Cross-language text classification

Text Analytics - Andrea Esuli

One world, many languages

English is the modern *lingua franca* of scientific communication, and a dominant language on the Web and in global communication in general.

English is also the most common test bed for NLP/IR/ML research.

Pros:

- **focus** on one of the most used language in the digital world.
- **many** shared resources (lexica, datasets).
- **many** shared tools, enabling the test new methods focusing only on the delta part.

One world, many languages

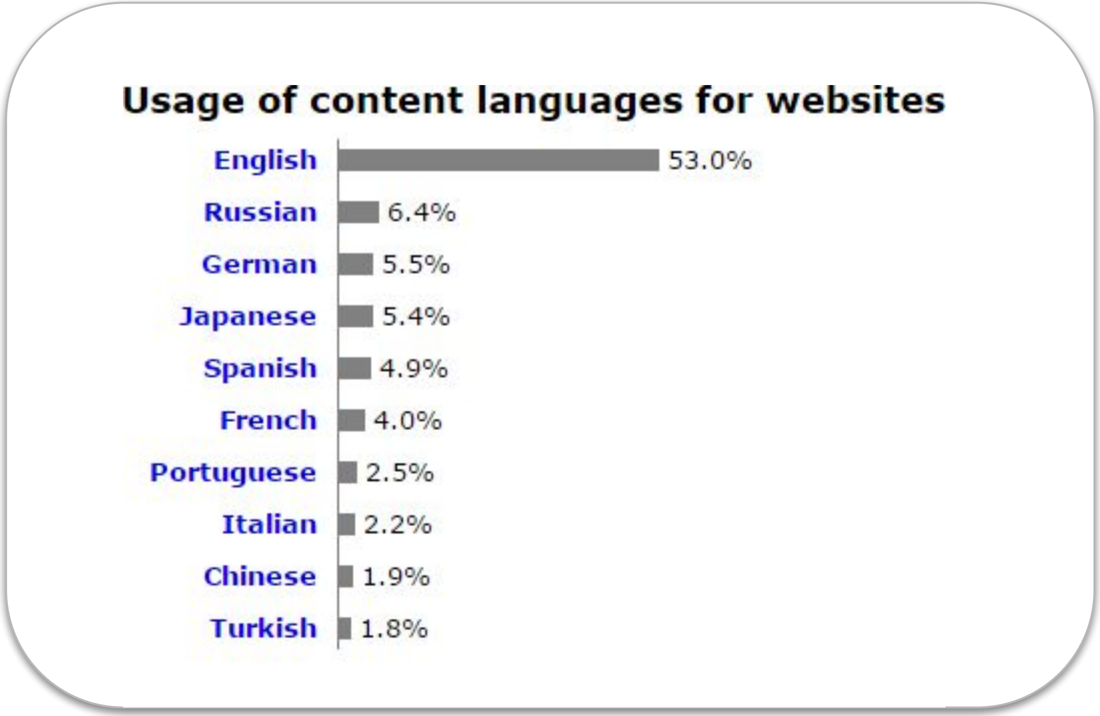
English is the modern *lingua franca* of scientific communication, and a dominant language on the Web and in global communication in general.

English is also the most common test bed for NLP/IR/ML research.

Cons:

- **less** research on many languages used by a large part of world population.
- **less** research on language-specific aspects of NLP.
- **less** resources (lexica, datasets) on other languages.

