EMNLP 2014
CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING
DOHA, QATAR. OCTOBER 25–29, 2014

SEMANTIC-BASED MULTILINGUAL DOCUMENT CLUSTERING VIA TENSOR MODELING

Salvatore Romeo[1], Andrea Tagarelli[1], and Dino Ienco[2]

[1] DIMES, University of Calabria, Rende, Italy
[2] IRSTEA - UMR TETIS, and LIRMM, Montpellier, France

UNIVERSITÀ DELLA CALABRIA
Dipartimento di INGEGNERIA INFORMATICA, MODELLISTICA, ELETTRONICA E SISTEMISTICA

TETIS

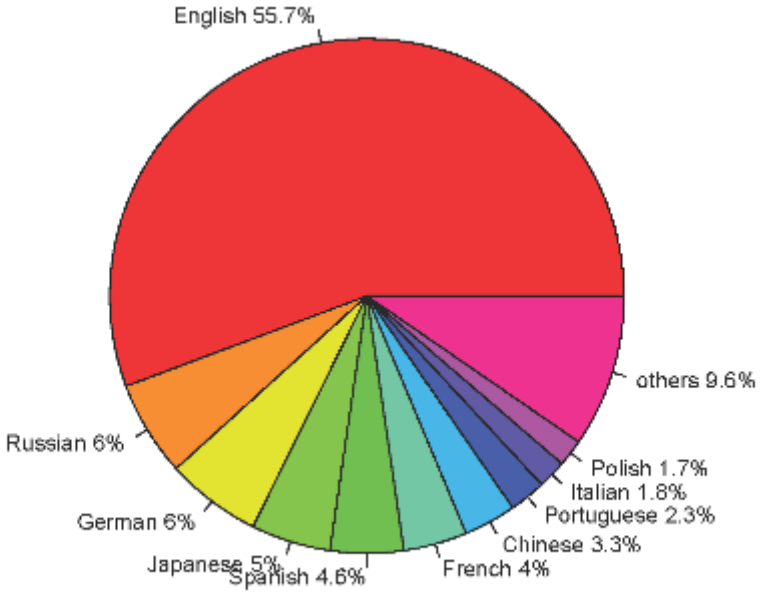
Laboratoire Informatique Robotique Microélectronique Montpellier

Multilingual information overload

- Increased popularity of systems for collaboratively editing through contributors across the world
- Massive amounts of text data written in different languages

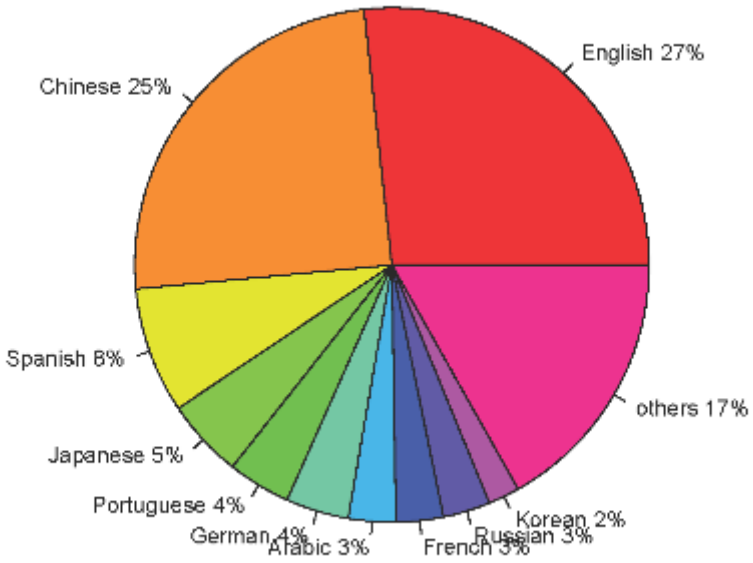
Multilingual information overload

Content languages for websites



Source: W3Techs.com (March 12, 2014)

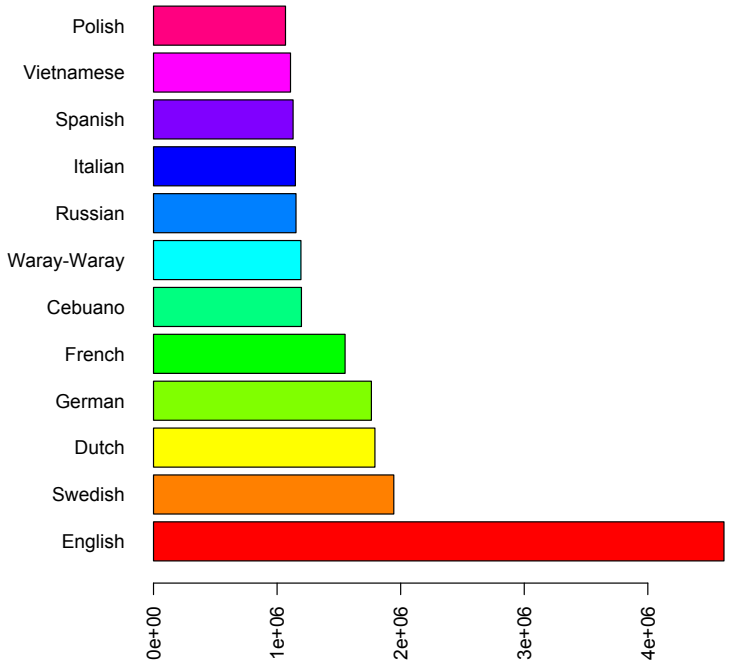
Internet users by language



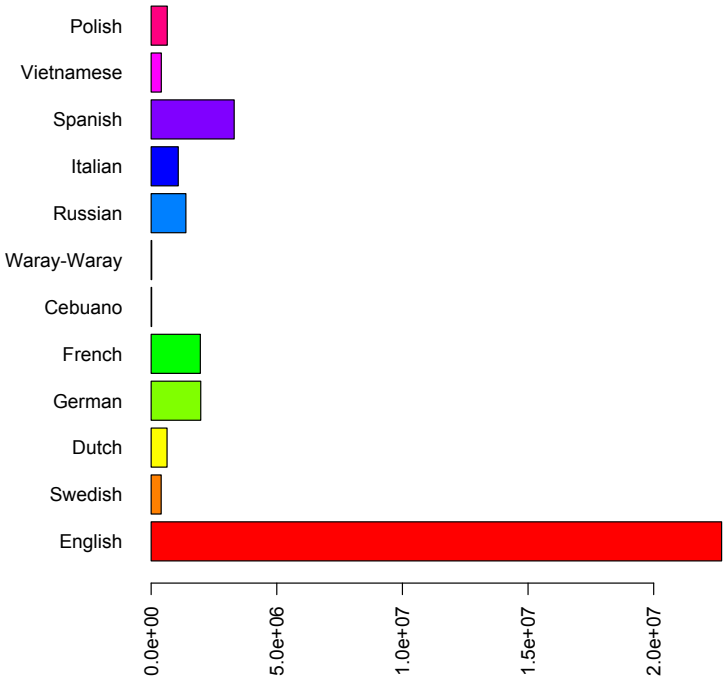
Source: Internet World Stats (May11, 2011)

