ECIR 2016 Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine '16)

Dino Ienco
IRSTEA, LIRMM, Montpellier, France
dino.ienco@irstea.fr

Mathieu Roche
CIRAD, LIRMM, Montpellier, France
mathieu.roche@cirad.fr

Salvatore Romeo
Qatar Computing Research Institute, Doha, Qatar
sromeo@qf.org.qa

Paolo Rosso
PRHLT, Universitat Politècnica de València, València, Spain
prosso@dsic.upv.es

Andrea Tagarelli
DIMES, University of Calabria, Rende, Italy
andrea.tagarelli@unical.it

Abstract

The First International Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine) was held in conjunction with the 2016 European Conference on Information Retrieval (ECIR), in Padua, Italy. This report presents an overview of the motivations and objectives underlying the establishment of this workshop. It also provides a summary of the contributing papers and of the main research topics and trends discussed among the participants.

1 Introduction

The increasing availability of text information coded in many different languages has led to a new phenomenon of *multilingual information overload*. Thanks also to the growing popularity of tools that are designed for collaboratively editing through contributors across the world,

there is an increasing demand for methods capable of effectively and efficiently searching, retrieving, managing and mining different language-written document collections. This has posed new challenges to modern information retrieval and mining systems in order to discover and exchange knowledge at a larger world-wide scale. These challenges correspond to a number of novel research questions that include (but are not limited to) the following: How can we define a translation-independent representation of the documents across many languages? Can existing solutions for comparable corpora be enhanced to deal with multiple languages without depending on bilingual dictionaries or incurring bias in merging language-specific results? How can we profitably exploit knowledge bases to enable translation-independent preserving and unveiling of content semantics? How can we define proper indexing structures and multidimensional data structures to better capture the multi-topic and/or multi-aspect nature of the documents in a multilingual context? How can we detect duplicate/redundant information among different languages or, conversely, novelty in the produced information? How can we enrich and update multi-lingual knowledge bases from documents? How can we exploit multi-lingual knowledge bases for question answering? How can we extend stochastic models (e.g., topic models, expert finding, language models) to deal with cross/multilingual documents? How can we evaluate and visualize multilingual and multimodal retrieval and mining results?

The above constitute key motivations for the First International Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine), which was held in conjunction with the 38th European Conference on Information Retrieval (ECIR) in Padua, Italy, on March 20, 2016. The original mission of this workshop was to establish a venue to discuss research advances in cross/multilingual retrieval and mining, with a particular emphasis on theoretical and experimental on-going works about novel representation models, learning algorithms, and knowledge-based methodologies for emerging trends and applications.

2 Keynote

The workshop hosted a keynote, given by prof. Nicola Ferro (University of Padua, Italy), focusing on multilingual information access and relating challenges in the evaluation of multilingual corpora. In his talk, Ferro discussed the growing interrelation between multilinguality and multimediality in today web-based systems, and how both represent key concerns for multilingual information access. He in fact pointed out that multilingual information access is a multimedia concern, and that text is the primary enabler for multilingual information access. Multilingual and multimedia information systems are increasingly complex: they need to satisfy diverse user needs and support challenging tasks. Their development calls for proper evaluation methodologies to ensure that they meet the expected user requirements and provide the desired effectiveness. Large-scale world-wide experimental evaluations provide fundamental contributions to the advancement of state-of-the-art techniques through common evaluation procedures, regular and systematic evaluation cycles, comparison and benchmarking of the adopted approaches, and spreading of knowledge [15].

Ferro introduced the base challenges and approaches to multilingual information access, including possible cross-lingual and multilingual IR frameworks (based on translation at query, document, or both levels). He then discussed what performance trends emerge from several years of evaluation at CLEF, the European forum for multilingual and multimodal information access evaluation. He raised the importance of standard practices for the sake

Figure 1: Wordcloud of titles and abstracts of the papers accepted at MultiLingMine 2016.

of comparability of the systems under evaluation, then he showed interesting statistics about how monolingual and bilingual systems behaved over the last few years.

3 Highlights from the accepted paper presentations

The workshop received ten submissions, of which six papers were selected through a pee-review process which involved a Program Committee comprised of twenty research scholars, coming from different countries and widely recognized as experts in cross/multi-lingual information retrieval.

Figure 1 shows a wordcloud of the most representative (stemmed) terms occurring in *title* and *abstract* of the accepted papers. We observe how the contributing works mainly covered the following research themes:

- translation-independent representation models,
- cross-lingual analysis via translation models,
- ontology-driven approaches (with application to disease symptom identification),
- improvement in the analysis of poor-resource languages (specifically, Catalan and Latin),
- cross-script question answering,
- discourse segmentation,
- multilingual expertise finding,
- evaluation of cross-lingual analysis.

