Effect of Data Imbalance on Unsupervised Domain Adaptation of Part-of-Speech Tagging and Pivot Selection Strategies

Xia Cui XIA.CUI@LIVERPOOL.AC.UK
Frans Coenen COENEN@LIVERPOOL.AC.UK
Danushka Bollegala DANUSHKA@LIVERPOOL.AC.UK
Department of Computer Science, University of Liverpool, Ashton Street, Liverpool L69 3BX, UK

Editors: Luís Torgo, Bartosz Krawczyk, Paula Branco and Nuno Moniz.

Abstract

Domain adaptation is the task of transforming a model trained using data from a source domain to a different target domain. In Unsupervised Domain Adaptation (UDA), we do not assume any labelled training data from the target domain. In this paper, we consider the problem of UDA in the contact of Part-of-Speech (POS). Specifically, we study the effect of data imbalance on UDA of POS, and compare different pivot selection strategies for accurately adapting a POS tagger trained using some source domain data to a target domain. We propose the use of F-score to select pivots using available labelled data in the source domain. Our experimental results on using benchmark dataset for cross-domain POS tagging, show that using frequency combined with F-scores for selecting pivots in the source labelled data produces the best results.
Keywords: Domain Adaptation, Data Imbalance, Part-of-Speech Tagging, Pivot Selection

1. Introduction

In many real-world applications involving machine learning methods we frequently encounter two important problems: (a) the training and testing data distributions being different (*data mismatch*) (Blitzer et al., 2006, 2007; Ben-David et al., 2009), and (b) large discrepancy in terms of the amount of training data available for the different target classes we would like to learn (*data imbalance*) (Provost, 2000; Guo and Viktor, 2004; Zheng et al., 2004).

A popular solution to the first problem is Domain Adaptation (DA). DA considers the problem of adapting a machine learning model from a *source* domain towards a different *target* domain. For example, we would like to train a sentiment classifier for classifying the sentiment on *iPads*. Let us further assume that we do not have any labelled training data expressing user sentiment associated with *iPads*. However, we might have some labelled training data expressing user sentiment on *iPhones*. Considering that *iPhones* and *iPads* have some resemblance in terms of their functionalities, we might be able to first use the available labelled data for *iPhones* and train a sentiment classifier. We could then *adapt* this trained iPhone sentiment classifier to classify the user reviews of *iPads*. In this example, we assumed the availability of unlabelled data for both *iPhone* source domain and *iPad* target domain, whereas labelled training instances were available only for the source domain. This particular DA setting is referred to as Unsupervised Domain Adaptation (UDA) (Daumé III, 2007). In contrast, if we had at least a few labelled training instances for the target domain, in addition to the labelled training instances we have for the source domain, then it is referred to as Supervised Domain Adaptation (SDA) (Daumé III et al., 2010). UDA is particularly challenging compared to SDA because of the lack of labelled training data for the target domain. In this paper we consider the effect of the distribution of the source domain's labelled data on UDA.

Data imbalance arises when we have unequal numbers of training instances for the different target classes we would like to learn (Chawla et al., 2004; Branco et al., 2016; Krawczyk, 2016). For example, in a sentiment classification setting, we might have a disproportionately large amount of positively labelled data to negatively labelled data. If we simply mix all available data and train a classifier, it might be incorrectly biased towards predicting the positive label by default. Under or oversampling methods that respectively select a subset of training instances from the majority class or take multiple samples from the minority class have been proposed to overcome data imbalance issues in machine learning (He and Garcia, 2009).

We study cross-domain part-of-speech (POS) tagging (Schnabel and Schütze, 2013; Schanbel and Schütze, 2014) in which we can encounter both the data mismatch and data imbalance problems discussed above. POS tagging is the task of assigning POS categories such as *noun*, *verb*, *adjective*, *adverb*, etc. to each word in a sentence. POS tagging is one of the fundamental steps in most natural language processing (NLP) applications such as dependency parsing, sentiment classification, machine translation and text summarisation. For example, adjectives are known to carry useful information related to the sentiment of a user who has written a review about a product. Consequently, using adjectives as features for training a classifier to predict sentiment has been an effective strategy. In the cross-domain POS setting, we would like to train a POS tagger using data from a source domain and apply the trained POS tagger on a different target domain. For example, we could train a POS tagger using manually annotated Wall Street Journal articles and adapt the learnt POS tagger to tag POS in social media such as tweets. In the UDA of POS taggers we do not assume any POS labelled training data for the target domain.

