Language processing components of the StaViCTA project

Skeppstedt, Maria; Kucher, Kostiantyn; Paradis, Carita; Kerren, Andreas

2017

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Skeppstedt, M., Kucher, K., Paradis, C., & Kerren, A. (2017). *Language processing components of the StaViCTA project*. Abstract from Logic and Algorithms in Computational Linguistics, Stockholm, Sweden.

Total number of authors:
4

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Language processing components of the StaViCTA project

The StaViCTA project is concerned with visualising the expression of stance in written text, and is therefore dependent on components for stance detection. These components are to (i) download and extract text from any HTML page and segment it into sentences, (ii) classify each sentence with respect to twelve different, notionally motivated, stance categories, and (iii) provide a RESTful HTTP API for communication with the visualisation components. The stance categories are CERTAINTY, UNCERTAINTY, CONTRAST, REQUIREMENT, VOLITION, PREDICTION, AGREEMENT, DISAGREEMENT, TACT, RUDENESS, HYPOTHETICALITY, and SOURCE OF KNOWLEDGE.

Since standard libraries (Requests, jusText, NLTK or Flask) could be used for (i) and (iii), most work was spent on the classifiers (ii). The approach of training machine learning classifiers was preferred over a rule-based approach, as there was a requirement to provide confidence estimates for the classifications. There exist previously constructed corpora that are annotated for categories similar to some of those studied within StaViCTA, but none that is similar enough to use as training data for the models. Therefore, new training data was created by manual text annotation.

The twelve categories are not trivial to determine by human annotators (as shown by low inter-annotator agreement scores), and some of them occur rarely in most types of text. This indicates that large resources in the form of annotated data would be required to train the classifiers, and for this reason active learning was applied. The unlabelled sample closest to the separating hyperplane of a support vector machine was actively selected, i.e., an approach which had previously been shown to reduce the amount of training data required to detect similar categories. This functionality was implemented by using the MongoDB database and Scikit-learn's SVC class.

An annotation tool, developed within StaViCTA, was used to manually categorise the actively selected sentences with respect to the categories studied. In addition to this sentence-level annotation, the words that were used for expressing the categories were also marked. These were first automatically pre-annotated using the PAL tool and then checked by an annotator. The word-level annotated data was then used for training a Scikit-learn LogisticRegression classifier for performing the stance-detection task (which in general led to better results than when using the SVC classifier). The probability scores of the logistic regression model could also be used to provide confidence estimates for the stance classification.