Domain Adaptation for Authorship Attribution: Improved Structural Correspondence Learning

Upendra Sapkota
University of Alabama at Birmingham
upendra@uab.edu

Thamar Solorio
University of Houston
solorio@cs.uh.edu

Manuel Montes-y-Gómez
Instituto Nacional de
Astrofísica, Optica y Electrónica
mmontesg@inaoep.mx

Steven Bethard
University of Alabama at Birmingham
bethard@uab.edu

Abstract

We present the first domain adaptation model for authorship attribution to leverage unlabeled data. The model includes extensions to structural correspondence learning needed to make it appropriate for the task. For example, we propose a median-based classification instead of the standard binary classification used in previous work. Our results show that punctuation-based character n-grams form excellent pivot features. We also show how singular value decomposition plays a critical role in achieving domain adaptation, and that replacing (instead of concatenating) non-pivot features with correspondence features yields better performance.

1 Introduction

Authorship Attribution (AA) can be used for historical purposes, such as disentangling the different authors contributing to a literary work. It can also help in understanding language evolution and change at the individual level, revealing a writer's changes in linguistic patterns over time (Hirst and Feng, 2012). Authorship attribution can also help to settle disputes over the original creators of a given piece of text. Or it can help build a prosecution case against an online abuser, an important application especially considering the rising trends in cyber-bullying and other electronic forms of teen violence[1]. The absorbing social media networks, together with the ever increasing use of electronic communications will require robust approaches to authorship attribution that can help to determine with certainty the author of a text, determine the provenance of a written sample, and in sum, help us determine the trustworthiness of electronic data.

[1] http://cyberbullying.org/

One of the scenarios that has received limited attention is cross-domain authorship attribution, when we need to identify the author of a text but all the text with known authors is from a different topic, genre, or modality. Here we propose to solve the problem of cross-domain authorship attribution by adapting the Structural Correspondence Learning (SCL) algorithm proposed by Blitzer et al. (2006). We make the following contributions:

- We introduce the first domain adaptation model for authorship attribution that combines labeled data in a source domain with unlabeled data from a target domain to improve performance on the target domain.
- We examine two sets of features that have previously been successful in cross-domain authorship attribution, explain how these can be used to select the "pivot" features required by SCL, and show that typed n-gram features (which differentiate between the *the* in ***the**ir* and the *the* in *brea**the***) produce simpler models that are just as accurate.
- We propose a new approach for defining SCL's pivot feature classification task so that it is able to handle count-based features, and show that this median-based approach outperforms the standard SCL approach.
- We examine the importance of the dimensionality reduction step in SCL, and show that the singular value decomposition increases robustness even beyond the robustness achieved by SCL's learned feature transformations.
- We propose an alternative approach to combining features within SCL, and show that excluding the non-pivot features from the final classifier generally improves performance.

Our experimental results show that using standard SCL for this domain adaptation authorship attribution task improves prediction accuracy by

only 1% over a model without any domain adaptation. In contrast, our proposed improvements to SCL reach an accuracy boost of more than 15% over the no domain adaptation model and of 14% over the standard SCL formulation. The extensions to SCL that we propose in this work are likely to yield performance improvements in other tasks where SCL has been successfully applied, such as part-of-speech tagging and sentiment analysis. We plan to investigate this further in the future.

2 Related Work

Cross-Domain Authorship Attribution Almost all previous authorship attribution studies have tackled traditional (single-domain) authorship problems where the distribution of the test data is the same as that of the training data (Madigan et al., 2005; Stamatatos, 2006; Luyckx and Daelemans, 2008; Escalante et al., 2011). However, there are a handful of authorship attribution studies that explore cross-domain authorship attribution scenarios (Mikros and Argiri, 2007; Goldstein-Stewart et al., 2009; Schein et al., 2010; Stamatatos, 2013; Sapkota et al., 2014). Here, following prior work, cross-domain is a cover term for cross-topic, cross-genre, cross-modality, etc., though most work focuses on the cross-topic scenario.

Mikros and Argiri (2007) illustrated that many stylometric variables are actually discriminating topic rather than author. Therefore, the authors suggest their use in authorship attribution should be done with care. However, the study did not attempt to construct authorship attribution models where the source and target domains differ.