As we later see in our analysis, the POS distribution of words is highly uneven. Some POS categories such as nouns and adjectives are highly frequent, whereas adverbs are much less frequent. Therefore, when we adapt a POS tagger to a new domain we must take into account the imbalance of training data for the different POS categories. Several heuristic methods have been proposed in prior work on cross-domain POS tagging for selecting pivots as we discuss later in Section 2. However, to the best of our knowledge, prior work on cross-domain POS tagging has largely ignored this data imbalance issue and have focused purely on the adaptation task. In this paper, we study the effect of data imbalance on UDA applied in cross-domain Part-of-Speech (POS) tagging. UDA methods first select a subset of features that are common to both source and target domains, which are referred to as *pivots*. Next, a projection is learnt from the source and target domains to the space spanned by the pivots. The source domain's labelled training data can then be used to learn a POS tagger in this shared pivot space. By using common features as pivots we can reduce the dissimilarity between the two domains, thereby improving the accuracy of POS tagging in the target domain.

Our contributions in this paper can be summarised as follows:

- We compare the effect of previously proposed pivot selection strategies for selecting pivots for UDA of POS tagging under data imbalance. Specifically, we compare frequency (FREQ), mutual information (MI), pointwise mutual information (PMI) and positive pointwise mutual information (PPMI) as heuristics for selecting pivots. These heuristics can be computed either using labelled data or unlabelled data giving rise to two flavours.

- We propose a pivot selection method using the F-score for UDA of POS tagging, aimed at the problem of high imbalance ratio in POS categories. This method prefers categories with lower performance, measured using F-score, when selecting pivots, thereby selecting more pivots to cover low performing categories. We use only labelled data from the source domain training instances when measuring F-scores. In our experiments, we see that the proposed F-score-based pivot selection method indeed improves the POS tagging accuracy of low-performing categories, thereby improving the overall performance.

2. Related Work

Blitzer et al. (2006) propose one Structural Correspondence Learning (SCL) (Blitzer et al., 2006) for adapting a POS tagger from domain to another. SCL uses the frequency of a word in the source and the target domain to determine its appropriateness as a pivot. A word that appears frequently in both the source and the target domain is likely to be independent of the domains and more suitable for domain adaptation. SCL train linear predictors to predict the presence of pivots using other features. These pivot predictors can then be used to predict the probability of a particular pivot in a sentence even if that pivot does not appear in that sentence. In effect, the pivot predictors can be seen as representing a projection from the source (or target) feature spaces to the common pivot space.

In addition to FREQ, various pivot selection strategies for DA have been proposed in the literature such as mutual information (MI), pointwise mutual information (PMI) and positive pointwise mutual information (PPMI). Blitzer et al. (2006) proposed to select features that frequently occurred in the two domains to be pivots for cross-domain POS tagging. Some other strategies were proposed for cross-domain sentiment classification. Blitzer et al. (2007) proposed to select features with higher MI between labels to be pivots. Pan et al. (2010) proposed to select features with lower MI between different domains to be pivots. Bollegala et al. (2015) and Bollegala et al. (2014) proposed to select pivots using PMI and PPMI respectively.

Although we focus on pivot selection strategies for domain adaptation in this paper, we note that there are alternative DA methods that do not require pivot selection. For example, prediction-based lower dimensional word embeddings have been used as features for reducing the mismatch between source and target sentences thereby adapting a POS tagger trained using source domain data to a different target domain (Schanbel and Schütze, 2014). Instance weighting methods emphasise source domain labelled data instances that are similar to the target domain during training (Jiang and Zhai, 2007). Autoencoders have also been used to learn domain-independent feature representations which can then be used for learning a classifier (Ziser and Reichart, 2016). We do not consider these pivotless DA methods in this paper.

3. Pivot Selection for Unsupervised Cross-domain Part-of-Speech Tagging

The POS tag of a word depends on the POS tags of the preceding words; sequence labellers such as hidden markov models (HMMs) and conditional random fields (CRFs) have been successfully used for learning accurate POS taggers (Kudo et al., 2004). However, by encoding structural features, it is possible to obtain comparable performance using sequence labellers as well as classifiers on POS tagging (Keerthi and Sundararajan, 2007). Therefore, in this work we model POS tagging as a multi-class classification problem where for a given word, we must select its correct POS tag from a pre-defined finite set of POS categories. This modelling assumption enables us to straightforwardly extend previously proposed pivot selection methods for cross-domain sentiment classification. However, sentiment classification is often modelled as a binary classification task (positive vs. negative sentiment) whereas POS tagging is a multi-class classification task. For example, the PennTreebank POS tag set contains 36 categories[1].

