Building a Synthetic Biomedical Research Article Citation Linkage Corpus

Sudipta Singha Roy, and Robert E. Mercer

Department of Computer Science, University of Western Ontario, London, Ontario, Canada

MOTIVATIONS

- **Creating a citation linkage framework for the biomedical research articles by means of utilizing semantic similarity in between citing and candidate cited text span**
- **Building an automatically generated silver standard corpus for training the model**

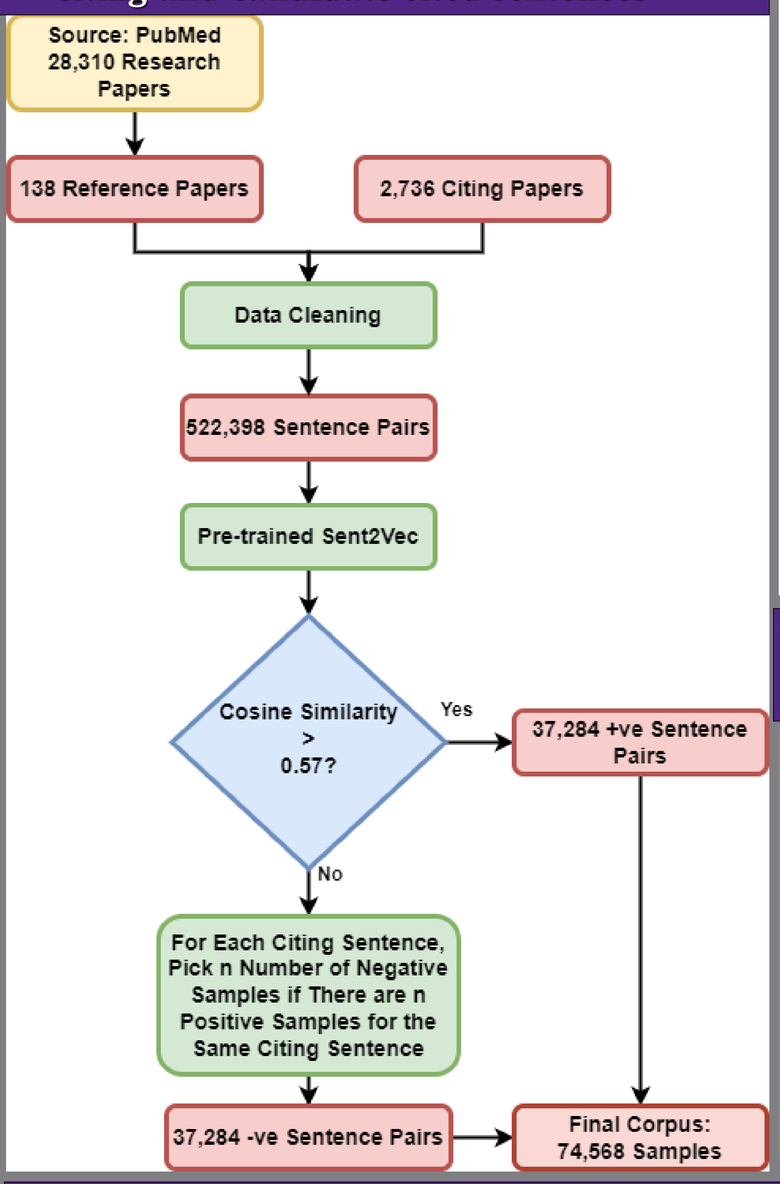
Citation Sentence	Formalin fixation, the most widely used fixative in histopathology, has many advantages such as the ease of tissue handling, the possibility of long-term storage, an optimal histological quality and its availability in large quantities at a low price.	Paraphrase
Cited Sentence	The advantages of formalin fixation are the ease of tissue handling, the possibility of long-term storage of wet material, and its low price.	
Citation Sentence	Sample DNA is often damaged by exposure to formaldehydeand a potentially extremely acidic environment	Different representation of the same condition
Cited Sentence	However, DNA is relatively stable in mildly acidic solutions, but at around pH4 the beta glycosidic bond is in the purine bases are hydrolysed.	
Citation Sentence	Different PCR buffer systems and/or different Taq poly- merases may yield different real time PCR results.	Interpretation
Cited Sentence	A significant difference can be seen between the results from the different DNA polymerase-buffer systems.	

