui Wu, Jingqi Wang, Xiao Dong, and Hua Xu. 2015. Citation sentiment analysis in clinical trial papers. In *AMIA Annual Symposium Proceedings*, volume 2015, page 1334. American Medical Informatics Association.

Abdallah Yousif, Zhendong Niu, James Chambua, and Zahid Younas Khan. 2019. Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing*, 335:195–205.

Guo Zhang, Ying Ding, and Staša Milojević. 2013. Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, 64(7):1490–1503.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. 2020. Attacks Which Do Not Kill Training Make Adversarial Learning Stronger. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11278–11287. PMLR.

Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427.