To extend the pivot selection methods proposed for binary classification tasks (i.e. sentiment classification) to multi-class classification tasks (i.e. POS tagging) we collate all training data for the categories to a single category, except for the POS category of interest. This is similar to building a one vs. rest binary classifier for each POS category. Specifically, the score function $\phi(x, \mathcal{D})$ for a feature x in a set of training instances $\mathcal{D}$ is computed by heuristic pivot selection methods such as: FREQ, MI, PMI and PPMI. The frequency of a feature x in a set of training instances $\mathcal{D}$ is denoted by $\mathrm{FREQ}(x, \mathcal{D})$. The mutual information between a feature x and a set of instances $\mathcal{D}$ is given by:

$$\mathrm{MI}(x, \mathcal{D}) = p(x, \mathcal{D}) \log \left(\frac{p(x, \mathcal{D})}{p(x)p(\mathcal{D})} \right) \tag{1}$$

We use "$*$" to denote the sum over the set of features or sets of instances for all the domains, and compute the probabilities in (1) using the frequency counts as follows:

$$\begin{aligned} p(x, \mathcal{D}) &= \mathrm{FREQ}(x, \mathcal{D})/\mathrm{FREQ}(*, *), \\ p(x) &= \mathrm{FREQ}(x, *)/\mathrm{FREQ}(*, *), \\ p(\mathcal{D}) &= \mathrm{FREQ}(*, \mathcal{D})/\mathrm{FREQ}(*, *) \end{aligned}$$

Similarly, we compute PMI and PPMI by:

$$\mathrm{PMI}(x, \mathcal{D}) = \log \left(\frac{p(x, \mathcal{D})}{p(x)p(\mathcal{D})} \right) \tag{2}$$

$$\mathrm{PPMI}(x, \mathcal{D}) = \max(\mathrm{PMI}(x, \mathcal{D}), 0) \tag{3}$$

1. https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

3.1. Pivot Selection for Unlabelled Data

Unlabelled pivot selection methods use unlabelled data from the source domain and target domain (we use notations $\mathcal{D}_{S_U}$ and $\mathcal{D}_{T_U}$ to denote unlabelled data in the source and the target domains respectively).

For example, FREQ_U can be computed using Eq. (4) for selecting top-ranked features by occurrence in both domains to be pivots. However, for labelled datasets, pivot selection methods are based on the number of classes, hence the selection process is under multi-class settings.

$$x_U = \min(\phi(x, \mathcal{D}_{S_U}), \phi(x, \mathcal{D}_{T_U})) \tag{4}$$

3.2. Pivot Selection for Labelled Data

As described above, we follow the idea of one vs. rest binary classification to select pivots based on each known tag for labelled datasets in the source domain. For each POS tag P in m POS tags, we split the labelled datasets into $\mathcal{D}_{P_+}$ ($\boldsymbol{x}$ is labelled as P) and $\mathcal{D}_{P_-}$ ($\boldsymbol{x}$ is NOT labelled as P), then compute the score $\phi(\mathcal{D}_P)$ for this POS tag as follows:

$$\phi(x, \mathcal{D}_P) = |\phi(x, \mathcal{D}_{P_+}) - \phi(x, \mathcal{D}_{P_-})| \tag{5}$$

$|\cdot|$ is the absolute value and used for measuring the difference between two sets of instances. The score for each feature x is then computed by the sum of scores from all POS categories:

$$x_L = \sum_{i=1}^{m} \phi(x, \mathcal{D}_{P_i}) \tag{6}$$

Under these scoring methods, features with higher score are more likely to be pivots because they occur frequently or they are more associated with labels.

3.3. Effect of the Label Distribution

In the training datasets (Figure 1), there are very popular POS categories (e.g., nouns (NN)) and less popular ones (e.g., symbols (SYM)). However, none of the above-mentioned pivot selection methods take into consideration this imbalance in data when computing the score when selecting a feature as a pivot. A straightforward method to incorporate the distributional information to the pivot selection process is to multiply the score $\phi(x, \mathcal{D}_{P_i})$ of a feature x as a pivot for representing the i-th POS category by the probability q_i of that category, thus:

$$q(x) = \sum_{i=1}^{m} q_i \phi(x, \mathcal{D}_{P_i}) \tag{7}$$

The pivot selection score $q(x)$ of a feature x given by (7) prefers frequent POS categories when selecting pivots.

