+

# MultiLingMine 2016 1st Int. Workshop on Modeling, Learning and Mining for Cross/Multilinguality

ecir PADUA 2016

38th European Conference on Information Retrieval

March 20, 2016
Padua, Italy

+ 
# Organizers

- **Dino Ienco**, TETIS-LIRMM, Montpellier, France
- **Mathieu Roche**, TETIS-LIRMM, Montpellier, France
- **Salvatore Romeo**, QCRI, Doha, Qatar
- **Paolo Rosso**, Universitat Politècnica de València, Valencia, Spain
- **Andrea Tagarelli**, Dipartimento di Ingegneria Informatica, Modellistica, Elettronica, e Sistemistica (DIMES), Università della Calabria, Italy

# + Scientific Advisors

- *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 Politècnica de València, 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

# Accepted papers - Program

- 14:30 – 14:35 — **Opening**
- 14:35 – 15:20 — **Invited Talk** (Nicola Ferro, Univ. Padua, Italy)
- 15:20 – 15:40 — "*Identification of Disease Symptoms in Multilingual Sentences: an Ontology-Driven Approach*", by Angelo Ferrando, Silvio Beux, **Viviana Mascardi** and Paolo Rosso
- 15:40 – 16:00 — "*Deep Level Lexical Features for Crosslingual Authorship Attribution*", by **Marisa Llorens-Salvador** and Sarah Jane Delany
- Coffee break
- 16:30 – 16:50 — "*Profile-based Translation in Multilingual Expertise Retrieval*", by Hossein Nasr Esfahani, Azadeh Shakery and Javid Dadashkarimi. Speaker: …..
- 16:50 – 17:10 — "*Extending Automatic Discourse Segmentation for Texts in Spanish to Catalan*", Iria Da Cunha, **Eric Sanjuan**, Juan-Manuel Torres-Moreno, Irene Castellón and Marina Lloberes
- 17:10 – 17.30 — "*A New Image Analysis Framework for Latin and Italian Language Discrimination*", by Darko Brodic, **Alessia Amelio** and Zoran N Milivojevic
- 17:30 – 17:50 — "*The First Cross-Script Code-Mixed Question Answering Corpus*", by Somnath Banerjee, Sudip Kumar Naskar, Paolo Rosso and Sivaji Bandyopadhyay
- 17:50 – 18:15 — **Panel Discussion and Closing**

# Invited Talk

- **Title**: Multilingual Information Access: What and How Well?
- **Speaker**: prof. *Nicola Ferro*, Univ. Padua, Italy
- **Abstract**: Measuring is a key to scientific progress. This is particularly true for research concerning complex systems, whether natural or human-built. 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 worldwide 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. This talk will thus introduce the base challenges and approaches to multilingual information access (MLIA) and discuss what performance trends emerge from several years of MLIA evaluation at CLEF, the European forum for multilingual and multimodal information access evaluation.

