A Comparison of Distributional Semantics Models for Polylingual Text Classication

Andrea Esuli[1]
Alejandro Moreo Fernández[1]
Fabrizio Sebastiani[2]

[1] Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche, 56124 Pisa, IT
andrea.esuli@isti.cnr.it
alejandro.moreo@isti.cnr.it

[2] Qatar Computing Research Institute
Qatar Foundation, PO Box 5825, Doha, QA
fsebastiani@qf.org.qa

May 2015

Outline

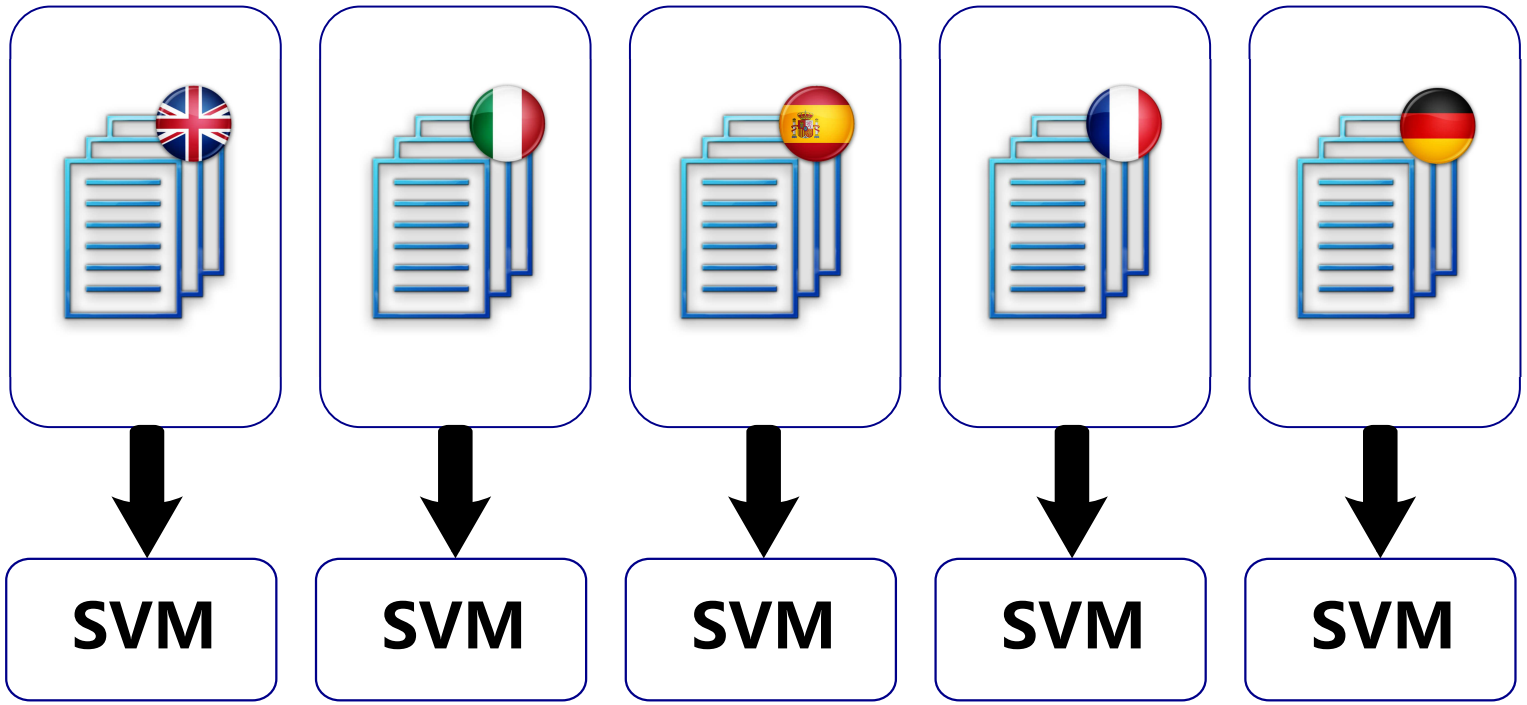
Polylingual Text Classification

Definition:

Polylingual Text Classification (PLTC) is a *supervised learning* task that consists of assigning class labels to documents belonging to *different languages*, assuming a representative set of training documents is available for each language.

