CONTRIBUTIONS TO MULTILANGUAGE INFORMATION MANAGEMENT USING WIKIPEDIA

Author: Mouriño-García, Marcos A.
marcos@gist.uvigo.es

Thesis Advisor: Anido-Rifón, Luis
lanido@gist.uvigo.es

GIST Group, Department of Telematics Engineering

Motivation

- Large amount of digital information in different languages [1]
- Information organized and easily accesible
  - Information Retrieval
  - Classification
- Common representation of documents:
  - Decisive for the performance of retrieval and classification algorithms
  - Traditionally => BoW model [2,3]
    - Only takes into account frequency of words => suboptimal
    - Redundancy, ambiguity, orthogonality, hiponymy, hypernymy [3,4,5]
- Hypothesis → To leverage Wikipedia concept-based representations alleviates the main drawbacks of BoW model.
  - Useful for classification and retrieval algorithms => ↑ performance
  - Wikipedia Miner semantic annotator to create Wikipedia concept-based representations of documents.

Objectives

- Main objective:
  - To validate the suitability and benefits of using Wikipedia knowledge to improve the performance of different management tasks such as classification and information retrieval.
- Specific objectives:
  - Review of the state of the art:
    - Machine learning techniques.
    - State-of-the-art document representation paradigms.
    - Classification and Information Retrieval algorithms
    - State-of-the-art datasets and algorithms to perform classification and information retrieval tasks.
  - Design and development of the benchmark to validate the suitability of the concept-based representation of documents.
  - Dissemination of results in international conferences and journals.

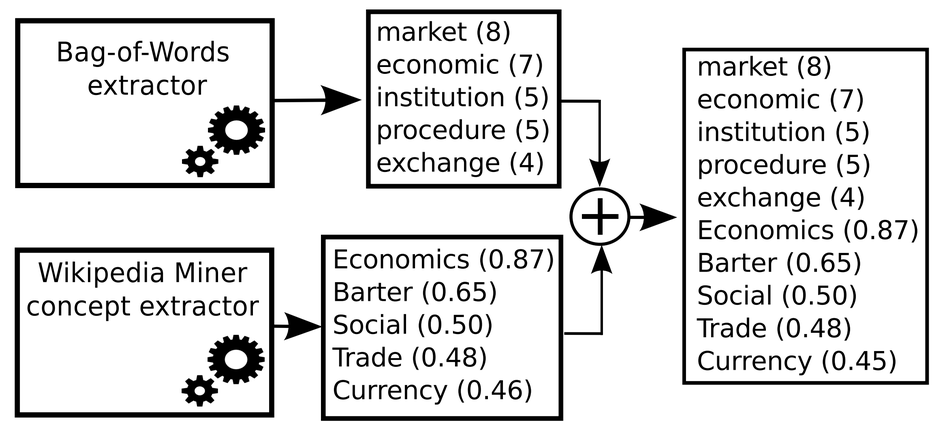
Research Plan

Past years results

- An information retrieval system can be effectively built on top of the world-knowledge provided by Wikipedia.
- The use of the Wikipedia concept-based representation offers a higher performance than the use of the traditional BoW paradigm.

2016 - 2017

Hybrid word-concept representation

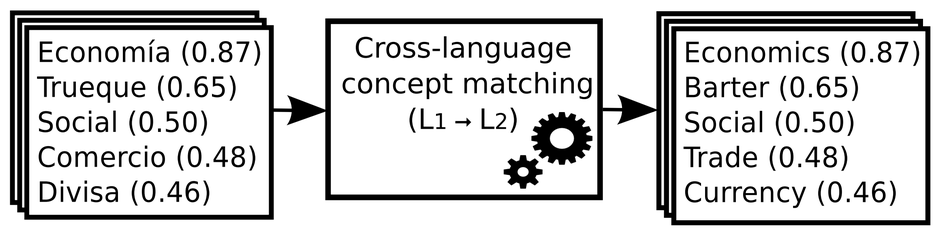


- Evolve benchmark.
- Select / create datasets to perform classification experiments.

Cross-language clasification

- Train a classifier with documents written in a language L_1.
- Classify documents written in a language L_2.

Cross-Language Concept Matching (CLCM)



L_1 and L_2 are any pair of languages that have Wikipedia.

- Evolve benchmark.
- Select / create datasets to perform classification experiments.

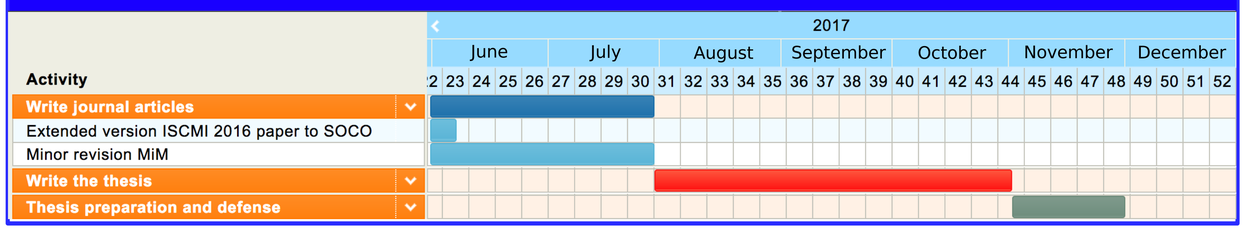
Results and Discussions

- 1 JCR (Information Sciences) [6]
- 2 conference papers
  - ISCMI 2016, Dubai, UAE [7] → Shortlisted for Soft Computing JCR journal
  - IEEE CIT 2016, Fiji [8]
- 1 JCR with minor revisions (Methods of Information in Medicine)
- Pure Wikipedia concept-based representations are advantageous in applications where training data is limited.
- Hybrid approaches ↑ performance.
  - Wikipedia concepts add valuable information to the classifier.
    - Especially with documents in the biomedical field.
- No semantic weight loss between languages in cross-language tasks

References

[1] Ferrández, S. et al. Exploiting wikipedia and eurowordnet to solve cross-lingual question answering. Information Sciences, 179(20):3473-3488. 2009.
[2] Salton, G. et al. A vector space model for automatic indexing. Communications of the ACM, 18(11):613-620, 1975.
[3] Täckström, O. An Evaluation of Bag-of-Concepts Representations in Automatic Text Classication. PhD thesis, 2005.
[4] Wang, P. et al. Using wikipedia knowledge to improve text classication. Knowledge and Information Systems, 19(3):265-281, 2009.
[5] Ming, Z.-Y. and Chua, T. S. (2015). Resolving polysemy and pseudonymity in entity linking with comprehensive name and context modeling. Information Sciences, 307:18–38.
[6] Mouriño-García, M.A. et al. Wikipedia-based cross-language text classification. Information Sciences, 406, 12-28. (2017)
[7] Mouriño-García, M.A. et al. Wikipedia-Based Hybrid Document Representation for Textual News Classification. In Proc. of ISCMI 2016 : 3rd Intl. Conference on Soft Computing & Machine Intelligence, Dubai, UAE, 2016. ISBN 978-1-5090-1696-7.
[8] Mouriño-García, M.A et al. Bag-of-Concepts Document Representation for Bayesian Text Classification. In Computer and Information Technology (CIT), 2016 IEEE International Conference on (pp. 281-288). IEEE.

Next Year Planning



Universida de Vigo