Type-Supervised Domain Adaptation for Joint Segmentation and POS-Tagging

Meishan Zhang[†], **Yue Zhang**[‡] , **Wanxiang Che**[†], **Ting Liu**[†*]
[†]Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{mszhang, car, tliu}@ir.hit.edu.cn
[‡]Singapore University of Technology and Design
yue_zhang@sutd.edu.sg

Abstract

We report an empirical investigation on type-supervised domain adaptation for joint Chinese word segmentation and POS-tagging, making use of domain-specific tag dictionaries and only unlabeled target domain data to improve target-domain accuracies, given a set of annotated source domain sentences. Previous work on POS-tagging of other languages showed that type-supervision can be a competitive alternative to token-supervision, while semi-supervised techniques such as label propagation are important to the effectiveness of type-supervision. We report similar findings using a novel approach for joint Chinese segmentation and POS-tagging, under a cross-domain setting. With the help of unlabeled sentences and a lexicon of 3,000 words, we obtain 33% error reduction in target-domain tagging. In addition, combined type- and token-supervision can lead to improved cost-effectiveness.

1 Introduction

With accuracies of over 97%, POS-tagging of WSJ can be treated as a solved problem (Manning, 2011). However, performance is still well below satisfactory for many other languages and domains (Petrov et al., 2012; Christodoulopoulos et al., 2010). There has been a line of research on using a tag-dictionary for POS-tagging (Merialdo, 1994; Toutanova and Johnson, 2007; Ravi and Knight, 2009; Garrette and Baldridge, 2012). The idea is compelling: on the one hand, a list of lexicons is often available for special domains, such as bio-informatics; on the other hand, compiling a lexicon of word-tag pairs appears to be less time-consuming than annotating full sentences.

However, success in type-supervised POS-tagging turns out to depend on several subtle factors. For example, recent research has found that the quality of the tag-dictionary is crucial to the success of such methods (Banko and Moore, 2004; Goldberg et al., 2008; Garrette and Baldridge, 2012). Banko and Moore (2004) found that the accuracies can drop from 96% to 77% when a hand-crafted tag dictionary is replaced with a raw tag dictionary gleaned from data, without any human intervention. These facts indicate that careful considerations need to be given for effective type-supervision. In addition, significant manual work might be required to ensure the quality of lexicons.

To compare type- and token-supervised tagging, Garrette and Baldridge (2013) performed a set of experiments by conducting each type of annotation for two hours. They showed that for low-resource languages, a tag-dictionary can be reasonably effective if label propagation (Talukdar and Crammer, 2009) and model minimizations (Ravi and Knight, 2009) are applied to expand and filter the lexicons. Similar findings were reported in Garrette et al. (2013).

Do the above findings carry over to the Chinese language? In this paper, we perform an empirical study on the effects of tag-dictionaries for domain adaptation of Chinese POS-tagging. We aim to answer the following research questions: (a) Is domain adaptation feasible with only a target-domain lexicon? (b) Can we further improve type-supervised domain adaptation using unlabeled target-domain sentences? (c) Is crafting a tag dictionary for domain adaptation more effective than manually annotating target domain sentences, given similar efforts?

Our investigations are performed under two Chinese-specific settings. First, unlike low-resource languages, large amounts of annotation

[*]Corresponding author.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 588–597,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

are available for Chinese. For example, the Chinese Treebank (CTB) (Xue et al., 2005) contains over 50,000 manually tagged news sentences. Hence rather than studying purely type-supervised POS-tagging, we make use of CTB as the source domain, and study domain adaptation to the Internet literature.

Second, one uniqueness of Chinese POS-tagging, in contrast to the POS-tagging of alphabetical languages, is that word segmentation can be performed jointly to avoid error propagation (Ng and Low, 2004; Zhang and Clark, 2008; Kruengkrai et al., 2009; Zhang and Clark, 2010). We adopt this approach for a strong baseline. Previous studies showed that unsupervised domain adaptation can give moderate improvements (Liu and Zhang, 2012). We show that accuracies can be much more significantly improved by using target-domain knowledge in the form of lexicons.

Both token-supervised and type-supervised domain adaptation rely on a set of source-domain annotations; while the former makes additional use of a small set of target annotations, the latter leverages a target-domain lexicon. We take a feature-based method, analogous to that of Daume III (2007), which tunes domain-dependent versions of features using domain-specific data. Our method tunes a set of lexicon-based features, so that domain-dependent models are derived from inserting domain-specific lexicons.

