Studies of Citation Sentences in the Scientific Literature

Anjali Ramaprasad, Kaivan Shah, Kathryn Ricci and **Purujit Goyal**
College of Information and Computer Sciences
University of Massachusetts Amherst

Abstract

Citations provide a formal reference to a published or unpublished source that you consulted and obtained information from while writing your research paper and thus, have a fundamental role in research. In the past, there have been studies to identify the genuineness of a citation sentence, but without considering the provenance of the citation. Therefore, in this project we propose to use these citation sentences to identify the claim referred by the citation sentence in the full text document. We use a transformer based encoder-decoder model to obtain multi-codebook embeddings of the sentences and calculate similarity scores. Finally, we rank candidate sentences based on these scores and eventually extract the claim sentences. However, since our domain is quite niche, we evaluate the performance of our model using alternative evaluation metrics of Mean Average Precision and Extractive Summarization.

1 Introduction

Citations in the scientific literature serve to ground new claims in previous evidence and enable scientific argument across papers. All papers contain claims, which may then be cited by other papers within citation sentences, and these papers may be cited in turn, thereby creating a trace of evidence. Citation sentences, in general, are paraphrases of the original claim, however, there is a big problem associated with that. The authors might refer to data in other papers but interpret it differently to the original authors. This can be misleading if the implication is that the reinterpretation is the view of the original authors. Greenberg (2009) describes this as claim distortion. This situation arises when the meaning of a claim is misrepresented, even slightly, when it is cited and when that citing paper is later cited, the change to the original claim can be compounded, and so on again and again through the citation network.

Here is an example of claim distortion. Consider the original claim to be the sentence

> "A **few** fibres of β-amyloid precursor was found in 2 of 5 patients with inclusion body myositis (IBM)."

However, this has been cited as

> "βAPP in s-IBM fibers has been **confirmed** by others."

This distortion could lead to further inaccuracies and the original claim is lost.

In the biomedical domain, claim distortion can divert the course of the research, delaying real results or producing incorrect ones. Moreover, this is pervasive and difficult to detect. The extent to which citations in biomedical articles support authors' claims or whether they are inaccurate, has been probed by a number of studies, an overview is provided by Jergas et al. (2015). What we propose is a first step towards tackling this problem, because having a trace of the evidence will allow claim distortions to be more easily detected.

Our primary task is to identify the claim that is being referred to in a citation sentence given the full text of the cited paper, which will enable the extraction of such evidence traces for further analysis. This could be modeled as an entailment task, however, since our domain is quite niche, obtaining annotated data is expensive. Therefore, we try to solve this problem in an unsupervised manner and model the problem as a sentence similarity task and for our evaluation metric, we propose a secondary task of extractive summarization, where we want to select the set of citation sentences that best summarizes the cited paper. Although the ground truth summaries are also not available for this task, abstract and the title of the cited paper can work as

a good substitute for the same.

Citation sentences generally cover multiple topics in one sentence, which are hard to analyze just using a single embedding without supervision. Hence, we try to learn the multi-mode representation of a sentence in a co-occurring word distribution space, where each embedding represent a cluster center of possible co-occurring words. To learn these multi-codebook embeddings, we deploy the sequence to embeddings approach from Chang and McCallum (2020), which attains strong performance on STS (Cer et al., 2017), for sentence similarity and CNN/Daily Mail dataset (Hermann et al., 2015) for extractive summarization. We train our model using pre-trained word2vec embeddings which outperforms the single mode based word2vec baselines while achieving similar performance to the sent2vec and BERT (Devlin et al., 2019) based baselines.

Finally, to evaluate the applicability of our approach, we are currently training our model on a different domain, the ACL corpus and eventually plan to submit our results to the Scholarly Document Processing @EMNLP 2020 Workshop.

2 Related Work

Citations play an integral role in scientific development. They help disseminate the new findings and they allow new works to be grounded on previous efforts. While several studies have investigated the problem of the intent behind a citation (Cohan et al., 2019; Jergas et al., 2015), significantly fewer have addressed citation provenance. Wan et al. (2009) examined researchers' literature browsing habits and reported that while encountering citations, readers would find it useful if there is an agent that identifies or extracts the relevant sentences in the cited paper that justifies the citation. This suggests the convenience of having an intelligent browsing tool with citation provenance support.