One world, many languages



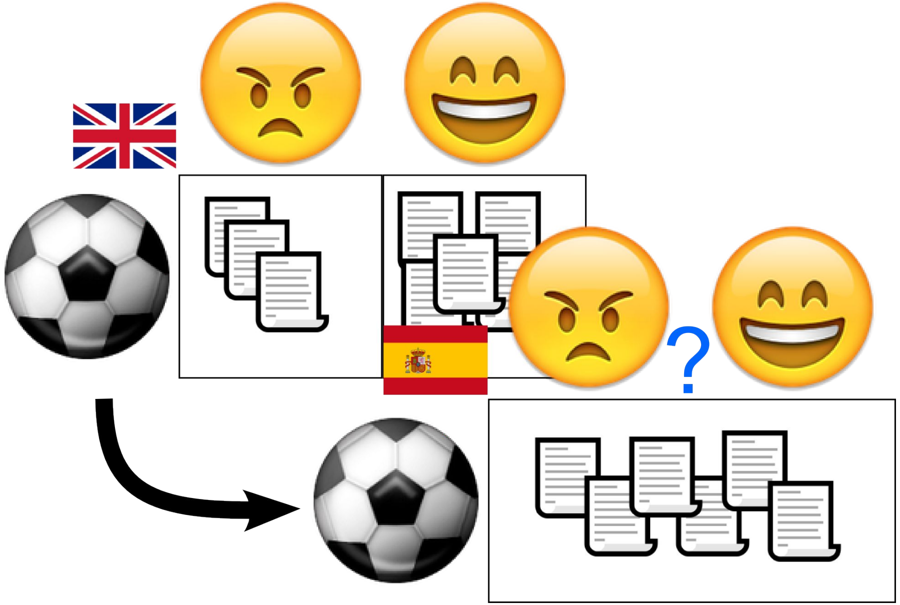
One world, many languages

Top Ten Languages Used in the Web - June 30, 2017 (Number of Internet Users by Language)					
TOP TEN LANGUAGES IN THE INTERNET	World Population for this Language (2017 Estimate)	Internet Users by Language	Internet Penetration (% Population)	Internet Users Growth (2000 - 2017)	Internet Users % of World Total (Participation)
English	1,434,937,438	**984,703,501**	68.6 %	599.6 %	25.3 %
Chinese	1,425,430,865	**770,797,306**	54.1 %	2,286.1 %	19.8 %
Spanish	510,380,423	**312,069,111**	61.1 %	1,616.4 %	8.0 %
Arabic	421,345,425	**184,631,496**	43.8 %	7,247.3 %	4.8 %
Portuguese	281,603,515	**158,399,082**	56.2 %	1,990.8 %	4.1 %
Indonesian / Malaysian	295,108,771	**157,580,091**	53.4 %	2,650.1 %	4.1 %
Japanese	126,045,211	**118,453,595**	94.0 %	151.6 %	3.0 %
Russian	143,375,006	**109,552,842**	76.4 %	3,434.0 %	2.8 %
French	405,644,599	**108,014,564**	26.6 %	800.2 %	2.8 %
German	94,943,848	**84,700,419**	89.2 %	207.8 %	2.2 %
TOP 10 LANGUAGES	5,138,815,101	**2,988,902,008**	58.2 %	907.2 %	76.9 %
Rest of the Languages	2,380,213,869	**896,665,611**	37.7 %	1,296.1 %	23.1 %
WORLD TOTAL	7,519,028,970	**3,885,567,619**	51.7 %	976.4 %	100.0 %

NOTES: (1) Top Ten Languages Internet Stats were updated in June 30 2017. (2) Internet Penetration is the ratio between the sum of Internet users speaking a language and the total population estimate that speaks that specific language. (3) The most recent Internet usage information comes from data published by Nielsen Online, International Telecommunications Union, GfK, and other reliable sources. (4) Population estimates are based mainly on figures from the United Nations Population Division and local official sources. (5) For definitions, methodology and navigation help, please see the Site Surfing Guide. (6) These statistics may be cited, stating the source and establishing an active link back to Internet World Stats.

Working across languages

Can we reuse labeled information for a **source** language on a different **target** language where such information is scarce or missing?



Cross-language learning

Cross-language learning methods are based on the idea of projecting documents into a common representation space.

Two possible approaches:

- Machine translation
  - straightforward, or using some tricks such as co-training
  - requires a good MT for the pair of languages involved...
  - ...and a good MT usually has a cost
- Vector space projection: a common vector space in which documents with similar content from the two languages end up in similar positions.
  - focused on the task, simpler than MT

