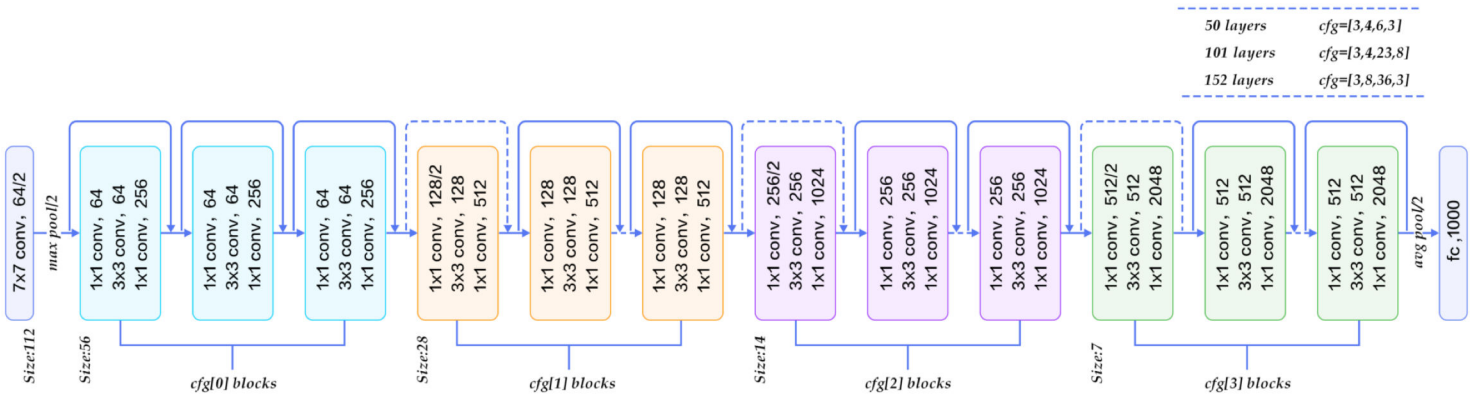
Distilling Knowledge for Search-based Structured Prediction

Yijia Liu*, Wanxiang Che, Huaipeng Zhao, Bing Qin, Ting Liu

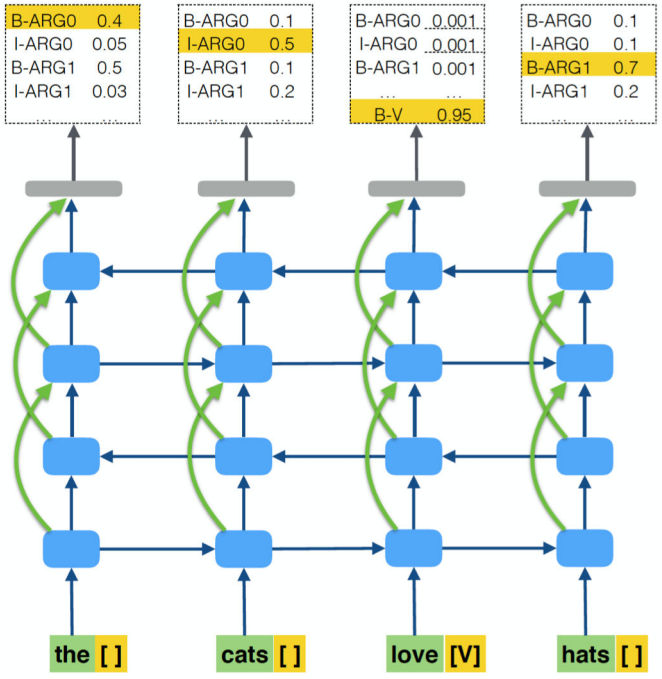
Research Center for Social Computing and Information Retrieval

Harbin Institute of Technology

Complex Model Wins



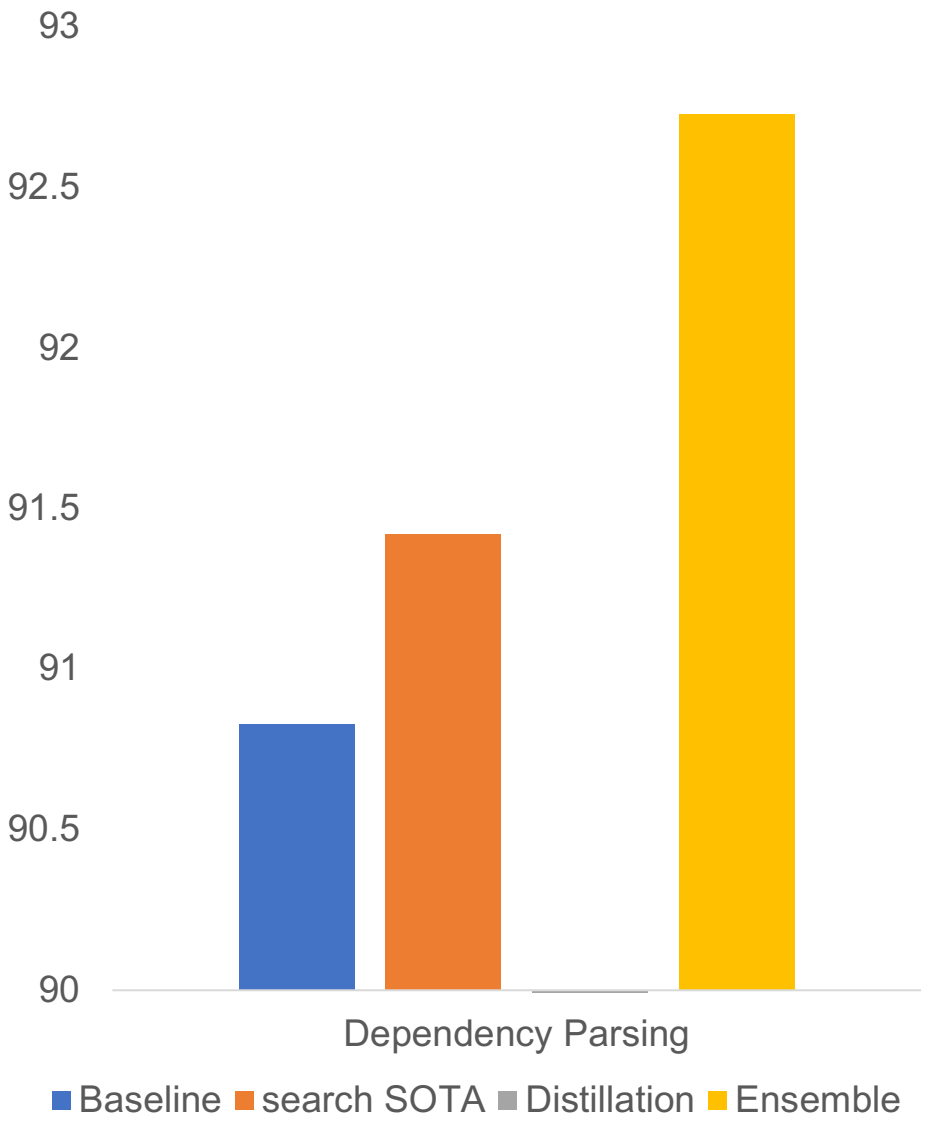
[ResNet, 2015]

