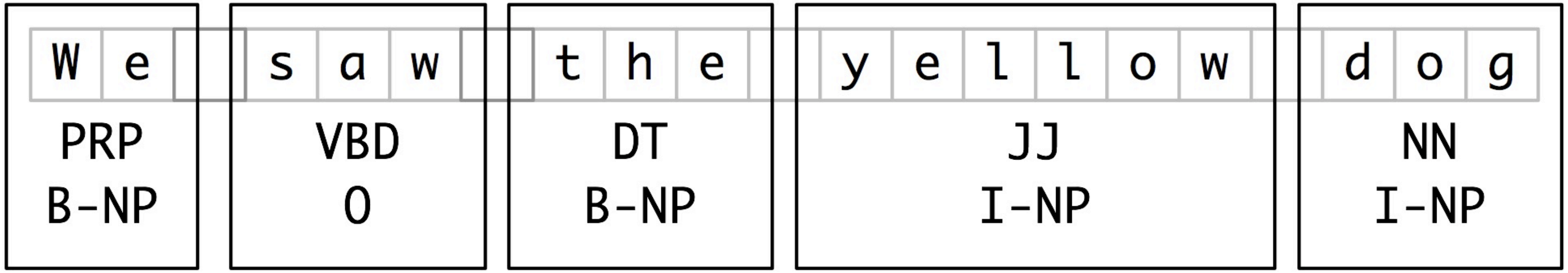
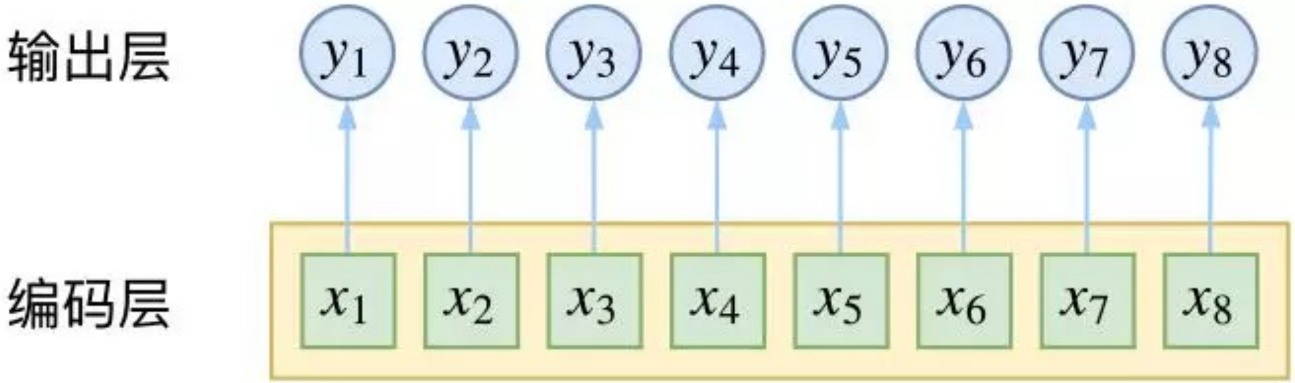
Uncertainty-Aware Sequence Labeling

主讲人：复旦大学 桂韬
导师：　张奇、黄萱菁

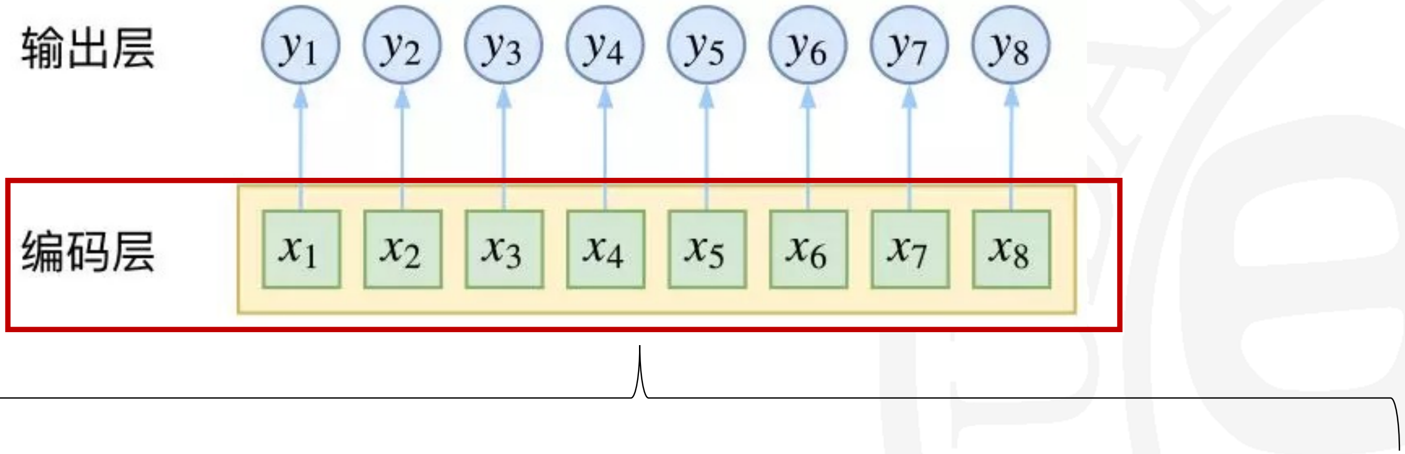
Introduction to Sequence Labeling

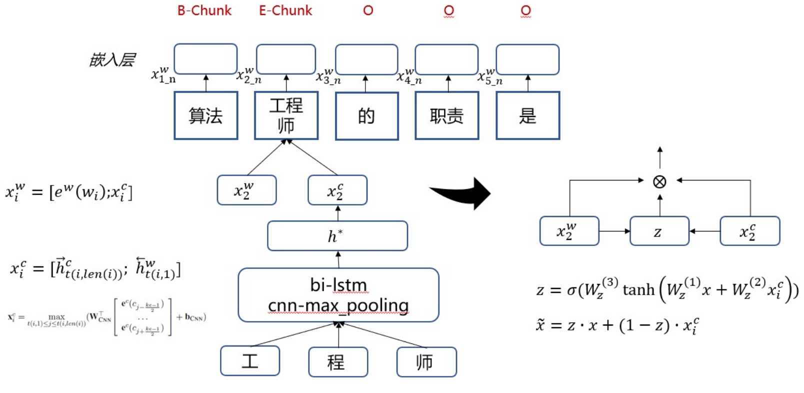




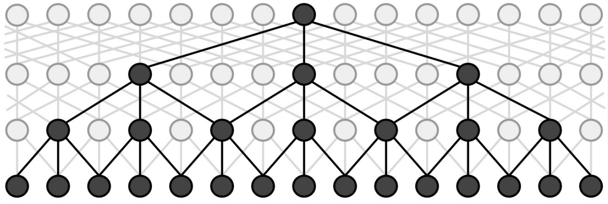
https://yiyibooks.cn/yiyi/nltk_python/ch07.html