3.4. Effect of the F-Score

The pivot selection method described in Section 3.3 is agnostic to the individual performance on a particular POS category. As we later see in our experiments (Figure 2), the frequency of a POS category is not correlating with the performance obtained for that category by a POS tagger. In other words, some low-frequent as well as high-frequent POS categories appear to be equally difficult for adapting a POS tagger to. Therefore, we need a pivot selection method that is aware of the performance on POS categories.

For this purpose, we propose a novel pivot selection method that uses F-score. We first train a POS tagger separately for each POS category P_i using a randomly selected sample from the labelled data from the source domain. Next, we evaluate its performance on a randomly selected (different) sample from the source domain. We compute the F-score for this POS tagger on the i-th POS category. Note that we *do not* use any labelled test data from the target domain for this purpose because in UDA we do not have any labelled data for the target domain. Let us denote the F-score for the i-th POS category to be F_i.

We would like to select pivots from POS categories that have low F_i values to encourage adaptation to those categories. We can consider the reciprocal of the F-scores, $1/F_i$ for this purpose. Unfortunately, $1/F_i$ is not a $[0, 1]$ bounded score such as a probability. Therefore, we compute such a bounded score r_i using the softmax function:

$$r_i = \frac{\exp(1/F_i)}{\sum_{j=1}^{N} \exp(1/F_j)} \tag{8}$$

Here, N is the total number of POS categories. Note that for pivot selection purposes it is sufficient to determine the relative ordering of the features according to their scores $r(x)$. Because (8) is monotonically increasing w.r.t. to the reciprocal of the F scores, we can simply use the reciprocal of the F score as r_i in (9) as follows:

$$r(x) = \sum_{i=1}^{m} r_i \phi(x, \mathcal{D}_{P_i}) \tag{9}$$

3.5. Nouns

Nouns is one of the most popular POS categories. In fact, in our datasets nouns are the majority POS category. As a baseline for selecting pivots from the majority category, we propose a score function for pivot selection that prefers features that occur frequently in the noun category. This baseline demonstrates the performance of a pivot selection method that considers only one POS category such as nouns (NN). This score function x_{NN} is defined as the score from only category NN.

$$x_{\mathrm{NN}} = |\phi(x, \mathcal{D}_{\mathrm{NN}+}) - \phi(x, \mathcal{D}_{\mathrm{NN}-})| \tag{10}$$

4. Experiments

To evaluate the different pivot selection methods described in Section 3, we use the selected pivots with SCL to perform cross-domain POS tagging.

4.1. Experimental Data

Following Blitzer et al. (2006), we use the Penn Treebank (Marcus et al., 1993) of the Wall Street Journal (WSJ) section 2-21 as the labelled data, and 100,000 WSJ sentences from

	Source	Target				
Domians	**wsj**	**newsgroups**	**weblogs**	**reviews**	**answers**	**emails**
#sentences	30,060	1,195	1,016	1,906	1,744	2,450
#tokens	731,678	20,651	24,025	28,086	28,823	29,131
#types	35,933	4,924	4,747	4,797	4,370	5,478
OOV	0.0%	23.1%	19.6%	29.5%	27.7%	30.7%

Table 1: Number of sentences, tokens and types in the source and target labelled data. OOV (Out-Of-Vocabulary) is the percentage of types that have not been observed in the source domain (**wsj**) (Petrov and McDonald, 2012).

	Unlabelled				
Domians	**newsgroups**	**weblogs**	**reviews**	**answers**	**emails**
#sentences	1,000,000	524,834	1,965,350	27,274	1,194,173
#tokens	18,424,657	10,356,284	29,289,169	424,299	17,047,731
#types	357,090	166,515	287,575	33,425	221,576

Table 2: Number of sentences, tokens and types in the target unlabelled data after sentence splitting and tokenisation (Petrov and McDonald, 2012).

1988 as unlabelled data in the source domain. Following Schnabel and Schütze (2013), we evaluate on 5 different target domains (newsgroups, weblogs, reviews, answers and emails) from SANCL 2012 shared task (Petrov and McDonald, 2012). The Penn treebank tag annotated Wall Street Journal (**wsj**) is considered as the source domain in all experiments. Table 1 and Table 2 are the statics of the experimental data. All the datasets have been tokenized during pre-processing. Tokens with the occurrence < 5 are removed.

4.2. Training

To train a POS tagger, we model this task as a multi-class classification problem. We represent each training instance (a POS labelled word in a sentence) by a feature vector. For this purpose, we use two types of features: (a) contextual words and (b) embeddings.

