News Analysis for Knowledge Based Policy Making

Milena Slavcheva

Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
milena@lml.bas.bg

Abstract. The paper presents the construction of highly multilingual text classifiers for the automatic detection of text documents within a corporate body setting. The language resource is applied in the customised use of a media monitoring system that provides an in-house multilingual knowledge service for policy analysis and policy making. The service includes the production of specific multilingual newsletters and their dissemination. The text classifiers are related to an extendable taxonomy of topics in the domain of research and innovation. The service design and performance have proved to be very useful and highly effective in a large scale, multilingual, dynamic environment. The interface for human-machine interaction is friendly and ensures the possibility for the active type of users to take part in the creation of linguistic structures and the provision of web sources that enhance the automatic detection of relevant information. The additional tuning of the system is done in a time and labour saving mode. The second type of users, the service recipients, benefit from obtaining targeted information after its double, machine and human, evaluation and selection. Multilinguality is a special asset of the monitoring system and associated knowledge service. Information is detected and provided in the languages of the countries taken into consideration. In this way it is possible to gain an insight from the authentic background of the analysed processes and to rely on evidence-from-the-source.

Keywords: Media Monitoring, Text Classification, Information Retrieval, Multilingual Application, Human-Machine Interaction, Knowledge Service

1 Introduction

The paper presents the construction of highly multilingual text classifiers for the automatic detection of text documents within a corporate body setting. The language resource is applied in the customised use of the *NewsDesk* module of the *Europe Media Monitor* (EMM)[1] system developed at the *Joint Research Centre* (JRC)[2] of the European Commission (EC) [1]. The *NewsDesk* provides an in-house multilingual knowledge service to the policy analysts and policy makers. The service includes the

[1] https://emm.newsbrief.eu/overview.html

[2] https://ec.europa.eu/jrc/en

production of specific multilingual newsletters and their dissemination to interested parties within the European Union (EU) institutions. The text classifiers are related to a taxonomy of topics in the domain of research and innovation. The service design and performance have proved to be very useful and highly effective in a large scale, multilingual, dynamic environment.

The *Europe Media Monitor* (EMM) system as a whole incorporates several media gathering and analysis applications [1], [2], [3], whose scale is comparable, for instance, to the SUMMA (Scalable Understanding of Multilingual Media) platform [4] aiming to support the work of large news organisations. Both systems incorporate NLP techniques into a media processing pipeline, certainly, each one having its history of development, specific users and real-world function.

Text classification is an everlasting challenge in text understanding tasks [5] employing various techniques. The EMM system uses intensively finite state techniques to match the predefined text category definitions. These techniques have proved to be well performing in a highly multilingual system operating in a very dynamic environment with continuous client requests for adding new text classes to be searched in "new" languages, with the active involvement of non-IT experts in various domains of interest. Such an approach of tailored automatic text analysis that employs manual effort in building classifiers relates to that of [6], where an extension of the DUALIST tool [7], [8] is described enabling social scientists to engage directly in media monitoring.

The paper is structured as follows. Section 2 describes the process of monitoring and analysis of media for selecting relevant information according to predefined text classifiers. Section 3 presents the design and operation of the knowledge service. Section 4 draws some conclusions of the workflow scenario.

2 Media Monitoring and Analysis

As pointed out above, the *Europe Media Monitor* (EMM) system consists of several tools that gather and classify news in around 70 languages. Most of them are accessible in the public domain. The *NewsDesk* module, described in this paper, is used internally within the EU institutions and is customised to automatically select news and other relevant documents on specifically defined topics of interest that serve the needs of particular units in the different EC Directorates and EU institutions. This paper presents a use case of the *NewsDesk* in the domain of research and innovation activities and policies in the 28 member states of the European Union. Following a topic taxonomy, information is retrieved for each EU country in the official language of that country. On the basis of the retrieved news items, a team of policy analysts produces newsletters for each EU member state in the respective language. The newsletters are disseminated to a number of clients in the EC directorates.

The EMM engine gathers news articles using two basic components: 1) hand-selected web sources; 2) manually defined text classifiers (called category definitions). The customised application necessitates the tuning of the web sources and of the text classification system of the EMM.

