Cross-lingual Text Classification Using Topic-Dependent Word Probabilities

Daniel Andrade **Akihiro Tamura** **Masaaki Tsuchida** **Kunihiko Sadamasa**
Knowledge Discovery Research Laboratories, NEC Corporation, Japan
{s-andrade@cj, a-tamura@ah,
m-tsuchida@cq, k-sadamasa@az}.jp.nec.com

Abstract

Cross-lingual text classification is a major challenge in natural language processing, since often training data is available in only one language (target language), but not available for the language of the document we want to classify (source language). Here, we propose a method that only requires a bilingual dictionary to bridge the language gap. Our proposed probabilistic model allows us to estimate translation probabilities that are conditioned on the whole source document. The assumption of our probabilistic model is that each document can be characterized by a distribution over topics that help to solve the translation ambiguity of single words. Using the derived translation probabilities, we then calculate the expected word frequency of each word type in the target language. Finally, these expected word frequencies can be used to classify the source text with any classifier that was trained using only target language documents. Our experiments confirm the usefulness of our proposed method.

1 Introduction

Text classification is ubiquitous in natural language processing. It's applications range from simple topic detection, like articles about sport vs articles about computers, to sentimental analysis, and subtle discrimination of Tweets that report the abuse of drugs or the metaphoric use of drugs ("love is like a drug"). Text classification hugely relies on manually annotated training data in one language.

However, creating training data for each language is expensive, and therefore, we are interested in using training data given in only one language (e.g. English, denoted as target language) to classify text written in a different language (e.g. Chinese, or Japanese, denoted as source language).

Our approach addresses this issue by using a simple bilingual dictionary. Bilingual dictionaries have the great advantage that they are available often for free[1], and have good coverage for major languages, like Chinese and Japanese. With the help of the dictionary, we calculate the expected frequency of each word in the target language. Finally, we create a feature vector in the target language that is used as input for the text classifier.

However, due to the translation ambiguity of a word in the source language, it is important to carefully choose the translation probability for calculating the expected frequencies of the target words. For example, consider a Japanese news article that contains the word 拘束 (restrict, restrain, in custody), and we want to find out whether the article is about "foreign policy" or not. The most simple method is to use all its English translations, and assume a uniform distribution over them, i.e. $\{0.33, 0.33$ and $0.33\}$. However, depending on the topic of the news article, the translation "in custody" is more appropriate. For example, if the article reports about a crime/crime suspect, the translation "in custody" is more likely than "restrict" and "restrain". Conversely, if the article is about "military", the translation "in custody" is less likely. Moreover, an article that is about the topic "military" is more likely to belong to the class "foreign policy". This example demonstrates the importance of estimating good translation probabilities in order to improve the clas-

[1]For example from Wikitionay.org under Creative Commons Licence.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1466–1471,
Denver, Colorado, May 31 – June 5, 2015.

sification of the source text.

Therefore, we propose a probabilistic model that uses latent document topics to help improve the translation probabilities for a source document. Our experiments, on three different pairs of corpora, confirm that our probabilistic model for estimating word translation probabilities is helpful for cross-lingual text classification.

2 Related Work

The work in (Wu et al., 2008) and (Shi et al., 2010) uses a bilingual dictionary for cross-lingual text classification. The method described in (Wu et al., 2008) is motivated by transfer learning to adjust the class probability $p(c)$ to account for the differences in distributions between source and target language. Similar to our work, in the first step, they generate a probabilistic bilingual lexicon that contain word translation probabilities $p(e|f)$. However, one main difference to our work is that they translate each source word f in source text F independently, without considering any topic or context information of F.

Instead of translating the source text into the target language, the method in (Shi et al., 2010) suggests to translate the target classification model into the source language. They directly estimate the translation probabilities $p(f|e, c)$ using the source and target language data. One limitation of their method is that it assumes that the class of the document, that we want to translate, is given.

