Appendix

Anonymous EMNLP-IJCNLP submission

A Details in Posterior Regularization

Expectation rewriting

$$\begin{aligned} R(C,q) &\leq r \\ \frac{\sum_k \sum_{(i,j):(k,i,j)\in C^+} q_k(i,j)}{\sum_k \sum_{(i,j):(k,i,j)\in C^+\cup C^-} q_k(i,j)} &\leq r \\ (1-r)\sum_{(k,i,j)\in C^+} q_k(i,j) - r\sum_{(k,i,j)\in C^-} q_k(i,j) &\leq 0 \\ \sum_{(k,i,j)} q_k(i,j)\phi(k,i,j) &\leq 0 \\ \mathbb{E}_{\mathbf{y}\sim q}[\phi(\mathbf{y})] &\leq 0 \end{aligned}$$

Let $\phi'(k,i,j) = -\phi(k,i,j)$, we have

$$\begin{aligned} R(C,q) &\geq r \\ \frac{\sum_k \sum_{(i,j):(k,i,j)\in C^+} q_k(i,j)}{\sum_k \sum_{(i,j):(k,i,j)\in C^+\cup C^-} q_k(i,j)} &\geq r \\ (1-r)\sum_{(k,i,j)\in C^+} q_k(i,j) - r\sum_{(k,i,j)\in C^-} q_k(i,j) &\geq 0 \\ \sum_{(k,i,j)} q_k(i,j)\phi(k,i,j) &\geq 0 \\ \sum_{(k,i,j)} q_k(i,j)\phi'(k,i,j) &\leq 0 \\ \mathbb{E}_{\mathbf{y}\sim q}[\phi'(\mathbf{y})] &\leq 0 \end{aligned}$$

Dual form solving This dual form of the optimization problem is given in Ganchev et al. (2010) as

$$q^*(\mathbf{y}) = \frac{p_\theta(\mathbf{y}|\mathbf{w})\exp(-\lambda^*\cdot\phi(\mathbf{y}))}{Z(\lambda^*)},$$

$$\lambda^* = \arg\max_{\lambda\geq 0} -\log Z(\lambda),$$

$$Z(\lambda) = \sum_{\mathbf{y}'} p_\theta(\mathbf{y}'|\mathbf{w})\exp(-\lambda^*\cdot\phi(\mathbf{y}')).$$

Noting that both the feature function ϕ and model p_θ can be easily factorized to arc-level, we can set

$$q_k^*(i,j) = p_\theta(y_k(i,j)|\mathbf{w}_k)\exp(-\lambda^*\cdot\phi(k,i,j))$$

as our model on target language.

The dual form is not easy to solve since the number of possible $\mathbf{y}'$ can be exponentially large. We need to do factorization to make it tractable.

$$\begin{aligned} & Z(\lambda) \\ =& \sum_{\mathbf{y}'} p_\theta(\mathbf{y}'|\mathbf{w})\exp(-\lambda\cdot\phi(\mathbf{w},\mathbf{y}')) \\ =& \sum_{\mathbf{y}'}\prod_k p_\theta(\mathbf{y}_k|\mathbf{w}_k)\exp(-\lambda\cdot\phi(\mathbf{w}_k,\mathbf{y}_k)) \\ =& \sum_{\mathbf{y}'}\prod_k \prod_{(i,j):\mathbf{y}_k(i,j)=1} p_\theta(y_k(i,j)|\mathbf{w}_k)\exp(-\lambda\cdot\phi(y_k(i,j)) \\ =& \sum_{\mathbf{y}'}\prod_k \prod_{(i,j):\mathbf{y}_k(i,j)=1} q_k^*(i,j) \\ =& \prod_{k\in[N]}\prod_{i\in[L_k]}\sum_{j\in[L_k]} q_k^*(i,j) \end{aligned}$$

$$\begin{aligned} & \log Z(\lambda) \\ =& \sum_{k\in[N]}\sum_{i\in[L_k]}\log\sum_{j\in[L_k]} q_k^*(i,j) \end{aligned}$$

$$\begin{aligned} & \frac{\partial \log Z(\lambda)}{\partial\lambda} \\ =& \sum_{k\in[N]}\sum_{i\in[L_k]}\log\frac{\sum_{j\in[L_k]} -\phi(k,i,j)q_k^*(i,j)}{\sum_{j\in[L_k]} q_k^*(i,j)} \end{aligned}$$

