Fast Easy Unsupervised Domain Adaptation with Marginalized Structured Dropout

Yi Yang and **Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
{yiyang, jacobe}@gatech.edu

Abstract

Unsupervised domain adaptation often relies on transforming the instance representation. However, most such approaches are designed for bag-of-words models, and ignore the structured features present in many problems in NLP. We propose a new technique called **marginalized structured dropout**, which exploits feature structure to obtain a remarkably simple and efficient feature projection. Applied to the task of fine-grained part-of-speech tagging on a dataset of historical Portuguese, marginalized structured dropout yields state-of-the-art accuracy while increasing speed by more than an order-of-magnitude over previous work.

1 Introduction

Unsupervised domain adaptation is a fundamental problem for natural language processing, as we hope to apply our systems to datasets unlike those for which we have annotations. This is particularly relevant as labeled datasets become stale in comparison with rapidly evolving social media writing styles (Eisenstein, 2013), and as there is increasing interest in natural language processing for historical texts (Piotrowski, 2012). While a number of different approaches for domain adaptation have been proposed (Pan and Yang, 2010; Søgaard, 2013), they tend to emphasize bag-of-words features for classification tasks such as sentiment analysis. Consequently, many approaches rely on each instance having a relatively large number of active features, and fail to exploit the structured feature spaces that characterize syntactic tasks such as sequence labeling and parsing (Smith, 2011).

As we will show, substantial efficiency improvements can be obtained by designing domain adaptation methods for learning in structured feature spaces. We build on work from the deep learning community, in which *denoising autoencoders* are trained to remove synthetic noise from the observed instances (Glorot et al., 2011a). By using the autoencoder to transform the original feature space, one may obtain a representation that is less dependent on any individual feature, and therefore more robust across domains. Chen et al. (2012) showed that such autoencoders can be learned even as the noising process is analytically marginalized; the idea is similar in spirit to feature noising (Wang et al., 2013). While the marginalized denoising autoencoder (mDA) is considerably faster than the original denoising autoencoder, it requires solving a system of equations that can grow very large, as realistic NLP tasks can involve 10^5 or more features.

In this paper we investigate noising functions that are explicitly designed for *structured feature spaces*, which are common in NLP. For example, in part-of-speech tagging, Toutanova et al. (2003) define several feature "templates": the current word, the previous word, the suffix of the current word, and so on. For each feature template, there are thousands of binary features. To exploit this structure, we propose two alternative noising techniques: (1) **feature scrambling**, which randomly chooses a feature template and randomly selects an alternative value within the template, and (2) **structured dropout**, which randomly eliminates all but a single feature template. We show how it is possible to marginalize over both types of noise, and find that the solution for structured dropout is substantially simpler and more efficient than the mDA approach of Chen et al. (2012), which does not consider feature structure.

We apply these ideas to fine-grained part-of-speech tagging on a dataset of Portuguese texts from the years 1502 to 1836 (Galves and Faria, 2010), training on recent texts and evaluating

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 538–544,
Baltimore, Maryland, USA, June 23-25 2014.

on older documents. Both structure-aware domain adaptation algorithms perform as well as standard dropout — and better than the well-known structural correspondence learning (SCL) algorithm (Blitzer et al., 2007) — but structured dropout is more than an order-of-magnitude faster. As a secondary contribution of this paper, we demonstrate the applicability of unsupervised domain adaptation to the syntactic analysis of historical texts.

2 Model

In this section we first briefly describe the denoising autoencoder (Glorot et al., 2011b), its application to domain adaptation, and the analytic marginalization of noise (Chen et al., 2012). Then we present three versions of marginalized denoising autoencoders (mDA) by incorporating different types of noise, including two new noising processes that are designed for structured features.

