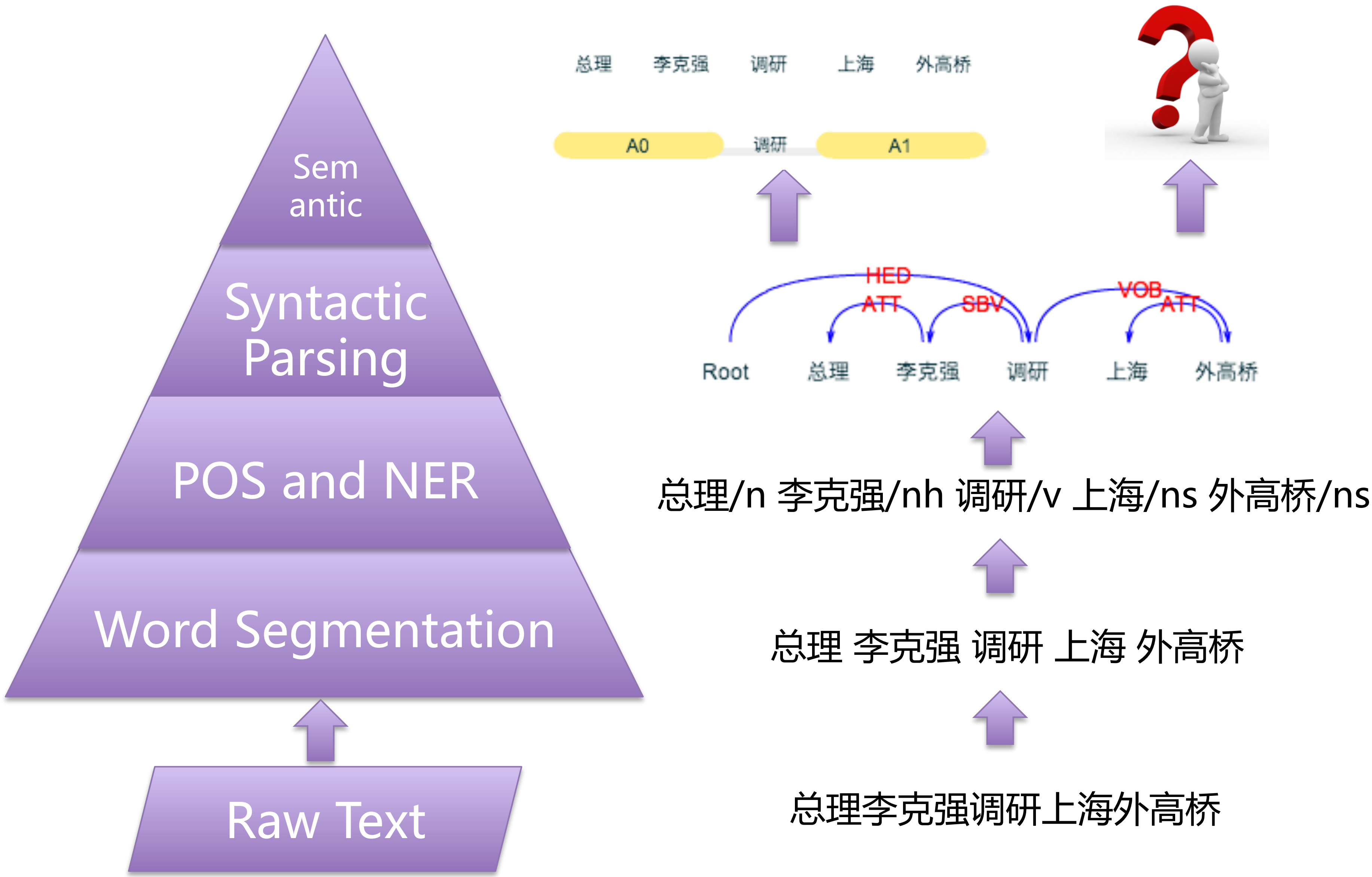
Cross-lingual based NLP

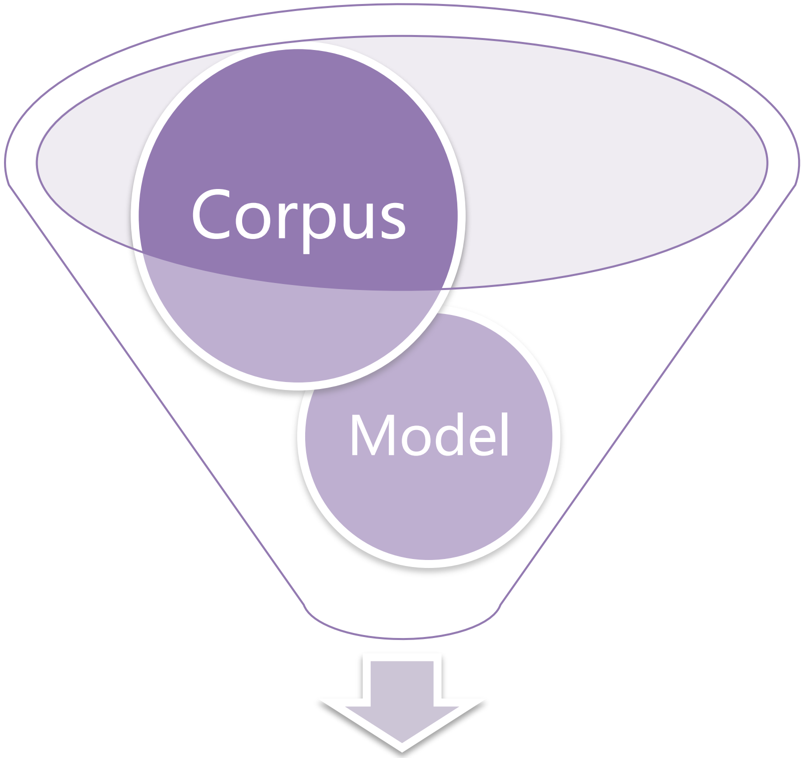
Wanxiang Che
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology
2015-4-16

Tasks of Basic NLP



Methodology

- Statistical NLP



NLP System

Challenges of NLP

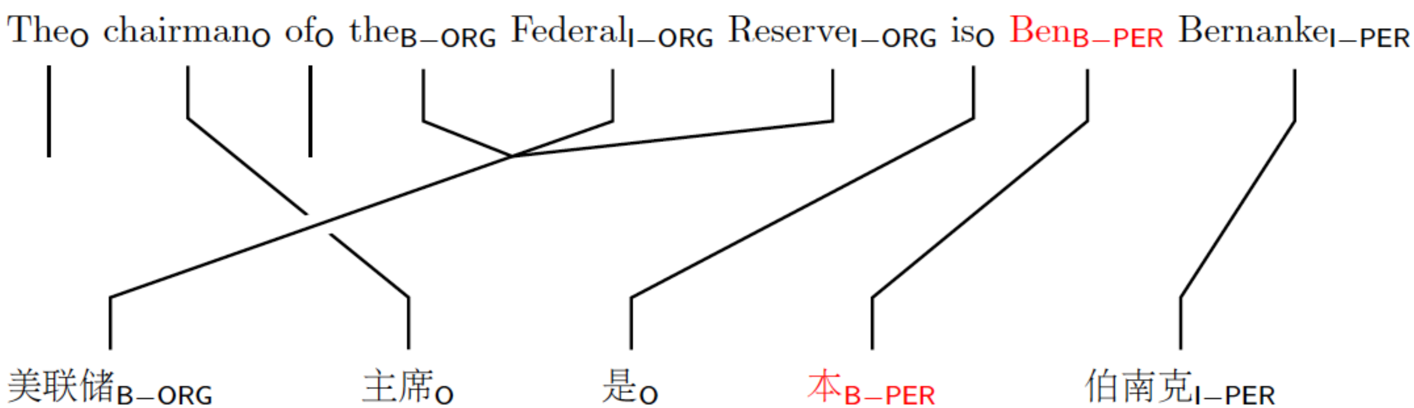
- Lack of Training Data
- Domain Adaptation
- Error Propagation
- Semantic Parsing

Challenges of NLP

- **Lack of Training Data**
- Domain Adaptation
- Error Propagation
- Semantic Parsing

A Solution to Lack of Training Data

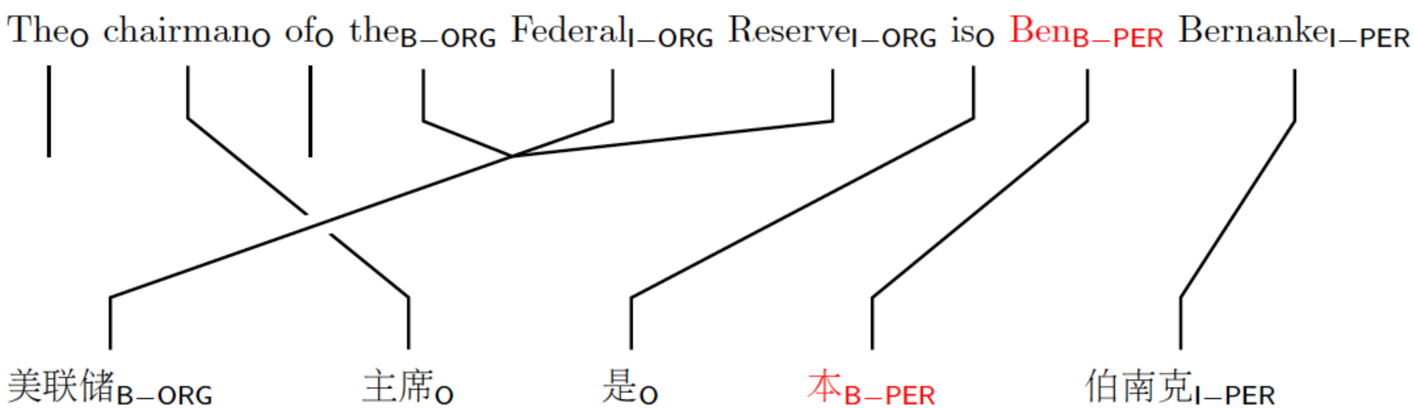
- Based on Cross-lingual Clue



- Methods
  - Cross-lingual Annotation Projection
  - Joint Bilingual Modeling
  - Cross-lingual Transfer

