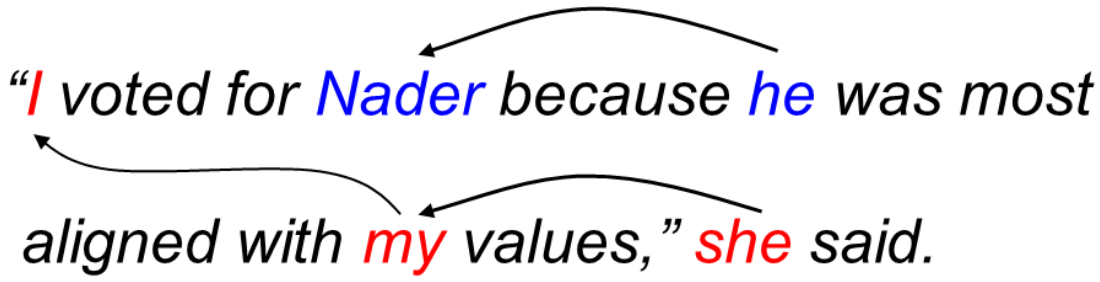
Stanford CoreNLP Coreference Resolution

Coreference resolution: What is it?

Coreference resolution is the process of linking references to the same entity in a document. For example, in this sentence:



(image © https://nlp.stanford.edu/projects/coref.shtml)

The red expressions (I, my, and she) refer to the same entity, and the blue expressions (Nader and he) refer to a different entity. It is necessary to determine which expressions refer to which entities for many different NLP tasks, such as summarization, question answering, and information extraction.

And why is it important?

Coreference resolution is an important first step for the accuracy of other NLP tasks. For instance, it is one of the first data cleaning steps involved in the SVO extraction pipeline, in order to have consistent subjects and objects. Without coreference resolution, a frequency distribution of subjects or objects, for instance, may give you a list of "he" "she" "they" that may refer to completely different entities.

Coreference Resolution with Stanford CoreNLP

We wrote a Java script – Stanford_CoreNLP_coreference_resolution.jar – that implements three different approaches to coreference resolution by Stanford CoreNLP:

Deterministic (fast rule based)

Statistical (machine learning requiring dependency parsing)

Neural network (most accurate and slowest)

The script replaces all expressions referring to the same entity in a text with one representative expression.

For example, the following sentence:

"Bill Cato attempted to assault Mrs. Vickers, but her husband stopped him."

Would become:

"Bill Cato attempted to assault Mrs. Vickers, but Mrs. Vickers's husband stopped Bill Cato."

Two types of coreference: nominal & pronominal

CoreNLP approaches implement both **pronominal** (i.e., pronouns referring to nouns, e.g., Barack Obama came to Boston; *he* said that…") and **nominal** (i.e., nouns referring to other nouns, e.g., Barack Obama came to Boston; *the President* said that…") coreference resolution. The algorithms do NOT resolve adverbial coreference (i.e., adverbs referring to nouns, e.g., Barack Obama came to Boston; *there* Obama said that…").

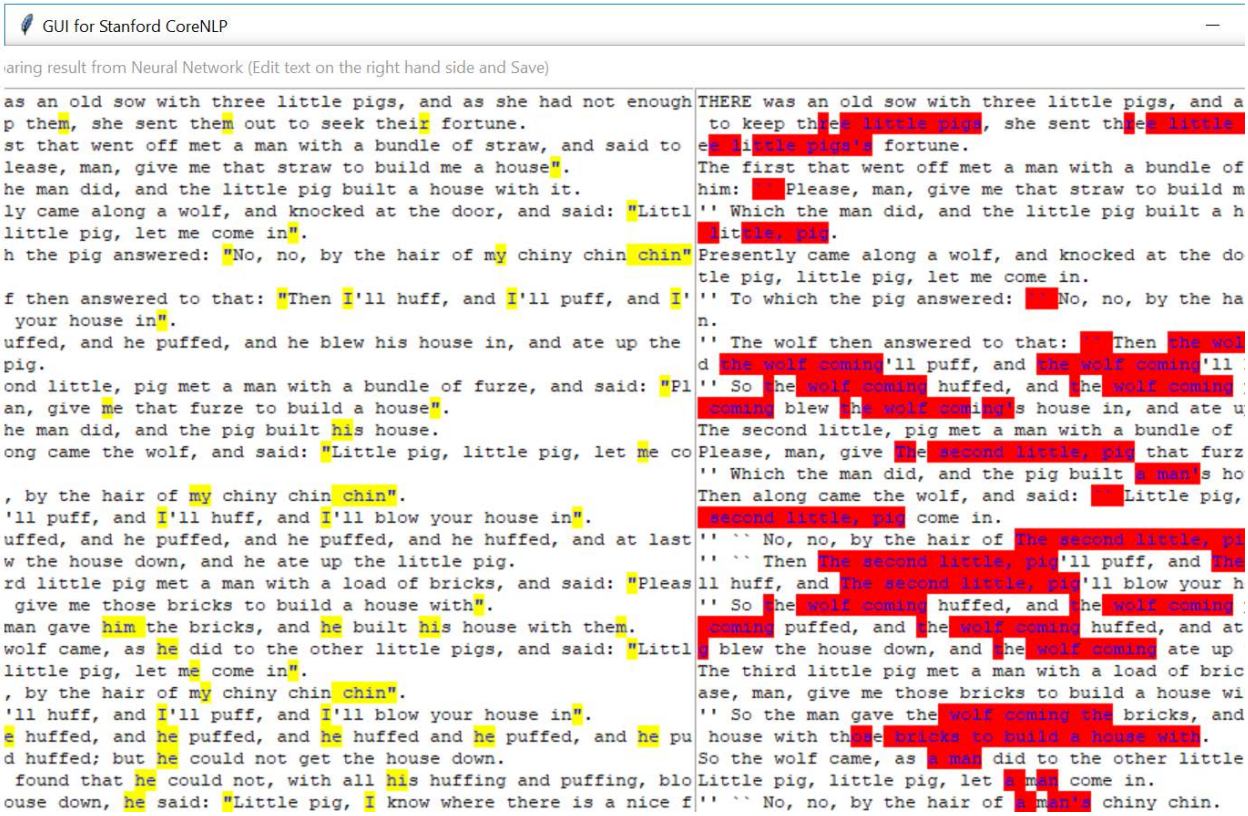
The NLP Suite SVO pipeline implementation of coreference filters out the nominal coreference and focuses on the pronominal coreference. Too many errors otherwise.