2.1 Denoising Autoencoders

Assume instances $\mathbf{x}_1, \ldots, \mathbf{x}_n$, which are drawn from both the source and target domains. We will "corrupt" these instances by adding different types of noise, and denote the corrupted version of $\mathbf{x}_i$ by $\tilde{\mathbf{x}}_i$. Single-layer denoising autoencoders reconstruct the corrupted inputs with a projection matrix $\mathbf{W} : \mathbb{R}^d \rightarrow \mathbb{R}^d$, which is estimated by minimizing the squared reconstruction loss

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{W}\tilde{\mathbf{x}}_i||^2. \quad (1)$$

If we write $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, and we write its corrupted version $\tilde{\mathbf{X}}$, then the loss in (1) can be written as

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2n} tr\left[\left(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}}\right)^\top \left(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}}\right)\right]. \quad (2)$$

In this case, we have the well-known closed-form solution for this ordinary least square problem:

$$\mathbf{W} = \mathbf{P}\mathbf{Q}^{-1}, \quad (3)$$

where $\mathbf{Q} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ and $\mathbf{P} = \mathbf{X}\tilde{\mathbf{X}}^\top$. After obtaining the weight matrix $\mathbf{W}$, we can insert nonlinearity into the output of the denoiser, such as $\tanh(\mathbf{WX})$. It is also possible to apply stacking, by passing this vector through another autoencoder (Chen et al., 2012). In pilot experiments, this slowed down estimation and had little effect on accuracy, so we did not include it.

High-dimensional setting Structured prediction tasks often have much more features than simple bag-of-words representation, and performance relies on the rare features. In a naive implementation of the denoising approach, both $\mathbf{P}$ and $\mathbf{Q}$ will be dense matrices with dimensionality $d \times d$, which would be roughly 10^{11} elements in our experiments. To solve this problem, Chen et al. (2012) propose to use a set of pivot features, and train the autoencoder to reconstruct the pivots from the full set of features. Specifically, the corrupted input is divided to S subsets $\tilde{\mathbf{x}}_i = \left[(\tilde{\mathbf{x}})_i^{1\top}, \ldots, (\tilde{\mathbf{x}})_i^{S\top}\right]^\top$. We obtain a projection matrix $\mathbf{W}^s$ for each subset by reconstructing the pivot features from the features in this subset; we can then use the sum of all reconstructions as the new features, $\tanh(\sum_{s=1}^{S} \mathbf{W}^s\mathbf{X}^s)$.

Marginalized Denoising Autoencoders In the standard denoising autoencoder, we need to generate multiple versions of the corrupted data $\tilde{\mathbf{X}}$ to reduce the variance of the solution (Glorot et al., 2011b). But Chen et al. (2012) show that it is possible to marginalize over the noise, analytically computing expectations of both $\mathbf{P}$ and $\mathbf{Q}$, and computing

$$\mathbf{W} = E[\mathbf{P}]E[\mathbf{Q}]^{-1}, \quad (4)$$

where $E[\mathbf{P}] = \sum_{i=1}^{n} E[\mathbf{x}_i\tilde{\mathbf{x}}_i^\top]$ and $E[\mathbf{Q}] = \sum_{i=1}^{n} E[\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top]$. This is equivalent to corrupting the data $m \rightarrow \infty$ times. The computation of these expectations depends on the type of noise.

2.2 Noise distributions

Chen et al. (2012) used dropout noise for domain adaptation, which we briefly review. We then describe two novel types of noise that are designed for structured feature spaces, and explain how they can be marginalized to efficiently compute $\mathbf{W}$.

Dropout noise In dropout noise, each feature is set to zero with probability $p > 0$. If we define the scatter matrix of the uncorrupted input as $\mathbf{S} = \mathbf{X}\mathbf{X}^\top$, the solutions under dropout noise are

$$E[\mathbf{Q}]_{\alpha,\beta} = \begin{cases} (1-p)^2\mathbf{S}_{\alpha,\beta} & \text{if } \alpha \neq \beta \\ (1-p)\mathbf{S}_{\alpha,\beta} & \text{if } \alpha = \beta \end{cases}, \quad (5)$$

and

$$E[\mathbf{P}]_{\alpha,\beta} = (1-p)\mathbf{S}_{\alpha,\beta}, \quad (6)$$

where α and β index two features. The form of these solutions means that computing $\mathbf{W}$ requires solving a system of equations equal to the number of features (in the naive implementation), or several smaller systems of equations (in the high-dimensional version). Note also that p is a tunable parameter for this type of noise.