Multilingual information overload

1million+ Wikipedia articles



…and corresponding registered users



Source: Wikipedia (October 6, 2014)

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

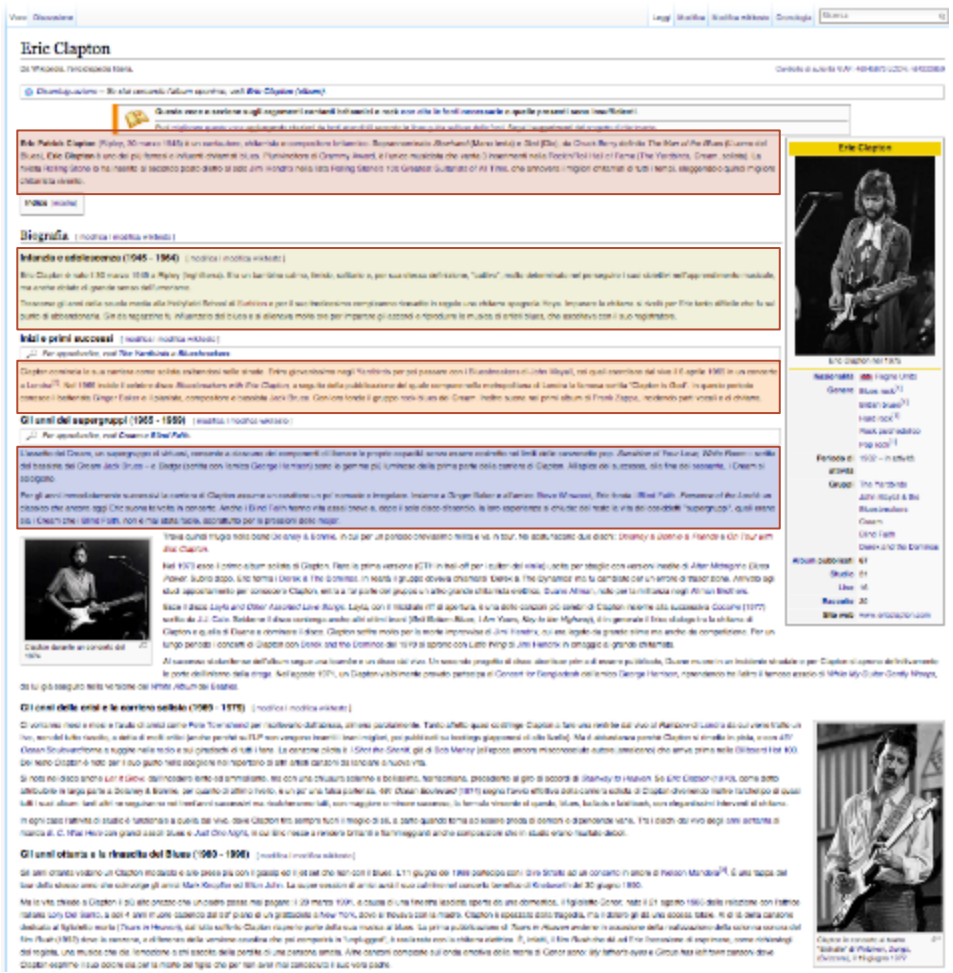
"The Tower of Babel", P. Bruegel (ca. 1563)

Multilingual document analysis

- Comparable corpora
  - Contain documents with non-aligned sentences, which are not exact translations of each other, but still thematically aligned
  - Usually available in abundance:
    - Wikipedia, Amazon, news sites, etc.
  - But often unstructured and noisy
    - Words/terms have multiple senses per corpus
    - Terms have multiple translations per corpus
    - Translations might not exist in the target document
    - Frequencies and positions are generally not comparable