Goldstein-Stewart et al. (2009) performed a study on cross-topic authorship attribution by concatenating the texts of an author from different genres on the same topics. Such concatenation allows some cross-topic analysis, but as each test document contains a mix of genres it is not representative of real world authorship attribution problems.

Stamatatos (2013) and Sapkota et al. (2014) explored a wide variety of features, including lexical, stopword, stylistic, and character n-gram, and demonstrated that character n-grams are the most effective features in cross-topic authorship attribution. Stamatatos (2013) concluded that avoiding rare features is effective in both intra-topic and cross-topic authorship attribution by training a SVM classifier on one fixed topic and testing on each of the remaining topics. Sapkota et al. (2014), rather than fixing a single training topic in advance, considered all possible training/testing topic combinations to investigate cross-topic authorship attribution. This showed that training on documents from multiple topics (thematic areas) improves performance in cross-topic authorship attribution (Sapkota et al., 2014), even when controlling the amount of training data.

However, none of these studies exploited domain adaptation methods that combine labeled data in a source domain with unlabeled data from a target domain to improve performance on the target domain. Instead, they focused on identifying relevant features and simply evaluating them when trained on source-domain data and tested on target-domain data. To our knowledge, we are the first to leverage unlabeled data from the target domain to improve authorship attribution.

Domain Adaptation Domain adaptation is the problem of modifying a model trained on data from a source domain to a different, possibly related, target domain. Given the effort and the cost involved in labeling data for a new target domain, there is a lot of interest in the design of domain adaptation techniques. In NLP related tasks, researchers have explored domain adaptation for part-of-speech tagging, parsing, semantic role labeling, word-sense disambiguation, and sentiment analysis (Li, 2012).

Daumé (2007) proposed a feature space transformation method for domain adaptation based on a simple idea of feature augmentation. The basic idea is to create three versions of each feature from the original problem: the general (domain-independent) version, the source specific version, and the target specific version. While generally successful, there are some limitations of this method. First, it requires labeled instances in the target domain. Second, since this method simply duplicates each feature in the source domain as domain-independent and domain-specific versions, it is unable to extract the potential correlations when the features in the two domains are different, but have some hidden correspondences.

In contrast, structural correspondence learning (SCL) is a feature space transformation method that requires no labeled instances from the target domain, and can capture the hidden correlations among different domain-independent features. SCL's basic idea is to use unlabeled data from both the source and target domains to obtain a common feature representation that is meaningful across

domains (Blitzer et al., 2006). Although the distributions of source and target domain differ, the assumption is that there will still be some general features that share similar characteristics in both domains. SCL has been applied to tasks such as sentiment analysis, dependency parsing, and part-of-speech tagging, but has not yet been explored for the problem of authorship attribution.

The common feature representation in SCL is created by learning a projection to "pivot" features from all other features. These pivot features are a critical component of the successful use of SCL, and their selection is something that has to be done carefully and specifically to the task at the hand. Tan and Cheng (2009) studied sentiment analysis, using frequently occurring sentiment words as pivot features. Similarly, Zhang et al. (2010) proposed a simple and efficient method for selecting pivot features in domain adaptive sentiment analysis: choose the frequently occurring words or word-bigrams among domains computed after applying some selection criterion. In dependency parsing, Shimizu and Nakagawa (2007) chose the presence of a preposition, a determiner, or a helping verb between two tokens as the pivot features. For part-of-speech tagging, Blitzer et al. (2006) used words that occur more than 50 times in both domains as the pivot features, resulting in mostly function words. In cross-lingual adaptation using SCL, semantically related pairs of words from source and target domains were used as pivot features (Prettenhofer and Stein, 2011). For authorship attribution, we propose two ways of selecting pivot and non-pivot features based on character n-grams.

Another important aspect of the SCL algorithm is associating a binary classification problem with each pivot feature. The original SCL algorithm assumes that pivot features are binary-valued, so creating a binary classification problem for each pivot feature is trivial: is the value 0 or 1? Most previous work on part-of-speech tagging, sentiment analysis, and dependency parsing also had only binary-valued pivot features. However, for authorship attribution, all features are count-based, so translation from a pivot feature value to a binary classification problem is not trivial. We propose a median-based solution to this problem.

