Blind Domain Transfer for Named Entity Recognition using Generative Latent Topic Models

Ramesh Nallapati, Mihai Surdeanu and Christopher Manning
Department of Computer Science
Stanford University
Stanford, CA 94305
{nmramesh,mihais,manning}@cs.stanford.edu

Abstract

State-of-the-art named entity recognizers (NER) are highly accurate at tagging documents with named-entity labels when the test documents are from the same domain as the training set, but performance drops significantly when switching to a novel domain. In this paper, we propose extensions to the state-of-the-art conditional random field model (CRF) based on features from generative unsupervised latent topic models such as Latent Dirichlet Allocation (LDA). In a transfer learning setting which we call *Blind Domain Transfer*, where no labeled data from the target domain is available for training, we show that this approach reduces the CRF's error rate by 3%. We also build a new supervised variant of LDA specifically for NER, and show that the CRF's error rate can be reduced by 5.5% on unseen domains, besides achieving consistent and statistically significant gains over the baseline.

1 Introduction

Named entity recognition (NER) is a fundamental natural language processing (NLP) component that is crucial for many applications, such as question answering or information extraction. Modern named entity recognizers based on conditional random fields (CRF)[7] have attained near-human performance on many research datasets. However, there is one caveat: their performance is high only when evaluated on the same domain that they are trained on, but rapidly degrades when evaluated on a new domain. For example, in a simple experiment using the six domains from the corpus provided in the Automatic Content Extraction (ACE) evaluation[1], we scored the Stanford NER system [3] at 78 F1 points when evaluated on domains it has seen in training, but only at 72 F1 points, on average, when evaluated on domains that were held out during training. In the recent past, many researchers have turned their attention to fixing this deficiency of NER systems [1, 4, 6]. This problem, called *domain adaptation* or *domain transfer*, is a specific type of transfer learning, where the system is evaluated on the same task that it is trained on, but on a domain different from the training domain.

In most of the past work, the focus has been on the setting where a large amount of training data is available in the source domain, but only limited training data is available in the target domain. In this work, we consider a setting that we call *blind domain transfer*, where we evaluate the performance of the NER system on a domain that is completely unseen by the system at training time. We believe this is an important and useful setting because it evaluates how well the model adapts and generalizes to completely novel settings.

2 Models

In the blind domain transfer setting, the basic CRF-based NER system has no way to predict the features or distribution of features in the target domain. Therefore, we rely on the family of generative latent topic models such as latent Dirichlet allocation (LDA) [2] to boost the generalizability of the model. These models help improve the adaptability of NER systems in two ways: (i) topic models

[1] http://www.itl.nist.gov/iad/mig/tests/ace/

project words to a lower dimensional latent topical space that helps the NER systems overcome data sparsity and improve generalizability to unseen words; and (ii) they help disambiguate the named-entity labels of ambiguous words, based on the document's overall topicality. For example, the word *Washington* may be tagged with the label 'LOCATION' in a document on contemporary politics, whereas in a document that discusses movies, the same word may be tagged as 'PERSON'.

2.1 Baseline

As baseline, we used the Stanford named entity recognizer [3], which implements a linear chain conditional random field [7]. We used a feature set previously tuned to maximize performance on the CoNLL-2003[2] dataset. The features model sequences of words, word shapes, part of speech tags, and cluster assignments for tokens constructed using distributional similarity.[3]

2.2 CRF with LDA features

The second model is a straight-forward extension of the CRF that takes into account topics of the words as additional features. This involves training an LDA model on the word tokens of the training dataset and performing inference on the test set. We perform LDA training and inference using Gibbs sampling as described in [5]. We train the CRF by adding for each word, topic IDs of the current word, previous word and the next word as additional features, and perform inference on the test set using these features as well.

Since the LDA model trains/infers on the whole document, it is able to factor in the topical context of the whole document while assigning topic IDs to each word in the document. On the other hand, the CRF has no explicit mechanism to incorporate document level global information while inferring the NE labels of each token.

2.3 CRF with external LDA features

Since LDA is a completely unsupervised model, there is no reason to restrict the training data of LDA to the training set used to train the CRF. Hence, we relaxed this condition and trained LDA on a much larger dataset consisting of a 10,000 document subset of the Gigaword corpus comprising newswire articles from New York Times.