Wikipedia: example comparable corpus

Eric Clapton: Italian Wikipage



Eric Clapton: English Wikipage

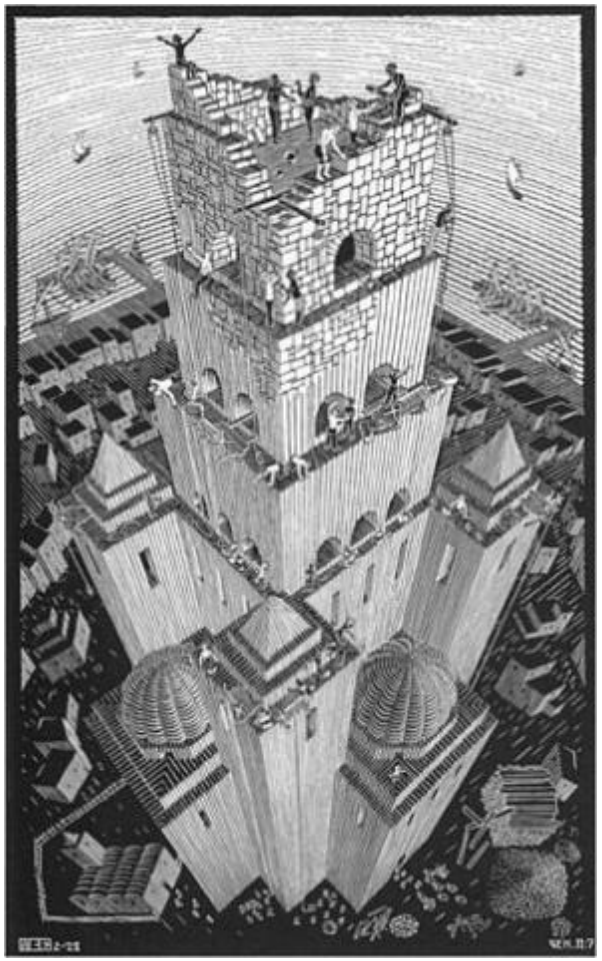
Why do CL approaches fail

- 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

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

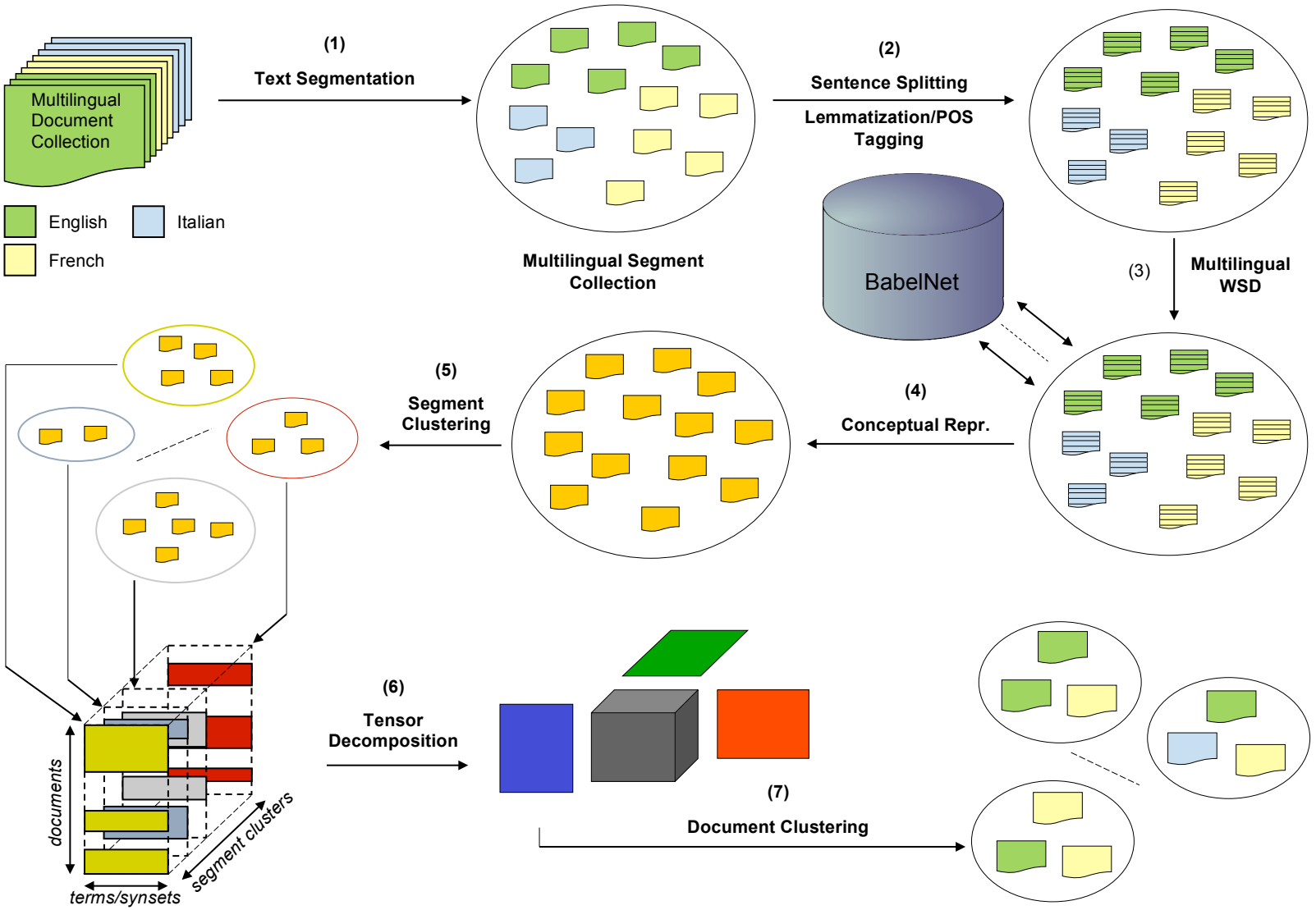
Knowledge-based multilingual document modeling: our proposal

- Key aspects:
  - Model the multilingual documents over a unified conceptual space
    - Generated through a large-scale multilingual knowledge base: BabelNet
    - Enables language-independent preserving of the content semantics
  - Decompose the multilingual documents into topically-coherent segments
    - Enables the grouping of linguistically different portions of documents by content
  - Describe the multilingual corpus under a multi-dimensional data structure
    - Third-order tensor model



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

Multilingual Document Clustering: Framework Overview



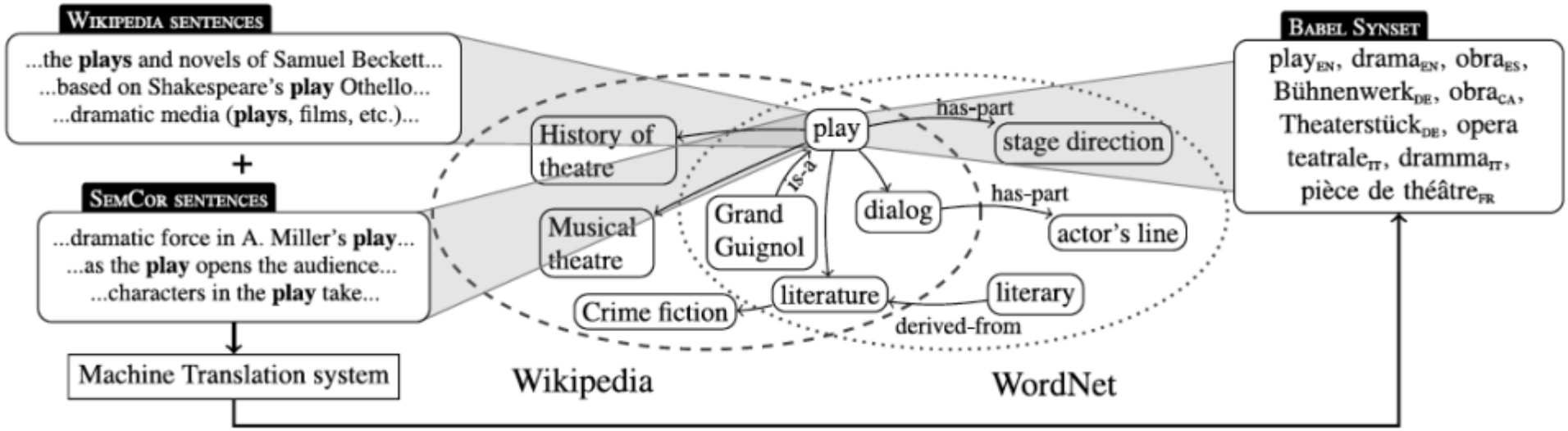
BabelNet (1/6)

