MediaGist: A cross-lingual analyser of aggregated news and commentaries

Josef Steinberger
University of West Bohemia
Faculty of Applied Sciences
Department of Computer Science and Engineering, NTIS Center
Univerzitni 8, 30614 Pilsen, Czech Republic
jstein@kiv.zcu.cz

Abstract

We introduce MediaGist, an online system for crosslingual analysis of aggregated news and commentaries based on summarization and sentiment analysis technologies. It is designed to assist journalists to detect and explore news topics, which are controversially reported or discussed in different countries. News articles from current week are clustered separately in currently 5 languages and the clusters are then linked across languages. Sentiment analysis provides a basis to compute controversy scores and summaries help to explore the differences. Recognized entities play an important role in most of the system's modules and provide another way to explore the data. We demonstrate the capabilities of MediaGist by listing highlights from the last week and present a rough evaluation of the system.

1 Introduction

News portals publish thousands of articles every day in various languages. Making sense out of such data without automated tools is impossible.

There are many news aggregators/analysers and each of them has its strengths. Google News aggregates headlines and displays the stories according to each reader's interests. IBM Watson News Explorer gives a more analytical way to read news through linked data visualizations. Europe Media Monitor (EMM) produces a summary of news stories clustered near realtime in various languages and compares how the same events have been reported in the media written in different languages.

However, there is another source of information at the news sites – commentaries – which contain very valuable public opinion about the news topics and has not been explored enough yet. Including commentaries opens many new use cases for journalists, agencies, which study public opinion, and partially also for readers. Controversial topics, such as the refugee crisis in Europe, or the Volkswagen's emission scandal, and their perception in different countries might be itself a source for reporting. Focusing on such topics should bring more traffic and rich discussions to the news portals. International agencies or political institutions will find useful the comparisons when studying particular public opinions. Crosslingually-organized news and commentaries will be useful for readers living in a multicultural environment, as they can quickly find and understand different views on the controversial topics.

MediaGist[1] builds on the ideas of news aggregators, but adds the comments' dimension. It continuously gathers metadata about news articles and their commentaries, currently in 5 languages. Articles from current week are clustered monolingually several times a day. It extracts entities, labels news and commentaries with sentiment scores and generates summaries on both sides. It also links the clusters across languages, similarly to EMM. Having aggregated news on one side and commentaries on the other side, it compares the information by sentiment analysis and summarization. A different sentiment of news and commentaries indicate a controversial topic and summaries help to identify the difference qualitatively. The crosslingual links allow to discover and explore topics, which are controversially reported or discussed in different countries.

The next section (2) relates MediaGist to the current news aggregation or analytics solutions. Section 3 describes MediGist from inside. The

[1]MediaGist is running at: http://mediagist.eu. A screencast video can be found at: https://www.youtube.com/watch?v=ONtKw_l6_X4.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations, pages 145–150,
Berlin, Germany, August 7-12, 2016.

overall architecture is followed by a description of the NLP pipeline. Section 4 gives an overview of the system's functionality and shows highlights from the last week, followed by a rough evaluation of the system, conclusions and future plans.

2 Related sites

Google News[2] is an automatic service that aggregates headlines from more than 50K news sources worldwide, groups similar stories together, and displays them according to each reader's interests. The content is selected and ranked using many factors, e.g. coverage, freshness, location, relevance and diversity of the story. There are more than 70 regional editions in many different languages.

IBM Watson News Explorer[3] gives a more analytical way to read news. It gathers 250k articles a day from 70k sources and converts the unstructured text into entities and concepts, and connects the dots through linked data visualizations.

EMM NewsBrief[4] is a summary of news stories (news clusters) from around the world, automatically classified according to thousands of criteria. It is updated every 10 minutes, and over 100k articles in 50+ languages run through it a day. It automatically detects the stories that are the most reported in each language at the moment. The Alert system presents the stories in many different classifications (Atkinson and van der Goot, 2009).

The second EMM's technology, NewsExplorer[5], allows to see the major news stories in various languages for any specific day and to compare how the same events have been reported in different languages (Steinberger et al., 2009). It shows the most mentioned names and other automatically derived information, eg. variant name spellings or a list of related entities (Pouliquen and Steinberger, 2009).

To summarize, the current systems gather masses of news articles and cluster them into stories. Some systems do it in many languages, and few link the stories across languages. Analytical solutions add information extraction (locations, entities, relations or categories). However, they do not integrate commentaries, which complement well the stories with public opinion. MediaGist adds the commentaries and uses them for various monolingual or crosslingual comparisons resulting in discovering and exploring controversies in the whole data.

