Structure-Level Knowledge Distillation For Multilingual Sequence Labeling

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, Kewei Tu

School of Information Science and Technology, ShanghaiTech University
DAMO Academy, Alibaba Group

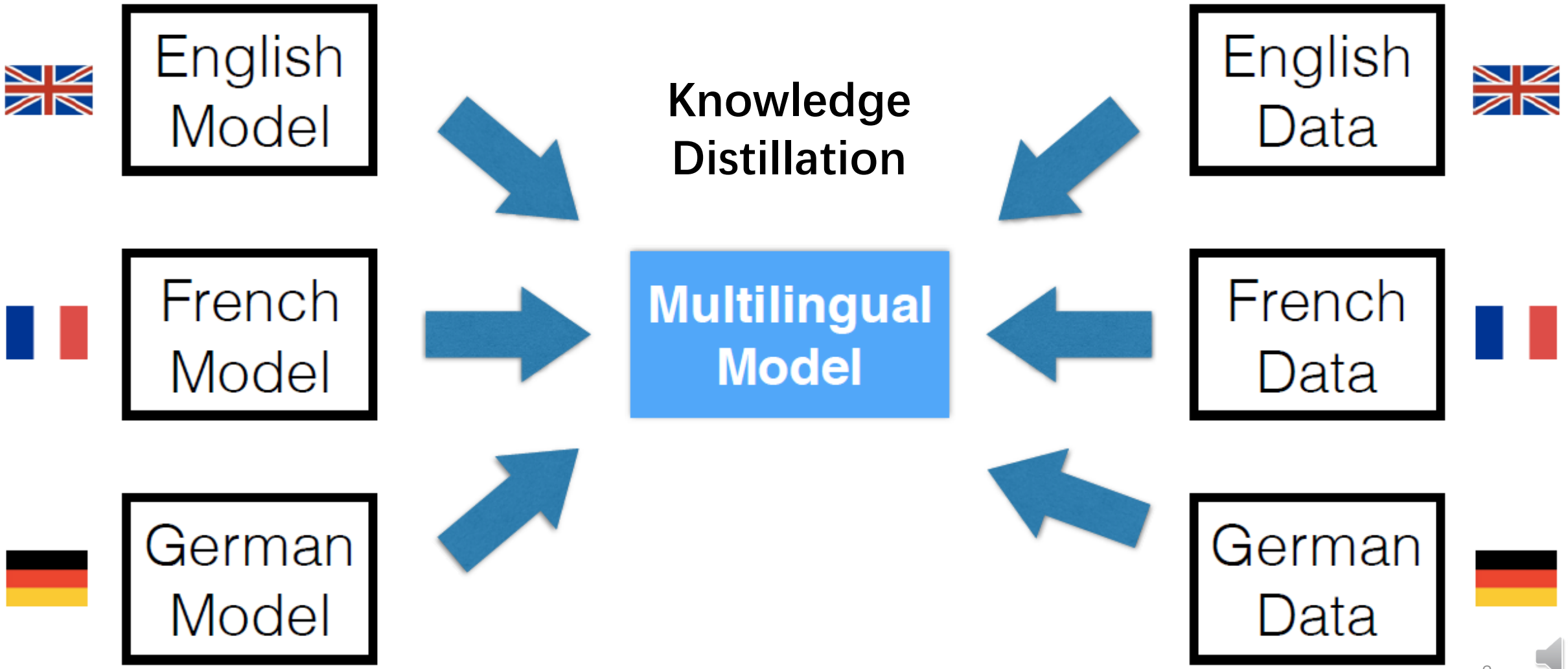
上海科技大学
ShanghaiTech University

DAMO
ALIBABA DAMO ACADEMY

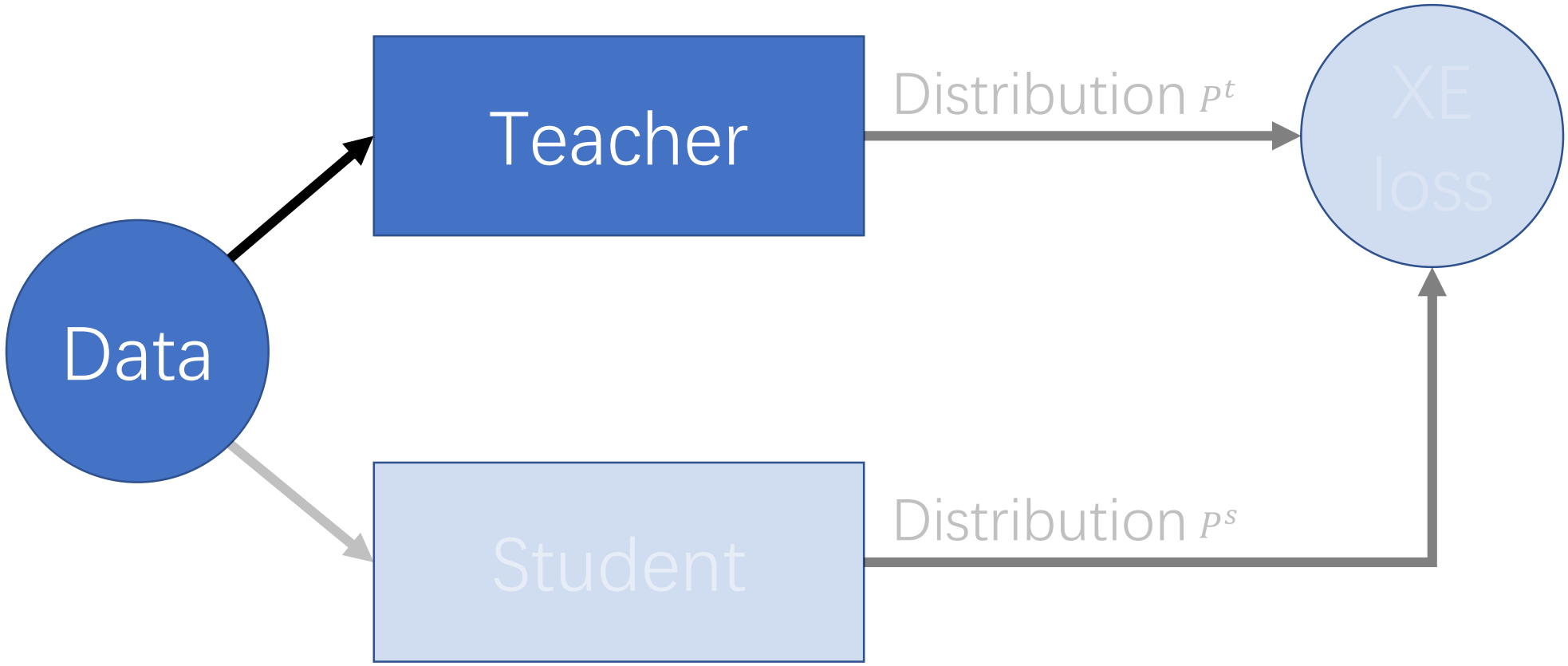
Motivation

- Most of the previous work of sequence labeling focused on monolingual models.
- It is resource consuming to train and serve multiple monolingual models online.
- A unified multilingual model: smaller, easier, more generalizable.
- However, the accuracy of the existing unified multilingual model is inferior to monolingual models.