A Solution to Lack of Training Data

- Based on Cross-lingual Clue

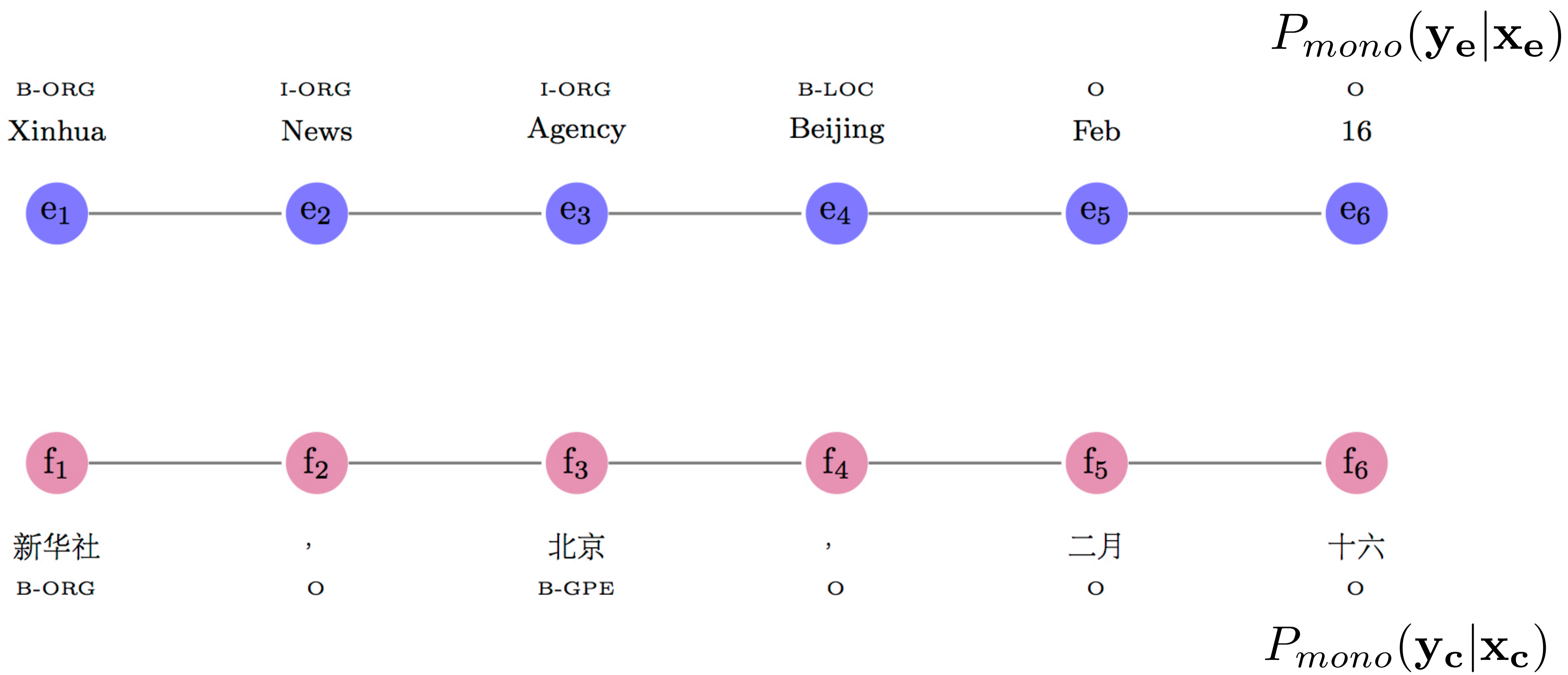


- Methods
  - Cross-lingual Annotation Projection
  - **Joint Bilingual Modeling**
  - **Cross-lingual Transfer**

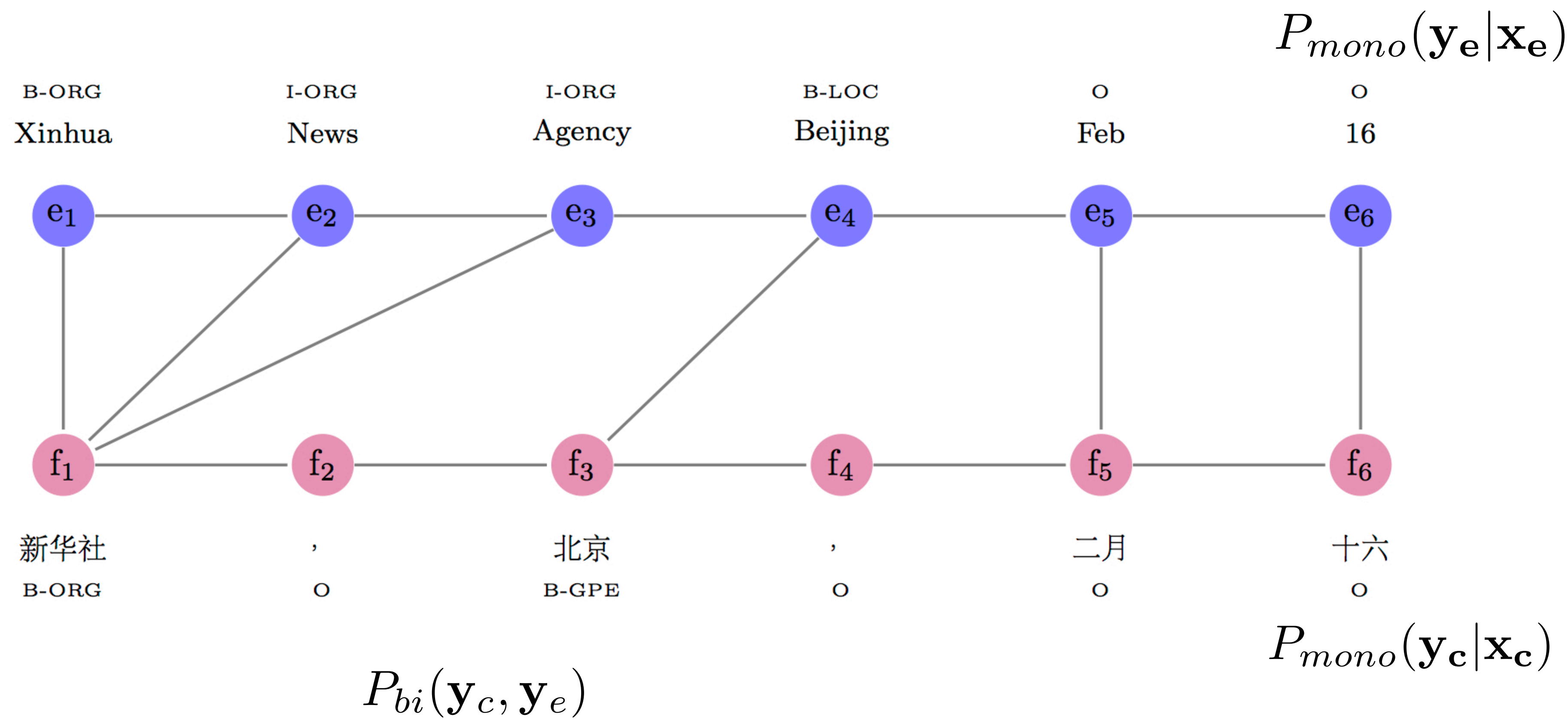
Joint Named Entity Recognition

- Named Entity
  - Person, Location, Organization, …
  - *The chairman of the Federal Reserve [ORG] is Ben Bernanke [PER]*
- State-of-the-art Methods
  - Sequence tagging, such as CRF
  - Require large amounts of annotated data
  - Difficult and expensive to annotate

