Distributional Correspondence Indexing for Cross-Language Text Categorization

Andrea Esuli and 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

Abstract. Cross-Language Text Categorization (CLTC) aims at producing a classifier for a target language when the only available training examples belong to a different source language. Existing CLTC methods are usually affected by high computational costs, require external linguistic resources, or demand a considerable human annotation effort. This paper presents a simple, yet effective, CLTC method based on projecting features from both source and target languages into a common vector space, by using a computationally lightweight distributional correspondence profile with respect to a small set of pivot terms. Experiments on a popular sentiment classification dataset show that our method performs favorably to state-of-the-art methods, requiring a significantly reduced computational cost and minimal human intervention.

Keywords: Cross-Language Text Categorization, Distributional Semantics, Sentiment Analysis.

1 Introduction

Automated Text Categorization methods usually rely on a *training set* of labeled examples to learn a classifier that will then predict the categories of unlabeled documents. The creation of a training set requires substantial human effort, and it is inherently language-dependent. Cross-Language Text Categorization (CLTC [1]) aims at using the labeled examples available for a *source* language to learn a classifier for a different *target* language, thus reducing, or completely avoiding, the need for human labeling of examples in the target language. A practical scenario for CLTC is to exploit the labeled examples freely available on the Web for the prevailing languages (e.g., English star-rated reviews) to build classifiers for languages for which the amount of labeled examples is much smaller.

A number of different approaches to CLTC have been presented in literature. The use of *Machine Translation* (MT) [8,10] to reduce all the documents to a single language is a straightforward solution, but it is bound to the availability of MT systems/services for the relevant languages, and it suffers from the cost, economical and of time, of translating a large number of documents.

Methods exploiting *parallel corpora* [3,5,11] are usually affected by the high computational costs derived from the use of a sophisticated statistical analysis, e.g.,

A. Hanbury et al. (Eds.): ECIR 2015, LNCS 9022, pp. 104–109, 2015.

Principal Component Analysis (PCA), and are bound to the availability of a parallel corpus between the relevant languages.

Structural Correspondence Learning (SCL [2]) was applied to the cross-language setting (CL-SCL [6,7]) by using a word-translator oracle in order to create a set of word pairs (dubbed *pivots*). The pivots are later used to discover structural analogies between the source and target languages through unlabeled corpora. Even though CL-SCL succeeded in alleviating the problems posed by the use of MT tools, it still has a considerable computational cost, deriving from the intermediate optimizations of the *structural problems* (i.e., pivot predictors), and from the use of Latent Semantic Analysis (LSA).

Our method takes the CL-SCL idea as an inspiration, but it follows a different, simpler approach, with a more direct application of the *distributional hypothesis*, which states that words with similar distributions of use in text are likely to have similar meanings. Given a small sets of pivots, textual features extracted from both languages are projected into a common vector space (feature representation transfer [4]) in which each dimension reflects the *distributional correspondence* between the feature being projected and a pivot. The distributional correspondence is efficiently estimated on sets of unlabeled documents for each language. There is no need for a parallel corpus, and computationally-expensive statistical techniques are avoided.

Despite being simple, this method compares favorably to the state of the art in experiments on a popular sentiment classification dataset, sporting a significantly reduced computational cost, and also requiring less human intervention.

2 Distributional Correspondence Indexing

In the traditional bag-of-words model each word is mapped into a dedicated dimension of the vector space. Without resorting to translation or other source of external knowledge, words like the English "beautiful" and its German equivalent "*schöne*" point to orthogonal directions in the vector space, while their vectorial representation should be aligned in order to model their correspondence.

Our *Distributional Correspondence Indexing* (DCI) method profiles each feature with respect to its distributional correspondence to the pivots. As word pairs defining the pivots are expected to behave similarly in their respective language, semantically related words from the source and target languages should present similar distributions to them, thus obtaining similar representations.

Pivot selection. Words from the source training set are ranked by their relevance with respect to the classification task by means of a supervised feature selection function; similarly to [7], we use mutual information. The 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 $p = \langle t_S, t_T \rangle$. Following [7] the set of pivots consists of the top-m pivots with a *support* (occurrences in the unlabeled corpora) greater than a given threshold ϕ.