Introduction to Sequence Labeling

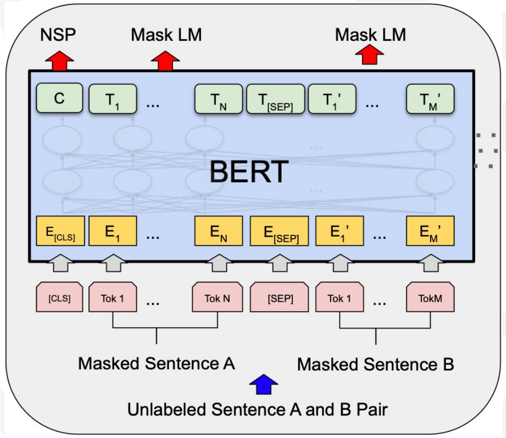




LSTM

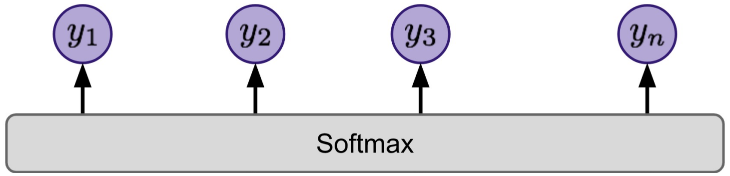


CNN

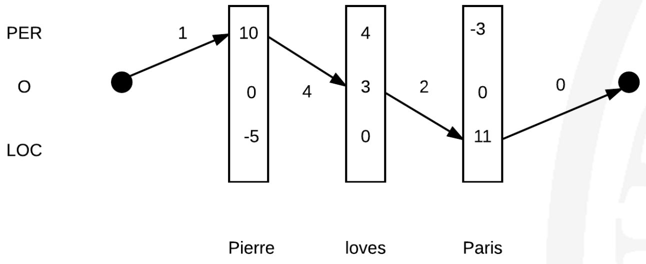


Pre-training

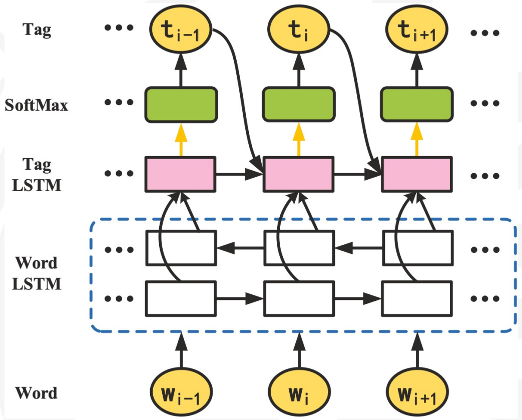
Introduction to Sequence Labeling



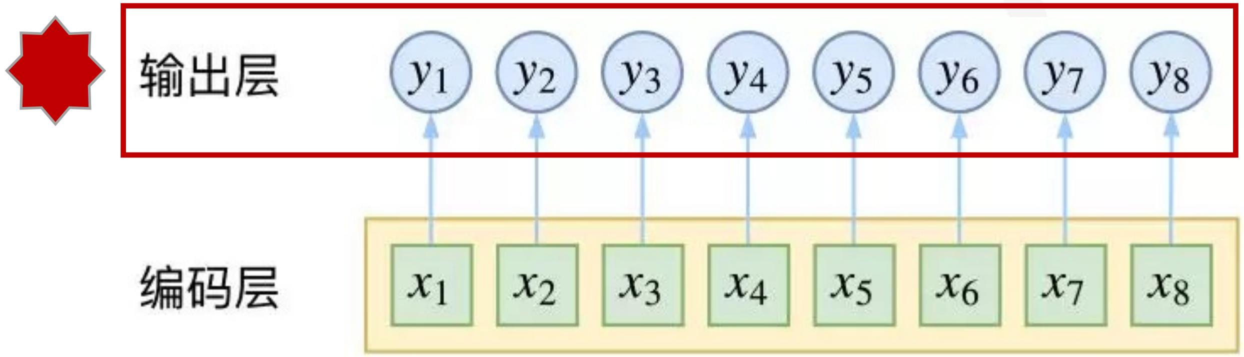
Softmax



CRF



Seq2seq



Motivation

Decoding Methods	Strength	Weakness
Softmax	parallel decoding	No label dependency
CRF	Local label dependency	Viterbi decoding
Seq2seq	Long-term label dependency	Sequence decoding

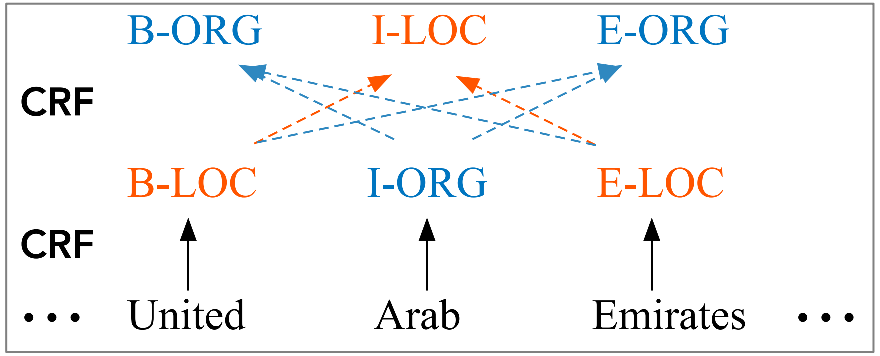
Comparison of different label decoding methods

Motivation

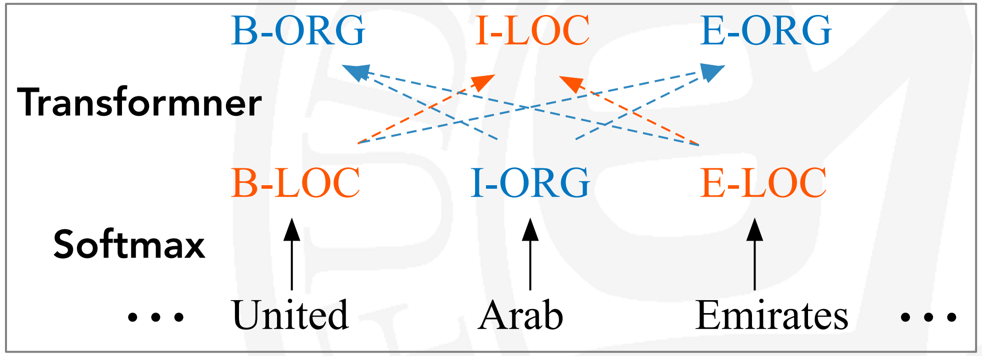
Decoding Methods	Strength	Weakness
Softmax	**parallel decoding**	No label dependency
CRF	Local label dependency	Viterbi decoding
Seq2seq	**Long-term label dependency**	Sequence decoding

What do we want?

Model Design



Manning’ s

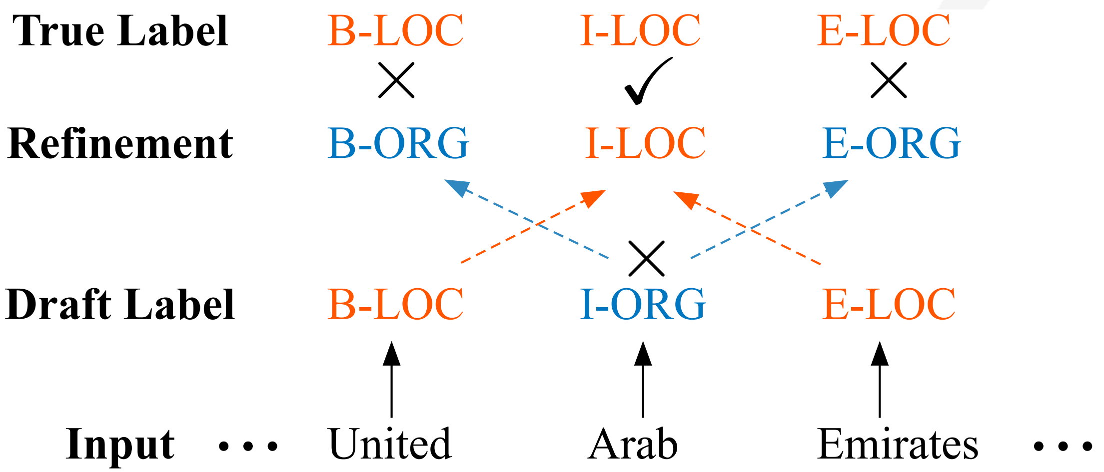


Ours

An intuitive way: Two Stage

Krishnan, Vijay, and Christopher D. Manning. "**An effective two-stage model for exploiting non-local dependencies in named entity recognition**." *ACL*, 2006.