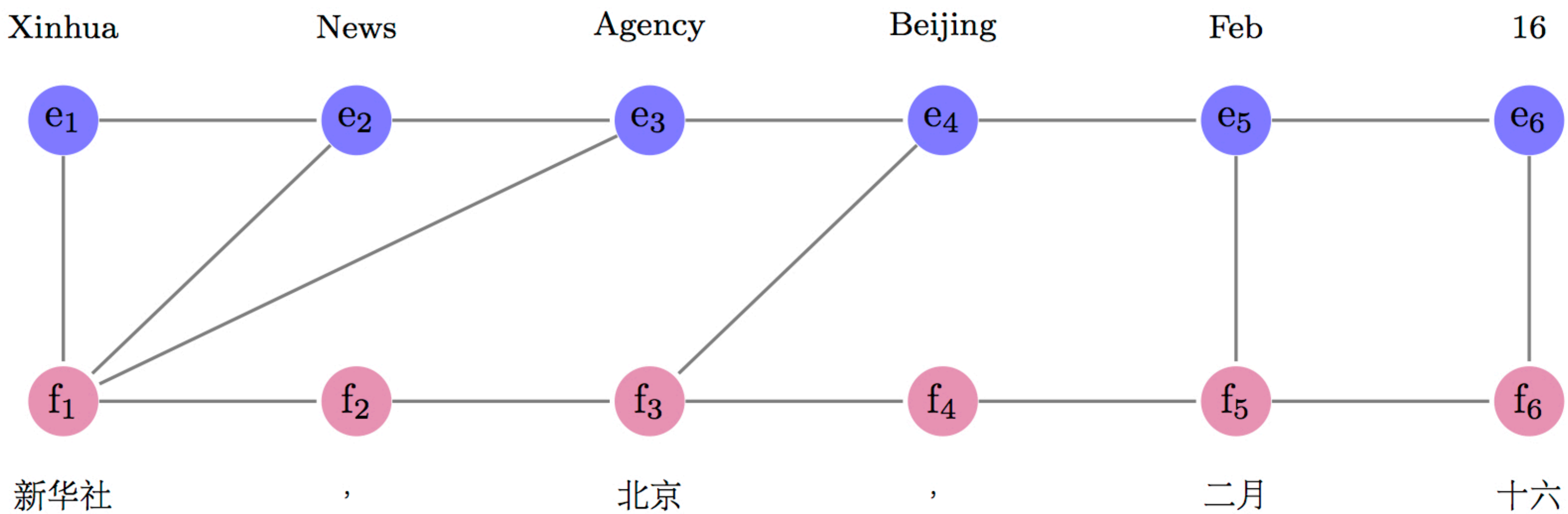
Bilingual NER w/o Constraints



Bilingual NER with Constraints



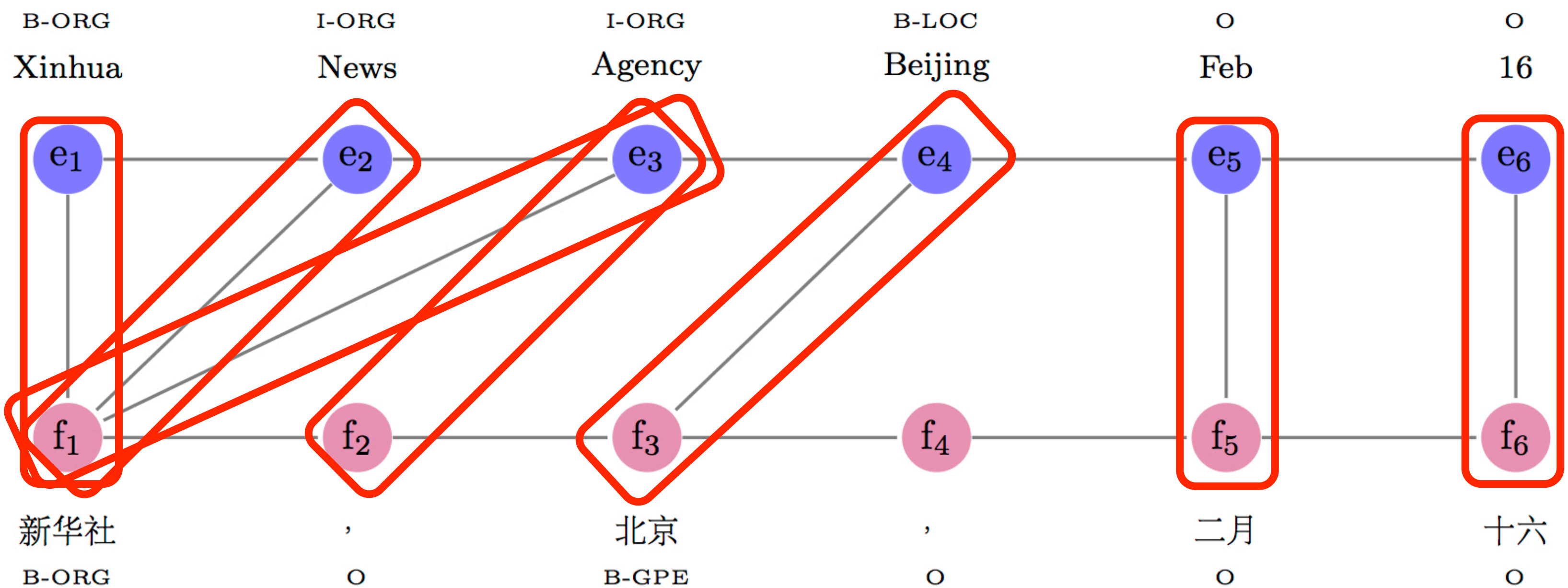
Bilingual Factored Model



$$P(\mathbf{y}_c, \mathbf{y}_e | \mathbf{x}_c, \mathbf{x}_e) = P_{mono}(\mathbf{y}_c | \mathbf{x}_c) P_{mono}(\mathbf{y}_e | \mathbf{x}_e) P_{bi}(\mathbf{y}_c, \mathbf{y}_e)$$

Bilingual NER with Constraints

$$P_{mono}(\mathbf{y_e}|\mathbf{x_e})$$

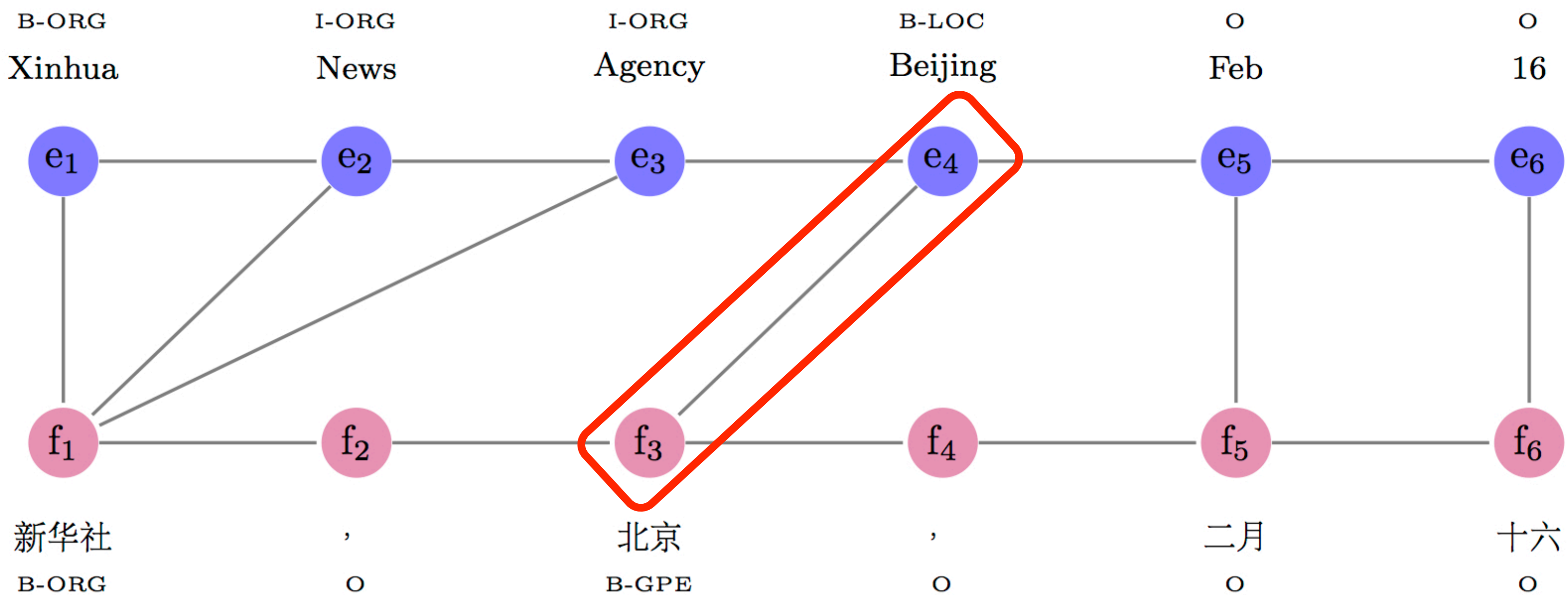


$$P_{mono}(\mathbf{y_c}|\mathbf{x_c})$$

$$P_{bi}(\mathbf{y}_c, \mathbf{y}_e) = \prod_{\boxed{A=\{a^c, a^e\}}} \mathbb{I}(y_{a^c}, y_{a^e})$$

Bilingual NER with Constraints

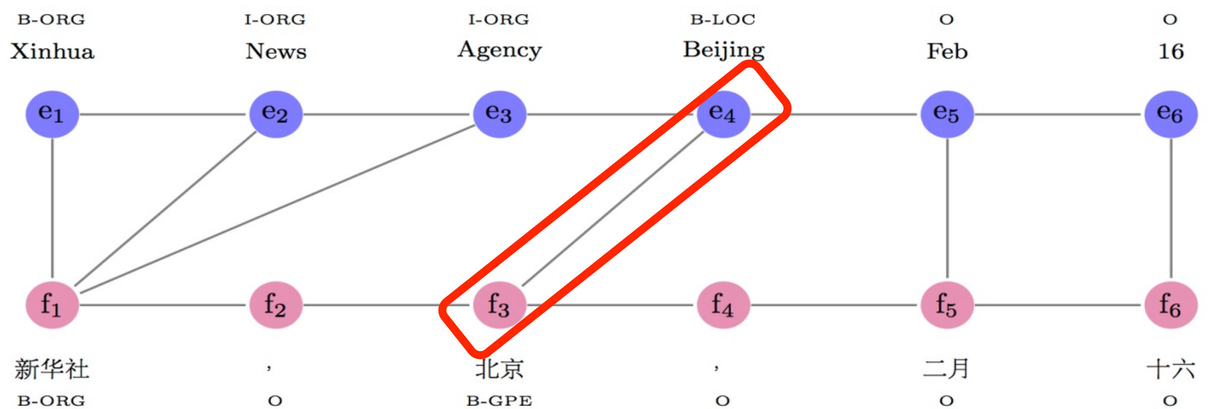
$$P_{mono}(\mathbf{y_e}|\mathbf{x_e})$$



$$P_{mono}(\mathbf{y_c}|\mathbf{x_c})$$

$$P_{bi}(\mathbf{y}_c, \mathbf{y}_e) = \prod_{A=\{a^c, a^e\}} \boxed{\mathbb{I}(y_{a^c}, y_{a^e})}$$

