5. Conclusion

5.1. Contributions

This dissertation proposed a new feature based on the class hierarchy of entities annotated on sentences to improve the performance of classifiers for the relation extraction task. This feature is based on the selection of the class that best represents entities.

To demonstrate the effectiveness of this feature, this dissertation presented experiments involving articles in the English Wikipedia and triples from DBpedia. A corpus of sentences with annotated instances representing examples of relations was heuristically labeled using the distant supervision method. We used our feature in combination with common lexical features used for this task and we showed a substantial gain of accuracy and recall for most of the relations used. Those gains were demonstrated by two sets of experiments: held-out experiments and human evaluation experiments.

This dissertation, in general, provides an example of the use of Semantic Web resources for natural language processing tasks.

5.2. Limitations and Future work

One of the main limitations of this work refers to the use of just a few lexical features. We intend to extend the feature vector extracted from sentences by adding more lexical features, such as dependencies path (Mintz et al. 2009).

We also intend to explore how sensible is the choice of the class that represents an entity. A broader analysis can be performed to evaluate if taking the class at the middle of the hierarchy is the optimal choice or if there are different approaches to improve results.

Also, as a future work, we intend to explore different strategies to extract more sentences from an article. In this dissertation, we used a simple heuristic that annotates entities that are related to the article's subject by a simple string match between the article title and the text. Different strategies, such as coreference resolution, may improve the number of annotations, generating more examples of sentences to be used.

We also intend to obtain a different annotated corpus with entities from DBpedia to re-run our experiments with sentences in a written style different from an encyclopedia and compare the results to verify the influence of the nature of the text.

Finally, we intend to compare our results with previous work that used distant supervision to label their dataset. Our aim is to empirically demonstrate that our semantic feature can improve their results by simply adding our feature to their proposed set of features. In order to compare results, human evaluation experiments need to be performed. On the other hand, for previous results that contains human evaluation experiments, it is necessary to compute the ontology mapping between the DBpedia and the ontology used in order to obtain a list of equivalent relations.