Low presented the first automated tool to identify citation provenance, which uses a two-step approach. First, it classifies citations into either general or specific and then identifies the relevant reference texts for citations marked as specific. This task started attracting more attention recently as a pre-processing step for generating multifaceted comprehensive summaries of scientific documents. In the CL SciSumm Shared Task 2016 (Jaidka et al., 2016), most systems used traditional features such as lexical and syntactic dependency cues (Aggarwal and Sharma, 2016), TF-IDF and Longest Common Subsequence (LCS) for the syntactic score (Prasad, 2017) among others. These systems show reasonable performance with a wide variance. More recently, the SciSummNet dataset (Yasunaga et al., 2019) was also released, focusing on NLP papers rather than biomedicine. However, the training dataset used by CL-SciSumm/SciSummNet for the subtask of identifying the spans of text in the Reference Paper that most accurately reflect the citation sentence is limited and noisy.

In addition to the above methods, a language model integrating word embeddings and domain knowledge was proposed to treat this problem as an information retrieval challenge (Cohan and Goharian, 2017). Further, there have been multiple studies that apply ensemble modeling to identify the cited text spans (Wang et al., 2019; Ma et al., 2018). Su et al. (2019) hypothesized that the two tasks of citation intention and citation provenance are closely related. They reformulated the problem as multi-task learning to exploit the relationship between the two tasks to enhance the overall performance.

Wadden et al. (2020), recently introduced SCIFACT, a dataset of 1.4K scientific claims paired with evidence-containing rationales coinciding with the biomedical domain our work is focused on. This is the most relevant work for our task. They employ the "BERT-to-BERT" approach for evidence retrieval and rationale selection and pretrain their model on the FEVER dataset (Thorne et al., 2018), before fine-tuning on SCIFACT.

However, all the aforementioned methods and studies are supervised or semi-supervised in nature, and to the best of our knowledge, there is no other study that tries to solve this problem in an unsupervised manner.

3 Model

Our model architecture (Chang and McCallum, 2020) is similar to the Transformer based seq2seq encoder-decoder model (Vaswani et al., 2017). The function of our encoder is to transform the input sequence into contextualized embeddings such that the sentences which are likely to have similar co-occurring word distribution are mapped closer to each other. While our encoder functions same as a typical Transformer encoder, our decoder differs in the sense that it does not need to output words