2.4 Joint CRF and NER-LDA

In this subsection, we present a new supervised topic model that trains on document words as well as their named entity labels and infers topics. We call this model NER-LDA. The generative process of the model is indicated in the standard plate representation in Figure 1 and is enumerated below:

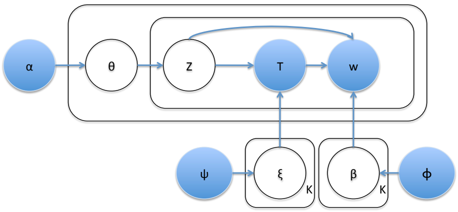

Figure 1: Graphical representation of the NER-LDA model

```
1. For each topic $k \in 1, \cdots, K$
2.   Generate multinomial over NE labels for topic $k$: $\xi_k \sim$ Dir$(\cdot|\Psi)$
3.   For each NE label $t$:
4.     Generate multinomial over words for topic $k$ and NE label $t$: $\beta_{k,t} \sim$ Dir$(\cdot|\Phi)$
5. For each document $d \in 1, \cdots, M$:
6.   Generate document's distribution over topics $\theta_d \sim$ Mult$(\cdot|\alpha)$
7.   For each position $n \in 1, \cdots, N_d$:
8.     Generate topic indicator $z_{dn} \sim$ Mult$(\cdot|\theta_d)$
9.     Generate NE label $t_{dn} \sim$ Mult$(\cdot|\xi_{z_{dn}})$
```

[2] http://www.cnts.ua.ac.be/conll2003/ner/

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

10. Generate word $w_{dn} \sim \text{Mult}(\cdot|\beta_{z_{dn},t_{dn}})$

In other words, the new model is allowed to have different distributions over NE labels for different topics, and the generation of each word is conditioned not only on its topic, but also on its NE label. We use collapsed Gibbs sampling for training, whose sampling probability is as follows:

$$p(z_{dn} = k|t_{dn} = t, w_{dn} = w, \mathbf{z}^{(-n)}) \propto (c_{dk}^{(-n)} + \alpha) \frac{c_{tk}^{(-n)} + \psi}{\sum_{t'} c_{t'k'}^{(-n)} + T\psi} \frac{c_{wkt}^{(-n)} + \phi}{\sum_{w'} c_{w'kt}^{(-n)} + V\phi}, \quad (1)$$

where $c_{dk}^{(-n)}$ is the count of tokens in document d that are assigned to topic k not including the position n, $c_{tk}^{(-n)}$ is the number of tokens assigned to NE label t and topic k discounting position n, and $c_{wkt}^{(-n)}$ is the number of tokens of type w assigned to topic k and NE label t, excluding position n. Also, T is the number of unique named entity labels, V is the vocabulary size, and K is the total number of topics. At testing time, we assume β and ξ to be given as computed from the empirical estimates based on averaging several Gibbs samples of the training data. Since the named-entity tags are not visible at test time, we alternately sample topic z and entity label t as follows:

$$p(z_{dn} = k|t_{dn} = t, w_{dn} = w, \mathbf{z}^{(-n)}) \propto (c_{dk}^{(-n)} + \alpha)\hat{\eta}_{tk}\hat{\beta}_{wkt} \quad (2)$$

$$p(t_{dn} = t|z_{dn} = k, w_{dn} = w, \mathbf{z}^{(-n)}) \propto \hat{\eta}_{tk}\hat{\beta}_{wkt} \quad (3)$$

We train the CRF and the NER-LDA models independently on the same data, but at testing time, for each token position in the test document, we combine the probability of the NE label of the NER-LDA model as given in Eq. 3 with the local conditional probability predicted by CRF using a weighted product approach as described in [3] and shown below:

$$P(T_n = t|W_n = w, d) \propto P_{\text{CRF}}(T_n = t|\mathbf{T}^{(-n)}, \mathbf{W})^{\gamma} P_{\text{NER-LDA}}(T_n = t|W_n = w, Z_n = k)^{(1-\gamma)}$$

where $0 \leq \gamma \leq 1.0$ and the NER-LDA probability above is averaged over multiple Gibbs samples.