Structured dropout noise In many NLP settings, we have several feature templates, such as previous-word, middle-word, next-word, etc, with only one feature per template firing on any token. We can exploit this structure by using an alternative dropout scheme: for each token, choose exactly one feature template to keep, and zero out all other features that consider this token (transition feature templates such as $\langle y_t, y_{t-1} \rangle$ are not considered for dropout). Assuming we have K feature templates, this noise leads to very simple solutions for the marginalized matrices $E[\mathbf{P}]$ and $E[\mathbf{Q}]$,

$$E[\mathbf{Q}]_{\alpha,\beta} = \begin{cases} 0 & \text{if } \alpha \neq \beta \\ \frac{1}{K}\mathbf{S}_{\alpha,\beta} & \text{if } \alpha = \beta \end{cases} \quad (7)$$

$$E[\mathbf{P}]_{\alpha,\beta} = \frac{1}{K}\mathbf{S}_{\alpha,\beta}, \quad (8)$$

For $E[\mathbf{P}]$, we obtain a scaled version of the scatter matrix, because in each instance $\tilde{\mathbf{x}}$, there is exactly a $1/K$ chance that each individual feature survives dropout. $E[\mathbf{Q}]$ is diagonal, because for any off-diagonal entry $E[\mathbf{Q}]_{\alpha,\beta}$, at least one of α and β will drop out for every instance. We can therefore view the projection matrix $\mathbf{W}$ as a row-normalized version of the scatter matrix $\mathbf{S}$. Put another way, the contribution of β to the reconstruction for α is equal to the co-occurence count of α and β, divided by the count of β.

Unlike standard dropout, there are no free hyper-parameters to tune for structured dropout. Since $E[\mathbf{Q}]$ is a diagonal matrix, we eliminate the cost of matrix inversion (or of solving a system of linear equations). Moreover, to extend mDA for high dimensional data, we no longer need to divide the corrupted input $\tilde{\mathbf{x}}$ to several subsets.[1]

For intuition, consider standard feature dropout with $p = \frac{K-1}{K}$. This will look very similar to structured dropout: the matrix $E[\mathbf{P}]$ is identical, and $E[\mathbf{Q}]$ has off-diagonal elements which are scaled by $(1-p)^2$, which goes to zero as K is large. However, by including these elements, standard dropout is considerably slower, as we show in our experiments.

Scrambling noise A third alternative is to "scramble" the features by randomly selecting alternative features within each template. For a feature α belonging to a template F, with probability p we will draw a noise feature β also belonging to F, according to some distribution q. In this work, we use an uniform distribution, in which $q_\beta = \frac{1}{|F|}$. However, the below solutions will also hold for other scrambling distributions, such as mean-preserving distributions.

Again, it is possible to analytically marginalize over this noise. Recall that $E[\mathbf{Q}] = \sum_{i=1}^{n} E[\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top]$. An off-diagonal entry in the matrix $\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top$ which involves features α and β belonging to different templates ($F_\alpha \neq F_\beta$) can take four different values ($\mathbf{x}_{i,\alpha}$ denotes feature α in $\mathbf{x}_i$):

- $\mathbf{x}_{i,\alpha}\mathbf{x}_{i,\beta}$ if both features are unchanged, which happens with probability $(1-p)^2$.
- 1 if both features are chosen as noise features, which happens with probability $p^2 q_\alpha q_\beta$.
- $\mathbf{x}_{i,\alpha}$ or $\mathbf{x}_{i,\beta}$ if one feature is unchanged and the other one is chosen as the noise feature, which happens with probability $p(1-p)q_\beta$ or $p(1-p)q_\alpha$.