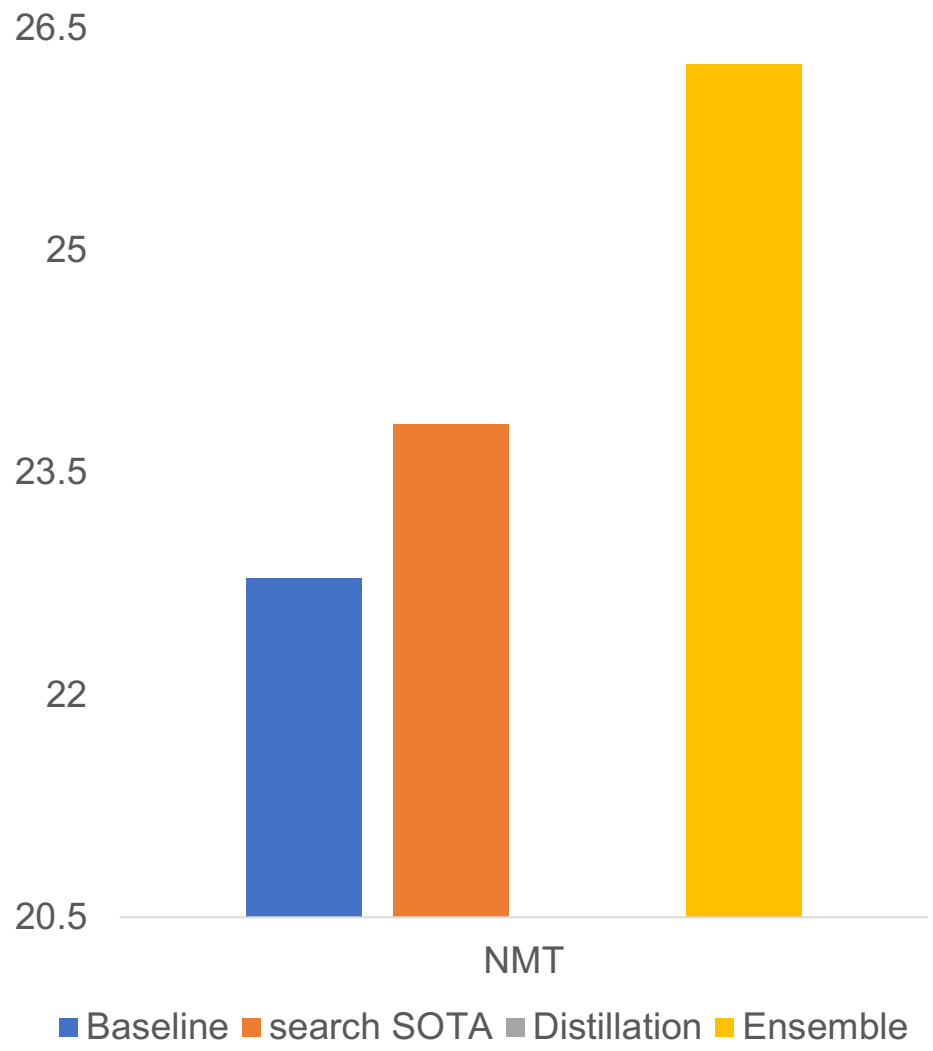


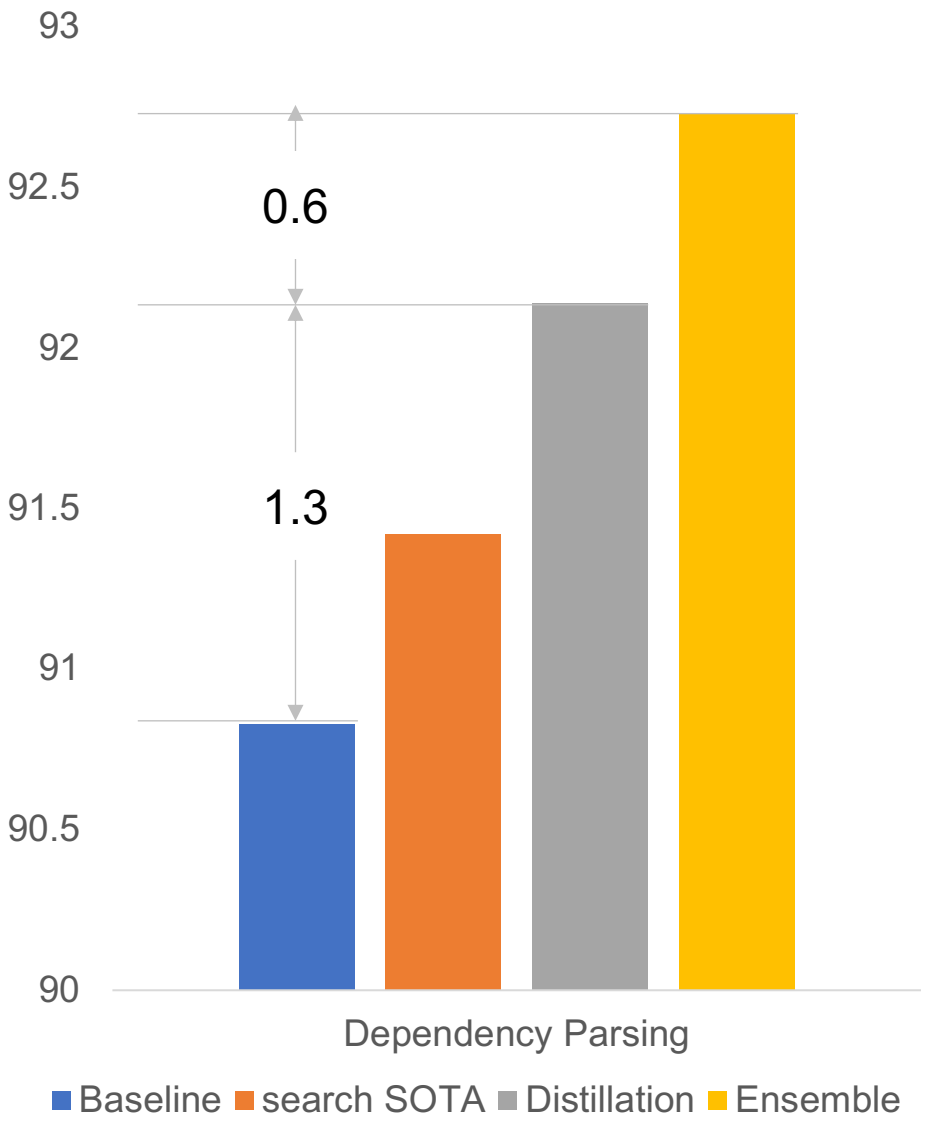
[He+, 2017]

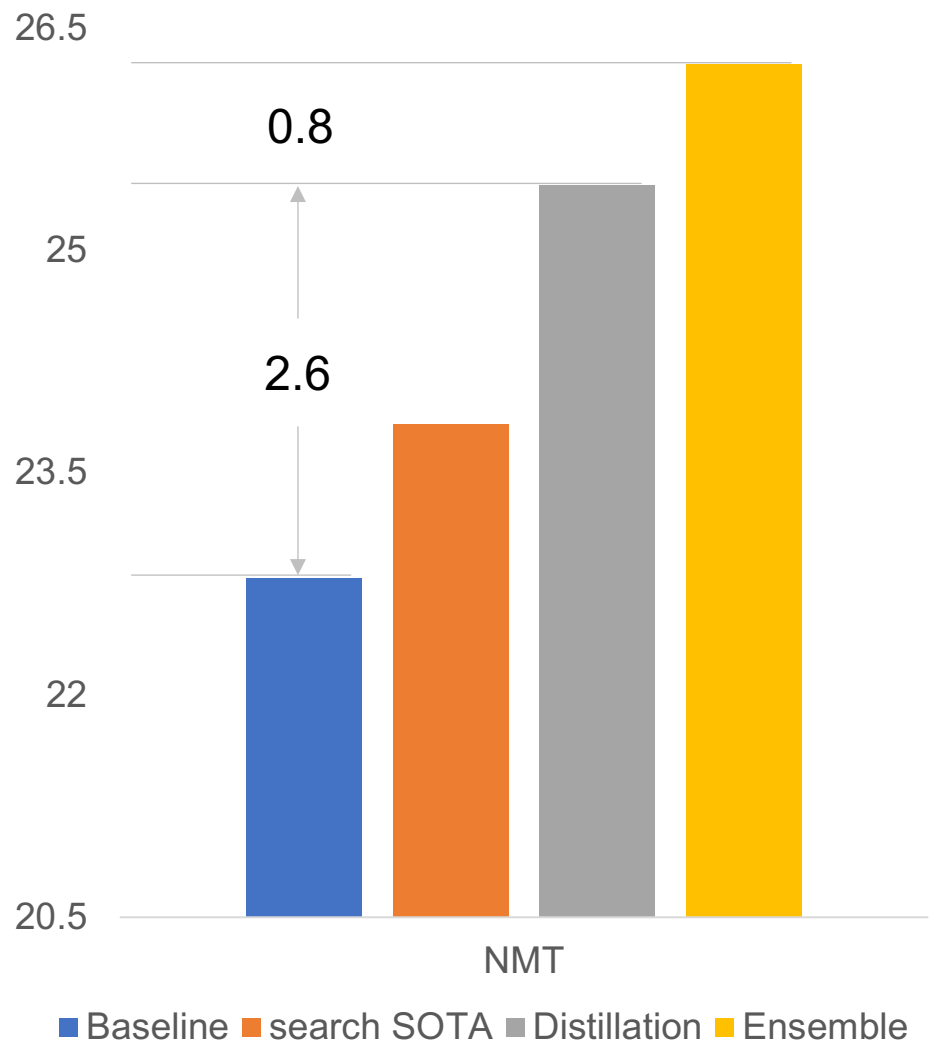
Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, Ting Liu


