Ephemeral Clustering of Microblogs for Future Multilingual Opinion Analysis
PhD Proposal
Janvier 2012 © HULTECH/GREYC - Tous droits réservés.

Stéphane Ferrari and Yann Mathet
HULTECH/GREYC, University of Caen Basse-Normandie (France)

Most research on opinion analysis has been dealing with the polarity and/or subjectivity of words [10] [23] [21] [6], sentences [11] [22] and texts [19] [18] [2]. More recently, the research field has been evolving to cross-domain classification of opinionated texts [1] [8] [15] [14] [9] [12]. Another research direction has also been proposed by [5], who introduces the writer bias in the opinion analysis. Some other researchers such as [24], propose semi-supervised techniques to learn subjectivity language based on translation equivalents.

In this research, we propose to study opinions in the Blogosphere along with their temporal dimension. Within this context, a very few works have been proposed such as [7] and [20]. However, in both cases, the temporal issue is about past events. In this case, the authors are interested in gathering the different evolutions of the opinions along a time line ending at the present time. But, we know that a great deal of rumours, and as a consequence opinions, are about future events as studied by [3], who propose different examples such as IDC predicts sales of mobile apps will be a $35 billion industry by 2014 or 2011 will be best year to buy a home, says BSA.

Predicting the future has always been the Holy Grail of Mankind. Although we cannot know the future, a lot can be inferred from it by mining huge collections of texts such as the Web (e.g. Web news stories) and Microblogs (e.g. Twitter, Facebook). Indeed, future events, prophecies or forecast analyses have traditionally been edited on written documents. As a consequence, retrieving texts with their future intent is likely to benefit a lot of Business Intelligence applications.

For example, in the specific context of politics, understanding the public opinion about future governmental decisions can be a crucial issue. Moreover, discovering future unplanned events can be of the utmost importance. A good example is the recent wave of revolutions in the Arabic countries, where the Blogosphere has played a crucial role in organising the strategy of the rebels. The same principle can be extended to the economical sector. For example, knowing that Dacia plans 8 new models by 2015 is likely to have a great impact on its competitors, as it is clearly a positive statement for Dacia. As a consequence, this research aims to provide press agencies or decision leaders with information about future planned or unplanned events and to what extent they provide positive or negative insights.
The first task of the research will aim to understand the expression of future in Web News articles and Microblogs. Within this context, we propose a semi-supervised learning strategy based on the extension of the work of [24] for the extraction of lexical elements characterizing the future tense in English, French and Arabic. Indeed, comparatively to most of the works on temporal information retrieval, which are either language-dependent [13] or only use year dates to characterize the future [3], we first propose to understand the correlation of lexical clues about the future across languages.

The second task will aim to understand the differences in topics between future statements for a given query by applying ephemeral clustering. The idea is to extend the work of [4], by proposing a soft clustering version of the HISGK-means algorithm. The intention of applying ephemeral clustering is to divide the data set of future statements in different topics. For example, for a query Tunisia, different statements may occur about different topics such as revolution, party, religion etc.

The third task is to understand the important information conveyed by future statements. For that purpose, we propose to first analyse the genre of the Microblogs or Web news stories to better predict their usefulness for decision makers (e.g. journalists, governmental leaders, etc.). Indeed, future statements are usually scheduled events, informative news or rumours, which can easily be classified as stated in [3]. As a consequence, understanding their class may provide useful information about their importance, i.e. if they convey strategical information.

Finally, the fourth task is to understand to what extent the future statements extracted from Microblogs and Web news stories can be opinionated texts. In order to propose a global architecture, which can deal with real-world texts (especially the Web), we aim to develop models, which learn the polar language (positive, negative or neutral) across domains. For that purpose, we will continue the work pursued by [16] based on Multi-view learning and in particular propose a new algorithm mixing the SAR framework [8] and the Class-Guided Multiview Learning Algorithm [17].

It is important to notice that the proposed work will be developed on a multilingual basis and will be tested especially for English, French and Arabic. As a consequence, the technologies and methodologies used for each one of the four steps will have to be thought to deal with mainly lexical information. Finally, the application will be tailored for a specific domain, which will depend on the company included in the project.

The thesis will certainly count with the collaboration of Adam Jatowt from the University of Kyoto (Japan).

References