Cross-language learning

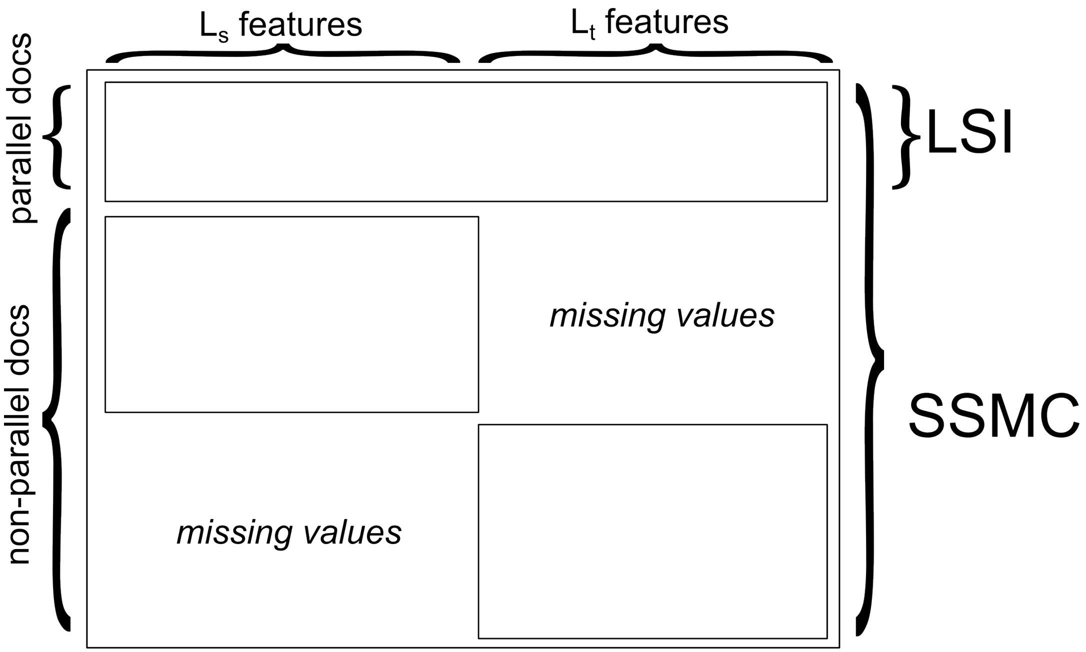
Vector space projection methods can be classified with respect to the type of data they need to build the projection:

- parallel corpora: documents with exactly the **same content** in both languages.
  - LSI (Dumais et al., 1997), Semi-Supervised Matrix Completion (SSMC)
  - parallel corpora are not easy to find (otherwise MT would be a cheap tool)
- comparable corpora: documents with **similar content** in the two languages
  - very easy to collect
  - methods may additionally require short lists of translated word pairs (e.g., "cat/gatto"), still much easier than doing full translation

LSI - SSMC

Latent Semantic Indexing use SVD to project features that have similar distributional properties across languages into the same positions of the projection space.

Semi-Supervised Matrix Completion extends this approach to include also documents that have no translation.



Structural Correspondence Learning

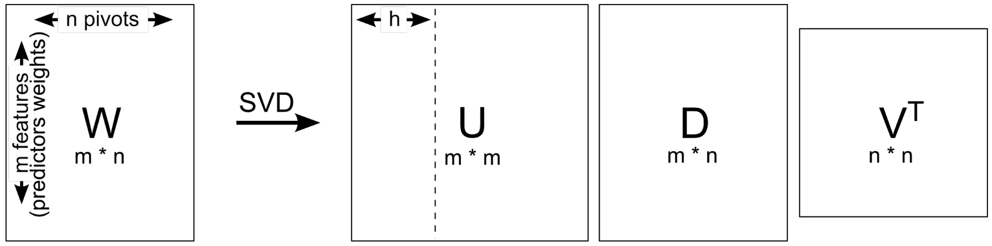
Structural Correspondence Learning leverages on a set of pivot terms:

`cat-gatto, run-correre, sun-sole, …`

to model the similarities of features from the two languages.

- Select n pivot features that are **frequent** in both domains and **informative** on labeled data (e.g., by their mutual information with labels)
- Build a **linear predictor** for each pivot based on non-pivot features.
- Group predictors weights in a matrix W and decompose it using SVD.

Structural Correspondence Learning



Pivots are typically in the order of hundreds

- There is a high cost to train all predictors and then doing the SVD

Can we use a simpler model of distributional similarity?