Our idea of learning word translation probabilities in context is related to the work in (Koehn and Knight, 2000). They describe an efficient method for learning word translation probabilities $p(f|e)$ using a bilingual dictionary and a pair of comparable corpora[2]. Like our approach, their method has the advantage that no parallel corpora are needed for translation. However, to solve the ambiguity of word-translation they considered only (local) bi-gram context. Moreover, their method assumes that the word order in the languages are the same. This is obviously not the case for language pairs like English and Japanese.

We note that the bilingual paired topic model, suggested in (Jagarlamudi and Gao, 2013), can also be used to disambiguate and select the appropriate word translations by using the topic associated with the given document. However, their model does not consider the use of a document class, and uses fixed word translation probabilities. In Section 3.2, we show that our model can also be used to learn the translation probabilities.

Alternatively, the multi-lingual topic model described in (Ni et al., 2011), and the use of a common low-dimensional projection described in (Platt et al., 2010) have also been applied to the cross-lingual text classification problem. However, both models require for training that cross-lingually aligned documents are available.

3 Proposed Method

Our proposed method does not use one translation of F, but implicitly generates all translations and weights them by the probability of each translation. More formally, let E be one translation of source text F. Moreover, let $count_E(e)$ denote the frequency of word e in E. Instead of using $count_E(e)$, we use the expected number of word occurrences denoted by $\mathbb{E}[count_E(e)|F]$ as features. When we use a simple uni-gram language model in the source language we get:

$$\mathbb{E}[count_E(e)|F] = \sum_{j=1}^{k} p(e_j = e|f_j) \qquad (1)$$

where we might write F as $(f_1, f_2, f_3, ...f_k)$, where f_j is the j-th word in F, and k is the number of words in source text F.[3] The random variable e_j denotes the translation of the j-th word in F. However, such a simple model translates each source word independently and ignores the context of the word.

In the following, we describe a probabilistic model that allows us to consider the whole document context F into account for translating one word f_j. The generative story is as follows:

1. For each document, we generate a class label c with probability υ_c. Here we consider only the binary classification task with class label "positive", or "negative".

[2]Two corpora written in different languages which do not need to be translations of each other

[3]Here "word" refers to a word occurrence (and not unique word). Therefore, k is the length of the source text F.

2. For each document, we generate a topic z with probability $\pi_{z|c}$.

3. Given topic z, we generate each word e in the target language document independently from a categorical distribution with probability $\vartheta_{e|z}$.

4. For each word e in the target language, we generate a word f in the source language independently from a categorical distribution with probability $\theta_{f|e}$.[4]

Under this model, for one target document $(e_1, ..., e_k)$ and its corresponding source document $(f_1, ..., f_k)$, the joint probability $p(z, c, e_1, ..., e_k, f_1, ..., f_k)$ is

$$\upsilon_c \pi_{z|c} \prod_{j=1}^{k} \vartheta_{e_j|z} \cdot \theta_{f_j|e_j} \,.$$

The parameter vector $\boldsymbol{\vartheta}_z$ specifies the target word probabilities $\vartheta_{e|z}$ that can be learned from the target language training data as described in Section 3.1. The parameter vector $\boldsymbol{\theta}_e$ specifies the word translation probability $\theta_{f|e}$ for a target word e into a source language word f. These word translation probabilities are determined with the help of the bilingual dictionary as described in Section 3.2.

Our goal is to estimate the translation probability $p(e|f_j, F)$, since this allows us to calculate

$$\mathbb{E}[count_E(e)|F] = \sum_{j=1}^{k} p(e_j = e|f_j, F) \,. \qquad (2)$$

Note, that under our proposed probabilistic model, it holds that

$$p(e_j|f_j, F) \;=\; \sum_z p(e_j|f_j, z) \cdot p(z|F) \,.$$

This can be interpreted as follows. First, the model determines a probability distribution over the latent topics, conditioned on the given input source document, i.e. $p(z|F)$. And then, second, the model uses the conditional probability $p(z|F)$ to determine the translation probability for each word in the source document, i.e. $p(e_j|f_j, z)$.