Our Solution

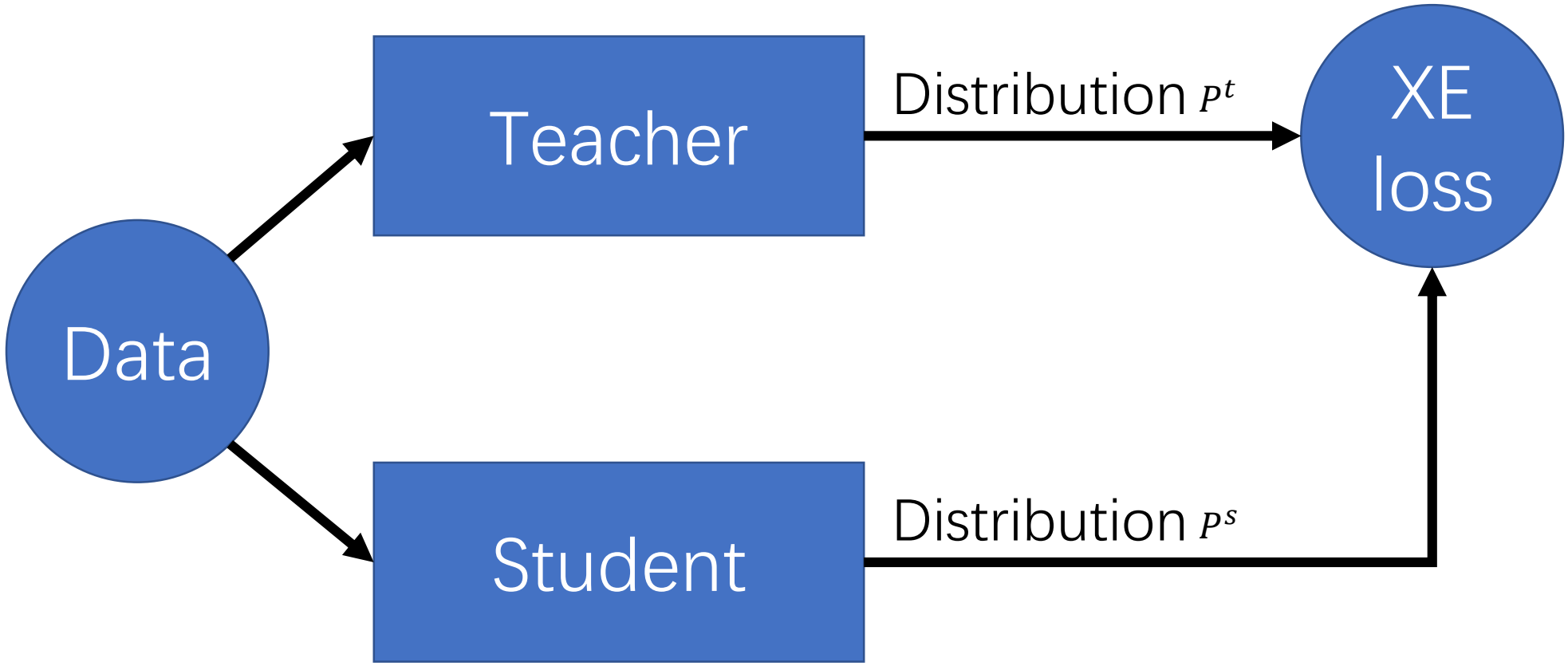


Background: Knowledge Distillation



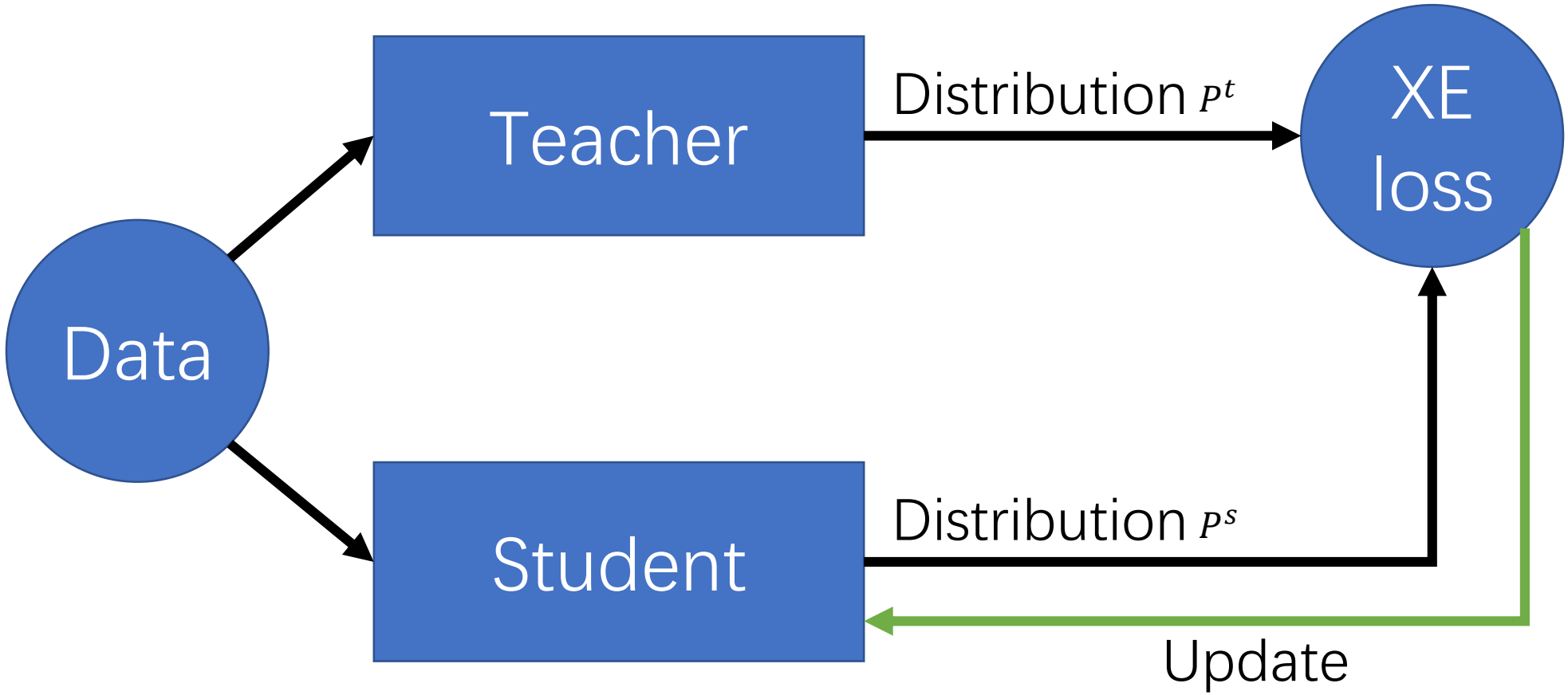
Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2014. Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop.