93
92.5
92
91.5
91
90.5
90
Dependency Parsing
Baseline
search SOTA
Distillation
Ensemble

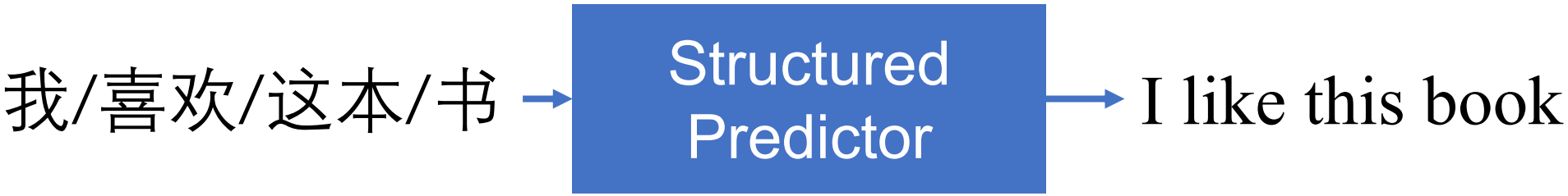

26.5
25
23.5
22
20.5
NMT
Baseline
search SOTA
Distillation
Ensemble


93
92.5
92
91.5
91
90.5
90
0.6
1.3
Dependency Parsing
Baseline
search SOTA
Distillation
Ensemble

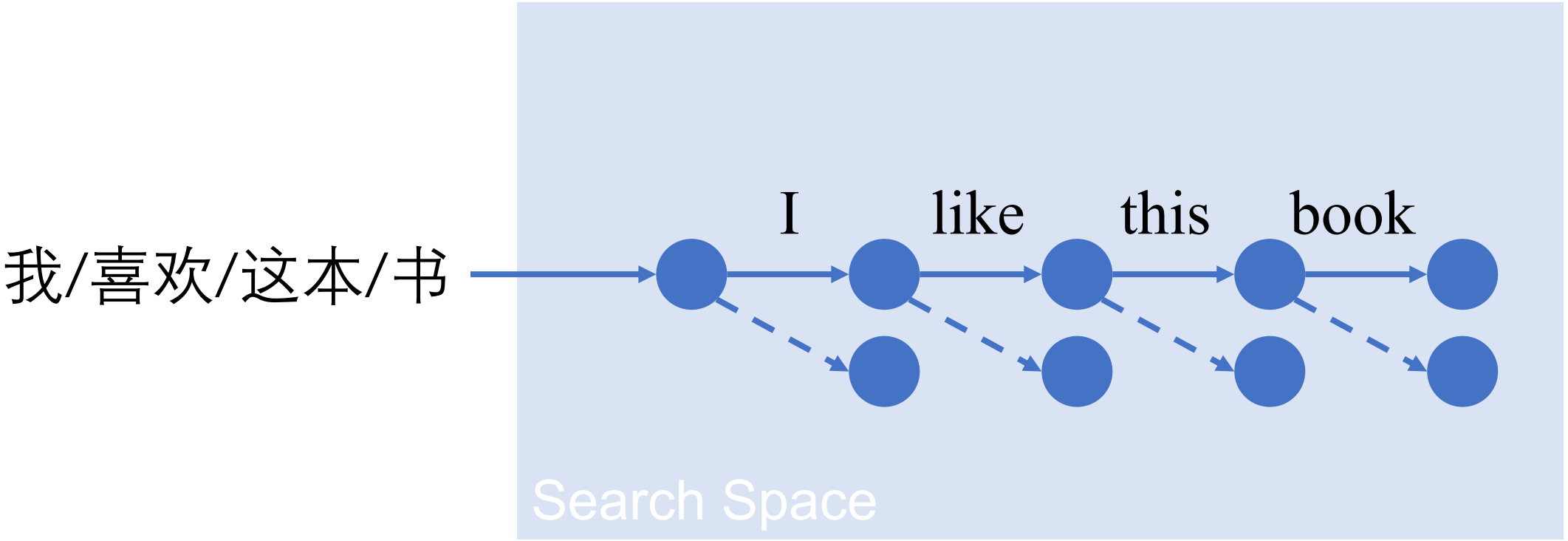

26.5
25
23.5
22
20.5
0.8
2.6
NMT
Baseline
search SOTA
Distillation
Ensemble

Classification vs. Structured Prediction

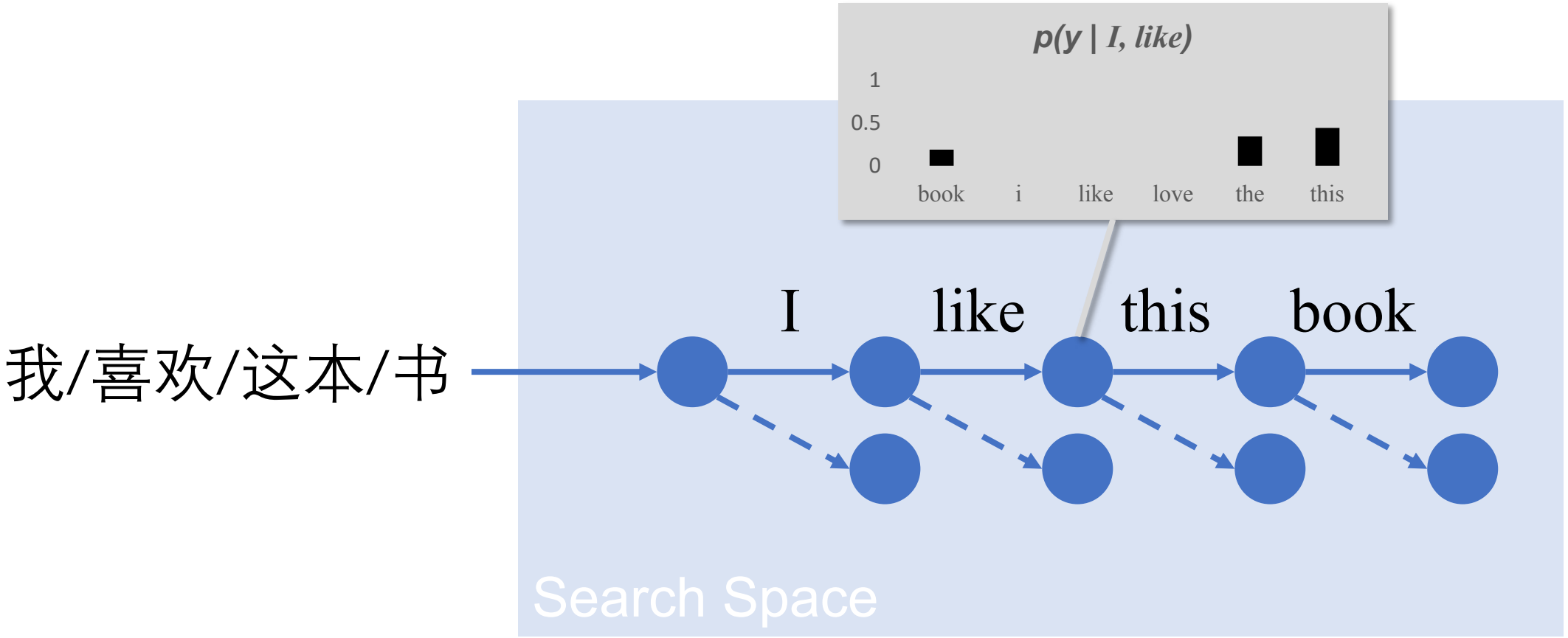
Classification vs. Structured Prediction



