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 
\[ p_{\text{Metamap}} = [p_1, p_2, \ldots p_{mp}] \] 
\[ h_{\text{Metamap}} = [p_1, p_2, \ldots p_{mh}] \] 
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 
\[ p_{\text{nltk}} = [w_1, w_2, \ldots w_{np}] \] 
\[ h_{\text{nltk}} = [w_1, w_2, \ldots w_{nh}] \] 
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\((p_{\text{nltk}}, h_{\text{nltk}})\) and MetaMap\((p_{\text{Metamap}}, h_{\text{Metamap}})\). 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 \(w_l, w_{l+1}, \ldots w_{l+k}\) where \(l\) is the starting position of the phrase in \(p_{\text{nltk}}\). In this case, we will align the phrase \(ph_i\) to each word in \(p_{\text{nltk}}\) corresponding to it. So, \(ph_i\) is copied to each word in \(p_{\text{nltk}}\). 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.