3 Methodology

Structural Correspondence Learning (Blitzer et al., 2006) uses only unlabeled data to find a common feature representation for a source and a target domain. The idea is to first manually identify "pivot" features that are likely to have similar behavior across both domains. SCL then learns a transformation from the remaining non-pivot features into the pivot feature space. The result is a new set of features that are derived from all the non-pivot features, but should be domain independent like the pivot features. A classifier is then trained on the combination of the original and the new features.

Table 1 gives the details of the SCL algorithm. First, for each pivot feature, we train a linear classifier to predict the value of that pivot feature using only the non-pivot features. The weight vectors learned for these linear classifiers, $\hat{w}_i$, are then concatenated into a matrix, W, which represents a projection from non-pivot features to pivot features. Singular value decomposition is used to reduce the dimensionality of the projection matrix, yielding a reduced-dimensionality projection matrix θ. Finally, a classifier is trained on the combination of the original features and the features generated by applying the reduced-dimensionality projection matrix θ to the non-pivot features $\mathbf{x}_{[p:m]}$.

3.1 Standard SCL parameter definitions

Standard SCL does not define how pivot features are selected; this must be done manually for each new task. However, SCL does provide standard definitions for the loss function (L), the conversion to binary values (B_i), the dimensionality of the new correspondence space (d), and the feature combination function (C).

L is defined as Huber's robust loss:

$$L(a, b) = \begin{cases} \max(0, 1 - ab)^2 & \text{if } ab \geq -1 \\ -4ab & \text{otherwise} \end{cases}$$

The conversion from pivot feature values to binary classification is defined as:

$$B_i(y) = \begin{cases} 1 & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

A few different dimensionalities for the reduced feature space have been explored (Prettenhofer and Stein, 2011), but most implementations have followed the standard SCL description (Blitzer et al., 2006) with d defined as:

$$d = 25$$

The feature combination function, C, is defined as simple concatenation, i.e., use all of the old pivot

Input:

- $S = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^m\}$, the labeled instances from source domain
- $U = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^m\}$, the unlabeled instances from both domains
- p and n such that $\mathbf{x}_{[0:p]}$ are the p pivot features and $\mathbf{x}_{[p:m]}$ are $n = m - p$ non-pivot features
- $f : S \rightarrow A$, the source domain labels, where A is the set of authors
- $L : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, a loss function
- $B_i : \mathbb{R} \rightarrow \{0, 1\}$ for $0 \leq i < p$, a conversion from a real-valued pivot feature i to binary classification
- d, the size of the reduced-dimensionality correspondence space to learn
- $C : \mathbb{R}^m \times \mathbb{R}^d \rightarrow \mathbb{R}^k$, a function for combining the original and new features

Output:

- $\theta \in \mathbb{R}^{n \times d}$, a projection from non-pivot features to the correspondence space
- $h : \mathbb{R}^{m+d} \rightarrow A$, the trained predictor

Algorithm:

1. For each pivot feature $i : 0 \leq i < p$, learn prediction weights $\hat{w}_i = \min_{\mathbf{w} \in \mathbb{R}^n} \sum_{\mathbf{x} \in U} L(\mathbf{w}^\top \mathbf{x}_{[p:m]}, B(x_i))$
2. Construct a matrix $W \in \mathbb{R}^{n \times p}$ using each $\hat{w}_i$ as a column
3. Apply singular value decomposition $W = U \Sigma V^\top$ where $U \in \mathbb{R}^{n \times n}$, $\Sigma \in \mathbb{R}^{n \times p}$, $V^\top \in \mathbb{R}^{p \times p}$
4. Select the reduced-dimensionality projection, $\theta = {U_{[0:d,:]}}^\top$
5. Train a classifier h from $\left\{ \left([C(\mathbf{x}, \mathbf{x}_{[p:m]}\theta), f(\mathbf{x}) \right) : \mathbf{x} \in S \right\}$

Table 1: The structural correspondence learning (SCL) algorithm

features, all the old non-pivot features, and all the new correspondence features:

$$C(\mathbf{x}, \mathbf{z}) = [\mathbf{x}; \mathbf{z}]$$

We call this the *pivot+nonpivot+new* setting of C.

The following sections discuss alternative parameter choices for pivot features, B_i, d, and C.

3.2 Pivot Features for Authorship Attribution

The SCL algorithm depends heavily on the pivot features being domain-independent features, and as discussed in Section 2, which features make sense as pivot features varies widely by task. No previous studies have explored structural correspondence learning for authorship attribution, so one of the outstanding questions we tackle here is how to identify pivot features. Research has shown that the most discriminative features in attribution and the most robust features across domains are character n-grams (Stamatatos, 2013; Sapkota et al., 2014). We thus consider two types of character n-grams used in authorship attribution that might make good pivot features.

