Extracting objective facts
from subjective and noisy sources

Kirill Tumanov & Panagiotis Chatzichristodoulou

June 2, 2014

Abstract

Some random abstract. (To be put in the end)

1 Introduction

Modern news agencies discuss the ongoing events with a great deal of subjectivity and dispute. This may be related to the general complexity of the causal relations in each situation nowadays, but also to the desire to put the events in a different light so that their story sells best. Whichever the reason is it is becoming more and more demanding for the reader to grasp what actually is going on. Especially when it comes to the analysis of the several information sources.

Today it is very popular to discover and evaluate polarity of the opinions on the web by emotional and sentiment mining. This analysis is useful specifically for uncovering of the (mostly author's) attitude towards the situation. However, it does not provide any insight in what exactly was criticized or admired. In other words, it does not provide the context for the overall result, it just presents the attitude.

When a more pragmatic and detailed analysis is needed, fact extraction comes into play. An underlying basis of the present work is the belief that the proper information presentation is free of the emotional flavor, and that each story is comprised of a set of interrelated facts and events. When it comes to understanding the situation, only facts matter.

An example of both subjective and noisy information sources would be having the articles with different values of the magnitude of an earthquake. In this case as soon as the earthquake occurs, the first articles have different magnitude of the earthquake but after some time, when the scientists agree on the number they all converge to it. An example of only the subjective data would be articles referring to casualties of war. These data would have an extra challenge being correctly mined as in such cases each side claims a different number of victims. Furthermore, those numbers do not change over time due to the fact that they are solely subjective and not noisy.

The purpose of this work is to extract the set of facts from a list of sources to be able track the situation regardless of the way it is presented. This means that objective facts from articles with subjective and noisy data are mined for the extraction.

2   Information Retrieval

2.1   Introduction to the Mined Topic

The hottest topic of the days of the work was a political crisis in Ukraine. Since a lot of the articles were written on the topic every day it was selected for investigation. It was observed that the Russian and the West policies were polar and that was another reason for mining those sources.

The eight following news agencies from Russia, the West, China and Middle East were used as sources of the info on the topic:

- Russia Today
- CNN
- Washington Post
- Reuters

- RIA Novosti
- ITAR-TASS
- Al-Jazeera
- Xinhua

2.2   Data Crawling and Parsing

It was decided to retrieve articles from the agencies' RSS feeds. For that purpose at first a custom built parser was implemented in Perl. The parser basic structure consists of:

1. The crawler (is installed on a crawl job to query the given list of feeds for all new articles)

2. The source parser (filters only the articles containing one of the keywords in the manually created dictionary (which in this work contained 35 words))

3. The HTML data parser (removes the HTML tags from the articles)

4. The meta-information handler (saves the information about the time article was published, its title and the publisher agency)

However, this manual approach was soon discarded mainly due to the poor quality of the HTML tag removal, which left a lot of garbage data from the web pages.

Since then, the parser built in KNIME was used for the same purpose. The major difference in the structure was in an addition of a scoring block which evaluated article "relativeness" to the topic. This block was built based on the same dictionary as discussed. The minimum acceptable threshold of matching

words was set to two. Other processing steps mainly were the same, but the article text retrieval quality was much better and disc space was almost unused since information was kept in the form of links as long as it was possible.

For the KNIME-built parser notable is the inability to extract the meta information about the articles. Therefore, the workaround for that problem had to be implemented. As a result of an information retrieval stage each article with its full corresponding meta-information was reconstructed in the form of a table entry. In total, the data table used consisted of 1892 entries.

3 Preprocessing

The conventional text preprocessing apart from the rest should include such steps as stop-word filtering and stemming. In this work it was observed that performing of any of those procedures dramatically decreases the system's performance, especially both in terms of the number of relations extracted and in their correspondence to the original ones in the text. Therefore, those steps were not performed to obtain the final results.

In the present work preprocessing stage mainly was limited to the proper named entity (NE) extraction from the text. Entities of types *Person*, *Organization* and *Location* were extracted from the articles. Later they were normalized to the same format and together they formed a set of NE dictionaries, which were essential at the mining stage.

4 Text Mining

Some stuff...

5 Results

Some stuff...

5.1 Visualization

Some stuff...

5.2 Discussion

Some stuff...

6 Conclusion

Some stuff...

References

[1] Aharon M, Elad M, Bruckstein A, Katz Y (2006) The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. IEEE Trans Signal Proces 54(11):43114322

[2] Buades A (2005) Image and film denoising by non-local means. PhD thesis, Uni. de les Illes Balears

[3] Buades A, Coll B, Morel JM (2005) A non-local algorithm for image denoising. Proc CVPR IEEE 2:6065