DISTRIBUTIONAL CORRESPONDENCE INDEXING FOR CROSS-LANGUAGE TEXT CATEGORIZATION

Andrea Esuli, Alejandro Moreo Fernández

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche, Pisa, Italy,
{andrea.esuli,alejandro.moreo}@isti.cnr.it

Distributional Correspondence Indexing is a very efficient feature-representation-transfer method for cross-language domain adaptation that directly applies the Distributional Hypothesis *to the concept of* Pivot features. *Our method performs favorably to state-of-the-art methods on a popular benchmark, requiring a significantly reduced computational cost and minimal human intervention.*

INTRODUCTION

Manually annotating labeled collections of documents requires substantial human effort, and it is inherently language-dependent. Cross-Language Text Categorization (CLTC [2]) aims at using the labeled examples available for a prevailing *source* language (typically English) to learn a classifier for a different resource-scarce *target* language. A practical scenario is to exploit the labeled opinions in English available on the Web to build sentiment classifiers for other languages.
Previous approaches to CLTC presented in literature make use of:

- Machine Translation tools: straightforward solution, bound however to the costs/availability of proper tools.
- Parallel Corpora: allow sophisticated statistical analysis, but are likely unavailable.

Structural Correspondence Learning (SCL [1]) was applied to the cross-language setting (CL-SCL [3]) by querying a word-translator oracle to create a set of word pairs (dubbed *pivots*).

- The pivots could be used to discover structural analogies between the source and target languages requiring only *unlabeled corpora* (easily obtainable).
- It still suffers from considerably high computational costs, deriving from the intermediate optimizations of the *structural problems* and from the use of Latent Semantic Analysis.

DISTRIBUTIONAL CORRESPONDENCE INDEXING

Our method is a direct application of the **distributional hypothesis** –words with similar distributions of use in text are likely to have similar meanings– to the concept of **pivots** –frequent and highly predictive features for the task that are expected to behave in a similar way in both domains.
Given a small sets of pivots, features from both languages are projected into a common vector space in which each dimension reflects the *distributional correspondence* between the feature being projected and a pivot. Semantically related words from the source and target languages should present similar distributions to pivots, thus obtaining similar representations.
DCI work-flow includes the following steps:

Pivot selection

Similar to [3], features with frequency greater than $\phi = 30$ are ranked by their relevance with respect to the classification task according to *mutual information*. The word-oracle is then requested to translate each source word t_S into its translation-equivalent word t_T in the target language, to form the pivot pairs $p = \langle t_S, t_T \rangle$. Top-$m$ features are selected as pivots.

Feature Profiles

Each source and target feature f (including pivots) is profiled as an m-dimensional vector:

$$\overrightarrow{f} = (\eta(f, p_1), \eta(f, p_2), ..., \eta(f, p_m)) \quad (1)$$

where p_i is the source or target word in the i^{th} pivot, and η denotes the *distributional correspondence function* (DCF) between the feature f and p_i. The DCF is efficiently estimated on sets of unlabeled documents for each language as (equation 2). All feature profile vectors $\overrightarrow{f_i}$ are then normalized to unit length.

$$\eta(f, p) = P(f|p) - P(f|\overline{p}) \quad (2)$$

Unification

As pivots are expected to behave similarly in both languages, we *unify* their feature profiles by averaging them. Unification is also applied to profiles of proper nouns or non-lexicalized terms (e.g., *Chopin* or *AC/DC*) common to the source and target languages.

Document Indexing

Train and test documents are projected into the cross-lingual space as the weighted sum of all profile vectors associated to their features.

$$\overrightarrow{d_j} = \sum_{f_i \in d_j} w_{ij} \cdot \overrightarrow{f_i} \quad (3)$$

where w_{ij} is the $tf \cdot idf$ weight of feature f_i in document d_j.

