A Regularization-based Framework for Bilingual Grammar Induction*

Yong Jiang[◇], **Wenjuan Han**[†], **Kewei Tu**[†]
[†]School of Information Science and Technology, ShanghaiTech University, Shanghai, China
[◇]Alibaba Group
yongjiang.jy@alibaba-inc.com
{hanwj,tukw}@shanghaitech.edu.cn

Abstract

Grammar induction aims to discover syntactic structures from unannotated sentences. In this paper, we propose a framework in which the learning process of the grammar model of one language is influenced by knowledge from the model of another language. Unlike previous work on multilingual grammar induction, our approach does not rely on any external resources, such as parallel corpora, word alignments or linguistic phylogenetic trees. We propose three regularization methods that encourage similarity between model parameters, dependency edge scores, and parse trees respectively. We deploy our methods on a state-of-the-art unsupervised discriminative parser and evaluate it on both transfer grammar induction and bilingual grammar induction. Empirical results on multiple languages show that our methods outperform strong baselines.

1 Introduction

Syntactic parsing is an important task in natural language processing. Supervised parsing requires manual labeling of gold parse trees, which is a very labor-intensive task. On the other hand, unsupervised parsing (a.k.a. grammar induction) does not require labeled data and can make use of large amounts of unlabeled data that are freely available. However, grammar induction is very challenging and its accuracy is still far below that of supervised parsing. To compensate the lack of supervision in grammar induction, some previous work considers multilingual grammar induction, i.e., simultaneously learning grammars of multiple languages (Snyder et al., 2009; Berg-Kirkpatrick and Klein, 2010; Liu et al., 2013). Existing multilingual approaches require external resources such as parallel corpora, word alignments, and linguistic phylogenetic trees.

Language	German	English	Spanish	French	Indonesian
Code	DE	EN	ES	FR	ID
C-MST	**60.2**	62.3	**68.8**	**72.3**	**69.7**
D-Tran	59.9	–	65.3	67.8	45.7
Δ	-0.3	–	-3.5	-4.5	-24.0
Language	Italian	Japanese	Korean	Portuguese	Swedish
Code	IT	JA	KO	PTBR	SV
C-MST	64.3	**57.5**	**59.0**	**68.3**	66.2
D-Tran	63.1	54.6	50.0	66.2	**67.8**
Δ	-1.2	-2.9	-9.0	-2.1	+1.6

Table 1: Directed dependency accuracy (DDA) on the universal treebanks with universal POS tags, on sentences of length $\leq$ 10. C-MST denotes the original Convex-MST model. D-Tran denotes direct transfer. Δ refers to the difference between C-MST and D-Tran.

In this paper, we aim at bilingual grammar induction without external resource. We are motivated by our observation that learning the unsupervised Convex-MST model (Grave and Elhadad, 2015) on the English corpus and then directly applying it to parse other languages produces surprisingly good results (Table 1). From the table, we can see that even with this simplistic method (which we call *direct transfer*), the dependency accuracy on each language is often very close to the accuracy of the model specifically trained on the corpus of that language. For the Swedish language, the accuracy of direct transfer is even better than that of the specifically trained model. This surprising result suggests that grammars of different languages, even those from different language families (e.g., English and Japanese), may have non-trivial similarity that can be helpful in bilingual grammar induction.

Inspired by this observation, we propose a regularization-based framework to bilingual grammar induction that encourage knowledge sharing between models learned on a language pair. We build our framework on top of Convex-MST, a state-of-the-art unsupervised dependency parser,

* Yong Jiang contributed to this work when at ShanghaiTech University. Kewei Tu is the corresponding author.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1423–1428,
Hong Kong, China, November 3–7, 2019.

and propose three regularization terms that encourage similarity between model parameters, edge scores, and parse trees respectively. We test our methods on ten languages on the tasks of transfer grammar induction and bilingual grammar induction and show that our methods can achieve a significant boost over strong baselines.