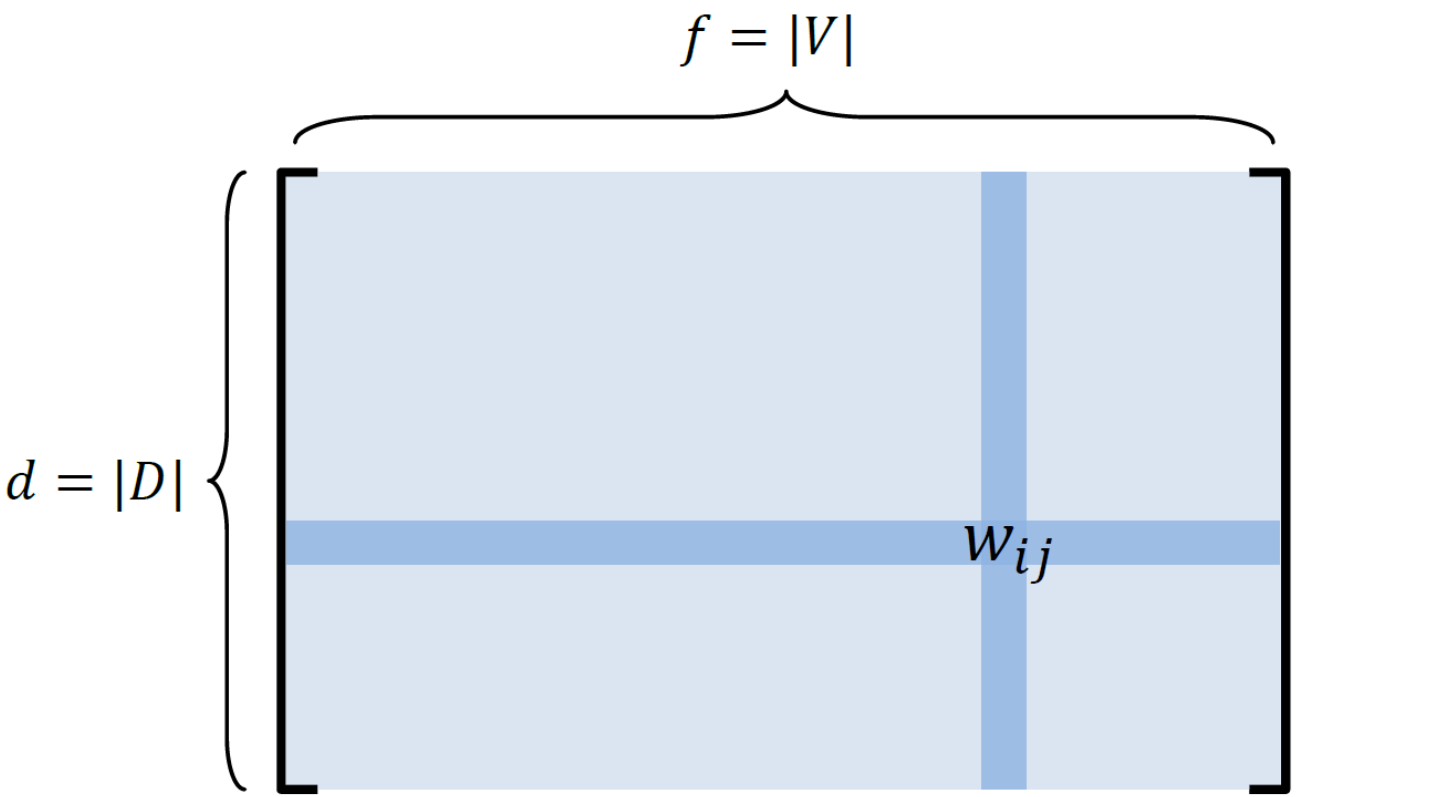
Example: News stories in international news media are written by different journalists in different languages, but likely referring to the same events.