The conceptually simple method worked highly effectively on a test set of 1,394 sentences from the Internet novel “Zhuxian”. Combined with the use of unlabeled data, a tag lexicon of 3,000 words gave a 33% error reduction when compared with a strong baseline system trained using CTB data. We observe that joint use of type- and token-supervised domain adaptation is more cost-effective than pure type- or token-supervision. With 10 hours of annotation, the best error reduction reaches 47%, with F-score increasing from 80.81% to 89.84%.

2 Baseline

We take as the baseline system a discriminative joint segmentation and tagging model, proposed by Zhang and Clark (2010), together with simple self-training (Liu and Zhang, 2012). While the baseline discriminative model gives state-of-the-art joint segmentation and tagging accuracies on CTB data, the baseline self-training makes use of unlabeled target domain data to find improved target domain accuracies over bare CTB training.

2.1 The Baseline Discriminative Chinese POS-Tagging Model

The baseline discriminative model performs segmentation and POS-tagging simultaneously. Given an input sentence $c_1 \cdots c_n$ (c_i refers to the ith character in the sentence), it operates incrementally, from left to right. At each step, the current character can either be appended to the last word of the existing partial output, or seperated as the start of a new word with tag p. A beam is used to maintain the *N-best* partial results at each step during decoding. At step i ($0 \leq i < n$), each item in the beam corresponds to a segmentation and POS-tagging hypothesis for the first $i-1$ characters, with the last word being associated with a POS, but marked as incomplete. When the next character c_i is processed, it is combined with all the partial results from the beam to generate new partial results, using two types of actions: (1) ***Append***, which appends c_i to the last (partial) word in a partial result; (2) ***Separate***$(\mathbf{p})$, which makes the last word in the partial result as completed and adds c_i as a new partial word with a POS tag $\mathbf{p}$.

Partial results in the beam are scored globally over all actions used to build them, so that the *N-best* can be put back to the agenda for the next step. For each action, features are extracted differently. We use the features from Zhang and Clark (2010). Discriminative learning with early-update (Collins and Roark, 2004; Zhang and Clark, 2011) is used to train the model with beam-search.

2.2 Baseline Unsupervised Adaptation by Self-Training

A simple unsupervised approach for POS-tagging with unlabeled data is EM. For a generative model such as HMM, EM can locally maximize the likelihood of training data. Given a good start, EM can result in a competitive HMM tagging model (Goldberg et al., 2008).

For discriminative models with source-domain training examples, an initial model can be trained using the source-domain data, and self-training can be applied to find a locally-optimized model using raw target domain sentences. The training process is sometimes associated with the EM algorithm. Liu and Zhang (2012) used perplexities of character trigrams to order unlabeled sentences, and applied self-training to achieve a 6.3% error

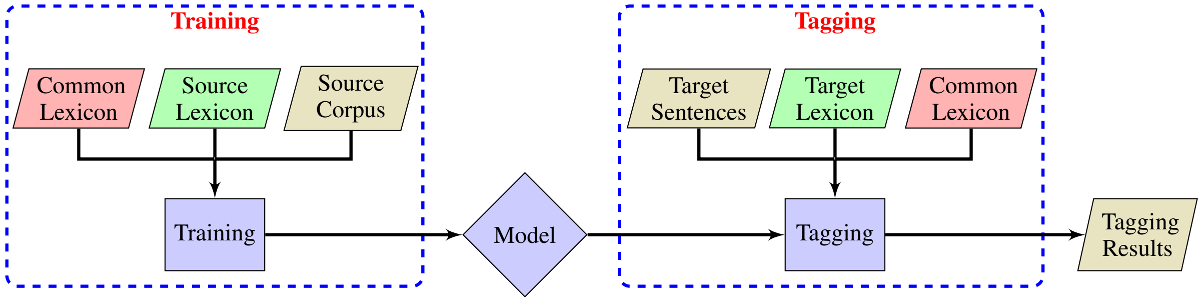

Figure 1: Architecture of our lexicon-based model for domain adaptation.

reduction on target-domain data when compared with source domain training. Their method is simple to implement, and we take it as our baseline.

