Towards Robust Cross-Domain Domain Adaptation for Part-of-Speech Tagging

Tobias Schnabel
CIS, University of Munich
tbs49@cornell.edu

Hinrich Schütze
CIS, University of Munich
inquiries@cislmu.org

Abstract

We investigate the robustness of domain adaptation (DA) representations and methods across target domains using part-of-speech (POS) tagging as a case study. We find that there is no single representation and method that works equally well for all target domains. In particular, there are large differences between target domains that are more similar to the source domain and those that are less similar.

1 Introduction

Domain adaptation (DA) is the problem of adapting a statistical classifier that was trained on a source domain (SD) to a target domain (TD) for which no or little training data is available. We present a case study that investigates the robustness of DA across six different TDs for POS tagging. Most prior work on DA has either been on a single TD, on two or more tasks – which results in an experimental setup in which two variables change at the same time, task and TD – or has not systematically investigated how robust different features and different DA approaches are.

The two main information sources in POS tagging are *context* – which POS's are possible in a particular syntactic context – and *lexical bias* – the prior probability distribution of POS's for each word. We address DA for lexical bias in this paper, focusing on unknown words; they are most difficult to handle in DA because no direct information about their possible POS is available in the SD training set. Since typical TDs contain a high percentage of unknown words, a substantial gain in the overall performance can be achieved by improving tagging for these words.

We address a problem setting where – in addition to *labeled SD data* – a large amount of *unlabeled TD data* is available, but *no labeled TD data*. This setting is often called unsupervised domain adaptation (cf. (Daumé III, 2007)).

We make three contributions in this paper. First, we systematically investigate the cross-TD robustness of different representations and methods. We show that there are some elements of DA setups used in the literature that are robust across TDs – e.g., the use of distributional information – but that many others are not, including dimensionality reduction and shape information.

Second, we present an analysis that shows that there are two important factors that influence cross-TD variation: (i) the magnitude of the difference in distributional properties between SD and TD – more similar TDs require other methods than less similar TDs and (ii) the evaluation measures used for performance. Since in unsupervised DA we optimize learning criteria on a SD that can be quite different from the TD, different TD evaluation measures can diverge more in DA than in standard supervised learning settings when comparing learning methods.

Our third contribution is that we show that if we succeed in selecting an appropriate DA method for a TD, then performance improves significantly. We establish baselines for unknown words for the five TDs of the SANCL 2012 shared task and present the best DA results for unknown words on the Penn BioTreebank. Our improvements on this data set (by 10% compared to published results) are largely due to a new DA technique we call training set filtering. We restrict the training set to long words whose distribution is more similar to unknown words than that of words in general.

The next section describes experimental data and setup and Section 3 experimental results. Section 4 presents analysis and discussion. Section 5 reviews related work. Section 6 concludes.

2 Experimental data and setup

Data. Our SD is the Penn Treebank (Marcus et

International Joint Conference on Natural Language Processing, pages 198–206,
Nagoya, Japan, 14-18 October 2013.

al., 1993) of Wall Street Journal (WSJ) text. Following Blitzer et al. (2006), we use sections 2-21 for training. We also use 100,000 WSJ sentences from 1988 as unlabeled data in training.

We evaluate on six different TDs. The first TD is the Penn BioTreebank data set distributed by Blitzer. It consists of development and test sets of 500 sentences each and an unlabeled set of 100,000 sentences of BIO text.

The remaining five TDs (newsgroups, weblogs, reviews, answers, emails) are from the SANCL shared task (Petrov and McDonald, 2012). We will use WEB to refer to these five TDs collectively. Each WEB TD has an unlabeled training set of 100,000 sentences and development and test sets of about 1000 labeled sentences each. WEB and BIO tag sets differ slightly; we use them as published without modifications to make our results directly comparable to the benchmarks.

We define the *target domain repository* (TD-R) for a TD as the union of development set and unlabeled data available for that TD. SD+TD-R is the union of the source data (labeled and unlabeled WSJ) and TD-R.

Classification setup. In contrast to most other work on POS DA, we adopt a simple approach of *word classification*. The objects to be classified are words and the classes are the POS's of the SD. The gold label of a word in training is the majority tag in the SD. A prediction for an unknown word is then made by computing its feature representation and applying the learned classifier.

We adopt word classification instead of the more common sequence labeling setup because word classification is much more efficient to train and allows us to run a large number of experiments efficiently. Our experiments demonstrate that word classification accuracies are comparable with or higher than sequence labeling in POS DA for unknown words (cf. Table 2).

