Transductive Learning of Neural Language Models for Syntactic and Semantic Analysis

Hiroki Ouchi[1,2] **Jun Suzuki**[2,1] **Kentaro Inui**[2,1]
[1] RIKEN Center for Advanced Intelligence Project [2] Tohoku University
hiroki.ouchi@riken.jp, {jun.suzuki,inui}@ecei.tohoku.ac.jp

Abstract

In transductive learning, an unlabeled test set is used for model training. While this setting deviates from the common assumption of a completely unseen test set, it is applicable in many real-world scenarios, where the texts to be processed are known in advance. However, despite its practical advantages, transductive learning is underexplored in natural language processing. Here, we conduct an empirical study of transductive learning for neural models and demonstrate its utility in syntactic and semantic tasks. Specifically, we fine-tune language models (LMs) on an unlabeled test set to obtain test-set-specific word representations. Through extensive experiments, we demonstrate that despite its simplicity, transductive LM fine-tuning consistently improves state-of-the-art neural models in both in-domain and out-of-domain settings.

1 Introduction

In supervised learning, a model is trained on a training set and its generalization performance is evaluated on an unseen test set. In this setting, the model has no access to the test set during training. However, the assumption of a completely unseen test set is not always necessary. In many cases, certain aspects of the test set are already known at training time. For example, a company may want to annotate a large number of existing documents automatically (Section 3). In such a scenario, the texts to be processed are known in advance, and using the model trained on the texts themselves to process them can be more efficient. Using an unlabeled test set in this way is the key idea behind *transductive learning*.

In transductive learning (Vapnik, 1998), an unlabeled test set is given in the training phase. That is, the inputs of the test set, i.e., the raw texts, can be used during training, but the labels are never used. In the test phase, the trained model is evaluated on the same test set. Despite its practical advantages, transductive learning has received little attention in natural language processing (NLP). After the pioneering work of Joachims (1999), who proposed a transductive support vector machine for text classification, transductive methods for linear models have been investigated in only a few tasks, such as lexical acquisition (Duh and Kirchhoff, 2006) and machine translation (Ueffing et al., 2007). In particular, transductive learning with neural networks is underexplored.

Here, we investigate the impact of transductive learning on state-of-the-art neural models in syntactic and semantic tasks, namely syntactic chunking and semantic role labeling (SRL). Specifically, inspired by recent findings that language model (LM)-based word representations yield large performance improvement (Devlin et al., 2019), we fine-tune Embeddings from Language Models (ELMo) (Peters et al., 2018) on an unlabeled test set and use them in each task-specific model. Typically, LMs are trained on a large-scale corpus whose word distributions are different from the test set. By contrast, transductive learning allows us to fit LMs directly to the distributions of the test set. Our experiments show the effectiveness of transductive LM fine-tuning.

In summary, our main contributions are:

- This work is the first to introduce an LM fine-tuning method to transductive learning[1].
- Through extensive experiments in both in-domain and out-of-domain settings, we demonstrate that transductive LM fine-tuning consistently improves state-of-the-art neural models in syntactic and semantic tasks.

[1]Our code and scripts are publicly available at https://github.com/hiroki13/transductive-language-models.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3665–3671, Hong Kong, China, November 3–7, 2019.

2 Related Work

Transductive learning. Vapnik advocated and formalized transductive learning (Vapnik, 1998; Gammerman et al., 1998), which has been applied to text classification (Joachims, 1999; Ifrim and Weikum, 2006) and image processing (Bruzzone et al., 2006; Sener et al., 2016; Liu et al., 2019). Although some studies have presented transductive methods for linear models in other tasks (Duh and Kirchhoff, 2006; Ueffing et al., 2007; Chen et al., 2008; Alexandrescu and Kirchhoff, 2009), transductive methods for neural models are underexplored in NLP.