- Links Wikipedia, i.e.,
  - the largest and most popular collaborative and multilingual resource of world knowledge
  - however lacking full coverage for lexicographic senses
- with WordNet, i.e.,
  - the most popular lexical ontology
  - computational lexicon of the English language, based on psycholinguistic principles
- via automatic mapping and filling in lexical gaps in resource-poor languages via MT

- BabelNet: encyclopedic dictionary [Navigli & Ponzetto, Artificial Intelligence, 2012]
  - Providing concepts and named entities in 6 (6 erano nella prima versione, ora sono di più) languages
  - Connected through (WordNet) semantic relations and (Wikipedia) topical associative relations

BabelNet (2/6)

- 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, Artificial Intelligence, 2012]

BabelNet (3/6)

Semantic network construction

1. Mapping WordNet senses and Wikipages
2. Harvesting multilingual lexicalizations of the available concepts (i.e., Babel synsets) by using
   - the human-generated translations provided by Wikipedia (i.e., *inter-language* links), and
   - a MT system to translate occurrences of the concepts within sense-tagged corpora
3. Establishing semantic relations between Babel synsets, and determining semantic relatedness

BabelNet (4/6)

Mapping algorithm:

- Each Wikipage, whose lemma is monosemous in both WordNet and Wikipedia, is mapped to a unique WordNet sense
- Each Wikipage, which is a redirection to a mapped Wikipage, is mapped to the pointed Wikipage's sense
- All remaining Wikipages are mapped to the WordNet sense which maximizes the conditional probability p(*w*|*s*), where *w* is the lemma of the particular Wikipage and *s* is a WordNet sense associated with *w*

- WSD process:
  - Graph-based algorithm
  - Disambiguation context for every concept (Wikipage or WordNet sense): set of words derived from the corresponding resource that are semantically related to the concept

BabelNet (5/6)

Translating BabelNet synsets

- After the mapping step, only English Wikipages are linked to WordNet senses
- Given a Wikipage *w* and related WordNet sense *s*, the corresponding Babel synset is comprised of:
  - The synset to which s belongs
  - The Wikipage *w*
  - The set of redirections to *w*
  - All pages linked by means of inter-language links
  - The set of the redirections to the Wikipages linked by the inter-lingual links

BabelNet (6/6)

Translating BabelNet synsets

- Issues:
  - A concept might be covered by only one of the two resources
  - The Wikipages related to a concept might not have inter-lingual links for the languages of interest
- … and solutions:
  1. For each English lexicalization of the Babel synset, retrieve
     - The occurrences in SemCor for a given WordNet sense
     - The sentences in Wikipedia which link the Wikipages of interest
  2. Translate the resulting set of sentences to all languages of interest
  3. For each term of the original Babel synset, keep the most frequent translation for each of the languages

Text segmentation

- No assumption based on paragraph boundaries
- Standard approach: Identify segment-boundaries by detecting thematic shifts in the text
  - TextTiling algorithm [Hearst, 1997]
    - Subdivides a text into multi-paragraph, contiguous, disjoint blocks
    - Terms discussing a topic tend to co-occur locally:
      - topic switch detected by the ending/beginning of co-occurrence of a given set of terms
    - Segment boundaries are inferred from min values in the sequence of cosine-sim values for all pairs of adjacent blocks
- Note that alternative text segmentation algorithms can be used

Bag-of-synsets model

- Semantic document features = BabelNet synsets

- 3-step procedure
  - Perform lemmatization and POS-tagging on every segment
  - Perform WSD to each pair (lemma, POS-tag) contextually to the sentence which the lemma belongs to
  - Model each segment as a BS-dimensional vector of BabelNet synset (BS is the no. of synsets retrieved)

Bag-of-synsets model

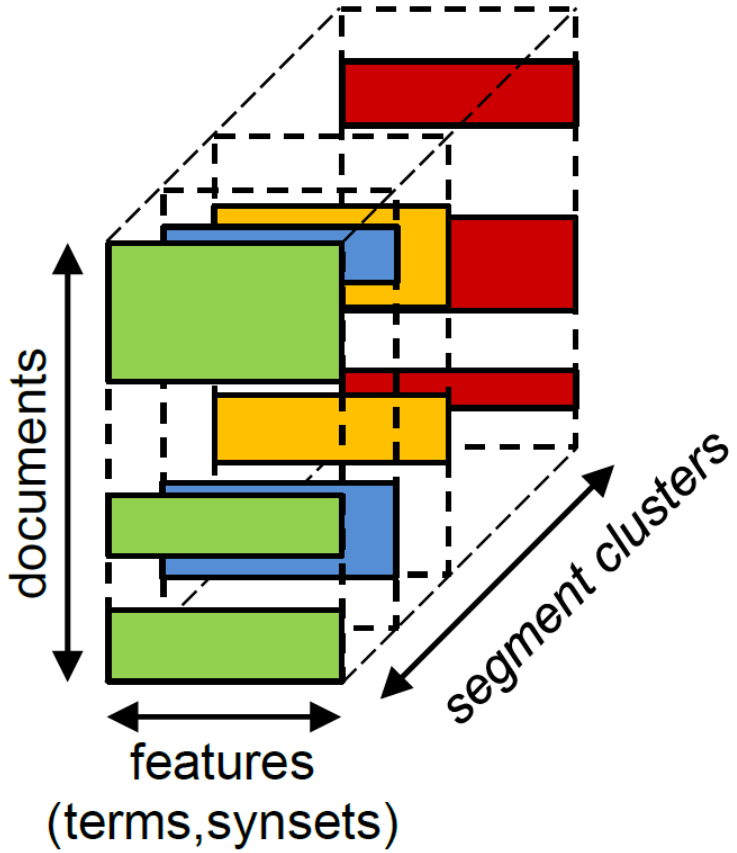
WSD step

- Graph-based eigenvector ranking methods
  - Idea: Apply over a lexical concept network (inferred from a plain text) to rank the word senses
  - Assumption: high-ranked meanings are “recommendations” by related meanings, and preferred recommendations are made by most influential meanings
  - Shown to improve knowledge-based WSD [Mihalcea et al., 2004; Agirre & Soroa, 2008, 2009]