Background: Knowledge Distillation



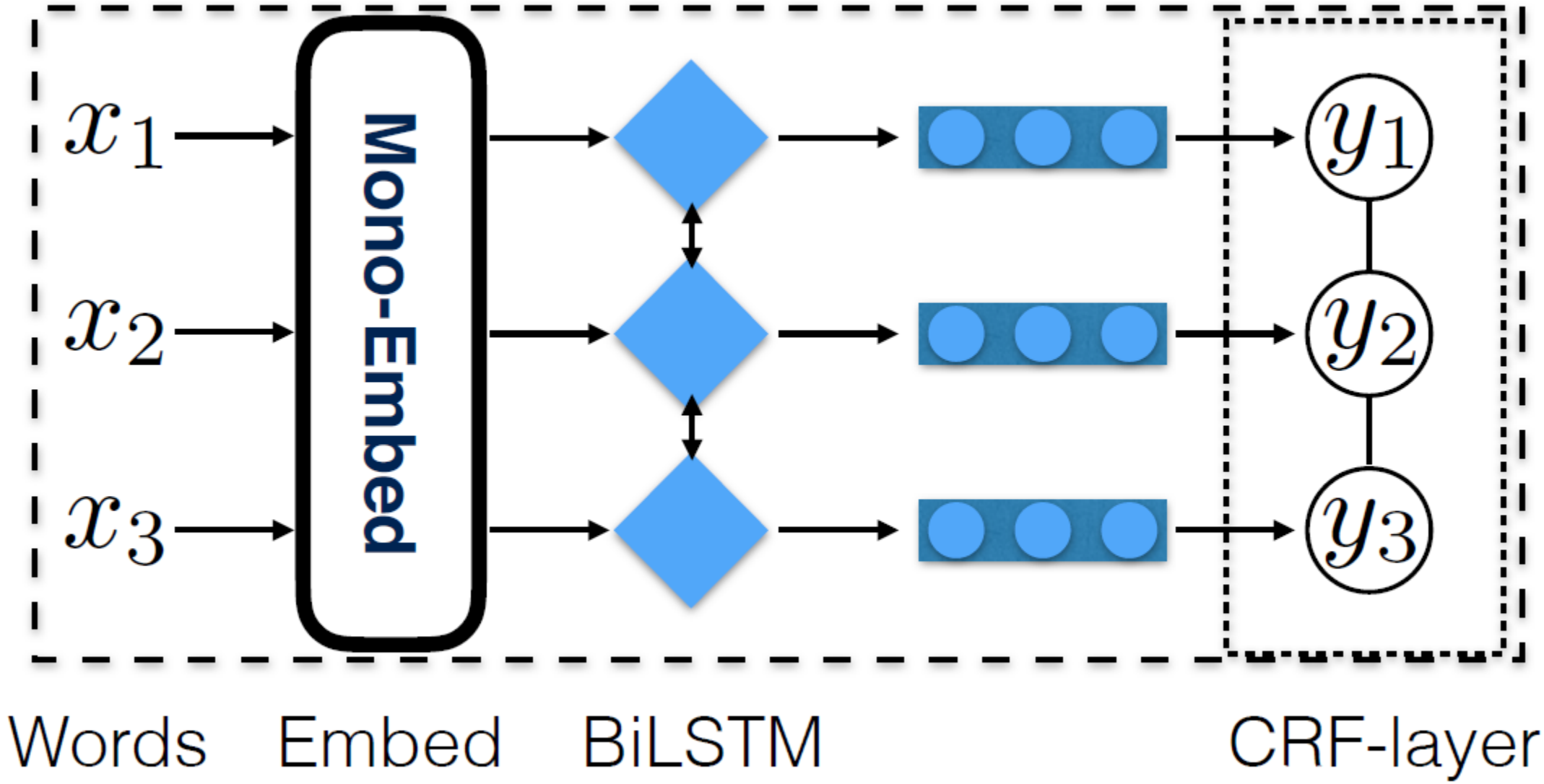
Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2014. Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop.