The Naïve Approach



Main Difficulties

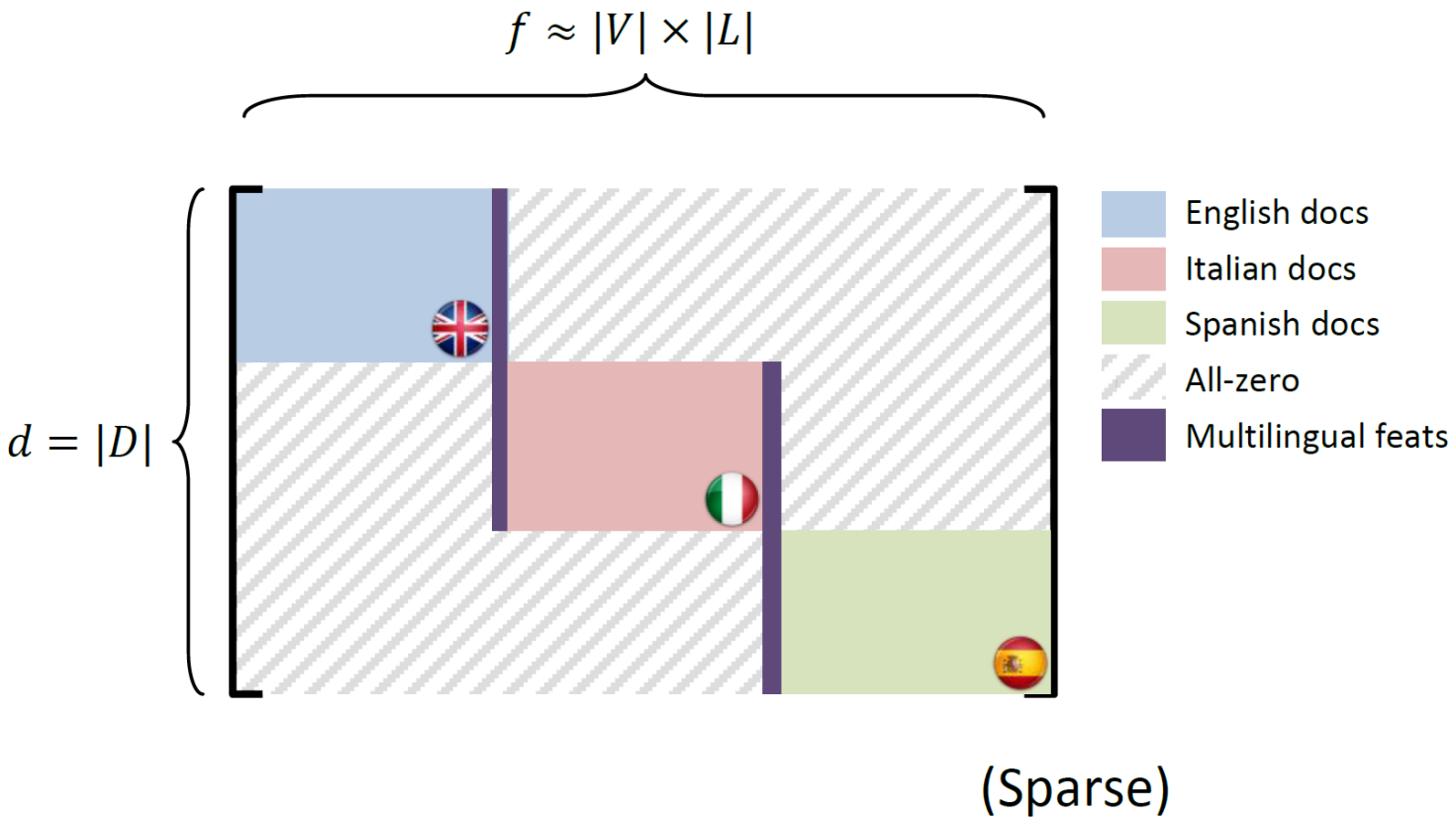
In the monolingual case



(Sparse)

Main Difficulties

In the polylingual case



Possible Solutions

Machine Translation

Why not simply translate all documents to one language?

- Automatically translated texts present different statistical properties.
- MT tools or dictionaries are not always available/public for all language pairs.
- MT implies considerable computational cost.

We will restrict our attention to **MT-free**, **Dictionary-free** methods.

Possible Solutions

Statistical Analysis

Why not use statistical analysis techniques such as *Principal Component Analysis* or *Canonical Correlation Analysis* to discover cross-lingual correlations?

- Statistical analysis is typically **computationally expensive**.
- Usually require a suited **parallel corpora** to mine the correlations.

However, **Random Indexing** [Kanerva et al., 2000] is considered a lighter approximation that could effectively reduce the dimensionality of the matrix!

Outline

Random Indexing

Theoretical foundations

Johnson-Lindenstrauss lemma:

"Distances in an Euclidean space are approximately preserved if projected into a lower dimensional random space."
[Johnson et al., 1986]

Hecht-Nielsen lemma:

"There are many more nearly-orthogonal directions than truly orthogonal directions in high-dimensional spaces."
[Hecht-Nielsen, 1994]

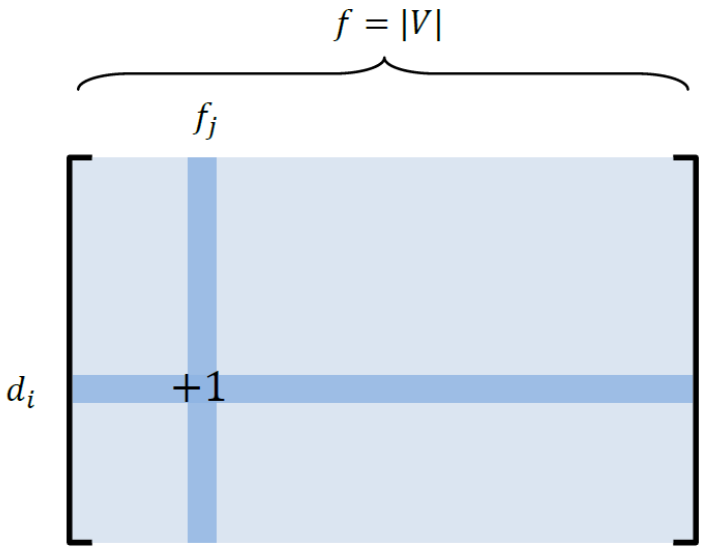
Achlioptas sufficient conditions:

"Any zero-mean with unit-variance distribution satisfies the lemma."
[Achlioptas, 2001]

Random Indexing

Vector cumulation (i)

Bag-of-Words as vector cumulation:



Each time $f_j \in d_i$ we add $\overrightarrow{d_i} \leftarrow \overrightarrow{d_i} + \overrightarrow{f_j}$

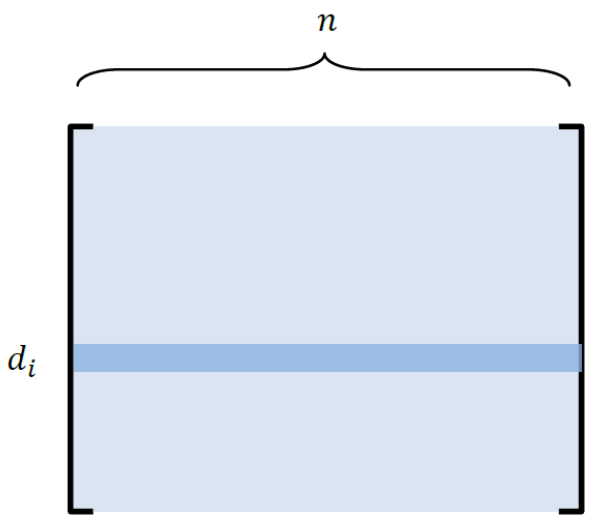
Where $\overrightarrow{f_j}$ is a "one-hot" orthogonal vector

$\overrightarrow{f_j} = (0, 0, 0, 0, \mathbf{1}, 0, 0, 0, \ldots, 0) \in \mathbb{R}^{|V|}$

Random Indexing

Vector cumulation (ii)

Random Indexing vector cumulation:



Each time $f_j \in d_i$ we add $\overrightarrow{d_i} \leftarrow \overrightarrow{d_i} + \overrightarrow{f_j}$

Where $\overrightarrow{f_j}$ is a nearly-orthogonal random vector

$\overrightarrow{f_j} = (0, +1, 0, 0, -1, 0, -1, 0, 0, 0, +1, 0, 0, \ldots) \in \mathbb{R}^n$

With k non-zero values evenly distributed between {+1, -1}

Lightweight Random Indexing

Parameters

Random Indexing depends on two parameters:

- n the *latent* dimensionality (typically 5,000 or 10,000)
- k the number of *non-zeros* values (typically $k = 1\%n$)

But, what is the rationale behind n and k?