Coreference resolution is still far from accurate, with perhaps 65% success.

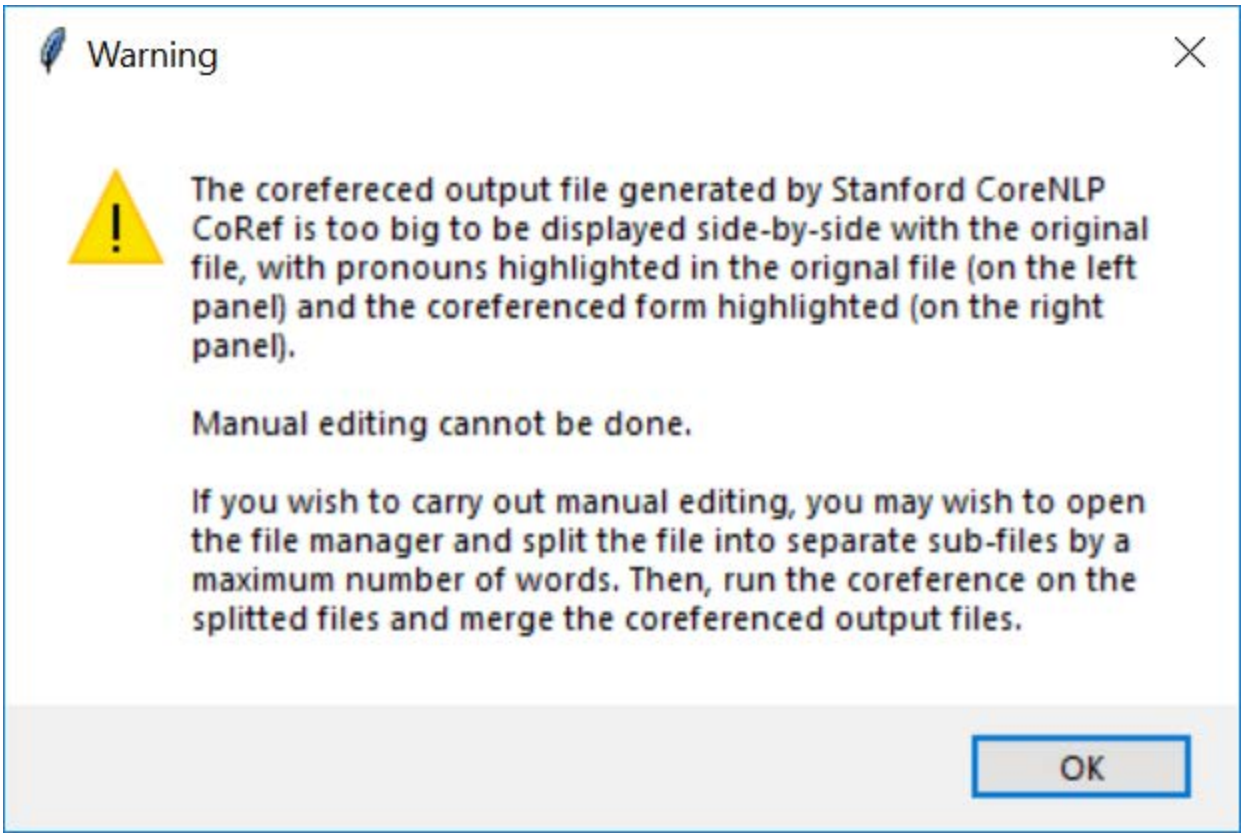
Manual coreference

Due to the relatively low success rate of coreference resolution, the NLP Suite also implements a manual approach on the coreferenced output, displaying the original and coreferenced documents on two panels, side-by-side, **original on the left** and **coreferenced on the right** with the relevant pronouns and coreferences highlighted. Although the highlighting is not perfect, it does provide

the user an immediate visual tool of comparison. The user can edit any unresolved or wrongly resolved pronominal cases.



Since the function works in memory, for large files memory this may not be an option. If that is the case, the script will warn the user.



The script similarly warns the user to deselect manual coreference when processing a directory.

References

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. "Deterministic coreference resolution based on entity-centric, precision-ranked rules." *Computational Linguistics*, 39(4).

Clark, Kevin and Christopher D. Manning. 2015. "Entity-Centric Coreference Resolution with Model Stacking." In *Proceedings of the ACL*.

Clark, Kevin and Christopher D. Manning. 2016. "Deep Reinforcement Learning for Mention-Ranking Coreference Models." In *Proceedings of EMNLP*.

Clark, Kevin and Christopher D. Manning. 2016. "Improving Coreference Resolution by Learning Entity-Level Distributed Representations." In *Proceedings of the ACL*.