MULTILINGUAL DOCUMENT CLASSICATION VIA TRANSDUCTIVE LEARNING

Salvatore Romeo – UNICAL
srome@dimes.unical.it

Dino Ienco - IRSTEA, LIRMM
dino.ienco@irstea.fr

Andrea Tagarelli – UNICAL
tagarelli@dimes.unical.it

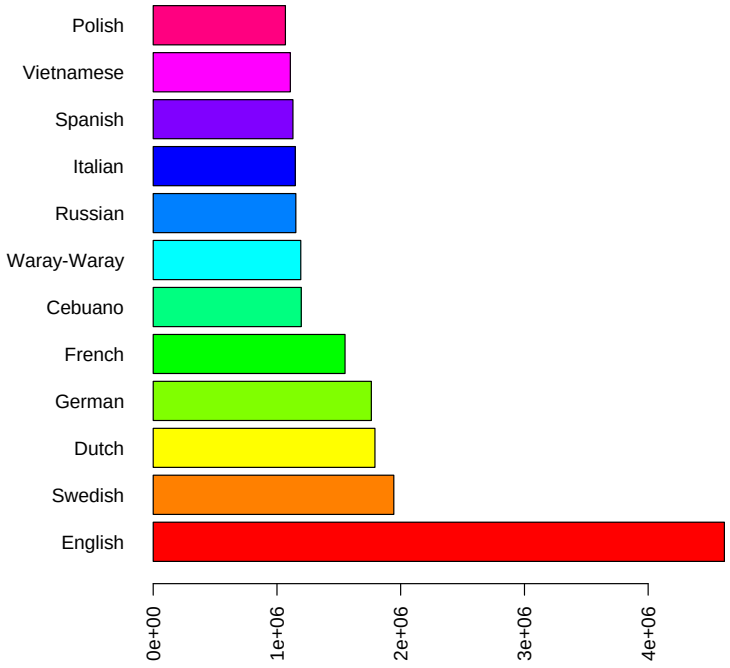
Introduction:
Multilingual information overload

- Increased popularity of systems for collaboratively editing through contributors across the world

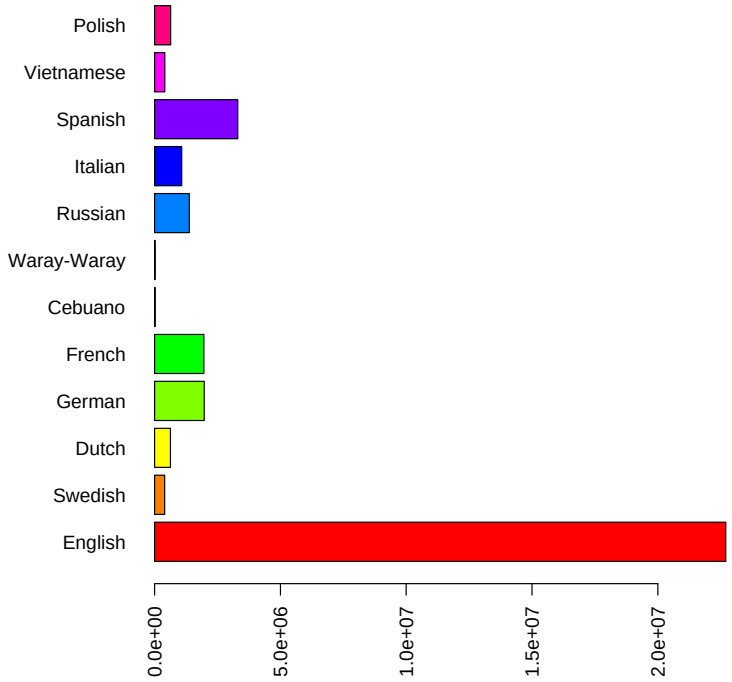
- Massive amounts of text data written in different languages

Introduction:
Multilingual information overload

1million+ Wikipedia articles



…and corresponding registered users



Source: Wikipedia (October 6, 2014)

Motivations & Issues:
From monolingual to multilingual analysis

- Discover and exchange knowledge at a larger world-wide scale

- Requires enhanced technology
  - Translation and multilingual knowledge resources
  - Cross-linguality tools
  - Topical alignment or sentence-alignment between document collections
    - Comparable vs. parallel corpora

"The Tower of Babel", P. Bruegel (ca. 1563)