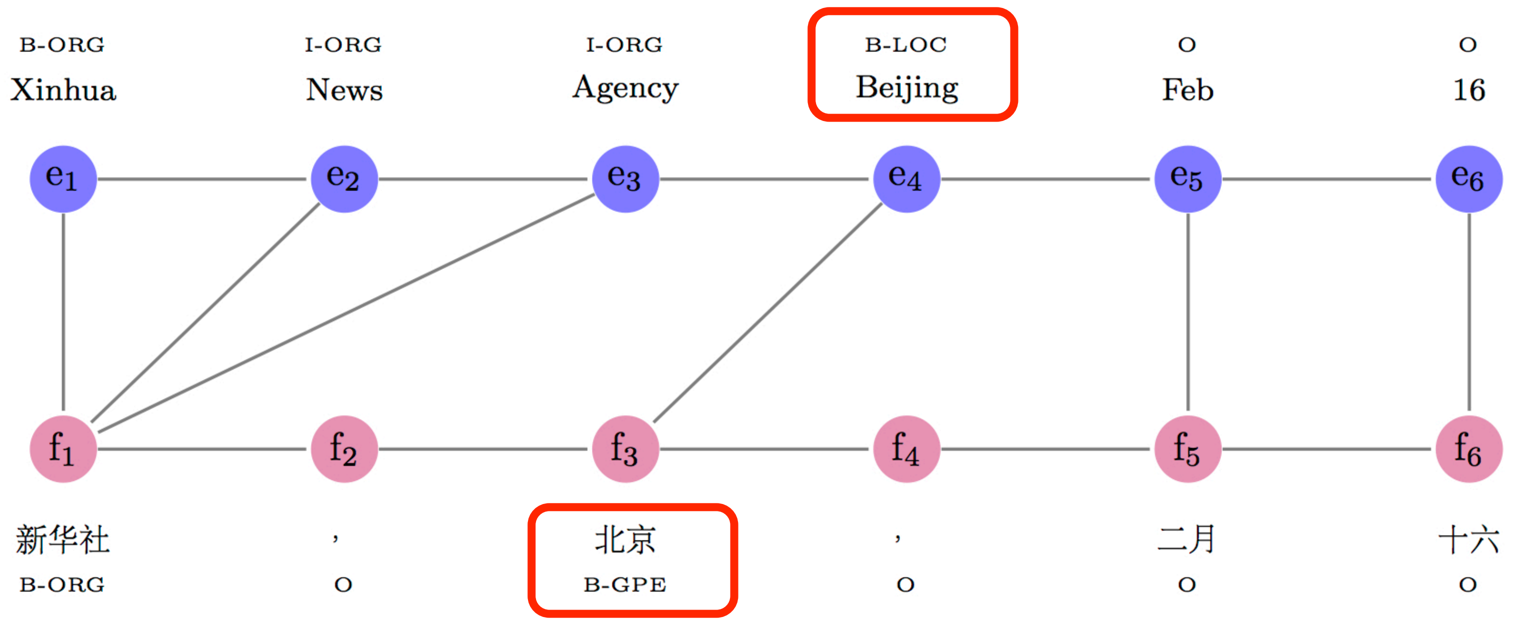
Hard Agreement Constraints



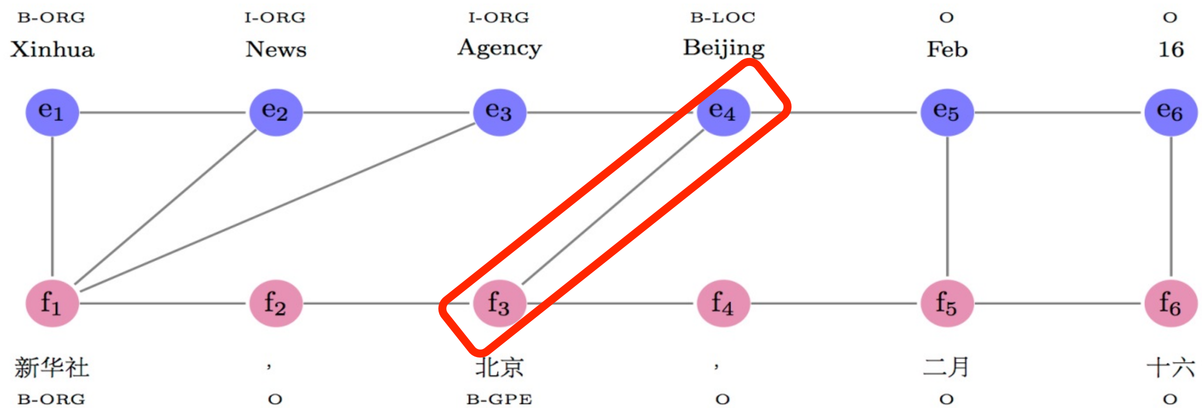
$$P_{bi}(\mathbf{y}_c, \mathbf{y}_e) = \prod_{A=\{a^c, a^e\}} \boxed{\mathbb{I}(y_{a^c}, y_{a^e})}$$

$$\mathbb{I}(y_{a^c}, y_{a^e}) = \begin{cases} 1, \text{if } y_{a^c} = y_{a^e} \\ 0, \text{else} \end{cases}$$

Inconsistency in Annotation Standards



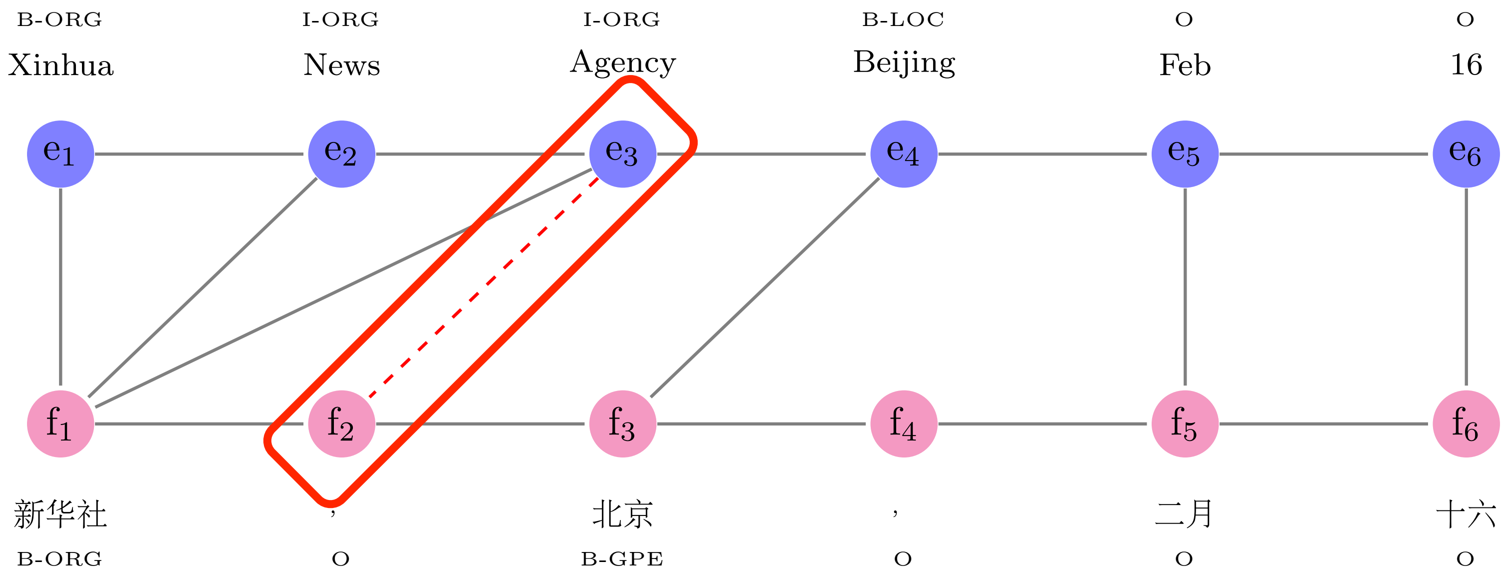
Soft Agreement Constraints



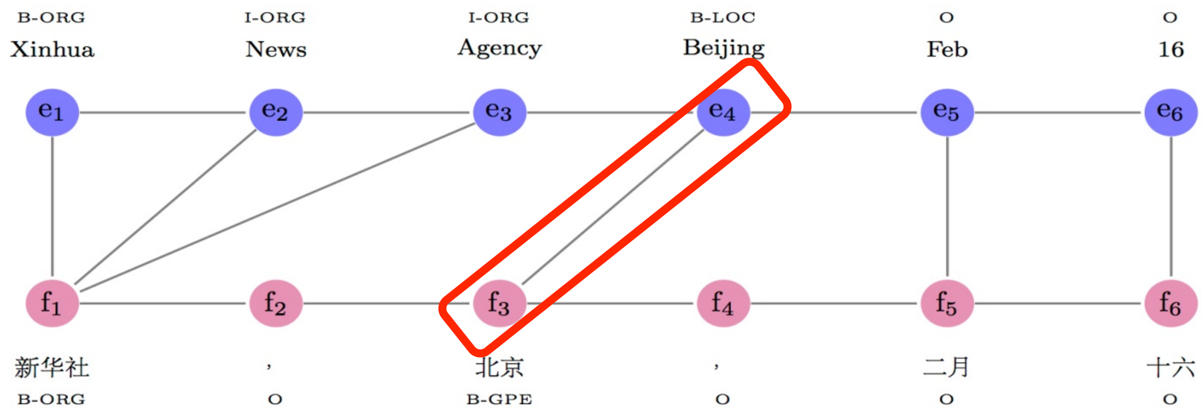
$$P_{bi}(\mathbf{y}_c, \mathbf{y}_e) = \prod_{A=\{a^c, a^e\}} \boxed{\mathbb{I}(y_{a^c}, y_{a^e})}$$

$$\mathbb{I}(y_{a^c}, y_{a^e}) = \mathrm{pmi}(y_{a^c}, y_{a^e})$$

Alignment Error



Modeling Alignment Uncertainty



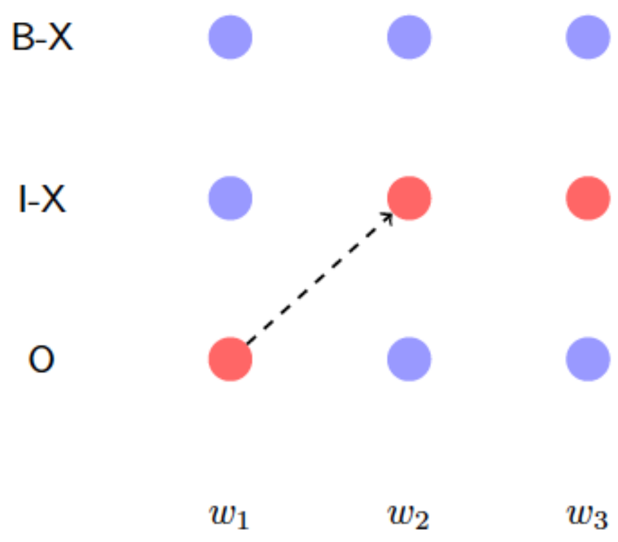
$$P_{bi}(\mathbf{y}_c, \mathbf{y}_e) = \prod_{A=\{a^c, a^e\}} \boxed{\mathbb{I}(y_{a^c}, y_{a^e})}$$

$$\mathrm{II}(y_{a^c}, y_{a^e}) = \mathrm{pmi}(y_{a^c}, y_{a^e})^{\boxed{P(y_{a^c}, y_{a^e})}}$$

Solve with Integer Linear Programing

- Monolingual NER Objective Function

$$\max \sum_{i=1}^{|\mathbf{x}|} \sum_{y \in Y} z_i^y \log P_i^y$$



- Constrains

$$\forall i : \sum_{y \in Y} z_i^y = 1$$

$$\forall i, \forall X : z_{i-1}^{\text{B-X}} + z_{i-1}^{\text{I-X}} - z_i^{\text{I-X}} \geq 0$$

Solve with Integer Linear Programing

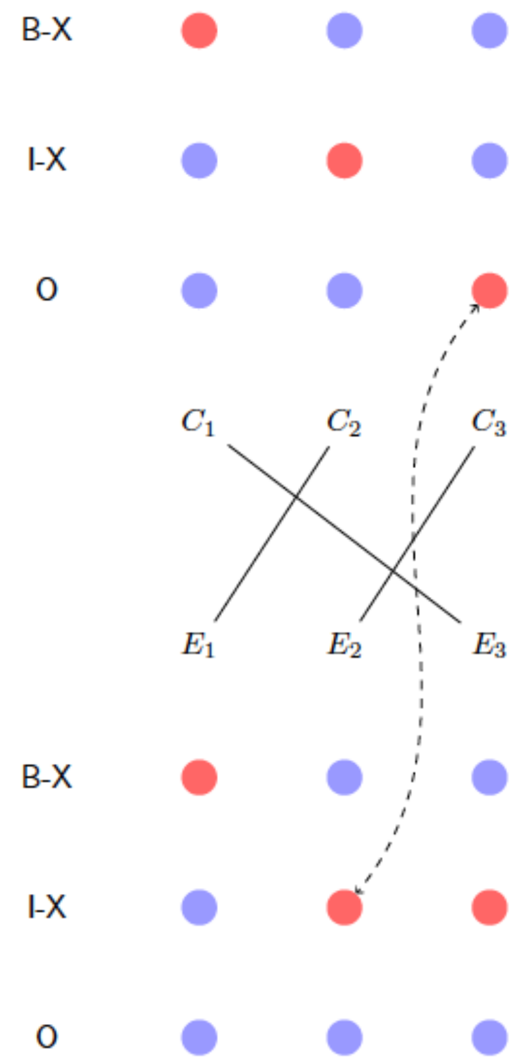
- Bilingual NER Objective Function

$$\max \sum_{i=1}^{|\mathbf{x}_c|} \sum_{y \in Y} z_i^y \log P_i^y + \sum_{j=1}^{|\mathbf{x}_e|} \sum_{y \in Y} z_j^y \log P_j^y + \sum_{a \in \mathscr{A}} \sum_{y_c \in Y} \sum_{y_e \in Y} z_a^{y_c y_e} P_a \log \lambda_a^{y_c y_e}$$