We use LIBSVM (Chang and Lin, 2011) to train $\binom{k}{2}$ one-vs-one classifiers on the training set, where k is the number of POS tags in the latter. The SVMs were trained with untuned default parameters; in particular, $C = 1$. For sequence classification, we use CRFSuite (Okazaki, 2007), a Conditional Random Field (CRF) toolkit. Apart from the word features described below, we use the base feature set of Huang and Yates (2009) for CRFs, including features for state, emission and transition probabilities. CRFs are trained until convergence with a limit of 300 iterations.

Features. There are in principle two sources of information to predict the POS of an unknown word in an unsupervised setting: the word itself (sequence of letters, shape etc) and the context(s) in which it occurs. For syntactic categorization, the immediate left and right neighbors of a word are the most informative aspect of context. Based on this reasoning, we create a feature representation for each word that has three components: left context information, right context information and shape information. We will refer to left/right context information as distributional information. Let f be the function that maps a word w to its (full) feature vector. We then define f as follows:

$$f(w) = \begin{bmatrix} f_{\text{left}}(w) \\ f_{\text{right}}(w) \\ f_{\text{shape}}(w) \end{bmatrix}$$

Based on the intuition that each of the three sources of information is equally important, each of the three component vectors is normalized to unit length.

For both distributional and shape features, we have a choice of either using *all possible features or a subset consisting of the most frequent features*. We directly compare these two possibilities, using recommended values from the literature for the subset condition: the 250 most frequent features (indicator words) for distributional vectors (Schütze, 1995) and the 100 most frequent features (suffixes) for shape vectors (Müller et al., 2012). Each component vector has an additional binary feature that is set to 1 if the rest of the vector is zero, and 0 otherwise to avoid numerical issues with zero vectors.

Distributional features. The i^{th} entry x_i of $f_{\text{left}}(w)$ is the number of times that the *indicator word* t_i occurs immediately to the left of w:

$$x_i = \text{freq}\left(\text{bigram}(t_i, w)\right)$$

where t_i is the word with frequency rank i in the corpus. $f_{\text{right}}(w)$ is defined analogously.

Many different ways of defining and transforming distributional features have been proposed in the literature. We systematically investigate the following variables: (i) weighting (ii) dimensionality reduction and (iii) selection of data that distributional vectors are based on.

We experiment with three different *weighting functions* that transform non-zero counts as follows. (i) tf: $w_{\text{tf}}(x) = 1 + \log(x)$, (ii) tf-idf:

$w_{\text{tf-idf}}(x) = (N/\log \text{df}_{t_i})(1 + \log(x))$ (where N is the total number of words and df_{t_i} the number of words that indicator word t_i is a non-zero feature of) and (iii) binary: $w_{\text{bin}}(x) = 1$.

Transformation operations like *dimensionality reduction* (Deerwester et al., 1990) can be effective in improving generalization in machine learning, in particular in nonstandard settings like DA where a labeled random sample of the TD is not available. We test singular value decomposition (SVD) here because it has been used in prior work on POS (Huang and Yates, 2009). We apply SVD to the matrix of all feature vectors and keep the dimensions corresponding to the $d = 100$ largest singular values.

We compute distributional vectors either on target data only (i.e., on TD-R) or on the union of source and target data (i.e., SD+TD-R). We compare these two alternatives and show in our experiments that SD distributional information does not consistently improve performance.

Shape features. Suffixes are likely to be helpful because regular processes of inflectional and derivational morphology do not change in English when going from one domain to the next. Many POS taggers incorporate information from suffixes to build robust features (Miller et al., 2007). For a selected suffix s, we simply set the dimension corresponding to s in $f_{\text{shape}}(w)$ to 1 if w ends in s and to 0 otherwise. We either select all suffixes or the top 100, depending on the experiment.

In addition to suffixes, we investigate two other representational variables related to shape: case and digits. For case, we compare keeping case information as is with converting all uppercase characters to lowercase characters. For digits, we compare keeping digits as is with converting all digits to the digit 0; e.g., \$1,643 is converted to \$0,000). We call these two transformations *case normalization* and *digit normalization*.

Training set filtering. The key challenge in DA is that the distributions of source and target are different. One simple trick we can apply to make the distributions more similar is to eliminate all short words from the training set. We call this (training set) *filtering*. The reason this is promising is that longer words are more likely to be examples of productive linguistic processes than short words – even if this is only a statistical tendency with many exceptions.

In future work, we would like to test other filtering options that are based on similar principles, including filtering based on word frequency and open/closed tag classes. Filtering on word length is simple and we show below that it is able to improve accuracy by several percentage points on one TD.