The diagonal entries take the first two values above, with probability $1-p$ and pq_α respectively. Other entries will be all zero (only one feature belonging to the same template will fire in $\mathbf{x}_i$). We can use similar reasoning to compute the expectation of $\mathbf{P}$. With probability $(1-p)$, the original features are preserved, and we add the outer-product $\mathbf{x}_i\mathbf{x}_i^\top$; with probability p, we add the outer-product $\mathbf{x}_i q^\top$. Therefore $E[\mathbf{P}]$ can be computed as the sum of these terms.

3 Experiments

We compare these methods on historical Portuguese part-of-speech tagging, creating domains over historical epochs.

3.1 Experiment setup

Datasets We use the Tycho Brahe corpus to evaluate our methods. The corpus contains a total of 1,480,528 manually tagged words. It uses a set of 383 tags and is composed of various texts from

[1] $E[\mathbf{P}]$ is an r by d matrix, where r is the number of pivots.

historical Portuguese, from 1502 to 1836. We divide the texts into fifty-year periods to create different domains. Table 1 presents some statistics of the datasets. We hold out 5% of data as development data to tune parameters. The two most recent domains (1800-1849 and 1750-1849) are treated as source domains, and the other domains are target domains. This scenario is motivated by training a tagger on a modern newstext corpus and applying it to historical documents.

Dataset	# of Tokens				
	Total	Narrative	Letters	Dissertation	Theatre
1800-1849	125719	91582	34137	0	0
1750-1799	202346	57477	84465	0	60404
1700-1749	278846	0	130327	148519	0
1650-1699	248194	83938	115062	49194	0
1600-1649	295154	117515	115252	62387	0
1550-1599	148061	148061	0	0	0
1500-1549	182208	126516	0	55692	0
Overall	1480528	625089	479243	315792	60404

Table 1: Statistics of the Tycho Brahe Corpus

CRF tagger We use a conditional random field tagger, choosing CRFsuite because it supports arbitrary real valued features (Okazaki, 2007), with SGD optimization. Following the work of Nogueira Dos Santos et al. (2008) on this dataset, we apply the feature set of Ratnaparkhi (1996). There are 16 feature templates and $372,902$ features in total. Following Blitzer et al. (2006), we consider pivot features that appear more than 50 times in all the domains. This leads to a total of 1572 pivot features in our experiments.

Methods We compare mDA with three alternative approaches. We refer to *baseline* as training a CRF tagger on the source domain and testing on the target domain with only base features. We also include *PCA* to project the entire dataset onto a low-dimensional sub-space (while still including the original features). Finally, we compare against Structural Correspondence Learning (*SCL*; Blitzer et al., 2006), another feature learning algorithm. In all cases, we include the entire dataset to compute the feature projections; we also conducted experiments using only the test and training data for feature projections, with very similar results.

Parameters All the hyper-parameters are decided with our development data on the training set. We try different low dimension K from 10 to 2000 for PCA. Following Blitzer (2008) we perform feature centering/normalization, as well as rescaling for SCL. The best parameters for SCL are dimensionality $K = 25$ and rescale factor $\alpha = 5$, which are the same as in the original paper. For mDA, the best corruption level is $p = 0.9$ for dropout noise, and $p = 0.1$ for scrambling noise. Structured dropout noise has no free hyper-parameters.

3.2 Results

Table 2 presents results for different domain adaptation tasks. We also compute the *transfer ratio*, which is defined as $\frac{\textit{adaptation accuracy}}{\textit{baseline accuracy}}$, shown in Figure 1. The generally positive trend of these graphs indicates that adaptation becomes progressively more important as we select test sets that are more temporally remote from the training data.

In general, mDA outperforms SCL and PCA, the latter of which shows little improvement over the base features. The various noising approaches for mDA give very similar results. However, structured dropout is orders of magnitude faster than the alternatives, as shown in Table 3. The scrambling noise is most time-consuming, with cost dominated by a matrix multiplication.