Feature profiles. Differently from [7], we propose to represent each source and target feature f (including pivots) as an m-dimensional profile 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* between the feature f and p_i, that we model with a probability-based linear function[1] that requires minimal computation:

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

where $P(f|p)$ denotes the conditional probability of finding f in documents containing p, and $P(f|\overline{p})$ is conditioned on documents not containing p. Both probabilities are estimated on the set of unlabeled documents for the pertinent language. All feature profile vectors $\overrightarrow{f_i}$ are then normalized to unit length.

Unification. As we assume pivot terms behave similarly in both languages, we *unify* their feature profiles by averaging them. Unification is also applied to profiles of words that the source and target languages have in common (e.g., proper nouns or non-lexicalized terms) having a support greater than ϕ.

Document indexing. Finally, train and test documents are represented into the cross-lingual space as the weighted sum of all profile vectors associated to their features. That is, document d_j is represented as the m-dimensional vector

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

where w_{ij} is the weight of feature f_i in document d_j. We used the normalized $tf \cdot idf$ weighting criterion in our implementation.

3 Experiments

We test our method[2] on the publicly available Webis-CLS-10 Cross-Lingual Sentiment collection[3] proposed in [6]. The dataset consists of Amazon product reviews written in four languages (**E**nglish, **G**erman, **F**rench, and **J**apanese), covering three product categories (**B**ooks, **D**VDs, and **M**usic). For each language-category pair there are 2,000 training documents, 2,000 test documents, and from 9,000 to 50,000 unlabeled documents depending on the language-category combination. Following [6], we consider English as the source language, and German, French, and Japanese as the target ones. Documents are either labeled as *Positive* or *Negative* (binary classification), and any train or test set contains an equal amount of positive and negative examples. The evaluation measure is *accuracy*, which is adequate since labels are always balanced in the dataset.

In our implementation we set $\phi = 30$, following the results of [6]. We test our method on three sizes for the pivot set: $m = 450$, which is the best-performing

[1] We also investigated other alternatives coming from information theory including Information Gain, χ^2, and Odds ratio, with negative or unstable results.

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

[3] `http://www.uni-weimar.de/en/media/chairs/webis/research/corpora/corpus-webis-cls-10/`

Table 1. Accuracy for cross-lingual sentiment analysis in the Webis-CLS-10 collection. Acronyms indicate source/target/product: "EGB" stands for English/German/Books.

	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

setup for SCL [6], $m = 100$, which is the minimal number of pivots tested in [6], and a minimal setup using just $m = 20$ pivots. To emulate the word-oracle – and for the sake of a fair comparison – we used the bilingual dictionary provided by [6]. We used the popular SVM^{light} implementation[4] of Support Vector Machines as the learning device, with default parameters.

In order to have an upper reference to accuracy, we implemented a method that trains the SVM classifier on the training set of the target language (Upper). We also report the MT baseline (MT) of [7], which first translates the target examples set into the source language. In Table 1 we compare DCI to the results published on the same dataset, same configuration, for five CLTC methods: structural correspondence learning (SCL [7]), latent semantic indexing (LSI [3]), kernel canonical correlation analysis (KCCA [9]), oriented principal component analysis (OPCA [5]), and semi-supervised matrix completion (SSMC [11]).

DCI_{450} obtains good results, performing better than the compared methods in four cases out of nine. DCI_{100} performs even better (five out of nine, and four highest results). DCI_{100} performs better than SCL in seven cases out of nine, with SCL requiring 450 calls to a word-oracle, 450 structural optimization problems, and LSA. DCI_{100} instead only needs 100 word-translations plus feature profile calculation and document indexing, which is extremely efficient[5]. SSMC performs better than DCI_{100} on German and French. SSMC algorithm requires however a parallel corpus, a double-sized source training set, and some labeled examples from the target language. Figure 1 shows how accuracy varies when varying m in the range between 15 and 500.

We noted that DCI performs much better than the other methods when Japanese is the target language. Given that DCI is applied to the same textual features used by all the other methods, and adopts the same SVM learner of Upper, with exactly the same parameters, we deem this difference to a better

[4] Available at `http://svmlight.joachims.org/`

[5] It took 22.2s, 15.3s, and 11.2s on average in the Books, DVDs, and Music tasks, respectively, to create the feature profiles and build the training index on a single threaded process on a 1.6GHz processor.