Unsupervised domain adaptation. Transductive learning is related to unsupervised domain adaptation, in which models are adapted to a target domain by using unlabeled target domain texts (Ben-David et al., 2010; Shi and Sha, 2012). This setting does not allow models to access the test set, which is the main difference between unsupervised domain adaptation and transductive learning. Various unsupervised adaptation methods have been proposed for linear models (Blitzer et al., 2006; Jiang and Zhai, 2007; Tsuboi et al., 2009; Søgaard, 2013). In the context of neural models, adversarial domain adaptation (Ganin and Lempitsky, 2015; Ganin et al., 2016; Guo et al., 2018), importance weighting (Wang et al., 2017), structural correspondence learning (Ziser and Reichart, 2017), self/tri/co-training (Saito et al., 2017; Ruder and Plank, 2018), and other techniques orthogonal to transductive LM fine-tuning have been applied successfully in unsupervised domain adaptation[2]. Integrating these methods with transductive LM fine-tuning is an interesting direction for future research.

LM-based word representations. Recently, LM-based word representations pre-trained on unlabeled data have gained considerable attention (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019). The most related method to ours is Universal Language Model Fine-tuning (ULMFiT), which pre-trains an LM on a large general-domain corpus and fine-tunes it on the target task (Howard and Ruder, 2018). Inspired by these studies, we introduce LM-based word representation in transductive learning.

[2]Feature augmentation is considered a *supervised* domain adaptation method (Daume III, 2007; Kim et al., 2016).

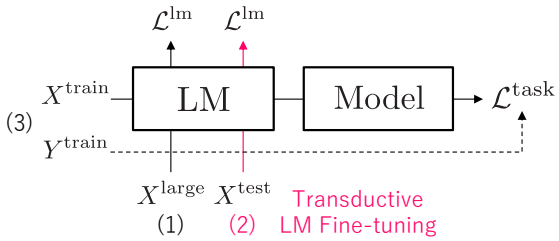

Figure 1: Training procedure. (1) LM pre-training: the LM is firstly pre-trained on the large-scale unlabeled corpus $\mathcal{D}^{\text{large}} = \{X_i^{\text{large}}\}_{i=1}^{N^{\text{large}}}$. (2) Transductive LM fine-tuning: the LM is then fine-tuned on the unlabeled test set $\mathcal{D}^{\text{test}} = \{X_i^{\text{test}}\}_{i=1}^{N^{\text{test}}}$. Note that the test set used for training is the identical one used in evaluation. (3) Task-specific model training: the task-specific model is trained on the training set $\mathcal{D}^{\text{train}} = \{(X_i^{\text{train}}, Y_i^{\text{train}})\}_{i=1}^{N^{\text{train}}}$. $\mathcal{L}$ denotes the loss function.

3 Neural Transductive Learning

Motivation. Suppose that a company has received a vast amount of customer reviews and wants to automatically process these reviews more accurately, even if it takes some time. For this purpose, they do not have to build a model that works well on new unseen reviews. Instead, they want a model that works well on only the reviews in hand. In this situation, using these reviews themselves to train a model can be more efficient. This is the key motivation for developing effective and practical transductive learning methods. Toward this goal, we develop transductive methods for state-of-the-art neural models.

Problem formulation. In the training phase, a training set $\mathcal{D}^{\text{train}} = \{(X_i^{\text{train}}, Y_i^{\text{train}})\}_{i=1}^{N^{\text{train}}}$ and an unlabeled test set $\mathcal{D}^{\text{test}} = \{X_i^{\text{test}}\}_{i=1}^{N^{\text{test}}}$ are used for model training, where X_i is an input, e.g., a sentence, and Y_i represents target labels, e.g., labels from a set of syntactic or semantic annotations. In the test phase, the trained model is used for predicting labels and is evaluated on the same test set $\mathcal{D}^{\text{test}}$.