# Panel Discussion

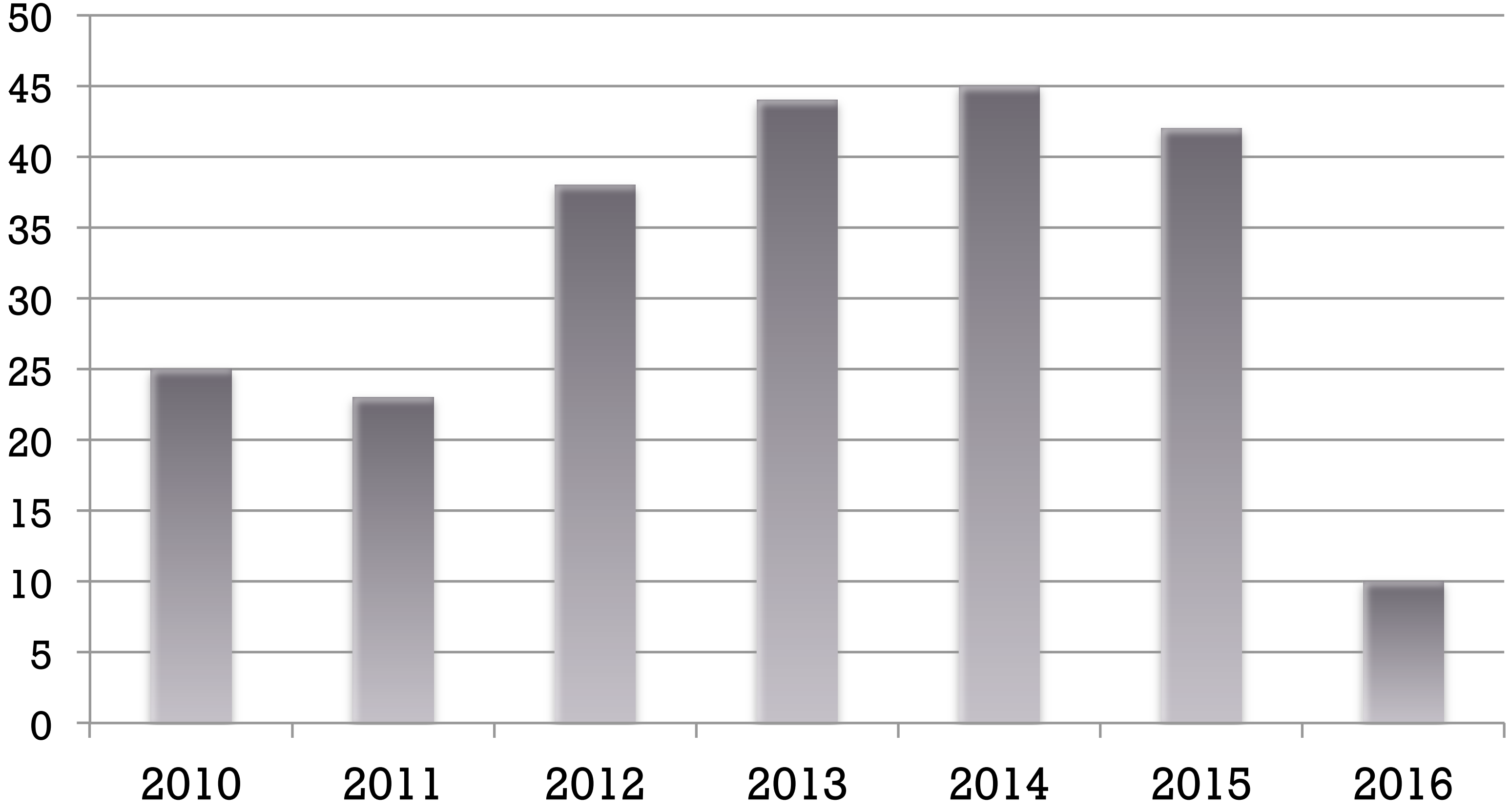
## MultiLingMine 2016 1st Int. Workshop on Modeling, Learning and Mining for Cross/Multilinguality