RESULTS

We test our method[1] on the publicly available Webis-CLS-10 Cross-Lingual Sentiment collection used in [3] on which several related methods reported results. The dataset consists of Amazon product reviews in four languages (**E**nglish, **G**erman, **F**rench, and **J**apanese), covering three product categories (**B**ooks, **D**VDs, and **M**usic). Acronyms will denote tasks, e.g., "EGB" refers to English-German Books reviews adaptation.
Requiring on average only 16.3s to run, our method performs better than the comparison methods (Table 1) in most cases – with statistical significance to LSI, KCCA, and OPCA – with $m = 450$ and $m = 100$, and still performs fine with just 20 pivots. Its performance tends to improve and stabilize as m increases (Figure 1). As embeddings, features profiles seem to capture cross-lingual semantics (Table 2).

	Upper	MT	SCL	LSI	KCCA	OPCA	SSMC	DCI_{450}	DCI_{100}	DCI_{20}
EGB	86.75	79.68	**83.34**	77.59	79.14	74.72	81.88	76.25	81.40	79.50
EGD	83.50	77.92	80.89	79.22	76.73	74.59	**82.25**	80.40	79.95	77.75
EGM	85.90	77.22	82.90	73.81	79.18	74.45	81.30	75.20	**83.30**	73.70
EFB	86.15	80.76	81.27	79.56	77.56	76.55	**83.05**	82.95	82.30	75.15
EFD	87.15	78.83	80.43	77.82	78.19	70.54	82.70	**84.10**	82.40	64.35
EFM	88.95	75.78	78.05	75.39	78.24	73.69	80.46	**81.90**	81.05	75.80
EJB	81.15	70.22	77.00	72.68	69.46	71.41	73.76	73.90	**79.10**	74.50
EJD	83.40	71.30	76.37	72.55	74.79	71.84	77.58	81.55	**82.25**	80.25
EJM	84.20	72.02	77.34	73.44	73.54	74.96	77.53	78.45	**82.00**	79.30

Table 1: Accuracy performance in the Webis-CLS-10 collection

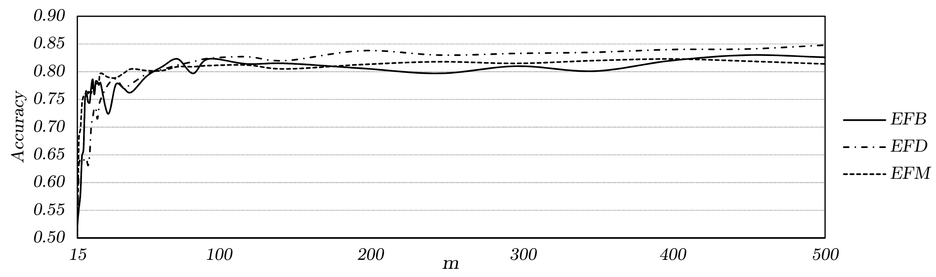

Figure 1: Variation of accuracy at the variation of the number of pivots for EF* setups.

beautifully	**classical**	**delightful**
schöne (beautiful) 0.635	*adagio* 0.767	魅力(attractive) 0.610
liebevoll (loving) 0.596	*Martenot* 0.746	描き出さ(portrayed) 0.546
sehnsucht (longing) 0.533	*Charles-Marie* 0.736	風景(scenes) 0.545
ungewöhnlich (unusual) 0.510	*violoncelle* (cello) 0.727	繊細(delicate) 0.542
phantastisch (fantastic) 0.507	*soliste* (soloist) 0.720	味わえる(taste) 0.538

Table 2: Five most similar (cosine similarity) words in a target language given a word in English.

CONCLUSIONS

Distributional Correspondence Indexing is an efficient feature-representation-transfer method for CLTC that creates feature profiles based on their distributional correspondence to a small set of pivots. Empirical evaluation demonstrated our method compares favorably to the state of the art in Webis-CLS-10 dataset. However, DCI has a much lower computational cost, and requires less human intervention.

REFERENCES

[1] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 120–128, 2006.

[2] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[3] P. Prettenhofer and B. Stein. Cross-lingual adaptation using structural correspondence learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(1):13, 2011.

ECIR 2015, 29 March - 2 April, 2015, Vienna, Austria

[1] The code to replicate our experiments is available at http://hlt.isti.cnr.it/dci/