- To increase the probability that any two feature-vectors are **orthogonal**.
- To be able to compactly **codify** many different features.

Random Indexing

Random Vectors

- Any two random vectors are likely orthogonal (dot product):

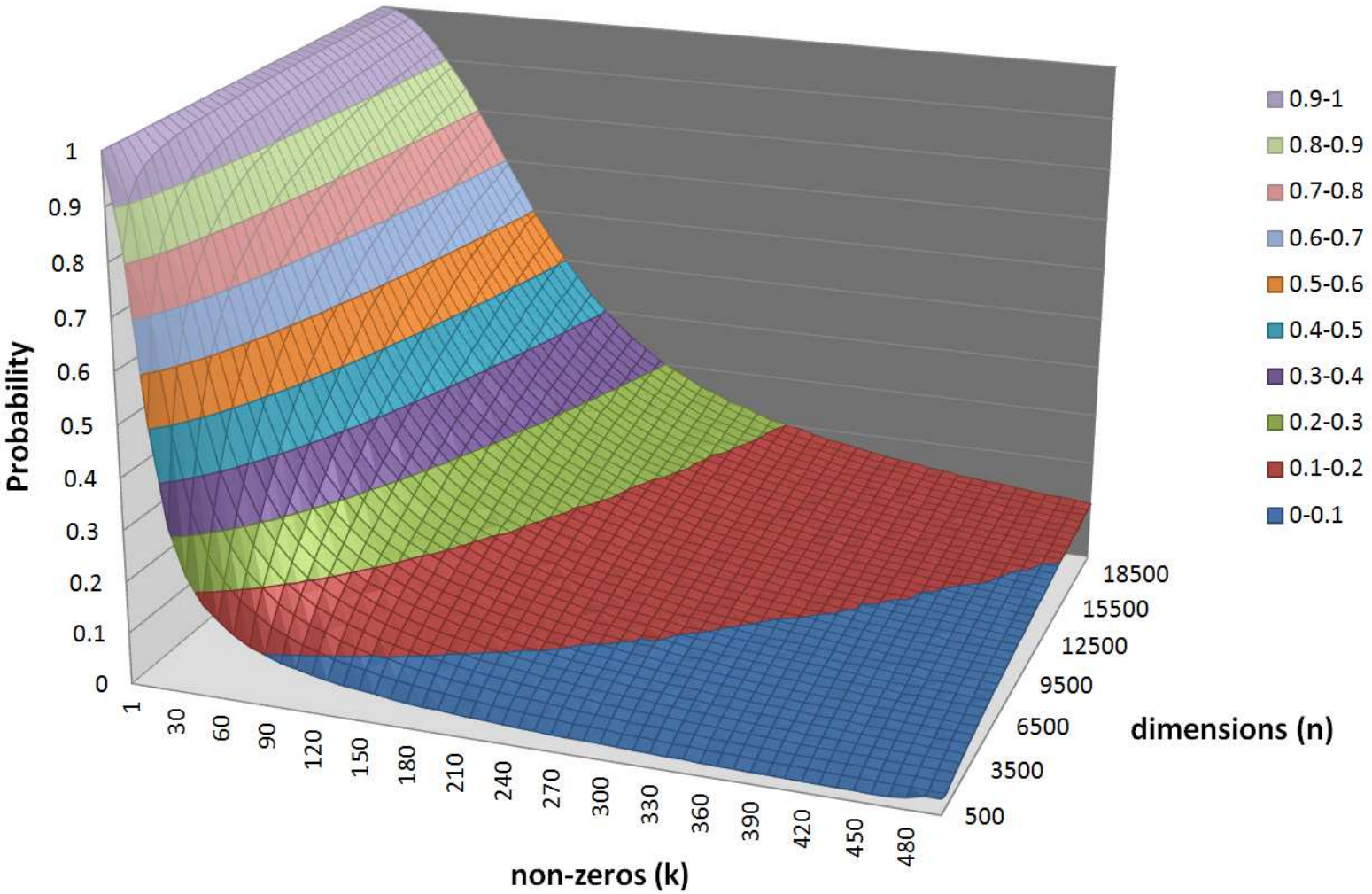
$$\overrightarrow{f_i} = (0, 0, 0, 0, +1, 0, -1, 0, 0, +1, 0, 0, -1, ...)$$
$$\overrightarrow{f_j} = (0, -1, 0, 0, 0, 0, +1, +1, 0, 0, -1, 0, 0, ...)$$

$$\langle \overrightarrow{f_i}, \overrightarrow{f_j} \rangle = \sum_{0 \leq x \leq n} f_i^x \cdot f_j^x \approx 0$$

- But, how does $P(\langle \overrightarrow{f_i}, \overrightarrow{f_j} \rangle = 0)$ depend on k and n?