Method. We present a simple transductive method for neural models. Specifically, we fine-tune an LM on an unlabeled test set. Figure 1 illustrates the training procedure that consists of the following steps: (1) LM pre-training, (2) Transductive LM fine-tuning and (3) task-specific model training. We first train an LM on a large-scale unlabeled corpus $\mathcal{D}^{\text{large}}$ and then fine-tune the LM on an unlabeled test set $\mathcal{D}^{\text{test}}$. Finally, we use the fine-tuned LM as the embedding layer of

	Training		Development		Test	
	Sents	Preds	Sents	Preds	Sents	Preds
BC	11.8k	28.9k	2.1k	5.0k	2,0k	5.4k
BN	10.6k	3.1k	1.2k	3,9k	1.2k	3.7k
MZ	6.9k	2.4k	0.6k	2.1k	0.7k	2.6k
NW	34.9k	96.6k	5.8k	16.6k	1.8k	5.8k
PT	21.5k	34.9k	1.7k	2.5k	1.2k	2.8k
TC	12.8k	16.2k	1.6k	2.0k	1.3k	1.7k
WB	16.9k	20.0k	2.3k	2.8k	0.9k	2.2k

Table 1: Dataset statistics on the CoNLL-2012 dataset. 1k = 1,000. Column "Sents" denotes the number of sentences in each dataset. Column "Preds" denotes the number of predicates in each dataset.

each task-specific model and train the model on a training set $\mathcal{D}^{\text{train}}$.

$$\Theta' \leftarrow \text{argmin}_{\Theta} \mathcal{L}^{\text{lm}}(\Theta|\mathcal{D}^{\text{large}}), \quad (1)$$

$$\Theta'' \leftarrow \text{argmin}_{\Theta'} \mathcal{L}^{\text{lm}}(\Theta'|\mathcal{D}^{\text{test}}), \quad (2)$$

$$\Phi' \leftarrow \text{argmin}_{\Phi} \mathcal{L}^{\text{task}}(\Phi|\Theta'', \mathcal{D}^{\text{train}}). \quad (3)$$

Here, $\mathcal{L}^{\text{lm}}$ and $\mathcal{L}^{\text{task}}$ are the loss functions for an LM and task-specific model, respectively.[3] In the LM pre-training and fine-tuning phases (Eqs. 1 and 2), we first train the initial LM parameters Θ and then fine-tune the pre-trained parameters Θ'. In the task-specific training phase (Eq. 3), we fix the fine-tuned LM parameters Θ'' used for the embedding layer of a task-specific model, and train only the task-specific model parameters Φ.

4 Experiments

Tasks. To investigate the effectiveness of transductive LM fine-tuning for syntactic and semantic analysis, we conduct experiments in syntactic chunking (Ramshaw and Marcus, 1999; Sang and Buchholz, 2000; Ponvert et al., 2011) and SRL (Gildea and Jurafsky, 2002; Palmer et al., 2005; Carreras and Màrquez, 2005)[4]. The goal of syntactic chunking is to divide a sentence into non-overlapping phrases that consist of syntactically related words. The goal of SRL is to identify semantic arguments for each predicate. For example, consider the following sentence:

	The	man	kept	a	cat
SynChunk	[NP]			[NP]	
SemRole	[A0]			[A1]	

In syntactic chunking, given the input sentence, systems have to recognize "The man" and "a cat" as noun phrases (NP). In SRL, given the input sentence and the target predicate "kept", systems have to recognize "The man" as the A0 argument and "a cat" as the A1 argument. For syntactic chunking, we adopted the experimental protocol by Ponvert et al. (2011) and for SRL, we followed Ouchi et al. (2018) (details in Appendix A).

Datasets. We perform experiments using the CoNLL-2012 dataset[5]. To investigate the performances under in-domain and out-of-domain settings, we use each of the seven domains in the CoNLL-2012 dataset. Table 1 shows the data statistics. Each test set contains at most 2,000 sentences. Compared with previous studies, such as Xiao and Guo (2013) that used 570,000 sentences as unlabeled data for unsupervised domain adaptation of syntactic chunking, our transductive experiments can be regarded as a low-resource adaptation setting. As a large-scale unlabeled raw corpus for LM training, we use the 1B word benchmark corpus (Chelba et al., 2013).