- Basic PageRank formula

$$PR(i) = \alpha \sum_{j \in B_i} \frac{PR(j)}{out(j)} + \frac{1-\alpha}{N}$$

Multi-dimensional representation

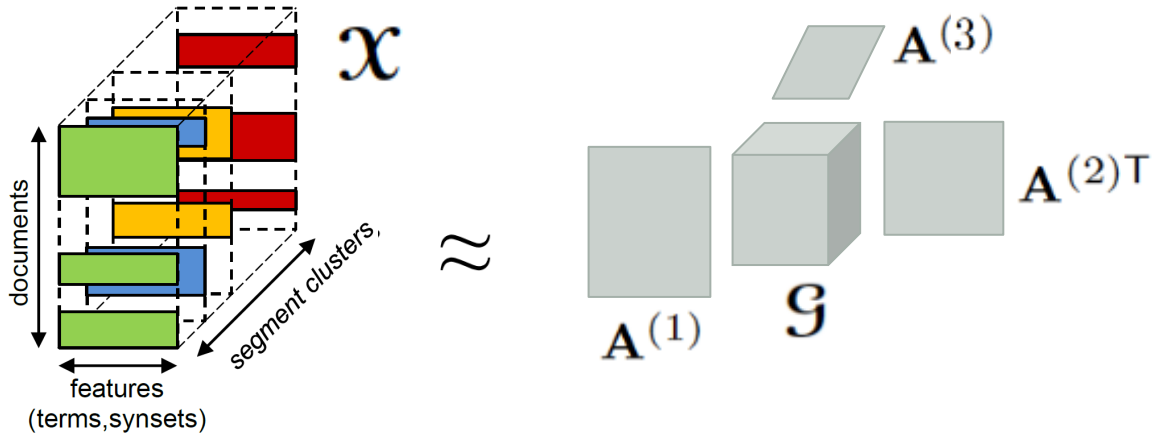


- Dimensions:
  - Mode-1: documents
  - Mode-2: features (of each segment cluster)
  - Mode-3: segment clusters
- Each segment cluster can be seen as a view of the document collection
- The document collection is described with a “non-flat” representation
- Tensor decompositions allow for the extraction of meaningful hidden information about the document collection

Tensor Decomposition

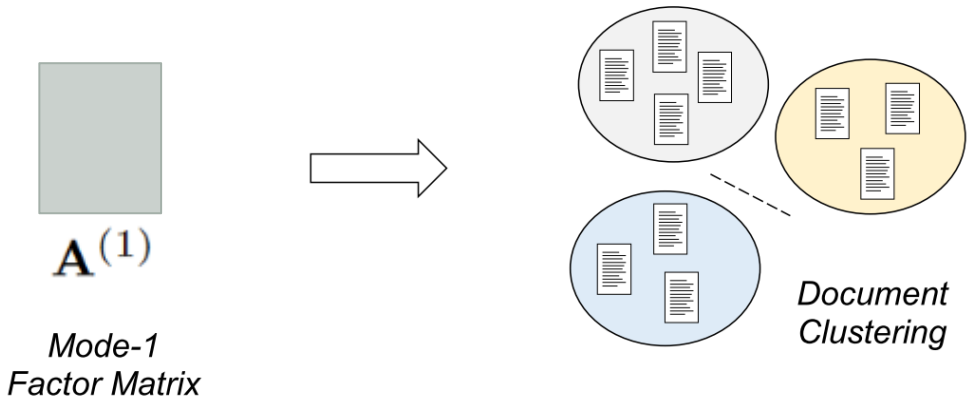
procedure HOSVD($\mathcal{X}, R_1, R_2, \ldots, R_N$)
  for $n = 1, \ldots, N$ **do**
    $\mathbf{A}^{(n)} \leftarrow R_n$ leading left singular vectors of $\mathbf{X}_{(n)}$
  end for
  $\mathcal{G} \leftarrow \mathcal{X} \times_1 \mathbf{A}^{(1)\mathsf{T}} \times_2 \mathbf{A}^{(2)\mathsf{T}} \cdots \times_N \mathbf{A}^{(N)\mathsf{T}}$
  return $\mathcal{G}, \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(N)}$

- The third-order tensor is decomposed into a core tensor and three factor matrices, one for each mode
  - Each mode is seen as one projection over the data via the tensor



Document clustering

- The mode-1 factor matrix is the input for a document clustering algorithm
- It's a low-dimensional representation of the documents
  - Embeds the view-oriented segment-clusters



SeMDocT algorithm

Algorithm 1 *SeMDocT* (**Se**gment-based **M**ultiLingual **Doc**ument Clustering via **T**ensor Modeling)

Input: A collection of multilingual documents $\mathcal{D}$, the number k of segment clusters, the number of tensorial components r.

Output: A document clustering solution $\mathcal{C}$ over $\mathcal{D}$.

1: Apply a text segmentation algorithm over each of the documents in $\mathcal{D}$ to produce a collection of document segments $\mathcal{S}$.
2: Represent $\mathcal{S}$ in either a bag-of-words (BoW) or a bag-of-synsets (BoS) space.
3: Apply any document clustering algorithm on $\mathcal{S}$ to obtain a segment clustering $\mathcal{C}^{\mathcal{S}} = \{C_i^s\}_{i=1}^k$.
4: Represent $\mathcal{C}^{\mathcal{S}}$ in either a bag-of-words (BoW) or a bag-of-synsets (BoS) space.
5: Model $\mathcal{S}$ as a third-order tensor $\boldsymbol{\mathcal{X}} \in \Re^{I_1 \times I_2 \times I_3}$, with $I_1 = |\mathcal{D}|$, $I_2 = |\mathcal{F}|$, and $I_3 = k$.
6: Decompose the tensor using a Truncated HOSVD.
7: Apply a document clustering algorithm on the mode-1 factor matrix to obtain the final clusters of documents $\mathcal{C} = \{C_i\}_{i=1}^K$.