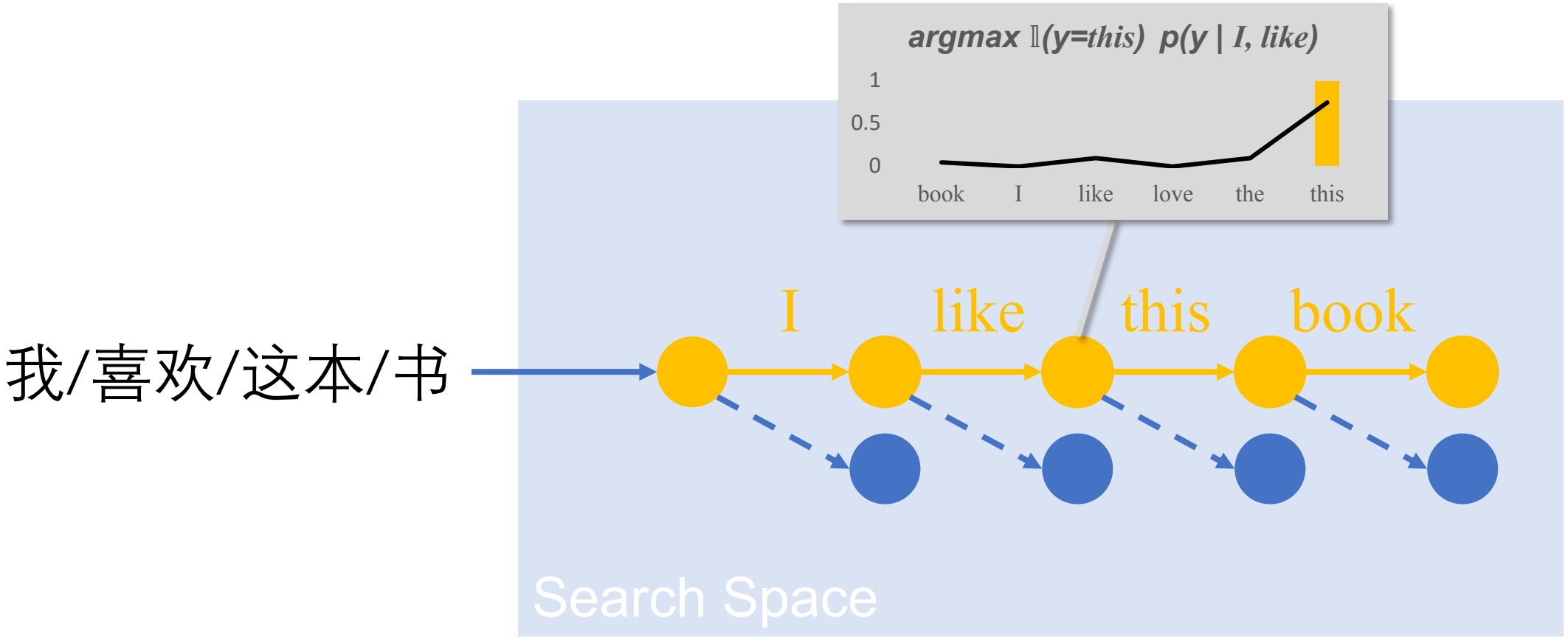
Search-based Structured Prediction



$p(a \mid s)$ that Controls Search Process

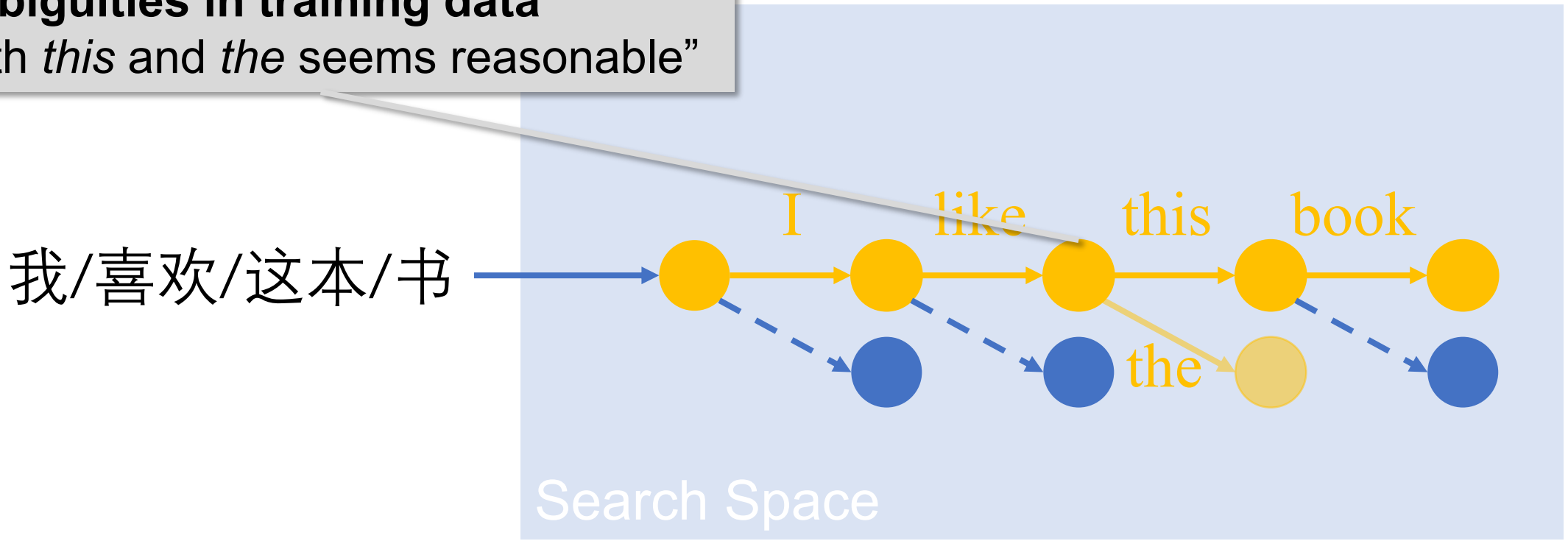


Generic $p(a \mid s)$ Learning Algorithm



Problems of the Generic Learning Algorithm

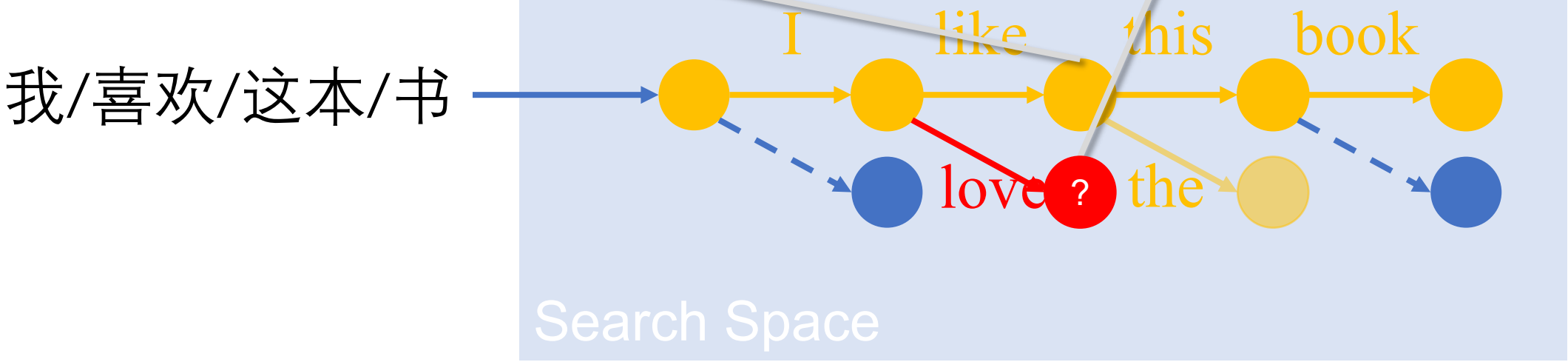
Ambiguities in training data
"both *this* and *the* seems reasonable"



Problems of the Generic Learning Algorithm

Ambiguities in training data
"both *this* and *the* seems reasonable"

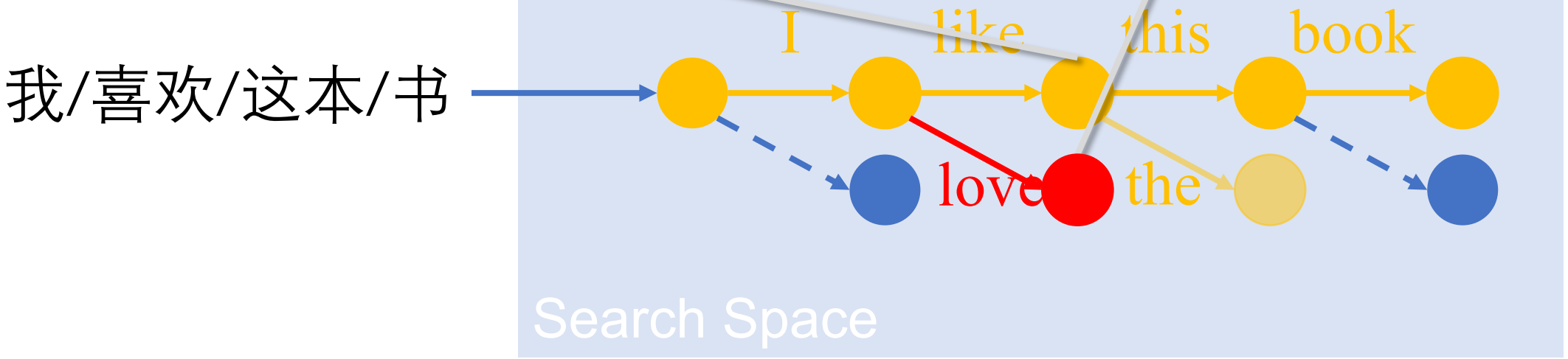
Training and test discrepancy
"What if I made wrong decision?"



