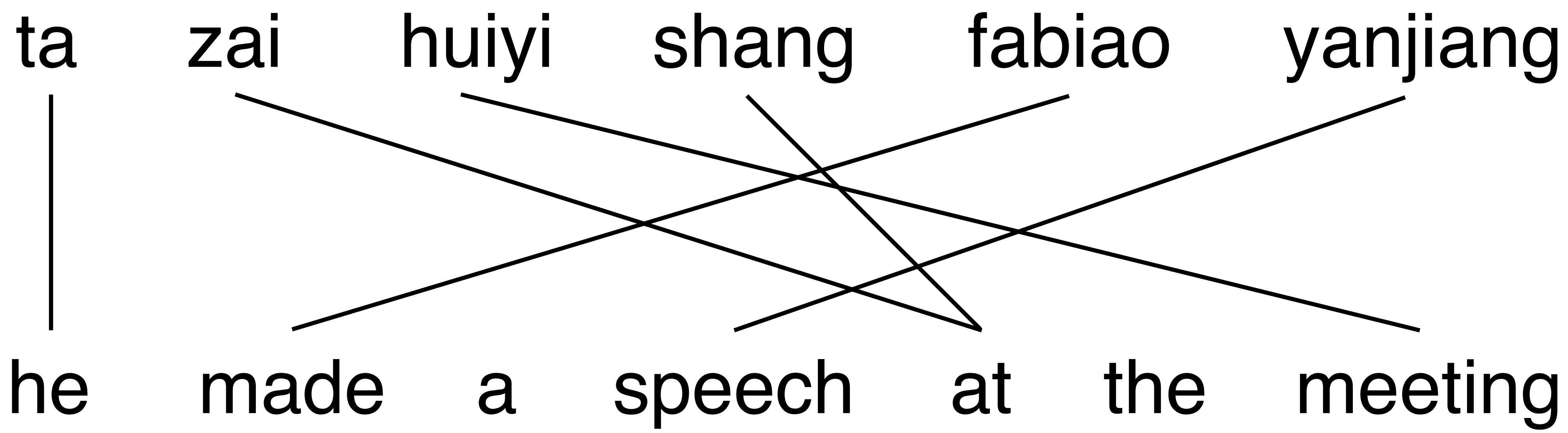
Contrastive Unsupervised Word Alignment with Non-Local Features

Yang Liu and Maosong Sun

Word Alignment

- Word alignment: aligning words between two languages



Approaches

- **Generative** [Brown et al., 1993; Vogel et al., 1996, Liang et al., 2006]
  - pros: no need for labeled data
  - cons: hard to extend
- **Discriminative** [Taskar et al., 2005; Moore et al., 2006; Liu et al., 2010]
  - pros: easy to extend
  - cons: rely on labeled data

Latent-Variable Log-Linear Models

sentence pair

alignment

parameters

$$P(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} P(\mathbf{x},\mathbf{y};\boldsymbol{\theta})$$

$$= \frac{\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \exp(\boldsymbol{\theta}\cdot\boldsymbol{\phi}(\mathbf{x},\mathbf{y}))}{Z(\boldsymbol{\theta})}$$

partition function

features

Challenge

training data $\{\mathbf{x}^{(i)}\}_{i=1}^{I}$

objective $$L(\boldsymbol{\theta}) = \sum_{i=1}^{I} \log \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(i)})} \exp(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\mathbf{x}^{(i)}, \mathbf{y})) - \log Z(\boldsymbol{\theta})$$

derivative $$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} = \sum_{i=1}^{I} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta}}[\boldsymbol{\phi}_k(\mathbf{x}^{(i)}, \mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y};\boldsymbol{\theta}}[\boldsymbol{\phi}_k(\mathbf{x}, \mathbf{y})]$$

intractable to calculate two feature expectations

[Smith and Eisner, 2005; Berg-Kirkpatrick et al., 2010; Dyer et al., 2011]

Idea

observation

ta	zai	huiyi	shang	fabiao	yanjiang	
he	made	a	speech	at	the	meeting

noise

zai	fabiao	huiyi	shang	wo	yanjiang	
talk	a	meeting	she	at	the	made

Intuition: observations have higher probabilities than noises

Contrastive Learning

training data $\{\langle \mathbf{x}^{(i)}, \tilde{\mathbf{x}}^{(i)} \rangle\}_{i=1}^{I}$