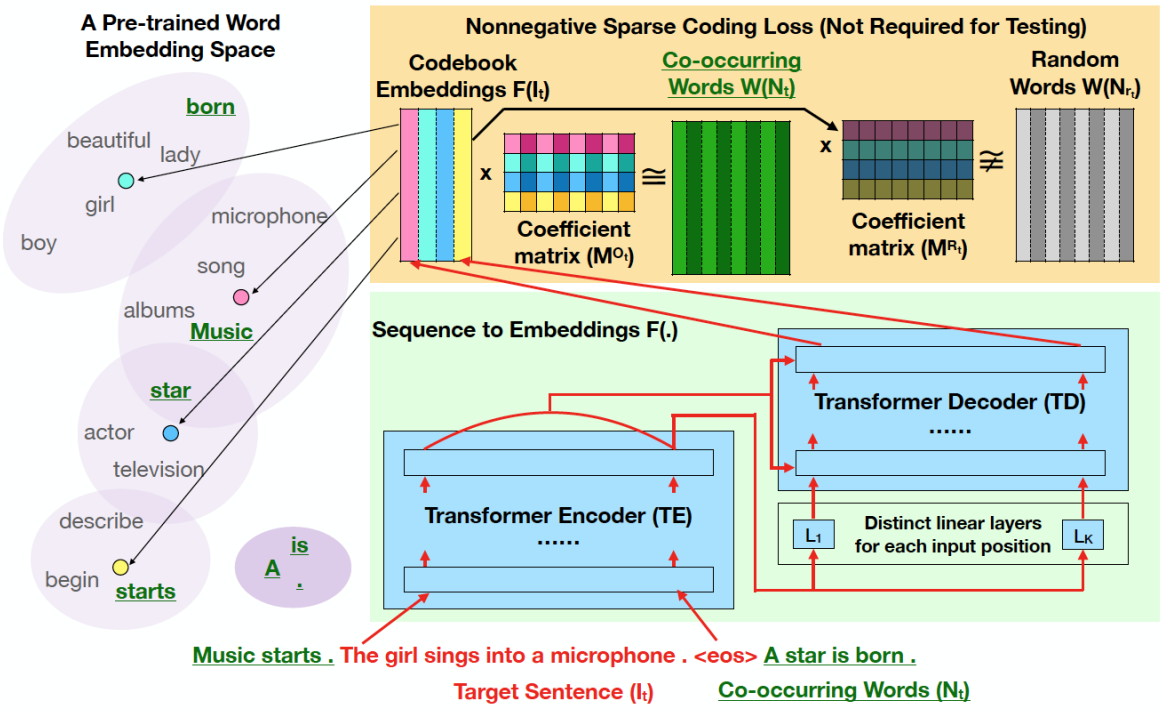

Figure 1: Our model for sentence representation. Our loss encourages the model to generate codebook embeddings whose linear combination can well reconstruct the embeddings of co-occurring words (e.g., Music), while not able to reconstruct the negatively sampled words (i.e., the co-occurring words from other sentences) to avoid predicting common topics which co-occur with every sentence (e.g., is in this example).

in an auto-regressive manner. Rather we only need a sequence of embeddings that capture different aspects of the input sentence. This allows us to predict all the codebook embeddings in a single pass. Fig. 1 gives an overview of our model.

Given an input sentence $I_t = w_{x_t} \dots w_{y_t}$<eos>, where x_t and y_t are the start and end position of the target sentence, we believe that the neighbouring words beside the input are related to some aspects of the sentence. Hence, our training objective is to reconstruct a set of neighbouring words or words that could possibly co-occur, $N_t = \{w_{x_t - d_1^t}, \dots, w_{x_t - 1}, w_{y_t + 1}, \dots, w_{y_t + d_2^t}\}$. For our experiments, we set N_t as the set of all words in the previous and the next sentence of the input sentence I_t. Finally, we use the non-negative sparse coding (NNSC) loss while training as explained in Chang and McCallum (2020). Further, to prevent the model from predicting the same global topics regardless of the input, we augment our loss function with a randomly selected sentence, I_{r_t}. Our final loss function is defined as

$$L_t(F) = Er(F(I_t), W(N_t)) - Er(F(I_{r_t}, W(N_{r_t}))$$

where F is our neural network mode, and $F(I_t)$ is the predicted cluster centers matrix, $W(N_t)$ defines the embeddings of co-occurring words N_t, N_{r_t} is a set of co-occurring words of a randomly sampled sequence I_{r_t} and $Er(F(I_t), W(N_t))$ is the reconstruction error using NNSC. We use SGD to optimize our model with respect to L_t.

4 Experiments

4.1 Datasets

We used Molecular Biology Open Access Pubmed Word and Sentence Representations (Burns et al., 2018) as a sample dataset to test our training pipeline, while CZI prepared the final dataset. This dataset contains 403,825 PMC open access documents focusing on molecular biology.

CZI has provided us with a dataset of around 23,500 open-access full-text papers on neurodegenerative diseases (tokenized on space and sentence-split). Along with the full text, we have been provided with the citation sentences in the aforementioned papers and the respective pubmedIDs for the cited papers. We also have the section title for many sentences in the dataset. Since this dataset is generated by CZI for this particular task, there is

no previous work to report.

Of the CZI corpus, the number of papers that were full-text and were cited by other papers in the corpus was 12,160. About 9,100 of these had at least two citations in the corpus, and this is the set that we used for our experiments.

For CL-scisumm summarization task, to train our model we are using ACL Anthology Reference Corpus, a corpus of scholarly publications about Computational Linguistics. The version in use has includes all ACL Anthology files whose copyright belongs to the ACL, before December 2015, consisting of 31,217 articles.

For testing, CL-scisumm has provided manually annotated dataset of 40 ACL articles and citing papers with 740 citation sentences. In the dataset, a topic consisting of a Reference Paper (RP) and upto 10 Citing Papers (CPs) that all contain citations to the RP. In each CP, the text spans (i.e., citances) have been identified that pertain to a particular citation to the RP. We modified this data for our purpose by generating a testing file with three sections divided by @highlight1 and @highlight2: section one has the citation sentence, section two after highlight 1 has the ground truth (reference text), and section three after highlight 2 has reference text as candidates in one setting and full text of the article as candidates in one setting

4.2 Baselines

Since our model is trained on word2vec embedding space, for direct comparison we chose two alternative ways to represent a sentences using the same word2vec embeddings. First is **Avg. Words**, which takes average of the word embeddings to get single-aspect sentence embedding. Second is **All Words** which concatenates the embeddings of all the words in the sentences to represent the different aspects of the sentence. Moreover, we also introduce variants of these baselines by normalizing the embeddings with the length of the sentence and the frequency of the word occurring in the corpus. We also compare our results with the sentence embeddings generated by BioSentVec (Chen et al., 2018) and add similar normalizing variants to this also.