Experimental evaluation

Data (1/2)

- Multilingual comparable corpus: RCV2
  - News articles in 13 languages
- Language selection:
  - English, French, and Italian
- Topic selection:
  - Conditioned to the document coverage in the various languages
- Balanced and unbalanced scenarios

RCV2 Topics	*English*	*French*	*Italian*
Balanced Corpus			
C15 - PERFORMANCE	850	850	850
C18 - OWNERSHIP CHANGES	850	850	850
E11 - ECONOMIC PERFORMANCE	850	850	850
E12 - MONETARY/ECONOMIC	850	850	850
M11 - EQUITY MARKETS	850	850	850
M13 - MONEY MARKETS	850	850	850
Total	5 100	5 100	5 100
Unbalanced Corpus			
C15 - PERFORMANCE	850	850	0
C18 - OWNERSHIP CHANGES	850	850	0
E11 - ECONOMIC PERFORMANCE	0	850	850
E12 - MONETARY/ECONOMIC	850	0	850
M11 - EQUITY MARKETS	0	850	850
M13 - MONEY MARKETS	850	0	850
Total	3 400	3 400	3 400

Statistics	*Balanced Corpus*	*Unbalanced Corpus*
# of docs	15 300	10 200
# of terms	58 825	44 535
# of synsets	16 395	14 339
BoW Density	1.5E-3	2.0E-3
BoS Density	2.6E-3	3.1E-3

Experimental evaluation

Data (2/2)

- Generally, more (resp. less) segments from English (resp. Italian) documents

RCV2 Topics	*English*	*French*	*Italian*
C15 - PERFORMANCE	3.41	3.67	3.27
C18 - OWNERSHIP CHANGES	3.20	3.32	2.40
E11 - ECONOMIC PERFORMANCE	4.89	3.17	2.07
E12 - MONETARY/ECONOMIC	5.22	3.69	2.05
M11 - EQUITY MARKETS	4.29	2.94	2.15
M13 - MONEY MARKETS	3.31	3.12	2.10

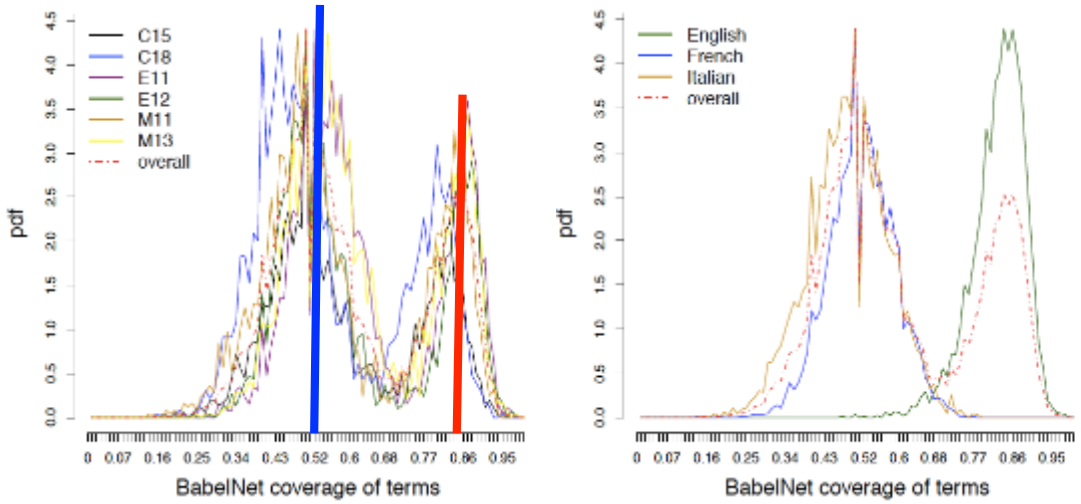
- BoS-modeled segments smaller than in the BoW space
  - BoS/BoW segment length ratio:
    - 2/3 on English, 1/4 on French, 1/3 on Italian

	English		*French*		*Italian*	
RCV2 Topics	avg *BoS* seg. leng.	avg *BoW* seg. leng.	avg *BoS* seg. leng.	avg *BoW* seg. leng.	avg *BoS* seg. leng.	avg *BoW* seg. leng.
C15	21.76	36.32	11.54	34.92	10.58	37.75
C18	20.94	36.87	10.94	35.62	11.24	41.20
E11	22.90	37.24	11.47	34.73	11.96	38.60
E12	22.70	37.70	11.50	37.44	12.59	43.63
M11	22.04	36.83	10.91	32.76	11.57	42.39
M13	22.22	36.97	11.34	34.75	11.72	39.36

Experimental evaluation

BabelNet coverage

- Analysis of the distribution of documents over different values of BabelNet coverage
  - i.e., fraction of words belonging to the document whose concepts are present as entries in BabelNet
- Per-topic distributions (left), per-language distribution (right)



