MultiLingMine 2016: Modeling, Learning and Mining for Cross/Multilinguality

Salvatore Romeo[1], Andrea Tagarelli[2], Dino Ienco[3], Mathieu Roche[4], and Paolo Rosso[5]

[1] Qatar Computing Research Institute, Doha, Qatar
[2] DIMES, University of Calabria, Rende, Italy
[3] IRSTEA, LIRMM, Montpellier, France
[4] CIRAD, LIRMM, Montpellier, France
[5] Universitat Politecnica de Valencia, Valencia, Spain

Abstract. The increasing availability of text information coded in many different languages poses new challenges to modern information retrieval and mining systems in order to discover and exchange knowledge at a larger world-wide scale. The 1st International Workshop on Modeling, Learning and Mining for Cross/Multilinguality (dubbed MultiLingMine 2016) provides a venue to discuss research advances in cross-/multilingual related topics, focusing on new multidisciplinary research questions that have not been deeply investigated so far (e.g., in CLEF and related events relevant to CLIR). This includes theoretical and experimental on-going works about novel representation models, learning algorithms, and knowledge-based methodologies for emerging trends and applications, such as, e.g., cross-view cross-/multilingual information retrieval and document mining, (knowledge-based) translation-independent cross-/multilingual corpora, applications in social network contexts, and more.

1 Motivations

In the last few years the phenomenon of multilingual information overload has received significant attention due to the huge availability of information coded in many different languages. We have in fact witnessed a growing popularity of tools that are designed for collaboratively editing through contributors across the world, which has led to an increased demand for methods capable of effectively and efficiently searching, retrieving, managing and mining different language-written document collections. The multilingual information overload phenomenon introduces new challenges to modern information retrieval systems. By better searching, indexing, and organizing such rich and heterogeneous information, we can discover and exchange knowledge at a larger world-wide scale. However, since research on multilingual information is relatively young, important issues still remain uncovered:

– how to define a translation-independent representation of the documents across many languages;

- whether existing solutions for comparable corpora can be enhanced to generalize to multiple languages without depending on bilingual dictionaries or incurring bias in merging language-specific results;
- how to profitably exploit knowledge bases to enable translation-independent preserving and unveiling of content semantics;
- how to define proper indexing and multidimensional data structures to better capture the multi-topic and/or multi-aspect nature of multi-lingual documents;
- how to detect duplicate or redundant information among different languages or, conversely, novelty in the produced information;
- how to enrich and update multi-lingual knowledge bases from documents;
- how to exploit multi-lingual knowledge bases for question answering;
- how to efficiently extend topic modeling to deal with multi/cross-lingual documents in many languages;
- how to evaluate and visualize retrieval and mining results.

2 Objectives, topics, and outcomes

The aim of the *1st International Workshop on Modeling, Learning and Mining for Cross/Multilinguality* (dubbed *MultiLingMine 2016*),[6] held in conjunction with the 2016 ECIR Conference, is to establish a venue to discuss research advances in cross-/multilingual related topics. MultiLingMine 2016 has been structured as a *full-day* workshop. Its program schedule includes invited talks as well as a panel discussion among the participants. It is mainly geared towards students, researchers and practitioners actively working on topics related to information retrieval, classification, clustering, indexing and modeling of multilingual corpora collections. A major objective of this workshop is to focus on research questions that have not been deeply investigated so far. Special interest is devoted to contributions that aim to consider the following aspects:

- Modeling: methods to develop suitable representations for multilingual corpora, possibly embedding information from different views/aspects, such as, e.g., tensor models and decompositions, word-to-vector models, statistical topic models, representational learning, etc.
- Learning: any unsupervised, supervised, and semi-supervised approach in cross/multilingual contexts.
- The use of knowledge bases to support the modeling, learning, or both stages of multilingual corpora analysis.
- Emerging trends and applications, such as, e.g., cross-view cross-/multilingual IR, multilingual text mining in social networks, etc.

Main research topics of interest in MultiLingMine 2016 include the following:

- Multilingual/cross-lingual information access, web search, and ranking

[6] http://events.dimes.unical.it/multilingmine/

- Multilingual/cross-lingual relevance feedback
- Multilingual/cross-lingual text summarization
- Multilingual/cross-lingual question answering
- Multilingual/cross-lingual information extraction
- Multilingual/cross-lingual document indexing
- Multilingual/cross-lingual topic modeling
- Multi-view/Multimodal representation models for multilingual corpora and cross-lingual applications
- Cross-view multi/cross-lingual information retrieval and document mining
- Multilingual/cross-lingual classification and clustering
- Knowledge-based approaches to model and mine multilingual corpora
- Social network analysis and mining for multilinguality/cross-linguality
- Plagiarism detection for multilinguality/cross-linguality
- Sentiment analysis for multilinguality/cross-linguality
- Deep learning for multilinguality/cross-linguality
- Novel validity criteria for cross-/multilingual retrieval and learning tasks
- Novel paradigms for visualization of patterns mined in multilingual corpora
- Emerging applications for multilingual/cross-lingual domains

The ultimate goal of the MultiLingMine workshop is to increase the visibility of the above research themes, and also to bridge closely related research fields such as information access, searching and ranking, information extraction, feature engineering, text mining and machine learning.