- Constrains
  - Monolingual
  - Bilingual

$$\forall a \in A : \sum_{y_c \in Y} \sum_{y_e \in Y} z_a^{y_c y_e} = 1$$

$$\forall a = (i, j) \in A : z_a^{y_c y_e} \le z_i^{y_c}, z_a^{y_c y_e} \le z_j^{y_e}$$



Experimental Results

	Chinese			English		
	P	R	F_1	P	R	F_1
CRF (No Cluster)	74.74	56.17	64.13	–	–	–
CRF (Word Cluster)	76.90	63.32	69.45	**82.95**	76.67	79.68
Monolingual ILP	76.20	63.06	69.01	82.88	76.68	79.66
Hard	74.38	65.78	69.82	82.66	75.36	78.84
Soft-tag (`Auto`)	77.37	71.14	74.13	81.36	**78.74**	80.03
Soft-align (`Auto`)	**77.71**	**72.51**	**75.02**	81.94	78.35	**80.10**

Semi-supervised Results

Method	#sent	P	R	F_1
CRF	∼16k	76.90	63.32	69.45
Semi	10k	77.60	66.51	71.62
	20k	77.28	67.26	71.92
	40k	77.40	67.81	72.29
	80k	77.44	68.64	72.77
	160k	78.04	69.83	73.71

Che et al. NAACL 2013

Modeling Global Consistency

- Global consistency: occurrences of the same word sequence within a given document are unlikely to take on different entity types
- Using **Gibbs sampling** to incorporate non-local constraints

Wang et al. AAAI 2013. Outstanding Paper Award Honorable Mention

Consistency Results

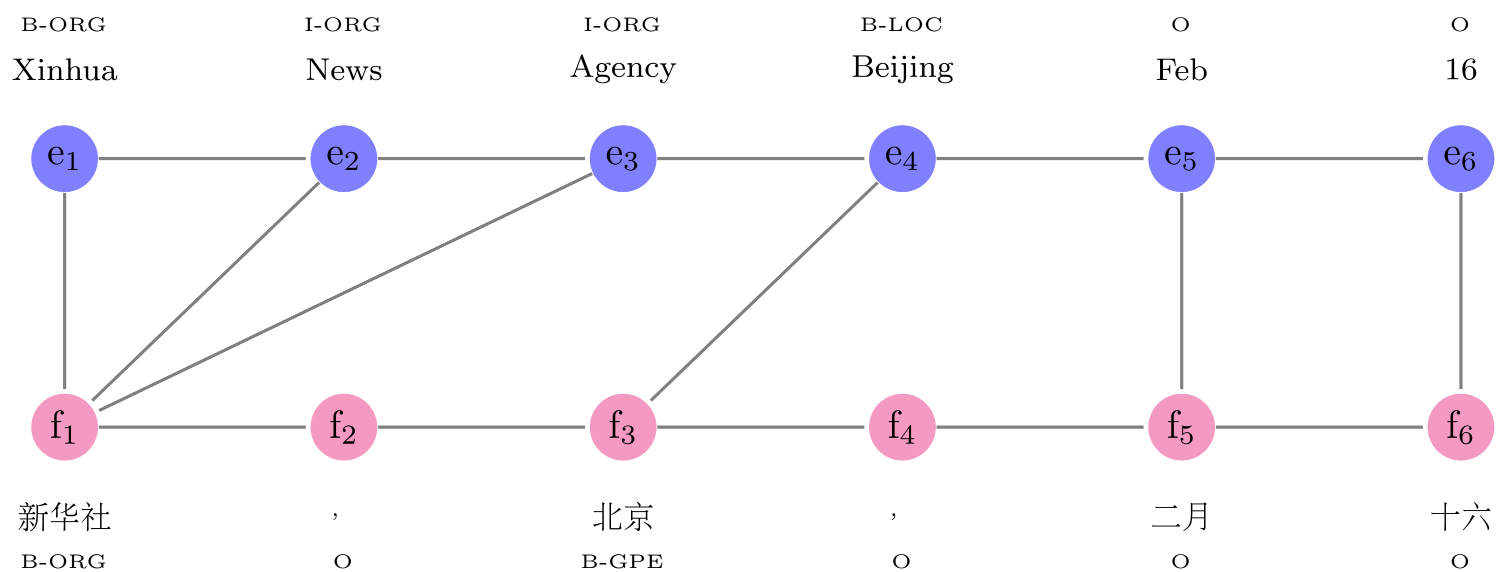
	Chinese			English		
	P	R	F_1	P	R	F_1
mono	76.89	61.64	68.42	81.98	74.59	78.11
+*global*	77.30	58.96	66.90	83.89	74.88	79.13
+*global-recall*	75.23	68.12	71.50	82.31	77.63	79.90
$\mathrm{PMI}^{\mathrm{alignProb}}$	79.17	68.46	73.43	82.05	75.56	78.67
+*global*	**79.31**	65.93	72.01	84.01	75.81	79.70
+*global-recall*	76.43	**72.32**	**74.32**	82.30	78.35	80.28

Joint NER and Word Alignment

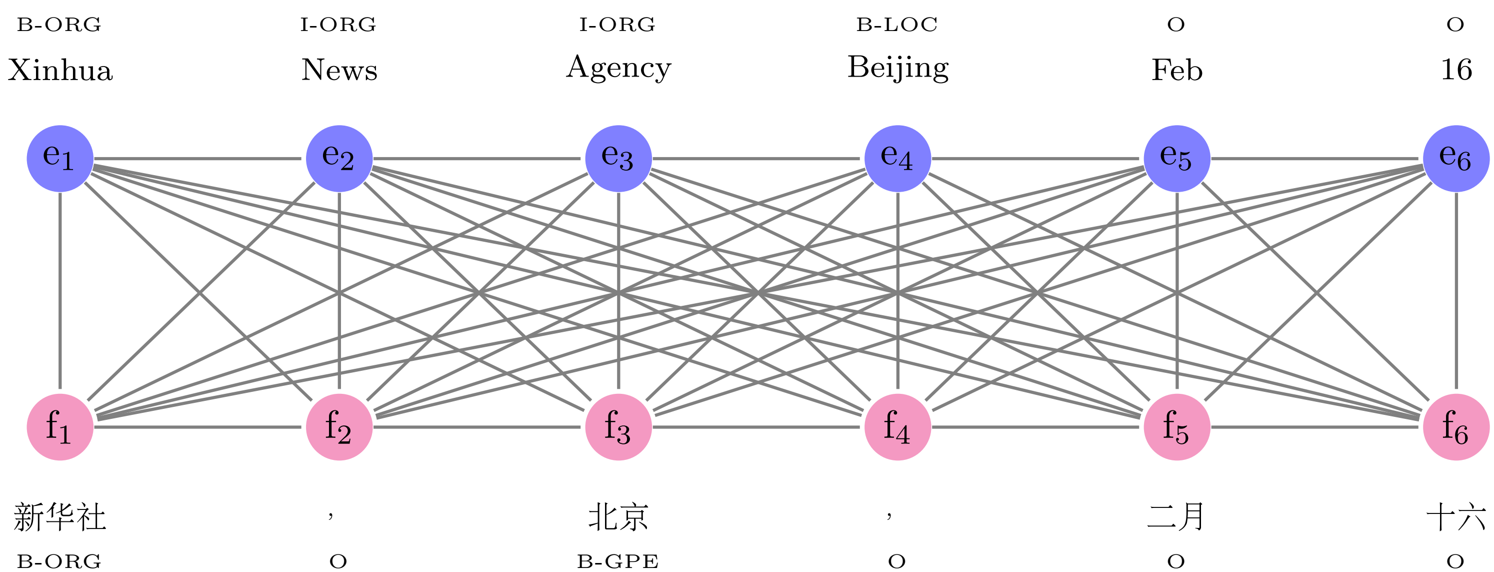
- In the hard and soft agreement models, we assumed that word alignment is given as **fixed** input
- We observe that NER labels can be used to help correct word alignment errors (e.g., functional words in EN should not be aligned to a word of entity type PERSON in CH)
- Idea: can we jointly decode word alignment with NER?

Wang et al. ACL 2013.

Edge Factors in Joint WA and NER



Edge Factors in Joint WA and NER



Joint NER and Word Alignment

NEREN NERCH

$$\max_{\mathbf{y}^{\mathbf{e}(\mathbf{k})}\mathbf{y}^{\mathbf{f}(\mathbf{l})}\mathbf{y}^{\mathbf{e}(\mathbf{h})}\mathbf{y}^{\mathbf{f}(\mathbf{h})}\mathbf{B}^e\mathbf{B}^f\mathbf{A}} \quad f(\mathbf{y}^{\mathbf{e}^{(\mathbf{k})}}) + g(\mathbf{y}^{\mathbf{f}^{(\mathbf{l})}}) +$$

$$m(\mathbf{B}^e) + n(\mathbf{B}^f) + \boxed{\sum_{(i\in|\mathbf{e}|, j\in|\mathbf{f}|)} q_{(i,j)}(y_i^{e^h}, y_j^{f^{(h)}}, a(i,j))}$$

WAEN WACH Edge Factor

$$\ni \forall (i,j) : \left(b^e(i,j){=}a(i,j)\right) \wedge \left(b^f(i,j){=}a(i,j)\right)$$

$$\wedge \text{ if } a(i,j) = 1 \text{ then } \left(y_i^{e^{(k)}}{=}y_i^{e^{(h)}}\right) \wedge \left(y_i^{f^{(l)}}{=}y_i^{f^{(h)}}\right)$$

$$q_{(i,j)}\left(y_i^e, y_j^f, a(i,j)\right) =$$

$$\log(P(a(i,j))) + \log\left(P(y_i^e, y_j^f | a(i,j))^{P(a(i,j))}\right)$$

$$\log(P(y_i^e, y_j^f | a(i,j))) = \begin{cases} a(i,j){=}1 : \text{pmi}(y_i^e, y_j^f) \\ a(i,j){=}0 : \text{pmi}(y_i^e \perp\!\!\!\perp y_j^f) = 0 \end{cases}$$

