Coreference Resolution

A Social Impact Assessment

Karan Praharaj

February 10th

In the paper assigned for this week, *End-to-end Neural Coreference Resolution* (Lee et al., 2017), the authors do discuss briefly the strengths and the weaknesses of their proposed approach. However, while they say that the "model does little in the uphill battle of making coreference decisions requiring world knowledge", and also provide some examples, they do not discuss the potential downstream social impacts – positive or negative – of deploying their model in real-world applications.

One strong positive social impact I could think of with this enhanced approach could be by using it in the clinical domain, to take full advantage of the information in clinical free text. Because this coreference resolution model has shown good performance in linking related information together, it could conceivably fill in the gap of a robust coreference resolution system in clinical document classification systems. In fact, I went around looking for some literature on this, and did come across a paper where Garla et al. (2011) identified that lack of coreference resolution contributed to misclassifications in a clinical document classification system. Extracting information from clinical narratives has been a problem that has attracted focus in the NLP community, and a strong general purpose coreference resolution framework *should* be able to bolster clinical NLP systems.

A negative impact that could emanate from this system is gender bias. The proposed model is trained on English coreference resolution data from the CoNLL-2012 shared task, which was drawn from the OntoNotes corpus. The OntoNotes corpus, which is used to train most datasets, has the major issue of significant under-representation of female entities. While there is no analysis provided, one could argue that this skewedness of representation can very likely be a source of bias. It is also well established that GloVe embeddings exhibit female/male gender stereotypes to a disturbing extent (Bolukbasi et al., 2016). Since the word embeddings used in this approach are GloVe embeddings, this can be another source of gender bias. Zhao et al. (2018) propose a way to overcome this bias by augmenting the training data with a gender-swapped version of the dataset where all male entities are replaced with female ones, and vice-versa. In addition to this, they also use debiased GloVe embeddings to retrain the new coreference system.

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Vijay Garla, Vincent Lo Re, Zachariah L. Dorey-Stein, Farah Kidwai, Matthew Scotch, Julie A. Womack, Amy C. Justice, and Cynthia A Brandt. The yale ctakes extensions for document classification: architecture and application. *Journal of the American Medical Informatics Association : JAMIA*, 18 5:614–20, 2011.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1018. URL https://aclanthology.org/D17-1018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.