Lightweight Random Indexing

Our Observation (i)



Lightweight Random Indexing

Our Observation (ii)

With (n, k) it is possible to codify $C(n, k) = \binom{n}{k} 2^k$ distinct features. E.g.,

$$C(5000, 50) \approx 2.5 \cdot 10^{135}$$

But even with small values of k we could codify many distinct features!

$$C(5000, 2) = 49,990,000$$

Lightweight Random Indexing

Our Hypothesis

Random Indexing for PLTC

RI as a representation mechanism for PLTC could help reduce the feature space, while the new latent features become potentially informative for all languages.

Lightweight Random Indexing: Setting $k = 2$ could be advantageous since:

- It increases the orthogonality of the random space.
- It suffices to represent large feature vocabularies.
- Each index vector has only two non-zero entries (efficiency).

Outline

Experiments

Comparison Methods

We confronted Lightweight Random Indexing (LRI) with the following baselines:

- The Naïve Approach (MonoBoW [García Adeva et al., 2005])
- Polylingual BoW matrix (PolyBoW [García Adeva et al., 2005])
- Random Indexing with $k = 1\%n$ (RI [Sahlgren and Cöster, 2004])
- Latent Semantic Analysis (LSA [Dumais et al., 1997])
- Achlioptas distribution (Ach [Achlioptas, 2001])
- Multilingual Domain Models (MDM [Gliozzo and Strapparava, 2005])

We used *SVM-light* with standard parameters as the classifier in all cases.

Experiments

Evaluation Measure

As the effectiveness measure we use the well-known F_1 measure, the harmonic mean of precision (π) and recall (ρ).

$$F_1 = \frac{2\pi\rho}{\pi + \rho} = \frac{2TP}{2TP + FP + FN}$$

We compute both micro-averaged F_1 (F_1^{μ}) and macro-averaged F_1 (F_1^{M}).

Experiments

Dataset

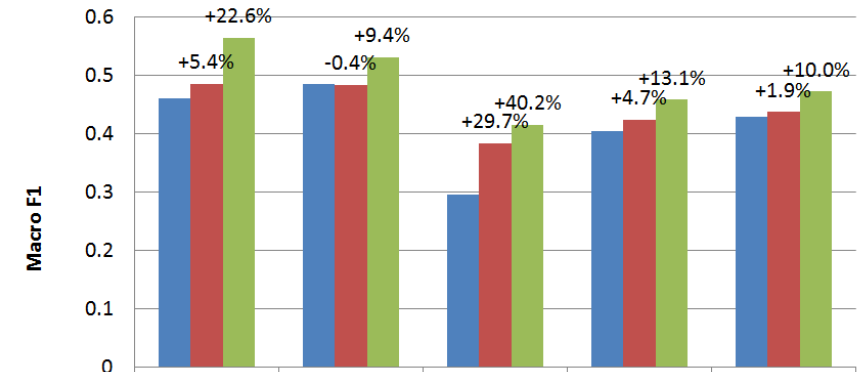
RCV1/RCV2 is a *comparable*, publicly available, Reuters collection of news stories in fourteen different languages.

languages:	English, Italian, Spanish, French, German.
#features:	123,258 distinct stems.
#categories:	67/103 categories with at least 1 positive doc in all languages.
#documents:	40,000 documents (8,000 per language).
split:	70%/30% (28,000 training / 12,000 test documents).

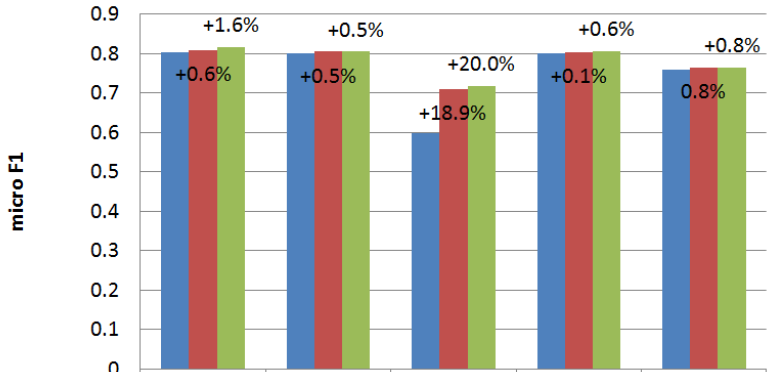
We preprocessed the corpus by removing stop words and by stemming terms using the Snowball Stemmers.

Results

Improving the performance of Monolingual Classifiers



	English	Italian	Spanish	French	German
MonoBoW	0.46	0.485	0.296	0.405	0.429
PolyBoW	0.485	0.483	0.384	0.424	0.437
LRI	0.564	0.531	0.415	0.458	0.472



	English	Italian	Spanish	French	German
MonoBoW	0.803	0.801	0.598	0.802	0.76
PolyBoW	0.808	0.805	0.711	0.803	0.766
LRI	0.816	0.805	0.718	0.807	0.766