Solutions in Previous Works

Ambiguities in training data
Ensemble *(Dietterich, 2000)*

Training and test discrepancy
Explore *(Ross and Bagnell, 2010)*

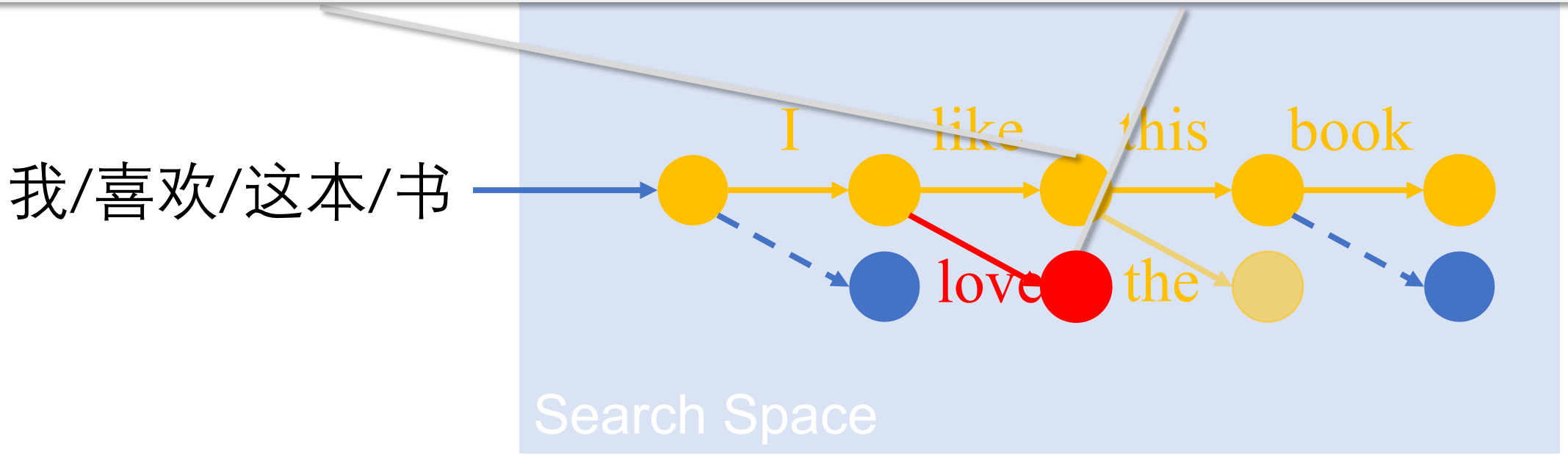


Where We Are

Knowledge Distillation

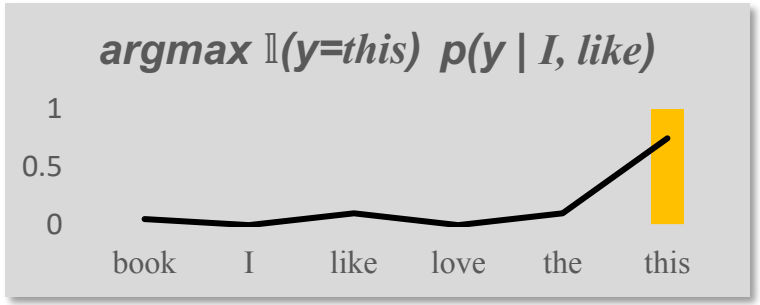
Ambiguities in training data

Training and test discrepancy

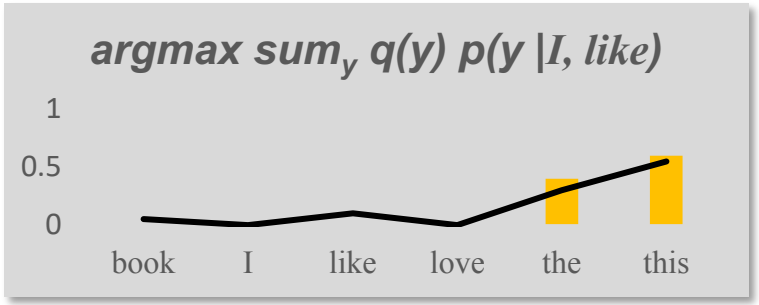


Knowledge Distillation

Learning from negative log-likelihood



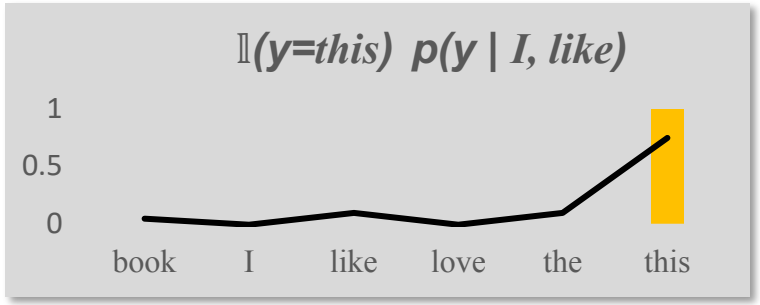
Learning from knowledge distillation



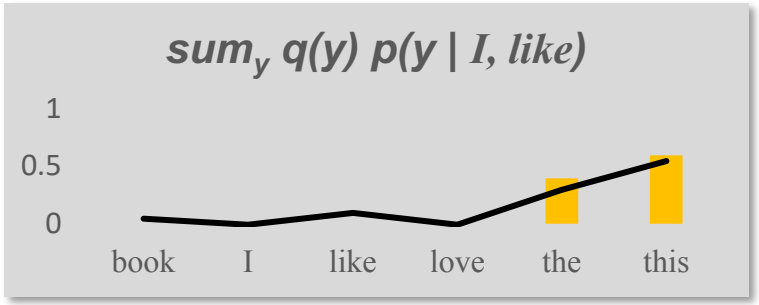
$q(y \mid I, like)$ is the output distribution of a ***teacher*** model (e.g. ensemble)