Finally, we believe it is hard to make a fair comparison with BERT, because BERT uses a word piece tokenization which might not be fair to other methods as all of them use standard word tokenization. Nonetheless, we still present the unsupervised performance of the SciBERT (Beltagy et al., 2019) as a reference. Similar to the word2vec baselines, we represent SciBERT embeddings in two ways, first to directly use the sentence embedding, labeled as **SciBERT Avg** and second to concat the word embeddings extracted from the hidden layers of the model, denoted as **SciBERT W**.

Target sentence: Pretreatment processes , which can disrupt cellulose crystallinity and increase the porosity of the biomass , have been shown to be effective in enhancing the hydrolysis process of corn stover . <eos>	
Codebook embeddings:	
1	soybeans 0.901, corn 0.890, sorghum 0.888
2	cellulose 0.947, pectin 0.812, Cellulose 0.800
3	biomass 0.907, soil 0.804, biomasses 0.780
4	porosity 0.879, porous 0.864, coating 0.830
5	strategies 0.739, innovative 0.731, applications 0.711
6	hydrolysis 0.905, hydrolytic 0.780, transacylation 0.770
7	glycerin 0.707, tween-80 0.678, 0.3M 0.678
8	nanostructural 0.771, crystallinity 0.763, microstructure 0.749
9	increase 0.880, decrease 0.868, increases 0.868
10	investigated 0.637, study 0.618, explored 0.588

Table 1: The 3 nearest neighboring words to each embedding and their cosine similarities to the embedding.

4.3 Tasks

In this section, we describe the experiments we ran to evaluate the embeddings learned by our model in an indirect way since we don't have annotated data for our primary task of claim extraction. We trained two versions of our model, one with 10 number of clusters (**Ours N10**) and the other with 100 number of clusters (**Ours N100**).

4.3.1 Mean Average Precision

First, we wanted to evaluate how precise the embeddings learned by our model are as compared to the other baselines. For this we used mean average precision score as used for Information Retrieval ranking results. Given a reference text, we want to select (rank) the correct candidate sentences (higher) from a set of the correct ones mixed with randomly selected sentences from other papers. We experimented with two different settings for CZI dataset:

- We use abstract of the paper as the reference text, while corresponding citation sentences mixed with some randomly selected citation sentences were used as candidate sentences
- In the other setting, we invert the roles i.e. we used the citation sentences of a paper as the reference text, while sentences from the abstract again mixed with other random abstract sentences were used as candidate sentences.

We select **top k** documents, where k varies from 1 to 10. The mean average precision can be defined as

$$MAP(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{m_i}\sum_{k=1}^{m_i} Precision(R_{ik})$$

where Q is the set of queries or papers in our case and R_{ik} is the set of ranked retrieval results until you get to k correct candidate sentences.

4.3.2 Extractive Summarization

Since we did not have annotated data to directly evaluate the similarity measure between the citation and extracted claim sentence, we decided to use the task of extractive summarization as an alternative. Even though, we don't have the annotated ground truth summaries, abstract/title of the paper can serve as a good alternative. We represent an article summary S as the union of the multi-faced embeddings of the sentences in the summary

$$R(S) = \bigcup_{t=1}^{T}\{\hat{F}_u(S_t)\}$$

where $\{\hat{F}_u(S_t)\}$ is the set of column vectors (normalized codebook embeddings for the sentence S_t). We define our objective as discovering a summary S whose multiple embeddings $R(S)$ best reconstruct the distribution of normalized word embedding w in the article to be summarized (D) (Kobayashi et al., 2015), i.e.

$$\arg\max_{S}\sum_{\underline{w}\in D}\frac{\alpha}{\alpha + p(w)}\max_{\underline{s}\in R(s)}\underline{w}^T\underline{s}$$

where $\frac{\alpha}{\alpha+p(w)}$ is the importance weight given to a word in each sentence. α is a constant here, which we set to 10^{-4} while $p(w)$ is the probability of seeing the word w in the corpus. Finally, we greedily select sentences to optimize the equation defined above. The results are compared using F-1 ROUGE Scores (Lin, 2004).