The accepted papers can broadly be organized according to two main trends: approaches that leverage translation models [4, 3] and approaches that introduce translation-independent

representations [5, 1, 10, 2] in order to deal with language heterogeneity. Some of the contributions also underline how multilingual analysis can support the investigation of resource-scarce languages [3]. It is also clear that classic information retrieval tasks, such as question answering and discourse segmentation, need to be revised for multilingual document collections [1, 3]. In the following we provide a summary of each of the contributing papers.

In [5], Ferrando et al. propose an ontology-rich approach to the identification and translation of disease symptoms in running text. At the heart of the approach lies a methodology which combines a knowledge-rich method for multilingual text classification (MOoD-TC) with a very large multilingual knowledge base (BabelNet [12]). The methodology proposed by Ferrando et al. features no requirement of training examples to build the classification model.

Llorens-Salvador and Delany present a feature extraction method for the task of cross-language author identification [10]. This method expands traditional bag-of-word (BOW) features with fragments and BOW of fixed size that are sampled from a portion of the document. The new representation space, at the end, is built on language-agnostic features that demonstrate to improve the final classification task. The various reviewers agreed on the fact that the use of language-agnostic features constitutes an interesting and novel direction in the context of cross-language classification.

Esfahani et al. address the problem of expert retrieval in the multilingual context [4]. A major contribution of this work lies in a cross-feeding of information between the multilingual information retrieval and the expert retrieval task. The multilingual documents of an expert are translated into the query language and then a ranking is produced based on the similarity between query and documents. Translation-based techniques are used to have a common representation space for the collection of documents. The use of translation tools allow to address domain specific disambiguation exploiting the fact that experts usually concentrate their activities on a particular topic.

The problem of resource-scarce language analysis is addressed in [3]. In this work, Cunha et al. propose a new technique to discourse segmentation for texts in Spanish to Catalan. An automatic discourse parser is developed based on rhetorical structure theory and translation rules between Spanish and Catalan. Evaluation on a novel Catalan dataset shows significant improvements over two baselines.

The work presented by Brodić et al. [2] introduces an approach for discriminating texts written in Latin from texts written in Italian, by combining feature engineering and clustering methods. An input text is transformed into a uniformly coded text, then eleven features are extracted for each text and a clustering method is applied to obtain a classification of the documents.

Banerjee et al. focus on question answering in a multilingual context [1]. The reference scenario the authors envisage is the development of question answering system capable to process questions posed in a mixed language and to retrieve answers from documents still written in the same mixed language setting. The proposed approach refers to an application in which Bengali speakers provide queries mixing their native language with English. Main contributions of this work include an encoding scheme, a public dataset, and an evaluation measure.

4 Panel discussion

The workshop finally hosted a 40 minute session which was devoted to a discussion among participants (above 30) concerning hot trends, open questions and challenges in cross/multilingual information retrieval.

Most participants first recognized the need for investing more in research on the development of suitable representations for multilingual corpora. On the one hand, there is a need for multidimensional data structures to embed and model information from different views/aspects/modes that are often present in (relatively long) documents, such as tensor models and decompositions [14, 9]: extending such existing methods to deal with multilingual documents is clearly made further difficult by the language heterogeneity. To cope with this issue, in literature, some works propose to use knowledge-based strategies in order to constitute a common representation space [14, 13, 7, 6]. On the other hand, the audience agreed that great opportunities also come from deep learning approaches combined with word embedding representation models: in this respect, some interesting suggestions were given that consider the application of the above approaches to handle (groups of) language semantic spaces [8].

Also related to the modeling aspect in cross/multilinguality is how to cope with multilingual corpora, and cross-lingual applications, in which there is an interleaving of usage of two or more languages within a reference language (e.g., usage of English terms in Latin language). Yet, the audience acknowledged the importance of using different models and learning methods depending on the length of the multilingual documents, which opens novel directions to the analysis of relatively long versus short and/or noisy text data. This makes sense not only in various web-based environments (e.g., encyclopaedic corpora versus microblogs) but also in more general, Internet-based contexts, such as mobile device environments. Another important point that stands up from the discussion is how to adapt (or extend) learning tools like topic models to a multilingual or cross-lingual scenario [11, 16, 17]. The point here is how to design more flexible statistical generative models without the necessity to have topic models customized for a specific number of languages.

Last but not least there was a discussion focused on the evaluation aspect. Particularly, the workshop attendees raised the need for building benchmarks of multilingual corpora to enable a controlled, repeatability/reproducibility-aware evaluation of various tasks of interest, including multilingual text summarization, question answering, document indexing, plagiarism detection and sentiment analysis.