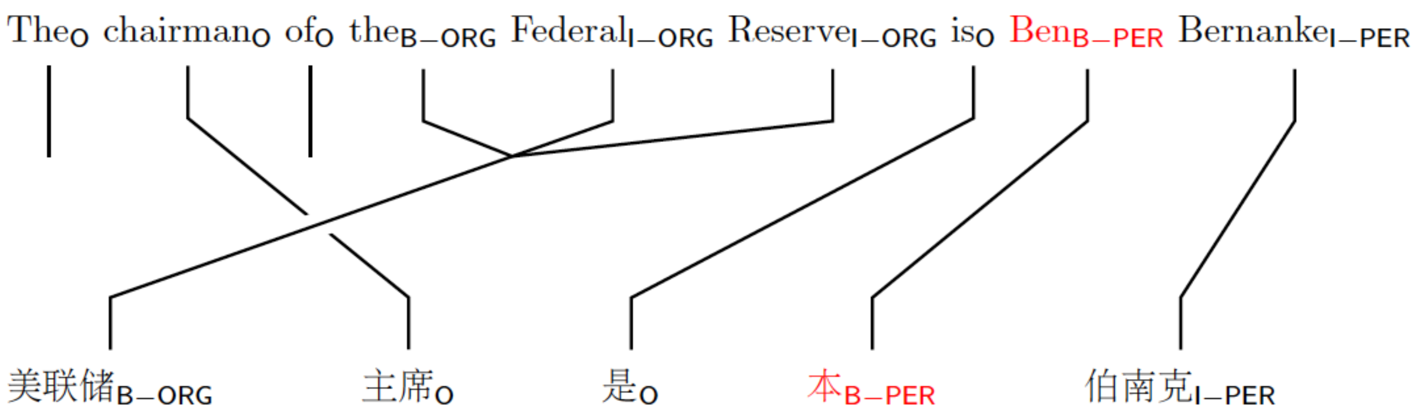
Joint NER and WA results

- Solve with DD (Dual Decomposition)

	NER-Chinese			NER-English			word alignment			
	P	R	F_1	P	R	F_1	P	R	F_1	AER
HMM-WA	-	-	-	-	-	-	**90.43**	40.95	56.38	43.62
Mono-CRF	82.50	66.58	73.69	84.24	78.70	81.38	-	-	-	-
Bi-NER	**84.87**	75.30	79.80	**84.47**	81.45	82.93	-	-	-	-
Bi-NER-WA	84.42	**76.34**	**80.18**	84.25	**82.20**	**83.21**	77.45	50.43	61.09	38.91
Bi-NER-WA+NC	84.25	75.09	79.41	84.28	82.17	**83.21**	76.67	**54.44**	**63.67**	**36.33**

A Solution to Lack of Training Data

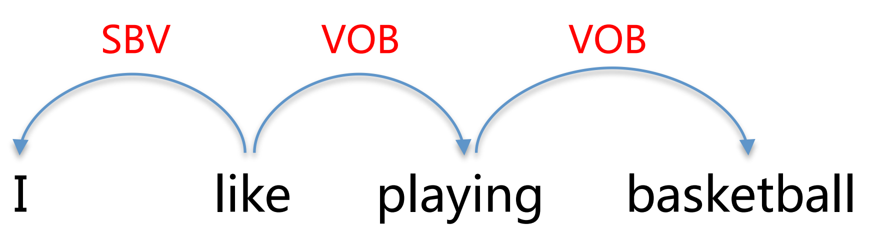
- Based on Cross-lingual



- Methods
  - Cross-lingual Annotation Projection
  - Joint Bilingual Modeling
  - **Cross-lingual Transfer (Dependency Parsing)**

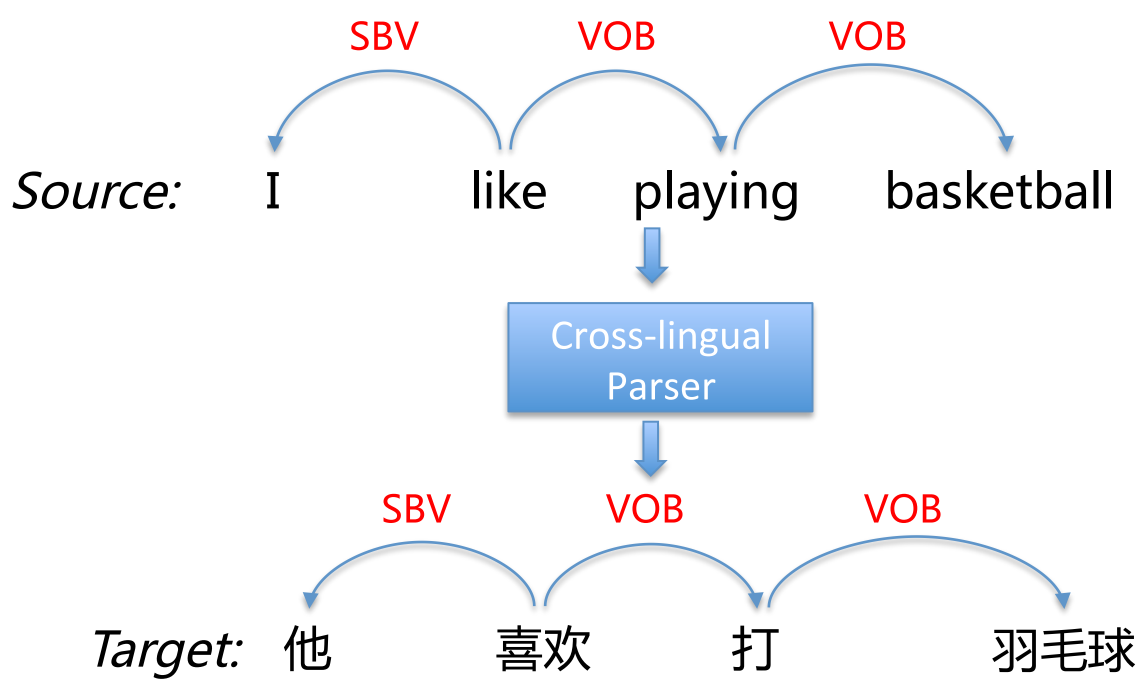
Dependency Parsing

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies



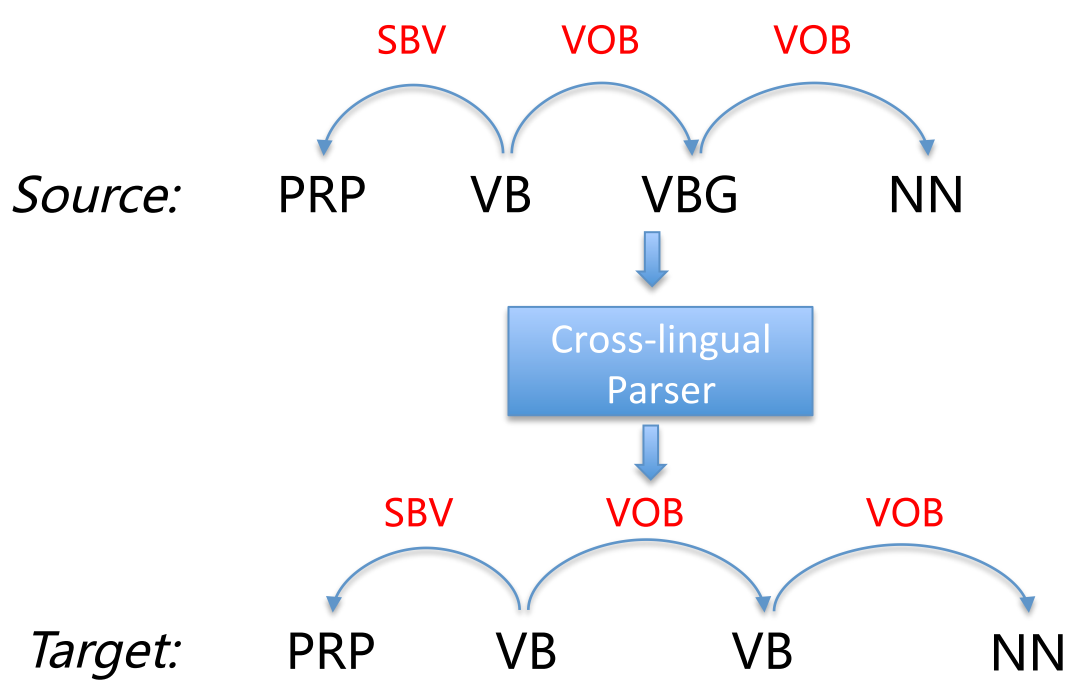
Cross-lingual Transfer DP

- No TreeBanks for low-resource (target) languages
- Transfer the parser of a rich-resource (source) language (e.g. English) to a low-resource language



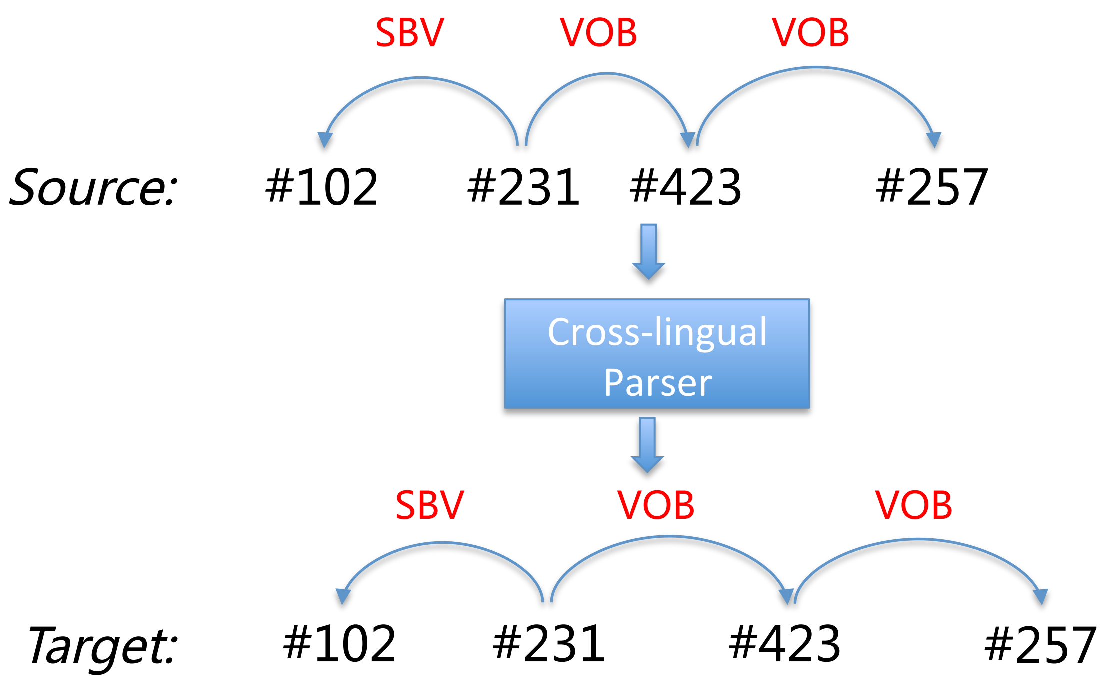
Previous Work

- Delexicalized Parser (McDonald et al. 2011)
  – Only use non-lexical features



Previous Work

- Cross-lingual Word Clustering (Tackstrom et al. 2012)
  - Coarse-grained word representation, which partially fills the *lexical feature gap*