On supervised data argmax_p

$(1-\alpha)$

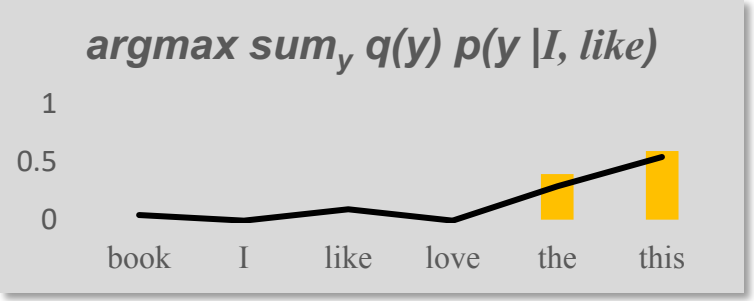


$+ \quad \alpha$



Knowledge Distillation: from Where

Learning from knowledge distillation

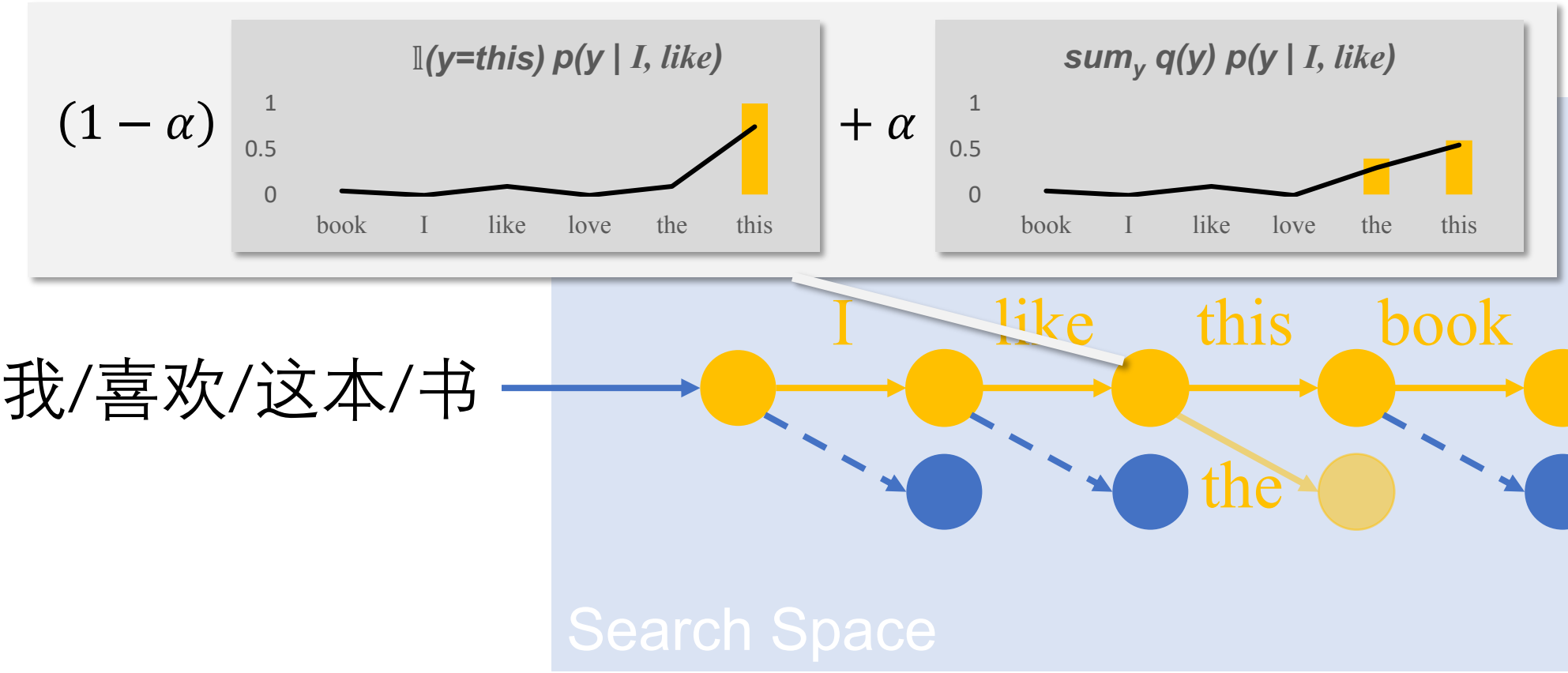


Ambiguities in training data

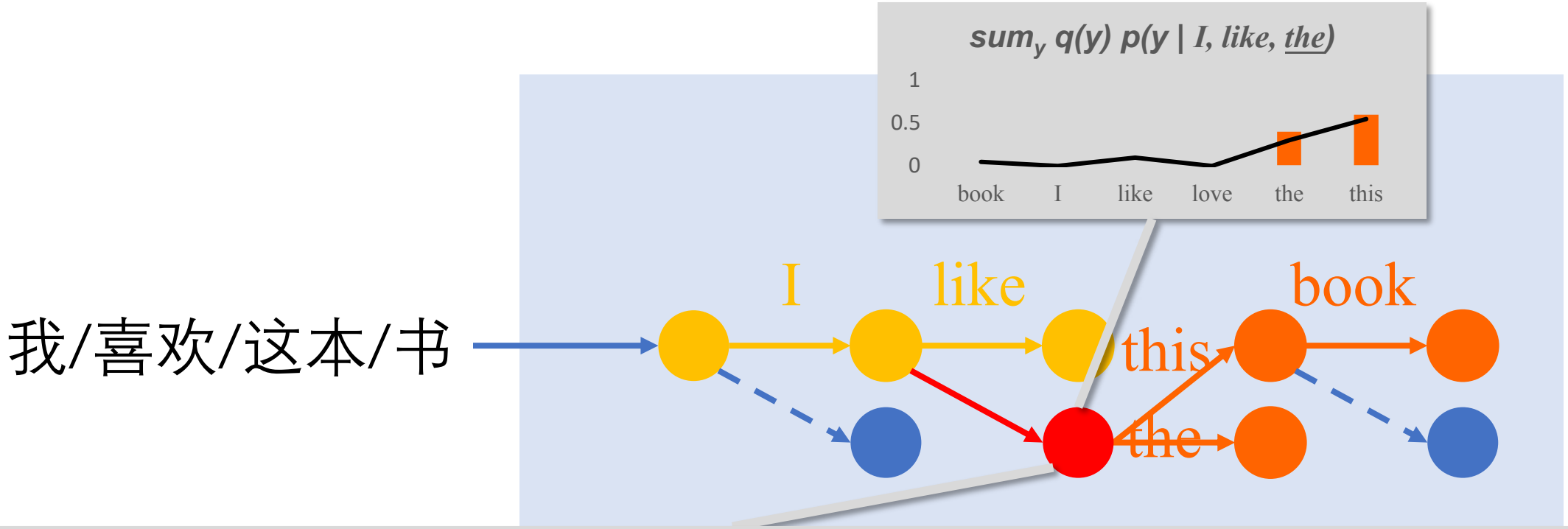
Ensemble *(Dietterich, 2000)*

We use ensemble of M structure predictor as the ***teacher q***

KD on Supervised (reference) Data



KD on Explored Data



Training and test discrepancy

Explore *(Ross and Bagnell, 2010)*

We use ***teacher q*** to explore the search space & learn from KD on the explored data

We combine KD on reference and explored data

$D \leftarrow \emptyset$;
for $n \leftarrow 1...N$ **do**
  $t \leftarrow 0$, $s_t \leftarrow s_0(\mathbf{x}^{(n)})$;
  while $s_t \notin \mathcal{S}_T$ **do**
    if *distilling from reference* **then**
      $a_t \leftarrow \pi_{\mathcal{R}}(s_t, \mathbf{y}^{(n)})$;
    else
      $a_t \leftarrow \pi_{\mathcal{E}}(s_t)$;
