論文の内容の要旨

On the Applicability and Representation of Lexical Prior Knowledge

(語彙的な事前知識の適用可能性とその表現について)

氏名　ステネートルプ　ポントス　ラース　エリック

In this thesis we investigate several methods for utilising lexical prior knowledge in the form of dictionaries and also unannotated text. We use the problem of out-of-vocabulary words as our main motivation for using lexical prior knowledge as a way for machine-learning based methods to generalise beyond the vocabulary observed in the training data.

Previous work, in particular for the task of Named Entity Recognition, commonly use lexical resources to alleviate some of the difficulty to generalise beyond the training data vocabulary. However, the usage of lexical resources poses a problem due to surface form variations, this is particularly true for domains such as scientific texts. In order to counter the problem of lexical variation we introduce a novel set of features that are generated using a recently introduced fast method for approximate string matching. Our task setting is to classify a textual span in its context, a task we refer to as Semantic Category Disambiguation. For evaluation we use six corpora of biomedical text. Using these novel features we are able to establish an advantage against a strong baseline consisting of a superset of commonly used features for the task of Named Entity Recognition.

Furthermore, we are able to generalise our findings to the newswire and clinical domain and investigate to which extent we are dependent on the availability of suitable lexical resources. We also motivate a task setting where a system is to act as a part of a pipeline and reduce the number of candidates passed on to another system or human. For such a task setting we argue that a system must achieve a very high level of recall in order to not cause frustration for human users and not to obfuscate potential candidates for a downstream system. We find that our method is applicable to this setting and can achieve close to perfect recall while maintaining a comparatively low level of ambiguity.

As a practical application of our theoretically motivated ambiguity re-

duction setting, we integrate our method with a recently introduced annotation tool. Doing so we reduce the amount of screen real estate needed to present the semantic categories to a human annotator and are able to show a reduction of 30.7 percent in annotator time spent selecting the appropriate semantic category and a 15.4 percent reduction in total annotation time.

Lastly we introduce the concept of word representations and investigate the domain dependence of these representations. We perform an extrinsic evaluation of a large set of different word representations induced from unannotated corpora from both the newswire and the biomedical domain for the tasks of Semantic Category Disambiguation and Named Entity Recognition. Our findings show clear benefits of using a majority of the representations and suggests that word representations, to a large extent, are domain dependent.