Motivations & Issues:
Cross-Lingual approaches

- Customized for a small set of languages (e.g., 2 or 3)
- Hard to generalize to many languages
  - Use of bilingual dictionaries
  - Sequential, pairwise language translation
- Bias due to merge of language-specific results independently obtained
- Noise introduced by machine translation
- Performance may vary depending on the source and target languages

- ➔ Emergence for
  - A language-independent representation of the documents across many languages, without using translation dictionaries

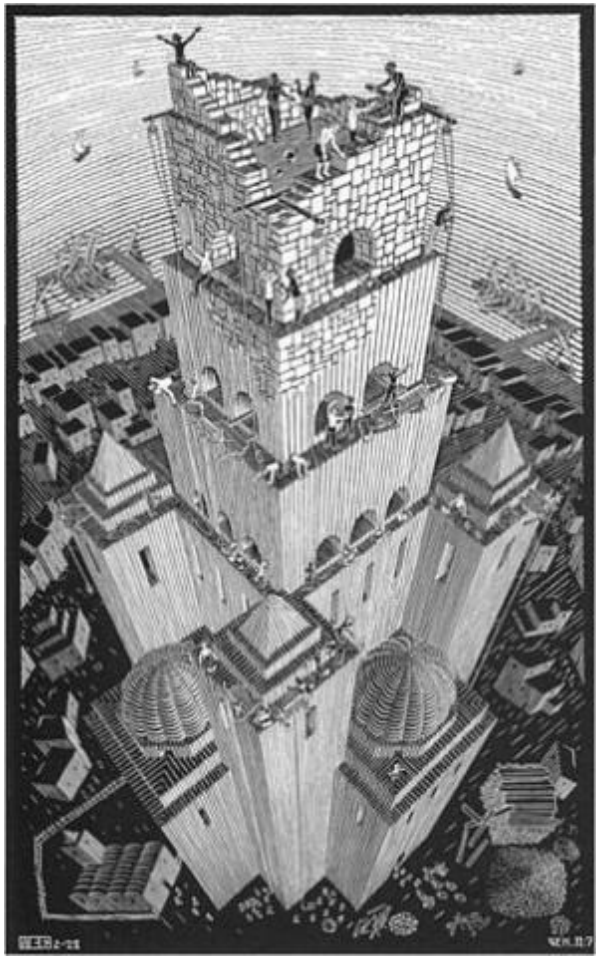
Motivations & Issues:
Issues in Multi-lingual Document Classification (MDC):

- Document labels might be more difficult to obtain
  - More language-specific experts need to be involved in the annotation process
- Test data can be available at the same time of training data, but
- It might be comprised of documents written in different languages than labeled documents

Our proposal:

Knowledge-based Representation for Transductive Multilingual Document Classification

- Key aspects:
  - Model the multilingual documents over a unified conceptual space
    - Generated through a large-scale multilingual knowledge base: **BabelNet**
    - Enables **translation-independent** preserving of the content semantics
  - Employ a **Transductive** Learning setting to perform MDC



"Tower of Babel", M. C. Escher (1928)

Our proposal:
Model the multilingual documents

- BabelNet: encyclopedic dictionary [Navigli & Ponzetto, 2012]
  - Providing concepts and named entities in different languages
  - Connected through (**WordNet**) semantic relations and (**Wikipedia**) topical associative relations

- BabelNet Structure:
  - Encoded as a labeled directed graph
    - Concepts and named entities, as nodes
    - Links between concepts, labeled with semantic relations, as edges
  - Babel synset (a node):
    - Contains a set of lexicalizations of the concept for different languages

[Navigli & Ponzetto, 2012] BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. 2012

Our proposal:

Model the multilingual documents

- Knowledge-based text representation widely used in monolingual contexts
  - e.g., [Ramakrishnanan and Bhattacharyya, 2003; Semeraro et al., 2007; Lops et al., 2007; de Gemmis et al., 2008]
- Semantic document features = BabelNet synsets
- 3-step procedure:
  - Perform **lemmatization** and **POS-tagging** on every document
  - Perform **WSD** to each pair (lemma, POS-tag) contextually to the sentence which the lemma belongs to
  - Model each document as a ***m*-dimensional vector** of BabelNet synset (*m* is the no. of synsets retrieved)

Transductive inference

- It needs **partial supervision**
  - a small portion of the documents needs to be labeled (labels difficult to obtain)
- Inference "**from particular to particular**"
  - Does not induce any general rule to classify new unseen docs (training and test data available together)
- Classification of unlabeled documents provided contextually to learning the currently labeled documents
  - Relevance feedback, filtering, document reorganization

[Joachims, 1999] Transductive Inference for Text Classification using Support Vector Machines. ICML, 1999.
[Joachims, 2003] Transductive learning via spectral graph partitioning..ICML, 2003.