3 System overview

MediaGist processing starts with a crawler (see figure 1). It gathers articles and their comments from predefined news sites[6]. It creates an RSS file for each article, which goes down the NLP pipeline. The pipeline first recognizes entities, in both the article and its comments, and assigns a crosslingual id to each mention. The next step is performed by the sentiment analyser, which assigns to each article and comment a tonality score[7]. The coreference resolver then enriches the list of entity mentions by name part references and definite descriptions. Each entity mention is then assigned a sentiment score and article comments are summarized[8]. These fully annotated article RSS files enter the clustering phase. Every four hours, for each language, the clusterer takes the articles published during the current week and creates monolingual clusters. Since this step, RSS files contain information about all articles in the cluster. The crosslingual linker then adds to each cluster links to the most similar cluster in other languages. The last step is creating a summary of clustered articles and a summary of cluster's comments (already summarized per article before). The RSS now contains all information needed by the presentation layer, the MediaGist website.

3.1 NER and coreference

The named entity recognizer is based on JRC-Names[9], which is a highly multilingual named entity resource for person and organisation names (Steinberger et al., 2011c). It consists of large lists of names and their many spelling variants (up to hundreds for a single person), including across scripts (Steinberger and Pouliquen, 2009).

Because the resource does not contain many morphological variants for Czech, it was extended

[2]https://news.google.com/

[3]http://news-explorer.mybluemix.net/

[4]EMM (Europe Media Monitor) is developed at Joint Research Centre, European Commission: http://emm.newsbrief.eu

[5]http://emm.newsexplorer.eu

[6]Currently, it gathers data from 7 sources in 5 languages: English (theguardian.com), Czech (idnes.cz, ihned.cz, novinky.cz), Italian (corriere.it), French (lemonde.fr) and German (spiegel.de).

[7]We call a document-level sentiment 'tonality'.

[8]There can be even thousands of comments attached to a single article. This summarization step largely reduces the size of the data sent further down the pipeline.

[9]https://ec.europa.eu/jrc/en/language-technologies/jrc-names

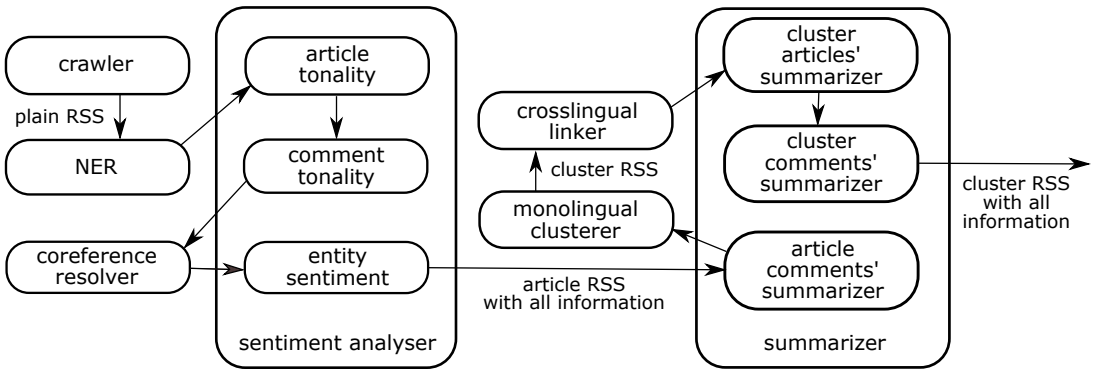

Figure 1: The architecture of MediaGist.

by an in-house rule-based morphological analyser.

Coreference resolution was inspired by (Steinberger et al., 2011a). In the cases of titles, it uses a list of person-title associations semi-automatically compiled over the past few years (Pouliquen and Steinberger, 2009).

3.2 Sentiment analysis

The sentiment analyser is used for 2 purposes. Assigning first a document-level tonality score $\langle -100; +100 \rangle$ to each article and comment, and second, a sentiment score $\langle -100; +100 \rangle$ to each entity mention. It uses highly multilingual and comparable sentiment dictionaries having similar sizes and based on a common specification, created by triangulation from English and Spanish (Steinberger et al., 2012). In the case of the tonality score, it counts subjective terms in an article, resp. a comment, and in the case of the entity score, it counts terms around entity mentions. It includes rules for using negation, intensifiers and diminishers (Steinberger et al., 2011b). Although machine learning approaches would produce better sentiment predictions, they require training data per language, and ideally per industry as well. And such data are currently expensive to create. With the rule-based approach, the system can easily process multiple languages.