3 Experiments and Results

We used the ACE 2005 corpus for our experiments, which consists of six domains such as newswire (NW), weblogs (WL), broadcast news (BN), etc [6]. Each domain is further split into official train and test splits. For our transfer learning experiments, we used a leave-one-out approach, where the models are trained on all but one held-out domain which we call *source* domains, and tested on the held out domain, called the *target* domain. This gave us 6 different transfer learning settings. For tuning the models' parameters in each setting, we further divided the training set of that setting into *dev-train* and *dev-test* subsets by pooling together all the official training partitions and the official test partitions of the source domains, respectively. During development, the models are trained on the *dev-train* set and the free parameters are tuned to optimize the performance on *dev-test* set. The baseline CRF model was previously tuned to maximize its performance on CONLL NER corpora. For the CRF with LDA features, we tuned the number of topics of the LDA model until we achieved optimal performance on the development-test set. For the joint CRF-NER-LDA model, we tuned the number of topics of the NER-LDA model, as well as γ, the mixture weight between CRF and NER-LDA. Note that both during development and actual training, the models have no access to any data from the target domain.

For evaluation, we use the CONLL evaluation software[4], which measures F1 accuracy on exact matches of entire named entity strings. The results of our experiments are summarized in Table 1. The table shows that adding LDA features to the CRF (third column in the table) improves performance over the baseline on four out of six domains. Performance increases further when LDA is trained on the larger Gigaword corpus with as much as 3% error reduction on the NW corpus (fourth column). This is especially significant because LDA delivers the improvement free of annotation cost due to its unsupervised nature. Finally, the joint CRF-NER-LDA model (last column) is also able to improve the performance over the basic CRF model consistently, and reduces the error rate by 5.5% on CTS. Between CRF with LDA features trained on the Gigaword corpus and the joint CRF-NER-LDA model there are no clear winners, as each of them significantly outperforms

[4] http://www.cnts.ua.ac.be/conll2003/ner/

Domain	CRF	CRF w. LDA features	CRF w. LDA features (Gigaword 10K)	Joint CRF-NER-LDA
BC	75.32	75.53^{+}	**76.13** (2.45%) $^{++}$	75.68^{+}
BN	64.16	64.71^{+}	**64.85**(1.93%)$^{+}$	64.66^{+}
CTS	84.65	84.54^{-}	84.10^{-}	**85.50**(5.53%)$^{++}$
NW	71.73	71.70^{-}	**72.56**(2.93%)$^{++}$	72.10^{+}
UN	69.94	70.23^{+}	69.93	**70.62**(2.26%)$^{++}$
WL	68.71	69.04^{+}	69.22^{+}	**69.64**(2.97%)$^{++}$
# Wins over baseline	-	4	4	6

Table 1: F1 scores on the six transfer learning tasks. Bold numbers indicate the best performing model on a given domain, and the numbers in brackets indicate the *% error reduction* compared to the baseline CRF. A single '+' / '-' next to a number indicates the corresponding model is significantly better/worse than the baseline CRF as per one-tailed paired T-test at 95% confidence measured using bootstrap resampling of the test data. The symbol '++' indicates the corresponding model is significantly better than the baseline as well as the nearest performing model in that row, using the same statistical test. All numbers are lower than the

the other on 3 domains. However. the joint CRF-NER-LDA model is the most consistent model, achieving statistically signifcant improvement in performance over the baseline CRF on all domains.

The performance numbers we obtained are lower than the ones reported in official CONLL evaluation[5], for the following reasons: (a) ACE corpus is harder because it has more NE-types than CONLL and has more variability in text due to multiple domains, and (b) there is less amount of training data in ACE than in CONLL.

4 Conclusion

In this work, we proposed two LDA based techniques that improve the generalizability of the CRF model for NER. Our work proves empirically that topic models can be effectively used to mitigate performance loss in domain transfer setups. As part of future work, we intend to check if one could improve the performance of CRF further by training LDA on larger unlabeled datasets. Another direction we would like to explore is the co-training of CRF *and* NER-LDA models. In this model, the CRF would infer NE tags on unlabeled data, which the NER-LDA model would use for training purposes, while the CRF would retrain using NER-LDA's topic assignments as additional features.

References

[1] Andrew Arnold, Ramesh Nallapati, and William Cohen. Exploiting feature hierarchy for transfer learning in named entity recognition. In *Association for COmputational Linguistics*, 2008.

[2] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[3] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370*, 2005.

[4] Jenny Rose Finkel and Christopher Manning. Hierarchical bayesian domain adaptation. In *NAACL*, 2009.

[5] Thomas Griffiths and Mark Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, 2004.

[6] Hal Daume III. Frustratingly easy domain adaptation. In *Association for Computational Linguistics*, 2007.

[7] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, 2001.

[5] http://www.cnts.ua.ac.be/conll2003/ner/