Model setup. We use ELMo (Peters et al., 2018) as an LM. For syntactic chunking, we use a variant of the Reconciled Span Parser (Joshi et al., 2018). For SRL, we use the span selection model (BiLSTM-Span model) (Ouchi et al., 2018). Each model is trained on a source domain training set and was evaluated on a target domain test set[6]. The development set is also the source domain, and it is used for hyperparameter tuning[7]. Consider the case where NW $\rightarrow$ BC, i.e., the source domain is the newswire NW and the target domain is the broadcast conversation BC. We first train ELMo on the large-scale raw corpus (one billion word benchmark corpus) and fine-tune it on the BC test set. We then train syntactic and semantic models that use the fine-tuned ELMo on the NW training set. During the task-specific model training, we freeze the fine-tuned ELMo. We select hyperparameters by using the NW development set. Finally, we evaluate the trained model on the BC test set. In the same way, we conduct training and evaluation for each domain pair.

[3] In our experiments (Section 4), both losses were given by the negative log-likelihood (Appendix A).

[4] This paper addresses span-based, PropBank-style SRL. Detailed descriptions on other lines of SRL research (e.g. dependency-based SRL and FrameNet-based SRL) can be found in Baker et al. (1998); Das et al. (2014); Surdeanu et al. (2008); Hajič et al. (2009).

[5] We used the version of OntoNotes downloaded at: http://cemantix.org/data/ontonotes.html.

[6] We used the official evaluation scripts downloaded at https://www.clips.uantwerpen.be/conll2000/chunking/ and http://www.lsi.upc.edu/ srlconll/soft.html.

[7] All models and hyperparameters are described in Appendices B, C, and D.

src → tgt	BC	BN	MZ	NW	PT	TC	WB	Averaged F1
				SYNTACTIC CHUNKING				
BC	93.0 / **93.5**	92.9 / **93.0**	90.0 / **90.6**	88.1 / **88.7**	94.1 / **94.9**	84.5 / **85.1**	89.4 / **89.8**	90.3 / **90.8**
BN	92.5 / **93.0**	94.7 / **95.0**	91.2 / **91.4**	90.0 / **90.6**	94.8 / **95.3**	84.0 / **84.9**	89.9 / **90.7**	91.0 / **91.6**
MZ	89.9 / **90.8**	91.6 / **92.3**	92.3 / **92.5**	89.2 / **90.0**	93.4 / **94.3**	80.5 / **82.2**	89.9 / **90.9**	89.5 / **90.4**
NW	91.4 / **92.0**	93.7 / **93.9**	92.2 / **92.6**	94.2 / **94.5**	95.7 / **96.1**	83.3 / **84.2**	92.3 / **92.9**	91.8 / **92.3**
PT	87.1 / **88.2**	86.9 / **87.5**	85.6 / **86.9**	81.0 / **82.7**	97.5 / **97.7**	79.0 / **80.0**	86.9 / **88.1**	86.3 / **87.3**
TC	87.3 / **88.2**	87.2 / **87.5**	84.1 / **85.4**	80.8 / **82.3**	93.0 / **93.7**	89.3 / **89.5**	85.4 / **86.5**	86.7 / **87.6**
WB	91.8 / **92.3**	93.4 / **93.7**	91.7 / **92.2**	91.0 / **91.5**	95.6 / **96.0**	83.6 / **85.1**	93.0 / **93.5**	91.4 / **92.0**
				SEMANTIC ROLE LABELING				
BC	83.3 / **83.9**	78.9 / **79.3**	74.2 / **74.8**	71.0 / **72.4**	82.8 / **84.4**	80.2 / **80.6**	78.7 / **79.8**	78.4 / **79.3**
BN	80.3 / **81.2**	83.3 / **83.5**	76.5 / **77.4**	75.0 / **75.7**	86.5 / **86.8**	77.1 / **78.0**	78.8 / **79.9**	79.6 / **80.4**
MZ	76.4 / **77.3**	76.6 / **77.3**	80.2 / **80.6**	73.8 / **74.8**	84.8 / **87.2**	72.8 / **73.3**	77.5 / **78.7**	77.4 / **78.5**
NW	79.2 / **80.1**	79.8 / **80.0**	79.5 / **80.0**	83.8 / **84.4**	88.3 / **89.0**	75.5 / **76.5**	81.1 / **81.8**	81.0 / **81.7**
PT	71.2 / **72.1**	67.4 / **67.8**	66.6 / **68.0**	64.7 / **66.0**	92.8 / **93.0**	72.6 / **73.9**	76.2 / **77.2**	73.1 / **74.0**
TC	73.8 / **74.1**	67.6 / **67.8**	64.5 / **64.9**	59.2 / **60.2**	79.0 / **80.4**	83.3 / **83.6**	71.3 / **72.5**	71.2 / **71.9**
WB	74.1 / **74.4**	71.7 / **72.4**	72.0 / **72.8**	71.4 / **72.0**	87.8 / **88.8**	76.3 / **76.7**	81.8 / **82.4**	76.4 / **77.1**