CONTRIBUTIONS

- Have created a synthetic corpus for the citation linkage task focused on the biomedical research articles.
- We propose an approach for creating synthetic corpus without using expert annotators' opinion.
- This introduced corpus can be used for both citation linkage and semantic similarity measurement task.
- The results for models trained with this data, and the statistical analysis proves the effectiveness of this corpus.

CORPUS CREATION

- **Sent2Vec for generating sentence representation**
- **Cosine similarity measured between the citing and candidate cited sentences**



Data Cleaning

Purpose	Regex command
Capture equations	[(((([\[\s]*([^\s(((([\[\)]))]\]%*+/:<=>\-\^|~¬±·×÷ΘθЋ←→↔↦≒⇔∀∇∈∉∈∌∏∑∓∘∨∝∧∩∪∮≁≤≥≰⊂⊄⊆⊇⊈⊑⊕⊙⊣⊤⊥∪·|╨▷○∘⟶⫫]+(\s*[(((([\[]+\s*)+)?(∃|∫|log|ln)?\s*([^\s(((([\[)]))]\]+|(\d+[\.,])+\d+)?[)]))]\]]*([(((([\[\s]*[%*+/:<=>\^|~¬±·×÷ΘθЋ←→↔↦≒⇔∀∇∈∉∈∌∏∑∓∘∨∝∧∩∪∮≁≤≥≰⊂⊄⊆⊇⊈⊑⊕⊙⊣⊤⊥∪·|╨▷○∘⟶⫫]+\s*-?[(((([\[\s]*(∃|∫|log|ln)?\s*([^\s(((([\[\)]))]\]]+|(\d+[\.,])+\d+)\s*([(((([\[\s]*[^\s(((([\[\)]))]\]%*+/:<=>\^|~¬±·×÷ΘθЋ←→↔↦≒⇔∀∇∈∉∈∌∏∑∓∘∨∝∧∩∪∮≁≤≥≰⊂⊄⊆⊇⊈⊑⊕⊙⊣⊤⊥∪·|╨▷○∘⟶⫫]+|(\d+[\.,])+\d+|-)?[\s)))}\]]*)+
Capture numbers with no prior symbol	(^\d+| \K\d+)((\.\d+)|\^\d+|e[\-+]?\d+(\.\d+)?)?(?![\-\w])
Capture numbers with prior symbol	(?<!(\w|\d))[\-+]\d+((\.\d+)|\^\d+|e[\-+]?\d+(\.\d+)?)?
Capture citations	\[\d+([,\-]\d+)*\]
Capture chemical names ending with ∝ and β	-\s*∝|-\s* β

VALIDATION OF THE AUTOMATICALLY GENERATED CORPUS

	Annotator Group 1	Annotator Group 2	The Automatically Generated Corpus
Positive samples (in total)	731	709	750
Negative Samples (in total)	769	791	750

Table 2: Analysis of the agreements among the expert annotators and the automatically generated corpus

	Between Annotator Groups 1 and 2	Between Annotator Group 1 and the Automatically Generated Corpus	Between Annotator Group 2 and the Automatically Generated Corpus
Agreed Positive Samples	706	715	701
Agreed Negative Samples	765	750	750
Cohen's κ	0.96	0.95	0.93

HYPER-PARAMETER SETTINGS USED FOR SENT2VEC TRAINING

Table 3: Hyper-parameter settings used for training Sent2Vec. The selected parameter values are marked as bold.

Hyper-parameters	Values
Embedding Dimension	700/600/**500**/400/300/200
Iterations	20/15/**10**/5
Window Size	**20**/10
Learning Rate	**0.2**/0.1/0.05/0.01
Negative Samples	**10**
Loss Function	softmax/ Hierarchical softmax/ **Negative sampling**
Sampling Threshold	**0.0001**

RESULTS

Table 4: Performance analysis of different models trained with the gold corpus (Houngbo and Mercer, 2017). The test set contains 400 samples from (Houngbo and Mercer, 2017). The performance metrics are TP: true positive; FP: false positive; TN: true negative; FN: false negative, P: precision, R: recall, F1: F1 score, MCC: Matthews correlation coefficient; Acc: accuracy, BAcc: balanced accuracy.