The Hessian matrix of $Z(\lambda)$ is given by

$$H(Z(\lambda)) = \sum_{\mathbf{y}'} p_\theta(\mathbf{y}'|\mathbf{w})\exp(-\lambda\cdot\phi)[\phi^T\phi].$$

Noting that $\phi^T\phi$ is positive semi-definite, and $p_\theta(\mathbf{y}'|\mathbf{w})\exp(-\lambda\cdot\phi) \geq 0$, so $H(Z(\lambda))$ is also positive semi-definite, which means $Z(\lambda)$ is convex, and $\log Z(\lambda)$ is quasi-convex. We can sample b instances from dataset and compute the gradient

Hyper-parameter	Value
Input word dimension	300
Input pos dimension	50
Encoder layer	6
Encoder d_{model}	350
Encoder d_{ff}	512
Arc MLP size	512
Label MLP size	128
Training dropout	0.2
Optimizer	Adam
Learning rate	0.0001
Batch size	80

Table 1: Hyper-parameters in our model.

Hyper-parameter	Lagrangian	PR
Initial learning rate	50	1
Learning rate decay	0.9	0.98
Maximal iteration	60	100
Batch size	full batch	128

Table 2: Hyper-parameters in our model.

to estimate the full gradient, and apply stochastic gradient descent to get the optimal λ^*. We can verify that $Z(\lambda)$ is convex, and $\log Z(\lambda)$ is quasi-convex. We can sample b instances from dataset and compute the gradient to estimate the full gradient, and apply stochastic gradient descent or Adam to get the optimal λ^*.

B Hyper-parameters

Model We follow the hyper-parameters used in Ahmad et al. (2019) shown in table 4.

Inference The hyper-parameters in inference algorithms, Lagrangian relaxation and posterior regularization, are shown in table 2.

C Language Family

The languages we selected and their language families are shown in Table 3

D Entire experiment results

The entire experiment results are shown in Table 4

E Efficientiveness of constraints figure

The figure is shown in Figure 1

Language Families	Languages
Afro-Asiatic	Arabic (ar), Hebrew (he)
Austronesian	Indonesian (id)
Dravidian	Tamil (ta)
Turkic	Turkish (tr)
IE.Celtic	Welsh (cy)
IE.Baltic	Latvian (lv)
IE.Germanic	Danish (da), Dutch (nl), English (en)
IE.Indic	Hindi (hi), Urdu (ur)
IE.Latin	Latin (la)
IE.Romance	Catalan (ca), French (fr)
IE.Slavic	Bulgarian (bg), Croatian (hr)
Korean	Korean (ko)
Uralic	Estonian (et), Finnish (fi)

Table 3: The selected languages for experiments from UD v2.2 (Nivre et al., 2018)).

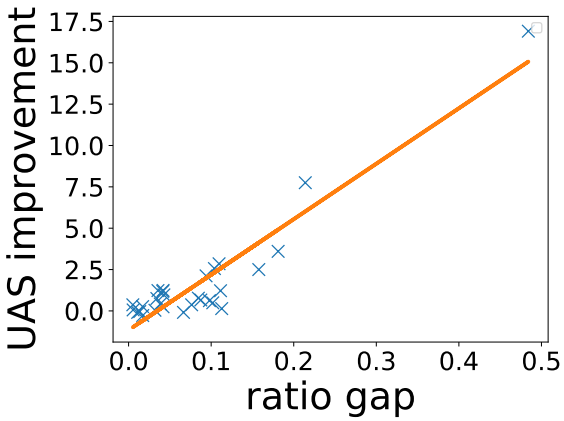

Figure 1: Ratio gap v.s. Δ perf.

Figure 2: The performance improvement is highly correlated to the difference in corpus linguistic statistics (estimated by weighted average ratio gaps in constraints) between target and source languages. (The Pearson Correlation Coefficient is 0.938.)