2 Background

2.1 Unsupervised Dependency Parsing

Dependency parsing is the task of mapping an input sentence $\mathbf{x} = x_1, x_2, ..., x_n$ of length n to an output dependency structure $\mathbf{y}$. A dummy root x_0 is typically added at the beginning of the sentence to denote the head of the dependency tree. There are several approaches to represent the parse tree $\mathbf{y}$. In transition based dependency parsers, the dependency tree can be regarded as a sequence of actions. In graph based dependency parsers, the dependency tree can be represented as a spanning tree in the graph. In chart based parsers (a.k.a., grammar based parsers), the dependency tree is denoted as a set of grammar rules. In unsupervised graph based dependency parsers, since gold trees are not available, carefully designed models and objective functions are required for learning a dependency parser. Regardless of model architectures, current unsupervised dependency models usually use the following form of objective function,

$$J(\mathbf{w}; \mathcal{X}) = \sum_{\mathbf{x} \in \mathcal{X}} \mathrm{O}_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \Big(D(\mathbf{w}, \mathbf{x}, \mathbf{y}) + R(\mathbf{w}) \Big)$$

where $\mathcal{Y}$ is the set of all possible dependency tree, $\mathbf{w}$ is the model parameter, $\mathcal{X}$ is the unlabeled training corpus, D is the measurement between the parse $\mathbf{y}$ and model prediction on sentence $\mathbf{x}$, $R(\mathbf{w})$ is the regularization term of parameter $\mathbf{w}$, $\mathrm{O} \in \{\min, \sum\}$ is an operator. Table 2 shows the choices of O, D and R for several widely used models.

2.2 Graph based Dependency Parsing

In this paper, we focus on graph based dependency parsers, though we believe that our approaches can be generalized to other types of parsers. Previous work on unsupervised graph based dependency parsing utilizes the autoencoder structure (Cai et al., 2017) or the discriminative clustering techniques (Grave and Elhadad, 2015).

Following (McDonald et al., 2005), we can use a discriminative model for dependency parsing with first order factorization such that the score of a dependency tree $\mathbf{y}$ is the sum of the scores of its dependency edges. The score of an edge from word h to word m , $\mathbf{s_w}(\mathbf{x}, h, m)$, can be computed as the inner product of a feature vector $\mathbf{f}(\mathbf{x}, h, m)$ and a parameter vector $\mathbf{w}$. The optimal dependency tree for sentence $\mathbf{x}$ be discovered in polynomial time (Eisner, 1996; McDonald et al., 2005).

3 Bilingual Knowledge Sharing

Given non-parallel corpora of two languages $\mathcal{X}_s$ and $\mathcal{X}_t$, our goal is to learn two models with parameters $\mathbf{w}_s$ and $\mathbf{w}_t$ for the two languages. The simplest learning objective function is,

$$J(\mathbf{w}_s, \mathbf{w}_t; \mathcal{X}_s, \mathcal{X}_t) = J(\mathbf{w}_s; \mathcal{X}_s) + J(\mathbf{w}_t; \mathcal{X}_t)$$

which contains no interaction between the two models.

As suggested by our empirical observation in Table 1, the model of one language may provide a useful inductive bias in learning the model of another language. Note that given a sentence, a graph-based dependency parser has three levels of representations: the model parameters, the scores of dependency edges computed from the parameters, and the parse tree computed from the edge scores. Therefore, we propose three different regularization terms to effectively encourage similarity of the two models. An example is shown in Figure 1.

Regularization of Weight Parameters (W-Reg) Motivated by the approach of Berg-Kirkpatrick and Klein (2010), we encourage the similarity between the two weight parameters $\mathbf{w}_s$ and $\mathbf{w}_t$ measured by l2 norm distance:

$$\begin{aligned} &J(\mathbf{w}_s, \mathbf{w}_t; \mathcal{X}_s, \mathcal{X}_t) = \\ &J(\mathbf{w}_s; \mathcal{X}_s) + J(\mathbf{w}_t; \mathcal{X}_t) + \lambda ||\mathbf{w}_s - \mathbf{w}_t||_2^2 \end{aligned}$$

in which λ is a hyper-parameter.