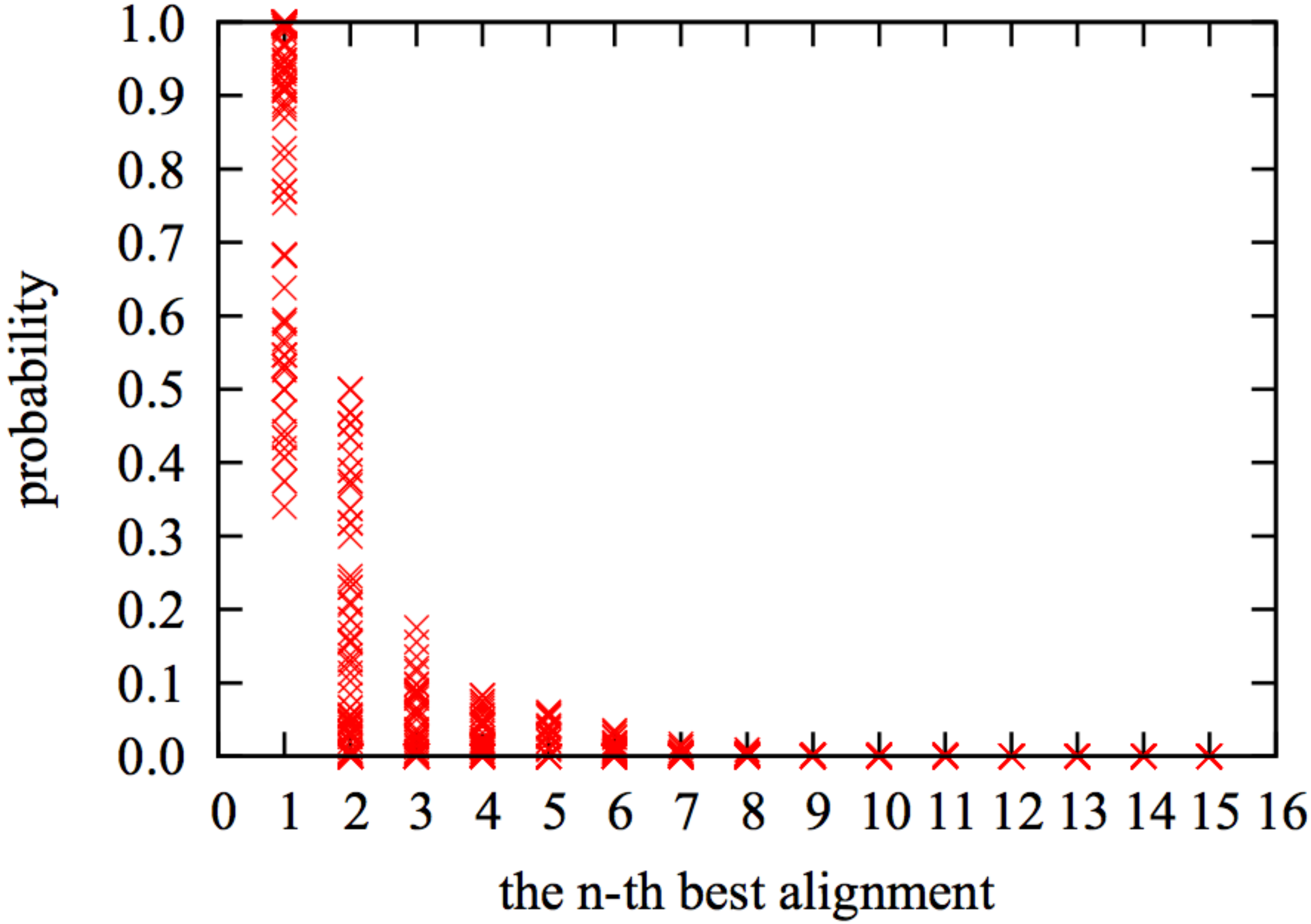
objective $$J(\boldsymbol{\theta}) = \log \prod_{i=1}^{I} \frac{P(\mathbf{x}^{(i)}; \boldsymbol{\theta})}{P(\tilde{\mathbf{x}}^{(i)}; \boldsymbol{\theta})}$$

derivative $$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} = \sum_{i=1}^{I} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta}}[\phi_k(\mathbf{x}^{(i)}, \mathbf{y})] - \mathbb{E}_{\mathbf{y}|\tilde{\mathbf{x}}^{(i)};\boldsymbol{\theta}}[\phi_k(\tilde{\mathbf{x}}^{(i)}, \mathbf{y})]$$

