NewsGist: A Multilingual Statistical News Summarizer

Mijail Kabadjov, Martin Atkinson, Josef Steinberger,
Ralf Steinberger, and Erik van der Goot

Joint Research Centre, European Commission,
Via E. Fermi 2749, Ispra (VA), Italy
firstname.lastname@jrc.ec.europa.eu

Abstract. In this paper we present NewsGist, a multilingual, multi-document news summarization system underpinned by the Singular Value Decomposition (SVD) paradigm for document summarization and purpose-built for the Europe Media Monitor (EMM). The summarization method employed yielded state-of-the-art performance for English at the Update Summarization task of the last Text Analysis Conference (TAC) 2009 and integrated with EMM represents the first online summarization system able to produce summaries for so many languages. We discuss the context and motivation for developing the system and provide an overview of its architecture. The paper is intended to serve as accompaniment of a live demo of the system, which can be of interest to researchers and engineers working on multilingual open-source news analysis and mining.

1 Introduction

On a daily basis, the Europe Media Monitor (EMM)[1] gathers over 100k news articles in several dozens of languages from thousands of on-line news sources worldwide [1, 8]. It clusters all these articles into major news stories and plots in real time news clusters' sizes along a time line to provide, as opposed to standard search engines, a visual overview of the current state of affairs. It also automatically identifies spikes on the graph and sends out relevant breaking-news alerts, where 'relevance' is user-defined by using a set of intuitive semantic categories (called EMM categories), to the thousands of subscribed users.

Additionally, EMM recognizes references to entities (locations, persons and organizations) in the news [4], detects sentiment, monitors the development of news stories over time and links news clusters across languages [7].

Currently, however, EMM does not provide succinct summaries for the, potentially large, news clusters. This is clearly a desirable feature since these clusters may contain hundreds of news articles which would be impossible to read in full within a short time frame. Yet, this is often the need of EU decision makers who make use of the EMM system on a daily- or even hourly-basis and based on the information they receive they must produce timely responses to complex issues. Therefore, providing high quality

[1] EMM's news analysis applications are NewsBrief, NewsExplorer, MedISys and EMM labs accessible from: http://emm.newsbrief.eu

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part III, LNAI 6323, pp. 591–594, 2010.

summaries would substantially improve the usability of EMM as a news aggregation and trend visualization system.

In this paper we describe the summarization system currently under development for EMM, which we have named NewsGist. In the search for a suitable summarization method we adopted a general processing model for summarization foreseeing three phases [5]: interpretation, transformation and generation, and we chose the Singular Value Decomposition (SVD) paradigm to underpin it [2, 6]. The SVD approach has the advantage of being language-independent and has proven to be an effective summarization method yielding state-of-the-art performance in international evaluation efforts such as those of the Text Analysis Conference[2] (TAC) [6]. Furthermore, high-performance implementations of SVD taking advantage of heavily parallel architectures such as GPUs are already available.

The rest of the paper is organized as follows. In the next section we provide a brief overview of the Europe Media Monitor. Then, we describe the SVD model to summarization, section 3. After that, in section 4, we present NewsGist. Finally, we conclude the paper with pointers to further work.

2 Europe Media Monitor

The Europe Media Monitor is a web-based multilingual news aggregation system that collects over 100k news articles per day in about 50 languages from more than 2500 web news sources. The system employs text mining techniques to provide a picture of the present situation in the World (as conveyed in the media). Every ten minutes it automatically clusters all the collected news articles and displays the ten largest clusters per language by plotting them on a time-by-size graph. It also provides all the necessary hyperlinks to navigate through the clusters and to go to the source for a detailed exploration. In addition, it applies some deeper information analysis techniques, as for example, to automatically detect violent events, derive reported social networks and analyze media impact.

The public website provides a user interface to all this information. This public website is visited on a regular basis by some 30000 human users, and gets some 1.2M hits per day.[3]

3 Multi-document Summarization Based on SVD

As mentioned above, we chose the SVD paradigm to build our summarizer on. Next, we describe how each one of the three processing phases of interpretation, transformation and generation are realized.

In SVD-based summarization the interpretation phase takes the form of building a term-by-sentence matrix $A = [A_1, A_2, \ldots, A_n]$, where $A_j = [a_{1j}, a_{2j}, \ldots, a_{nj}]^T$ represents the weighted term-frequency vector of sentence j in a given set of documents.

[2] `http://www.nist.gov/tac/`

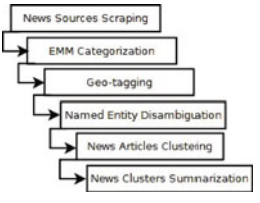
[3] For more details on EMM see [1, 8].