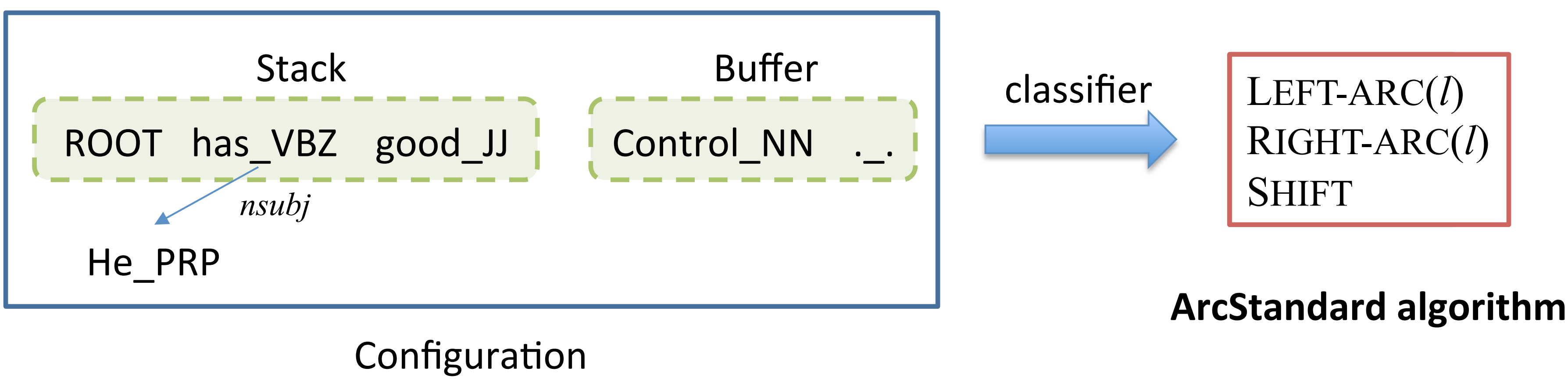
Our Motivation

- Bridging the *lexical feature gap* with distributed features representations (Embeddings)
  - Cross-lingual words
  - POS, Label histories
  - Word clusters

Transition-based Dependency Parsing

- Greedily predict a transition sequence from an initial parser state to some terminal states
- State (configuration)

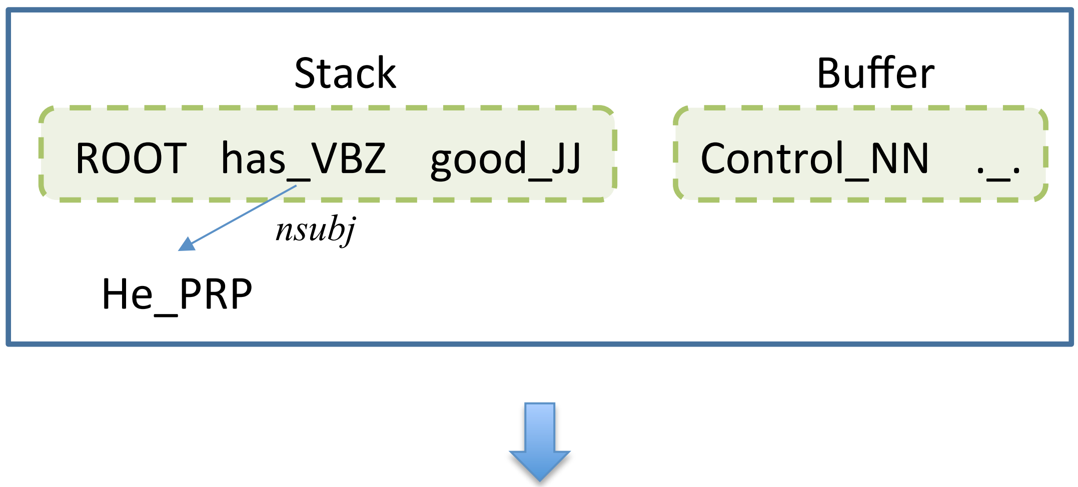
  = Stack + Buffer + Dependency Arcs



Neural Network Classifier

- Learn a **dense** and **compact** feature representation (Chen and Manning, 2014)

Configuration

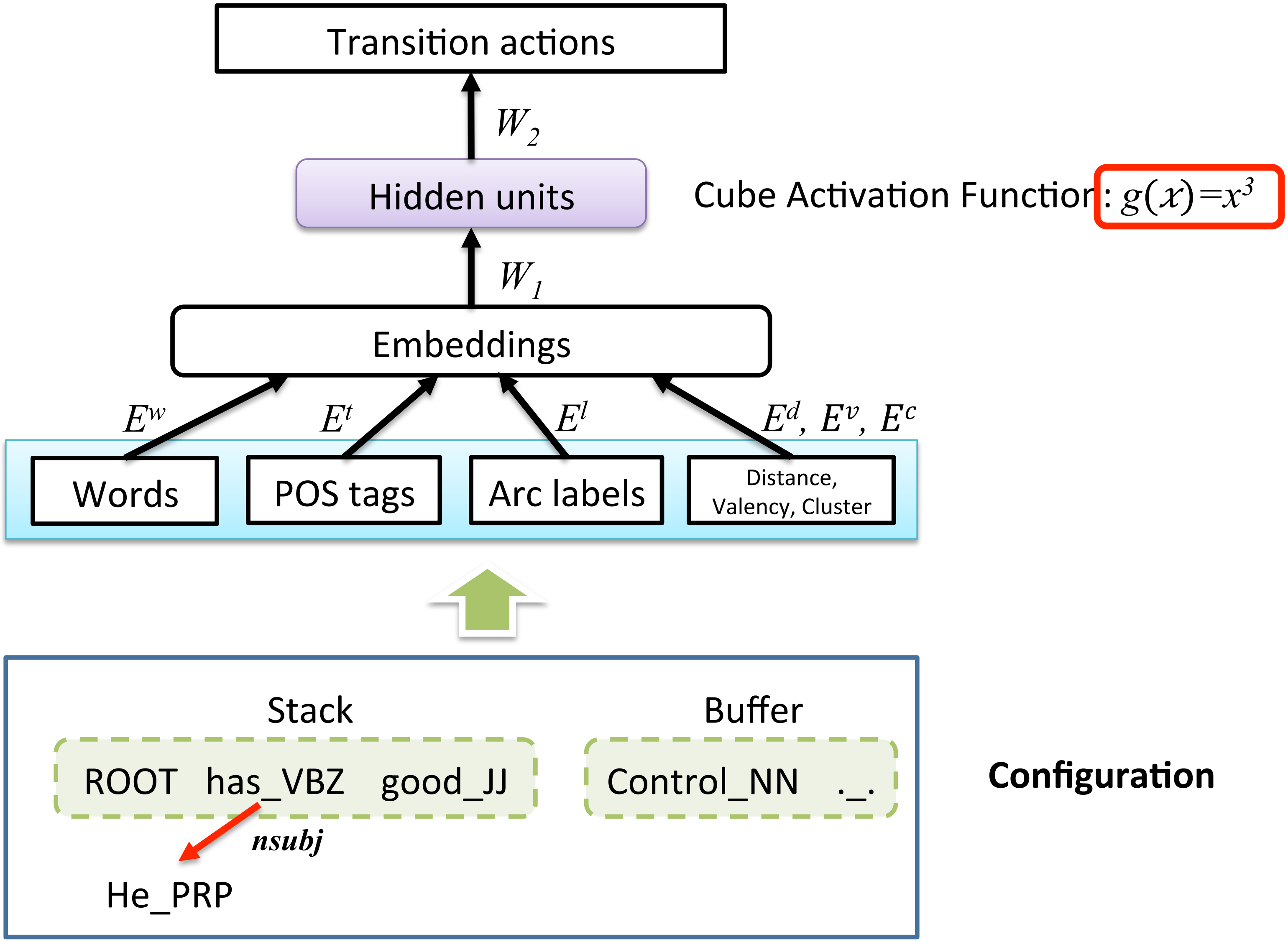


Feature Vector:

- Continuous
- Dense
- Low-dimensional

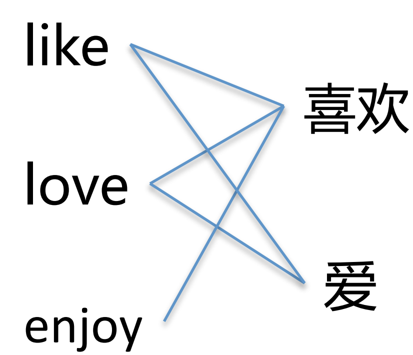
0.1	0.9	-0.3	1.2	0.2	...	-0.1	-0.6

Model Architecture



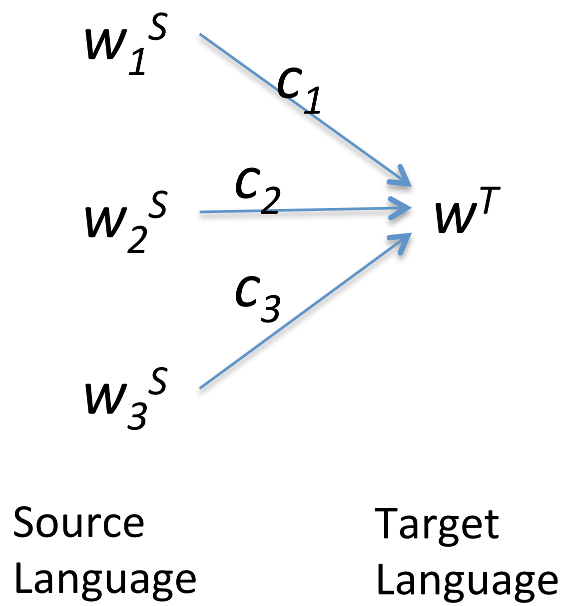
Cross-lingual Representation Transfer

- Non-lexical Features
  - POS, Label, Distance, …
  - One to one mapping: directly transfer
- Lexical Features
  - Word
  - Many to many mapping: ?