3 Advisory board

The scientific significance of the workshop is assured by a Program Committee which includes 20 research scholars, coming from different countries and widely recognized as experts in cross/multi-lingual information retrieval:
Ahmet Aker, Univ. Sheffield, United Kingdom
Rafael Banchs, I2R Singapore
Martin Braschler, Zurich Univ. of Applied Sciences, Switzerland
Philipp Cimiano, Bielefeld University, Germany
Paul Clough, Univ. Sheffield, United Kingdom
Andrea Esuli, ISTI-CNR, Italy
Wei Gao, QCRI, Qatar
Cyril Goutte, National Research Council, Canada
Parth Gupta, Universitat Politcnica de Valncia, Spain
Dunja Mladenic, Jozef Stefan International Postgraduate school, Slovenia
Alejandro Moreo, ISTI-CNR, Italy
Alessandro Moschitti, Univ. Trento, Italy; QCRI, Qatar
Matteo Negri, FBK - Fondazione Bruno Kessler, Italy
Simone Paolo Ponzetto, Univ. Mannheim, Germany
Achim Rettinger, Institute AIFB, Germany
Philipp Sorg, Institute AIFB, Germany
Ralf Steinberger, JRC in Ispra, Italy
Marco Turchi, FBK - Fondazione Bruno Kessler, Italy

Vasudeva Varma, IIIT Hyderabad, India
Ivan Vulic, KU Leuven, Belgium

4 Related events

A COLING'08 workshop [1] was one of the earliest events that emphasized the importance of analyzing multilingual document collections for information extraction and summarization purposes. The topic also attracted attention from the semantic web community: in 2014, [2] solicited works to discuss principles on how to publish, link and access mono and multilingual knowledge data collections; in 2015, another workshop [3] took place on similar topics in order to allow researchers continue to address multilingual knowledge management problems. A tutorial on Multilingual Topic Models was presented at WSDM 2014 [4] focusing on how statistically model document collections written in different languages. In 2015, a WWW workshop aimed at advancing the state-of-the-art in Multilingual Web Access [5]: the contributing papers covered different aspects of multilingual information analysis, leveraging attention on the lack of current information retrieval techniques and the necessity of new techniques especially tailored to manage, search, analysis and mine multilingual textual information.

The main event related to our workshop is the CLEF initiative [6] which has long provided a premier forum for the development of new information access and evaluation strategies in multilingual contexts. However, differently from MultiLingMine, it does not have emphasized research contributions on tasks such as searching, indexing, mining and modeling of multilingual corpora.

Our intention is to continue the lead of previous events about multilingual related topics, however from a broader perspective which is relevant to various information retrieval and document mining fields. We aim at soliciting contributions from scholars and practitioners in information retrieval that are interested in Multi/Cross-lingual document management, search, mining, and evaluation tasks. Moreover, differently from previous workshops, we would emphasize some specific trends, such as cross-view cross/multilingual IR, as well as the growing tightly interaction between knowledge-based and statistical/algorithmic approaches in order to deal with multilingual information overload.

References

1. Bandyopadhyay, S., Poibeau, T., Saggion, H., Yangarber, R. (2008). Procs. of the Workshop on Multi-source Multilingual Information Extraction and Summarization (MMIES). ACL.
2. Chiarcos, C., McCrae J. P., Montiel E., Simov, K., Branco, A., Calzolari, N., Osenova, P., Slavcheva, M., Vertan, C. (2014). Procs. of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and NLP (LDL).
3. McCrae, J. P., Vulcu G. (2015). CEUR Procs. of the 4th Workshop on the Multilingual Semantic Web (MSW4), Vol. 1532.

4. Moens, M.-F., Vulié, I. (2014). Multilingual Probabilistic Topic Modeling and Its Applications in Web Mining and Search. In Procs. of the 7th ACM WSDM Conf.
5. Steichen, B., Ferro, N., Lewis, D., Chi, E. E. (2015). Procs. of the Int. Workshop on Multilingual Web Access (MWA).
6. The CLEF Initiative. http://www.clef-initiative.eu/.