3.2.1 Untyped Character N-grams

Classical character n-grams are simply the sequences of characters in the text. For example, given the text:

> The structural correspondence

character 3-gram features would look like:

```
"The", "he ", "e s", " st",
"str",  "tru", "ruc", "uct", ...
```

We propose to use as pivot features the p most frequent character n-grams. For non-pivot features, we use the remaining features from prior work (Sapkota et al., 2014). These include both the remaining (lower frequency) character n-grams, as well as stop-words and bag-of-words lexical features. We call this the *untyped* formulation of pivot features.

3.2.2 Typed Character N-grams

Sapkota et al. (2015) showed that classical character n-grams lose some information in merging together instances of n-grams like *the* which could be a prefix (***the**sis*), a suffix (*brea**the***), or a standalone word (***the***). Therefore, untyped character n-grams were separated into ten distinct categories. Four of the ten categories are related to affixes: prefix, suffix, space-prefix, and space-suffix. Three are word-related: whole-word, mid-word, and multi-word. The final three are related to the use of punctuation: beg-punct, mid-punct, and end-punct. For example, the character n-grams from the last section would instead be replaced with:

```
"whole-word:The", "space-suffix:he ",
"multi-word:e s", "space-prefix: st",
"prefix:str",  "mid-word:tru",
"mid-word:ruc", "mid-word:uct", ...
```

Sapkota et al. (2015) demonstrated that n-grams starting with a punctuation character (the *beg-punct* category) and with a punctuation character in the middle (the *mid-punct* category) were the most effective character n-grams for cross-domain authorship attribution. We therefore propose to use as pivot features the $p/2$ most frequent character n-grams from each of the *beg-punct* and *mid-punct* categories, yielding in total p pivot features. For non-pivot features, we use all of the remaining features of Sapkota et al. (2015). These include both the remaining (lower frequency) *beg-punct* and *mid-punct* character n-grams, as well as all of the character n-grams from the remaining eight categories. We call this the *typed* formulation of pivot features.[2]

3.3 Pivot feature binarization parameters

Authorship attribution typically relies on count-based features. However, the classic SCL algorithm assumes that all pivot features are binary, so that it can train binary classifiers to predict pivot feature values from non-pivot features. We propose a binarization function to produce a binary classification problem from a count-based pivot feature by testing whether the feature value is above or below the feature's median value in the training data:

$$B_i(y) = \begin{cases} 1 & \text{if } y > median(\{\mathbf{x}_i : \mathbf{x} \in S \cup U\}) \\ 0 & \text{otherwise} \end{cases}$$

The intuition is that for count-based features, "did this pivot feature appear at least once in the text" is not a very informative distinction, especially since the average document has hundreds of words, and pivot features are common. A more informative distinction is "was this pivot feature used more or less often than usual?" and that corresponds to the below-median vs. above-median classification.

3.4 Dimensionality reduction parameters

The reduced dimensionality (d) of the low-rank representation varies depending on the task at hand, though lower dimensionality may be preferred as it will result in faster run times. We empirically compare different choices for d: 25, 50, and 100.

We also consider the question, how critical is dimensionality reduction? For example, if there are only $p = 100$ pivot features, is there any need to run singular-value decomposition? The goal here is to determine if SCL is increasing the robustness across domains primarily through transforming non-pivot features into pivot-like features, or if the reduced dimensionality from the singular-value decomposition contributes something beyond that.

[2]Because the **untyped** and **typed** feature sets are designed to directly replicate Sapkota et al. (2014) and Sapkota et al. (2015), respectively, both include character n-grams, but only **untyped** includes stop-words and lexical features.

Topics	4
Authors	13
Documents/author/topic	10
Average sentences/document	53
Average words/document	1034

Table 2: Statistics of the Guardian dataset.