Table 2: Main results under cross-domain settings, src ("source", training set) → tgt ("target", test set). Cells show the F1 scores of the baseline model (before the slash) and the transductive model (after the slash). Column "Averaged F1" represents the F1 scores averaged across the target domains. Domains are as follows: BC = Broadcast Conversation, BN = Broadcast News, MZ = Magazine, NW = Newswire, PT = New Testament, TC = Telephone Conversation, and WB = Weblogs and Newsgroups.

Results. Table 2 shows the F1 scores on each test set. All reported F1 scores are the average of five distinct trials using different random seeds. In each cell, the left-hand side denotes the F1 score of the baseline (using a base LM without fine-tuning) and the right-hand side denotes F1 of the transductive models (using a fine-tuned LM on each test set). In in-domain (same source/target domains, e.g., BC→BC) and out-of-domain (different source/target domains, e.g., BC→NW) settings, all transductive models consistently outperformed the baselines, which suggests that transductive LM fine-tuning improves performance of neural models. Although the improvements were undramatic (around 1.0 F1 gain), these consistent improvements can be regarded as valuable empirical results because of the difficulty of *unsupervised* and *low-resource* adaptation settings.

5 Analysis

Comparison between unsupervised domain adaptation and transduction. In unsupervised domain adaptation, target domain unlabeled data (the texts whose domain is the same as that of a test set) is used for adaptation. Although the domain is identical between target domain data and a test set, their word distributions are somewhat different. In transductive learning, because an unlabeled test set can be used for training, it is possible to adapt LMs directly to the word distributions of the test set. Here, we investigate whether adapting LMs directly to each test set is more effective than adapting LMs to each target domain unlabeled data. Similarly to our transductive method shown in Figure 1, we first train LMs on the large-scale unlabeled corpus (the 1B word benchmark corpus) and then fine-tune them on the unlabeled target domain data[8]. In addition, we control the sizes of the target domain unlabeled data and test sets. That is, we use the same number of sentences in the unlabeled data of each target domain as in each test set. Table 3 shows the F1 scores averaged across all the target domains. The transductive models (T) consistently outperformed the domain-adapted models (CU). This demonstrates that adapting LMs directly to test sets is more effective than adapting them to target domain unlabeled data.

	Syntactic chunking		Semantic role labeling	
	CU	T	CU	T
BC	90.4	**90.8**	78.6	**79.3**
BN	91.1	**91.6**	79.8	**80.4**
MZ	90.0	**90.4**	77.9	**78.5**
NW	92.1	**92.3**	81.1	**81.7**
PT	87.1	**87.3**	73.5	**74.0**
TC	87.1	**87.6**	71.3	**71.6**
WB	91.8	**92.0**	76.6	**77.1**

Table 3: Performance comparison between LM fine-tuning on target domain unlabeled data of the same size as each test set, "Controlled Unlabeled data (CU)," and transductive LM fine-tuning on each test set (T). Cells show the F1 scores averaged across the target domains.

[8] As target domain unlabeled data, we use the CoNLL-2012 training set of each domain.