Distributional Correspondence Indexing

Distributional Correspondence Indexing steps:

- Select n pivot features that are informative on labeled data and **similarly frequent in both domains**.
- Represent any feature with a **profile vector** that measures the **distributional similarities** between the feature and **each of the pivots**, by using a **distributional correspondence function** (DCF).
    - DCF functions are **fast** to compute
- Use the n-dimensional profile vectors to index documents.
    - DCF values directly define the projection, **no matrix decomposition or additional modeling required**.

Distributional Correspondence Indexing

A Distributional Correspondence Function measures the correlation between a feature f_i and a pivot f_j (which is itself a feature) by comparing how they appear in a collection of documents.

DCFs can use a probabilistic model:

Probability-based DCFs	Mathematical form
Linear	$P(f^i \vert f^j) - P(f^i \vert \overline{f^j})$
Pointwise Mutual Information	$\log_2 \frac{P(f^i, f^j)}{P(f^i)P(f^j)}$
Asymmetric Mutual Information	$\rho(f^i, f^j) \sum_{x \in \{f^i, \overline{f^i}\}} \sum_{y \in \{f^j, \overline{f^j}\}} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$

Distributional Correspondence Indexing

...or a kernel-based model:

Kernel-based DCFs	Mathematical form
Cosine	$\dfrac{\langle \mathbf{f}^i, \mathbf{f}^j \rangle}{\|\mathbf{f}^i\| \, \|\mathbf{f}^j\|} - \sqrt{p_i p_j}$
Polynomial	$(a + \langle \mathbf{f}^i, \mathbf{f}^j \rangle)^b - (a + \sqrt{p_i p_j})^b$
RBF	$\exp\{-\gamma \|\mathbf{f}^i - \mathbf{f}^j\|^2\} - \exp\left\{-4\gamma \left(1 - \sqrt{p_i p_j}\right)^2\right\}$

Kernel-based DCFs have a normalization term, so that the expected value of $DCF(f_i, f_j)$ is zero for a uniform distribution of vectors with the same prevalence* p_i and p_j of f_i and f_j

*portion of values different from zero in f

Distributional Correspondence Indexing

A feature is represented as an n-dimensional vector of DCF value w.r.t. pivots (using the matching translation of the pivot).

$$e(f) = (DCF(f,p_1), DCF(f,p_2), \ldots, DCF(f,p_n))$$

Documents are directly indexed in the DCI space as a weighted sum of all profile vectors associated to their features:

$$e(d) = \sum_{f \in d} w_{fd}\, e(f)$$

where w_{fd} is the weight of feature f in document d according to a weighting function