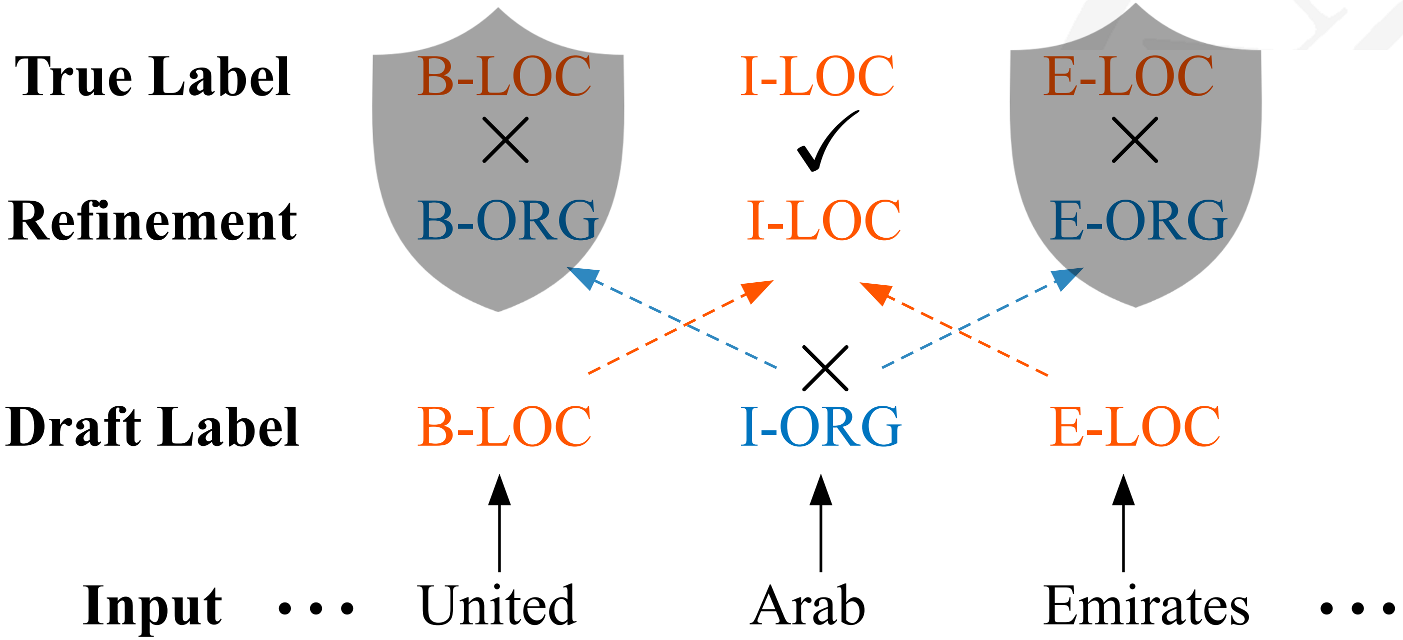
Model Design



Refinement	#Tokens
✔ → ✘	**39**
✘ → ✔	54

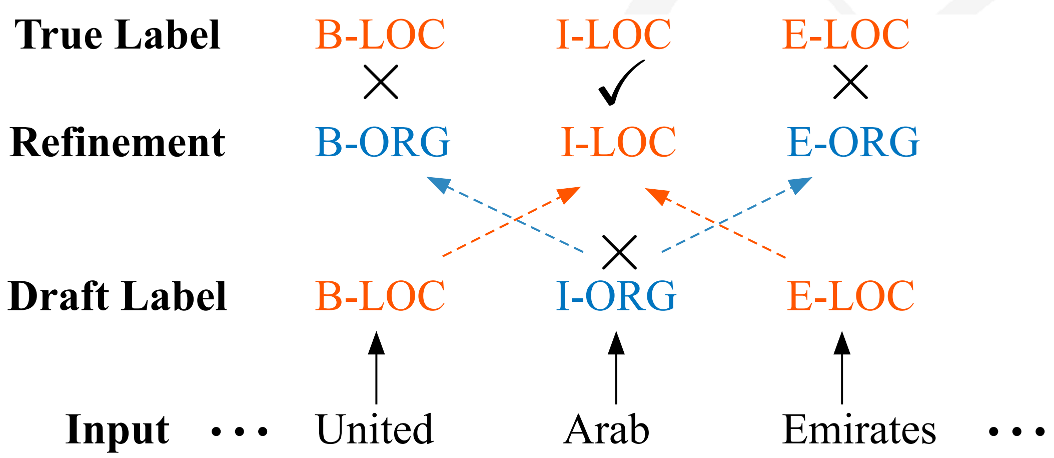
Table 1: Results of LAN with uncertainty estimation evaluated on CoNLL2003 test dataset. ✔ refers to the correct prediction, and ✘ refers to the wrong prediction.

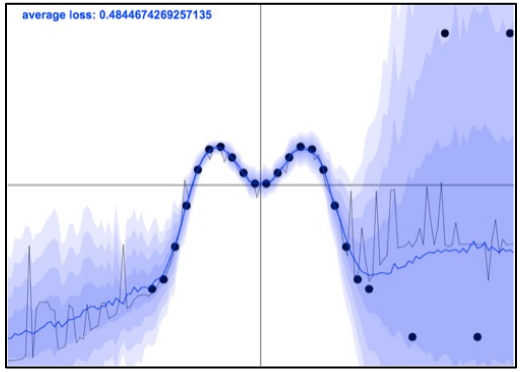
Model Design



Can we fine an indicator?