2.1 Web Sources

The web sources monitored for the customer specific tasks described here belong to two main groups – newspaper and institution websites. They augment the sets of media websites for scraping, that is, a channel directory in XML format is filled in with newly selected web sources, for example:

```
<channel id="hamagbicro.hr-website">
     <dc:format>site</dc:format>
     <dc:type>webnews</dc:type>
     <dc:subject>Science</dc:subject>
     <dc:description>Business Innovation Agency of Croatia
(BICRO)</dc:description>
     <dc:identifier>http://www.hamagbicro.hr/</dc:identifier>
     <ocs:encoding>UTF-8</ocs:encoding>
     <iso:country>HR</iso:country>
     <region>European Union</region>
     <category>National</category>
     <ranking>0</ranking>
     <iso:language>hr</iso:language>
     <ocs:schedule>
       <ocs:updatePeriod>daily</ocs:updatePeriod>
       <ocs:updateFrequency>1</ocs:updateFrequency>
     </ocs:schedule>
     <feed title="hamagbicro hr website News"
url="http://www.hamagbicro.hr/kategorija/vijesti/"/>
  </channel>
```

The EMM uses RSS feeds when provided by the websites, otherwise it extracts information from the HTML formatted web pages. The search engine also looks for relevant information in other machine readable text files, including PDF documents. The websites for monitoring have to be carefully chosen in respect to regular update and relevance of the information. The quality of the web sources impacts the obtained results. Besides the technicalities involved in the search process, there are interesting behavioural aspects that influence the results for the different countries - the availability of websites of the different institutions and media, how much they talk in the press on topics related to research and innovation, to what extent the institutional information is publicly available, etc.

2.2 Category Definitions

Using a predefined set of text classifiers, the so called category definitions, the EMM system estimates the inflowing information and classifies it on the fly. The category definition has two sections (either of them is optional): a word-weight list with a defined threshold, which consists of set phrases used by the EMM finite-state machine pattern matcher. The keywords in the word-weight list belong mainly to two groups –

generic patterns like *research*, *research and innovation*, *innovation*, *open access*, and country specific expressions referring to institutions like *Bulgarian Academy of Sciences*, *Ministry of Science and Education*, *Hellenic Research Foundation*, *Baltic Innovation Fund*. The second section of the category definition consists of one or more combinations of keywords connected with logical operators. These combinations are used for broader terms or concepts with looser connection between the words that denote them.

Wildcards are made use of in the category definitions: the "+" symbol is used in multiword expressions to indicate that several words form a search string; the "%" symbol stands for 0, 1 or more characters and is used mainly to replace word inflections; the "_" (underscore) symbol matches exactly one character within the word. A category definition can contain a *proximity* attribute, which defines a word context size within which the keywords in the combination have to occur. Below is an example of a combination of lists of keywords written as a logical formula, which means that at least one element of each of the OR lists within the brackets must occur for a category to be triggered (the combination of keywords belongs to the category *tax incentives* as part of the topic i*nnovation union*):

```
(tax+incentive% OR subsidy OR subsidies OR SBRI OR R&D+tax+incentive% OR
innovation+incentive%) AND (healthy+aging OR   bioeconomy OR
bio+economy OR biomass% OR biofuel% OR bio+fuel% OR photovoltaic% OR
energy+effiecien% OR solar OR electric+car% OR electric+vehicle% OR e-
vehicle%)
```

It should be noted that the EMM system also uses a multilingual, classified geospatial information base of place names, provinces, regions and countries in order to perform geo-tagging [9]. In the country-specific information retrieval the geospatial information base is used in a filter in the categorisation of news items. The geospatial information base is a built-in module created and maintained by the “central” EMM developers’ team.

The country-specific classifiers have been created by the team of policy analysts in an iterative process of definition, verification and tuning. Although a unified approach has been implemented to the construction of the text classifiers for the different countries, they still have a varied composition depending on the conceptualisation of the different languages and the discourse typical for the respective countries. The keywords have to be carefully chosen in respect to the balance between generic concepts and specific names of entities in order to avoid the abundance of noise in the retrieved news items and not to miss relevant articles.