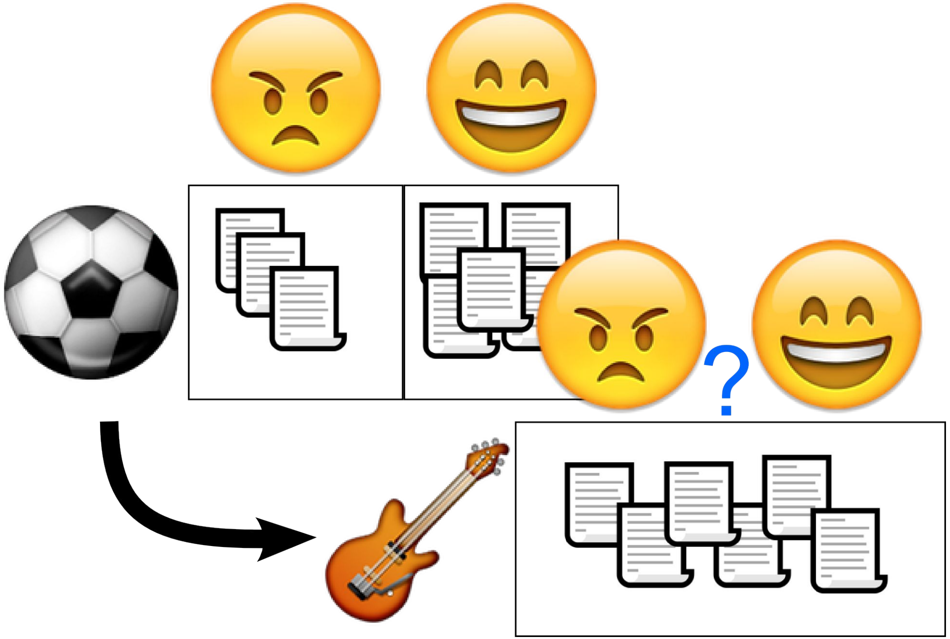
Cross-lingual classification

Results on the Webis-CLS-10 dataset

Task	Upper	MT	SCL-MI	LSI	SSMC	Linear	PMI	AMI	Cos	Poly	RBF
EB → GB	0.868	0.808	0.833	0.776	0.819	0.798	0.714	0.797	0.827	**0.837**	0.829
ED → GD	0.835	0.800	0.809	0.796	0.823	0.826	0.819	0.800	0.822	**0.833**	0.788
EM → GM	0.859	0.791	0.829	0.727	0.813	0.844	0.850	0.837	**0.856**	0.844	0.801
EB → FB	0.862	0.821	0.813	0.792	0.831	0.746	0.761	0.768	0.842	0.819	**0.844**
ED → FD	0.872	0.795	0.804	0.778	0.827	0.823	0.823	0.801	0.827	0.806	**0.846**
EM → FM	0.890	0.765	0.781	0.726	0.805	0.816	0.827	0.818	**0.844**	0.840	0.803
EB → JB	0.812	0.692	0.770	0.738	0.738	0.779	0.731	0.711	0.758	0.754	**0.782**
ED → JD	0.834	0.722	0.764	0.754	0.776	**0.822**	0.768	0.797	0.801	0.795	0.761
EM → JM	0.842	0.714	0.773	0.734	0.775	0.826	0.816	0.807	**0.839**	0.832	0.826
German	0.854	0.800	0.824	0.766	0.754	0.823	0.794	0.811	0.835	**0.838**	0.806
French	0.875	0.794	0.799	0.765	0.766	0.795	0.804	0.796	**0.838**	0.822	0.831
Japanese	0.829	0.709	0.769	0.742	0.770	**0.809**	0.772	0.772	0.799	0.794	0.790
Average	0.852	0.767	0.797	0.758	0.763	0.809	0.790	0.793	**0.824**	0.818	0.809

Sentiment across domains

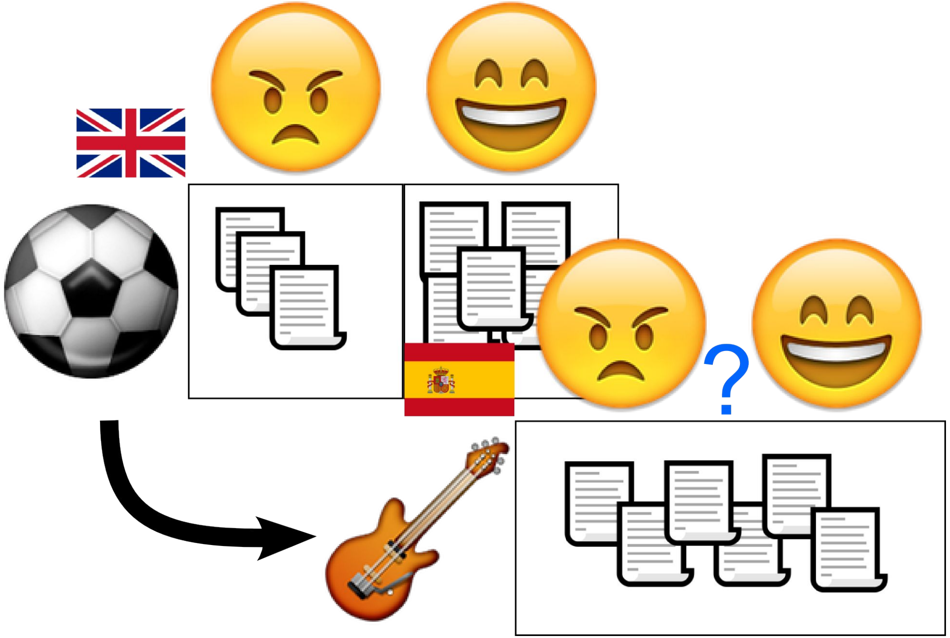
Sentiment has shared semantics across domain, can we exploit sentiment data on a topic to perform sentiment classification for a different one?