3 Experimental results

We train $\binom{k}{2}$ binary SVM classifiers on the feature representations we just defined. The training set consists of all words that occur in the WSJ training set (in condition SD+TD-R) or all words that occur in both the WSJ training set and TD-R (in condition TD-R). An unknown word is classified by building its feature vector, running the classifiers on it and then assigning it to the POS class returned by the LIBSVM one-vs-one setup.

We divide our experiments into two parts. In the *basic experiment*, we investigate four parameters of the model that are likely to interact with each other: dimensionality of shape vectors (ALL vs. 100 most frequent suffixes), dimensionality of distributional vectors (ALL vs. 250 most frequent indicator words), use of dimensionality reduction (SVD: yes or no) and weighting of distributional vectors (bin, tf, tf-idf).

In the *extended experiment*, we then investigate the effect of other parameters on the best performing model from the basic experiment: distributional vectors based on SD+TD-R vs TD-R, case normalization, digit normalization, completely omitting either shape or distributional information and training set filtering. For the basic experiment, these parameters are set to the following values: distributional vectors are computed on TD-R, case normalization is used, digit normalization is not used, and the training set is not filtered (i.e., all words are included in the training set).

Basic experiment. Table 1 gives the results of the basic experiment: the 24 possible combinations of number of shape features, number of distributional features, use of dimensionality reduction and weighting scheme. In each column, the best three accuracies are underlined and the best accuracy is doubly underlined; the results significantly different from the best result are marked with a dagger.[1]

The goal of the basic experiment is to exhaus-

[1] $p < .05$, 2-sample test for equality of proportions with continuity correction. We use the same test and level for all significance results in this paper.

	shape	dist	svd	wght	grp	rev	blog	ans'r	em'l	BIO
1	100	250	n	bin	56.88	63.92	67.13 †	52.14	63.30	65.64†
2				tf	56.50	65.67	70.33	52.47	64.37	63.14†
3				tf-idf	57.14	65.83	70.23	51.86 †	64.14	64.94†
4			y	bin	52.52†	54.68†	62.74 †	47.81 †	60.08 †	70.29†
5				tf	54.42	58.18†	68.01 †	48.14 †	61.70 †	69.70†
6				tf-idf	54.73	57.44†	68.75 †	48.93 †	61.38 †	70.95†
7		ALL	n	bin	55.98	63.60	68.70 †	52.14	62.87	68.92†
8				tf	56.58	64.67	70.82	51.02 †	63.52	65.72†
9				tf-idf	56.15	63.50	68.85 †	50.09 †	61.87 †	68.61†
10			y	bin	52.05†	52.82†	60.67 †	41.95 †	59.82 †	68.57†
11				tf	53.65†	57.23†	66.24 †	43.02 †	61.22 †	69.82†
12				tf-idf	54.21†	55.47†	64.17 †	42.50 †	58.52 †	69.11†
13	ALL	250	n	bin	56.02	65.04	70.77	54.05	64.37	68.45†
14				tf	55.59	66.05	72.45	55.03	64.43	64.82†
15				tf-idf	55.93	65.99	72.10	54.98	63.98	65.87†
16			y	bin	52.48†	56.16†	65.50 †	43.48 †	59.79 †	70.64†
17				tf	53.26†	59.46†	68.95 †	48.51 †	60.60 †	68.68†
18				tf-idf	54.16†	59.56†	68.70 †	44.18 †	60.66 †	69.35†
19		ALL	n	bin	56.06	63.55	68.85 †	54.38	59.85 †	66.22†
20				tf	56.62	64.61	71.86	54.28	61.05 †	65.64†
21				tf-idf	56.15	63.07	69.74	52.65	59.95 †	65.25†
22			y	bin	52.35†	55.74†	62.89 †	41.95 †	58.68 †	71.07†
23				tf	53.99†	59.83†	68.16 †	43.62 †	60.37 †	69.93†
24				tf-idf	54.81	58.98†	68.65 †	41.95 †	58.68 †	74.39

Table 1: Accuracy of unknown word classification in the basic experiment. The performance of the best (three best) parameter combinations per column are doubly (singly) underlined. A dagger indicates a result significantly worse than the column's best result.

tively investigate combinations of the four parameters that we suspect to have the strongest interaction with each other and then find a parameter combination that is a good basis for testing the remaining parameters in the extended experiment. The guiding principle in this investigation is that when in doubt, we select the simpler or default setting for the extended experiment in order to make as few assumptions as possible.