Regularization on Edge Scores (E-Reg) Directly encouraging weight similarity might result in an inductive bias that is too strong, because the difference between the two languages (e.g., different word orders) may lead to different meanings of each feature dimension. Therefore, we propose to encourage similarity between the scores computed by the two models for each dependency edge of

Parsers	O	D	R
DMV (Klein and Manning, 2004)	$\sum$	negative log likelihood	-
Convex-MST (Grave and Elhadad, 2015)	min	ℓ_2 distance	ℓ_2 norm
LC-DMV (Noji et al., 2016)	$\sum$	negative log likelihood	ℓ_2 norm
NDMV (Jiang et al., 2016)	$\sum$, min	negative log likelihood	-
CRFAE (Cai et al., 2017)	min	negative conditional log likelihood	ℓ_1 norm
D-NDMV (Han et al., 2019)	$\sum$, min	negative (conditional) log likelihood	-

Table 2: Notations of several widely used models.

each sentence, which can be seen as a soft version of weight regularization.

$$J(\mathbf{w}_s, \mathbf{w}_t; \mathcal{X}_s, \mathcal{X}_t) = J(\mathbf{w}_s; \mathcal{X}_s) + J(\mathbf{w}_t; \mathcal{X}_t) + \lambda \sum_{\mathbf{x}\in\mathcal{X}'} \sum_{(h,m)\in\mathcal{G}(\mathbf{x})} ||\mathbf{s}_{\mathbf{w}_s}(\mathbf{x}, h, m) - \mathbf{s}_{\mathbf{w}_t}(\mathbf{x}, h, m)||_2^2$$

where $\mathcal{X}' = \mathcal{X}_s \cup \mathcal{X}_t$. $\mathcal{G}(\mathbf{x})$ is the weighted dependency graph of sentence $\mathbf{x}$.

Regularization on Parse Trees (T-Reg) Another alternative is to encourage similarity between the parse trees predicted by the two models. Motivated by the idea of knowledge distillation (Kim and Rush, 2016), in the learning objective of each model, we add a fourth term to encourage the parse tree to be close to the prediction of the other model. Below we show the objective function for $\mathbf{w}_s$.

$$J'(\mathbf{w}_s, \mathbf{w}_t; \mathcal{X}_s) = \sum_{\mathbf{x}\in\mathcal{X}_s} \mathrm{O}_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \Big(D(\mathbf{w}_s, \mathbf{x}, \mathbf{y}) + \lambda \underbrace{D(\mathbf{w}_t, \mathbf{x}, \mathbf{y})}_{\textbf{T-Reg term}} + R(\mathbf{w}_s) \Big)$$

$$J'(\mathbf{w}_t, \mathbf{w}_s; \mathcal{X}_t) = \sum_{\mathbf{x}\in\mathcal{X}_t} \mathrm{O}_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \Big(D(\mathbf{w}_t, \mathbf{x}, \mathbf{y}) + \lambda \underbrace{D(\mathbf{w}_s, \mathbf{x}, \mathbf{y})}_{\textbf{T-Reg term}} + R(\mathbf{w}_t) \Big)$$

$$J(\mathbf{w}_s, \mathbf{w}_t; \mathcal{X}_s, \mathcal{X}_t) = J'(\mathbf{w}_s, \mathbf{w}_t; \mathcal{X}_s) + J'(\mathbf{w}_t, \mathbf{w}_s; \mathcal{X}_t)$$

We apply these regularization method to the Convex-MST model (Grave and Elhadad, 2015). Our three objective functions can be optimized with coordinate descent in a similar way to Convex-MST. In each iteration, we first fix parse

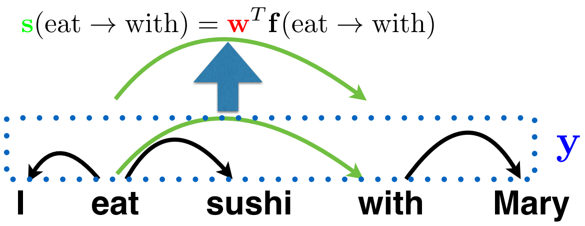

Figure 1: Three levels of representations of the parser: the parameter $\mathbf{w}$, the edge score $\mathbf{s}$, and the parse $\mathbf{y}$.