Method	PCA	SCL	mDA		
			dropout	structured	scambling
Time	7,779	38,849	8,939	**339**	327,075

Table 3: Time, in seconds, to compute the feature transformation

4 Related Work

Domain adaptation Most previous work on domain adaptation focused on the supervised setting, in which some labeled data is available in the target domain (Jiang and Zhai, 2007; Daumé III, 2007; Finkel and Manning, 2009). Our work focuses on unsupervised domain adaptation, where no labeled data is available in the target domain. Several representation learning methods have been proposed to solve this problem. In structural correspondence learning (SCL), the induced representation is based on the task of predicting the presence of pivot features. Autoencoders apply a similar idea, but use the denoised instances as the latent representation (Vincent et al., 2008; Glorot et al., 2011b; Chen et al., 2012). Within the context of denoising autoencoders, we have focused

Task	baseline	PCA	SCL	mDA		
				dropout	structured	scrambling
from 1800-1849						
→ 1750	89.12	89.09	89.69	**90.08**	**90.08**	90.01
→ 1700	90.43	90.43	91.06	91.56	**91.57**	91.55
→ 1650	88.45	88.52	87.09	88.69	**88.70**	88.57
→ 1600	87.56	87.58	88.47	89.60	**89.61**	89.54
→ 1550	89.66	89.61	90.57	**91.39**	**91.39**	91.36
→ 1500	85.58	85.63	86.99	**88.96**	88.95	88.91
from 1750-1849						
→ 1700	94.64	94.62	94.81	**95.08**	**95.08**	95.02
→ 1650	**91.98**	90.97	90.37	90.83	90.84	90.80
→ 1600	92.95	92.91	93.17	**93.78**	**93.78**	93.71
→ 1550	93.27	93.21	93.75	**94.06**	94.05	94.02
→ 1500	89.80	89.75	90.59	**91.71**	**91.71**	91.68

Table 2: Accuracy results for adaptation from labeled data in 1800-1849, and in 1750-1849.

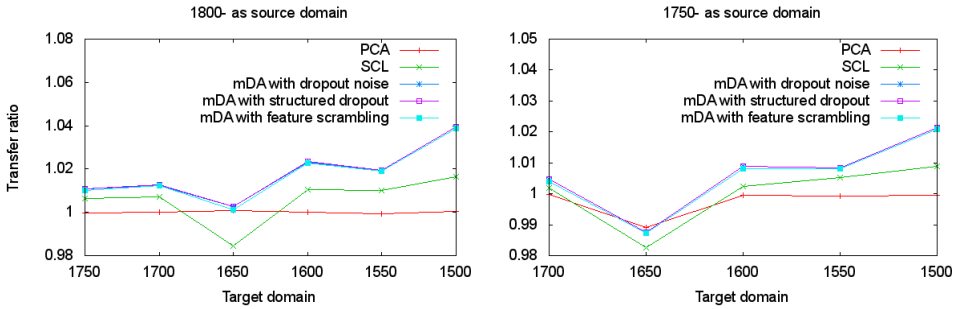

Figure 1: Transfer ratio for adaptation to historical text

on dropout noise, which has also been applied as a general technique for improving the robustness of machine learning, particularly in neural networks (Hinton et al., 2012; Wang et al., 2013).

On the specific problem of sequence labeling, Xiao and Guo (2013) proposed a supervised domain adaptation method by using a log-bilinear language adaptation model. Dhillon et al. (2011) presented a spectral method to estimate low dimensional context-specific word representations for sequence labeling. Huang and Yates (2009; 2012) used an HMM model to learn latent representations, and then leverage the Posterior Regularization framework to incorporate specific biases. Unlike these methods, our approach uses a standard CRF, but with transformed features.