For the number of shape features, ALL generally does better than 100. Five TDs have their best result for ALL: rev, blog, answer, email (line 14) and BIO (line 24). The exception is grp (best result on line 3). The reason seems to be that the newsgroups TD contains a larger number of unknown words with suffixes that do not support POS generalization well. E.g., the suffixes -ding, -eding, -eeding, -breeding of a newsgroup name like "alt.animals.horses.breeding" (mistagged as VBG, gold tag: NN) are misleading. Despite these problems, the best 100 result for newsgroups is not significantly better than the best ALL result (lines 3 vs. 20). This argues for using the setting ALL for the extended experiment.

For the number of distributional features, there is a similar tendency for the WEB TDs (grp, rev, blog, answer, email) to do slightly better for fewer features (250) than ALL features. However, BIO clearly benefits from using the full dimensionality of the distributional feature space: all 250 results are statistically worse than the best ALL result and the gap to the best 250 result is large (line 24 vs line 6, a difference of $74.39 - 70.95 = 3.44$). The gap between best 250 result and best ALL result is smaller for the other five TDs (although only slightly smaller for email) and for each of the five TDs there is an ALL result that is statistically indistinguishable from the best 250 result. For this reason, we choose dist=ALL for the extended experiment. Simply using ALL indicator words also has the advantage of eliminating the need to optimize an additional parameter, the number of indicator words selected.

In a way similar to distributional features, the behaviors of WEB and BIO TDs also diverge for dimensionality reduction. The top three results for the WEB TDs are always achieved without SVD (lines 1, 3, 13, 14, 15, 19, 20), the top three results for the BIO TD are all SVD results (lines 6, 22, 24). We opt for the simpler option (no SVD) for the extended experiment in the absence of strong consistent cross-TD evidence for the need of dimensionality reduction. We will also see in the extended experiment that we can recover and surpass the best BIO result (74.39, line 24) by optimizing other parameters.

The results on weighting argue against using binary weighting: the six best results in the table all use tf weighting, either by itself or in conjunction with idf (lines 3, 14, 24). Apparently, the distinction between lower and higher frequencies of indicator word occurrences is beneficial for unknown word classification. Whether tf or tf-idf is better, is less clear. For two TDs, tf-idf yields the best result (grp on line 3, BIO on line 24), for four TDs tf (rev, blog, answer, email: line 14). The difference between best tf-idf and best tf result is not significant for grp; we will get tf results for BIO that are better than the best tf-idf result of 74.39 in Table 1. For this reason, we choose the setting tf for the extended experiment. Again, we are selecting the simpler of two options (tf vs tf-idf) when faced with somewhat mixed evidence.

In summary, based on the results of the base experiment, we choose the following settings for the extended experiment: shape = ALL, dist = ALL, svd = n, wght = tf. For shape, dist, and svd this is the simpler of two possible settings. For

weighting, we choose tf (instead of the simpler binary option) because of clear evidence that some form of frequency weighting is beneficial across TDs. These settings correspond to line 20 in Table 1. This line is repeated as the baseline on line 1 in Table 2. Admittedly, choosing this as a baseline setting is somewhat arbitrary as one could always weigh the optimization criteria – peak performance, robustness, simplicity – differently.

	grp	rev	blog	ans'r	em'l	BIO
1 baseline	56.62	64.61	71.86	54.28	61.05†	65.64†
2 CRF	58.18	64.51	70.48	56.52	63.10	56.62†
3 SD+TD-R	55.50	64.13	72.50	55.31	62.91	65.17†
4 no case NRM	52.83†	64.45	70.68	52.00†	59.27†	67.51†
5 digit NRM	56.80	64.61	72.01	54.05	63.88	68.61†
6 shape only, ALL	48.77†	45.32†	56.58†	39.90†	49.19†	52.52†
7 shape only, 100	47.69†	39.16†	51.90†	36.17†	47.24†	50.14†
8 dist only, ALL	52.05†	63.34	68.21†	47.07†	53.06†	73.41†
9 dist only, 250	51.49†	64.13	66.34†	45.76†	54.13†	72.86†
10 $\|w\| > 1$	56.58	64.67	71.81	54.84	60.83†	65.99†
11 $\|w\| > 2$	57.06	64.61	71.56	54.38	63.17	68.61†
12 $\|w\| > 3$	55.33	60.89†	69.69	48.79†	62.39	73.84†
13 $\|w\| > 4$	52.87†	60.10†	67.67†	47.53†	53.06†	77.66
14 $\|w\| > 5$	53.09†	59.35†	66.58†	44.37†	51.69†	77.66
15 $\|w\| > 6$	52.27†	58.55†	66.93†	43.25†	49.74†	77.74
16 $\|w\| > 7$	51.96†	56.64†	63.18†	40.46†	47.17†	78.41
17 $\|w\| > 8$	49.59†	56.16†	58.26†	39.06†	44.31†	79.77
18 $\|w\| > 9$	46.87†	52.82†	55.54†	33.94†	42.69†	74.58†
19 $\|w\| > 10$	43.42†	51.22†	52.54†	33.33†	39.24†	76.10†

Table 2: Extended experiment. The effect of various parameter changes on accuracy of unknown word classification. "NRM" = "normalization.