    $D \leftarrow D \cup \{s_t\}$, $s_{t+1} \leftarrow \mathcal{T}(s_t, a_t)$, $t \leftarrow t + 1$;
if *distilling from reference* **then**
  optimize $\alpha\mathcal{L}_{KD} + (1 - \alpha)\mathcal{L}_{NLL}$;
else
  optimize $\mathcal{L}_{KD}$;

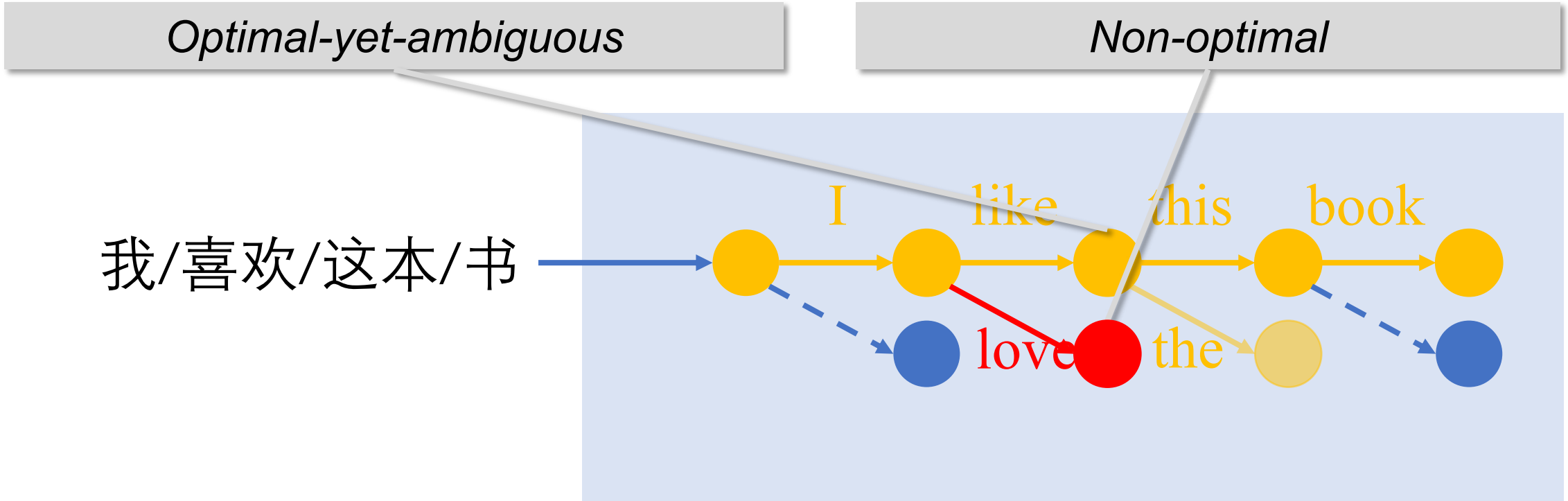
Experiments

Transition-based Dependency Parsing *Penn Treebank (Stanford dependencies)*	LAS
Baseline	90.83
Ensemble (20)	92.73
Distill (reference, $\alpha = 1.0$)	91.99
Distill (exploration)	92.00
Distill (both)	92.14
Ballesteros et al. (2016) (dyn. oracle)	91.42
Andor et al. (2016) (local, B=1)	91.02

Neural Machine Translation *IWSLT 2014 de-en*	BLEU
Baseline	22.79
Ensemble (10)	26.26
Distill (reference, $\alpha = 0.8$)	24.76
Distill (exploration)	24.64
Distill (both)	25.44
MIXER (Ranzato et al. 2015)	20.73
Wiseman and Rush (2016) (local B=1)	22.53
Wiseman and Rush (2016) (global B=1)	23.83

Analysis: Why the Ensemble Works Better?

- Examining the ensemble on the "**problematic**" states.

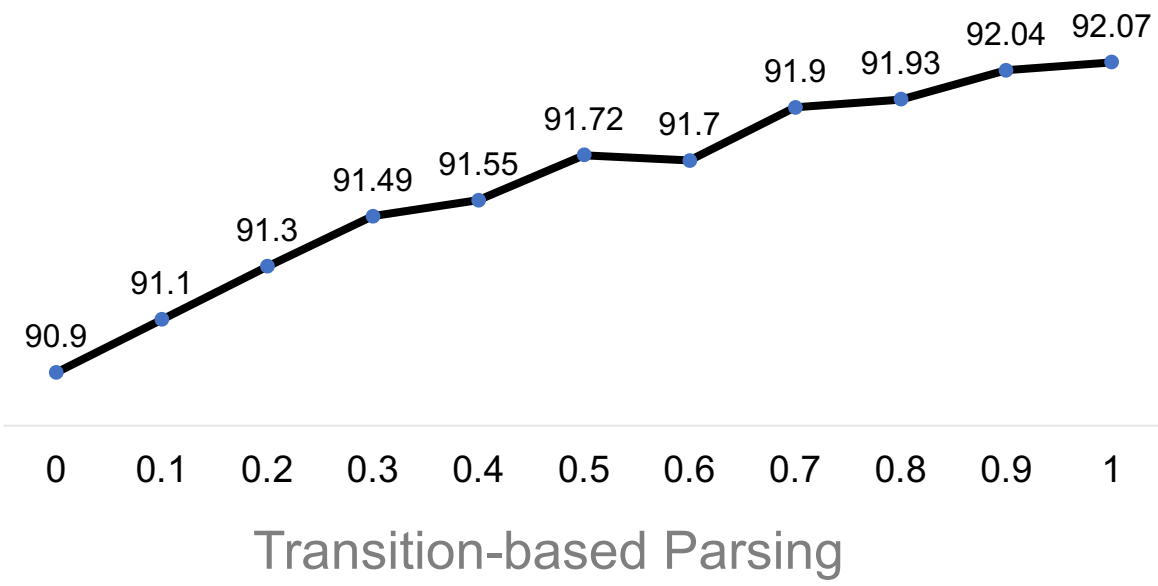


Analysis: Why the Ensemble Works Better?