3 Type-Supervised Domain Adaptation

To give a formal definition of the domain adaptation tasks, we denote by C_s a set of annotated source-domain sentences, C_t a set of annotated target-domain sentences, and $\mathfrak{L}_t$ an annotated target-domain lexicon. The form of $\mathfrak{L}_t$ is a list of target-domain words, each associated with a set of POS tags. *Token-supervised* domain adaptation is the task of making use of C_s and C_t to improve target-domain performances, while *type-supervised* domain adaptation is to make use of C_s and $\mathfrak{L}_t$ instead for the same purpose.

As described in the introduction, type-supervised domain adaptation is useful when annotated sentences are absent, but lexicons are available. In addition, it is an interesting question which type of annotation is more cost-effective when neither is available. We empirically compare the two approaches by proposing a novel method for type-supervised domain adaptation of a discriminate tagging model, showing that it can be a favourable choice in practical situation.

In particular, we split Chinese words into domain-independent and domain-specific categories, and define unlexicalized features for domain-specific words. We train lexicalized domain-independent and unlexicalized domain-specific features using the source domain annotated sentences and a source-domain lexicon, and then apply the resulting model to the target domain by replacing the source-domain lexicon with a target domain lexicon. Combined with unsupervised learning with unlabeled target-domain of sentences, the conceptually simple method worked highly effectively. Following Garrette and Baldridge (2013), we address practical questions on type-supervised domain adaptation by comparison with token-supervised methods under similar human annotation efforts.

3.1 System Architecture

Our method is based on the intuition that domain-specific words of certain types (e.g. proper names) can behave similarly across domains. For example, consider the source-domain sentence "江泽民|NR (Jiang Zemin) 随后|AD (afterwards) 访问|VV (visit) 上汽|NR (Shanghai Automobiles Corp.)" and the target-domain sentence "碧瑶|NR (Biyao) 随后|AD (afterwards) 来到|VV (arrive) 大竹峰|NR (the Bamboo Mountains)". "江泽民 (Jiang Zemin)" and "碧瑶 (Biyao)" are person names in the two domains, respectively, whereas "上汽 (Shanghai Automobiles Corp.)" and "大竹峰 (the Bamboo Mountains)" are location names in the two domains, respectively. If the four words are simply treated as domain-specific nouns, the two sentences both have the pattern "⟨domain-NR⟩ AD VV ⟨domain-NR⟩", and hence source domain training data can be useful in training the distributions of the lexicon-based features for both domains.

Further, we assume that the syntax structures and the usage of function words do not vary significantly across domains. For example, verbs, adjectives or proper nouns can be different from domain to domain, but the subject-verb-object sentence structure does not change. In addition, the usage of closed-set function words remains stable across different domains. In the CTB tagset, closed-set POS tags are the vast majority. Under this assumption, we introduce a set of unlexicalized features into the discriminative model, in order to capture the distributions of domain-specific dictionary words. Unlexicalized features trained for source domain words can carry over to the target domain. The overall architecture of our sys-

Action	Lexicon Feature templates
Separate	in-lex(w_{-1}), $l(w_{-1}) \circ$ in-lex(w_{-1}), in-lex(w_{-1}, t_{-1}), $l(w_{-1}) \circ$ in-lex(w_{-1}, t_{-1})

Table 1: Dictionary features of the type-supervised model, where w_{-1} and t_{-1} denote the last word and POS tag of a partial result, respectively; $l(w)$ denotes the length of the word w; in-lex(w, t) denotes whether the word-tag pair (w, t) is in the lexicon.

tem is shown in Figure 1, where lexicons can be treated as "plugins" to the model for different domains, and one model trained from the source domain can be applied to many different target domains, as long as a lexicon is available.

The method can be the most effective when there is a significant amount of domain-independent words in the data, which provide rich lexicalized contexts for estimating unlexicalized features for domain-specific words. For scientific domains (e.g. the biomedical domain) which share a significant proportion of common words with the news domain, and have most domain specific words being nouns (e.g. "糖尿病 (diabetes)"), the method can be the most effective. We choose a comparatively difficult domain pair (e.g. modern news v.s. ancient style novel), for which the use of many word types are quite different. Results on this data can be relatively more indicative of the usefulness of the method.

3.2 Lexicon-Based Features

Table 1 shows the set of new unlexicalized features for the domain-specific lexicons. In addition to words and POS tags, length information is also encoded in the features, to capture different distributions of different word sizes. For example, a one-character word in the dictionary might not be identified as confidently using the lexicon as a three-character word in the dictionary.