3.5 Feature combination parameters

It's not really clear why the standard formulation of SCL uses the non-pivot features when training the final classifier. All of the non-pivot features are projected into the pivot feature space in the form of the new correspondence features, and the pivot feature space is, by design, the most domain independent part of the feature space. Thus, it seems reasonable to completely replace the non-pivot features with the new pivot-like features. We therefore consider a *pivot+new* setting of C:

$$\textit{pivot+new: } C(\mathbf{x}, \mathbf{z}) = [\mathbf{x}_{[0:p]}; \mathbf{z}]$$

We also consider other settings of C, primarily for understanding how the different pieces of the SCL feature space contribute to the overall model.

$$\begin{aligned} \textit{pivot: } C(\mathbf{x}, \mathbf{z}) &= \mathbf{x}_{[0:p]} \\ \textit{nonpivot: } C(\mathbf{x}, \mathbf{z}) &= \mathbf{x}_{[p:m]} \\ \textit{new: } C(\mathbf{x}, \mathbf{z}) &= \mathbf{z} \\ \textit{pivot+nonpivot: } C(\mathbf{x}, \mathbf{z}) &= \mathbf{x} \end{aligned}$$

Note that the *pivot+nonpivot* setting corresponds to a model that does not apply SCL at all.

4 Dataset

To explore cross-domain settings of authorship attribution, we need datasets containing documents from a number of authors from different domains (different topics, different genres). We use a corpus that consists of texts published in The Guardian daily newspaper that is actively used by the authorship attribution community in cross-domain studies (Stamatatos, 2013; Sapkota et al., 2014;

Sapkota et al., 2015). The Guardian corpus contains opinion articles written by 13 authors in four different topics: World, U.K., Society, and Politics. Following prior work, to make the collection balanced across authors, we choose at most ten documents per author for each of the four topics. Table 2 presents some statistics about the datasets.

5 Experimental Settings

We trained support vector machine (SVM) classifiers using the Weka implementation (Witten and Frank, 2005) with default parameters. For the *untyped* features, we used character 3-grams appearing at least 5 times in the training data, a list of 643 predefined stop-words, and the 3,500 most frequent non-stopword words as the lexical features. For the *typed* features, we used the top 3,500 most frequent 3-grams occurring at least five times in the training data for each of the 10 character n-gram categories. In both cases, we selected $p = 100$ pivot features as described in Section 3.2.

We measured performance in terms of accuracy across all possible topic pairings. That is, we paired each of the 4 topics in the Guardian corpus with each of the 3 remaining topics: train on Politics, test on Society; train on Politics, test on UK; train on Politics, test on World; etc. For each such model, we allowed SCL to learn feature correspondences from the labeled data of the 1 training topic and the unlabeled data of the 1 test topic. This resulted in 12 pairings of training/testing topics. We report both accuracy on the individual pairings and an overall average of the 12 accuracies.

We compare performance against two state-of-the-art baselines: Sapkota et al. (2014) and Sapkota et al. (2015), as described in Section 3.2, and whose features are denoted as **untyped** and **typed**, respectively. We replicate these models by using the *pivot+nonpivot* setting of C, i.e., not including any of the new SCL-based features.

6 Results

The following sections explore the results of our innovations in different areas: pivot features, feature binarizations, dimensionality reduction, and feature combination. For each section, we hold the other parameters constant and vary only the one parameter of interest. Thus, where not otherwise specified, we set parameters to the best values we observed in our experiments: we set the feature set to *typed*, the binarization $B_i(y)$ to the median, the reduced dimensionality d to 50, and the feature combination $C(\mathbf{x}, \mathbf{z})$ to *pivot+new* (i.e., we use the old pivot features alongside the new correspondence features). All reports of statistical significance are based on paired, two-tailed t-tests over the 12 different topic pairings.

Dataset	**untyped**	**typed**
Politics-Society	61.29	**67.74**
Politics-UK	**66.67**	63.33
Politics-World	58.97	**64.10**
Society-Politics	**62.96**	**62.96**
Society-UK	**72.50**	**72.50**
Society-World	**56.62**	48.08
UK-Politics	**68.75**	60.71
UK-Society	66.13	**67.74**
UK-World	57.27	**58.97**
World-Politics	**62.50**	59.82
World-Society	61.29	**62.90**
World-UK	46.67	**54.44**
Average	61.80	**61.94**

Table 3: Accuracy of **untyped** and **typed** feature sets. The difference between the averages is not statistically significant (p=0.927).