RMGT

- Transductive learning: “from particular to particular”
- Natural implementation in case-based learning algorithms

- Robust Multi-class Graph Transduction (RMGT) [Liu & Chang, 2009]
  - State-of-the-art transductive learner [de Sousa et al., ECML-PKDD, 2013]
  - Implements a graph-based label propagation approach
    - i.e., exploits a kNN graph built over the entire document collection to propagate the class information from the labeled to the unlabeled documents

[Liu & Chang, 2009] W. Liu, S.-F. Chang: Robust multi-class transductive learning with graphs. CVPR 2009
[de Sousa et al, 2014] C. A. R. de Sousa, S.O. Rezende, G. E. A. P. A. Batista: Influence of Graph Construction on Semi-supervised Learning. ECML/PKDD, 2013

Our proposal:

Transductive Multiglingual Document classification

Key steps:

1. Bag of Synsets representation for multilingual documents
2. Graph-Based transductive learner (RMGT) upon BoS model.

Algorithm 1. Transductive classification of multilingual documents

Input: A collection of multilingual documents $\mathcal{D}$, with labeled documents $\mathcal{L}$ and unlabeled documents $\mathcal{U}$ (with $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$ and $\mathcal{L} \cap \mathcal{U} = \emptyset$); a set of labels $\mathcal{C} = \{C_j\}_{j=1}^{M}$ assigned to the documents in $\mathcal{L}$; a positive integer k for the neighborhood selection.

Output: A classification over $\mathcal{C}$ for the documents in $\mathcal{U}$.

1. Model each document in $\mathcal{D}$ using BoS or alternative representations.
2. Build the similarity graph $\mathcal{G}$ for the document collection $\mathcal{D}$.
3. Extract the k-nearest neighbor graph $\mathcal{G}_k$ from $\mathcal{G}$.
4. Build the matrix $\mathbf{W}$ from $\mathcal{G}_k$, which represents the symmetry-favored k-nearest neighbor graph.
5. Compute the normalized Laplacian of $\mathbf{W}$.
6. Compute the $RMGT$ solution $\mathbf{F}$.
7. Assign document $d_i \in \mathcal{U}$ to the class C_{j^*} that maximizes the class likelihood, $j^* = \arg\max_j \mathbf{F}_{ij}$.

Experimental evaluation

Data and setting (I)

- RCV2 and Wikipedia *balanced* datasets
  - English, French, and Italian documents
  - Cover six different topics

	RCV2	Wikipedia
# of docs	15 300	18 000
# of terms	12 698	15 634
# of synsets	10 033	10 247
BoW density	4.56E-3	1.61E-2
BoS density	3.87E-3	1.81E-2

- Both are *comparable* corpora, but
  - In RCV2, different language-written documents belonging to the same topic-class do not share the content subjects,
  - In Wikipedia, different language-specific versions of articles discussing the same Wiki concept

Experimental evaluation

Data and setting (II)

Different Document Representations:

a) **Machine Translation**: MT-fr, MT-it, MT-en
b) **Bag of Words (BoW)**: union of language-specific term vocabularies
c) **BoW-LSA**: Latent Semantic Analysis over the BoW space
d) **Bag of Synsets (BoS)**

- RMGT setup
  - k = 10 (to build the KNN graph)