Extended experiment. In the extended experiment, we investigate the effect of additional parameters. Results are shown in Table 2. Underlining conventions and statistical test setup are identical to Table 1. The CRF baseline used a parameter setting similar to word classification with two exceptions: we set dist=250 because we were not able to run dist=ALL due to memory limitations; and we convert all features to binary due to space restrictions.

Using sequence classification instead of word classification for unknown word prediction does not consistently improve results (line 2). For grp and answer, the CRF achieves the best overall accuracy, but the difference to the baseline is not significant. For the other four TDs, the best result occurs in a different parameter setting. For BIO, a large drop in performance occurs (from 65.64 to 56.62), perhaps suggesting that word classification is more robust than sequence classification for unknown words.

Calculating distributional vectors on both source and target (as opposed to target only) has similarly inconsistent effects (line 3). Performance compared to the baseline decreases for four TDs and increases for two. Based on this evidence, SD distributional information is not robust cross-TD and should probably not be used.

Omitting case normalization (line 4) consistently hurts for WEB TDs, but helps for BIO. In other words, for BIO it is better not to case-normalize words. This result is plausible because case conventions vary considerably in different TDs. Whether keeping case distinctions is helpful or not depends on how similar source and target are in this respect and is therefore not stable in its effect across TDs.

Digit normalization (line 5) has a minor positive or negative effect on the first four TDs, but increases accuracy by more than 2% in the last two, email and BIO. The makeup of the WSJ tag set makes it unlikely that differences between digits could result in POS differences that are predictable in unsupervised DA. This argues for using digit normalization when WSJ is the SD.

The clearest result of the table is that distributional information is necessary for good performance. Performance compared to the baseline drops in all cases and all accuracies on lines 6&7 are significantly worse than the best result. Moreover, distributional features seem to encode more meaningful information for POS tagging than shape features; results on lines 6&7 are consistently lower than results on lines 8&9.

The evaluation is similarly consistent for shape information in the WEB TDs (lines 8 and 9). All accuracies are below the baseline, with some of the drops being quite large, e.g., about 7% for answer and email. Surprisingly, omitting shape information results in a large *increase* of accuracy for the BIO TD. We will further investigate this puzzling result below.

Finally, training set filtering – only training the classifier on words above a threshold length k – is beneficial for all TDs except for blog; and even for blog, moderate filtering has only a negligible negative effect on accuracy (lines 10–11). In principle, the idea of restricting training to longer words because they are most likely to be representative of unknown words seems to be a good one. However, the effect of filtering is sensitive to the threshold length k. We leave it to future work to find properties of the TD that could be used as diagnostics for finding a good value for k.

The motivation of splitting the experiments into

basic experiment and extended experiments was to find a stable point in parameter space for the parameters that are most likely to interact and then look at the effect of the remaining parameters using this stable point as starting point. In Table 2, we see that for the WEB TDs, all variations of experimental conditions either hurt performance or produce only small positive changes in accuracy in comparison to the baseline. This is evidence that our strategy of splitting experiments into basic and extended was sound for these TDs.

		BIO
1	baseline	73.41†
3	SD+TD-R	67.94†
4	no case NRM	72.39†
5	digit NRM	74.15†
10	$\|w\| > 1$	73.96†
11	$\|w\| > 2$	75.24†
12	$\|w\| > 3$	81.30†
13	$\|w\| > 4$	81.88†
14	$\|w\| > 5$	82.98
15	$\|w\| > 6$	82.47
16	$\|w\| > 7$	84.46
17	$\|w\| > 8$	83.09
18	$\|w\| > 9$	79.03†
19	$\|w\| > 10$	80.52†

Table 3: Extended experiment for BIO without shape information. Dist=ALL.

However, the situation for BIO is different. Two parameter changes result in large performance gains for BIO: omitting shape information (increase by 8%, lines 1 vs 8) and filtering out short training words (increase by 14%, lines 1 vs 17). This indicates that the base configuration of the extended experiment is not a good starting point for exploring parameter variation for BIO.

For this reason, we repeat parts of the extended experiment without any shape information. As we would expect, we obtain results for WEB TDs that are consistently worse than those in Table 2 (not shown), with one exception: a slight increase for $|w| > 8$ in email. However, the results for BIO are much improved as shown in Table 3.