Model Design





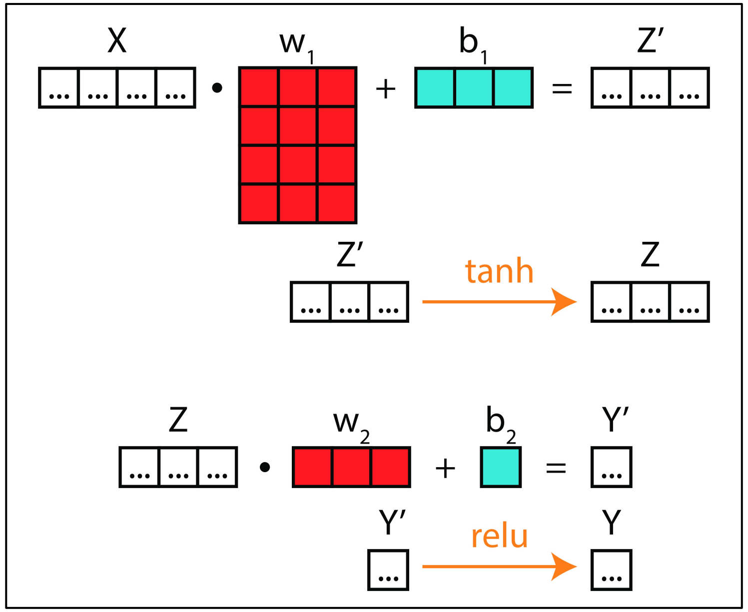
Bayesian NNs for
Uncertainty Estimation

Draft	Uncertainty	Refinement	#Tokens
✔	0.018	✔ → ✘	**39**
✘	**0.524**	✘ → ✔	54

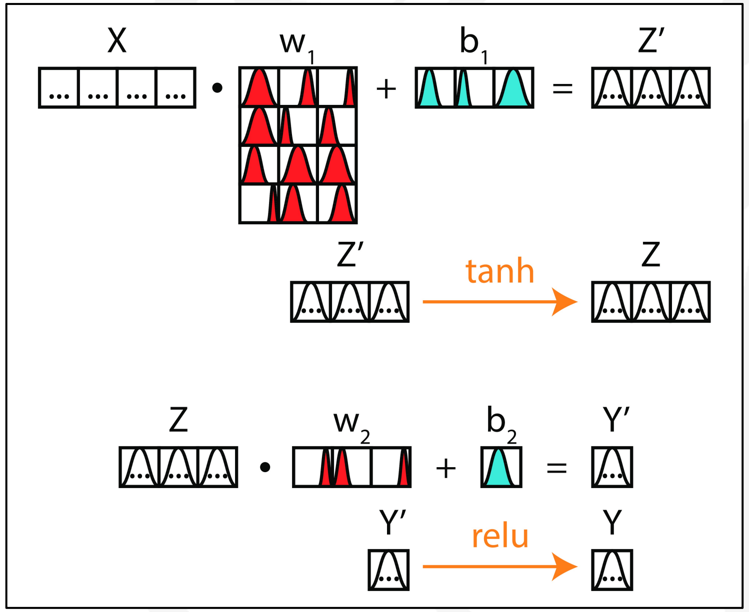
Table 1: Results of LAN with uncertainty estimation evaluated on CoNLL2003 test dataset. ✔ refers to the correct prediction, and ✘ refers to the wrong prediction.

Model Design

Neural Network



Bayesian Neural Network



$$\hat{y} = f_w(x)$$

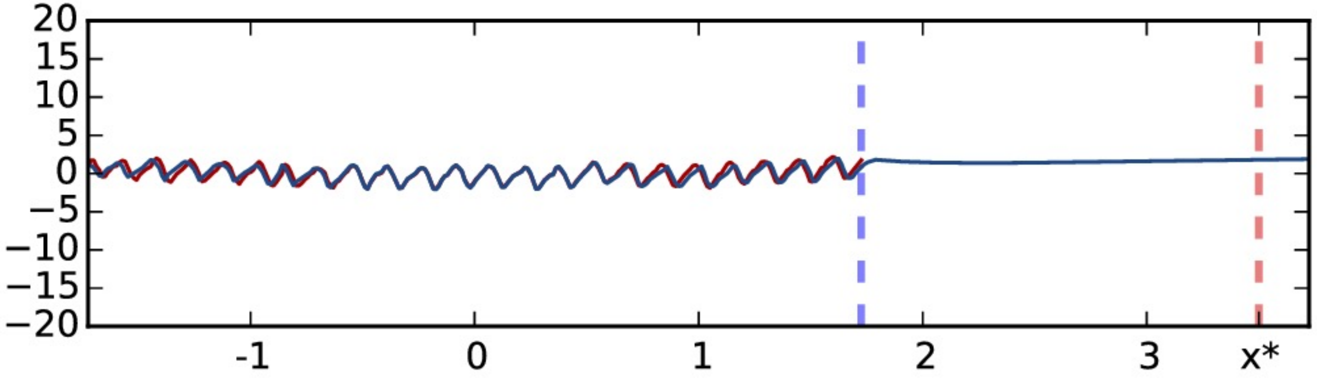
$$p(y^*|x^*, D) = \int p(y^*|W, x^*)p(W|D)dW$$

Source: ericmjl/bayesian-deep-learning-demystified

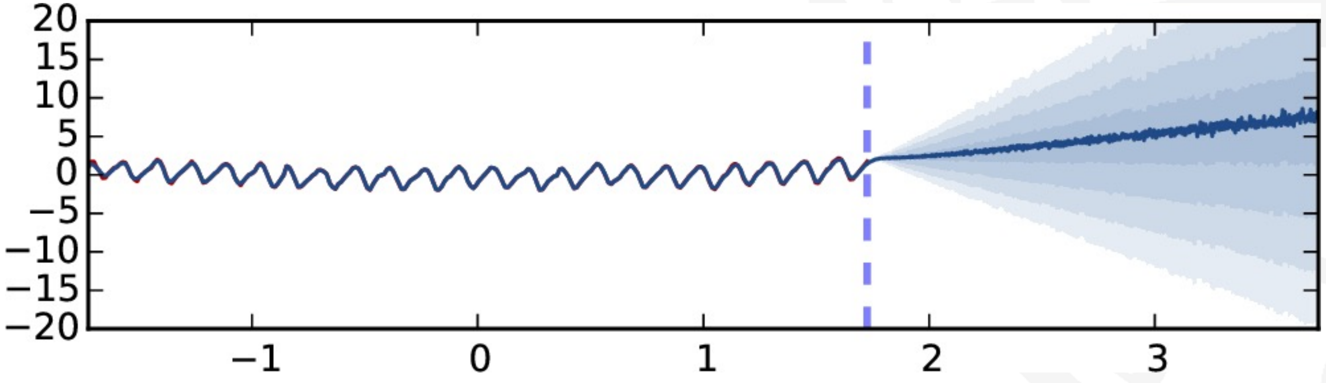
Model Design

Regression

Deep learning



Bayesian deep learning



Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. 2016.

Model Design

$$p(y^*|x^*, D) = \int p(y^*|W, x^*)\boxed{p(W|D)}dW$$

Learning

$$KL[q(W)||p(W|X,Y)]$$

$$-\int q(W)\log p(Y|X,W)\,dW + KL[q(W)||p(W)]$$

Bernoulli **Dropout**

$$\mathcal{L}(\theta, p) = -\frac{1}{N}\sum_{i=1}^{N}\log p(\mathbf{y}_i|\mathbf{f}^{\widehat{\mathbf{W}}_i}(\mathbf{x}_i)) + \frac{1-p}{2N}||\theta||^2$$

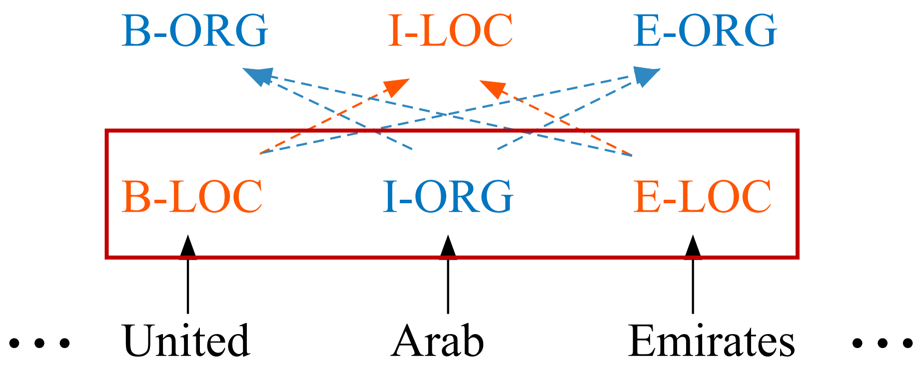
Inference

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathcal{D}) \approx \sum_{j=1}^{T} p(\mathbf{y}^*|\mathbf{W}_j, \mathbf{x}^*)q_\theta^*(\mathbf{W}_j)$$

