In Proceedings of the Workshop on Logic and Algorithms in Computational Linguistics
(LACompLing '17), pages 137-138, Stockholm, Sweden, 2017.

Language Processing Components of the StaViCTA Project

Maria Skeppstedt[†,*], Kostiantyn Kucher[*],
Carita Paradis[‡], Andreas Kerren[*]

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 [3], and (iii) provide a RESTful HTTP API for communication with the visualisation components. The stance categories are CERTAINTY, UNCERTAINTY, CONTRAST, RECOMMENDATION, VOLITION, PREDICTION, AGREEMENT, DISAGREEMENT, TACT, RUDENESS, HYPOTHETICALITY, and SOURCE OF KNOWLEDGE.

Since standard libraries (jusText, Flask[1] and NLTK [1]) could be used for (i) and (iii), most work was spent on constructing machine learning classifiers (ii). These were trained on data that was created by manual text annotation of political blogs.

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 [2]. 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 [8]. 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 [7]. The approach was implemented with the MongoDB[2] database and Scikit-learn's SVC class [5].

An annotation tool, developed within StaViCTA [4], 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 [6] and then imported into

[†]Corresponding author, maria.skeppstedt@lnu.se
[*]Department of Computer Science, Linnaeus University, Växjö, Sweden
[‡]Centre for Languages and Literature, Lund University, Lund, Sweden
[1]`corpus.tools/wiki/Justext` and `flask.pocoo.org`
[2]`www.mongodb.com`

BRAT [9] and checked by an annotator. The word-level annotated data was then used for training a Scikit-learn LogisticRegression classifier to perform 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.[3]

References

[1] S. Bird. Nltk: The natural language toolkit. In *Proceedings of the Workshop on Effective Tools and Methodologies for Teaching NLP and CL*, Stroudsburg, PA, USA, 2002. ACL.

[2] Forthcoming. Annotating speaker stance in discourse: the Brexit Blog Corpus. Submitted for review, 2017.

[3] D. Glynn and M. Sjölin. *Subjectivity and epistemicity : corpus, discourse, and literary approaches to stance*. Centre for Languages and Literature, Lund University, Lund, 2014.

[4] K. Kucher, A. Kerren, C. Paradis, and M. Sahlgren. Visual Analysis of Text Annotations for Stance Classification with ALVA. In *Proceedings of EuroVis 2016 - Posters*, pages 49–51, Geneva, Switzerland, 2016. Eurographics.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] M. Skeppstedt, C. Paradis, and A. Kerren. PAL, a tool for Pre-annotation and Active Learning. *JLCL*, 31(1):91–110, 2016.

[7] M. Skeppstedt, M. Sahlgren, C. Paradis, and A. Kerren. Active learning for detection of stance components. In *Proceedings of PEOPLES*, pages 50–59, Stroudsburg, PA, USA, December 2016. ACL.

[8] M. Skeppstedt, V. Simaki, C. Paradis, and A. Kerren. Detection of stance and sentiment modifiers in political blogs. In *Proceedings of SPECOM (accepted)*, 2017.

[9] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J. Tsujii. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of EACL*, pages 102–107, Stroudsburg, PA, USA, 2012. ACL.

[3] StaViCTA is funded by the framework grant "the Digitized Society – Past, Present, and Future" with No. 2012-5659 from the Swedish Research Council (Vetenskapsrådet).