To conclude, we found that shape information is helpful for the WEB TDs, but it decreases performance by about 10% for BIO. We will analyze the reason for this discrepancy in the next section.

As a last set of experiments, we run the optimal parameter combination ($|w| > 7$ in Table 3, 84.46) on the BIO test set and obtained an accuracy of 88.13. This is more than 10% higher than the best number for unknown word prediction on BIO published up to this point (76.3 by Huang and Yates (2010)). For the experimental conditions with the best WEB results in Table 2 (double underlining), we get the following test accuracies: grp=56.66, rev=67.79, blog=64.80, answer=66.51, email=65.51. These are either better than dev or slightly worse except for blog; the blog result can be explained by the fact that the blog base model (line 1) also is a lot worse on test than on dev (66.08 vs 71.86). We interpret these test set results as indicating that we did not overfit to the development set in our experiments.

Summary. We have investigated the cross-TD robustness of a number of configurational choices in DA for POS tagging. Based on our results, the following choices are relatively robust across TDs: using ALL indicator words (as opposed to a subset) for distributional features, no dimensionality reduction, tf weighting, digit normalization, target-only distributional features, and formalization of the problem of unknown word prediction as word classification (as opposed to sequence classification).

We found other choices to be dependent on the TD, in particular the use of shape features, case normalization and training set filtering.

The most important lesson from these results is that many aspects of DA are highly dependent on the TD. Given our results, it is unlikely that a single DA setup will work in general. Instead, criteria need to be developed that allow us to predict which features and methods work for different TDs.

4 Analysis and discussion

The biggest TD differences we found in the experiments are those between WEB and BIO: they behave differently with respect to dimensionality reduction (bad for WEB, good for BIO), shape information (good for WEB, bad for BIO) and sequence classification (neutral for WEB, bad for BIO).

One hypothesis that could explain these results is that the difference between BIO and WSJ is larger than the difference between WEB and WSJ. For example, dimensionality reduction creates more generalized representations, which may be appropriate for TDs with large source-target differences like BIO; and WSJ suffixes may not be helpful for BIO because biomedical terminology has suffixes specific to scientific vocabulary and is rare in newspaper text. In contrast, WEB suffixes may not diverge as much from WSJ since both are "non-technical" genres.

One way to assess the difference between two

TD	tags	suffixes	transitions
grp	.009	.275	.068
rev	.057	.352	.212
blog	.009	.295	.074
answer	.048	.337	.158
email	.036	.273	.139
BIO	.096	.496	.385

Table 4: JS divergences between WSJ and TDs.

		grp	rev	blog	ans'r	em'l	BIO
1	baseline	32.77	38.89	43.48	30.52	34.26	40.06
2	CRF	38.74	42.71	46.63	38.08	36.21	39.03
3	SD+TD-R	32.87	38.55	44.75	33.19	35.30	41.42
4	no case NRM	27.08	39.82	39.54	25.80	27.33	39.98
5	digit NRM	32.80	39.09	43.68	30.47	34.69	37.72
6	shape only, ALL	18.02	21.25	24.61	16.25	16.37	26.55
8	dist only, ALL	27.70	38.39	34.38	22.11	29.71	37.01
10	$\|w\| > 1$	32.73	39.48	43.54	30.60	34.20	35.32
11	$\|w\| > 2$	33.33	37.38	43.52	30.02	34.66	35.05
13	$\|w\| > 4$	26.37	28.92	37.68	22.33	24.14	37.55

Table 5: Selected conditions of the extended experiment (Table 2), evaluated using macroaveraged F_1.

domains is to compare various characteristic probability distributions. The distance of two domains under a representation R has been shown to be important for DA (Ben-David et al., 2007). Similar to Huang and Yates (2010), we use Jensen-Shannon (JS) divergence as a measure of divergence. Table 4 shows the JS divergences between WSJ and the six TDs for different distributions.

The results confirm our hypothesis. BIO is indeed more different from WSJ than the other TDs. Tag distribution divergence is 0.096 for BIO and ranges from 0.009 to 0.057 for WEB. Suffix distribution divergence of BIO is 0.496, almost 50% more than rev, the WEB TD with highest suffix divergence. The underlying probability distributions here are $P(\text{suffix}|t)$, where $t \in \{\text{NN, NNP, JJ}\}$ – most unknown words are in these three classes and accuracy is therefore mostly a measure of accuracy on NN, NNP and JJ. Finally, transition probability divergence of BIO for NN, NNP, JJ is also much larger than for WEB. The distribution investigated here is $P(t_{i-1}|t_i)$; we compute the divergence between, say, BIO and WSJ for the three tags and then average the three divergences.