Results

Dimensionality reduction experiments on RCV2/RCV1

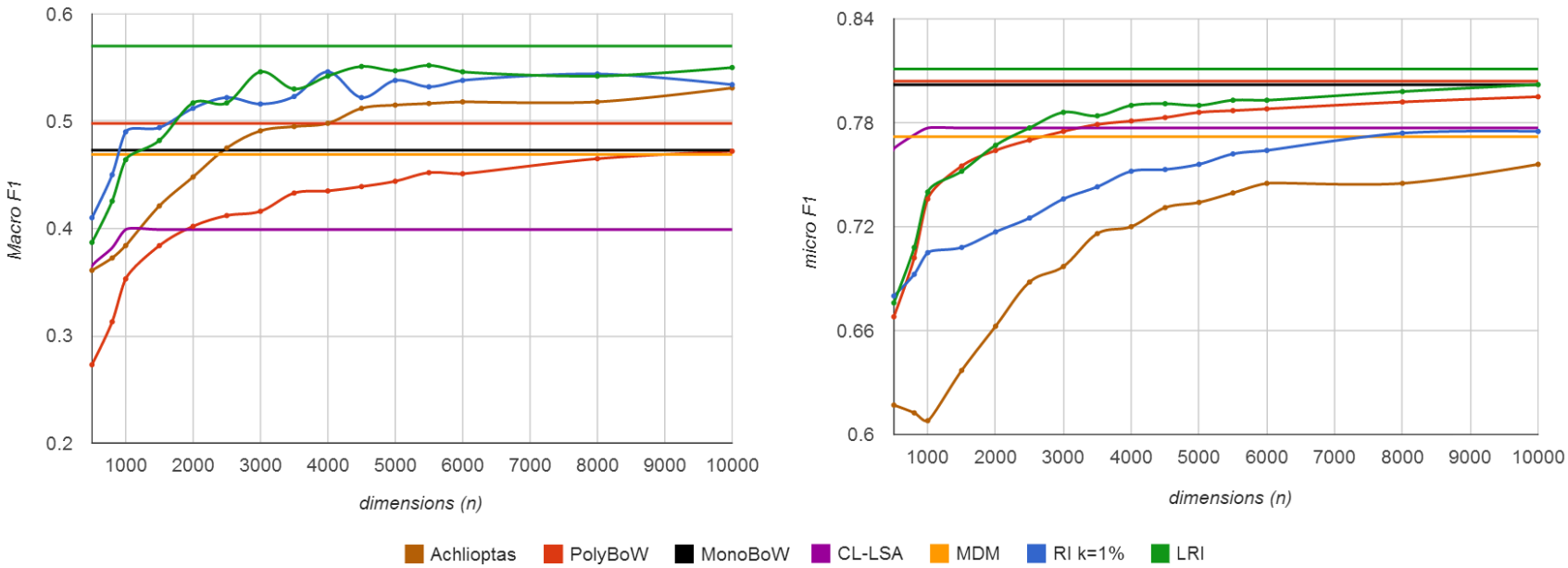


Figure: Dimensionality reduction on RCV2/RCV1, only **English and Italian**.

Results

Dimensionality reduction experiments on RCV2/RCV1

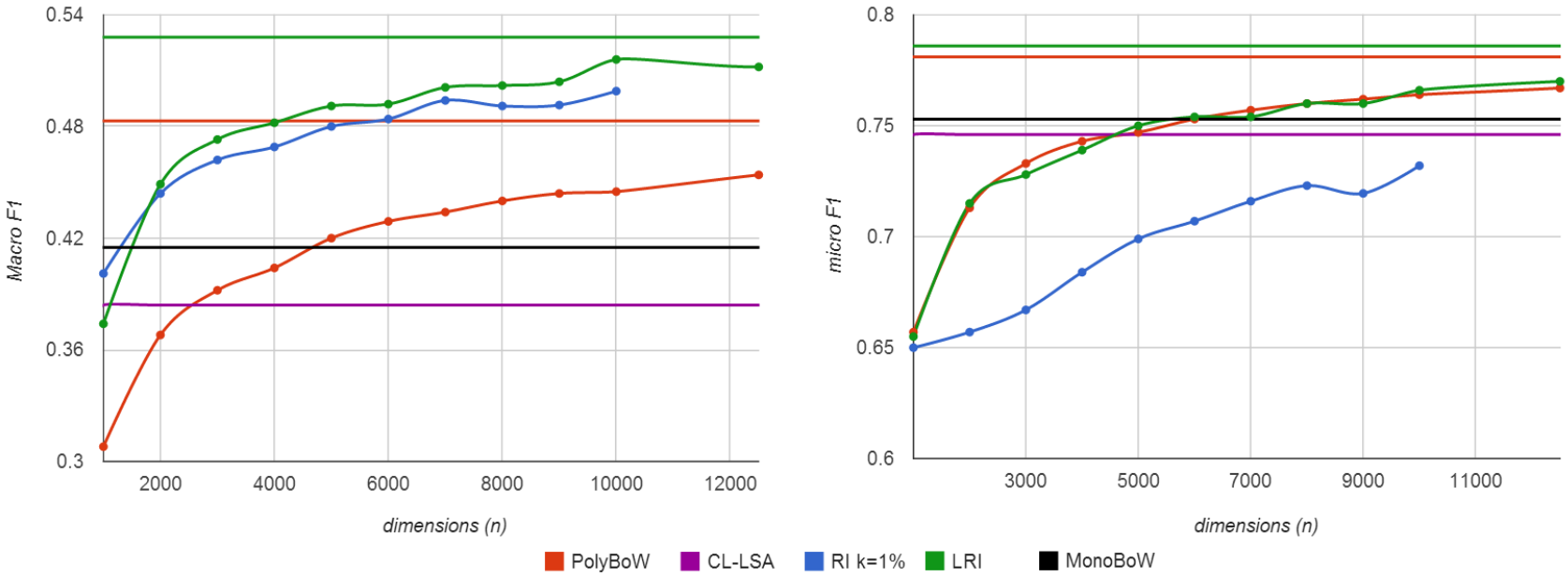
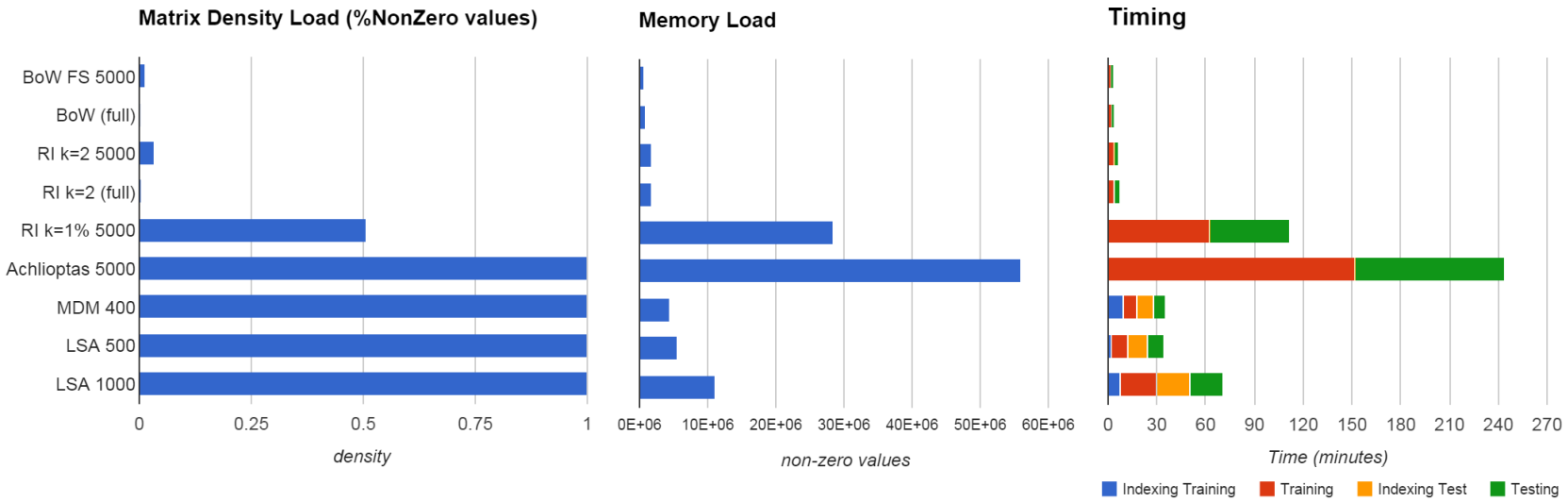