➔ BabelNet provides a more complete coverage for English documents

Experimental evaluation

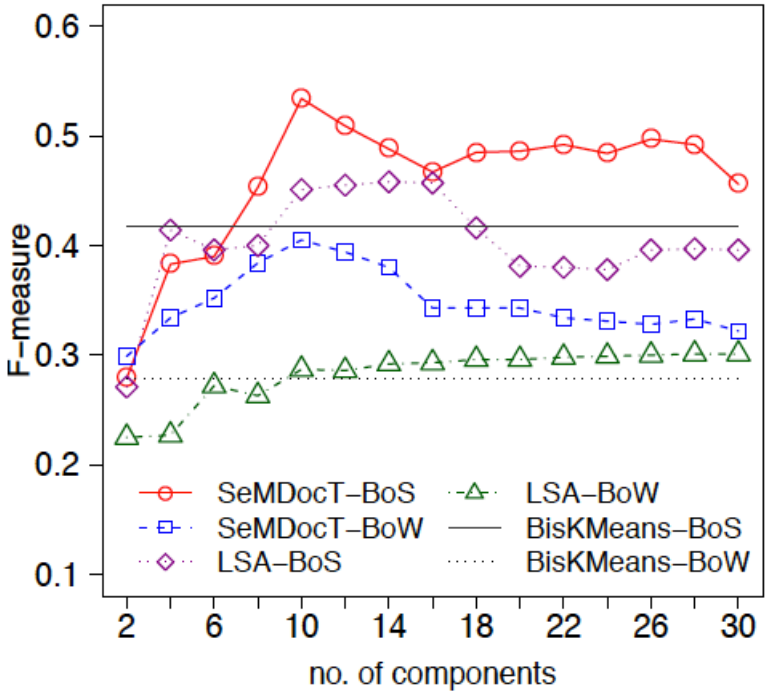
Methods and settings

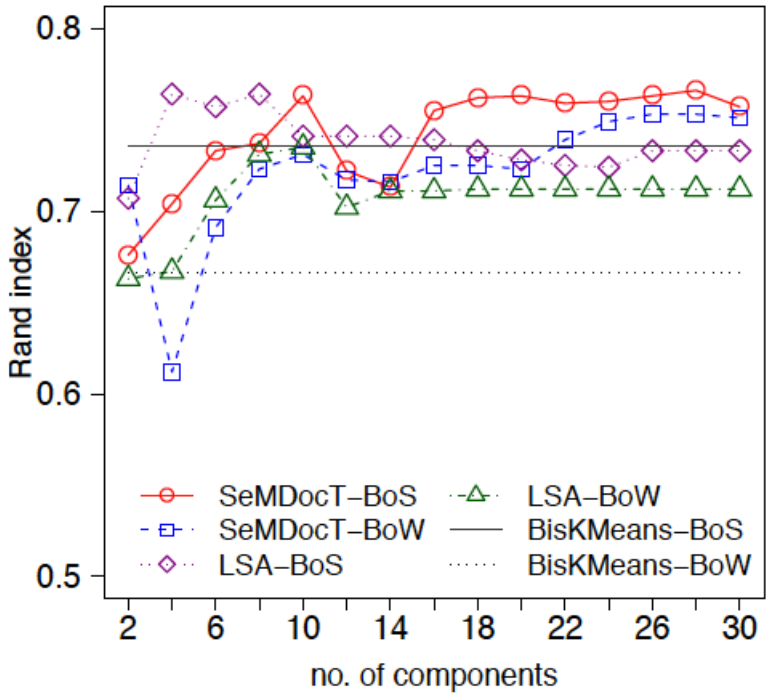
- Competing methods (over BoW or BoS space):
  - Bisecting k-Means
  - LSA based document clustering
    - i.e., Bisecting k-Means upon SVD representation of the collection
- Number of components (for SeMDocT and LSA)
  - From 2 to 30, with increment step 2
- Number of segment clusters (for SeMDocT)
  - Evaluation of within-cluster cohesion change by varying *k* (from 2 to 50)
  - Balanced corpus: 22 (BoS), 25 (BoW)
  - Unbalanced corpus: 23 (BoS), 11 (BoW)

Experimental evaluation

Evaluation on Balanced corpus

- BoS is beneficial for all document clustering approaches
- SeMDocT outperforms Bisecting k-Means and LSA-DocClust with #components ≥ 10 (FM, on average for RI)

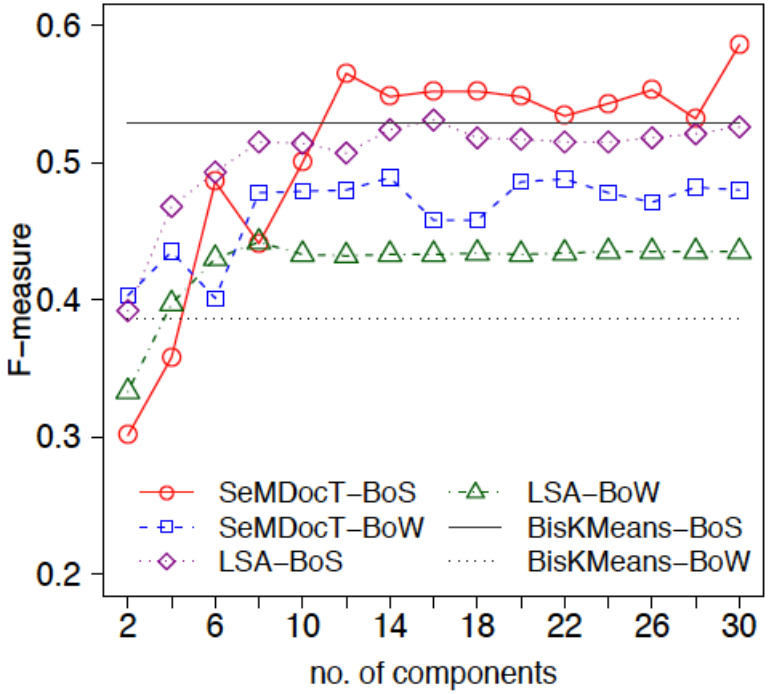


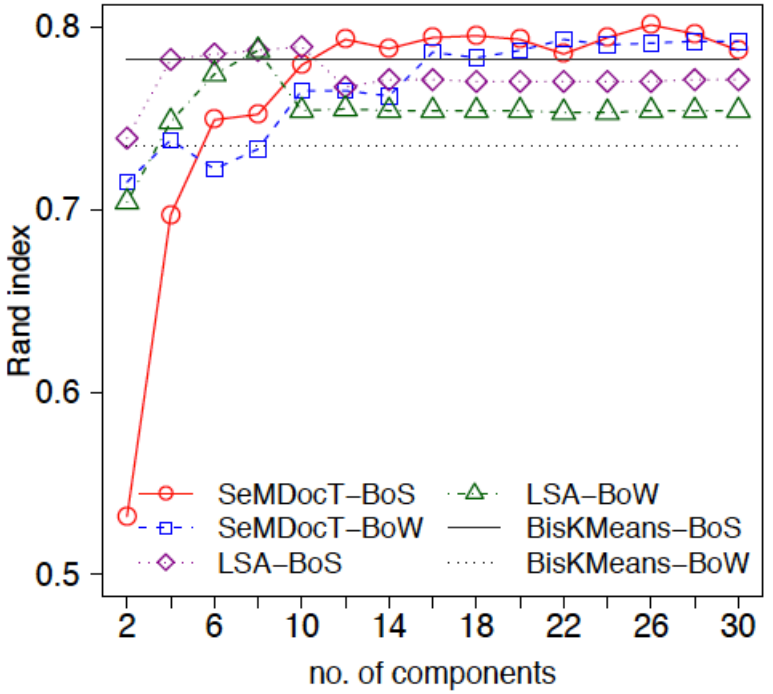


Experimental evaluation

Evaluation on Unbalanced corpus

- Again, BoS increases document clustering performance
- SeMDocT outperforms Bisecting k-Means and LSA-DocClust with #components ≥ 12 (FM, on average for RI)