Percentage of labeled documents from 1% to 20%

Results are averaged over 30 runs

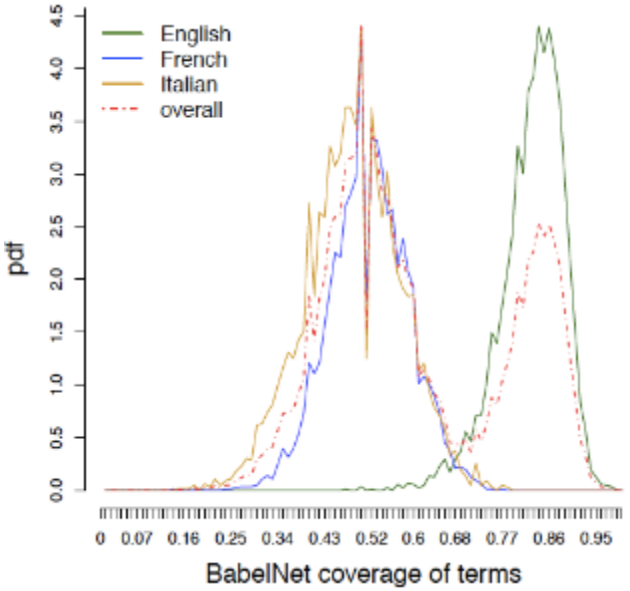
Experimental evaluation

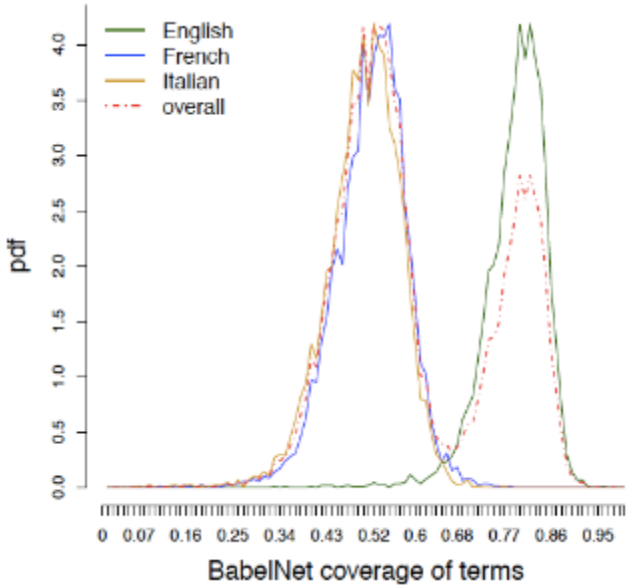
BabelNet coverage

- Per-language distributions of BabelNet Coverage:

  fraction of words belonging to the document whose concepts are present as entries in BabelNet

RCV2



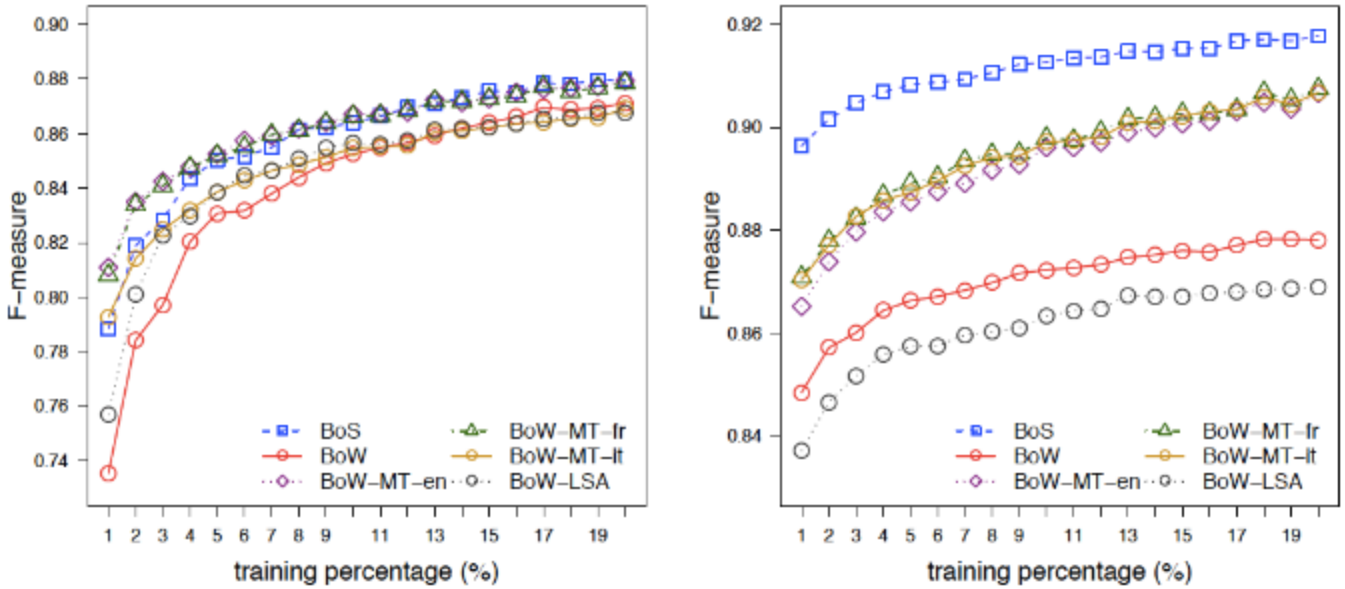


Wikipedia

- French and Italian documents determine the left peak of the overall distribution, whereas
- English documents correspond to negatively skewed distributions