ecir PADUA 2016
38th European Conference on Information Retrieval

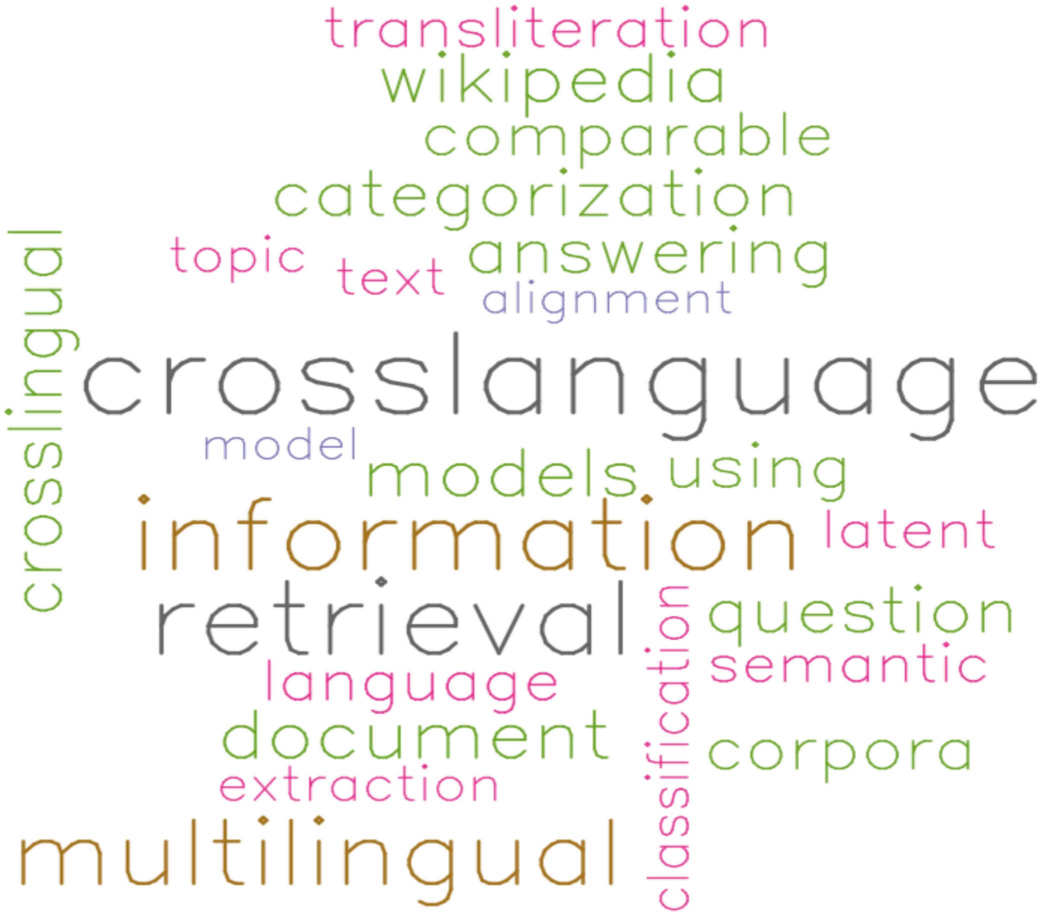
March 20, 2016
Padua, Italy

# Last 6-7 years in CL/ML IR

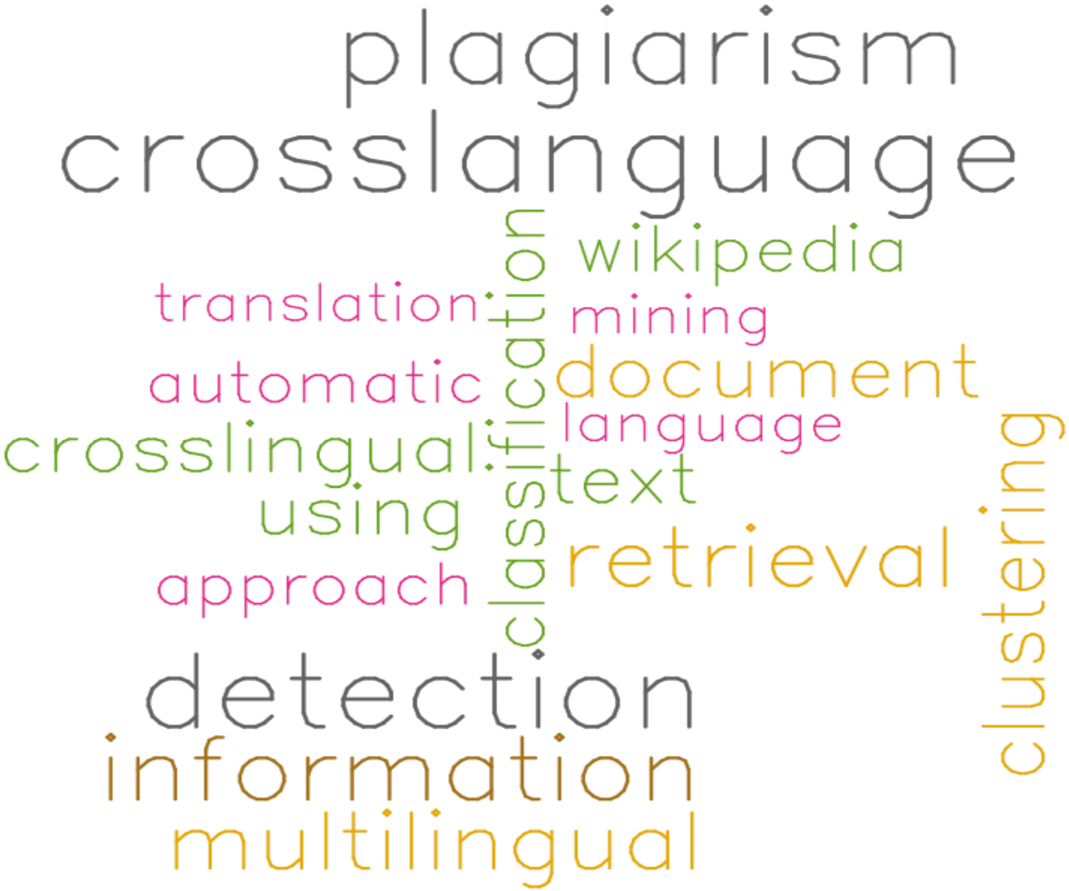
# + Nr. Papers in Cross/Multi-Lang. from 2010 to 2016 (source DBLP)