Family	Lang.	Features	Baseline	Lagrangian Relaxation			Posterior Regularization		
				Oracle	WALS	ΔWALS	Oracle	WALS	ΔWALS
IE.Indic	ur	-1,-1,1	18.3	35.2	34.0	+15.7	35.0	33.7	+15.4
IE.Indic	hi	-1,-1,1	34.3	52.4	53.4	+19.1	51.3	49.1	+14.8
Dravidian	ta	-1,-1,1	36.1	42.8	43.4	+7.3	43.1	43.0	+6.9
Turkic	tr	-1,-1,1	31.2	35.2	37.1	+5.9	35.1	36.3	+5.1
Afro-Asiatic	ar	1, 1,-1	38.5	47.3	45.3	+6.8	45.8	43.7	+5.2
Afro-Asiatic	he	1, 1, 1	55.7	58.8	57.6	+1.9	58.3	57.6	+1.9
Austronesian	id	1, 1, 1	49.3	53.1	52.3	+3.0	52.3	51.9	+2.6
Korean	ko	-1,-1,1	34.0	37.1	37.2	+3.2	36.3	36.4	+2.4
IE.Celtic	cy	1, 1,-1	47.3	54.2	51.7	+4.4	53.8	50.0	+2.7
IE.Slavic	hr	1, 1, 1	62.2	63.7	63.2	+1.0	63.6	63.4	+1.2
IE.Slavic	bg	1, 1, 1	79.6	79.7	79.2	+0.0	79.7	79.7	+0.1
IE.Slavic	cs	1, 1, 1	63.0	63.9	64.0	+1.0	63.8	63.6	+0.6
IE.Slavic	pl	1, 1, 1	74.6	74.8	73.6	-1.0	75.0	74.8	+0.2
IE.Slavic	ru	1, 1, 1	60.6	61.6	61.2	+0.6	61.4	61.4	+0.8
IE.Slavic	sk	?, ?, ?	66.8	66.3	67.9	+1.1	67.0	66.8	+0.0
IE.Slavic	sl	1, 1, 1	67.8	67.9	67.9	+0.1	67.9	67.9	+0.1
IE.Slavic	uk	1, 1, 1	59.9	62.1	60.9	+1.0	61.9	61.1	+0.2
IE.Romance	ca	1, 1,-1	73.9	74.9	73.8	-0.1	74.9	74.7	+0.8
IE.Romance	fr	1, 1,-1	77.8	79.1	78.7	+0.9	79.0	79.0	+1.2
IE.Romance	it	1, 1,-1	80.9	82.0	80.3	-0.6	81.8	81.4	+0.5
IE.Romance	pt	1, 1,-1	76.8	77.5	76.0	-0.8	77.4	77.6	+0.8
IE.Romance	ro	1, 1,-1	65.8	67.7	66.3	+0.5	67.7	66.9	+1.1
IE.Romance	es	1, 1,-1	74.6	75.8	74.2	-0.4	75.6	74.2	-0.4
IE.Baltic	lv	1, 1, 1	70.3	70.7	69.5	-0.8	70.5	69.9	-0.4
IE.Latin	la	?, ?, ?	47.4	48.0	45.6	-1.8	48.1	47.3	-0.1
Uralic	et	1,-1, 1	65.3	65.5	65.8	+0.5	65.7	66.0	+0.7
Uralic	fi	1,-1, 1	66.7	67.1	67.0	+0.3	66.9	67.1	+0.4
IE.Germanic	da	1, 1, 1	76.6	76.6	76.5	-0.1	76.6	76.6	+0.0
IE.Germanic	nl	0, 1, 1	67.5	67.6	67.5	+0.0	67.9	67.9	+0.4
IE.Germanic	de	0, 1, 1	70.6	70.8	70.6	+0.0	70.6	70.6	+0.0
IE.Germanic	no	1, 1, 1	80.5	80.4	80.5	+0.0	80.5	80.5	+0.0
IE.Germanic	sv	1, 1, 1	80.3	80.5	80.5	+0.2	80.5	80.5	+0.2
Average Performance		61.1	63.8	63.2	+2.2	63.6	63.1	+2.0	

Table 4: Proposed constraints performance compared with baseline, corpus-statistics constraints compiled from WALS features and model utilize the same annotation efforts data.

References

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(Jul):2001–2049.

Joakim Nivre, Mitchell Abrams, Željko Agić, and et al. 2018. Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.