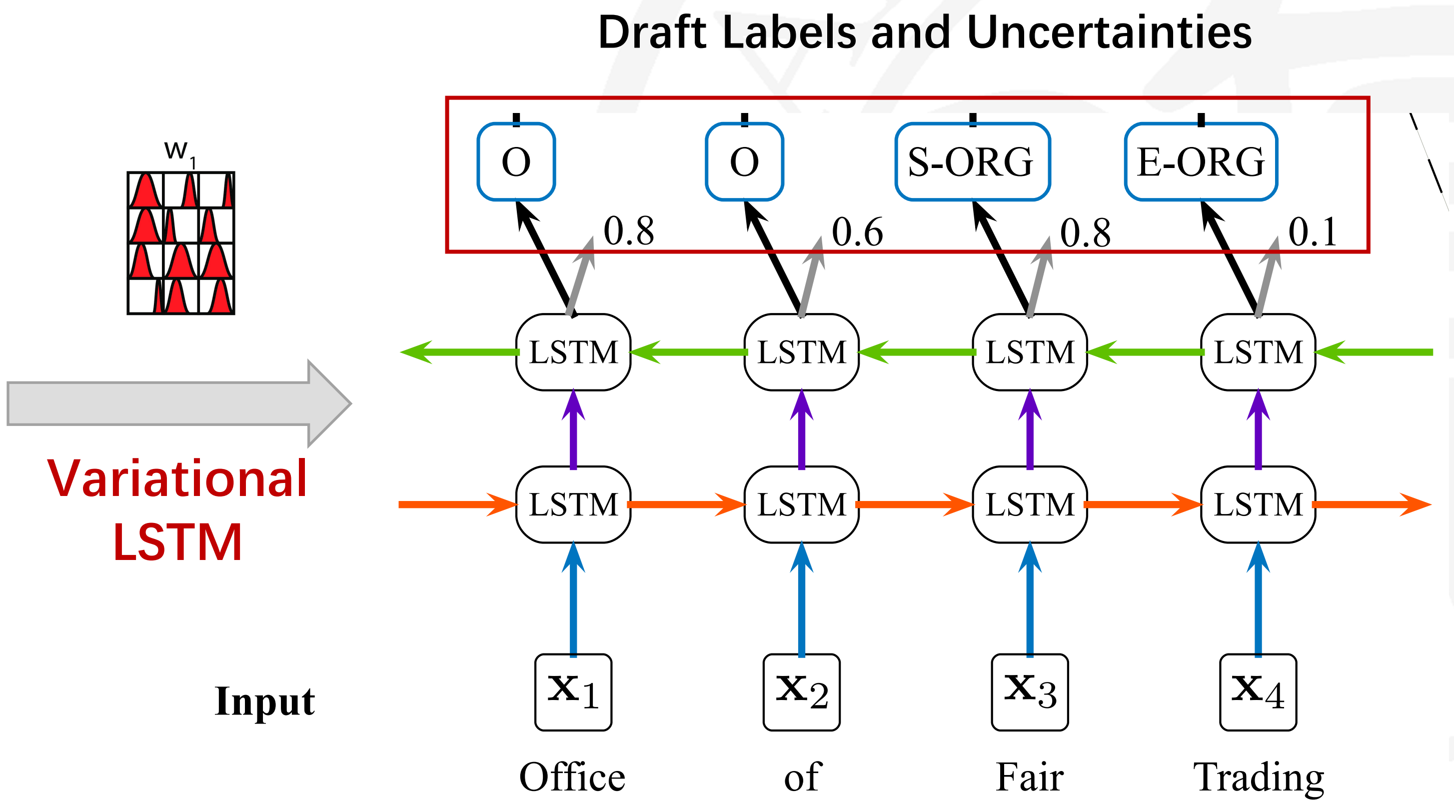
$$u_i = H(\mathbf{p}_i) = -\sum_{c=1}^{C} p_c \log p_c$$

Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. 2016.