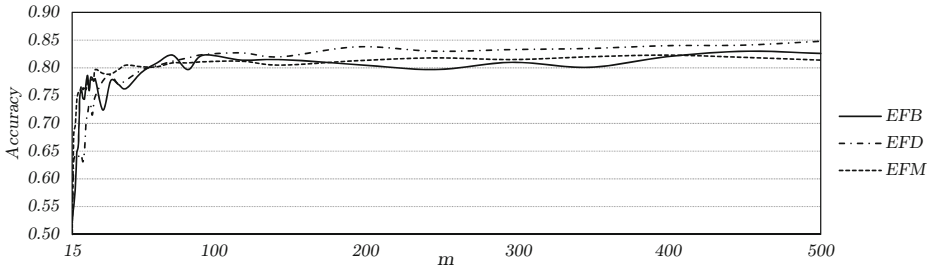

Fig. 1. Variation of accuracy at the variation of the number of pivots for EF* setups

Table 2. Five most similar words in a target language given a word in English

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

ability of DCI to embed the dispersed knowledge contained in less informative features, though this is a point left open to future investigation.

Statistical significance tests (paired t-test on the accuracy values) report that both DCI_{100} and DCI_{450} are significantly better, respectively with $p < 0.001$ and $p < 0.05$, than LSI, KCCA, and OPCA. There are no statistically significant differences between DCI, SCL and SSMC, so the comparison substantially ends with a tie, which is already a good result for a method so lightweight as DCI.

DCI obtains good results with just $m = 20$ pivots. For this value the list of source words to be translated is so small and composed by common-use words that even a user with an average proficiency in the foreign language could translate them without requiring external knowledge sources[6].

As a final note, we explored the ability of our feature profiles to capture the semantic relatedness of words, considering them as "cheap" word embeddings [12]. Table 2 illustrates the semantic properties captured by our feature profiles; it lists the most similar (cosine similarity) target words to a given source word.

4 Conclusions and Future Work

We have proposed Distributional Correspondence Indexing, an efficient feature-representation-transfer method for CLTC that creates feature profiles based on their distributional correspondence to a small set of pivots. The method indexes

[6] For example, for the EJD task the words to be translated were: great, worst, bad, awful, horrible, disappointed, terrible, love, wonderful, worse, disappointing, why, favorite, fun, performance, poor, collection, money, please, and enjoy.

documents in different languages into a common vector space where they become comparable. Empirical evaluation demonstrated our method performs comparably, and even better in some cases, to state-of-the-art methods. However, DCI has a much lower computational cost, and requires less human intervention.

DCI is a promising method, with many aspects worth being investigated: e.g., more sophisticated distributional correspondence functions; how to determine the optimal pivot set; testing DCI on imbalanced classes.

References

1. Bel, N., Koster, C.H.A., Villegas, M.: Cross-lingual text categorization. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 126–139. Springer, Heidelberg (2003)
2. Blitzer, J., McDonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 120–128 (2006)
3. Dumais, S.T., Letsche, T.A., Littman, M.L., Landauer, T.K.: Automatic cross-language retrieval using latent semantic indexing. In: AAAI Spring Symposium on Cross-language Text and Speech Retrieval, p. 21 (1997)
4. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)
5. Platt, J.C., Toutanova, K., Yih, W.T.: Translingual document representations from discriminative projections. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 251–261 (2010)
6. Prettenhofer, P., Stein, B.: Cross-language text classification using structural correspondence learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1118–1127 (2010)
7. Prettenhofer, P., Stein, B.: Cross-lingual adaptation using structural correspondence learning. ACM Transactions on Intelligent Systems and Technology (TIST) 3(1), 13 (2011)
8. Rigutini, L., Maggini, M., Liu, B.: An EM-based training algorithm for cross-language text categorization. In: Proceedings of the 3rd IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 529–535 (2005)
9. Vinokourov, A., Shawe-Taylor, J., Cristianini, N.: Inferring a semantic representation of text via cross-language correlation analysis. In: Proceedings of the 16th Annual Conference on Neural Information Processing Systems (NIPS), pp. 1473–1480 (2002)
10. Wan, X.: Co-training for cross-lingual sentiment classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, pp. 235–243 (2009)
11. Xiao, M., Guo, Y.: Semi-supervised matrix completion for cross-lingual text classification. In: Twenty-Eighth AAAI Conference on Artificial Intelligence (2014)
12. Zou, W.Y., Socher, R., Cer, D.M., Manning, C.D.: Bilingual word embeddings for phrase-based machine translation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1393–1398 (2013)