$\mathbf{y}$ for each training sentence and update parameters $\mathbf{w}_s$ and $\mathbf{w}_t$ by stochastic gradient descent; then we fix $\mathbf{w}_s$ and $\mathbf{w}_t$ and update $\mathbf{y}$ of each sentence by the Frank-Wolfe algorithm.

While our three methods are applicable to any pair of languages, intuitively one may use weight regularization only for similar languages, and use edge regularization and tree regularization for an arbitrary language pair.

4 Experiments

To enable direct comparison with the Convex-MST model, we use the dataset used in their paper (Grave and Elhadad, 2015), the universal treebanks version 2.0[1], introduced by McDonald et al. (2013). The dataset contains ten different languages, which belong to five diverse families. In additional, we test our methods on twelve languages from the more recent UD Treebank 1.4[2], which is also used in previous grammar induction work (Jiang et al., 2017; Li et al., 2019). Following previous work, we train all the models on the gold POS tags of sentences no longer than ten. We tune hyper-parameters on the development dataset and report the DDA on sentences no longer than ten and all the sentences in the test dataset. As our goal is to investigate the benefits of our regularization methods, the two hyper-parameters μ and β

[1]https://github.com/ryanmcd/uni-dep-tb. The version is not consistent with recent releases of UD Treebanks.

[2]http://universaldependencies.org/

CODE	C-MST	D-TRAN	W-REG	E-REG	T-REG
DE	60.2	-0.3	-0.2	**+0.2**	-0.2
ES	68.8	-3.5	-3.5	**+0.8**	+0.3
FR	72.3	-4.5	-3.8	**+0.3**	**+0.3**
ID	**69.7**	-24.0	-21.4	-0.6	-1.2
IT	64.3	-1.2	-0.4	+0.3	**+1.2**
JA	57.5	-2.9	-3.4	+0.9	**+2.3**
KO	59.0	-9.0	-9.5	+1.3	**+1.9**
PTBR	68.3	-2.1	-2.1	+0.2	**+0.3**
SV	66.2	+1.6	+1.6	**+2.7**	+2.6
Avg	65.14	-5.10	-4.74	+0.68	**+0.83**
Avg-All	56.16	-4.63	-4.29	+0.47	**+0.56**
UD 1.4*	52.83	-0.30	-1.92	+1.06	**+1.37**

Table 3: Transfer grammar induction from English to the other languages. We show accuracies on the test sentences of length $\leq$ 10 (except the Avg-All row which shows the average accuracies on all the sentences). *: for the UD 1.4 dataset we show the average results.

of Convex-MST are tuned on the English development dataset and then fixed ($\mu = 0.1, \beta = 0.001$) while λ is selected from $\{10, 5, 1, 5e-1, 1e-1, 5e-2, 1e-2, 5e-3, 1e-3, 1e-4\}$ for each language pair.

4.1 Experiments on Transfer Grammar Induction

In transfer grammar induction, we train the first model on the first language independent of the second language; then, with the first model fixed, we optimize our knowledge sharing objective with respect to the second model; finally, we evaluate the second model on the test set of the second language. In this way, we want to test whether our methods can transfer useful linguistic knowledge from the first language to the second language. We report the results of transfer grammar induction from English to the other nine languages in Table 3. Our edge regularization and tree regularization methods outperform the Convex-MST baseline in almost all the cases. The weight regularization method achieves worse results than Convex-MST except for the Swedish language, which demonstrates that directly regularizing weight parameters may not work well in the transfer grammar induction task. For the Swedish language, although direct transfer already achieves better performance than Convex-MST, our regularization methods can further boost the performance by a large margin. The Indonesian language is the only language for which transfer grammar induction provides no benefit, possibly because of its significant syntactic difference from the English language. Our additional experimental results on UD treebank 1.4 show a similar trend.