Following Schanbel and Schütze (2014), we imply a window of $2l + 1$ for tagging token x to take the contextual words into account:

$$\boldsymbol{x} = \{x_{-l}, x_{-l+1}, \ldots, x_0, \ldots, x_{l-1}, x_l\} \tag{11}$$

In SCL, original features are projected by the binary classifiers θ learnt from pivots and non-pivots (i.e. pivot predictors) after applied singular value decomposition (SVD). These projected features $\theta\boldsymbol{x}$ are influenced by the different sets of pivots selected by the different pivot selection methods. We follow Sapkota et al. (2016) to train the final adaptive classifier f only by projected features to reduce the dimensionality, where $\theta\boldsymbol{x} \in \mathbb{R}^h$.

We use $d = 300$ dimensional GloVe (Pennington et al., 2014) embeddings (trained using 42B tokens from the Common Crawl) as word representations. By applying the window, each word $\boldsymbol{w}$ is defined by:

$$\boldsymbol{w} = \boldsymbol{w}_{-l} \oplus \boldsymbol{w}_{-l+1} \oplus \ldots \oplus \boldsymbol{w}_0 \oplus \ldots \oplus \boldsymbol{w}_{l-1} \oplus \boldsymbol{w}_l \tag{12}$$

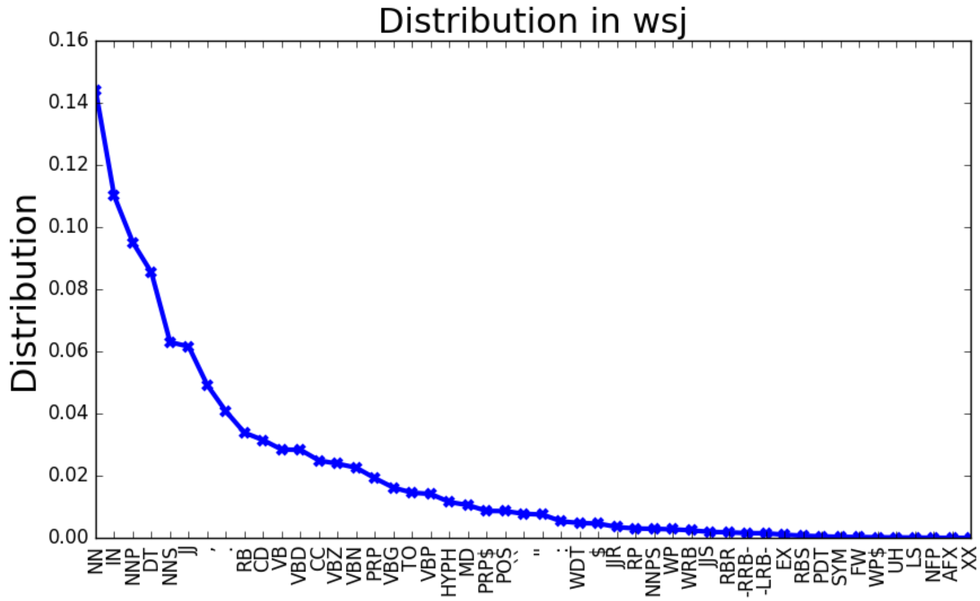


Figure 1: Distribution of the 48 PennTreebank POS tags in training data (**wsj**).

where $\oplus$ is vector concatenation and $\boldsymbol{w} \in \mathbb{R}^d$.

We combine two types of features by introducing a mixing parameter γ, so that adaptive classifier f is trained on $[\gamma\theta\boldsymbol{x}, \boldsymbol{w}]$.

4.3. Classification Accuracy

Accuracy (the percentage of correct predictions) is not a suitable measurement for datasets with large numbers of labels, as it cannot show the effect on imbalanced data from the various labels. Therefore, we use the F-score to measure the classification accuracy for each POS tag when a particular pivot selection strategy is applied to SCL. Here, the F-scores are computed using the target domain's test labelled instances as follows:

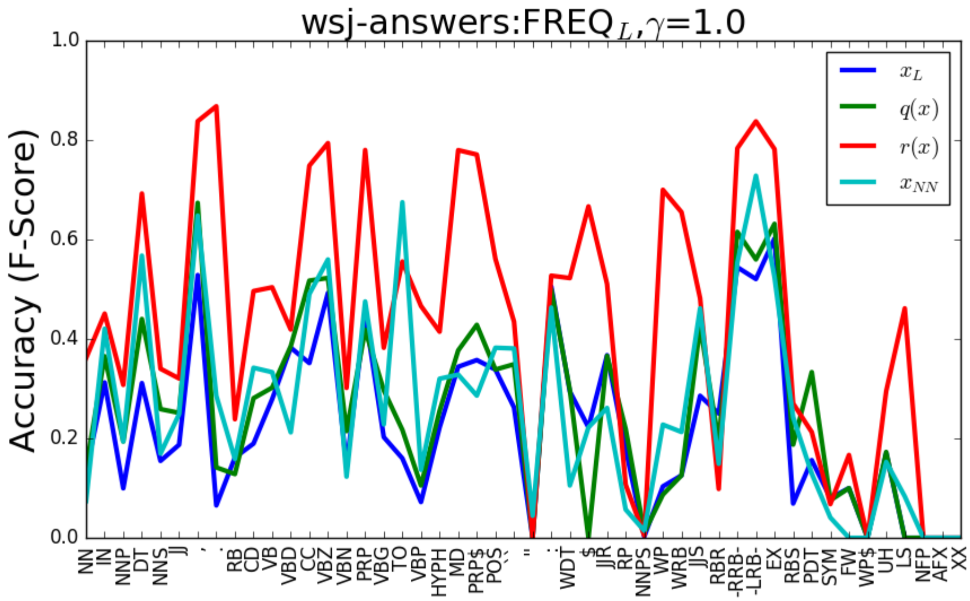
$$\text{Precision}(P_i) = \frac{\text{no. of correctly predicted words as category } P_i}{\text{total no. of test words in the target domain}} \tag{13}$$

$$\text{Recall}(P_i) = \frac{\text{no. of correctly predicted words as category } P_i}{\text{total no. of test words belonging to category } P_i} \tag{14}$$

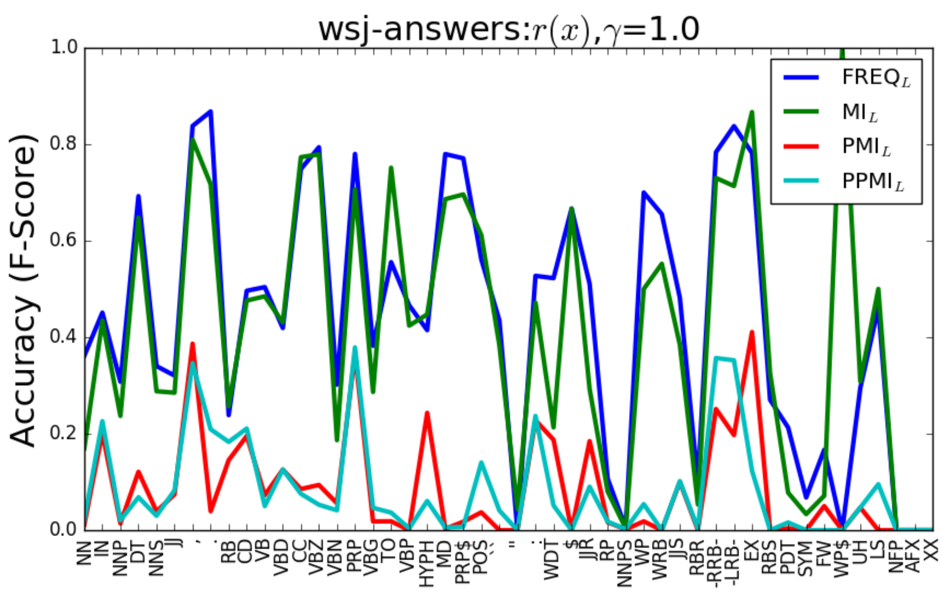
$$\text{F-score}(P_i) = \frac{2 \times \text{Precision}(P_i) \times \text{Recall}(P_i)}{\text{Precision}(P_i) + \text{Recall}(P_i)} \tag{15}$$

5. Results

In the Figure 2(a), we show the F-scores for the different POS tags obtained by adapting a POS tagger from **wsj** source domain to the **answers** target domain. Here, we select pivots using the FREQ_L method. x_L denotes the level of performance we obtain if we had simply used the pivots selected by FREQ_L without adjusting for the imbalance of data. $q(x)$, $r(x)$ and x_{NN} correspond to the pivot selection methods described respectively in Sections 3.3,



(a) Different labelled data strategies using FREQ$_L$.



(b) Different pivot selection methods using $r(x)$.