Background: Knowledge Distillation



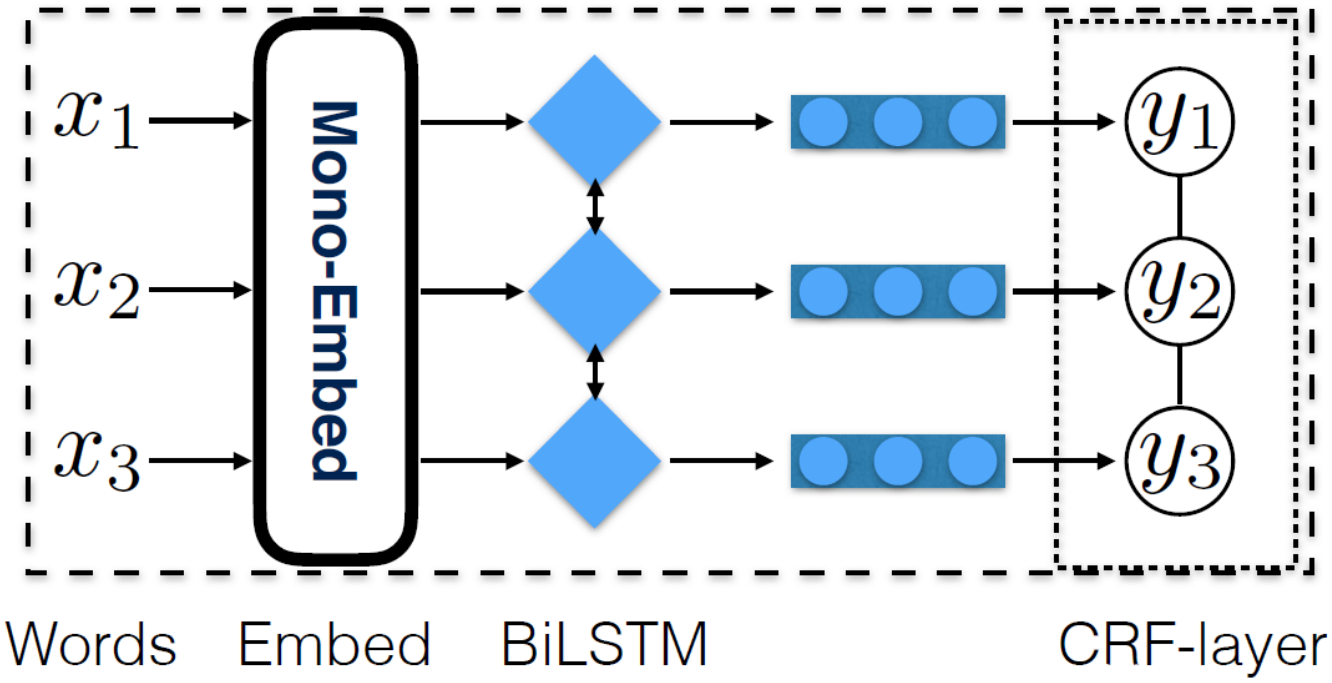
Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2014. Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop.

Background: Sequence Labeling



Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. NAACL 2016. Neural architectures for named entity recognition.