Model Design

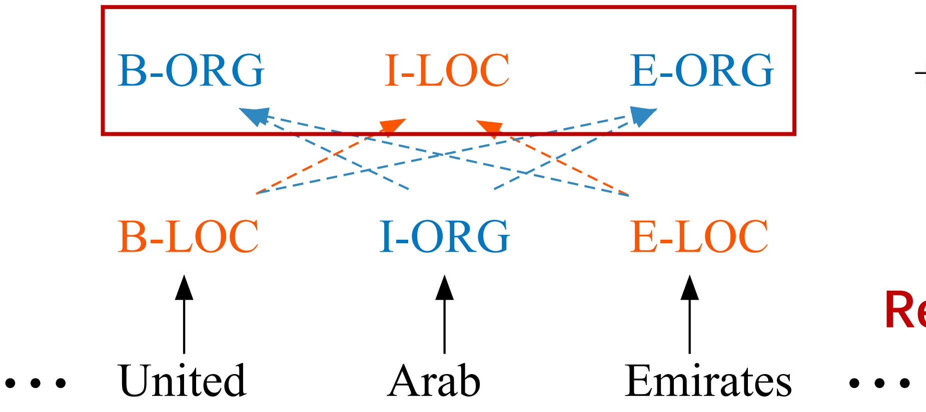


Variational LSTM



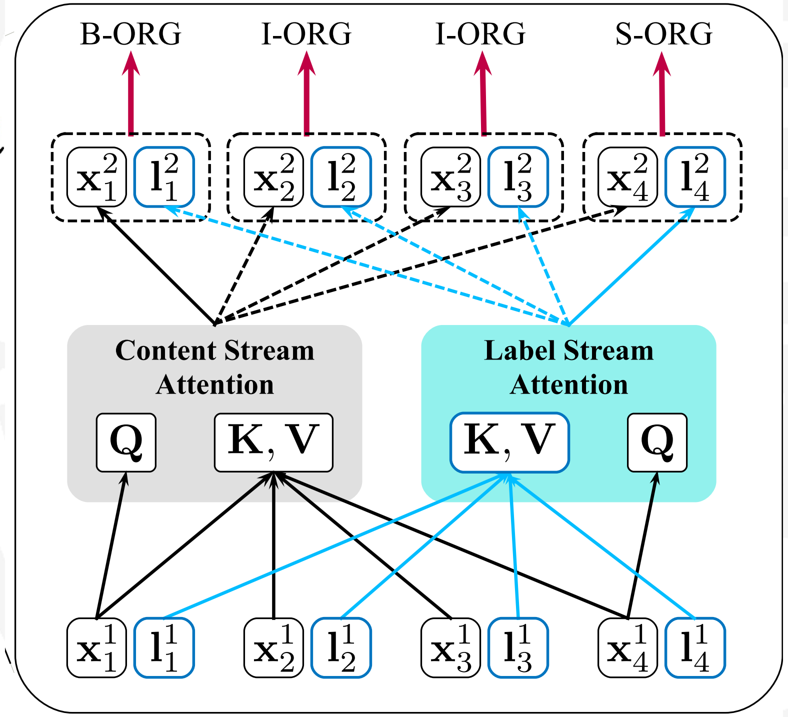
Variational LSTM for encoder

Model Design



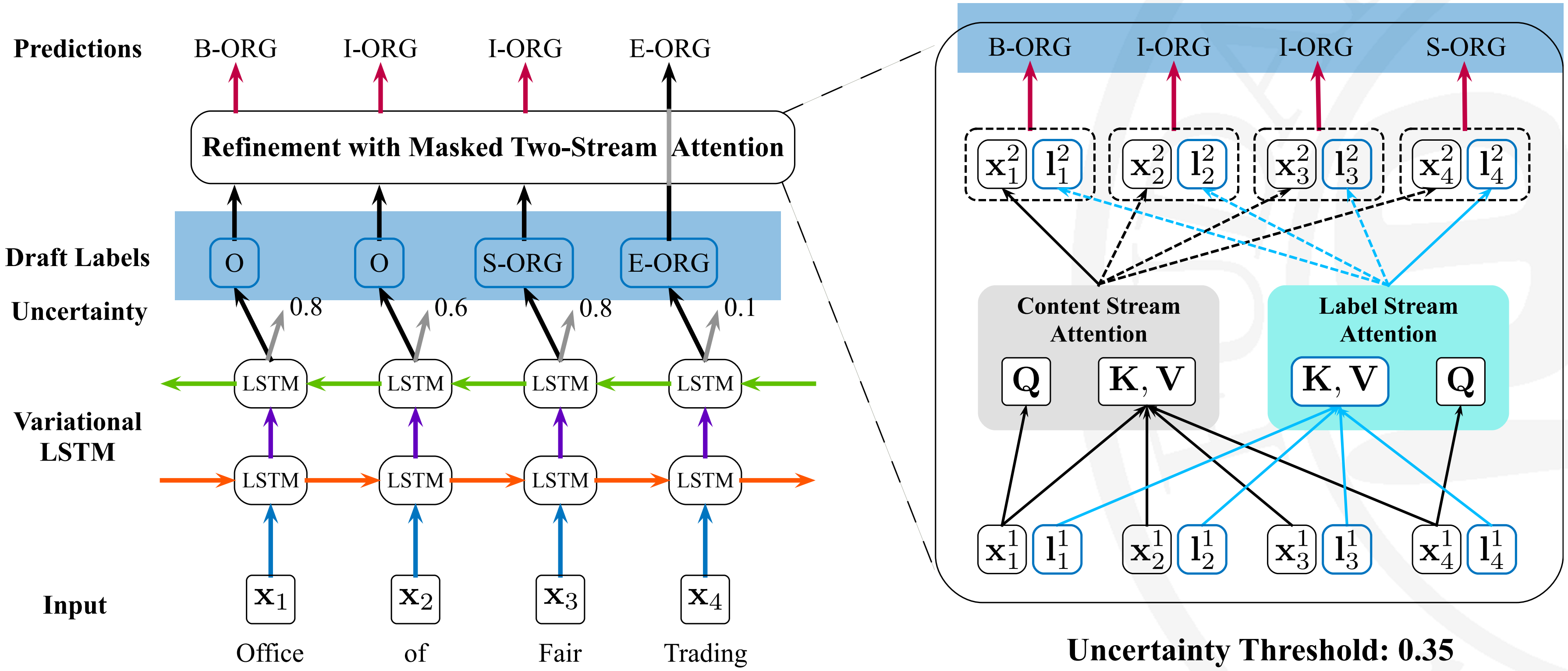
$$\mathbf{A}^{\mathrm{rel}}_{i,j} = \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_{k,E}\mathbf{E}_{x_j}}_{(a)} + \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_{k,R}\mathbf{R}_{i-j}}_{(b)} + \underbrace{u^\top \mathbf{W}_{k,E}\mathbf{E}_{x_j}}_{(c)} + \underbrace{v^\top \mathbf{W}_{k,R}\mathbf{R}_{i-j}}_{(d)}.$$

Relative Position Encoding



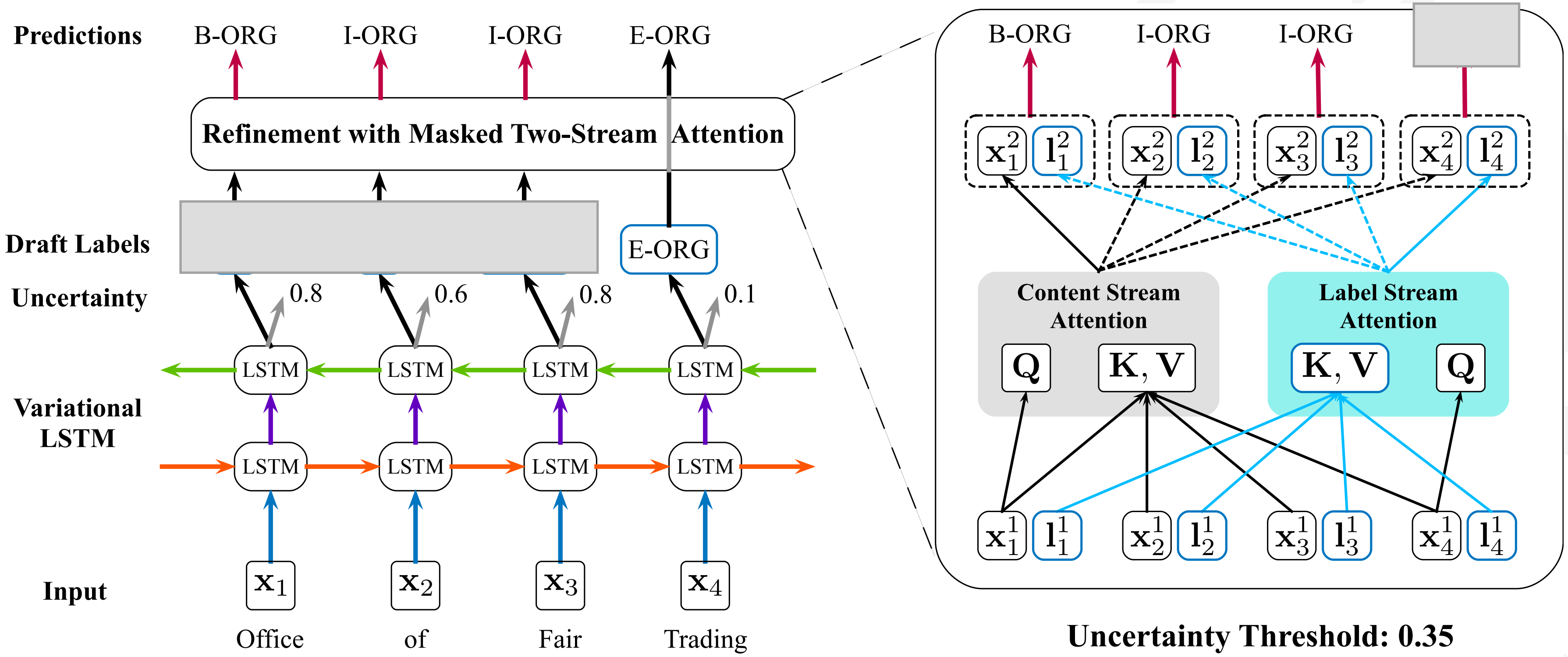
Two-stream attention for label refinement

Model Design



Draft labels and **refined labels**

Model Design



Uncertainty Threshold: 0.35

Setting a threshold

Experiments

Models	CoNLL2003	OntoNotes	WSJ
Chiu and Nichols (2016)	90.91	86.28	-
Strubell et al. (2017)	90.54	86.84	-
Liu et al. (2018)	91.24	-	97.53
Chen et al. (2019)	91.44	87.67	-
BiLSTM-CRF (Ma and Hovy, 2016)	91.21	86.99	97.51
BiLSTM-Softmax (Yang et al., 2018)	90.77	83.76	97.51
BiLSTM-Seq2seq (Zhang et al., 2018)	91.22	-	97.59
Rel-Transformer (Dai et al., 2019)	90.70	87.45	97.49
BiLSTM-LAN (Cui and Zhang, 2019)	90.77*	88.16	97.58
BiLSTM-UANet ($M = 8$)	**91.60**	**88.39**	**97.62**

Main results

Models	F_1
IntNet + BiLSTM-Softmax (Xin et al., 2018)	91.43
IntNet + BiLSTM-CRF	91.64
IntNet + UANet	**91.80**
BERT-Softmax (Devlin et al., 2019)	91.62
BERT-CRF	91.71
BERT + UANet	**92.02**

Results with complex representations

Experiments

	CoNLL2003	OntoNotes	WSJ
Average Sentence Length	13	18	24
BiLSTM-CRF	1,433	950	801
BiLSTM-LAN	949	773	943
BiLSTM-Seq2seq	1,084	842	751
BiLSTM-UANet ($M = 1$)	1,630	1,262	1,192
BiLSTM-UANet ($M = 8$)	1,474	1,129	1,044

Table 6: Comparison of inference speed. We show how many sentences the model can process per second.