To acquire a domain-specific lexicon for the source domain, we use HowNet (Dong and Dong, 2006) to classify CTB words into domain-independent and domain-specific categories. Consisting of semantic information for nearly 100,000 common Chinese words, HowNet can serve as a resource of domain-independent Chinese words. We choose out of all words in the source domain training data those that also occur in HowNet for domain-independent words, and out of the remaining words those that occur more than 3 times for words specific to the source domain. We assume that the domain-independent lexicon applies to all target domains also. For some target domains, we can obtain domain-specific terminologies easily from the Internet. However, this can be a very small portion depending on the domain. Thus, it may still be necessary to obtain new lexicons by manual annotation.

3.3 Lexicon and Self-Training

The lexicon-based features can be combined with unsupervised learning to further improve target-domain accuracies. We apply self-training on top of the lexicon-based features in the following way: we train a lexicon-based model M using a lexicon $\mathfrak{L}_s$ of the source domain, and then apply M together with a target-domain lexicon $\mathfrak{L}_t$ to automatically label a set of target domain sentences. We combine the automatically labeled target sentences with the source-domain training data to obtain an extended set of training data, and train a final model M_{self}, using the lexicon $\mathfrak{L}_s$ and $\mathfrak{L}_t$ for source- and target-domain data, respectively.

Different numbers of target domain sentences can be used for self-training. Liu and Zhang (2012) showed that an increased amount of target sentences do not constantly lead to improved development accuracies. They use character perplexity to order target domain sentences, taking the top K sentences for self-training. They evaluate the optimal development accuracies using a range of different Kvalues, and select the best K for a final model. This method gave better results than using sentences in the internet novel in their original order (Liu and Zhang, 2012). We follow this method in ranking target domain sentences.

4 Experiments

4.1 Setting

We use annotated sentences from the CTB5 for source-domain training, splitting the corpus into training, development and test sections in the same way as previous work (Kruengkrai et al., 2009; Zhang and Clark, 2010; Sun, 2011).

Following Liu and Zhang (2012), we use the free Internet novel "Zhuxian" (henceforth referred to as ZX; also known as "Jade dynasty") as our target domain data. The writing style of the novel is in the literature genre, with the style of Ming and Qing novels, very different from news in CTB. Ex-

CTB sentences	ZX sentences
乔石会见俄罗斯议员团 (Qiaoshi meets the Russian delegates.)	天下之大，无奇不有，山川灵秀，亦多妖魔鬼怪。 (The world was big. It held everything. There were fascinating landscapes. There were haunting ghosts.)
李鹏强调要加快推行公务员制度 (Lipeng stressed on speeding the reform of official regulations.)	时间无多，我去请出诛仙古剑。 (No time left. Let me call out Zhuxian, the ancient sword.)
中国化学工业加快对外开放步伐 (Chinese chemistry industry increases the pace of opening up.)	忽听得狂笑风起，法宝异光闪动。(There came suddenly a gust of wind, out of which was laughters and magic flashes.)

Table 2: Example sentences from CTB and ZX to illustrate the differences between news and novel.

Data Set		Chap. IDs	# sents	# words
CTB5	Train	1-270, 400-931, 1001-1151	10,086	493,930
	Devel	301-325	350	6,821
	Test	271-300	348	8,008
ZX	Train	6.6-6.10, 7.6-7.10, 19	2,373	67,648
	Devel	6.1-6.5	788	20,393
	Test	7.1-7.5	1,394	34,355

Table 3: Corpus statistics.

ample sentences from the two corpora are shown in Table 2. Liu and Zhang (2012) manually annotated 385 sentences as development and test data, which we download from their website.[1] These data follow the same annotation guidelines as the Chinese Treebank (Xue et al., 2000).

To gain more reliable statistics in our results, we extend their annotation work to a total 4,555 sentences, covering the sections 6, 7 and 19 of the novel. The annotation work is based on the automatically labeled sentences by our baseline model trained with CTB5 corpus. It took an experienced native speaker 80 hours, about one minute on average to annotate one sentence. We use chapters 1-5 of section 6 as the development data, chapters 1-5 of section 7 as the test data, and the remaining data for target-domain training,[2] in order to compare type-supervised methods with token-supervised methods. Under permission from the author of the novel, we release our annotation for future reference. Statistics of both the source and the target domain data are shown in Table 3. The rest of the novel is treated as unlabeled sentences, used for type-annotation and self-training.

We perform the standard evaluation, using F-scores for both the segmentation accuracy and the overall segmentation and POS tagging accuracy.