We do not have space to show detailed results on all tags, but the divergences are more similar for closed class POS. E.g., there is virtually no difference in transition probability divergence for modals between BIO and WEB. This observation prompted us to investigate whether some TD differences might depend on the evaluation measure used. Accuracy – a type of microaveraging – is mostly an evaluation of the classes that are frequent for unknown words: NN, NNP, JJ. If most of the higher divergence of BIO is caused by these categories, then a macroaveraged evaluation, which gives equal weight to each POS tag, should show less divergence.

This is indeed the case as the macroaveraged results in Table 4 show. These results are more consistent across TDs than those evaluated with accuracy. Removing shape and distributional information now hurts performance for all TDs (lines 6&8). WEB and BIO behave more similarly with respect to training set filtering: the large outliers for BIO we obtained in the accuracy evaluation are gone. SD distributional information has a more beneficial effect on F_1 than on accuracy, probably because the classification of POS that are more stable across TDs like verbs and adverbs benefits from SD information. The CRF produces the best result for all WEB TDs. For less frequent POS classes (those that dominate the macroaveraged measure, especially verbal POS), sequence information and "long-distance" context is probably more stable and can be exploited better than for NN, NNP and JJ. However, there is still a drop-off from the baseline for BIO; we attribute this to the larger differences in the transition probabilities for BIO vs WEB (Table 4); the sequence classifier is at a disadvantage for BIO, even on a macroaveraged measure, because the transition probabilities change a lot.

It is important to note that even though F_1 results are more consistent for DA, accuracy is the appropriate measure to use for POS tagging: the usefulness of a tagger to downstream components in the processing pipeline is better assessed by accuracy than by F_1.

5 Related work

Most work on POS tagging takes a standard supervised approach and assumes that source and target are the same (e.g., (Toutanova et al., 2003)). At the other end of the spectrum is the unsupervised setting (e.g., (Schütze, 1995; Goldwater and Griffiths, 2007)). Other researchers have addressed the task of adapting a known tagging dictionary to a TD (e.g., (Merialdo, 1994; Smith and Eisner, 2005)), which we view as complementary to methods for words about whose tags nothing is known. Subramanya et al. (2010) perform DA without using any unlabeled TD text. All of these applica-

tions scenarios are reasonable; however, it can be argued that the scenario we address is – apart from standard supervised learning – perhaps more typical of what occurs in practice: there is labeled SD text available for training; there is plenty of unlabeled TD text available; and there is a substantial number of TD words that do not occur in the SD. Frequently, researchers make the assumption that a small labeled target text has been created (e.g., (Daumé III, 2007)); in the process, a small number of unknown words may also be labeled, but this is not an alternative to handling unknown words in general.

Work by Das and Petrov (2011) is also a form of DA for POS tagging, using universal POS tag sets and parallel corpora. It is likely that best performance for TDs without training data can be achieved by combining our approach with a multilingual approach if appropriate parallel data is available. Ganchev et al. (2012) use another source of additional information, search logs. Again, it should be possible to integrate search-log based features into our framework.

Blitzer et al. (2006) learn correspondences between features in source and target. Our results suggest that completely ignoring source features (and only using source labels) may be a more robust approach for unknown words.

Cholakov et al. (2011) point out that improving tagging accuracy does not necessarily improve the performance of downstream elements of the processing pipeline. However, improved unknown word classification will have a positive impact on most downstream components.

Choi and Palmer (2012) perform DA by training two separate models on the available data, a generalized one and a domain-specific one. During tagging, an input sentence is tagged by the model that is most similar to the sentence. Since their approach is not conditioned on the underlying tagging model, it would be interesting to integrate their approach with ours.

Huang and Yates (2009) evaluate CRFs with distributional features. Besides raw feature vectors, they examine lower dimensional feature representations using SVD or a special HMM-based method. In our experiments, we did not find an advantage to using SVD.

Huang and Yates (2010) use sequence labeling to predict POS of unknown words. Huang and Yates (2012) extend this work by inducing latent states that are shown to improve prediction. As we argued above, a word classification approach has several advantages compared to a sequence labeling approach. Since latent sequence states can be viewed as a form of dimensionality reduction, it would be interesting to compare them to the non-sequence-based dimensionality reduction (SVD) we have investigated in our experiments.

Zhang and Kordoni (2006) use a classification approach for predicting POS for in-domain unknown words. They achieve an accuracy of 61.3%. Due to differences in the data sets used, these results are not directly comparable with ours.