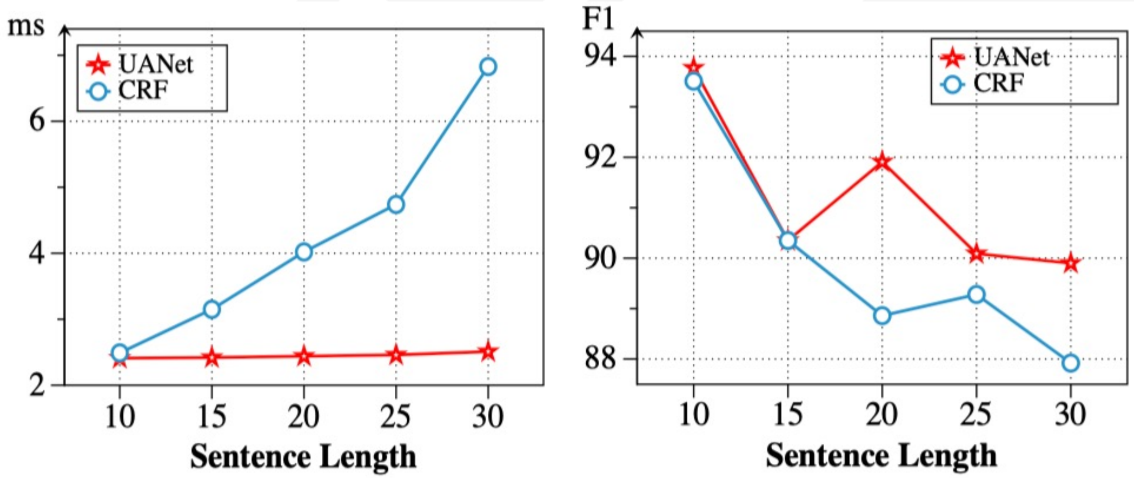
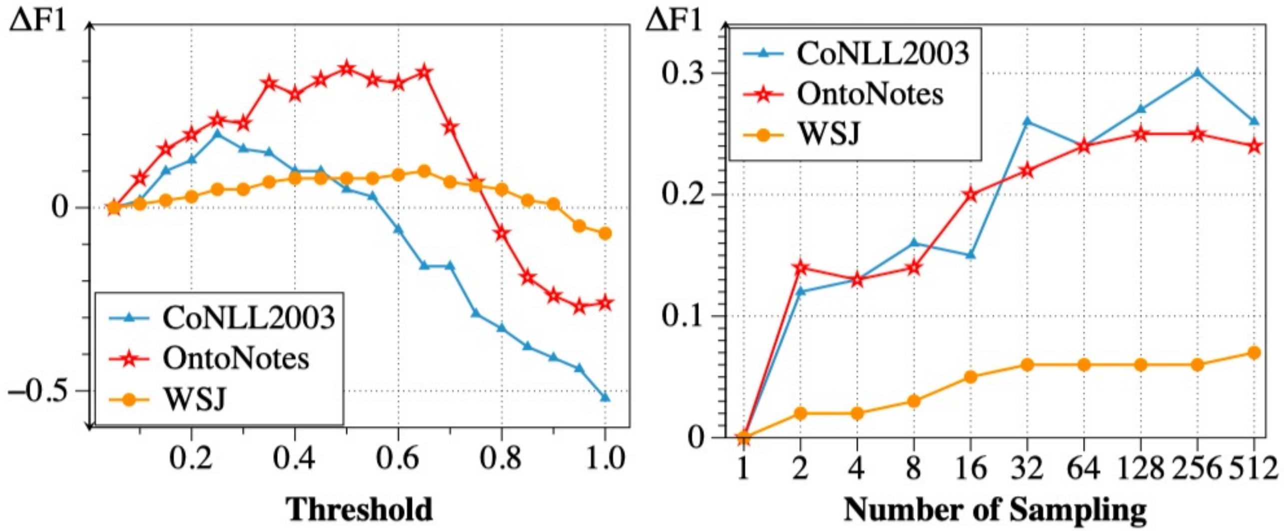


Figure 3: Speed and F1 against sentence length.

Experiments

Models	CoNLL2003	OntoNotes	WSJ
BiLSTM-UANet	**91.60**	**88.39**	**97.62**
- Label information	91.23	87.84	97.57
- Variational LSTM			
Rel-Transformer-Softmax	90.70	87.45	97.49
Rel-Transformer-CRF	91.22	87.77	97.56
- Two-stream self-attention			
Variational LSTM-Softmax	90.83	87.11	97.46
Variational LSTM-CRF	91.20	87.63	97.55

Table 4: Ablation study of UANet.



Influence of threshold and sampling

Experiments

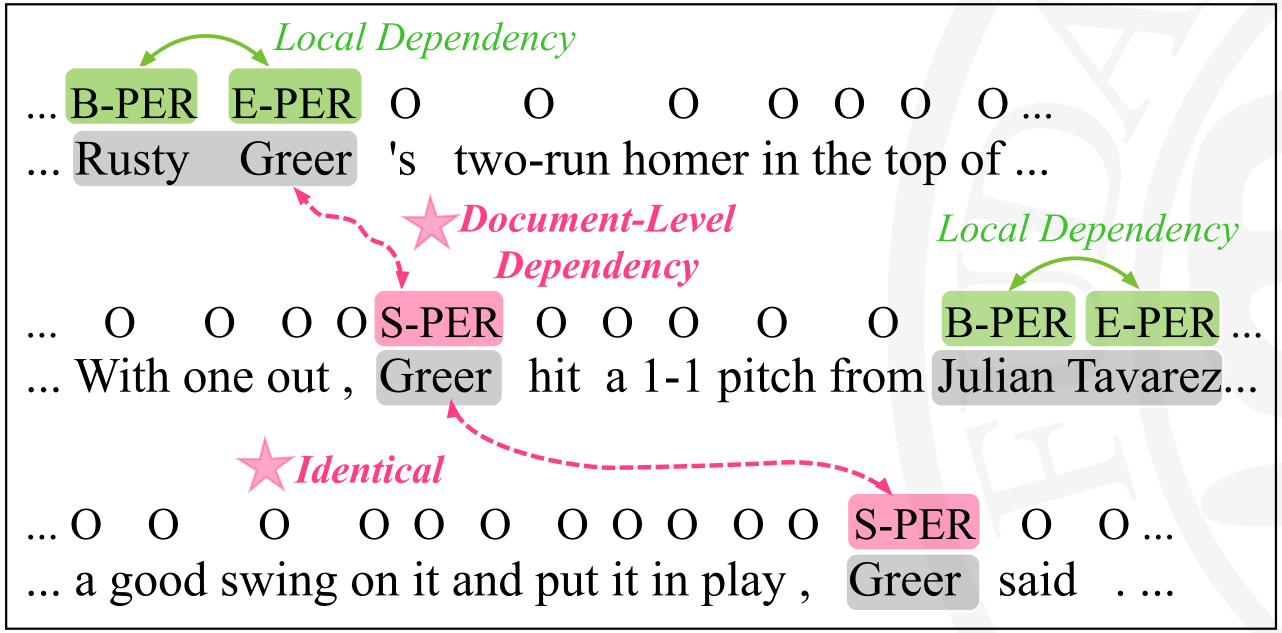
Text	... striker	Viorel	Ion	of	Otelul	Galati	and	defender	Liviu	Ciobotariu	of	National	Bucharest	...
BiLSTM-CRF	... O	B-PER	E-PER	O	B-PER	E-PER	O	O	B-PER	E-PER	O	B-LOC	E-LOC	...
Draft Label	... O	B-PER	E-PER	O	B-PER	E-PER	O	O	B-PER	E-PER	O	B-ORG	E-ORG	...
Refinement	... O	B-PER	E-PER	O	B-ORG	E-ORG	O	O	B-PER	E-PER	O	B-ORG	E-ORG	...
Uncertainty	... 0.001	0.005	0.047	0.004	0.532	0.605	0.000	0.000	0.001	0.014	0.001	0.818	0.927	...
Final Prediction	... O	B-PER	E-PER	O	B-ORG	E-ORG	O	O	B-PER	E-PER	O	B-ORG	E-ORG	...