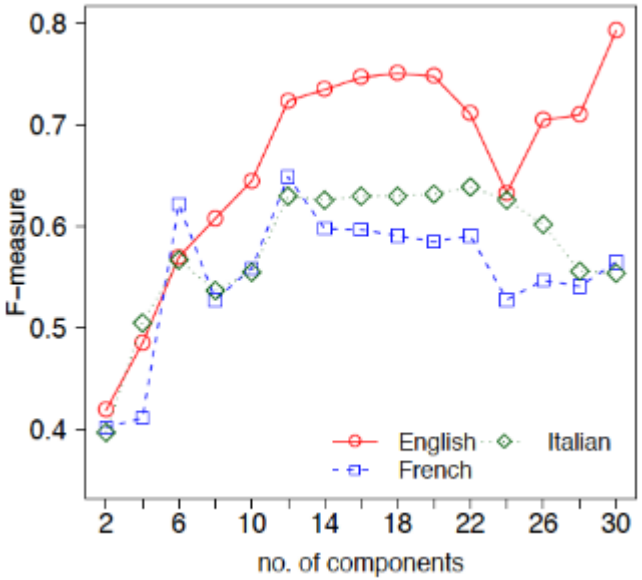


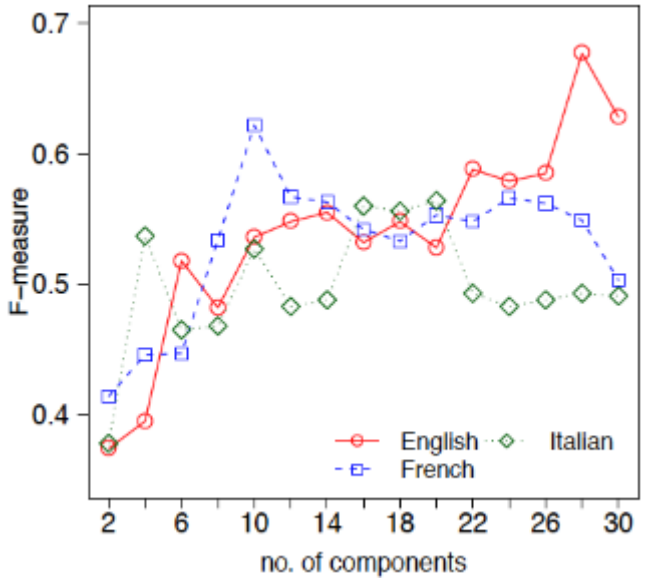


Experimental evaluation

Per language evaluation of SeMDocT-BoS

- Language-specific projections of clustering solutions





- Unbalanced case (left) vs. Balanced case (right)
  - higher performance in general
  - clearer evidence of better behavior for English documents

… needs explanation

Experimental evaluation

Per language evaluation of SeMDocT-BoS

- Focus on the avg #synsets per lemma
  - Always below 1
  - Higher for English than for French and Italian
  - Difference more evident in the Unbalanced case

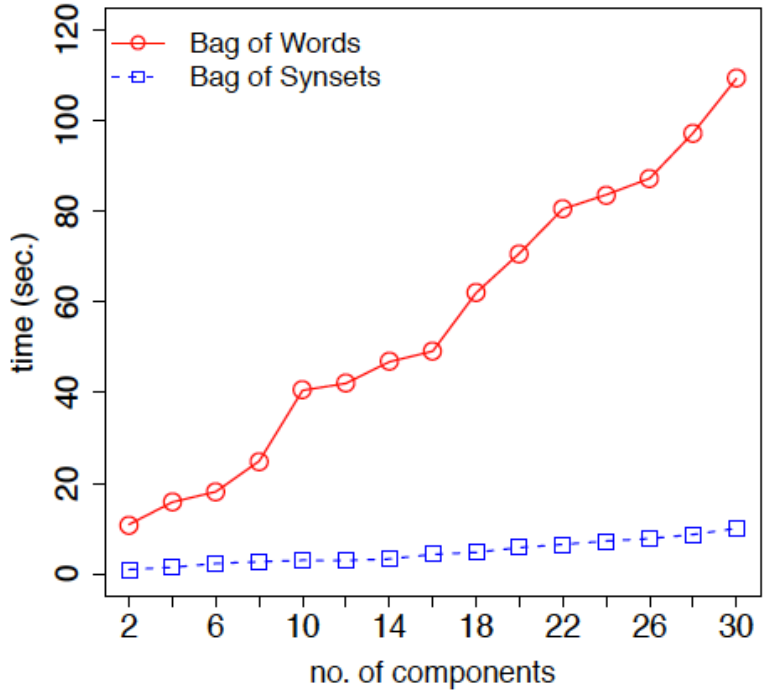
Dataset	*Language*	*BoW size*	*BoS size*	*avg # synsets per term* (β)
Balanced	English	29 999	12 065	0.4021
	French	17 826	5 310	0.2978
	Italian	16 951	4 471	0.2637
Unbalanced	English	19 432	10 387	0.5345
	French	14 439	4 431	0.3068
	Italian	14 743	4 012	0.2721

- SeMDocT performance improves with BabelNet coverage ability

Experimental evaluation

Runtime of tensor decomposition

- Execution time of SVD over the mode-1 matricization (Balanced corpus)
- BoS scales linearly with the no. of components,
- and better than BoW
  - thanks to higher dimensionality reduction



Summary of results

- SeMDocT: first MDC framework that integrates multidimensional, multi-topic-aware data structure with multilingual knowledge base

- SeMDocT requires a higher number of components than LSA-DocClust…
- …but ends with outperforming it (and conventional Bisecting k-Means) using few (i.e., 10-20) components
- Semantic coverage by BabelNet impacts on the SeMDocT performance
- SeMDocT scales linearly with the no. of components, and faster when using BoS

Future work

- BabelNet
  - Integrate more types of information (i.e., relations between synsets) to define richer multilingual document models
- Tensor modeling
  - Regularization of factor matrices and core tensor
  - Heuristics for the selection of number of components
  - Weighting of the components by means of Frobenius norm of core tensor slices
- Applications:
  - Multilingual Question Answering
  - Sentiment Analysis
  - Network analysis
    - Relation prediction
    - Topic and user popularity evolution
    - (SN) user language recognition