Background: Sequence Labeling

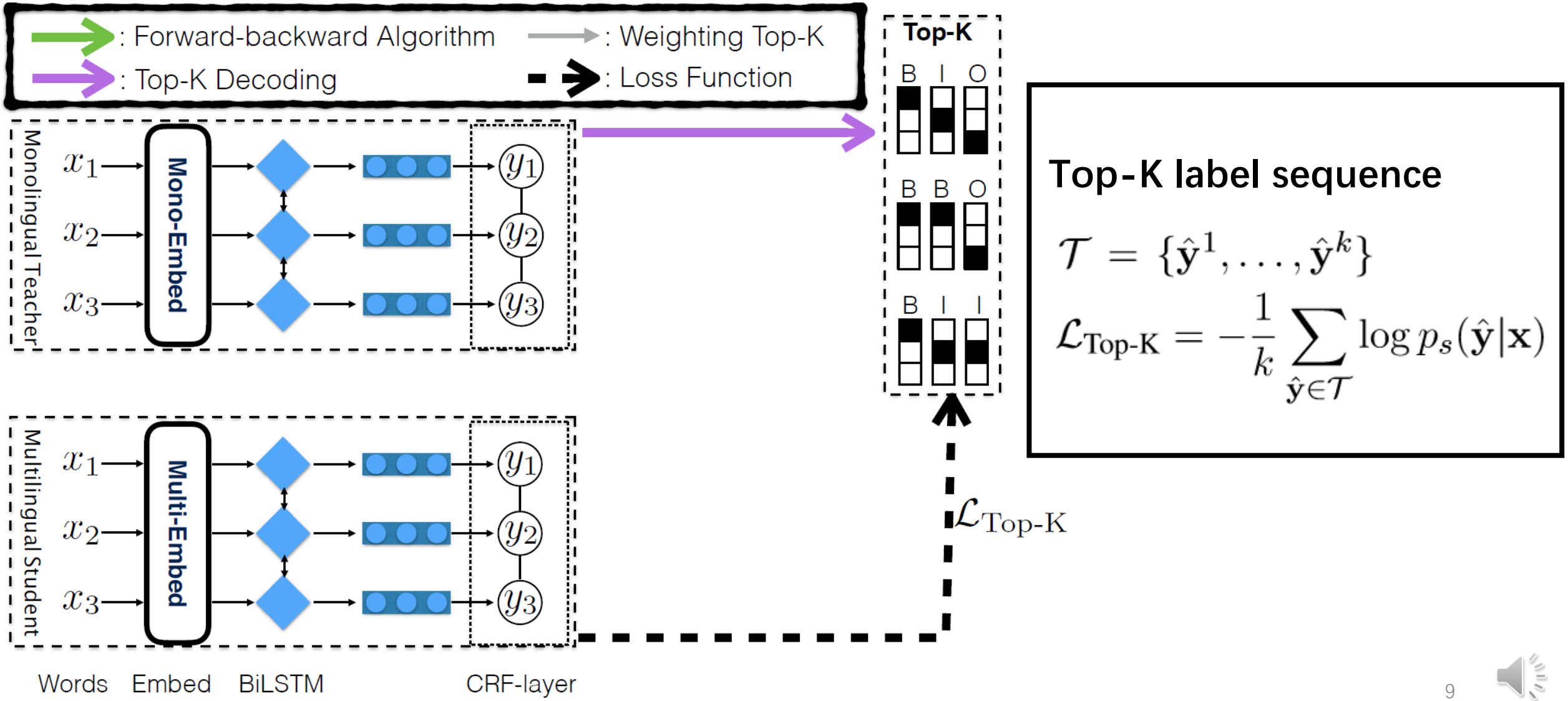


$$\mathcal{L}_{\text{Str}} = -\sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} p_t(\mathbf{y}|\mathbf{x}) \log p_s(\mathbf{y}|\mathbf{x})$$

Exponentially number of possible labeled sequences

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. NAACL 2016. Neural architectures for named entity recognition.

Top-K Distillation



Top-K label sequence

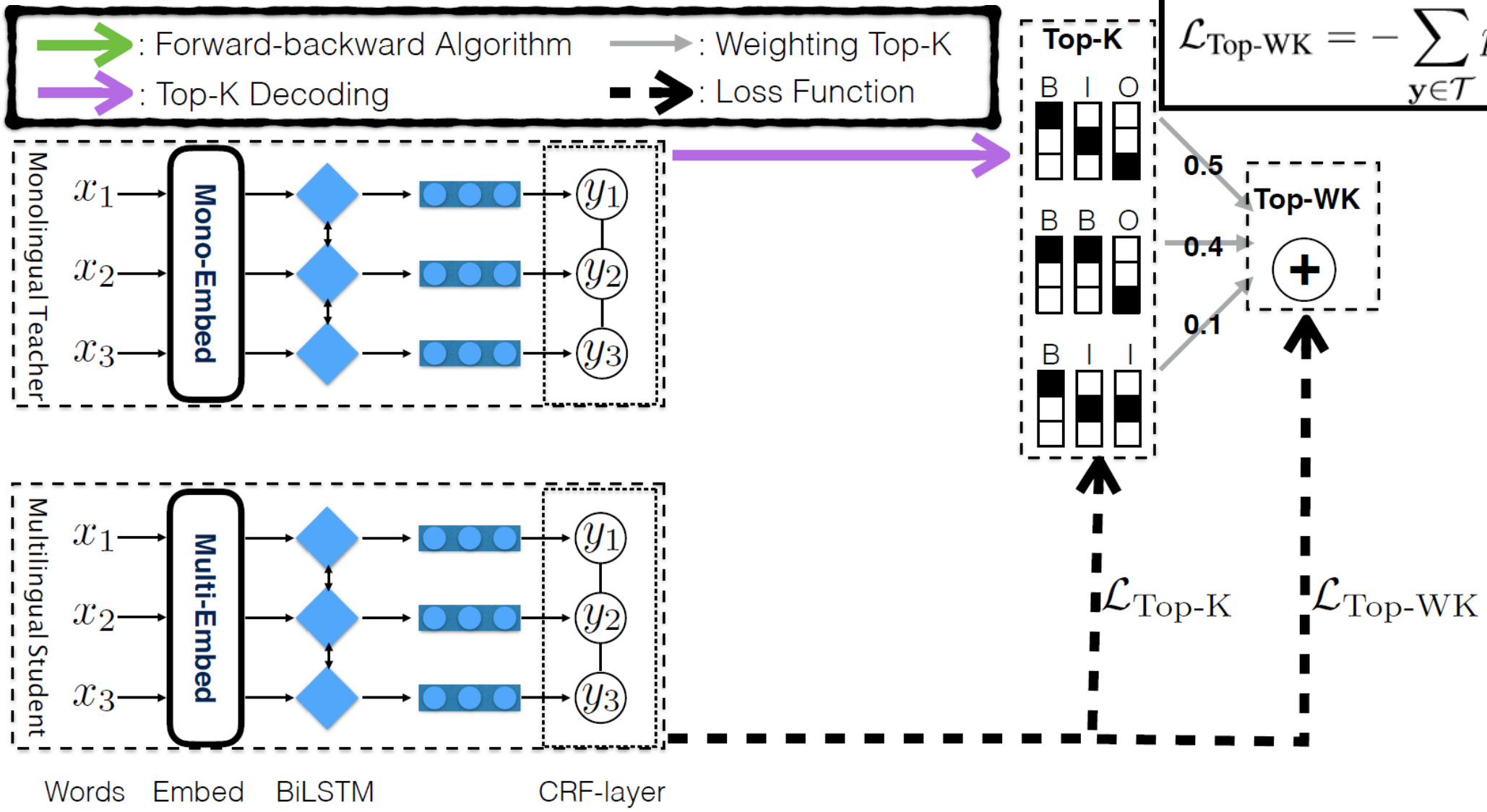
$$\mathcal{T} = \{\hat{\mathbf{y}}^1, \ldots, \hat{\mathbf{y}}^k\}$$