The actual calculation of $p(e_j = e|f_j, F)$ can be derived as follows.[5]

$$\begin{aligned} p(e_j|f_j, F) &= p(e_j|f_1, ..., f_k) \\ &\propto p(e_j, f_1, ..., f_k) \\ &= \sum_c \sum_z p(e_j, f_1, ..., f_k|z) p(z|c) p(c) \,, \end{aligned}$$

where the probability $p(e_j, f_1, ..., f_k|z)$ can be efficiently calculated using

$$\begin{aligned} & \sum_{e_{l_1} \in V} ... \sum_{e_{l_{k-1}} \in V} p(e_1, \ldots, e_k, f_1, \ldots, f_k|z) \\ = & \sum_{e_{l_1} \in V} ... \sum_{e_{l_{k-1}} \in V} \prod_{j'=1}^{k} \theta_{f_{j'}|e_{j'}} \cdot \vartheta_{e_{j'}|z} \\ = & \; \theta_{f_j|e_j} \cdot \vartheta_{e_j|z} \prod_{j' \in \{l_1 ... l_{k-1}\}} \sum_{e_{j'} \in V} \theta_{f_{j'}|e_{j'}} \cdot \vartheta_{e_{j'}|z} \,, \end{aligned}$$

where the indexes $l_1 \ldots l_{k-1}$ correspond to $1, \ldots, j-1, j+1, \ldots k$.

3.1 Learning υ_c, $\pi_{z|c}$, and $\vartheta_{e|z}$

Note that under our model, class c and topic z are independent from $f_1, ..., f_k$ given document $e_1, ..., e_k$ in the target language. Therefore, the parameters υ_c, $\pi_{z|c}$, and $\vartheta_{e|z}$ can be learned solely using the training documents in the target language. Given a collection of training documents with known classes $D = \{(E_1, c_1)..., (E_n, c_n)\}$, we can estimate the parameters as follows.

Parameter υ_c is estimated using the maximum-likelihood (ML), which is

$$\upsilon_c^* = \frac{\sum_{i=1}^{n} 1_c(c_i)}{n} \,, \qquad (3)$$

where $1_x(y)$ is the indicator function which is 1, if $x = y$, otherwise 0.

The optimal ML-estimate of $\vartheta_{e|z}$ and $\pi_{z|c}$ can be found by maximizing $\log p(D|\vartheta, \pi)$, for which, however, an analytic solution cannot be derived. Therefore, instead, we use the EM-algorithm

[4] It might seem that we need cross-lingually aligned documents, or documents of same length in both languages. However, both is not the case, since in our experiments the translations will always be unobserved, and therefore sum over all possible translations.

[5] When it is clear from the context, we write $p(e_j)$ instead of $p(e_j = e)$.

(Dempster et al., 1977), deriving for the E-step: setting the probability distribution q to

$$p(z_i|D,\vartheta,\pi) \propto \pi_{z_i|c_i} \prod_{j=1}^{k_i} \sum_{e_j} \vartheta_{e_j|z_i}\,, \tag{4}$$

and in the M-step:

$$\vartheta^*_{e|z} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i} 1_e(e_j)\cdot q(z_i=z)}{\sum_{i=1}^{n}\sum_{j=1}^{k_i} q(z_i=z)} \tag{5}$$

and

$$\pi^*_{z|c} = \frac{\sum_{i=1}^{n} 1_c(c_i)\cdot q(z_i=z)}{\sum_{i=1}^{n} q(z_i=z)}\,. \tag{6}$$