	Syntactic chunking		Semantic role labeling	
	U	U + T	U	U + T
BC	90.5	**91.0**	79.0	**79.4**
BN	91.3	**91.6**	80.1	**80.6**
MZ	90.2	**90.6**	78.3	**78.7**
NW	92.1	**92.5**	81.5	**81.9**
PT	87.3	**87.7**	73.6	**74.3**
TC	87.2	**87.6**	71.4	**72.0**
WB	91.8	**92.2**	76.8	**77.2**

Table 4: Performance comparison between LM fine-tuning on target domain unlabeled data (U) and on the combination of the unlabeled data and test sets (U + T). Cells show the F1 scores averaged across the target domains.

CoNLL	2000	2005		2012
		WSJ	Brown	
BASE	96.6	87.7	78.3	86.2
TRANS	**96.7**	**87.9***	**79.5***	**86.6***
Clark et al. (2018)	97.0	-	-	-
Peters et al. (2017)	96.4	-	-	-
Hashimoto et al. (2017)	95.8	-	-	-
Wang et al. (2019)	-	88.2	79.3	86.4
Li et al. (2019)	-	87.7	80.5	86.0
Ouchi et al. (2018)	-	87.6	78.7	86.2
He et al. (2018)	-	87.4	80.4	85.5

Table 5: Standard benchmark results. Cells show the F1 scores on each test set. The CoNLL-2000 and CoNLL-2005/2012 datasets are used for syntactic chunking and SRL, respectively. Results of the transductive models (TRANS) marked with * are statistically significant compared to the baselines (BASE) using the permutation test ($p < 0.05$).

Combination of unsupervised domain adaptation and transduction. In real-world situations, large-scale unlabeled data of target domains is sometimes available. In such cases, LMs can be trained on both the target domain unlabeled data and the test sets. Here, we investigate the effectiveness of using both datasets. Table 4 shows the F1 scores averaged across all the target domains. Fine-tuning the LMs on the target domain unlabeled data as well as each test set (U + T) showed better performance than fine-tuning them only on the target domain unlabeled data (U). This combination of tranduction with unsupervised domain adaptation further improves performance.

Effects in standard benchmarks. Some studies indicated that when promising new techniques are only evaluated on very basic models, determining how much (if any) improvement will carry over to stronger models can be difficult (Denkowski and Neubig, 2017; Suzuki et al., 2018). Motivated by such studies, we provide the results in standard benchmark settings. For syntactic chunking, we use the CoNLL-2000 dataset (Sang and Buchholz, 2000) and follow the standard experimental protocol (Hashimoto et al., 2017). For SRL, we use the CoNLL-2005 (Carreras and Màrquez, 2005) and CoNLL-2012 datasets (Pradhan et al., 2012) and follow the standard experimental protocol (Ouchi et al., 2018). Table 5 shows the F1 scores of our models and those of existing models. The results of the baseline model were comparable with those of the state-of-the-art models, and the transductive model consistently outperformed the baseline model[9]. Note that we cannot fairly compare the transductive and existing models due to the difference in settings. These results, however, demonstrate that transductive LM fine-tuning improves state-of-the-art chunking and SRL models.

6 Conclusion

In this study, we investigated the impact of transductive learning on state-of-the-art neural models in syntactic and semantic tasks. Specifically, we fine-tuned an LM on an unlabeled test set. Through extensive experiments, we demonstrated that, despite its simplicity, transductive LM fine-tuning contributes to consistent performance improvement of state-of-the-art syntactic and semantic models in cross-domain settings. One interesting line of future work is to explore effective transductive methods for task-dependent (neural) layers. For instance, as some unsupervised domain adaptation methods can be applied to transductive learning, integrating them with transductive LM fine-tuning may further improve their performance. Another line of our future work is to apply these transductive methods to various NLP tasks and investigate their performance.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number JP19H04162 and JP19K20351. We would like to thank Benjamin Heinzerling, Ana Brassard, Sosuke Kobayashi, Hitomi Yanaka, and the anonymous reviewers for their insightful comments.

[9]While the improvements in SRL were statistically significant compared to the baseline, the improvement in syntactic chunking was not. One reason for this is that the F1 score of the baseline in syntactic chunking is already high and there is less room for improvement. Since Clark et al. (2018) achieved 97.0 F1 with multi-task learning, missing information for further improvement might be derived from other tasks.

