Annotating textual data for machine learning

Elisabetta Jezek[1], Ilaria Poggiolini[2], Silvia Figini[3], Flavio Ceravolo[4]

Dipartimento di Studi Umanistici[1,2], Dipartimento di Scienze Economiche e Aziendali[3], Dipartimento di Scienze Politiche e Sociali[4]

Università di Pavia

In order to analyse texts automatically with the goal of extracting information that is useful for a variety of purposes ranging from academic research to computational tasks such as opinion mining, it is nowadays standard procedure in natural language processing (NLP) to rely on machine learning (MA) instead of rule-based systems. In our contribution, after introducing traditional MA approaches (supervised, unsupervised and semi-supervised, cf. Manning and Schütze 1999), and touching upon current deep MA approaches, we will focus on supervised methods and review the methodological steps and choices that need to be made to prepare the training data for the algorithm. This step is particularly important as it involves choosing the appropriate features for the task (cf. Pustejovsky and Stubbs 2012). We will discuss examples of selected features at different levels of linguistic analysis (ranging from morphosyntax to semantics and discourse patterns), and show how they are being used in concrete case studies in both the Humanities and the Social Sciences. Finally, we will also briefly touch on the issue of evaluation, i.e. what are the best practises in the evaluation of ML systems.

References

Manning, C.D. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge MA: The MIT press.

Pustejovsky, J. and Stubbs, A. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.