6.1 Untyped vs. Typed features

Table 3 compares the **untyped** feature set to the **typed** feature set. Both feature sets perform reasonably well, and substantially better than a model without SCL, where the performance of **untyped** is 56.43 and **typed** is 53.62 (see the *pivot+nonpivot* columns of Table 6 and Table 7, discussed in Section 6.4). Recall that the **typed** formulation includes only character n-gram features, while the **untyped** formulation includes stopwords and lexical features as well. Thus, given their very similar performance in Table 3, **typed** being slightly better, we select the simpler **typed** feature formulation for the remaining experiments.

6.2 Greater-than-zero vs. Median Binarization

Table 4 compares choices for $B_i(y)$, the function for converting a pivot feature value into a binary classification problem. In every single train/test scenario, and for both **untyped** and **typed** feature sets, our proposed median-based binarization function yielded performance greater than or equal to that of the traditional SCL greater-than-zero binarization function. This confirms our hypothesis that count-based features were inadequately modeled

Dataset	**untyped**		**typed**	
	>0	>med	>0	>med
Politics-Society	58.06	**61.29**	61.29	**67.74**
Politics-UK	**66.67**	**66.67**	**63.33**	**63.33**
Politics-World	55.56	**58.97**	63.81	**64.10**
Society-Politics	61.81	**62.96**	62.67	**62.96**
Society-UK	72.50	**72.50**	71.00	**72.50**
Society-World	51.92	**56.62**	46.00	**48.08**
UK-Politics	59.82	**68.75**	60.00	**60.71**
UK-Society	59.68	**66.13**	64.52	**67.74**
UK-World	47.86	**57.27**	57.27	**58.97**
World-Politics	56.25	**62.50**	56.50	**59.82**
World-Society	50.00	**61.29**	61.52	**62.90**
World-UK	42.22	**46.67**	50.00	**54.44**
Average	56.11	**61.80**	59.83	**61.94**

Table 4: Accuracy of greater-than-zero and median formulations of the $B_i(y)$ binarization function. Median is significantly better than greater-than-zero in both **untyped** (p=0.0007) and **typed** (p=0.003).

Dataset	d=25	d=50	d=100	no SVD
Politics-Society	66.13	67.74	**72.58**	50.00
Politics-UK	62.22	63.33	**66.67**	48.89
Politics-World	63.25	**64.10**	**64.10**	47.01
Society-Politics	**64.81**	62.96	55.56	57.41
Society-UK	67.50	**72.5**	67.5	70.00
Society-World	**48.08**	**48.08**	44.23	46.15
UK-Politics	**60.71**	**60.71**	58.93	51.79
UK-Society	64.52	**67.74**	56.45	59.68
UK-World	**60.68**	58.97	58.12	49.57
World-Politics	**62.50**	59.82	51.79	55.36
World-Society	59.68	62.90	**67.74**	62.90
World-UK	54.44	54.44	**55.56**	51.11
Average	61.21	**61.94**	59.94	54.16

Table 5: Accuracy of different choices for dimensionality reduction with **typed** features. The pattern is similar for **untyped**. $d = 50$ is significantly better than no SVD (p=0.0009), but not significantly different from $d = 25$ (p=0.291) or $d = 100$ (p=0.211).

in standard SCL and that the median-based binarization function improves the modeling of such features.

6.3 Dimensionality Reduction Choices

Table 5 compares different choices for the dimensionality reduction parameter d, as well as the possibility of not performing any dimensionality reduction at all ("No-SVD"). While each value of d yields the best performance on some of the train/test scenarios, $d = 50$ achieves the highest average accuracy (61.94). Removing the SVD entirely generally performs worse, and though on a small number of train/test scenarios it outperforms $d = 25$ and $d = 100$, it is always worse than $d = 50$.

This shows that SCL's feature correspondences alone are not sufficient to achieve domain adaptation. Without the SVD, performance is barely above a model without SCL: 54.16 vs. 53.62 (see Section 6.4). Much of the benefit appears to be coming from the SVD's basis-shift, since $d = 100$ outperforms no-SVD by more than 5 points[3], while $d = 50$ only outperforms $d = 100$ by 2 points. These results are consistent with SCL's origins in alternating structural optimization (Ando and Zhang, 2005), where SVD is derived as a necessary step for identifying a shared low-dimensional subspace.