3.3 Clustering and crosslingual linking

The monolingual clustering algorithm is based on agglomerative hierarchical clustering with the group average strategy (Hastie et al., 2009). The articles are represented by log-likelihood vectors of its terms and the similarity function is Cosine.

Crosslingual linking uses two kinds of features: entities and descriptors from EuroVoc[10]. EuroVoc is a multilingual, multidisciplinary thesaurus covering the activities of the EU, the European Parliament in particular. It contains terms organized in a hierarchy in 23 EU languages. Using Eurovoc features ensures that the linked clusters share the same topic. If at the same time the clusters share the same entities[11], it is very likely that the clusters are about the same story. A similar approach as in (Steinberger, 2013).

3.4 Summarization

The summarizer is used for three steps of the pipeline. First, it summarizes article comments, then articles in the cluster and finally comments of the cluster. We use an extractive approach based on latent semantic analysis, which uses both lexical and entity features (Kabadjov et al., 2013). This approach performed well in the Multiling evaluation campaigns[12].

4 Functionality

The systems has two main views to explore the media data: cluster view and entity view. We can select a language, a period (=week) and sort the data by different criteria[13]. Each view contains highlights of the selected period in the left panel.

4.1 The cluster view

It displays title and description, taken from the central article of the cluster (see figure 2). The left part shows information about articles and the right part about commentaries. On both sides, it displays generated summaries and aggregated tonal-

[10] http://eurovoc.europa.eu

[11] The entity ids are unified across languages.

[12] There were already 3 editions of MultiLing's multilingual multi-document summarization shared task: 2011 (Giannakopoulos et al., 2011), 2013, and 2015 (Giannakopoulos et al., 2015).

[13] The system currently holds data from the last 24 weeks.

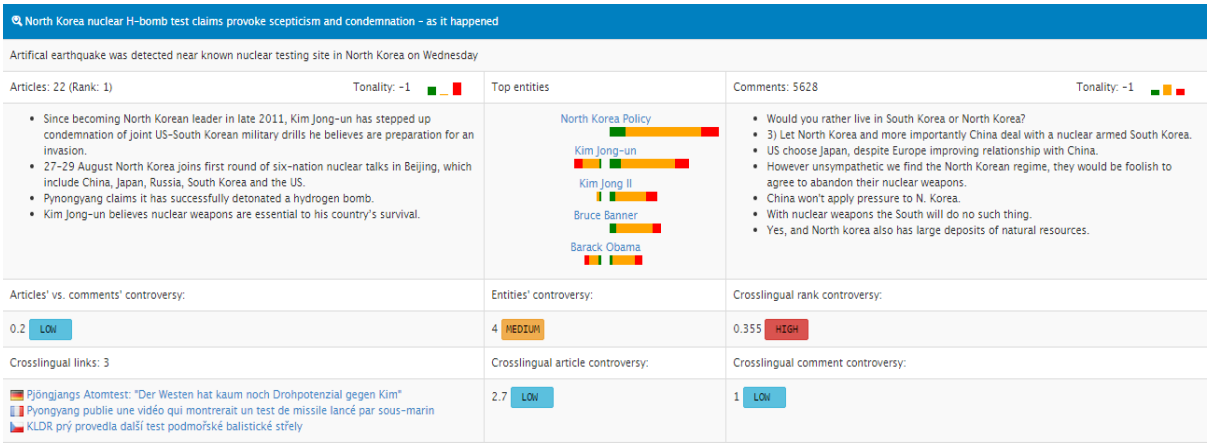

Figure 2: The top English cluster from the first week of 2016 (Jan 4th-10th). The screenshot does not include the page header, the left bar with highlights and the footer. More at http://mediagist.eu

ity figures. The central part shows entities and their sentiment in articles and comments.[14] At the bottom, you can see links to the related clusters in other languages.

MediaGist computes several controversy scores for each cluster. *Articles' vs. comments' controversy* compares tonality of articles and comments. The value correspond to the standard deviation of the two values. *Entities' controversy* compares sentiment of entity mentions in articles and comments. The value is a macro-average of standard deviations of each entity sentiment. *Crosslingual rank controversy* compares ranks of the cluster in different languages. Clusters are ranked for each language based on the number of articles. The value is a standard deviation of logarithms of the ranks. Logarithms give larger weights to the top ranks. This controversy is large if the topic is ranked at the top for some languages based on the coverage, while in other languages it is mentioned only marginally. Large *Crosslingual article controversy* indicates a large difference in articles' tonality among languages. The value is a standard deviation of average article tonalities across languages. This score says whether the topic is reported with the same tonality in different languages or not. And finally, a large *Crosslingual comment controversy* indicates topics which are discussed with different tonality across languages. The score compares average comment tonalities across languages by the standard deviation.

