Project plan

The project will be developed during a period of 6 months. The first task will be dedicated to establishing a baseline for the identification and classification of interactions using word embeddings and WordNet. The second subtask will be dedicated to the development of a prototype of the component where biomedical ontologies such as ChEBI are used instead of WordNet. Milestone 1 corresponds to this first prototype. Finally, after tuning the parameters and evaluating on the test corpus, the tender prototype will be published, along with a technical report and a tutorial (Milestone 2).

The software component will be shared via Docker image to the OpenMinTED platform, while its code will be published on the LaSIGE GitHub page (https://github.com/lasigeBioTM). This way, the component can be integrated into a larger text mining pipeline. The Docker image will contain all the dependencies of the component, making it easier to run the trained models and train new models. As such, the software component will have two modes: training, where the required input is an annotated corpus and ontology in compatible formats, and testing, where the required input is a model and the text to be annotated. The supported format will be the XML format used by the DDI Extraction gold standard, with the possibility of expanding to other popular formats. The software component will be thoroughly documented to improve its reusability.

During the first task, an approach based on word embeddings and WordNet will be explored. For each co-ocurrent entity pair in the text, we will extract the Shortest Dependency Path (SDP) and WordNet class of each word. For each word in that fragment we will create a vector based on the word embedding model (word2vec) proposed by Mikolov et. al., and a WordNet class vector, using the same method. WordNet will be used to extend the representations of each word with related ones, and therefore improve the probability of finding relations.

The second task will consist in developing the tender prototype that will build a interaction classification model based on given a OBO ontology (Open Biomedical Ontology) and a corpus with interactions between terms from the given ontology. Based on the model proposed by Xu 2015, we will adapt the channel architecture by using their word embeddings and WordNet channels and incorporating a OBO ontology channel. In their work, they used a WordNet channel, where each word was replaced by its WordNet hypernym. Instead, our OBO channel consists of the path of each entity of the pair to their ancestors. We consider two types of representations of an entity pairs based on their ancestors: first the concatenation of the sequence of ancestors of each entity; and second, the common ancestors between both entities. Each OBO concept will be associated with a vector of fixed length. While the training algorithm iterates through the training data, these vectors will be optimized to predict the target classes. We will analyze how each of these paths influences the performance of the model in classifying interactions, as well as when combining with the WordNet classes. We will study the effect of the vector size, as well as the difference between using the concatenation or the intersection of ancestors.

The third task will focus in applying the system to ChEBI and the DDI Corpus, which is a gold standard for Drug-Drug interactions extraction. The parameters of the system will be tuned using a partition of the training set, and training will be executed in cloud GPU services to reduce runtime and maximize the number of parameters that can be explored. The final configuration will be evaluated on the test set of the DDI corpus.