DuOMAn

Dutch Online Media Analysis

The media landscape is changing, and it is doing so in two important ways. First, online news sources are playing an increasingly important role—today's offering of online news sources is increasingly web specific, that is, without being a derivative of traditional media counterpart. The second challenge comes from the rise of the participatory web: the traditional consumers of media are turning around to produce it themselves. Blogs, sharing sites, and internet fora all provide ordinary people the opportunity to express their opinions to a global audience. These streams of user-generated content provide an opportunity—and challenge—for media analysts: an enormous amount of new data to analyze and a glimpse into the minds of the masses.

Media analysis on a web scale—on both news sources and user-generated content—may be impossible without tools that can facilitate or even partially automate the process. The primary research goal of the DuOMAn project is to use technologies from the fields of Language Technology and Information Retrieval to develop a sentiment mining toolkit for media analysts who are interested in discovering and measuring the sentiment on the web—both in news sources and in user generated content.

Media analysis is undertaken on behalf of a client to determine what the media coverage of the client's domain of activity is and how the client is being portrayed. Media analysts collect output from sources relevant to the domain and look at the themes and the actors being covered, as well as the nature of the coverage. TrendLight analysts use a coding method based on the 'net method' of Kleinnijenhuis (VU) to build a network of support and criticism relations between actors, with regard to specific topics. These relations, together with the roles of the actors involved, are used within the reputation model of Van Riel (EUR) to determine the reputation of the client. At its core, media analysis is concerned with filling the template: "stakeholder X supports/is critical of Y on topic Z," where X and Y are actors (individuals or groups) and Z can be just about anything. This information is coded in a database using a dynamic coding scheme that depends on the domain being investigated and current media trends.

The main scientific contributions of the DuOMAn project may be summarized as follows:

- Lexical resources for Dutch sentiment analysis. Domain-independent Dutch sentiment lexicon, built on top of the Dutch WordNet, and algorithms for extending this lexicon with sentiment orientation for domain-specific terms using data-driven methods.
- Algorithms for information retrieval and information extraction tasks. Transforming the output of professional media analysts into training data for machine learning algorithms for sentiment-oriented document classification and information extraction tasks; classifying documents and snippets according to sentiment orientation; extracting actors and opinion relations from documents; aggregating mined opinion relations; detecting topical trends related to mined opinion relations.
- Test sets to evaluate classification and extraction algorithms.
- A data set consisting of Dutch language blogs crawled over a three-year period.
- New language technology tasks. (Extraction of source-target sentiment relations; aggregation of sentiment relations from multiple documents; detecting themes related to aggregated relations.)

Results

All targeted results were achieved. Lexical resources and data sets are, or will soon be, available via the Dutch HLT-agency, together with descriptions of implementations of algorithms. Presentations and publications were realized at leading international conferences in the area. A demo[1] is online and has been used to mine the relation between news and user-generated content. DuOMAn has given rise to a number of follow-up projects, both national and international.

The lexical resources created by the project have been and are being used in a series of projects. A small pilot project with a commerical partner examined the use of these resources for tracking down user experiences with consumer products. The lexical resources are being used in ongoing work within the Political Mashup project (led by Dr. M. Marx) to mine opinions in Dutch language political documents.

Implementations of the algorithms developed within the project are being used with Fietstas, an open source text analysis pipeline. Fietstas is currently being used to support both academic and applied projects. Partners benefitting from Fietstas include Talking Trends BV, the Netherlands Institute for Sound and Vision, Politie Gelderland-Zuid. Academic projects building on Fietstas include a CATCH project, CLARIN-NL projects, and the Infiniti project on information retrieval for information services within the large-scale public-private COMMIT program.

University of Amsterdam
Prof. dr. M. de Rijke (project leader), Dr. D. Ahn, B. Ernsting Msc, Dr. V. Jijkoun, F. Laan

University of Groningen
Prof. dr. G.J. van Noord

University College Ghent
Dr. V. Hoste

Gridline BV
Drs. T. Spaan

TrendLight BV
R. Franz

- Jijkoun, V., Hofmann, K. (2009). Generating a Non-English Subjectivity Lexicon: Relations That Matter, *Proceedings of EACL 2009*.
- Jijkoun, V., Rijke, M. de, Weerkamp, W., Ackermans, P., Geleijnse, P. (2010). Mining User Experiences from Online Forums: An Exploration, *Proceedings NAACL HLT 2010*.
- Jijkoun, V., Rijke, M. de, Weerkamp, W. (2010). Generating Focused Topic-specific Sentiment Lexicons, *Proceedings of ACL 2010 (Uppsala, Sweden)*.
- Jijkoun, V., Rijke, M. de (2011). Bootstrapping Subjectivity Detection, Proceedings of SIGIR 2011 (Beijing, China).
- Tsagkias, E., Rijke, M. de, Weerkamp, W. (2009). Predicting the Volume of Comments on Online News Stories, *Proceedings of CIKM 2009 (Hong Kong)*.

1 http://peilend.nl