Figure 2: F-score for the 48 PennTreebank POS tags (left to right: high to low distribution in training data, as shown in Figure 1) for adapting from **wsj** to **answers** under mixing parameter $\gamma = 1.0$.

3.4 and 3.5. The POS tags are arranged in the horizontal axis in the descending order of their frequency in the source domain. The mixing parameter γ is fixed to 1 in this experiment and we later study its effect on the performance.

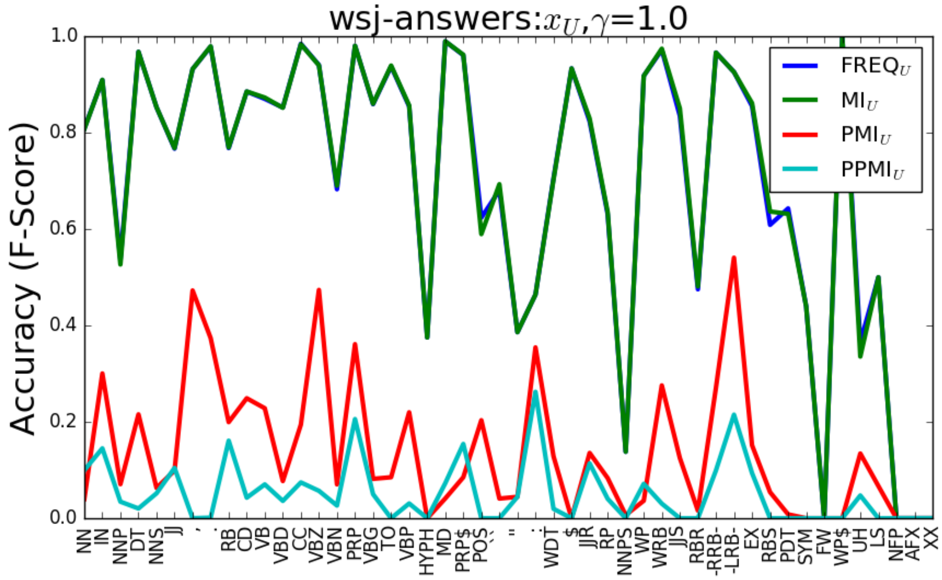


Figure 3: F-score for different pivot selection methods using unlabelled datasets.

Figure 2(*a*) shows that $r(x)$ is the best multi-label strategy for FREQ$_L$. Similar results were obtain when $r(x)$ was combined with other pivot selection methods (MI, PMI and PPMI), and on other target domains. Because of space limitation, we use show the results for the **wsj-answers** adaptation setting. We see that probability of a POS tag ($q(x)$), or selecting pivots from the majority category (x_{NN}), performs at a similar level to not performing any adjustments due to data imbalance (x_L).

Next, we study the effect of the proposed F-score-based pivot selection method, $r(x)$, with different labelled pivot selection methods. Figure 2(*b*) shows that F-score by FREQ is consistently better than others for all labelled methods. Figure 3 shows that FREQ is also one of the good pivot selection methods for unlabelled datasets, MI$_U$ is closely following FREQ$_U$. These two results agree with the observation made by Blitzer et al. (2007) that FREQ works better for POS tagging as a pivot selection strategy. Overall, PMI or PPMI with any multi-class pivot selection strategy proposed in this paper do not work well on datasets with large numbers of categories. A possible reason is that PMI and PPMI do not weight the amount of information obtained about one random event by observing another by the joint probability of the two events (Bollegala et al., 2015).

5.1. Effect on Mixing Parameter

In Section 4.2, we defined a mixing parameter γ for the combination of two types of features. Table 3 shows that all labelled pivot selection methods share the same trend for $\gamma = \{0.01, 0.1, 1, 10, 100\}$. The highest F-score is obtained with 0.01. These F-scores are closer to each other for different pivot selection methods when γ towards zero because we reduce the weight of pivot predictors from SCL and pretrained word embeddings are not influenced by the pivot selection method. All unlabelled pivot selection methods also follow this trend (not shown in Table 3 due to space limitations). The differences between F-scores reported