References

Andrei Alexandrescu and Katrin Kirchhoff. 2009. Graph-based learning for statistical machine translation. In *Proceedings of HLT-NAACL*, pages 119–127.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of ACL-COLING*, pages 86–90.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of ACL*, page 120.

Lorenzo Bruzzone, Mingmin Chi, and Mattia Marconcini. 2006. A novel transductive svm for semisupervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3363–3373.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL*, pages 152–164.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Yaodong Chen, Ting Wang, Huowang Chen, and Xishan Xu. 2008. Semantic role labeling of chinese using transductive svm and semantic heuristics. In *Proceedings of IJCNLP*, pages 919–924.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of EMNLP*, pages 1914–1925.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*, pages 256–263.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Kevin Duh and Katrin Kirchhoff. 2006. Lexicon acquisition for dialectal arabic using transductive learning. In *Proceedings of EMNLP*, pages 399–407.

Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik. 1998. Learning by transduction. In *Proceedings of Uncertainty in artificial intelligence*, pages 148–155.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of ICML*, pages 1180–1189.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of EMNLP*, pages 4694–4703.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*, pages 1–18.

Kazuma Hashimoto, Yoshimasa Tsuruoka, Richard Socher, et al. 2017. A joint many-task model: Growing a neural network for multiple nlp tasks. In *Proceedings of EMNLP*, pages 1923–1933.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of ACL*, pages 364–369.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of ACL*, pages 328–339.

Georgiana Ifrim and Gerhard Weikum. 2006. Transductive learning for text classification using explicit knowledge models. In *Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery*, pages 223–234.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of ACL*, pages 264–271.

Thorsten Joachims. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of ICML*, pages 200–209.

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of ACL*, pages 1190–1199.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016. Frustratingly easy neural domain adaptation. In *Proceedings of COLING*, pages 387–396.

Zuchao Li, Shexia He, Hai Zhano, Yiqing Zhang, Zhuosheng Zhang, Zhou. Xi, and Xiang Zhou. 2019. Dependency or span, end-to-end uniform semantic role labeling. In *Proceedings of AAAI*.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, and Yi Yang. 2019. Transductive propagation network for few-shot learning. In *Proceedings of ICLR*.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2018. A span selection model for semantic role labeling. In *Proceedings of EMNLP*, pages 1630–1642.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of ACL*, pages 1756–1765.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*, pages 2227–2237.

Elias Ponvert, Jason Baldridge, and Katrin Erk. 2011. Simple unsupervised grammar induction from raw text with cascaded finite state models. In *Proceedings of ACL*, pages 1077–1086.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Proceedings of CoNLL*, pages 1–40.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of ACL*, pages 1044–1054.

Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of ICML*, pages 2988–2997.

Erik F Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking.

Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. 2016. Learning transferrable representations for unsupervised domain adaptation. In *Proceedings of NIPS*, pages 2110–2118.

Yuan Shi and Fei Sha. 2012. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *Proceedings of ICML*, pages 1275–1282.

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

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of CoNLL*, pages 159–177.

Jun Suzuki, Sho Takase, Hidetaka Kamigaito, Makoto Morishita, and Masaaki Nagata. 2018. An empirical study of building a strong baseline for constituency parsing. In *Proceedings of ACL*, pages 612–618.

Yuta Tsuboi, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. 2009. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Transductive learning for statistical machine translation. In *Proceedings of ACL*, pages 25–32.

Vladimir Vapnik. 1998. *Statistical learning theory*. Wiley.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of EMNLP*, pages 1482–1488.

Yufei Wang, Mark Johnson, Stephen Wan, Yifang Sun, and Wei Wang. 2019. How to best use syntax in semantic role labelling. In *Proceedings of ACL*, pages 5338–5343.

Min Xiao and Yuhong Guo. 2013. Domain adaptation for sequence labeling tasks with a probabilistic language adaptation model. In *Proceedings of ICML*, pages 293–301.

Yftah Ziser and Roi Reichart. 2017. Neural structural correspondence learning for domain adaptation. In *Proceedings of CoNLL*, pages 400–410.