3 The Knowledge Service

The described customised use of the EMM system includes the production and dissemination of country specific newsletters through the *NewsDesk* application. The newsletters are in the official language of each country. The EMM engine detects automatically news items and other relevant documents, which are represented in the form of RSS feeds within the *Workspace* – a tool for creating newsletters. Each news item representation contains a title, a link to the web source, and a short three-four-line description. Information about the triggers of the particular news category is also displayed. The human user can navigate through the pages of the machine-selected articles and can make his/her own selection for the particular newsletter. By a drag-and-drop action the human-selected articles are inserted in the Editing Area of the window. It is possible to hand-edit the title and the other specifications of the chosen article if needed. The Editing Area offers the useful option of creating a taxonomy of the article topics (or sub-topics) related to the general topic of interest *Research and innovation*, for example:

```
Research Infrastructures, Labour market for researchers, Researchers'
mobility, Knowledge transfer, Public-private cooperation, Start-ups,
Open access, Intellectual property rights, Higher education, Funding,
Framework programmes, Innovation ecosystems, R&I evaluation, etc.
```

The elements of this taxonomy are represented as titles of sections or sub-sections in a user-defined newsletter content structure.

When compiling the newsletter, the human users have the possibility to add manually news articles and institution document references, which they happen to find “outside” the EMM operation, and which they consider useful.

Once the team of policy analysts have completed the selection of news items for the newsletters, the manager of the customised use of the EMM system checks the quality of the newsletters, generates the newsletter files and delivers them to the list of clients. All those actions are performed within the *NewsDesk* software environment.

The entire process whose output is the newsletters is in fact a multilingual knowledge service that provides evidence, outlines the social and political context of the Research and Innovation developments in a given country, backs up the argumentation for certain analyses and decisions. The use of the EMM system supports the production of analytical reports.

3.1 Impact of the Knowledge Service

As can be seen from the discussion so far, the *NewsDesk* application can have two types of users:

- users who actively interact with the software system building text classifiers, following the information stream offered by the machine, and producing newsletters for clients;
- users who benefit from receiving targeted information after its double, machine and human, evaluation and selection.

The interface for human-machine interaction is friendly and ensures the possibility for the first type of user to take part in the creation of linguistic structures and the provision of web sources that enhance the automatic detection of relevant information. The additional tuning of the system is done in a time and labour saving mode. The friendly interface is there also for the human monitoring of the EMM automatic output. Such a usage scenario ensures the active role of the service user where he/she becomes a service enhancer.

A survey has been carried out where users of the second type have been asked for feedback on the newsletters usefulness. The feedback can be summarised as very positive. The newsletters recipients found the service very useful and expressed their interest in additional topics and continuity of the service. The newsletters provide a highly targeted machine-aided selection of articles in the languages of the respective countries. The selected materials provide a vivid picture of the situation in the different countries, which is a useful additional knowledge for the analytical work.

The features below qualify the knowledge service.

Multilinguality. Information is detected and provided in the languages of the countries taken into consideration [10]. In this way it is possible to gain an insight from the authentic background of the analysed processes and to rely on evidence-from-the-source.

Support to the Analytical Work. The knowledge service ensures information supply that supports the production of analytical reports. The flow of filtered information provided in reasonably sized portions constantly updates the expert knowledge of the policy analysts and policy makers.

Flexibility and Expandability of the Search Topics. The EMM system serves institutions where it is necessary to react quickly to constantly emerging topics and discussions, as well as to related events. There is a straightforward possibility for adding language resources for new topics, “new” languages, “new” countries. The adjustment of the system can be performed by the system users themselves on demand by the community of the service beneficiaries.

Performance of the Automatic System. The performance of the automatic news detection is overall estimated as good. However, the results vary for the different countries. As pointed out above, the performance depends on “external” conditions related to the quality of the web sources, the availability and the quality of the information, the social environment in a given country. Those objective conditions interact in a complex way with the analytical work on creating the category definitions for the automatic detection of relevant news items.

Organisation and Maintenance. The customised use of the EMM system is a manifestation of large scale collaboration and team work, which requires skilful management.

4 Conclusion

The paper presented a media monitoring application for automatic multilingual news items detection, as well as multilingual newsletters production and dissemination within a large scale organization.