Model	TP	FP	TN	FN	P	R	F1	MCC	Acc (in %)	BAcc (in %)
hCNN	2	0	390	8	1	0.2	0.33	0.44	98	60
Bi-LSTM & Max-Pooling	1	0	390	9	1	0.1	0.18	0.31	97.75	55
Bi-LSTM & Inner Attention	1	2	398	9	0.33	0.1	0.15	0.17	97.25	54.74
Bi-LSTM & Hierarchical Attention	1	2	398	9	0.33	0.1	0.15	0.17	97.25	54.74
CT-Transformer	2	1	389	8	0.67	0.2	0.31	0.36	97.75	59.87
DT-Transformer	1	2	398	9	0.33	0.1	0.15	0.17	97.25	54.74

Table 5: Performance analysis of different models trained with the synthetic silver corpus. The test set contains 400 samples from (Houngbo and Mercer, 2017). The performance metrics are the same as for Table 4.

Model	TP	FP	TN	FN	P	R	F1	MCC	Acc (in %)	BAcc (in %)
hCNN	7	9	381	3	0.44	0.7	0.54	0.54	97	83.85
Bi-LSTM & Max-Pooling	7	7	383	3	0.5	0.7	0.58	0.58	97.5	84.10
Bi-LSTM & Inner Attention	8	6	384	2	0.57	0.8	0.67	0.67	98	89.23
Bi-LSTM & Hierarchical Attention	8	5	385	2	0.62	0.8	0.69	0.69	98.25	89.35
CT-Transformer	9	5	385	1	0.64	0.9	0.75	0.75	98.5	94.36
DT-Transformer	9	3	387	1	0.75	0.9	0.82	0.82	99	94.62

Table 6: Performance analysis of different models trained with the silver standard synthetic corpus. The test set contains 3057 sentence pairs from (Houngbo and Mercer, 2017). The performance metrics are the same as for Table 4.

Model	TP	FP	TN	FN	P	R	F1	MCC	Acc (in %)	BAcc (in %)
hCNN	46	576	2420	15	0.07	0.75	0.13	0.20	80.69	78.09
Bi-LSTM & Max-Pooling	53	359	2637	8	0.13	0.87	0.22	0.31	88.02	87.45
Bi-LSTM & Inner Attention	54	349	2647	7	0.13	0.89	0.23	0.32	88.38	88.43
Bi-LSTM & Hierarchical Attention	56	339	2657	5	0.14	0.92	0.25	0.34	88.75	90.24
CT-Transformer	57	315	2681	4	0.15	0.93	0.26	0.35	89.56	91.46
DT-Transformer	57	301	2695	4	0.16	0.93	0.27	0.36	90.02	91.70

CONCLUSIONS

In this paper, we introduce a synthetic silver standard corpus for the citation linkage task in the biomedical domain and also a method to annotate such a corpus without any human help or expert opinion. Performance of the models trained with this dataset reflects the effectiveness of this corpus. As we started this project a couple of years ago, we used Sent2Vec for the sentence embedding. In future work, different BERT-based models can be utilized. One limitation of this work is that the considered citation text span is limited to a single sentence only. However, in real application scenarios, the referenced text may span over multiple sentences. Keeping this in mind, we are trying to build a gold and a silver standard corpus for the citation linkage task where the text span can be single to multiple sentences.

REFERENCES

- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.
- Mahtab Ahmed, Muhammad Rifayat Samee, and Robert E Mercer. 2019. You only need attention to traverse trees. In Proceedings of the 57th Annual Meeting of the Association for Computational Lin- guistics, pages 316–322.
- Hospice Houngbo and Robert E Mercer. 2017. In- vestigating citation linkage with machine learning. In Canadian Conference on Artificial Intelligence, pages 78–83. Springer.

BIONLP 2022 @ ACL 2022