4.3.3 Claim Sentence Extraction

For our primary task of extracting claim sentences, we have asked CZI to annotate a list of candidate sentences (sentences chosen from the Results/Abstract section of the paper) for 100 randomly selected citation sentences according to their relevance for the corresponding citation sentence.

Method	MAP Score (k=1)	MAP Score (k=10)
SciBERT W	99.2	94.5
SciBERT Avg	94.1	84.6
SentVec	97.8	92.2
Ours N100	98.0	93.7
Ours N10	98.1	93.6
Avg. Words	94.7	89.1
All Words	97.8	92.0
Random	50.1	37.2

Table 2: MAP scores given abstract, select correct citation sentences

We use a similar objective function as extractive summarization, as in we compare the similarity between the citation sentence and the candidate sentences and rank them accordingly. The similarity score is calculated as

$$\sum_{w\in S_c}\frac{\alpha}{\alpha + p(w)}\max(0, W(S_c)^T\hat{F}_u(S_c))$$

Finally, we plan to evaluate the correlation score between the annotated rankings and the rankings generated by our model.

4.4 Main Results

4.4.1 Experiment on CZI dataset:

To evaluate the precision of the embeddings learned by our model, in the first setting we used abstract of the paper as reference text and corresponding citation sentences mixed with an equal number of randomly selected citation sentences from other papers in the corpus as candidate sentences. All the results reported are for the frequency normalized and sentence length normalized (for all words baselines) embeddings. Table 2 include the results for this setting.

In the second setting we use citation sentences of a paper as the reference text and sentences from the abstract mixed with other random abstract sentences as candidate sentences. Again all the results are reported for frequency normalized and sentence length normalized (for all words baselines) embeddings. Table 3 displays the results for this setting.

Results for the extractive summarization task, where given an abstract, our model selects citation

Method	MAP Score (k=1)	MAP Score (k=10)
SciBERT W	99.2	88.1
SciBERT Avg	97.3	78.1
SentVec	98.9	88.7
Ours N100	98.3	88.7
Ours N10	98.4	89.2
Avg. Words	97.3	80.9
All Words	98.7	85.5
Random	50.9	28.4

Table 3: MAP scores given citation sentences, select correct abstract sentences

Method	k	ROUGE-2 F1
SciBERT Avg	1	0.052
SciBERT W	2	0.049
SentVec	1	0.054
Ours N100	1	0.051
Avg. Words	1	0.048
All Words	2	0.047
Random	1	0.033

Table 4: ROUGE-2 F1 scores given abstract, select citation sentences that best summarize it, with title as ground truth

sentences that best summarize the title of the paper. Table 4 reports the results. We display the best rouge score achieved by a method and k represents the number of sentences selected to achieve that best score.

4.4.2 Experiment on CL-SciSumm dataset:

Since, CZI dataset does not have annotated data and ground truth values we trained the model on ACL Anthology files that date before 2015. Then tested the trained model on SciSumm corpus that has 40 ACL papers which are annotated and has ground truth values under two settings. In the first setting, all cited sentences in reference paper are taken as candidate sentences. All the results reported are for the frequency normalized and sentence length normalized (for all words baselines) embeddings. Table 5 and 6 include the results for this setting.

In the second setting, all sentences in reference paper are taken as candidate sentences. All the results reported are for the frequency normalized and sentence length normalized (for all words baselines) embeddings. We also show the best ROUGE score achieved by a method in the summarization of a citation sentence by a selection from all sentences in the reference paper, with the ground truth cited reference sentences as the reference summary. k represents the number of sentences selected to achieve that best score. Table 7 and 8 include the results for this setting.