4.2 Baseline Performances

The baseline discriminative model can achieve state-of-the-art performances on the CTB5, with a 97.62% segmentation accuracy and a 93.85% on overall segmentation and tagging accuracy. Using the CTB model, the performance on ZX drops significantly, to a 87.71% segmentation accuracy and a 80.81% overall accuracy. Applying self-training, the segmentation and overall F-scores can be improved to 88.62% and 81.94% respectively.

4.3 Development Experiments

In this section, we study type-supervised domain adaptation by conducting a series of experiments on the development data, addressing the following questions. First, what is the influence of tag-dictionaries through lexicon-based features? Second, what is the effect of type-supervised domain adaptation in contrast to token-supervised domain adaptation under the same annotation cost? Third, what is the interaction between tag-dictionary and self-training? Finally, what is the combined effect of type- and token-supervised domain adaptation?

4.3.1 The Influence of The Tag Dictionary

We investigate the effects of two different tag dictionaries. The first dictionary contains names of characters (e.g. 鬼厉 (Guili)) and artifacts (e.g. swords such as 斩龙 (Dragonslayer)) in the novel, which are obtained from an Internet Encyclopedia,[3] and requires little human effort. We extracted 159 words from this page, verified them, and put them into a tag dictionary. We associate every word in this tag dictionary with the POS "NR (proper noun)", and name the lexicon by ***NR***.

The second dictionary was constructed manually, by first employing our baseline tagger to tag the unlabeled ZX sentences automatically,

[1] http://faculty.sutd.edu.sg/˜yue_zhang/emnlp12yang.zip

[2] We only use part of the training sentences in our experiments, and the remaining can be used for further research.

[3] http://baike.baidu.com/view/18277.htm

Model	Target-Domain Resources	Cost	Supervised		+Self-Training		
			SEG	POS	SEG	POS	ER
Baseline	—	0	89.77	82.92	90.35	83.95	6.03
Type-Supervision	NR(T)	0	89.84	83.91	91.18	85.22	8.14
	3K(T)	5h	91.93	86.53	92.86	87.67	8.46
	ORACLE(T)	∞	93.10	88.87	94.00	89.91	9.34
Token-Supervision	300(S)	5h	92.59	86.86	93.33	87.85	7.53
	600(S)	10h	93.19	88.13	93.81	89.01	7.41
	900(S)	15h	93.53	88.53	94.15	89.33	6.97
Combined Type- and Token-Supervision	3K(T) + 300(S)	10h	93.49	88.54	94.00	89.21	5.85
	3K(T) + 600(S)	15h	93.98	89.27	94.61	89.87	5.59

Table 4: Development test results, where **Cost** denotes the cost of type- or token-annotation measured by person hours, **ER** denotes the error reductions of overall performances brought by *self-training*, **T** denotes type-annotation and **S** denotes token-annotation.

and then randomly selecting the words that are not domain-independent for an experienced native speaker to annotate. To facilitate comparison with token-supervision, we spent about 5 person hours in annotating 3,000 word-tag pairs, at about the same cost as annotating 300 sentences. Finally we conjoined the 3,000 word-tag pairs with the ***NR*** lexicon, and name the resulting lexicon by ***3K***.

For the target domain, we mark the words from both ***NR*** and ***3K*** as the *domain-specific* lexicons. In all experiments, we use the same domain-independent lexicon, which is extracted from the source domain training data by HowNet matching.

The accuracies are shown in Table 4, where the ***NR*** lexicon improved the overall F-score slightly over the baseline, and the larger lexicon ***3K*** brought more significant improvements. These experiments agree with the intuition that the size and the coverage of the tag dictionary is important to the accuracies. To understand the extent to which a lexicon can improve the accuracies, we perform an oracle test, in which lexicons in the gold-standard test outputs are included in the dictionary. The accuracy is 88.87%.

4.3.2 Comparing Type-Supervised and Token-Supervised Domain Adaptation

Table 4 shows that the accuracy improvement by 3,000 annotated word-tag pairs (86.53%) is close to that by 300 annotated sentences (86.86%). This suggest that using our method, type-supervised domain adaptation can be a competitive choice to the token-supervised methods.

The fact that the token-supervised model gives slightly better results than our type-annotation method under similar efforts can probably be explained by the nature of domain differences. Texts in the Internet novel are different with CTB news in not only the vocabulary, but also POS n-gram distributions. The latter cannot be transferred from the source-domain training data directly. Texts from domains such as modern-style novels and scientific articles might have more similar POS distributions to the CTB data, and can potentially benefit more from pure lexicons. We leave the verification of this intuition to future work.