The transformation phase is done by applying singular value decomposition (SVD) to the initial term-by-sentence matrix and is defined as $A = U\Sigma V^T$.

Finally, the generation phase takes place in the form of prominent sentence selection.[4]

4 NewsGist

In this section we present NewsGist.[5] We provide a brief overview of its architecture and show screenshots for several languages.



(a) EMM's main processing phases.

(b) NewsGist's overlayed screenshots for English, French and German.

Fig. 1. EMM and NewsGist

As mentioned earlier, NewsGist was developed as part of EMM which is built on a pipeline architecture, where an input text document undergoes several processing phases during which the source is augmented with several layers of metadata such as named entities recognized in the text and semantic categories triggered by the text. The data interchange format between processing phases is RSS, a light-weight type of XML typically used by on-line news providers.

Thus, the input to NewGist is an RSS file enriched with information acquired by previous processing phases. Most importantly, by the time the RSS file reaches NewsGist, it already contains the outcome of the clustering of news articles and as output NewsGist produces a summary for each distinct news cluster (see fig. 1(a)).

The core system is pretty compact and is implemented as a Java servlet, running on top of Apache's Tomcat web server[6].

Language-specific tokenization and sentence splitting is provided by CORLEONE [3]. Reading and writing of RSS files is provided by EMM utility libraries. Matrix

[4] See [6] for full details of the method.

[5] Online demo of the system is available at `http://emm-labs.jrc.it/EMMLabs/NewsGist.html`

[6] `http://tomcat.apache.org/`

operations, and in particular singular value decomposition, is provided by the matrix-toolkits-java libraries[7] and also, alternatively, by the Java Matrix Package (JAMA)[8].

In figure 1(b) we show overlayed screenshots of NewsGist's online demo (`http://emm-labs.jrc.it/EMMLabs/NewsGist.html`) for three languages of the European Union: English, French and German.

5 Conclusion

In this paper we presented NewsGist, a multilingual multi-document summarization system purpose-built for the Europe Media Monitor (EMM). We provided an overview of EMM, briefly discussed the underlying summarization method based on the SVD paradigm and described the architecture of the system.

In future work we intend to carry out a comprehensive evaluation of the quality of the summaries for languages other than English.

Acknowledgments

We are grateful to the EMM development team for their support.

References

[1] Atkinson, M., Van der Goot, E.: Near real-time information mining in multilingual news. In: Proceedings of the 18th International World Wide Web Conference (WWW 2009), Madrid, Spain, pp. 1153–1154 (April 2009)

[2] Kabadjov, M., Steinberger, J., Pouliquen, B., Steinberger, R., Poesio, M.: Multilingual statistical news summarisation: Preliminary experiments with English. In: Proceedings of IAP-WNC at the IEEE/WIC/ACM WI-IAT (2009)

[3] Piskorski, J.: CORLEONE - core linguistic entity online extraction. Tech. Rep. EN 23393, Joint Research Centre of the European Commission (2008)

[4] Pouliquen, B., Steinberger, R.: Automatic construction of multilingual name dictionaries. In: Goutte, C., Cancedda, N., Dymetman, M., Foster, G. (eds.) Learning Machine Translation. NIPS series, MIT Press, Cambridge (2009)

[5] Spärck-Jones, K.: Automatic summarising: Factors and directions. In: Mani, I., Maybury, M. (eds.) Advances in Automatic Text Summarization. MIT Press, Cambridge (1999)

[6] Steinberger, J., Kabadjov, M., Pouliquen, B., Steinberger, R., Poesio, M.: WB-JRC-UT's participation in TAC 2009: Update Summarization and AESOP tasks. In: National Institute of Standards and Technology (eds.) Proceedings of TAC, Gaithersburg, MD (November 2009)

[7] Steinberger, R., Pouliquen, B., Ignat, C.: Using language-independent rules to achieve high multilinguality in text mining. In: Fogelman-Soulié, F., Perrotta, D., Piskorski, J., Steinberger, R. (eds.) Mining Massive Data Sets for Security. IOS-Press, Amsterdam (2009)

[8] Steinberger, R., Pouliquen, B., van der Goot, E.: An introduction to the europe media monitor family of applications. In: Gey, F., Kando, N., Karlgren, J. (eds.) Proceeding of the SIGIR Workshop on Information Access in a Multilingual World (SIGIR-CLIR 2009), Boston, USA (July 2009)

[7] `http://code.google.com/p/matrix-toolkits-java/`

[8] `http://math.nist.gov/javanumerics/jama/`