6.4 Replacing vs. Concatenating Features

Table 6 and Table 7 compare the performance of different choices for the feature combination function $C(\mathbf{x}, \mathbf{z})$ on **untyped** and **typed** features, respectively. Our proposed *pivot+new* combination function, which replaces the non-pivot features with the new correspondence features, performs better on average than the two state-of-the-art baselines with no domain adaptation (*pivot+nonpivot*) and than the two state-of-the-art baselines augmented with classic SCL (*pivot+nonpivot+new*): 61.80 vs. 56.43 and 56.93 for untyped, and 61.94 vs. 53.62 and 54.23 for typed). These 5-8 point performance gains confirm the utility of our proposed *pivot+new* combination function, which replaces the old non-pivot features with the new correspondence features. These gains are consistent with (Blitzer et al., 2006), who included both pivot and non-pivot features, but found that they had to give pivot features a weight "five times that of the [non-pivot] features" to see improved performance.

While our approach is better on average, in some individual scenarios, it performs worse than classic SCL or no domain adaptation. For example, on Politics-Society, Politics-UK, and World-UK, using **typed** features, *pivot+new* performs worse than no domain adaptation (*pivot+nonpivot*). Our results suggest a rule for predicting when this degradation will happen: *pivot+new* will outperform

[3]Recall that $p = 100$, so $d = 100$ means the full matrix.

Dataset	*pivot*	*nonpivot*	*new*	*pivot+nonpivot*	*pivot+nonpivot+new*	*pivot+new*
Politics-Society	54.84	75.81	62.9	75.81	**77.42**	61.29
Politics-UK	63.33	68.89	58.89	70.00	**71.11**	66.67
Politics-World	58.12	63.25	53.85	64.96	**65.41**	58.97
Society-Politics	61.11	46.30	48.15	46.30	46.30	**62.96**
Society-UK	67.5	45.00	60.00	47.50	47.50	**72.50**
Society-World	50.00	42.31	53.85	46.15	46.15	**56.62**
UK-Politics	62.50	42.86	59.82	42.86	44.64	**68.75**
UK-Society	59.68	43.55	55.83	45.16	45.16	**66.13**
UK-World	45.30	38.46	48.72	39.32	39.32	**57.27**
World-Politics	55.36	69.64	56.25	68.75	**69.64**	62.5
World-Society	46.77	67.74	53.23	**69.35**	**69.35**	61.29
World-UK	43.33	**61.11**	50.00	**61.11**	**61.11**	46.67
Average	55.65	55.41	55.12	56.43	56.93	**61.80**

Table 6: Accuracy of different **untyped** feature combinations. The best performance for each dataset is in bold. The performance of *pivot+new* is not significantly different from *pivot+nonpivot* (p=0.258) or *pivot+nonpivot+new* (p=0.305).

Dataset	*pivot*	*nonpivot*	*new*	*pivot+nonpivot*	*pivot+nonpivot+new*	*pivot+new*
Politics-Society	48.39	70.97	59.68	**72.58**	**72.58**	67.74
Politics-UK	52.22	68.89	66.67	71.11	**72.22**	63.33
Politics-World	46.15	61.54	61.54	63.25	**64.10**	**64.10**
Society-Politics	55.56	48.15	61.11	48.15	50.00	**62.96**
Society-UK	65.00	45.00	65.00	45.00	45.00	**72.50**
Society-World	38.46	46.15	**53.85**	44.23	46.15	48.08
UK-Politics	48.21	44.64	55.36	45.54	45.54	**60.71**
UK-Society	51.61	41.94	66.13	41.94	41.94	**67.74**
UK-World	44.44	33.33	45.30	35.90	35.90	**58.97**
World-Politics	50.89	51.79	**61.39**	57.14	57.14	59.82
World-Society	54.84	59.68	43.55	59.68	61.29	**62.9**
World-UK	44.44	56.67	50.00	**58.89**	**58.89**	54.44
Average	50.02	52.40	57.47	53.62	54.23	**61.94**

Table 7: Accuracy of different **typed** feature combinations. The best performance for each dataset is in bold. The performance of *pivot+new* is significantly better than *pivot+nonpivot* (p=0.041) but not significantly different from *pivot+nonpivot+new* (p=0.059).

both *pivot+nonpivot* and *pivot+nonpivot+new* iff the *new* features alone outperform the *nonpivot* features alone. This rule holds in all 12 of 12 train/test scenarios for **untyped** features and 11 of 12 scenarios for **typed** features (failing on only World-Society). Intuitively, if the *new* correspondence features that result from SCL aren't better than the features they were meant to replace, then it is unlikely that they will result in performance gains. This might happen if the pivot features are not strong enough predictors, either because they have been selected poorly or because there are too few of them.