Method	x_L	$q(x)$	$r(x)$	x_{NN}	x_L	$q(x)$	$r(x)$	x_{NN}
γ	FREQ$_L$				MI$_L$			
0.01	**0.6993**	**0.6982**	**0.6985**	**0.6992**	**0.6986**	**0.6993**	**0.6993**	**0.7006**
0.1	0.6927	0.6910	0.6975	0.6877	0.6857	0.6890	0.6930	0.6977
1	0.2246	0.2604	0.4370	0.2649	0.2407	0.2461	0.3533	0.3689
10	0.4366	0.4328	0.4824	0.4290	0.4317	0.4314	0.4589	0.4725
100	0.6890	0.6909	0.6957	0.6959	0.6900	0.6892	0.6860	0.6931
	PMI$_L$				PPMI$_L$			
0.01	**0.7001**	**0.7025**	**0.6966**	**0.7034**	**0.6996**	**0.6977**	**0.6939**	**0.7002**
0.1	0.5270	0.6775	0.5005	0.5113	0.6992	0.4118	0.5133	0.4666
1	0.1254	0.1492	0.0846	0.1151	0.6955	0.0907	0.0836	0.0956
10	0.3296	0.4359	0.3225	0.3423	0.6811	0.3085	0.3198	0.3236
100	0.6732	0.6906	0.6779	0.6636	0.6609	0.6621	0.6611	0.6839

Table 3: F-score for pivot selection strategies with mixing parameter $\gamma = \{0.01, 0.1, 1, 10, 100\}$. Highest F-score for each strategy is bolded. x_L, $q(x)$, $r(x)$ and x_{NN} denote data imbalance strategies by (6), (7), (9) and (10) respectively.

by the different pivot selection methods with the optimal value of γ for that method are not statistically significant, which indicates that pretrained word embeddings can be used to overcome any disfluencies introduced by the pivot selection methods if the mixing parameter is carefully selected. We differ the study of learning the best combinations of pretrained word embedding-based features and pivot predictors to future work.

6. Conclusion

We compare the effect of previously proposed pivot selection strategies for UDA of POS tagging under data imbalance. We propose a combination of pivot selection method and labelled data strategy (FREQ$_L$ + $r(x)$) that works better than other combinations in the our experiments. We also show that the classification accuracy on a single category does not improve using a single category strategy (e.g. x_{NN}).

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2009.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proc. of EMNLP*, pages 120–128, 2006.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proc. of ACL*, pages 440–447, 2007.

Danushka Bollegala, David Weir, and John Carroll. Learning to predict distributions of words across domains. In *Proc. of ACL*, pages 613 – 623, 2014.

Danushka Bollegala, Tingting Mu, and John Y. Goulermas. Cross-domain sentiment classification using sentiment sensitive embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):398–410, Feb 2015. ISSN 1041-4347.

Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, 49(2):31:1–31:50, August 2016. ISSN 0360-0300.

Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1 – 6, June 2004.

Hal Daumé III. Frustratingly easy domain adaptation. In *Proc. of ACL*, pages 256–263, 2007.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. Frustratingly easy semi-supervised domain adaptation. In *Proc. of the Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59, 2010.

Hongyu Guo and Herna L. Viktor. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *SIGKDD Newsletters*, 6:30 – 39, 2004.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Proc. of ACL*, pages 264 – 271, 2007.

S. Sathiya Keerthi and S. Sundararajan. Crf versus svm-struct for sequence labelling. Technical report, Yahoo Research, 2007.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, Nov 2016.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *EMNLP'04*, 2004.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993. ISSN 0891-2017.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proc. of WWW*, pages 751–760, 2010.

Jeffery Pennington, Richard Socher, and Christopher D. Manning. Glove: global vectors for word representation. In *Proc. of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Slav Petrov and Ryan McDonald. Overview of the 2012 shared task on parsing the web. In *Notes of the first workshop on syntactic analysis of non-canonical language (sancl)*, volume 59, 2012.

Foster Provost. Machine learning from imbalanced data sets. In *AAAI 2000 Workshop on Imbalanced Data Sets*, 2000.

Upendra Sapkota, Thamar Solorio, Manuel Montes-y Gómez, and Steven Bethard. Domain adaptation for authorship attribution: Improved structural correspondence learning. pages 2226—2235, 2016.

Tobias Schanbel and Hinrich Schütze. Flors: Fast and simple domain adaptaton for part-of-speech tagging. *Transactions of Association for Computational Linguistics*, pages 15–26, 2014.

Tobias Schnabel and Hinrich Schütze. Towards robust cross-domain domain adaptation for part-of-speech tagging. In *Proc. of IJCNLP*, pages 198–206, 2013.

Zhaohui Zheng, Xiaoyun Wu, and Rohini Srihari. Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter*, 6(1):80 – 89, June 2004.

Yftah Ziser and Roi Reichart. Neural structural correspondence learning for domain adaptation. *arXiv*, 2016.