[14]Tonality/sentiment range is: $\langle -100; +100 \rangle$, green column = positive, orange = neutral, red = negative.

4.2 The entity view

The entity view displays variants of the entity found in the data (e.g. for David Bowie in week Jan 11-17, 2016: Bowie (3816 mentions), David Bowie (914), David (74), singer (60), star (46), musician (33), popstar (5), etc.). It shows the aggregated entity sentiment in articles and comments, which is compared by *Articles' vs. comments' controversy*. The sentiment is summarized by the most frequent subjective terms on both sides. Because we have also the entities linked across languages, we can compute their crosslingual controversy in articles and in comments. We can then easily find, which entities are reported or discussed with different sentiment across languages. As an example, Volkswagen is discussed negatively in Czech but positively in German (when all periods are selected).

4.3 Highlights from the last week

The most international topic during week (Mar 21-27, 2016) was *Fayal Cheffou charged over core role in Brussels bomb attacks* – covered well in all 5 languages. The English summary:

> At least 31 dead and more than 200 injured in bombings claimed by Islamic State. The attackers Brothers Khalid and Ibrahim el-Bakraoui have been identified as suicide bombers at the metro station and airport respectively. Before the Brussels attacks, Belgian prosecutors said DNA evidence had identified Moroccan-born Laachraoui as an accomplice of Paris attacker Salah Abdeslam. He was one of several men detained in police raids on Thursday. "What we feared has happened," said the Belgian prime minister, Charles Michel, at a press conference.

The following story was controversial in coverage: *Ukrainian pilot given 22-year jail sentence by Russian court* – one of the top clus-

ters in Czech but only few articles in English and French. The same topic was seen as controversially reported as well – the tonality of Czech articles was much more negative than English and French ones. A controversially discussed topic: *Sanders: 'We have a path towards victory' after win Washington caucuses* – while positive in English, negative in Czech. Reasons of the controversy can be found in the summaries.

Controversial entity in articles: *Donald Trump* – negative in English, close to neutral in Italian and French and positive in German and Czech. Difference between sentiment in articles and comments: *John Key* – positive in articles but negative in comments (English). Controversial entity in comments: *George W. Bush* – while the sentiment is balanced in English, it is negative in Czech and positive in German. The most frequent sentiment terms indicate the reasons: English: *good, helped, better, evil, violence*; German: *liebesmüh* (love effort), *deutlich besser* (clearly better), Czech: *zločiny* (crimes), *odsuzovat* (accusing), *špatný* (bad).

5 Evaluation

We present a rough evaluation of the key modules of the system. We discus results of NER, coreference, sentiment analysis and summarization obtained in the previous research. In the case of clustering, crosslingual linking and controversy predictions we validated the system output to get the first insight of their accuracy.

5.1 NER and coreference

The precision of the applied NER and coreference was measured in Steinberger et al. (2011a). From the current MediaGist's languages, person recognition performs best for French (98.4%) and worst for Italian (92.1%). The coreference module resolves name parts at precision of 98% and person title references at 70%. As the title references have not been continuously updated yet, several wrong references are caused by the missing temporal dimension.

5.2 Sentiment analysis

The accuracy of the sentiment analyser in all MediaGist's languages was measured in Steinberger et al. (2011b). For news sentences and entity targets, we got the best accuracy for English (74%) and the worst for Italian (66%). However, in the case of aggregating the polarities per entity and considering only entities with a larger difference between positive and negative mentions (extremely polar entities), 78% of entity classifications across all languages were correct.

5.3 Summarization

The LSA-based summarizer was evaluated during the last edition of the Multiling's multi-document summarization shared task (Giannakopoulos et al., 2015) as the top performing system overall (it received the lowest sum of ranks over all 10 languages). From the MediaGist's languages, it performed best in Czech, English and French. German and Italian was not included.

5.4 Clustering and crosslingual linking

In the case of clustering and crosslingual linking, we asked two annotators to validate the output of the system. The annotators were not fluent speakers in all 5 languages, but they had enough knowledge to judge the task. We selected the top 5 English clusters from the first 4 full weeks of 2016. The clusters were ranked based on the number of crosslingual links. The task of the clustering validation was to check whether the components of the cluster are relevant to the cluster's topic identified by the title of its central article. In the case of the crosslingual linking, the task was to check the similarity of the linked clusters based on their article titles. Clustering validation was found not to be that subjective, the inter-annotator kappa was .89. The validation of crosslingual links was more difficult, the annotators did not always agree (kappa was .63), mainly because of a different view on the right granularity of the topic (e.g. the clusters were both discussing the regugee crisis, but in one language it was about closing the borders and in the other about a disorder in Germany). From the total of 235 cluster components, 96% were judged as correct and from the 59 crosslingual links, 76% were pointing to the right cluster of the other language.