7 Discussion

To the best of our knowledge, we are the first to introduce a domain adaption model for authorship attribution that combines labeled data in a source domain with unlabeled data from a target domain to improve performance on the target domain. We proposed several extensions to the popular structural correspondence learning (SCL) algorithm for domain adaptation to make it more amenable to tasks like authorship attribution. The SCL algorithm requires the manual identification of domain independent *pivot* features for each task, so we proposed two feature formulations using charac-

ter n-grams as the pivot features, and showed that both yielded state-of-the-art performance. We also showed that for the binary classification task that is used by SCL to learn the feature correspondences, replacing the traditional greater-than-zero classification task with a median-based classification task allowed the model to better handle our count-based features. We explored the dimensionality reduction step of SCL and showed that singular value decomposition (SVD) over the feature correspondence matrix is critical to achieving high performance. Finally, we introduced a new approach to combining the original features with the learned correspondence features, and showed that replacing (rather than concatenating) the non-pivot features with the correspondence features generally yields better performance.

In the future, we would like to extend this work in several ways. First, though our median-based approach was successful in converting pivot feature values to binary classification problems, learning a regression model might be an even better approach for count-based features. Second, since the SVD basis-shift seems to be the source of much of the gains, we would like to explore replacing the SVD with other algorithms, such as independent component analysis. Finally, we would like to explore further our finding that the performance of the overall model seems to be predicted by the difference in performance between the non-pivot features and the new correspondence features, especially to see if this can be predicted at training time rather than as a post-hoc analysis.

8 Acknowledgments

This work was supported in part by CONACYT Project number 247870, and by the National Science Foundation award number 1462141.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853, December.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2006, pages 120–128, Stroudsburg, PA, USA.

Hal Daumé. 2007. Frustratingly easy domain adaptation. In *Annual meeting-association for computational linguistics*, volume 45, pages 256–256, 2007.

H. J. Escalante, T. Solorio, and M. Montes-y Gomez. 2011. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 288–298, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jade Goldstein-Stewart, Ransom Winder, and Roberta Evans Sabin. 2009. Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 336–344, Stroudsburg, PA, USA. Association for Computational Linguistics.

Graeme Hirst and Vanessa Wei Feng. 2012. Changes in style in authors with alzheimer's disease. *English Studies*, 93(3):357–370.

Qi Li. 2012. Literature survey: Domain adaptation algorithms for natural language processing. Technical report, February.

Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 513–520, Manchester, UK, August.

D. Madigan, A. Genkin, S. Argamon, D. Fradkin, and L. Ye. 2005. Author identification on the large scale. In *Proceedings of CSNA/Interface 05*.

George K. Mikros and Eleni K. Argiri. 2007. Investigating topic influence in authorship attribution. In *Proceedings of the International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*, pages 29–35.

Peter Prettenhofer and Benno Stein. 2011. Cross-lingual adaptation using structural correspondence learning. *ACM Trans. Intell. Syst. Technol.*, 3(1):13:1–13:22, October.

Upendra Sapkota, Thamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. 2014. Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1228–1237, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 93–102, Denver, Colorado, May–June. Association for Computational Linguistics.

Andrew I. Schein, Johnnie F. Caver, Randale J. Honaker, and Craig H. Martell. 2010. Author attribution evaluation with novel topic cross-validation. In *KDIR '10*, pages 206–215.

Nobuyuki Shimizu and Hiroshi Nakagawa. 2007. Structural correspondence learning for dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1166–1169, Prague, Czech Republic, June. Association for Computational Linguistics.

E. Stamatatos. 2006. Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence tools*, 15(5):823–838.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law & Policy*, 21(2):421 – 439.

Songbo Tan and Xueqi Cheng. 2009. Improving SCL model for sentiment-transfer learning. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 181–184, Boulder, Colorado, June. Association for Computational Linguistics.

I. H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kauffmann, 2nd edition.

Yanbo Zhang, Youli Qu, and Junsan Zhang. 2010. A new method of selecting pivot features for structural correspondence learning in domain adaptive sentiment analysis. In *Database Technology and Applications (DBTA), 2010 2nd International Workshop on*, pages 1–3.