Miller et al. (2007) and Cucerzan and Yarowsky (2000) have both investigated the use of suffixes for DA. Miller et al. characterized words by a list of hand-built suffix classes that they appear in. They then used a 5-NN classifier along with a custom similarity measure to find initial lexical probabilities for all words. We also ran extensive experiments with kNN, but found that one-vs-one SVM performs better.

Cucerzan and Yarowsky (2000) use distribution as a backoff strategy if no helpful suffix information is available. They address unknown word prediction for new languages. We have found that for within-language prediction, distributional information is generally more robust than shape information, including suffixes.

Van Asch and Daelemans (2010) find that DA performance is the higher, the more similar the unigram distribution of the TD is to that of the SD. However, we cannot compute unigram distributions in the case of unknown words.

6 Conclusions and Future Work

In this paper, we have investigated the robustness of DA representations and methods for POS tagging and shown that there are large differences in robustness across TDs that need to be taken into account when performing DA for a TD. We found that the divergence between source and target is an important predictor of what elements of DA will work; e.g., higher divergence makes it more likely that generalization mechanisms like dimensionality reduction will be beneficial.

In future work, we would like to develop statistical measures of source-target divergence that accurately predict whether a feature type or DA technique supports high-performance DA for a particular TD.

References

Shai Ben-David, John Blitzer, Koby Crammer, and Marina Sokolova. 2007. Analysis of representations for domain adaptation. In *NIPS 19*, pages 137–144.

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

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM TIST*, 2(3):27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Jinho D. Choi and Martha Palmer. 2012. Fast and robust part-of-speech tagging using dynamic model selection. In *Proc. of the ACL: Short Papers - Vol. 2*, pages 363–367.

Kostadin Cholakov, Gertjan van Noord, Valia Kordoni, and Yi Zhang. 2011. An empirical comparison of unknown word prediction methods. In *Proc. of the IJCNLP*, pages 767–775.

Silviu Cucerzan and David Yarowsky. 2000. Language independent, minimally supervised induction of lexical probabilities. In *Proc. of the ACL*, pages 270–277.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. of the ACL*, pages 600–609.

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

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, 41(6):391–407.

Kuzman Ganchev, Keith Hall, Ryan McDonald, and Slav Petrov. 2012. Using search-logs to improve query tagging. In *Proc. of the ACL: Short Papers - Vol. 2*, pages 238–242.

Sharon Goldwater and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proc. of the ACL*, pages 744–751.

Fei Huang and Alexander Yates. 2009. Distributional representations for handling sparsity in supervised sequence-labeling. In *Proc. of the Joint Conf. of the ACL and the IJCNLP*, pages 495–503.

Fei Huang and Alexander Yates. 2010. Exploring representation-learning approaches to domain adaptation. In *Proc. of the DANLP Workshop*, pages 23–30.

Fei Huang and Alexander Yates. 2012. Biased representation learning for domain adaptation. In *Proc. of the EMNLP-CoNLL*, pages 1313–1323.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The Penn treebank. *Comp. Linguistics*, 19(2):313–330.

Bernard Merialdo. 1994. Tagging english text with a probabilistic model. *Comp. Linguistics*, 20(2):155–171.

John Miller, Manabu Torii, and Vijay K. Shanker. 2007. Building domain-specific taggers without annotated (domain) data. In *Proc. of the EMNLP-CoNLL*, pages 1103–1111.

Thomas Müller, Hinrich Schütze, and Helmut Schmid. 2012. A comparative investigation of morphological language modeling for the languages of the european union. In *Proc. of the NAACL-HLT*, pages 386–395.

Naoaki Okazaki. 2007. CRFsuite: A fast implementation of conditional random fields (CRFs). Available at: `http://www.chokkan.org/software/crfsuite/`.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 Shared Task on Parsing the Web. Notes of the 1st SANCL Workshop.

Hinrich Schütze. 1995. Distributional part-of-speech tagging. In *Proc. of the EACL*, pages 141–148.

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: training log-linear models on unlabeled data. In *Proc. of the ACL*, pages 354–362.

Amarnag Subramanya, Slav Petrov, and Fernando Pereira. 2010. Efficient graph-based semi-supervised learning of structured tagging models. In *Proc. of the EMNLP*, pages 167–176.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of the NAACL-HLT - Vol. 1*, pages 173–180.

Vincent Van Asch and Walter Daelemans. 2010. Using domain similarity for performance estimation. In *Proc. of the DANLP Workshop*, pages 31–36.

Yi Zhang and Valia Kordoni. 2006. Automated deep lexical acquisition for robust open texts processing. In *Proc. of the LREC*, pages 275–280.