

### ***Identifying for concepts in a noise prone environment: Looking up Obesity and its 15 co-morbidities In patient discharged summaries***

Developing a tool for identifying clinical terms and concepts within a noise prone collection of clinical notes has its own requirements and issues. The specific nature of a noisy data collection raises at least two major issues. The first issue comes from a scattered matrix of evidence for a specific concept in which however has many common thereby confounding attributes with other concepts. Considering more features or patterns with the hope of covering more rare situations may lead to the absorption of more noise by the system and impact the identification of other major terms and concepts, and therefore the overall performance of the system. The second issue comes from the nature of the data collection and the necessary process for gathering evidence about existence/absent of a specific concept. Assuming 4 possible answers for a search concept namely Exists/Not Exists/Questionable/Unmentioned, biases the decision algorithm towards the two more frequent classes: Unmentioned and Exists which have unlike characteristics. The Unmentioned label has to be identified based on lack of evidence for a given search concept while an Exists label should only be assigned in presence of clear indication of a given concept. Adding more features to the feature list in the machine learner leads the system to a better and more confident classification for Exists class but at same time may lead to inaccurate results for the Unmentioned answer due to an increase in the level of the noise. We designed a customized system to address the common challenge of both issues, which is Noise reduction. Using a mixture of rules, different techniques in language processing algorithms, a decision tree classifier and some innovative solutions, a system was developed specifically for these types of noise prone corpora. We kept the number of features to monitor as low as possible based on the proposition that concepts are best defined in a few features and many features would add noise to the classifier. In a second stage, an effective noise reduction algorithm which filtered suspicious noisy features was applied to the dataset to suppress possible noise. The primary goal was to evaluate a proposed approach for processing a collection of 724 discharge summaries with a noise prone nature. Evaluation has been done against given human performance as a gold standard with precision and recall of 0.969 and 0.969 respectively.