Script for Presentation

One of the challenges that we encountered in our research was how to incorporate existing medical knowledge into NLP++ in order to more effectively extract information from radiology reports.

After investigating several options, we settled on a lexicon of radiology terms published by the Radiological Society of North America called RadLex. RadLex was created in 2006 by domain experts to standardize the vocabulary of radiology reports with the aim of facilitating computer interpretation.

RadLex was chosen for three reasons: it is openly available, it specific to the domain of radiology, so as to limit extraneous information, and it is exhaustive, in that it aims to supply a comprehensive list of terms included in radiology reports.

On top of this, RadLex is a medical ontology, which means that in addition to providing standardized terms, it defines relationships between them.

So after choosing an ontology, the next step was to get a better understanding of RadLex by performing some exploratory analysis.

My first aim was to create a collapsible tree diagram to provide a more intuitive interface for navigating the ontology. RadLex has predefined parent-child relationships, thus creating a hierarchy which groups terms according to classes, for example, anatomical entities or imaging modalities.

Since this was mainly implemented as a tool to explore the structure and terms of RadLex, I added a tooltip to display the definition of each of each term, if available, along with details about each subtree, including depth, number of children, and number of descendants.

The next approach I took to understanding RadLex was taking a look at individual words in the terms and analyzing how they occur across the lexicon. The first step here was generating a word cloud. Word clouds are very simple visualizations, but can offer a lot of insight into large textual datasets. Here I created a word cloud of the 100 most common words in RadLex, filtering out the 100 most common English words, which are mostly prepositions and articles, and give limited insight. Word frequency is mapped to font size, so that larger words in the cloud correspond to more frequent appearance in the lexicon.

The next step was to examine word distribution. I found that the median term length was five words and terms ranged from 1 to 18 words in length, with the longest term being “medial antebrachial cutaneous nerve component of trunk of anterior division of anterior ramus of left eighth cervical nerve”.

This, however, does not tell us anything about the distribution of words within the individual RadLex terms. To this end, I attempted to visualize the word co-occurences, or appearances in the same term, as a network where each connection represented a co-occurrence and connection weights were determined by frequency of co-occurrence. In the graphic on the right, you can see a simplification of this concept in a single term. Note that only words from the top 100 list are being included in this network. This concept was then extended to the entire RadLex lexicon, and an interactive chord diagram was created, which we can see in the next slide.

As you can see, the outer labels correspond to one of the top 100 words, ordered alphabetically, and the arc size corresponds to total word count. A color gradient was then added for visibility. Each co-occurrence is represented by a chord within the circle and the weight of each chord is relative to its frequency. Take, for example, the term from the last slide: “meningeal branch of left seventh cervical nerve.” Notice that the word “branch” occurs before the word “nerve”. As it turns out, this is a common trend in the lexicon, which can be identified by examining the visualization.

Note that the chord from branch to nerve is visibly wider at the intersection of the “branch” arc than the “nerve” arc. This is because “branch” more commonly precedes “nerve” in RadLex terms than vice versa, by a factor of almost 75.

As mentioned earlier, we are still the preliminary stages of our research. One of the design challenges we face going forward is how best to incorporate RadLex into NLP++ and, more generally, how to leverage the knowledge in the lexicon to better understand radiology reports. Some considerations in addressing this challenge include the design of knowledge bases to best accommodate our goals, as well as dealing with information outside the scope of RadLex, like non-standardized terms.

Nevertheless, NLP++ has given us the toolkit to effectively tackle these challenges moving forward.