PAIR	CODE	BASE	COMB	W-REG	E-REG	T-REG
EN-DE	EN	62.3	+0	**+0.5**	+0.1	+0.2
	DE	60.2	+0.1	-0.8	**+0.5**	+0.4
EN-ES	EN	62.3	+0.2	**+0.7**	+0.4	**+0.7**
	ES	68.8	-1.8	**+1.4**	+1.1	+0.8
EN-FR	EN	62.3	+0.4	**+0.5**	+0.1	+0.1
	FR	72.3	-3.6	-0.6	**+1.4**	**+1.4**
EN-ID	EN	62.3	+0.7	**+1.3**	+0.1	+0.7
	ID	69.7	-19.2	-2.1	+0.5	**+0.9**
EN-IT	EN	62.3	+0.7	+0.1	+0.6	**+1.0**
	IT	64.3	+0.7	**+1.8**	+0.4	+0.8
EN-JA	EN	62.3	+1.0	**+0.9**	-0.1	-0.2
	JA	57.5	-4.3	-0.2	**+2.7**	+2.4
EN-KO	EN	62.3	+0.3	**+1.4**	+0.5	+1.1
	KO	59.0	-3.8	+0.1	**+1.0**	+0.4
EN-PTBR	EN	62.3	+1.0	**+0.8**	+0.6	+0.7
	PT-BR	68.3	-0.9	**+1.4**	**+1.4**	+1.2
EN-SV	EN	62.3	-0.5	-0.9	**+0.1**	**+0.1**
	SV	66.2	+1.7	**+1.9**	+1.0	+1.1
Avg	EN	62.30	+0.42	**+0.58**	+0.23	+0.49
	Other	65.14	-3.46	+0.32	**+1.11**	+1.04
Avg-All	EN	52.10	-0.48	-0.40	+0.11	**+0.42**
	Other	56.16	-3.05	+0.64	**+1.55**	+1.21
UD 1.4*	EN	53.50	-0.76	-0.09	+ 0	**+0.71**
	Other	52.83	-0.09	+0.58	**+1.92**	+1.52

Table 4: Results of bilingual grammar induction on test sentences no longer than 10 (except the Avg-All row which shows the average accuracies on all the sentences). BASE refers to the individually trained baseline. COMB refers to learning a single model from the combined training set of the two languages. *: for the UD 1.4 dataset we show the average results.

We perform transfer grammar induction from English to Swedish with different values of λ and show the results in Figure 2. We can see that the impact of different hyper-parameter values on the accuracy generally follows the same tendency for our three methods.

4.2 Experiments on Bilingual Grammar Induction

In bilingual grammar induction, we jointly train two models on two languages. In our experiments, we pair English with each of the other nine languages. The results are reported in Table 4. It can be seen that in most cases joint training leads to better accuracies than the individually trained models as well as the single model learned from the combined training set. By comparing table 4 with table 3, we can also see that bilingual joint training leads to better accuracies than transfer grammar induction, which shows the benefit of training two models simultaneously rather than sequentially. Again, our additional experimental results on UD treebank 1.4 show a similar trend.

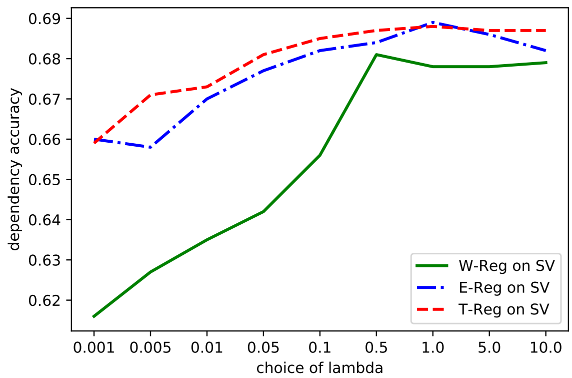

Figure 2: Transfer grammar induction on the SV language with different hyper-parameter values for the three regularization methods

5 Related Work

Our work is related to many previous work.