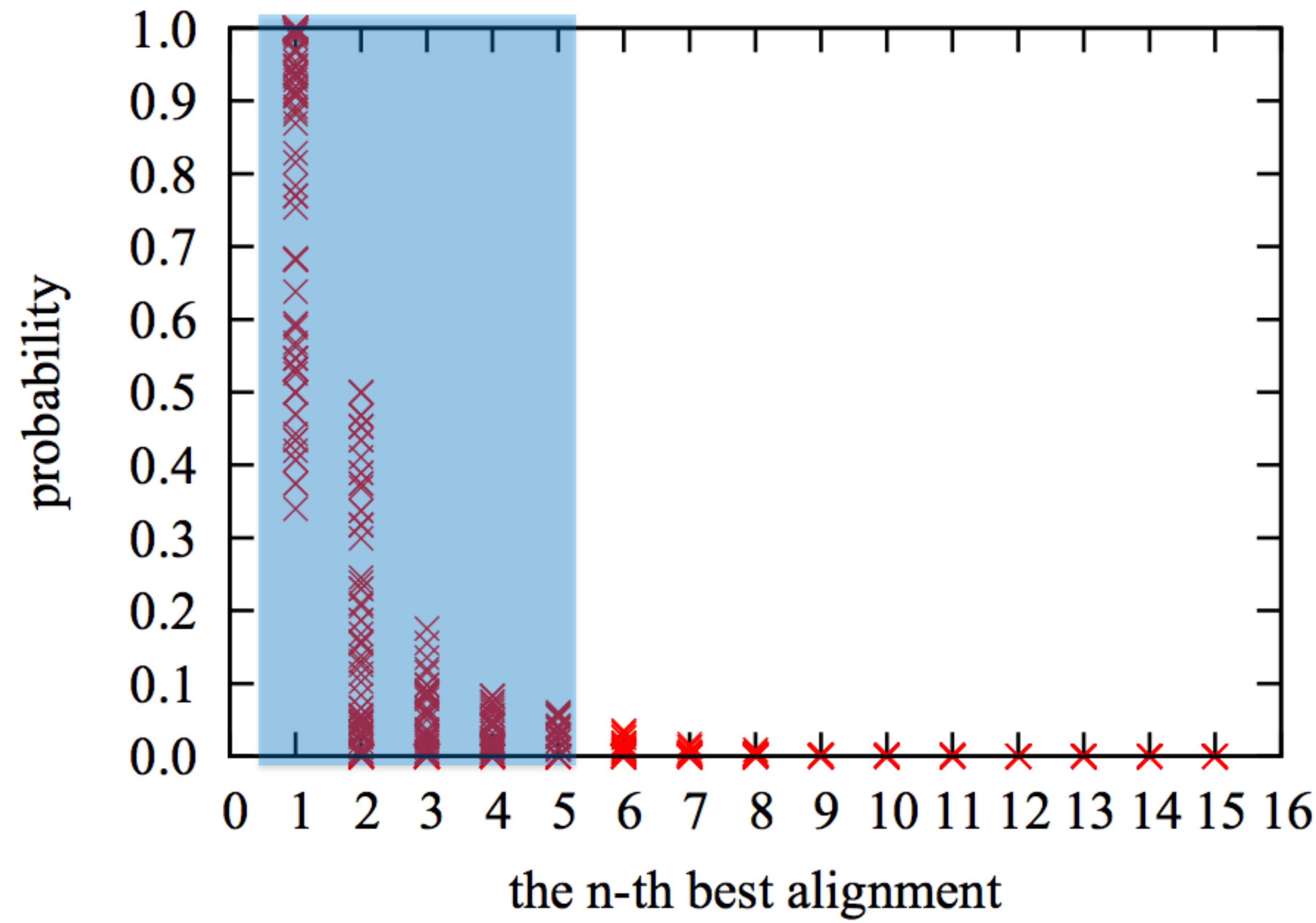
partition function canceled out

Concentration

- Alignments with higher probabilities are more important in calculating expectations



Top-n Sampling



$$\mathbb{E}_{\mathbf{y}|\mathbf{x};\boldsymbol{\theta}}[\phi_k(\mathbf{x},\mathbf{y})] \approx \frac{\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x};\boldsymbol{\theta})} \exp(\boldsymbol{\theta}\cdot\boldsymbol{\phi}(\mathbf{x},\mathbf{y}))\phi_k(\mathbf{x},\mathbf{y})}{\sum_{\mathbf{y}'\in\mathcal{N}(\mathbf{x};\boldsymbol{\theta})} \exp(\boldsymbol{\theta}\cdot\boldsymbol{\phi}(\mathbf{x},\mathbf{y}'))}$$

Comparison with Gibbs Samping

# samples	Gibbs	Top-n
1	1.5411	0.1653
5	0.7410	0.1477
10	0.6550	0.1396
50	0.5498	0.1108
100	0.5396	0.1086
500	0.5180	0.0932

Comparison with Gibbs sampling in terms of average approximation error

Effect of Noise Generation

noise generation	French-English	Chinese-English
SHUFFLE	8.93	21.05
DELETE	9.03	21.49
INSERT	12.87	24.87
REPLACE	13.13	25.59

Effect of noise generation in terms of alignment error rate

Final Result

system	**model**	**supervision**	**algorithm**	**French-English**	**Chinese-English**
GIZA++	IBM model 4	unsupervised	EM	6.36	21.92
Berkeley	joint HMM	unsupervised	EM	5.34	21.67
fast_align	log-linear model	unsupervised	EM	15.20	28.44
Vigne	linear model	supervised	MERT	4.28	19.37
this work	log-linear model	unsupervised	SGD	5.01	20.24

Comparison with state-of-the-art aligners

Conclusion

- Word alignment is important for multilingual NLP tasks
- Unsupervised learning of latent-variable log-linear models combines the merits of generative and discriminative approaches
- We have proposed an efficient and accurate learning algorithm for unsupervised word alignment with arbitrary features
- We will apply our approach to other NLP tasks

Thank You

Source code and data sets are freely available at:
http://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html