3.2 Learning $\theta_{f|e}$

Here we propose to chose the translation probabilities $\theta_{f|e}$ with highest probability, under our current model, and such that the probability of observing the source documents (without labels) is maximized. Formally, given a collection of source documents $D' := F_1, ..., F_m$, the optimal translation probability $\theta^*_{f|e}$ is

$$\underset{\theta_{f|e}}{\operatorname{argmax}}\, p(D'|\theta_{f|e}, \upsilon^*_c, \pi^*_{z|c}, \vartheta^*_{e|z})\,,$$

where $\upsilon^*_c, \pi^*_{z|c}, \vartheta^*_{e|z}$ are the parameters learned in the previous section. Unfortunately, the exact optimization is intractable, and therefore, we resort again to an EM-approximation, analogously to before.

The E-step corresponds to setting for each source document i, the probability $q(e_{i,1}, ..., e_{i,k_i})$ to

$$\begin{aligned} & p(e_{i,1}, ..., e_{i,k_i}|f_{i,1}, \ldots f_{i,k}, \theta_{f|e}) \\ \propto & \sum_{c_i}\sum_{z_i} p(c_i)p(z_i|c_i) \prod_{j=1}^{k} p(e_{i,j}|z_i)\theta_{f_{i,j}|e_{i,j}}\,. \end{aligned}$$

In the M-step, we update $\theta_{f|e}$ to

$$\theta^*_{f|e} \;=\; \frac{\sum_{i=1}^{m}\sum_{j=1}^{k_i} 1_f(f_{j,i})\cdot q(e_{i,j}=e)}{\sum_{i=1}^{m}\sum_{j=1}^{k_i} q(e_{i,j}=e)}\,.$$

4 Experiments

For our experiments we use three pair of corpora denoted by NEWS, WEB, and TWEETS. The corpora NEWS contains news articles in English and Japanese crawled from Internet news sites during 2012-2013, and were annotated as being related to "foreign policy" or not related. The corpora WEB contains web pages in English and Chinese that are categorized either as "sport" or "computer" in the Open Directory Project (ODP)[6] crawled in 2013. TWEETS contains tweets in English and Chinese gathered during 2013, classified as related to "violence", or not related.[7]

Method	NEWS	WEB	TWEETS
Co	**0.687** (0.68)	0.842 (0.84)	**0.430** (0.18)
Co (freq)	0.668 (0.68)	0.849 (0.83)	0.424 (0.20)
Co (uni)	0.666 (0.68)	0.842 (0.83)	0.426 (0.22)
Wu et al.	0.632 (0.56)	0.849 (0.74)	0.391 (0.13)
Freq	0.635 (0.58)	0.842 (0.76)	0.376 (0.13)
Uniform	0.628 (0.53)	**0.856** (0.76)	0.407 (0.13)
CN/JA only	0.816 (0.81)	0.893 (0.90)	0.894 (0.89)
EN only	0.718 (0.67)	0.967 (0.97)	0.682 (0.67)

Table 1: Shows the break-even point (f1-score) of the proposed method Co and three baselines for each pair of corpora. Co (freq) and Co (uni) denote the proposed method without estimation of dictionary probabilities, but instead using word frequency and uniform distribution, respectively.

We tokenize and stem the words in the English corpora using Senna (Collobert et al., 2011). For Chinese and Japanese we use the morphological analyzers described in (Qiu et al., 2013), and an in-house analyzer, respectively. The Chinese to English dictionary, and the Japanese to English dictionary contains translations for 94351 and 1483440 words, respectively.

For the classification we use LIBSVM (Chang and Lin, 2011) with linear kernel, and the feature representation as suggested in (Rennie et al., 2003).

For the parameter estimation of our proposed model we use EM, as described in Section 3.1 and 3.2.[8] The number of topics was determined by optimizing the f1-measure using only the English training data when applying the probabilistic model to monolingual text classification. In order to prevent non-zero probabilities, we use a symmetric Dirichlet

[6] www.dmoz.com

[7] The number of documents in the corpora pairs for source/target language are 2472/2289, 1302/6294, and 2005/1499 for NEWS, WEB and TWEETS, respectively.

