Methods for Analysing Clinical Text

There are two main approaches to computational methods in natural language processing, rule-based, and machine learning. Both approaches are used and one could combine them at different stages of the pipeline. Currently, machine learning techniques used in natural language processing are mostly supervised learning techniques. An example of a supervised learning technique, is support vector machines. A support vector machine can be thought of as a large margin classifier. It is a vector space based machine learning method where the goal is to find a decision boundary between two classes that is extremely far from any point in the training data. The support vector machine algorithm has been widely applied in biological and other sciences. For example, it has been used to classify proteins with a large percentage of the compounds being classified correctly. However, unsupervised learning techniques are now also being considered in natural language processing. Neural networks, and deep learning, which we discussed in last weeks machine learning topic, are becoming more popular in natural language processing. A particular application of deep learning methods is word embedding. Word embeddings are techniques used to help machines learn natural language. It is where words or phrases from the vocabulary are mapped to vectors, which means text is turned into numbers. This is necessary, as many machine learning algorithms require their input to be vectors of continuous values, and won't work on strings of plain text. Finally, I'd like to briefly mention the Hidden Markov model, as it is a common method used in natural language processing. The Hidden Markov model assigns a label or class to each unit, such as words, letters or sentences in a sequence, then maps a sequence of observations to a sequence of labels. A hidden Markov model is a probabilistic sequence model. This means when given a sequence of units, the probability distribution over a possible sequence of labels is calculated, and the best label sequence is chosen. Sequence labeling tasks are common throughout speech and language processing. Particularly, with part-of-speech tagging and named entity recognition. For many years, natural language processing and the life sciences focused on medical literature, as this was the only data available. In the last 5 to 10 years, there have been more initiatives to look at clinical data, as it has become possible to access hospital data. In the UK, this requires consent, and we must keep in mind that this data contains sensitive information. So anonymization is often needed. Medical expert input is also crucial in order to make sense of this data. For instance, to create training data when employing Machine Learning methods. In the context of precision medicine, natural language processing can filter and extract clinically relevant information from unstructured patient notes. This can be done relatively quickly and efficiently without the need for manual review. For example, natural language processing can identify the stage of a patient's cancer from a pathology report. Natural Language Processing also enables unstructured patient data to be analyzed in real time, which can support clinical decisions and patient treatment options. As the importance of precision medicine continues to grow in future, natural language processing will be an essential tool in helping to make sense of the growing volume of relevant data. This will hopefully enable further advances in personalized treatment methods for each individual.