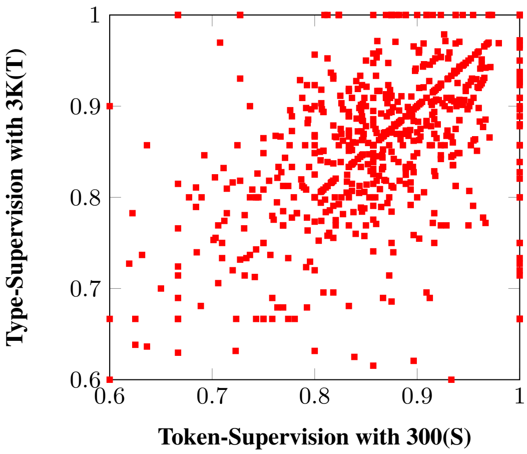

Figure 2: Sentence accuracy comparisons for type- and token-supervision with equal cost.

4.3.3 Making Use of Unlabeled Sentences

Both type- and token-supervised domain adaptation methods can be further improved via unlabeled target sentences. We apply self-training to both methods, and find improved results across the board in Table 4. The results indicate that unlabeled data is useful in further improving both type- and token-supervised domain adaptation.

Interestingly, the effects of the two methods on self-training are slightly different. The error reduction by self-training improves from 6.0% (baseline) to averaged 7.3% and 8.6% for token- and type-supervised adaptation, respectively. The better effect for the type-supervised method may result from comparatively more uniform coverage of the lexicon on sentences, since the target-domain lexicon is annotated by selecting words from much more than 300 sentences.

4.3.4 Combined Model of Type- and Token-Supervision

Figure 2 shows the F-scores of each development test sentence by type- and token-supervised domain adaptation with 5 person hours, respectively. It indicates that the two methods make different types of errors, and can potentially be used jointly for better improvements. We conduct a set of experiments as shown in Table 4, finding that the combined type- and token-supervised model with lexicon ***3K*** and 300 labeled sentences achieves an overall accuracy of 88.54%, exceeding the accuracies of both the type-supervised model with lexicon ***3K*** and the token-supervised model with 300 labeled sentences. Similar observation can be found for the combined model with lexicon ***3K*** and 600 labeled sentences. If combined with self-training, the same fact can be observed.

More interestingly, the combined model also exceeds pure type- and token-supervised models with the same annotation cost. For example, the combined model with ***3K*** and 300 labeled sentences gives a better accuracy than the token-supervised model with 600 sentences, with or without self-training. Similar observations hold between the combined model with ***3K*** and 600 labeled sentences and the token-supervised model with 900 sentences. The results suggest that the most cost-effective approach for domain adaptation can be combined type- and token-supervision: after annotating a set of raw sentences, one could stop to annotate some words, rather than continuing sentence annotation.

4.4 Final Results

Table 5 shows the final results on test corpus within ten person hours' annotation. With five person hours (lexicon ***3K***), the type-supervised model gave an error reduction of 32.99% compared with the baseline. The best result was obtained by the combined type- and token-supervised model, with an error reduction of 47.06%, higher than that the token-supervised model with the same cost under the same setting (the model of 600 labeled sentences with an error reduction of 41.53%). The results confirm that the type-supervised model is a competitive alternative for joint segmentation and POS-tagging under the cross-domain setting. Combined type- and token-supervised model yields better results than single models.

	SEG	POS	ER	Time
Baseline	87.71	80.81	0.00	0
Baseline+Self-Training	88.62	81.94	5.89	0
Type-Supervision				
NR(T)	88.34	82.54	9.02	0
NR(T)+ Self-Training	89.52	83.93	16.26	0
3K(T)	91.11	86.04	27.25	5h
3K(T)+Self-Training	92.11	87.14	32.99	5h
Token-Supervision				
300(S)	92.44	86.87	31.58	5h
300(S)+Self-Training	93.24	87.48	34.76	5h
600(S)	93.09	88.05	37.73	10h
600(S)+Self-Training	93.77	88.78	41.53	10h
Combined Type- and Token-Supervision				
3K(T)+300(S)	93.27	89.03	42.83	10h
3K(T)+300(S)+Self-Training	**93.98**	**89.84**	**47.06**	10h

Table 5: Final results on test set within ten person hours' annotation, where **ER** denotes the overall error reductions compared with *the baseline model*, **Time** denotes the cost of type- or token-annotation measured by person hours, **T** denotes type-annotation and **S** denotes token-annotation.