5.5 Controversy scores

We selected the most interesting controversy score, crosslingual comment controversy, to be judged by two annotators. For each crosslingual link evaluated in 5.4, we took the corresponding comment summaries (each in a different language) and showed them to an annotator. Her task was to

assess whether the view of the topic/entities is different (controversial) in the two languages or not. The task definition was rather shallow, but still there was a fair agreement (kappa was .48). We then produced a gold controversy scores: for instance if we had a topic linked across 5 languages, there were 10 combinations judged twice. The Boolean judgements were aggregated and normalized, resulting in a score between 0 and 1. These golden scores were then compared against the system's crosslingual comment controversy scores by Pearson correlation: .51. Although the correlation is not perfect, the measure can already be useful to indicate controversy.

6 Conclusion

MediaGist uses language technology to detect controversy in world news. Sentiment analysis helps to identify controversial topics and entities across languages, and via summarization it is possible to explore them in detail. The controversy scores are much dependent on the quality of sentiment analysis. Improving the sentiment module will directly lead to better predictions. Future plans include increasing the data volume on on both vertical (sources) and horizontal (historical data) axes. This will allow to study the evolution of a news thread or of a person name. The system currently consumes raw commentaries. Representing a precise opinion of real Internet users will require to fight trolls and filter the conversations (Mihaylov et al., 2015).

Acknowledgments

This work was supported by project MediaGist, EUs FP7 People Programme (Marie Curie Actions), no. 630786. MediaGist.

References

M. Atkinson and E. van der Goot. 2009. Near real time information mining in multilingual news. In *Proceedings of the 18th International World Wide Web Conference (WWW 2009)*, pages 1153–1154, Madrid, Spain.

G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, and V. Varma. 2011. TAC2011 MultiLing Pilot Overview. In *TAC 2011 Workshop*.

G. Giannakopoulos, J. Kubina, J. Conroy, J. Steinberger, B. Favre, M. Kabadjov, U. Kruschwitz, and M. Poesio. 2015. Multiling 2015: Multilingual summarization of single and multi-documents, online fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274. ACL.

T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning*. Springer-Verlag.

M. Kabadjov, J. Steinberger, and R. Steinberger. 2013. Multilingual statistical news summarization. In *Multilingual Information Extraction and Summarization*, volume 2013 of *Theory and Applications of Natural Language Processing*, pages 229–252. Springer.

T. Mihaylov, G. Georgiev, and P. Nakov. 2015. Finding opinion manipulation trolls in news community forums. In *Proceedings of the 19th CoNLL*, pages 310–314. ACL.

B. Pouliquen and R. Steinberger. 2009. Automatic construction of multilingual name dictionaries. In *Learning Machine Translation*. MIT Press.

R. Steinberger and B. Pouliquen. 2009. Cross-lingual named entity recognition. In *Named Entities - Recognition, Classification and Use*, volume 19 of *Benjamins Current Topics*, pages 137–164. John Benjamins Publishing Company.

R. Steinberger, B. Pouliquen, and C. Ignat. 2009. Using language-independent rules to achieve high multilinguality in text mining. In *Mining Massive Data Sets for Security*. IOS-Press, Amsterdam, Holland.

J. Steinberger, J. Belyaeva, J. Crawley, L. Della-Rocca, M. Ebrahim, M. Ehrmann, M. Kabadjov, R. Steinberger, and E. Van der Goot. 2011a. Highly multilingual coreference resolution exploiting a mature entity repository. In *Proceedings of the 8th RANLP Conference*, pages 254–260. Incoma Ltd.

J. Steinberger, P. Lenkova, M. Kabadjov, R. Steinberger, and E. van der Goot. 2011b. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *Proceedings of the 8th RANLP Conference*, pages 770–775.

R. Steinberger, B. Pouliquen, M. Kabadjov, J. Belyaeva, and E. van der Goot. 2011c. Jrc-names: A freely available, highly multilingual named entity resource. In *Proceedings of the International RANLP Conference*. Incoma Ltd.

J. Steinberger, M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, P. Lenkova, R. Steinberger, H. Tanev, S. Vzquez, and V. Zavarella. 2012. Creating sentiment dictionaries via triangulation. *Decision Support Systems*, 53(4):689 – 694.

R. Steinberger. 2013. Multilingual and cross-lingual news analysis in the europe media monitor (emm). In *Multidisciplinary Information Retrieval*, volume 8201 of *LNCS*, pages 1–4. Springer.