$$\mathcal{L}_{\text{Top-K}} = -\frac{1}{k} \sum_{\hat{\mathbf{y}} \in \mathcal{T}} \log p_s(\hat{\mathbf{y}}|\mathbf{x})$$

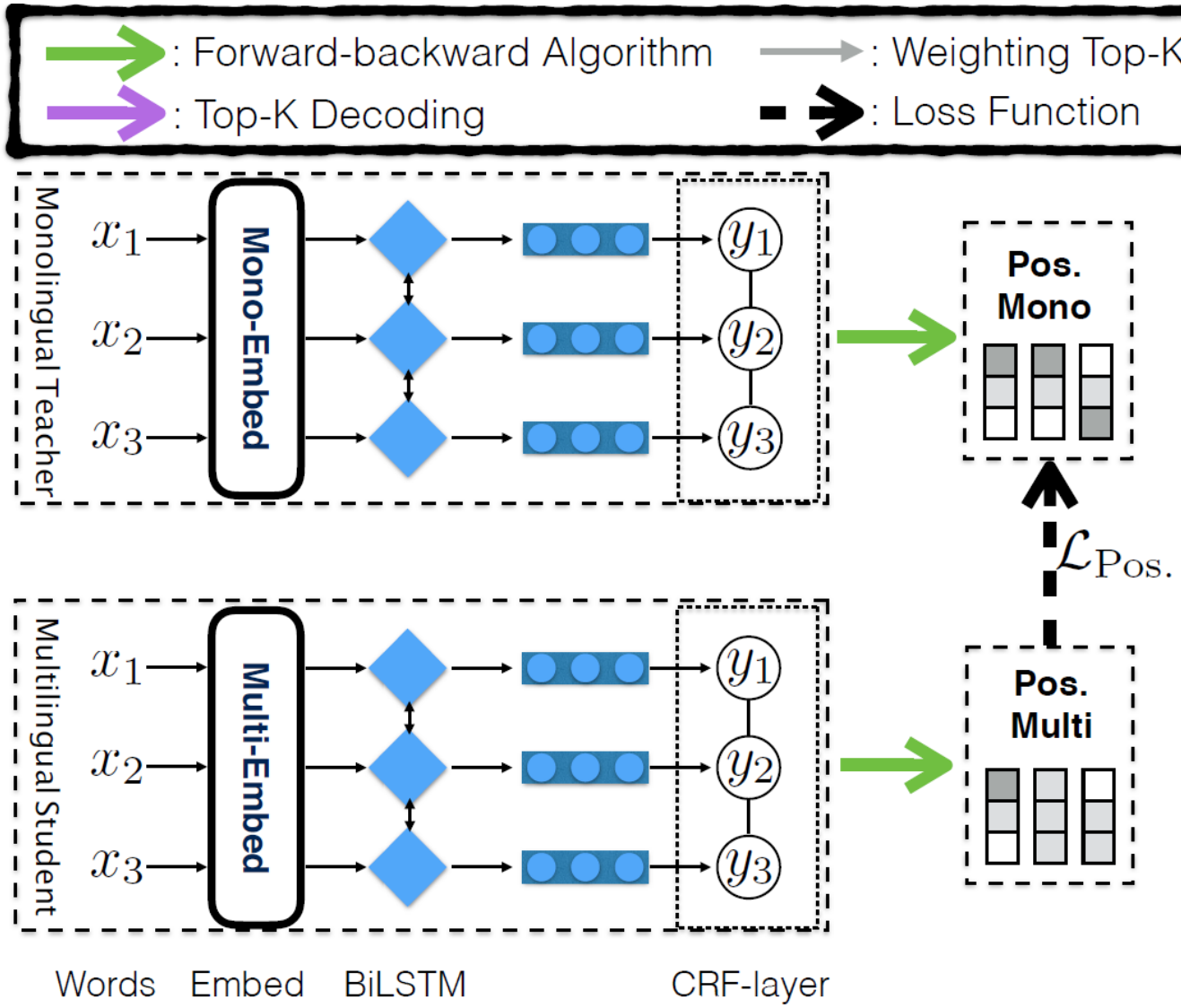
Top-WK Distillation

$$p'_t(\mathbf{y}|\mathbf{x}) = \begin{cases} \frac{p_t(\mathbf{y}|\mathbf{x})}{\sum_{\hat{\mathbf{y}}\in\mathcal{T}} p_t(\hat{\mathbf{y}}|\mathbf{x})} & \mathbf{y} \in \mathcal{T} \\ 0 & \mathbf{y} \notin \mathcal{T} \end{cases}$$

