Team:13 NLP Project Team
Aryamaan Jain (2019121002), Avani Gupta (2019121004)

Project: Scientific Document Summarization shared task

Introduction
Summarization of scientific articles is a widely studied problem. Incorporating the additional information like facets of research information along with the summaries of papers written by other papers citing it can be useful for the task of summarization. Citations contain meta-commentary and provide additional contextual information about a reference paper. Since the citances contain some specific excerpts of the paper depending on their use-case identifying which exert in reference paper a particular citance is referring to is helpful. Thus defining the summarization task of the reference paper based on its citation papers along with using abstract of the reference paper helps by providing additional contextual information ([1],[10],[11])

Shared Task
Shared task as defined by CL-SciSumm Shared Task [1] is as follows:
Given: A topic with a Reference Paper (RP) and up to ten Citing Papers (CPs), all of which include citations to the RP. The text spans (i.e. citances) that belong to a specific citation to the RP are identified in each CP.
Task 1a: For each citance, find the text spans in the RP that most properly reflect the citance (cited text spans).
Task 1b: Determine which facet of the article each referenced text span belongs to from a list of facets which are Method, Aim, Implication, Results or Hypothesis.
Taks 2: Create a structured summary of the RP using the cited text spans of the RP.

Dataset
CL-SciSumm Shared Task 2019 [1]
It consists of a manually annotated training set of 40 articles (Tasks 1a and b) and citing papers, human-written summaries (1040 documents) for them, and a further 1000 document corpus (ScisummNet[1]), an auto-annotated noisy dataset with thousands of article-citing RP.

Evaluation
Task 1: Scored by the overlap of text spans.
Task 2: ROUGE metrics between system output and ground truth(summary from the reference span and abstract of RP.

Challenges
- Citing sentences might be more biased or noisy, compared to the corresponding cited text spans extracted from a reference paper itself.
- The manually annotated dataset provided by the ScisummNet is small and the auto-annotated dataset is noisy since it is automatically generated by the DL models.

Techniques Planned
We plan to explore BERT-based encoders for the task of identification of cited text spans as well as summarization. We plan on experimenting with encodings of BERT and its variations along with ELMo. We would explore CNN's with BERT features as an additional task depending on the available time. We also plan to implement the basic similarity methods like TF-IDF, LSA, LDA, SVM for text span

identification and use other sequence-to-sequence models like RNN's, LSTM's as well as CNN-based architectures for summarization.

Timeline

Implementation/work	**Tentative Time**
Preprocess data, try task 1a	20th Nov
Implement task 1b (will be done based on the availability of time)	23rd Nov
Implement summarization: task2	27th Nov

Literature review

Early systems used similarity measures for modeling the relationship between citing and cited sentences like TF-IDF, LSA(latent semantic analysis)[5], informativeness measures such as point-wise mutual information (PMI), Jaccard similarity. Other methods used Word Movers Distance (WMD) in combination with LDA to infer the relevance between two text spans followed by linear classifiers like Random forest, SVM, ensemble, etc.

The Neural networks and embeddings were introduced by [7] which used the TF-IDF model with an ANN. CNN's, LSTM's, embeddings of Word2Ve and Glove, pre-trained deep bidirectional transformers and document level encodes based on BERT and its variants have been widely used([6],[4],[8]). BERT [6], XLNet [7], CNN+BERT, CNN+SciBERT, CTRL [9], BiMPM [7] have been overserved to give good results[4] for summarization as well as cited text spans identification.

References:

1. Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A. R., Li, I., Friedman, D., & Radev, D. R. (2019, July). Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 7386-7393).
2. Zerva, Chrysoula, et al. "Cited text span identification for scientific summarisation using pre-trained encoders." *Scientometrics* 125.3 (2020): 3109-3137.
3. Yeh, J. Y., Hsu, T. Y., Tsai, C. J., & Cheng, P. C. (2017, February). Reference scope identification for citances by classification with text similarity measures. In *proceedings of the 6th international conference on software and computer applications* (pp. 87-91).
4. Zerva, C., Nghiem, MQ., Nguyen, N.T.H. *et al.* Cited text span identification for scientific summarisation using pre-trained encoders. *Scientometrics* 125, 3109–3137 (2020). https://doi.org/10.1007/s11192-020-03455-z
5. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.
6. Nomoto, T. (2016, June). NEAL: A neurally enhanced approach to linking citation and reference. In *Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)* (pp. 168-174).
7. Wang, Z., Hamza, W., & Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*.
8. Miller, D. (2019). Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.
9. Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
10. Chandrasekaran, M. K., Yasunaga, M., Radev, D., Freitag, D., & Kan, M. Y. (2019). Overview and results: Cl-scisumm shared task 2019. *arXiv preprint arXiv:1907.09854*.
11. Jaidka, K., Yasunaga, M., Chandrasekaran, M. K., Radev, D., & Kan, M. Y. (2019). The cl-scisumm shared task 2018: Results and key insights. *arXiv preprint arXiv:1909.00764*.