Method	MAP (k=1)	MAP (k=10)
SciBERT W	24.5	41.3
Ours	23.6	39.7
Avg. Words	18.6	36.2
All Words	21.8	39.4
Random	7.3	17.2

Table 5: MAP scores when all cited sentences in reference paper are candidates

Method	k	ROUGE-2 F1
Upper Bound	1	0.637
SciBERT W	3	0.295
Ours	2	0.299
Avg. Words	1	0.290
All words	1	0.289
Random	1	0.165

Table 6: ROUGE-2 F1 scores when all cited sentences in reference paper are candidates

Method	MAP (k=1)	MAP (k=10)
SciBERT W	9.8	13.3
Ours	7.3	11.4
Ours + Avg. Words	10	15.1
Ours + All Words	6.6	10.6
Avg. Words	9.5	14.1
All Words	9.8	15.2
All Words + Avg. W	9.9	15.2
Random	0.6	1.4

Table 7: MAP scores when all sentences in reference paper are candidates

Method	k	ROUGE-2 F1
Upper Bound	2	0.86
SciBERT W	6	0.126
Ours	3	0.110
Avg. Words	3	0.107
All Words	5	0.108
Random	7	0.04

Table 8: ROUGE-2 F1 scores when all sentences in reference paper are candidates

4.5 Analysis

4.5.1 Analysis of CZI experiment

We expect our model to do well in general compared to the other word2vec methods because Avg. Words may suffer from information loss and may not pick sentences that have extra aspects, and All Words may prefer longer sentences because extra aspects may score some similarity to unimportant aspects in the citation sentence. By having multiple embeddings per sentence but fixing the number, our model aims to overcome these disadvantages.

On the task of selecting the correct citation sentences to match the abstract, SciBERT W achieves the best MAP scores, followed by our method. On the task of selecting the correct abstract sentences to match all of the citation sentences, SciBERT W does the best at picking the first closest match (followed by SentVec, All Words, then our method). However, SciBERT W lags behind our method and SentVec when it comes to picking the first 10. The reason our model does not outperform All Words consistently might be addressed by combining our model with Avg. Words, as is done in the ACL experiment and explained in that analysis below.

SentVec, followed by SciBERT Avg, does best by the ROUGE metric, and our model, which does the best of those in word2vec space, follows. One reason ROUGE scores are not higher might be due to adopting the paper's title as the reference summary in the absence of a ground truth. The abstract will necessarily contain more aspects than the title (which also sometimes is a description of the topic of the paper rather than the findings), therefore a citation sentence may be selected to match those abstract aspects and that may have less bigram overlap with the title.

4.5.2 Analysis of CL-SciSumm experiment

In the first setting, our model trained on word2vec embedding space outperforms all the other word2vec baselines with MAP score of 39.7%. While the performance is slightly worse compared to SciBERT that has a MAP score of 41.3% but it uses a different embedding space. This means that we may expect to see an improvement in our results by modifying our model to use BERT-space as the embedding space. In terms of ROUGE-2 F1 scores, our model performs the best, compared to all the other baselines.

In the second setting (the most difficult setting, as the model must pick the correct claim sentences as judged by humans from all the sentences in the paper), our model does not perform as well, so we combined our model with sentence embedding baseline (Avg. Words), which boosts the performance. The MAP scores show that our model performance improves from 11% to 15.1%, which is close to the best-performing baseline, all word embeddings (All Words). All Words gets only a very small boost from combining with Avg. Words, showing that not all methods can be boosted this way. The reason for this might be due to the fact that our model is trained to generate embeddings corresponding to the co-occurring word distribution in adjacent sentences. It is possible that, because of the complexity of the transformer model, the model overfits the training data. For instance, consider the following sequence of three sentences:

> "Our beam search method is as follows.
> f(x) = A + B + C.
> We choose this formula to increase the fluency of our decoder."

The model might memorize the sentence "f(x) = A + B + C" so that when it appears as input it will output the word embeddings for "beam", "search", "formula", "fluency", "decoder". The consequence is that if a citation sentence contains these aspects, the model will choose "f(x) = A + B + C" as the best-matching candidate, although it is the nearby sentences of "f(x) = A + B + C" which are the good match. So by combining our method with Avg. Words, we ensure that the sentence itself is a good match. On the other hand, combining our method with All Words, the best-performing baseline in this setting, leads to a a drop in performance relative to both individual methods.

In terms of ROUGE-2 F1 scores, our model per-

forms slightly worse as compared to SciBERT but beats all the other word2vec baselines.

An example of CL-SciSumm experiment is shown in the Table 9. It demonstrates an instance of our model selecting the ground truth while Avg. Words selects an unrelated sentence, which may be due to information loss from the averaging.