Experimental evaluation

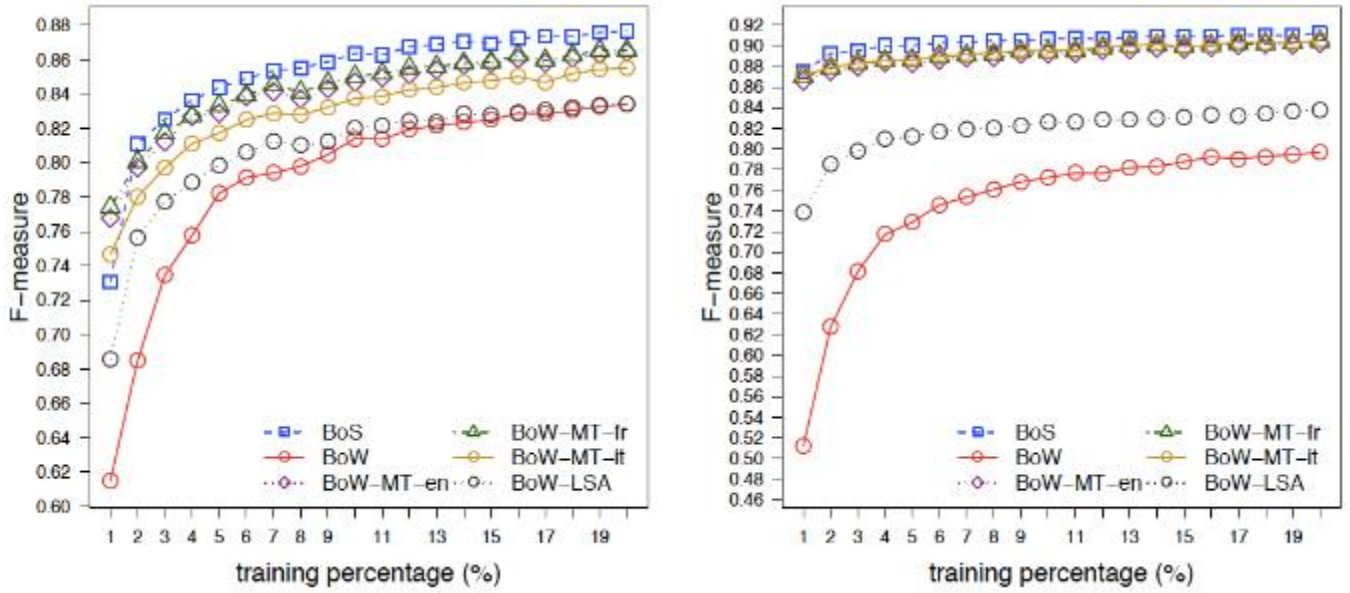
Classification performance



- On RCV2 (left), BoS comparable to the best competitors (BoW-MT-en, BoW-MT-fr)
- On Wikipedia (right), BoS outperforms the others
- BoS performance trend is not affected by language-specificity issues (unlike MT-based models)

Experimental evaluation

Classification performance (language unbalanced)



- On RCV2 (left), BoS behaves now better than the MT-based models (which have decreased their performance w.r.t. the balanced case)
- On Wikipedia (right), no change in the relative performance between BoS and MT-based models

Summary of results

- Effective and robust approach to multilingual document classification

- Bag-of-synsets model
  - achieves, in general, better results than various language-dependent models,
  - preserves its performance on both balanced and unbalanced datasets

- Transductive learning framework performs well using a very small (5%) portion of the available labeled documents

Future work

- BabelNet
  - Integrate more types of information (i.e., relations between synsets) to define richer multilingual document models
- Transductive & Active learning
  - Aid solicit user interaction in order to guide the labeling process
  - Applications to document reorganization tasks
- Consider the Multi-Topic nature of documents
  - Long documents usually contains more than one topic
  - Model document as complex structure (segment set)

Thank you for your attention

Datasets available at

uweb.dimes.unical.it/tagarelli/data

Questions?