Acknowledgements

The success of MultiLingMine 2016 would not have been possible without the valuable support of the members of the Program Committee—a complete list is available on the workshop website.[1]

This work is partially funded by research project PON 2007-2013 “BA2Know - Business Analytics to Know”, funded by Italian Ministry of Instruction, University, and Research, and by the research project TIN2015-71147-C2-1-P of the Spanish Ministry of Economy and Competitiveness.

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

References

[1] S. Banerjee, S. Kumar Naskar, P. Rosso, and S. Bandyopadhyay. The first cross-script code-mixed question answering corpus. In *ECIR 2016 Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine), Padua, Italy, March 20, 2016*, pages 56–65, 2016.

[2] D. Brodić, A. Amelio, and Z. N. Milivojević. A new image analysis framework for latin and italian language discrimination. In *ECIR 2016 Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine), Padua, Italy, March 20, 2016*, pages 46–55, 2016.

[3] I. Cunha, E. SanJuan, J.M. Torres-Moreno, I. Castellon, and M. Lloberes. Extending automatic discourse segmentation for texts in spanish to catalan. In *ECIR 2016 Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine), Padua, Italy, March 20, 2016*, pages 36–45, 2016.

[4] H. N. Esfahani, J. Dadashkarimi, and A. Shakery. Profile-based translation in multilingual expertise retrieval. In *ECIR 2016 Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine), Padua, Italy, March 20, 2016*, pages 26–35, 2016.

[5] A. Ferrando, S. Beux, V. Mascardi, and P. Rosso. Identification of disease symptoms in multilingual sentences: an ontology-driven approach. In *ECIR 2016 Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine), Padua, Italy, March 20, 2016*, pages 6–15, 2016.

[6] M. Franco-Salvador, F. L. Cruz, J. A. Troyano, and P. Rosso. Cross-domain polarity classification using a knowledge-enhanced meta-classifier. *Knowl.-Based Syst.*, 86:46–56, 2015.

[7] M. Franco-Salvador, P. Rosso, and R. Navigli. A knowledge-based representation for cross-language document retrieval and categorization. In *Proc. of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), April 26-30, 2014, Gothenburg, Sweden*, pages 414–423, 2014.

[8] J. Kim, J. Nam, and I. Gurevych. Learning semantics with deep belief network for cross-language information retrieval. In *Proc. of the 24th International Conference on Computational Linguistics (COLING), December 8-15, 2012, Mumbai, India*, pages 579–588, 2012.

[9] Y.-M. Kim, M.-R. Amini, C. Goutte, and P. Gallinari. Multi-view clustering of multilingual documents. In *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), July 19-23, 2010, Geneva, Switzerland*, pages 821–822, 2010.

[10] M. Llorens-Salvador and S. J. Delany. Deep level lexical features for cross-lingual authorship attribution. In *ECIR 2016 Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine), Padua, Italy, March 20, 2016*, pages 16–25, 2016.

[11] M.-F. Moens and I. Vulic. Multilingual probabilistic topic modeling and its applications in web mining and search. In *Proc. of the 17th ACM International Conference on Web Search and Data Mining (WSDM), February 24-28, 2014, New York, NY, USA*, pages 681–682, 2014.

[12] R. Navigli and S. P. Ponzetto. Babelnet: Building a very large multilingual semantic network. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 216–225, 2010.

[13] S. Romeo, D. Ienco, and A. Tagarelli. Knowledge-based representation for transductive multilingual document classification. In *Proc. of the 37th European Conference on IR Research (ECIR), March 29 - April 2, 2015, Vienna, Austria*, pages 92–103, 2015.

[14] S. Romeo, A. Tagarelli, and D. Ienco. Semantic-based multilingual document clustering via tensor modeling. In *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), October 25-29, 2014, Doha, Qatar*, pages 600–609, 2014.

[15] B. Steichen, N. Ferro, D. Lewis, and E. H. Chi. 1st international workshop on multilingual web access (MWA 2015). *SIGIR Forum*, 49(2):137–140, 2015.

[16] I. Vulic and M.-F. Moens. A unified framework for monolingual and cross-lingual relevance modeling based on probabilistic topic models. In *Proc. of the 35th European Conference on Information Retrieval Research (ECIR), March 24-27, 2013, Moscow, Russia*, pages 98–109, 2013.

[17] T. Zhang, K. Liu, and J. Zhao. Cross lingual entity linking with bilingual topic model. In *Proc. of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), August 3-9, 2013, Beijing, China*, 2013.