NLP Techniques

Scott Patterson

June 1, 2020

1 Exploratory: Automated Complex Tokenization

June 1, 2020
I have an idea for an automated approach to complex tokenization. It is a *supervised* machine learning approach that we could test after we build the UNGA complex word dictionary this week. We could then test the complex auto-tokenizer out of sample on a set of nuclear weapons diplomacy speeches that I already have. In theory, its usefulness would grow with more training data, so it could be an appreciating asset to add to our toolbox.

The complex word tokenizer would use Parts of Speech (POS) tagging to represent the probability that a given n-gram is a complex word. POS tagging is a process whereby 1-gram tokens are labelled as a part of speech (i.e. verb, adverb, noun, etc.). I haven't done this myself, but I know that open-sourced POS-tagging tools are available. POS-tagging makes it possible to represent a string of text as a string of *parts of speech.* To illustrate, here is the textual representation of sample 6-gram complex word:

heavily indebted poor countries debt initiative

Here is the same text represented represented as a POS string:

ADV + ADJ + ADJ + NOUN + NOUN + NOUN.

From the UNGA complex word dictionary, we know that this 6-gram is semantically-unified. Therefore, we can represent this text in POS form as:

$$ADV + ADJ + ADJ + NOUN + NOUN + NOUN = 1$$

where 1 indicates semantic unity.

Similarly, here is the textual representation of a sample nonsense 6-gram:

security united nation peacekeeping operations continue

As a POS string:

NOUN ADJ NOUN NOUN NOUN VERB

and therefore:

NOUN + ADJ + NOUN + NOUN + NOUN + VERB = 0

Iterating this form of probabilistic representation over the corpus will allow us to compute the probability that a given POS-string indicates semantic unity. We could then automatically extract POS-strings that fulfil a minimum probability threshold at a specified n-gram level. After extraction, the same process could be repeated at n=5,...,2.

Again, this is a supervised approach to machine learning that uses the UNGA speeches as training data. But it could be a good way to stretch the value of the manual complex tokenization I'll be doing this week. Plus, I already have another diplomatic speech database (for nukes) on hand, so we should be able to gauge its usefulness right away.