Historical text Our evaluation concerns syntactic analysis of historical text, which is a topic of increasing interest for NLP (Piotrowski, 2012). Pennacchiotti and Zanzotto (2008) find that part-of-speech tagging degrades considerably when applied to a corpus of historical Italian. Moon and Baldridge (2007) tackle the challenging problem of tagging Middle English, using techniques for projecting syntactic annotations across languages. Prior work on the Tycho Brahe corpus applied supervised learning to a random split of test and training data (Kepler and Finger, 2006; Dos Santos et al., 2008); they did not consider the domain adaptation problem of training on recent data and testing on older historical text.

5 Conclusion and Future Work

Denoising autoencoders provide an intuitive solution for domain adaptation: transform the features into a representation that is resistant to the noise that may characterize the domain adaptation process. The original implementation of this idea produced this noise directly (Glorot et al., 2011b); later work showed that dropout noise could be analytically marginalized (Chen et al., 2012). We take another step towards simplicity by showing that structured dropout can make marginalization even easier, obtaining dramatic speedups without sacrificing accuracy.

Acknowledgments : We thank the reviewers for useful feedback. This research was supported by National Science Foundation award 1349837.

References

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 '06, pages 120–128, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, Prague, Czech Republic.

John Blitzer. 2008. *Domain Adaptation of Natural Language Processing Systems*. Ph.D. thesis, University of Pennsylvania.

Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 767–774. ACM, New York, NY, USA, July.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *ACL*, volume 1785, page 1787.

Paramveer S Dhillon, Dean P Foster, and Lyle H Ungar. 2011. Multi-view learning of word embeddings via cca. In *NIPS*, volume 24, pages 199–207.

Cícero Nogueira Dos Santos, Ruy L Milidiú, and Raúl P Rentería. 2008. Portuguese part-of-speech tagging using entropy guided transformation learning. In *Computational Processing of the Portuguese Language*, pages 143–152. Springer.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL*, Atlanta, GA.

Jenny Rose Finkel and Christopher D Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610. Association for Computational Linguistics.

Charlotte Galves and Pablo Faria. 2010. Tycho Brahe Parsed Corpus of Historical Portuguese. http://www.tycho.iel.unicamp.br/ tycho/corpus/en/index.html.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011a. Deep sparse rectifier networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, volume 15, pages 315–323.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011b. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Fei Huang and Alexander Yates. 2009. Distributional representations for handling sparsity in supervised sequence-labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 495–503. Association for Computational Linguistics.

Fei Huang and Alexander Yates. 2012. Biased representation learning for domain adaptation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1313–1323. Association for Computational Linguistics.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *ACL*, volume 2007, page 22.

Fábio N Kepler and Marcelo Finger. 2006. Comparing two markov methods for part-of-speech tagging of portuguese. In *Advances in Artificial Intelligence-IBERAMIA-SBIA 2006*, pages 482–491. Springer.

Taesun Moon and Jason Baldridge. 2007. Part-of-speech tagging for middle english through alignment and projection of parallel diachronic texts. In *EMNLP-CoNLL*, pages 390–399.

Cícero Nogueira Dos Santos, Ruy L. Milidiú, and Raúl P. Rentería. 2008. Portuguese part-of-speech tagging using entropy guided transformation learning. In *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*, PROPOR '08, pages 143–152, Berlin, Heidelberg. Springer-Verlag.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

Marco Pennacchiotti and Fabio Massimo Zanzotto. 2008. Natural language processing across time: An empirical investigation on italian. In *Advances in Natural Language Processing*, pages 371–382. Springer.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, April 16.

Noah A Smith. 2011. Linguistic structure prediction. *Synthesis Lectures on Human Language Technologies*, 4(2):1–274.

Anders Søgaard. 2013. Semi-supervised learning and domain adaptation in natural language processing. *Synthesis Lectures on Human Language Technologies*, 6(2):1–103.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.

Sida I. Wang, Mengqiu Wang, Stefan Wager, Percy Liang, and Christopher D. Manning. 2013. Feature noising for log-linear structured prediction. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Min Xiao and Yuhong Guo. 2013. Domain adaptation for sequence labeling tasks with a probabilistic language adaptation model. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 293–301. JMLR Workshop and Conference Proceedings.