# CL/ML-IR research keywords: 2010

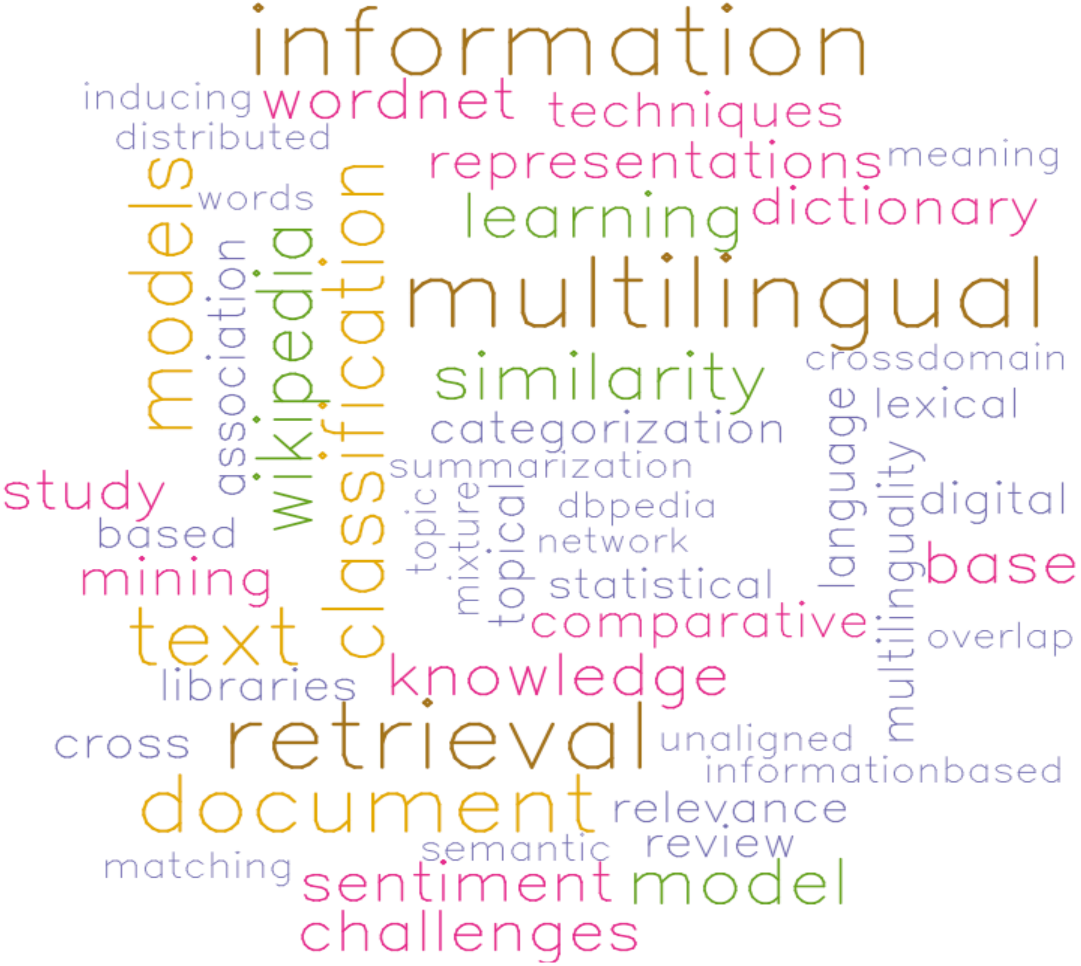


# + CL/ML-IR research keywords: 2011

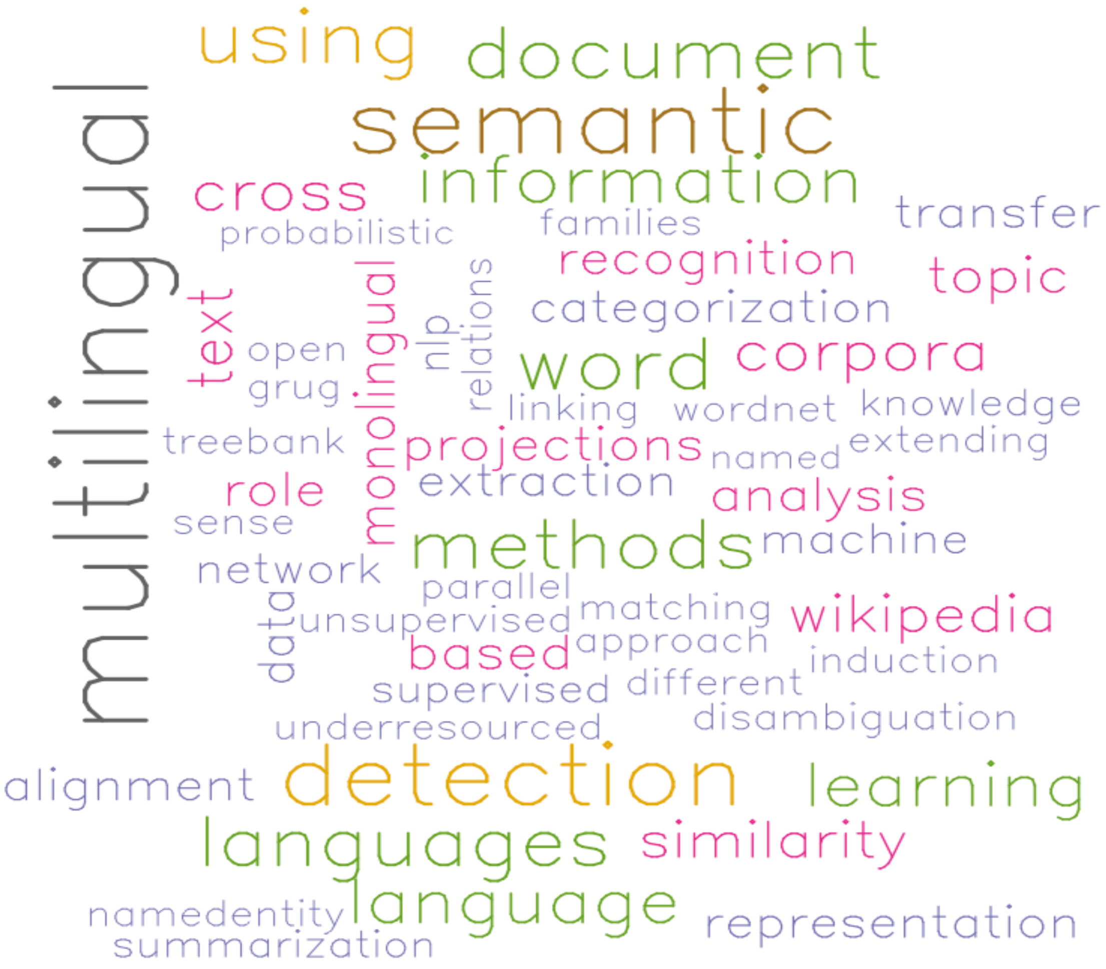


+

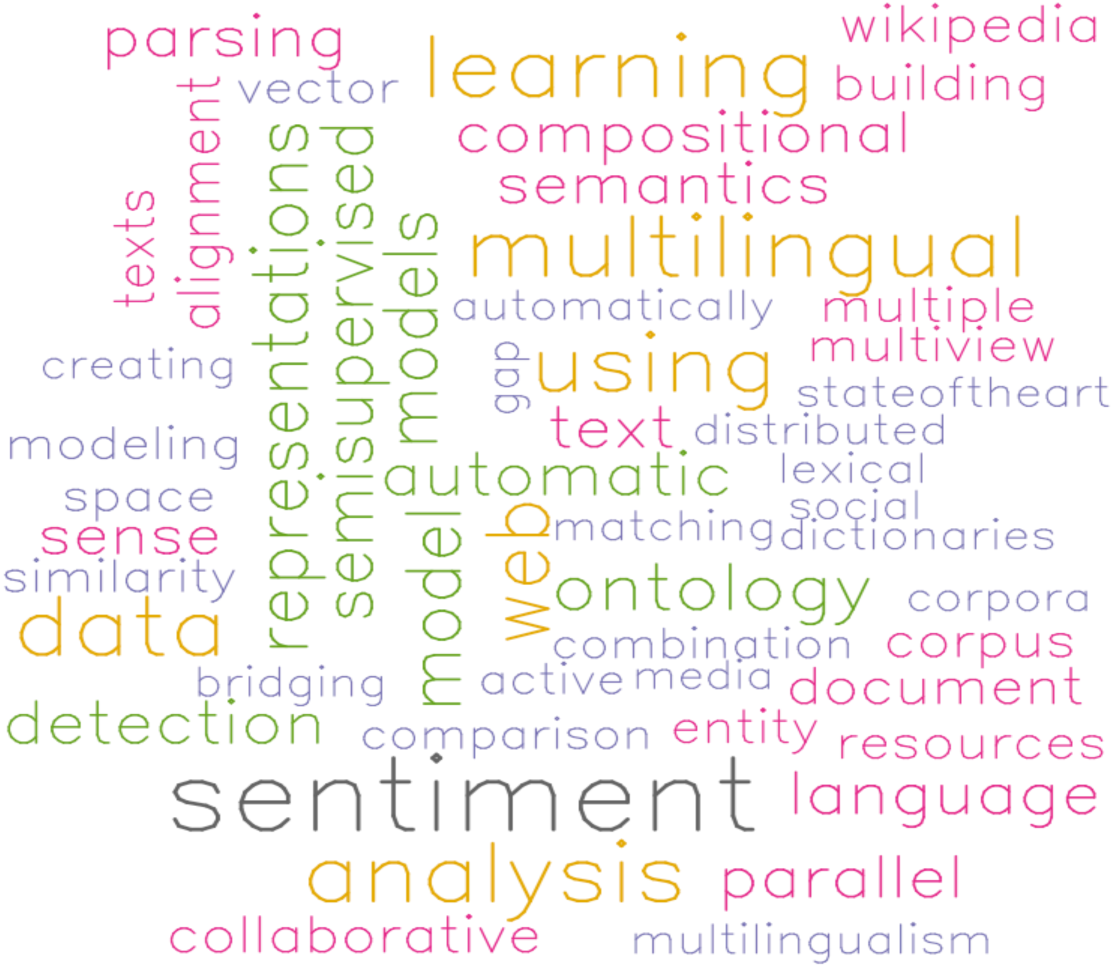
# CL/ML-IR research keywords: 2012



# CL/ML-IR research keywords: 2013

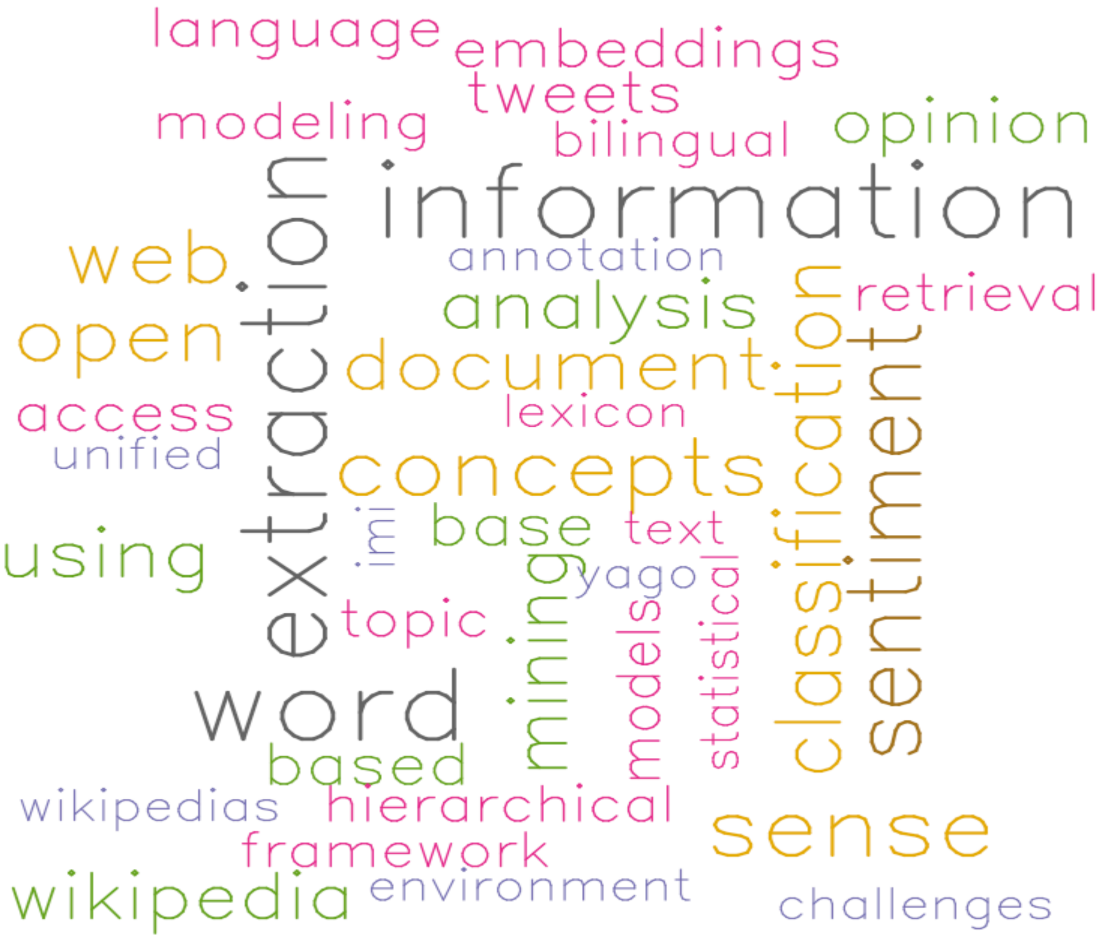


# CL/ML-IR research keywords: 2014

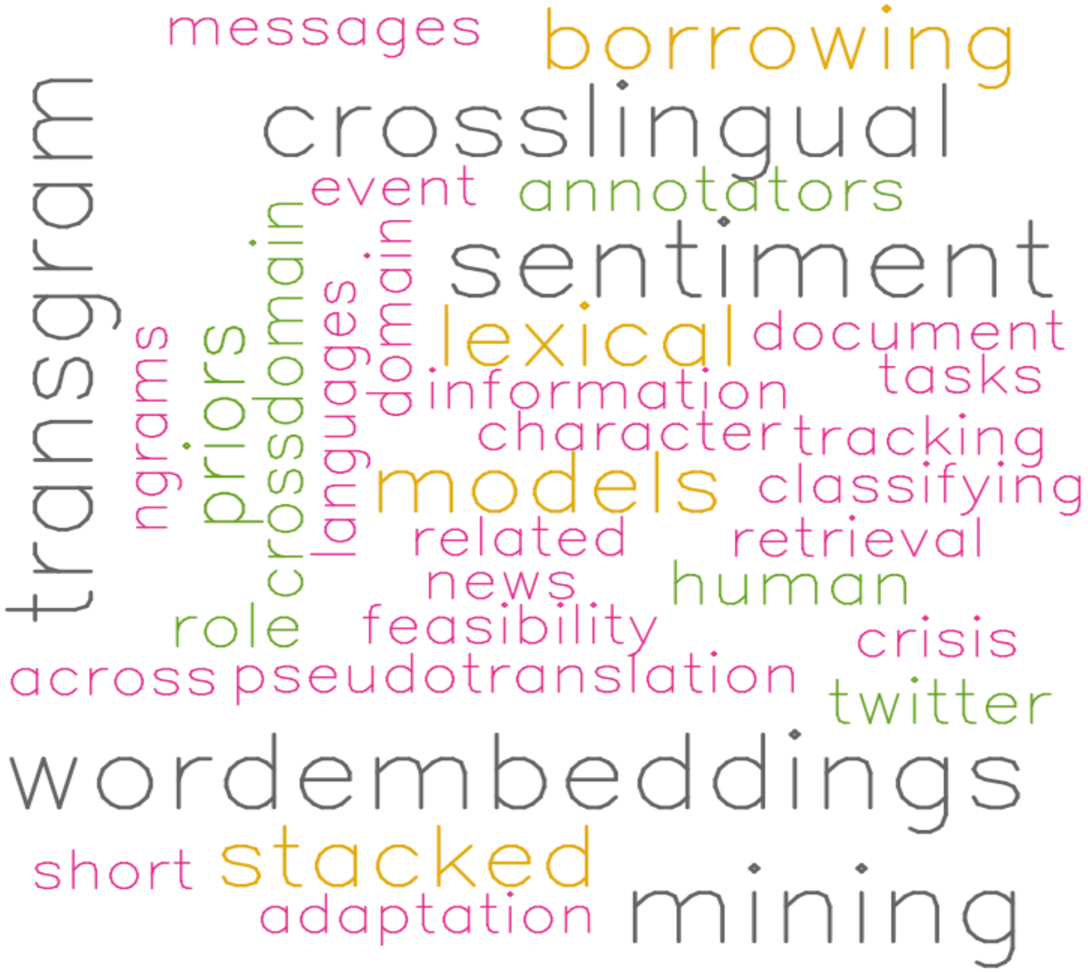


+

# CL/ML-IR research keywords: 2015



# CL/ML-IR research keywords: 2016



# + MultiLingMine: overview of results