Unsupervised Transfer Learning There has been previous work aiming at solving an unsupervised learning task of a target domain with the help of knowledge learned from a source domain (Dai et al., 2008; Wang et al., 2008; Pan and Yang, 2010). There is no labeled data in both the source and the target domains during training. Our transfer grammar induction setting can be seen as an instance of unsupervised transfer learning.

Cross-lingual Supervised Dependency Parsing This task focuses on learning a parser with unlabeled training data and additional labeled training data of a second language (McDonald et al., 2011; Naseem et al., 2012; Guo et al., 2015). The main difference between our approach and theirs is that our approach is fully unsupervised. and do not utilize external information like word alignments or cross-lingual word embeddings.

Other Approaches to Multilingual Grammar Induction To the best of our knowledge, this task is first proposed by Kuhn (2004). They assume that the syntax trees induced from parallel sentences share structured regularities and utilize the word alignments to guide parsing. From then on, many approaches are proposed on both constituency grammar induction and dependency grammar induction (Snyder et al., 2009; Berg-Kirkpatrick and Klein, 2010). We differ from these approaches in that we do not make use of any external rules or knowledge.

6 Conclusion

In this paper, we propose three regularization-based knowledge sharing methods to bilingual grammar induction problems. We test our methods on transfer grammar induction and bilingual grammar induction and show that our methods achieve better performance than the baselines. In future work, we plan to investigate the effectiveness of our approach in other types of induction tasks.

Acknowledgments

This work was supported by the Major Program of Science and Technology Commission Shanghai Municipal (17JC1404102).

References

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1288–1297, Uppsala, Sweden. Association for Computational Linguistics.

Jiong Cai, Yong Jiang, and Kewei Tu. 2017. Crf autoencoder for unsupervised dependency parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1638–1643, Copenhagen, Denmark. Association for Computational Linguistics.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. 2008. Self-taught clustering. In *Proceedings of the 25th international conference on Machine learning*, pages 200–207. ACM.

Jason M Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 340–345. Association for Computational Linguistics.

Edouard Grave and Noémie Elhadad. 2015. A convex and feature-rich discriminative approach to dependency grammar induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1375–1384, Beijing, China. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1234–1244.

Wenjuan Han, Yong Jiang, and Kewei Tu. 2019. Enhancing unsupervised generative dependency parser with contextual information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5315–5325, Florence, Italy. Association for Computational Linguistics.

Yong Jiang, Wenjuan Han, and Kewei Tu. 2016. Unsupervised neural dependency parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 763–771, Austin, Texas. Association for Computational Linguistics.

Yong Jiang, Wenjuan Han, and Kewei Tu. 2017. Combining generative and discriminative approaches to unsupervised dependency parsing via dual decomposition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Dan Klein and Christopher D Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 478. Association for Computational Linguistics.

Jonas Kuhn. 2004. Experiments in parallel-text based grammar induction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 470. Association for Computational Linguistics.

Bowen Li, Jianpeng Cheng, Yang Liu, and Frank Keller. 2019. Dependency grammar induction with a neural variational transition-based parser. In *AAAI 2019*.

Kai Liu, Yajuan Lü, Wenbin Jiang, and Qun Liu. 2013. Bilingually-guided monolingual dependency grammar induction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072, Sofia, Bulgaria. Association for Computational Linguistics.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 91–98. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*, pages 62–72. Association for Computational Linguistics.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637. Association for Computational Linguistics.

Hiroshi Noji, Yusuke Miyao, and Mark Johnson. 2016. Using left-corner parsing to encode universal structural constraints in grammar induction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 33–43, Austin, Texas. Association for Computational Linguistics.

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

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. 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 73–81, Suntec, Singapore. Association for Computational Linguistics.

Zheng Wang, Yangqiu Song, and Changshui Zhang. 2008. Transferred dimensionality reduction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 550–565. Springer.