Citation sentence: "The decoding algorithm employed for this chunk + weight ŏ0c3 ŏ2014 j f req(EA j , J j) based statistical translation is based on the beam search algorithm for word alignment statistical in which Ptm(J—E) and Plm (E) are translation model and language model probability , respec translation presented in (Tillmann and Ney , 2000) , tively1 , freq(EA j , J j) is the frequency for the which generates outputs in left - to - right order by consuming input in an arbitrary order"
Ground Truth: "[165] We apply a beam search concept as in speech recognition."
Ours:: "[165] We apply a beam search concept as in speech recognition."
Average Words:: "[28] The inverted alignment probability p(bijbi0001 ; I ; J) and the lexicon probability p(fbi jei) are obtained by relative frequency estimates from the Viterbi alignment path after the final training iteration ."

Table 9: SciSumm example of our model finding the correct reference text

5 Conclusion and Future Work

We showed that the multi-mode embeddings do help in alleviating the limitations of the single aspect embeddings for multiple tasks, and that it can be eventually used for the task of claim extraction. The qualitative analysis of claim sentence extraction tasks indicate that the multi-facet embeddings can be used to find the important aspects of a sentence while ignoring the redundant information which might be confusing for single-aspect representation. Similarly, qualtitative analysis of summarization experiments reveal that multi-mode embeddings seem to capture diverse aspects rather than focusing on limited number of major topics covered in the article. However, for the MAP and summarization experiment tasks, SciBERT outperforms our model on the evaluation metrics, which can be attributed to the much greater model size and parameters of SciBERT. Nevertheless, it would be interesting to see whether we can improve on SciBERT by training our model on its embeddings, which remains one of the future directions we tend to explore.

For proper evaluation of claim sentence extraction, we are currently training our model on ACL corpus and plan to submit our results to the Scholarly Document Processing @EMNLP 2020 Workshop. Furthermore, Wadden et al. (2020) recently released an annotated dataset pertaining to our domain of biomedical science. This dataset matches a citation sentence to the corresponding claim sentences in the referenced paper. Fine-tuning our model on this dataset might help us evaluate the performance and the limitations of our model in a much better and a systematic way.

6 Acknowledgements

We would like to thank our industry mentors Gully Burns and Boris Veytsman for their supervision and encouragement throughout this project. We would also like to thank our PhD mentor Haw-Shiuan Chang for his endless support and guidance. Finally, We would like to thank Professor Andrew McCallum, Rajarshi Das and Xiang (Lorraine) Li for providing their input and support in this project.

References

Peeyush Aggarwal and Richa Sharma. 2016. Lexical and syntactic cues to identify reference scope of citance. In *BIRNDL@JCDL*.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *ArXiv*, abs/1903.10676.

Gully Burns, Xiangci Li, and Nanyun Peng. 2018. Molecular Biology Open Access Pubmed Word and Sentence Representations.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Haw-Shiuan Chang and Andrew McCallum. 2020. Learning multi-facet embeddings of phrases and sentences using sparse coding for unsupervised semantic applications.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2018. Biosentvec: creating sentence embeddings for biomedical texts. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *NAACL-HLT*.

Arman Cohan and Nazli Goharian. 2017. Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries*, 19:287–303.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Steven A. Greenberg. 2009. How citation distortions create unfounded authority: analysis of a citation network. In *BMJ*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. 2016. Overview of the cl-scisumm 2016 shared task. In *BIRNDL@JCDL*.

Hannah Jergas, Christopher Baethge, and Elizabeth Wager. 2015. Quotation accuracy in medical journal articles—a systematic review and meta-analysis. In *PeerJ*.

Hayato Kobayashi, Masaki Noguchi, and Taichi Yatsuka. 2015. Summarization based on embedding distributions. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1984–1989.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shutian Ma, Jin Xu, and Chengzhi Zhang. 2018. Automatic identification of cited text spans: a multi-classifier approach over imbalanced dataset. *Scientometrics*, 116:1303–1330.

Animesh Prasad. 2017. Wing-nus at cl-scisumm 2017: Learning from syntactic and semantic similarity for citation contextualization. In *BIRNDL@SIGIR*.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *ArXiv*, abs/2004.14974.

Stephen Wan, Cécile Paris, Michael Muthukrishna, and Robert Dale. 2009. Designing a citation-sensitive research tool: An initial study of browsing-specific information needs.

Pancheng Wang, Shasha Li, Haifang Zhou, Jintao Tang, and Ting Wang. 2019. Cited text spans identification with an improved balanced ensemble model. *Scientometrics*, 120:1111 – 1145.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Richard Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *AAAI*.