The workflow scenario is an example of effective human-machine interaction that benefits the analytical work and the policy making at European level. The multilinguality is of special importance for such an application. The creation of text classifiers in a very big number of languages that trigger good results of the automatic text analysis is a challenge, which can be overcome by multiple experiments. The performance of the automatic news detection is estimated as good and very good for the different countries but further fine tuning of the language resources is always needed. The results depend on the interplay of the skilfully created language resources used for the automatic operation of the system, and objective conditions like the quality of the websites and the availabity and quality of relevant information.

The use case of the *NewsDesk* module of the EMM system is an example of team work in international environment, where specialists in different fields collaborate effectively.

5 Acknowledgements

The work presented in this paper was done during my six-year mandate as a policy analyst and IT researcher in Directorate General *Joint Research Centre* (JRC) of the European Commission in Brussels, in the *Knowledge for Finance, Innovation and Growth* Unit. EMM and its module *NewsDesk* have been very effectively used within the *Research and Innovation Observatory (RIO) – Horizon 2020 Policy Support Facility*[3] project of the *Knowledge for Finance, Innovation and Growth* Unit. The described service has been the result of the collaboration between the EMM developers' team in JRC-Ispra and the *Knowledge for Finance, Innovation and Growth* Unit in JRC-Brussels. I had the honour and responsibility to be the manager and the administrator of the customised use of the EMM system in JRC-Brussels. I would like to acknowledge the multiannual effort of the EMM team in JRC-Ispra for creating and maintaining the Europe Media Monitor system and to thank them for their cooperation and friendliness. I would also like to thank the team in JRC-Brussels for their cooperation and accuracy.

References

1. Steinberger, R., Podavini, A., Balahur, A., Jacquet, G., Tanev, H., Linge, J., Atkinson, M., Chinosi, M., Zavarella, V., Steiner, Y., van der Goot, E.: Observing trends in automated

[3] https://rio.jrc.ec.europa.eu/

multilingual media analysis. In: Proceedings of the Symposium on New Frontiers of Automated Content Analysis in the Social Sciences (ACA 2015), Zürich, Switzerland (2015).

2. Steinberger, R., Pouliquen, B., van der Goot, E.: An introduction to the Europe Media Monitor family of applications. In: Proceedings of the SIGIR 2009 Workshop on Information Access in a Multilingual World, pp. 1-8, Boston, Massachusetts, USA (2009).
3. Steinberger, R.: A survey of methods to ease the development of highly multilingual text mining applications. Language Resources and Evaluation 46, 155-176 (2012).
4. Germann, U., Liepins, R., Barzdins, G., Gosko, D., Miranda, S., Nogueira, D.: The SUMMA platform: a scalable infrastructure for multi-lingual multi-media monitoring. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics – System Demonstrations, pp. 99-104. Association for Computational Linguistics, Melbourne, Australia (2018).
5. Sachan, D. S., Zaheer, M., Salakhutdinov, R.: Investigating the working of text classifiers. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2120-2131. Santa Fe, New Mexico, USA (2018).
6. Wibberley, S., Reffin, J., Weir, D.: Language technology for agile social media science. In: Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 36-42. Association for Computational Linguistics, Sofia, Bulgaria (2013).
7. Settles, B.: Closing the loop: fast, interactive semi-supervised annotation with queries on features and instances. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1467-1478. Association for Computational Linguistics, Edinburgh, Scotland, UK (2011).
8. Settles, B., Zhu, X.: Behavioral factors in interactive training of text classifiers. In: Fosler-Lussier, E., Riloff, E., Bangalore, S. (eds.) Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 563-567. Association for Computational Linguistics, Montreal, Canada (2012).
9. Pouliquen B., Kimler M., Steinberger R., Ignat C., Oellinger, T., Blackler K., Fuart F., Zaghouani W., Widiger A., Forslund A-C., Best C. Geocoding multilingual texts: Recognition, Disambiguation and Visualisation. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 53-58. Genoa, Italy (2006).
10. Steinberger, R., Ehrmann, M., Pajzs, J., Ebrahim, M., Steinberger, J., Turchi, M.: Multilingual media monitoring and text analysis – challenges for highly inflected languages. In: Habernal, I., Matousek, V. (eds.), Text, Speech and Dialogue (TSD 2013), LNAI, vol. 8082, pp. 22--33. Springer, Heidelberg (2013).