[8] We observed convergence for less than 50 iterations.

prior.

We compare our proposed method "Co" to four different baselines that also use solely a bilingual dictionary. For all methods (baselines and proposed), we use Equation (2) to estimate the expected word frequencies. The baseline "Wu et al." refers to the method proposed in (Wu et al., 2008). The baseline "Freq" sets the probability $p(e|f)$ to be proportional to the word frequency in the training data. Analogously, the baseline "Uniform" assumes a uniform probability over all translations of f.

For measuring the performance of each text classifier we use precision and recall. The break-even point[9] and the f1-measure of our proposed method and all baselines are shown in Table 1. As can be seen, our method performs favorable for the NEWS and TWEETS corpora. For the WEB corpora pair and our proposed method is at par with the baseline "Wu et al.", and looses slightly to the "Uniform" baseline. For reference, we also show the upper bounds "CN/JA only" and "EN only" that train and test in the same source and target language, respectively.[10]

We also analyzed the contribution of using the word translation probabilities learned in Section 3.2. The method "Co (freq)" is the same as our proposed method, except that the translation probabilities $p(f|e)$ are not estimated using the method described in Section 3.2, but instead simply uses the word-frequency distribution. Analogously, the method "Co (uni)" is the same as our proposed method, except that $p(f|e)$ is set to the uniform probability for all translations of e. Limiting the discussion to break-even points, we see, in Table 1, an improvement of around 2 percent points for NEWS, but only minor changes in performance for the other two corpora (WEB and TWEETS).

Finally, we give an example which shows the translation probabilities for the word 拘束 (restrict, restrain, custody) for two different source documents in NEWS. The first source document F_1 reports a military action, and is labeled as "foreign policy". The second document F_2 is a news article about terror, and is labeled as "not foreign policy". The results shown in Table 2, confirm our intuition, that the translation "custody" is more likely in documents related to crime.

	e = restrict	**e = restrain**	**e = custody**
$p(e\|f, F_1)$	0.33	0.10	0.57
$p(e\|f, F_2)$	0.02	0.00	0.98

Table 2: Shows the translation probabilities for the source word f = 拘束, within document F_1 (military related, class is "foreign policy") and document F_2 (terror related, class is not "foreign policy").

5 Conclusions

In contrast, to most previous work, we focused on the word translation problem, rather than the domain-adaptation problem for cross-lingual text classification. We have proposed a probabilistic model that allows us to estimate word-translation probabilities that are conditioned on the whole source document. Our experiments on three different pairs of corpora, show that our estimated translation probabilities can improve text classification accuracy, and that our estimated word translation probabilities are able to reflect the topic of a text.

References

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Arthur P Dempster, Nan M Laird, Donald B Rubin, et al. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38.

Jagadeesh Jagarlamudi and Jianfeng Gao. 2013. Modeling click-through based word-pairs for web search. In *Proceedings of the ACM SIGIR Conference*, pages 483–492. ACM.

P. Koehn and K. Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *Proceedings of the National Conference on Artificial Intelligence*, pages 711–715. Association for the Advancement of Artificial Intelligence.

[9] That is the point where precision and recall are equal.

[10] These results were acquired using cross-validation.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2011. Cross lingual text classification by mining multilingual topics from wikipedia. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 375–384. ACM.

John C Platt, Kristina Toutanova, and Wen-tau Yih. 2010. Translingual document representations from discriminative projections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 251–261. Association for Computational Linguistics.

Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Fudannlp: A toolkit for chinese natural language processing. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*.

Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the International Conference on Machine Learning*, volume 3, pages 616–623.

Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067. Association for Computational Linguistics.

Ke Wu, Xiaolin Wang, and Bao-Liang Lu. 2008. Cross language text categorization using a bilingual lexicon. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 165–172.