Sentiment across domains

Task	NoTrans	Upper	SCL-MI	Linear	PMI	AMI	Cos	Poly	RBF
ED → EB	0.803	0.829	0.839	0.840	0.843	0.831	0.851	**0.855**	0.848
EM → EB	0.783	0.829	0.823	0.828	0.838	0.826	0.840	**0.841**	0.838
EB → ED	0.798	0.831	0.810	0.798	0.812	0.788	**0.818**	**0.818**	0.806
EM → ED	0.778	0.831	0.797	0.802	**0.821**	0.798	0.821	**0.822**	0.816
EB → EM	0.786	0.845	0.804	0.825	0.835	0.816	**0.838**	0.836	0.831
ED → EM	0.804	0.845	0.823	0.831	**0.833**	0.815	0.829	0.832	0.827
Books	0.793	0.829	0.831	0.834	0.841	0.829	0.846	**0.848**	0.843
DVDs	0.788	0.831	0.804	0.800	0.817	0.793	0.819	**0.820**	0.811
Music	0.795	0.845	0.814	0.828	**0.834**	0.816	**0.834**	**0.834**	0.829
Average	0.792	0.835	0.816	0.821	0.830	0.812	0.833	**0.834**	0.828

Cross-language + cross-domain



Cross-language + cross-domain

Task	Upper	MT	SCL-MI	Linear	PMI	AMI	Cos	Poly	RBF
ED → GB	0.868	0.789	0.823	0.823	0.764	0.811	**0.824**	0.818	**0.824**
EM → GB	0.868	0.751	**0.825**	0.791	0.821	0.705	0.812	0.791	0.800
EB → GD	0.835	0.774	0.784	0.790	0.796	0.788	**0.827**	0.825	0.783
EM → GD	0.835	0.773	0.792	0.778	0.829	0.772	**0.834**	0.814	0.808
EB → GM	0.859	0.768	0.811	0.786	0.812	0.793	**0.843**	0.833	0.807
ED → GM	0.859	0.768	0.824	**0.844**	**0.844**	0.828	0.816	0.835	0.832
ED → FB	0.862	0.788	0.790	0.744	0.798	0.747	0.848	0.846	**0.852**
EM → FB	0.862	0.765	0.784	0.810	0.833	0.785	**0.845**	0.843	0.789
EB → FD	0.872	0.783	0.780	0.810	0.816	0.788	0.823	0.793	**0.841**
EM → FD	0.872	0.780	0.745	0.798	0.822	0.761	**0.841**	0.829	0.775
EB → FM	0.889	0.771	0.762	0.822	0.753	0.794	**0.833**	0.824	0.829
ED → FM	0.889	0.745	0.757	0.836	0.826	0.827	0.847	0.849	**0.855**
ED → JB	0.812	0.700	0.725	0.738	0.675	0.715	**0.761**	0.741	0.741
EM → JB	0.812	0.642	0.708	0.711	0.621	0.636	0.721	0.689	**0.722**
EB → JD	0.834	0.708	0.742	**0.813**	0.663	0.710	0.805	0.789	0.782
EM → JD	0.834	0.693	0.756	0.792	**0.828**	0.721	0.790	0.763	0.711
EB → JM	0.842	0.673	0.742	0.826	0.699	0.811	**0.831**	0.826	0.827
ED → JM	0.842	0.710	0.776	**0.817**	0.804	0.762	0.816	**0.817**	0.804
German	0.854	0.771	0.810	0.802	0.811	0.783	**0.826**	0.819	0.809
French	0.874	0.772	0.770	0.803	0.808	0.784	**0.840**	0.831	0.824
Japanese	0.829	0.688	0.742	0.783	0.715	0.726	**0.787**	0.771	0.765
Books	0.847	0.739	0.776	0.770	0.752	0.733	**0.802**	0.788	0.788
DVDs	0.847	0.752	0.767	0.797	0.792	0.757	**0.820**	0.802	0.783
Music	0.863	0.739	0.779	0.822	0.790	0.803	**0.831**	**0.831**	0.826
Average	0.852	0.743	0.774	0.796	0.778	0.768	**0.818**	0.807	0.799