Case study 1

Text	... University	of	Yangon ...
BiLSTM-CRF	... O	O	S-LOC ...
Draft Label	... B-ORG	I-ORG	E-LOC ...
Refinement	... B-LOC	I-ORG	E-ORG ...
Uncertainty	... 0.302	0.816	0.800 ...
Final Prediction	... B-ORG	I-ORG	E-ORG ...

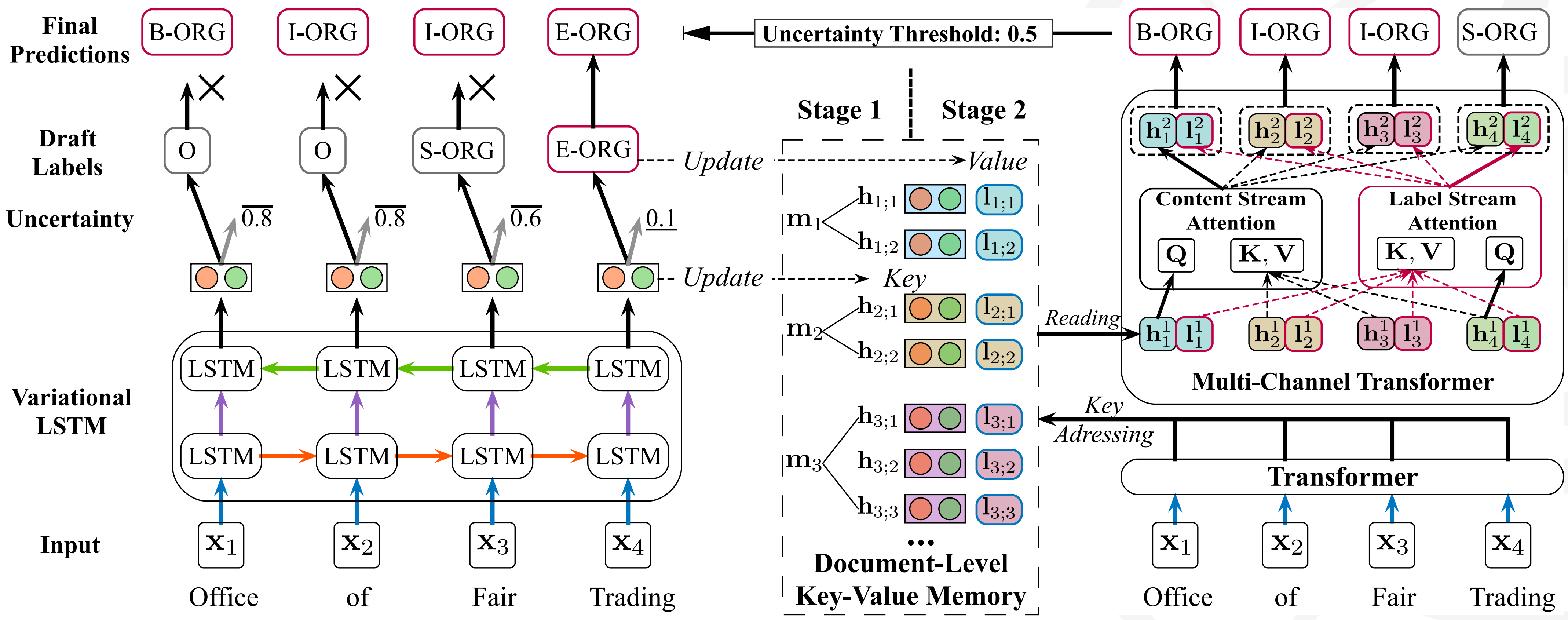
Case study 2

Extension to Document NER

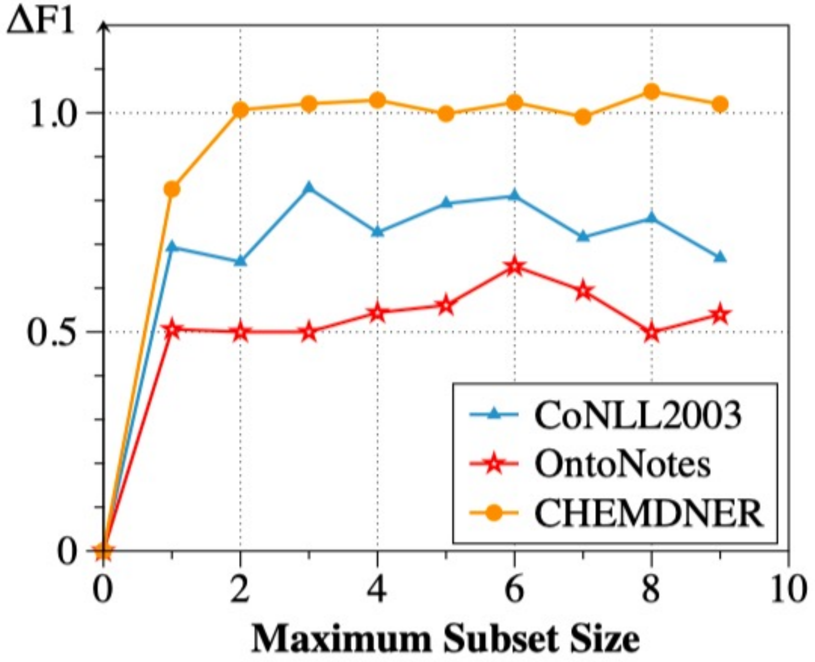


Document-level label consistency

Extension to Document NER



Extension to Document NER



Positive effects of label co-occurrence

Extension to Document NER

Models	F_1
IntNet + BiLSTM-Softmax (Xin et al., 2018)	91.43
IntNet + BiLSTM-CRF	91.64
IntNet + UANet	**91.80**
BERT-Softmax (Devlin et al., 2019)	91.62
BERT-CRF	91.71
BERT + UANet	**92.02**

UANet

Models	F_1
BERT-base [Devlin *et al.*, 2019]	91.82*
BERT-base + DocL-NER	**92.92**
ELMo [Peters *et al.*, 2018]	92.64*
ELMo + DocL-NER	**93.05**

DocL-NER

Conclusions

1. A novel two-stage label refinement framework
2. Bayesian neural networks to indicate the label with a high probability of being wrong
3. Two-stream self-attention networks for modeling long-term labal dependency and word-label interaction

Q & A