Pushing the frontiers of Information Extraction:
A Strategy for Identifying Calls for Action in Texts about War and Conflict

In many disciplines in the social sciences and humanities, research is increasingly dealing with large scale text corpora that require the use of advanced NLP tools to process them and extract information. However, to date, most existing algorithms focus on the topical organization (e.g., bag-of-words based LDA) and relatively simple grammatical (e.g., subjects and objects of the discussion, adverbials of place, time, etc.) and semantic structures (e.g., sentiment) of the text, while more complex meanings remain difficult to extract. This talk introduces the EU-funded research project INFOCORE's strategy to algorithmically detecting calls for action, and discusses the involved challenges, available tools and further developments in the case of conflict discourse. At the same time, it sets out a more general approach to integrating existing algorithms to enable the detection of implicit, more complex semantic structures.

In the focus of INFOCORE's research is the role media plays in violent conflicts and whether it can be used to predict phases of peace and violence. As escalation or de-escalation depends on coordinated social action, media texts calling for specific courses of action play a critical role in understanding the dynamics of violent conflict. To capture what agendas are presented in various kinds of media, we aim to extract calls for action from news coverage – patterns of semantic information that expresses dissatisfaction with the current state of affairs and calls, hopes, orders, etc. to change the status quo.

This information can be expressed in texts explicitly, using lexical (modal verbs) or grammatical (imperative sentences) markers. However, many calls for action are also formulated implicitly, when a piece of text can be interpreted as calling for action only after taking cotext and context into account. Identifying explicit calls for action is therefore not sufficient to detect what agendas are being formulated, but must be augmented with strategies capable of detecting meanings that are merely implied.

Furthermore, to be able to predict the development of the conflict, it is important to know what exactly is being called for. In this study, we distinguish between various kinds of calls for cooperative solutions, such as ceasefire, peace talks, humanitarian help, financial support, etc.; restrictive solutions, such as fighting, aggression, sanctions or prosecution; calls for not doing something; calls to change the current state of affairs with no certain actions specified. Thus our classification routine can be broken down into two steps: extracting calls for action in the first, and classifying them in accordance with the developed taxonomy in the second.

Considering the complexity of the identification and classification task, our strategy was to first survey existing tools and solutions, and integrate suitable packages into our own tools. The integration was done within the framework of the open source, python based content analysis platform AmCAT, and draws upon open source machine learning libraries (e.g., WEKA, scikit-learn) and various NLP tools (e.g., Stanford CoreNLP). Throughout the development process, we recurrently validated the classification results for various algorithms with different parameters (e.g., Naive Bayes, k-nearest neighbours, SVM). The classifiers with default settings turned out to produce sizeable amounts of misclassification, which were effectively removed by adding tailored features that represent grammar and lexical peculiarities of calls for action to the model, resulting into overall accuracy of 82.2% (however, improving only modestly over the baseline of tf-idf weighted n-grams so far). The analytic algorithm thus provides a platform to which additional features can be added, integrating further information from other existing tools to increase the model's power for structuring and analyzing the considerable variance of natural discourse text.

The suggested approach, as well as our algorithm, can be easily transferred and adjusted for use in other domains, e.g., classifying emails and notes as tasks or requests, extracting recommendations and suggestions from user reviews, etc. In this vein, we discuss new venues for using open source software for research purposes and applying its results to solve concrete challenges in the industry.