# Main stated objectives

- **Modeling**: methods to develop suitable representations for multilingual corpora, possibly embedding information from different views/aspects
- **Learning**: any unsupervised, supervised, and semi-supervised approach in cross/multilingual contexts
- Use of **knowledge bases** to support the modeling, learning, or both stages of multilingual corpora analysis
- **Emerging trends** and **applications**

# + Some research opportunities

- **Define a translation-independent representation of the documents across many languages**
- **Exploit knowledge bases to enable translation-independent preserving and unveiling of content semantics**
- **Exploit multi-lingual knowledge bases for Q/A**
- Enhance existing solutions for comparable corpora to handle multiple languages (w/o depending on bilingual dictionaries or incurring bias in merging language-specific results)
- Define indexing and multidimensional data structures to better capture the multi-topic/ aspect nature of multi-lingual documents
- Detect duplicate/redundant and/or novel information among different languages
- Enrich and update multi-lingual knowledge bases from documents
- Efficiently extend topic modeling to deal with multi/cross-lingual documents in many languages
- Evaluate and visualize retrieval and mining results

# Main covered topics

- Translation-independent representation
- Cross-lingual analysis via translation model
- Ontology-Driven approaches
- Improvement in the analysis of poor-resources language
- Text segmentation
- Question Answering
- Evaluation of Cross-lingual analysis

# Pointers for further research

- Word embeddings/deep learning for (groups of) language semantic spaces
- Mixed languages
  - Mobile environment
  - Recognition of Twitter messages

    -- > long vs. short texts
- Noisy
- Benchmarks
- Living labs

+

# Thank you for joining us

MultiLingMine 2016 1st Int. Workshop on Modeling, Learning and Mining for Cross/Multilinguality

ecir PADUA 2016
38th European Conference on Information Retrieval

March 20, 2016
Padua, Italy