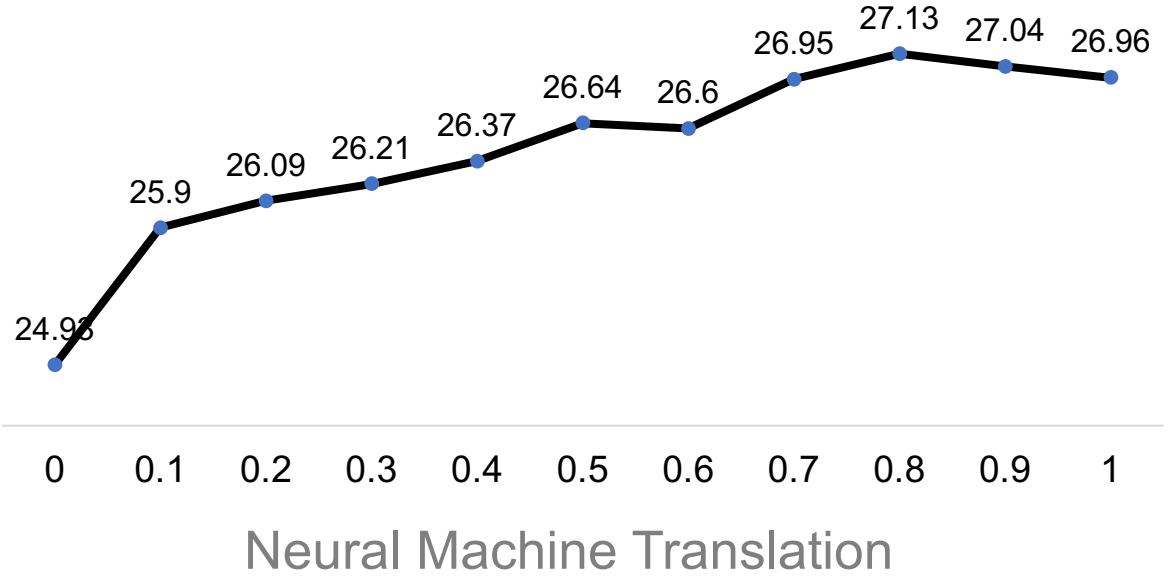
- Examining the ensemble on the "**problematic**" states.
- Testbed: *Transition-based dependency parsing*.
- Tools: **dynamic oracle**, which returns a set of reference actions for one state.
- Evaluate the output distributions against the reference actions.

	optimal-yet-ambiguous	non-optimal
Baseline	68.59	89.59
Ensemble	74.19	90.90

Analysis:
Is it Feasible to Fully Learn from KD w/o NLL?



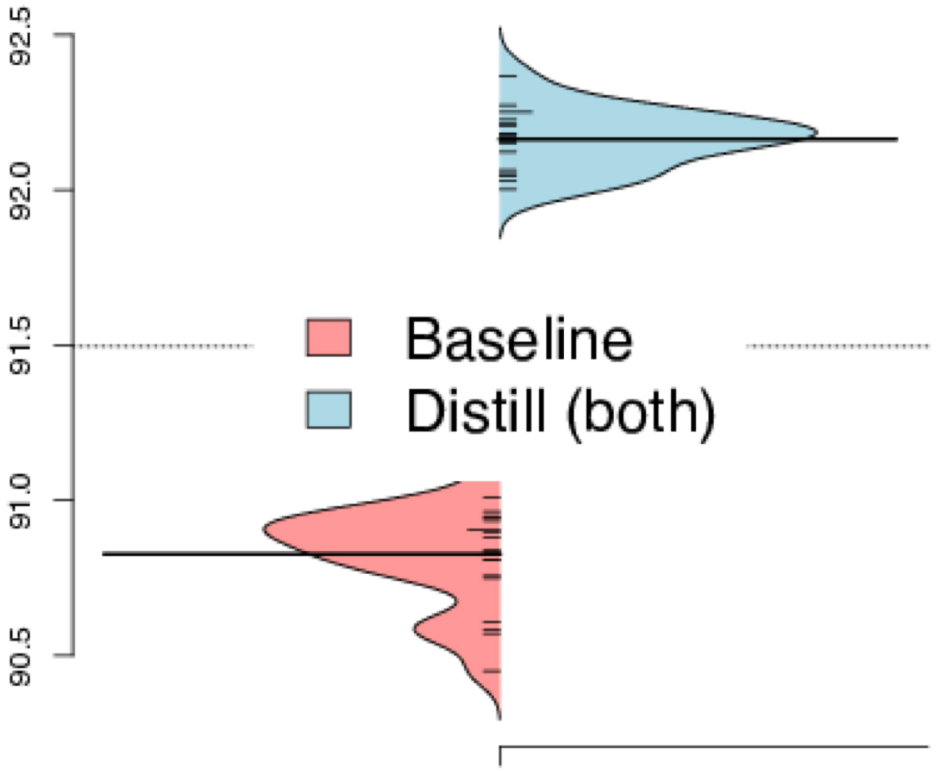
Transition-based Parsing



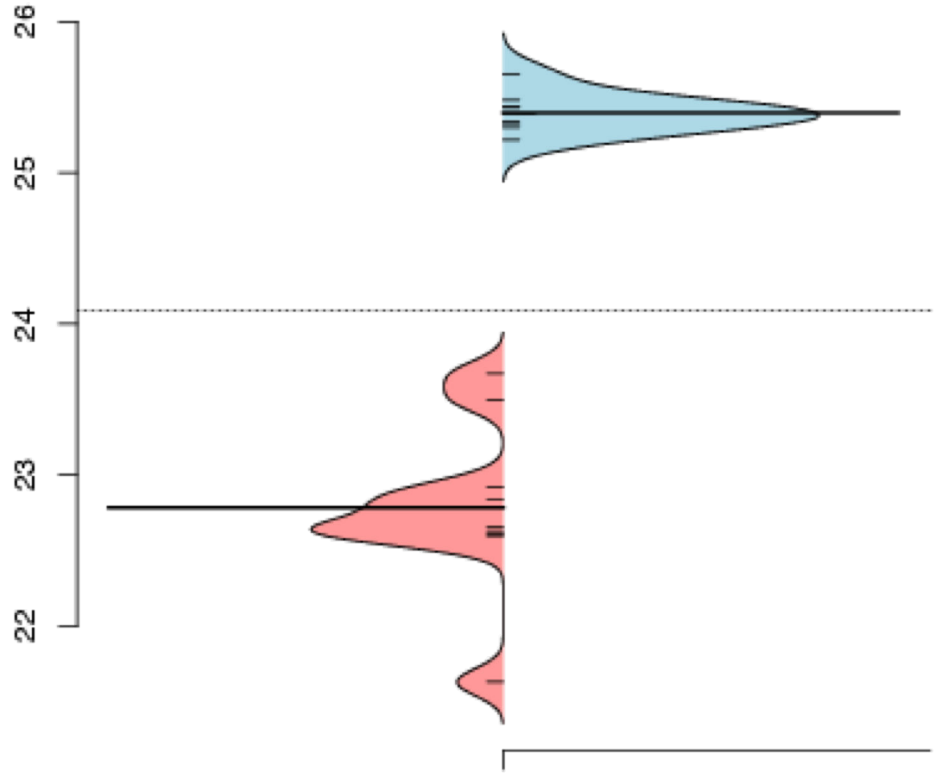
Neural Machine Translation

Fully learning from KD is feasible

Analysis:
Is Learning from KD Stable?



Transition-based Parsing



Neural Machine Translation

Conclusion

- We propose to distill an ensemble into a single model both from reference and exploration states.
- Experiments on transition-based dependency parsing and machine translation show that our distillation method significantly improves the single model's performance.
- Analysis gives empirically guarantee for our distillation method.

Thanks and Q/A