5 Related Work

As mentioned in the introduction, tag dictionaries have been applied to type-supervised POS tagging of English (Toutanova and Johnson, 2007; Goldwater and Griffiths, 2007; Ravi and Knight, 2009; Garrette and Baldridge, 2012), Hebrew (Goldberg et al., 2008), Kinyarwanda and Malagasy (Garrette and Baldridge, 2013; Garrette et al., 2013), and other languages (Täckström et al., 2013). These methods assume that lexicon can be obtained by manual annotation or semi-supervised learning, and use the lexicon to induce tag sequences on unlabeled sentences. We study type-supervised Chinese POS-tagging, but under the setting of domain adaptation. The problem is how to leverage a target domain lexicon and an available annotated resources in a different source domain to improving POS-tagging. Consistent with Garrette et al. (2013), we also find that the type-supervised method is a competitive choice to token-supervised adaptation.

There has been a line of work on using graph-based label propagation to expand tag-lexicons for POS-tagging (Subramanya et al., 2010; Das and Petrov, 2011). Similar methods have been applied to character-level Chinese tagging (Zeng et al., 2013). We found that label propagation from neither the source domain nor auto-labeled target domain sentences can improve domain adaptation. The main reason could be significant domain differences. Due to space limitations, we omit this negative result in our experiments.

With respect to domain adaptation, existing methods can be classified into three categories. The first category does not explicitly model differences between the source and target domains, but use standard semi-supervised learning methods with labeled source domain data and unlabeled target domain data (Dai et al., 2007; Raina et al., 2007). The baseline self-training approach (Liu and Zhang, 2012) belongs to this category. The second considers the differences in the two domains in terms of features (Blitzer et al., 2006; Daume III, 2007), classifying features into domain-independent source domain and target domain groups and training these types consistently. The third considers differences between the distributions of instances in the two domains, treating them differently (Jiang and Zhai, 2007). Our type-supervised method is closer to the second category. However, rather than splitting features into domain-independent and domain-specific types, we use domain-specific dictionaries to capture domain differences, and train a model on the source domain only. Our method can be treated as an approach specific to the POS-tagging task.

With respect to Chinese lexical analysis, little previous work has been reported on using a tag dictionary to improve joint segmentation and POS-tagging. There has been work on using a lexicon in improving segmentation in a Chinese analysis pipeline. Peng et al. (2004) used features from a set of Chinese words and characters to improve CRF-based segmentation; Low et al. (2005) extracted features based on a Chinese lexicon from Peking University to help a maximum segmentor; Sun (2011) collected 12,992 idioms from Chinese dictionaries, and used them for rule-based pre-segmentation; Hatori et al. (2012) collected Chinese words from HowNet and the Chinese Wikipedia to enhance segmentation accuracies of their joint dependency parsing systems. In comparison with their work, our lexicon contain additional POS information, and are used for word segmentation and POS-tagging simultaneously. In addition, we separate domain-dependent lexicons for the source and target lexicons, and use a novel framework to perform domain adaptation.

Wang et al. (2011) collect word-tag statistics from automatically labeled texts, and use them as features to improve POS-tagging. Their word-tag statistics can be treated as a type of lexicon. However, their efforts differ from ours in several aspects: (1) they focus on in-domain POS-tagging, while our concern is cross-domain tagging; (2) they study POS-tagging on segmented sentences, while we investigate joint segmentation and POS-tagging for Chinese; (3) their tag-dictionaries are not tag-dictionaries literally, but statistics of word-tag associations.

6 Conclusions

We performed an empirical study on the use of tag-dictionaries for the domain adaptation of joint Chinese segmentation and POS-tagging, showing that type-supervised methods can be a competitive alternative to token-supervised methods in cost-effectiveness. In addition, combination of the two methods gives the best cost-effect. Finally, we release our annotation of over 4,000 sentences in the Internet literature domain online at `http://faculty.sutd.edu.sg/~yue_zhang/eacl14meishan.zip` as a free resource for Chinese POS-tagging.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. We gratefully acknowledge the support of the National Key Basic Research Program (973 Program) of China via Grant 2014CB340503 and the National Natural Science Foundation of China (NSFC) via Grant 61133012 and 61370164, the National Basic Research Program (973 Program) of China via Grant 2014CB340503, the Singaporean Ministration of Education Tier 2 grant T2MOE201301 and SRG ISTD 2012 038 from Singapore University of Technology and Design.

