Supplemental Material

A Combining NLTK tokenized and MetaMap tokenized sentences

Each of the sentences, premise (p) or hypothesis (h), is broken into phrases where each phrase is either a biomedical concept obtained from UMLS or is not. Let each sentence become $p_{Metamap} = [ph_1, ph_2...ph_{mp}]$ and $h_{Metamap} = [ph_1, ph_2...ph_{mh}]$ where ph stands for a phrase which may be biomedical or not. mp and mh stand for max number of phrases in premise and hypothesis respectively. Using the NLTK tokenizer on premise (p) or hypothesis (h), we also get $p_{nltk} = [w_1, w_2...w_{np}]$ and $h_{nltk} = [w_1, w_2...w_{nh}]$ where w stands for a word, np and nh stand for max number of words in premise and hypothesis respectively.

We then align each tokenized sentence obtained from NLTK(p_{nltk}, h_{nltk}) and MetaMap($p_{Metamap}, h_{Metamap}$). Let us consider the case of aligning the premise with lengths np and mp for nltk tokenized and metamap tokenized premises respectively in all cases ($np \geq mp$). Let a phrase ph_i be made of words $w_l, w_{l+1}, ...w_{l+k}$ where l is the starting position of the phrase in p_{nltk}. In this case, we will align the phrase ph_i to each word in p_{nltk} corresponding to it. So, ph_i is copied to each word in p_{nltk}. This has been illustrated in Figure 1. Through this alignment, we get the same length tokenized sentences.

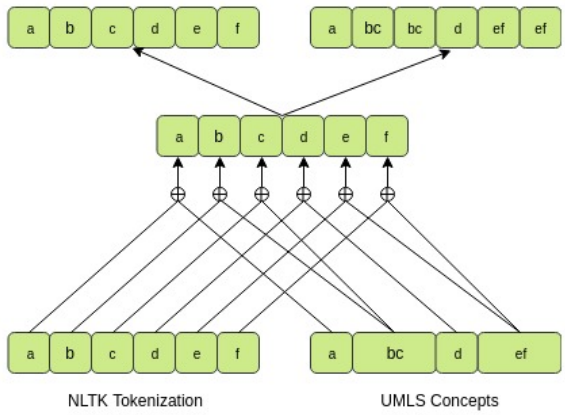

Figure 1: **Sentence Aligner**. Takes as input NLTK tokenization of a sentence (p or h) and MetaMap tokenized form of the sentence (p or h) and outputs an aligned tokenizations for both.