$$\mathcal{L}_{\text{Top-WK}} = -\sum_{\mathbf{y}\in\mathcal{T}} p'_t(\mathbf{y}|\mathbf{x}) \log p_s(\mathbf{y}|\mathbf{x})$$



Posterior Distillation



Posterior Distribution

$$q(y_k|\mathbf{x}) = \sum_{\{y_1,\ldots,y_n\}\setminus y_k} p(y_1,\ldots,y_n|\mathbf{x})$$

$$= \frac{\sum_{\{y_1,\ldots,y_n\}\setminus y_k} \prod_{i=1}^{n} \psi(y_{i-1}, y_i, \mathbf{r}_i)}{\mathcal{Z}}$$

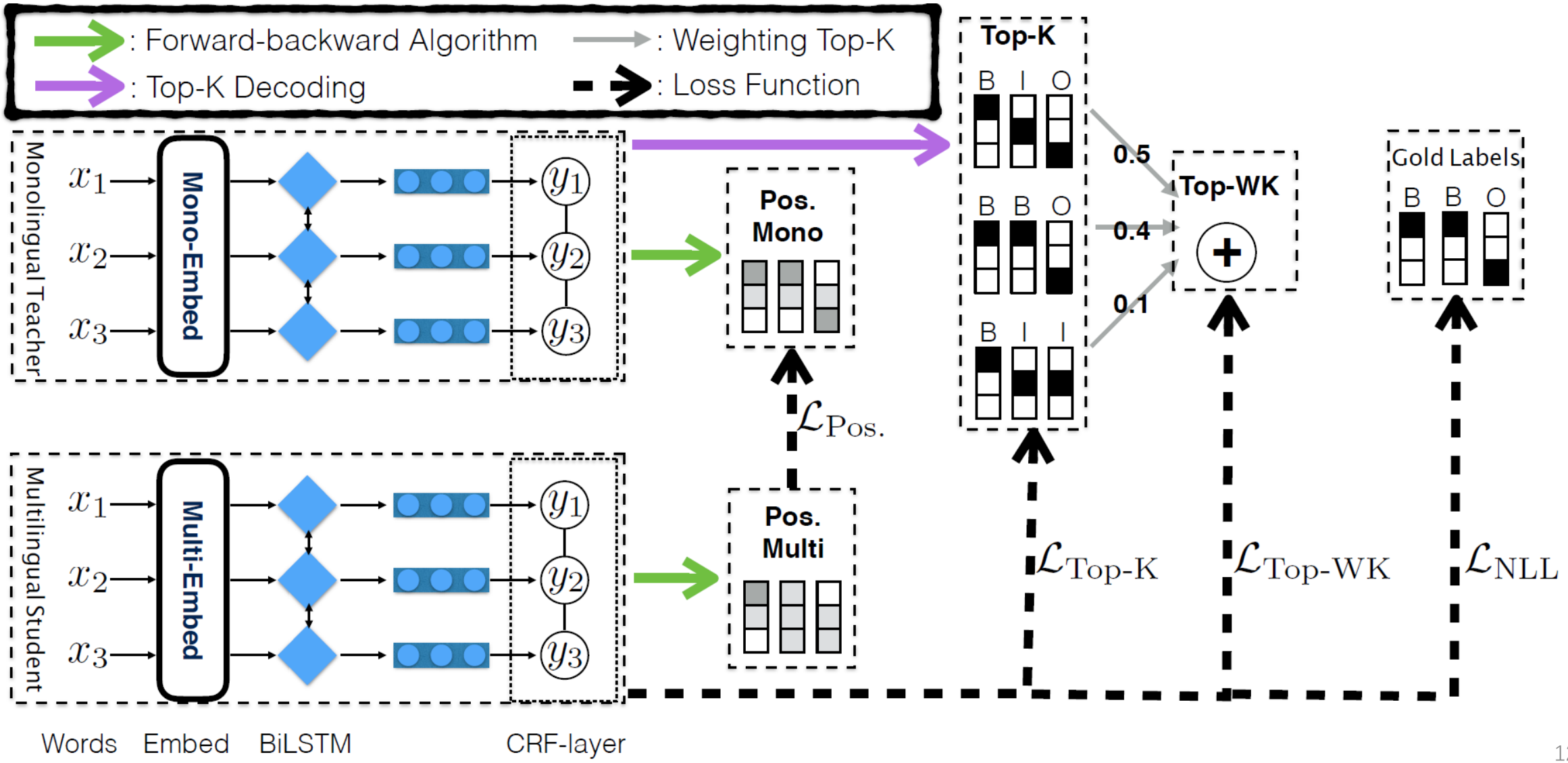
$$\propto \alpha(y_k) \times \beta(y_k)$$

$$\alpha(y_k) = \sum_{\{y_0,\ldots,y_{k-1}\}} \prod_{i=1}^{k} \psi(y_{i-1}, y_i, \mathbf{r}_i)$$

$$\beta(y_k) = \sum_{\{y_{k+1},\ldots,y_n\}} \prod_{i=k+1}^{n} \psi(y_{i-1}, y_i, \mathbf{r}_i)$$

$$\mathcal{L}_{\text{Pos.}} = -\sum_{i=1}^{n}\sum_{j=1}^{|\mathcal{V}|} q_t(y_i = j|\mathbf{x}) \log q_s(y_i = j|\mathbf{x})$$