References

Michele Banko and Robert C. Moore. 2004. Part-of-speech tagging in context. In *COLING*.

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*, pages 120–128, Sydney, Australia, July. Association for Computational Linguistics.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584, Cambridge, MA, October. Association for Computational Linguistics.

Michael Collins and Brian Roark. 2004. Incremental parsing with the perceptron algorithm. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 111–118, Barcelona, Spain, July.

Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2007. Transferring Naive Bayes Classifiers for Text Classification. In *AAAI*, pages 540–545.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA, June. Association for Computational Linguistics.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Zhendong Dong and Qiang Dong. 2006. *Hownet And the Computation of Meaning*. World Scientific Publishing Co., Inc., River Edge, NJ, USA.

Dan Garrette and Jason Baldridge. 2012. Type-supervised hidden markov models for part-of-speech tagging with incomplete tag dictionaries. In *EMNLP-CoNLL*, pages 821–831.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia, June. Association for Computational Linguistics.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of pos-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–592, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yoav Goldberg, Meni Adler, and Michael Elhadad. 2008. EM can find pretty good HMM POS-taggers (when given a good start). In *Proceedings of ACL-08: HLT*, pages 746–754, Columbus, Ohio, June. Association for Computational Linguistics.

Sharon Goldwater and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, Prague, Czech Republic, June. Association for Computational Linguistics.

Jun Hatori, Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2012. Incremental joint approach to word segmentation, pos tagging, and dependency parsing in chinese. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1045–1053, Jeju Island, Korea, July. Association for Computational Linguistics.

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An error-driven word-character hybrid model for joint chinese word segmentation and pos tagging. 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*, pages 513–521, Suntec, Singapore, August. Association for Computational Linguistics.

Yang Liu and Yue Zhang. 2012. Unsupervised domain adaptation for joint segmentation and POS-tagging. In *Proceedings of COLING 2012: Posters*, pages 745–754, Mumbai, India, December. The COLING 2012 Organizing Committee.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161–164.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceeding of CICLing'11*.

Bernard Merialdo. 1994. Tagging english text with a probabilistic model. *Computational Linguistics*, 20(2).

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 277–284, Barcelona, Spain, July. Association for Computational Linguistics.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of Coling 2004*, pages 562–568, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*, May.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. 2007. Self-taught learning: transfer learning from unlabeled data. In *ICML*, pages 759–766.

Sujith Ravi and Kevin Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *ACL/IJCNLP*, pages 504–512.

Amarnag Subramanya, Slav Petrov, and Fernando Pereira. 2010. Efficient graph-based semi-supervised learning of structured tagging models. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 167–176, Cambridge, MA, October. Association for Computational Linguistics.

Weiwei Sun. 2011. A stacked sub-word model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1385–1394, Portland, Oregon, USA, June. Association for Computational Linguistics.

Oscar Täckström, Dipanjan Das, Slav Petrov, McDonald Ryan, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. In *Transactions of the ACL*. Association for Computational Linguistics, March.

Partha Pratim Talukdar and Koby Crammer. 2009. New regularized algorithms for transductive learning. In *ECML/PKDD (2)*, pages 442–457.

Kristina Toutanova and Mark Johnson. 2007. A bayesian lda-based model for semi-supervised part-of-speech tagging. In *NIPS*.

Yiou Wang, Jun'ichi Kazama, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, and Kentaro Torisawa. 2011. Improving chinese word segmentation and pos tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 309–317, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Nianwen Xue, Fei Xia, Shizhe Huang, and Tony Kroch. 2000. The bracketing guidelines for the chinese treebank. Technical report, University of Pennsylvania.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Xiaodong Zeng, Derek F. Wong, Lidia S. Chao, and Isabel Trancoso. 2013. Graph-based semi-supervised model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 770–779, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yue Zhang and Stephen Clark. 2008. Joint word segmentation and POS tagging using a single perceptron. In *Proceedings of ACL-08: HLT*, pages 888–896, Columbus, Ohio, June. Association for Computational Linguistics.

Yue Zhang and Stephen Clark. 2010. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 843–852, Cambridge, MA, October. Association for Computational Linguistics.

Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105–151.