Figure: Dimensionality reduction on RCV2/RCV1, **five languages**.

Outline

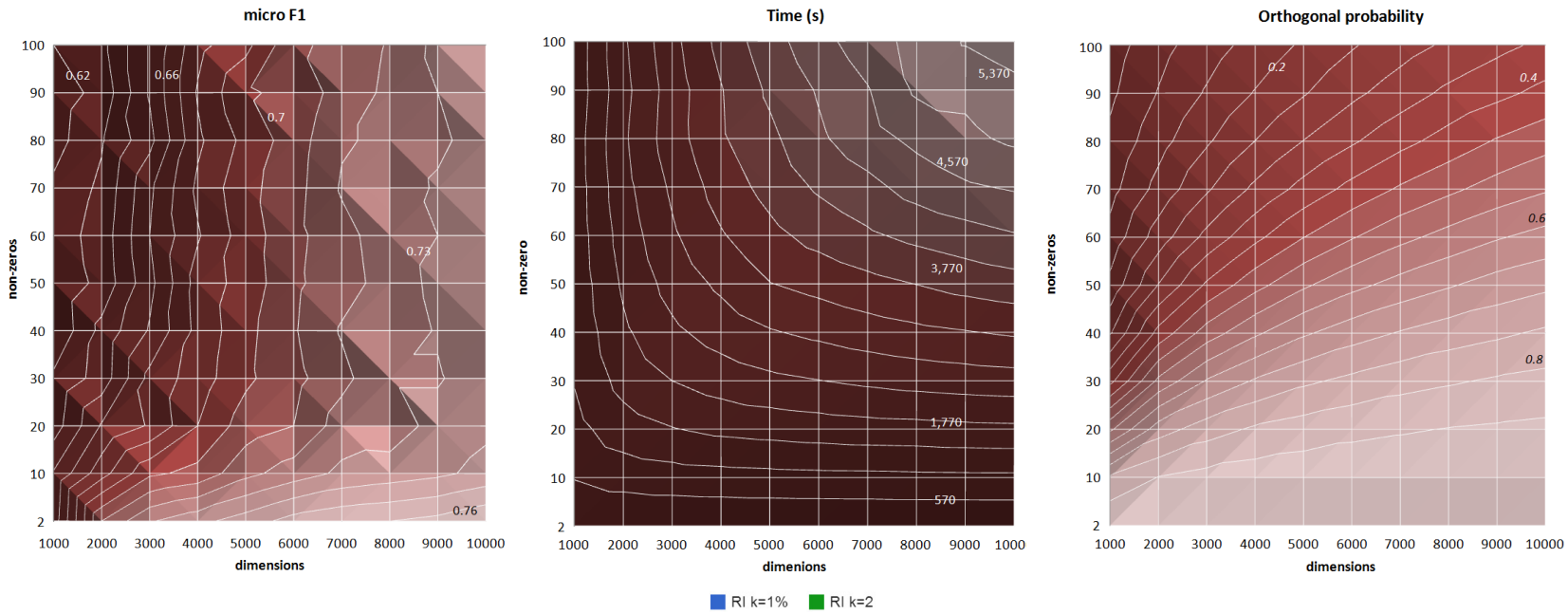
Analysis

Sparseness and Efficiency



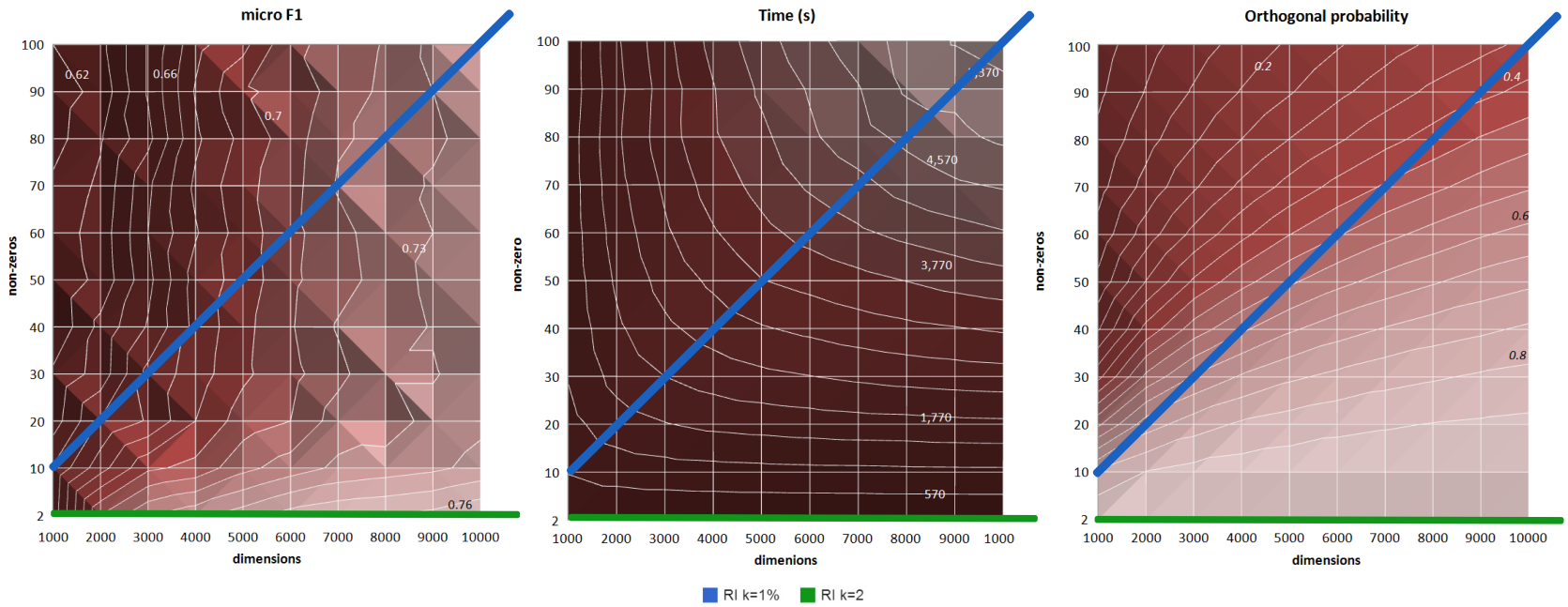
Analysis

Orthogonality and Efficacy



Analysis

Orthogonality and Efficacy



Conclusions

We have proposed a modification of Random Indexing for Polylingual Text Classification:

- **Resources**: Machine translation-free, Dictionary-free method.
- **Efficacy**: All latent features become potentially useful for any language, and the random space is "more orthogonal".
- **Efficiency**: Preserves sparsity, memory load and training times are not penalized.

Bibliography I

Achlioptas, D. (2001).

Database-friendly random projections.

In *Proceedings of the 20th ACM Symposium on Principles of Database Systems (PODS 2001)*, pages 274–281, Santa Barbara, US.

Dumais, S. T., Letsche, T. A., Littman, M. L., and Landauer, T. K. (1997).

Automatic cross-language retrieval using latent semantic indexing.

In *Working Notes of the AAAI Spring Symposium on Cross-language Text and Speech Retrieval*, pages 18–24, Stanford, US.

García Adeva, J. J., Calvo, R. A., and López de Ipińa, D. (2005).

Multilingual approaches to text categorisation.

European Journal for the Informatics Professional, 5(3):43–51.

Gliozzo, A. and Strapparava, C. (2005).

Cross-language text categorization by acquiring multilingual domain models from comparable corpora.

In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 9–16, Ann Arbor, US.

Bibliography II

Hecht-Nielsen, R. (1994).

Context vectors: General-purpose approximate meaning representations self-organized from raw data.

In Zurada, J., Marks, R., and Robinson, C., editors, *Computational Intelligence: Imitating Life*, pages 43–56. IEEE Press.

Johnson, W. B., Lindenstrauss, J., and Schechtman, G. (1986).

Extensions of Lipschitz maps into Banach spaces.

Israel Journal of Mathematics, 54(2):129–138.

Kanerva, P., Kristofersson, J., and Holst, A. (2000).

Random indexing of text samples for latent semantic analysis.

In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 1036–1037, Philadelphia, US.

Sahlgren, M. and Cöster, R. (2004).

Using bag-of-concepts to improve the performance of support vector machines in text categorization.

In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, CH.

Thank you!

Questions?