Highly Multilingual News Analysis Applications

Ralf Steinberger

European Commission - Joint Research Centre, Italy

Abstract. The publicly accessible Europe Media Monitor (EMM) family of applications (http://press.jrc.it/overview.html) gather and analyse an average of 80,000 to 100,000 online news articles per day in up to 43 languages. Through the extraction of meta-information in these articles, they provide an aggregated view of the news; they allow to monitor trends and to navigate the news over time and even across languages. EMM-NewsExplorer additionally collects historical information about persons and organisations from the multilingual news, generates co-occurrence and quotation-based social networks, and more. All EMM applications were entirely developed at, and are being maintained by, the European Commission's Joint Research Centre (JRC) in Ispra, Italy.

The applications make combined use of a variety of text analysis tools, including clustering, multi-label document classification, named entity recognition, name variant matching across languages and writing systems, topic detection and tracking, event scenario template filling, and more. Due to the high number of languages covered, linguistics-poor methods were used for the development of these text mining components. See the site http://langtech.jrc.it/ for technical details and a list of publications.

The speaker will give an overview of the various applications and will then explain the workings of selected text analysis components.

W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, p. 5, 2009.