Structure-Level Knowledge Distillation



Results

Task	CoNLL NER	Aspect Extraction	WikiAnn NER	UD POS
TEACHERS	89.38	70.20	88.97	96.31
BASELINE	87.36	66.54	87.48	94.06
EMISSION	87.55	65.79	87.43	94.13
TOP-K	87.62	67.18	87.53	94.12
TOP-WK	87.64	67.22	87.57	94.14
POSTERIOR	87.72	**67.49**	**87.83**	**94.29**
POS.+TOP-WK	**87.77**	67.34	87.71	94.20

- Monolingual teacher models outperform multilingual student models
- Our approaches outperform the baseline model
- Top-WK+Posterior stays in between Top-WK and Posterior

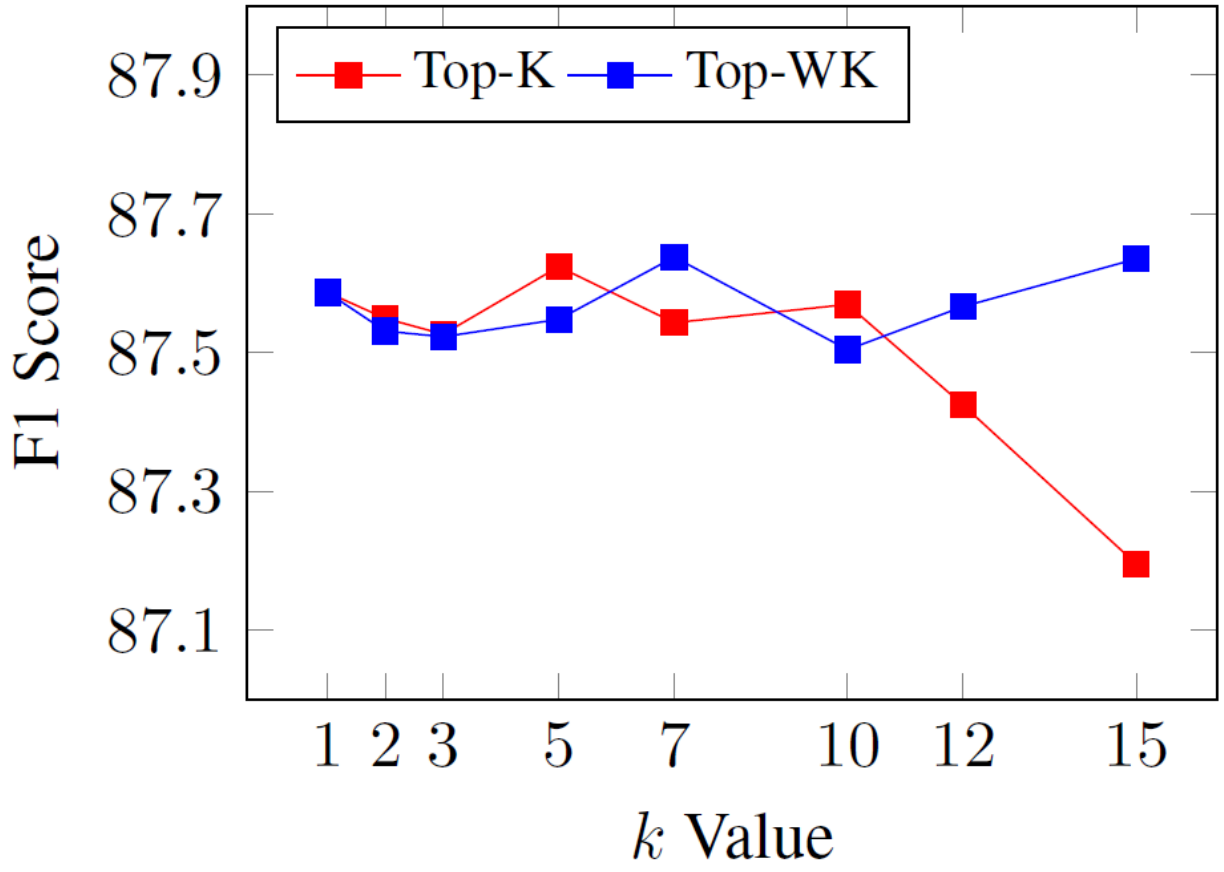
Zero-shot Transfer

	NER	POS
TEACHERS	41.85	56.01
BASELINE	50.86	84.11
EMISSION	50.19	84.17
POSTERIOR	**51.43**	**84.28**
POSTERIOR+TOP-K	51.14	84.24

KD with weaker teachers

	English	Dutch	Spanish	German	Avg.
TEACHERS	90.63	89.65	88.05	81.81	87.54
BASELINE	90.13	89.11	88.06	**82.16**	87.36
POSTERIOR	**90.57**	**89.17**	**88.61**	**82.16**	**87.63**

k Value in Top-K



Conclusion

- Two structure-level KD methods: Top-K and Posterior distillation
- Our approaches improve the performance of multilingual models over 4 tasks on 25 datasets.
- Our distilled model has stronger zero-shot transfer ability on the NER and POS tagging task.