Robust Alignment-based Projection

- w_i^S aligns with w^T in c_i times



$$v(w^T) = \sum_i \frac{c_i}{|c|} v(w_i^S)$$

$$v(w^{OOV}) = \underset{w' \in C}{Avg}(v(w'))$$

$$C = \{w \mid EditDist(w^{OOV}, w) = 1\}$$

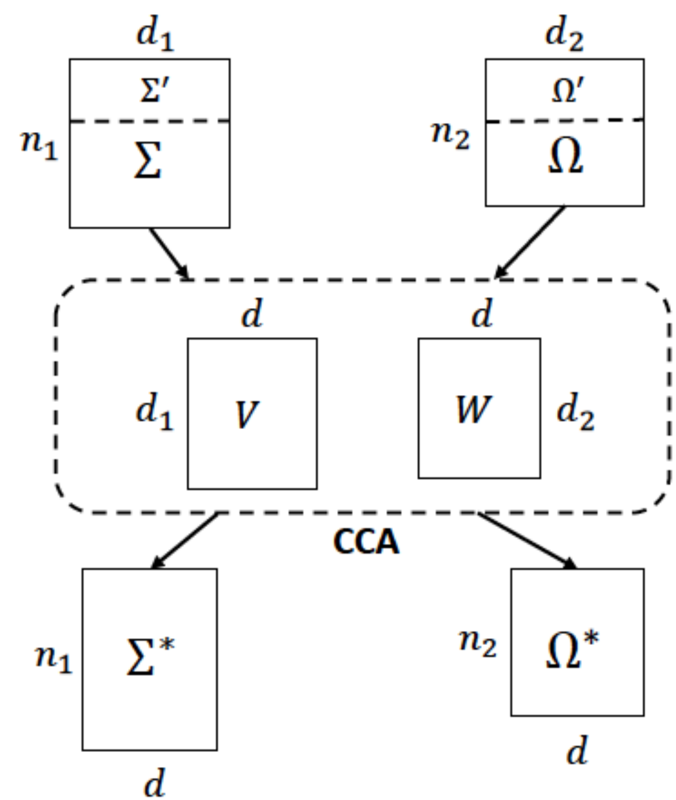
Canonical Correlation Analysis

- Canonical Correlation Analysis (CCA)
  - Measuring the linear relationship between multidimensional variables

$$V, W = CCA(\Sigma', \Omega')$$

$$\Sigma^* = \Sigma V, \quad \Omega^* = \Omega W$$

- Advantages
  - High word coverage
  - Encode the information of target language



Experiments

- Universal Dependency Treebanks v1 (Google)
  - Languages
    - Source: English (EN)
    - Target: German (DE), Spanish (ES), Franch (FR)
  - Universal Dependencies (42 relations)
  - Universal POS (12 tags)

Main Results

	Unlabeled Attachment Score (UAS)					Labeled Attachment Score (LAS)				
	EN	DE	ES	FR	AVG	EN	DE	ES	FR	AVG
Delexicalized	83.67	57.01	68.05	68.85	*64.64*	79.42	47.12	56.99	57.78	*53.96*
PROJ	91.96	60.07	71.42	71.36	*67.62*	90.48	49.94	61.76	61.55	*57.75*
PROJ+Cluster	92.33	60.35	**71.90**	**72.93**	***68.39***	90.91	**51.54**	**62.28**	**63.12**	***58.98***
CCA	90.62†	59.42	68.87	69.58	*65.96*	88.88†	49.32	59.65	59.50	*56.16*
CCA+Cluster	92.03†	**60.66**	71.33	70.87	*67.62*	90.49†	51.29	61.69	61.50	*58.16*
McD13	83.33	58.50	68.07	70.14	*65.57*	78.54	48.11	56.86	58.20	*54.39*
McD13*	84.44	57.30	68.15	69.91	*65.12*	80.30	47.34	57.12	58.80	*54.42*
McD13*+Cluster	90.21	60.55	70.43	72.01	*67.66*	88.28	50.20	60.96	61.96	*57.71*

Effect of Robust Projection

- Edit Distance for OOV words

		Simple	Robust	Δ
DE	coverage	91.37	94.70	+3.33
	UAS	59.74	60.35	+0.61
	LAS	50.84	51.54	+0.70
ES	coverage	94.51	96.67	+2.16
	UAS	70.97	71.90	+0.93
	LAS	61.34	62.28	+0.94
FR	coverage	90.83	97.60	+6.77
	UAS	71.17	72.93	+1.76
	LAS	61.72	63.12	+1.40

Effect of Fine-tuning Word Embedding

- **Projection** method over CCA lies in the fine-tuning of word embeddings while training the parser

		Fix	Fine-tune	Δ
DE	UAS	59.74	60.07	+0.33
	LAS	49.44	49.94	+0.50
ES	UAS	70.10	71.42	+1.32
	LAS	61.31	61.76	+0.45
FR	UAS	70.65	71.36	+0.71
	LAS	60.69	61.50	+0.81

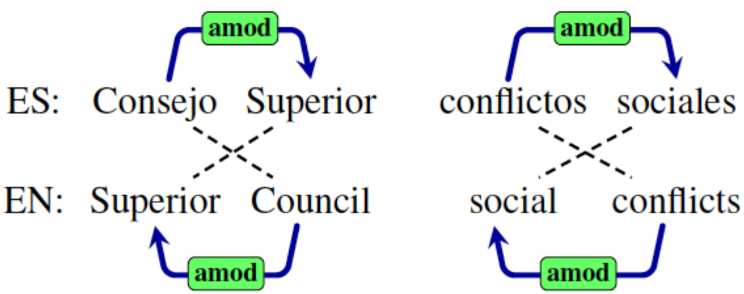
Target Minimal Supervision

- Cross-lingual approaches can only learn the **common** dependency structures shared between the source and target languages

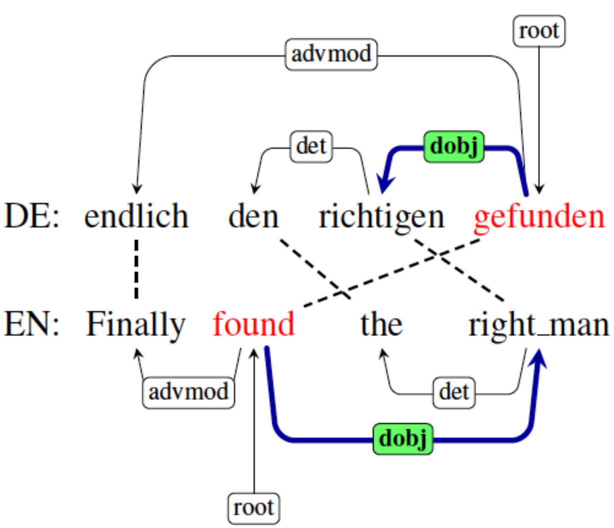
- For many languages, there are some **special** syntactic characteristics that are can only be learned from data in the target language

Target Minimal Supervision

- For example



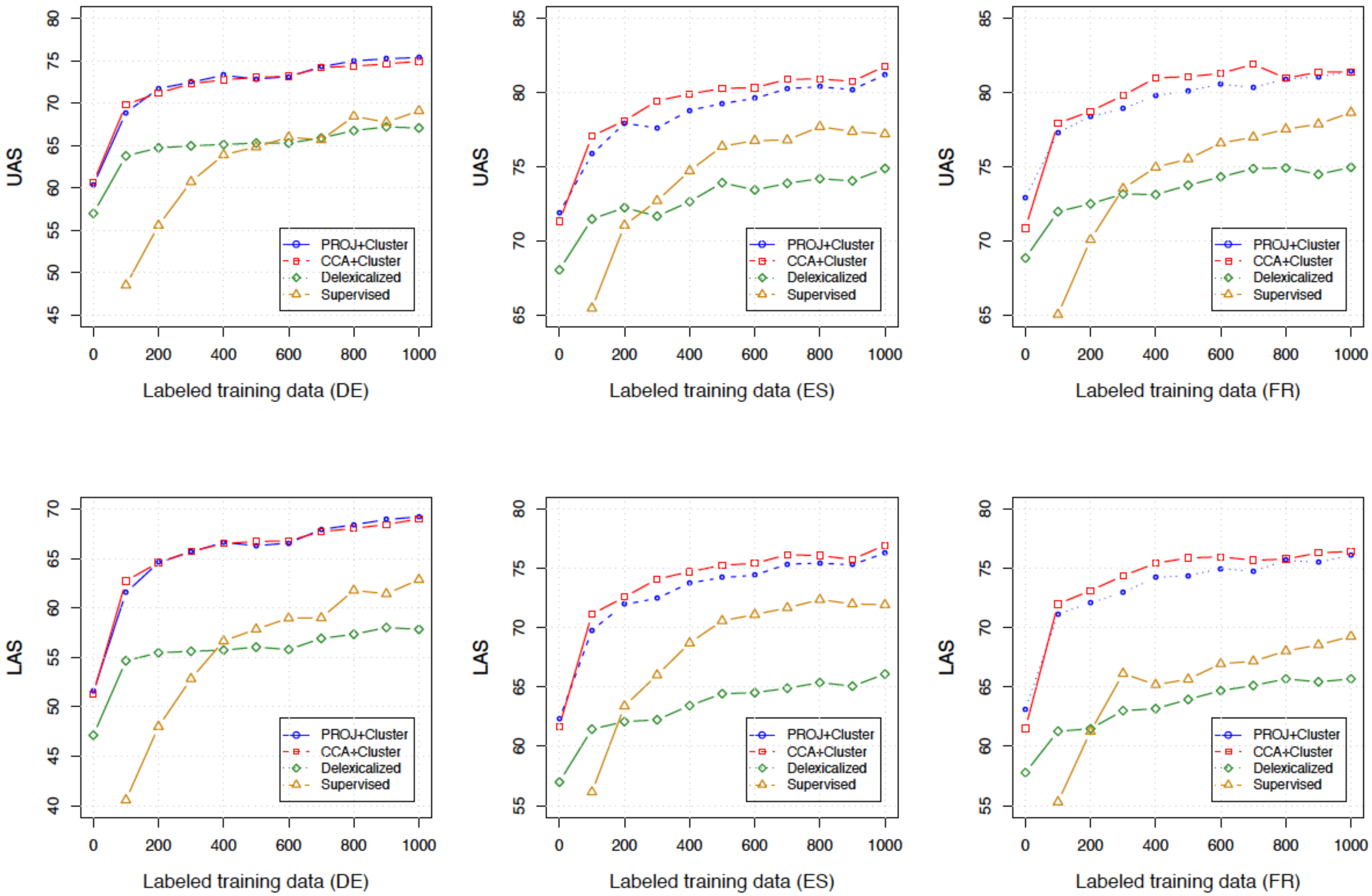
Relation: *amod*; Language: EN *vs.* ES, FR			
	$amod_{\curvearrowright}$	$amod_{\curvearrowleft}$	ratio
EN	1,667	57,864	1 : 34.7
ES	14,876	5,205	2.9 : 1
FR	12,919	4,910	2.6 : 1



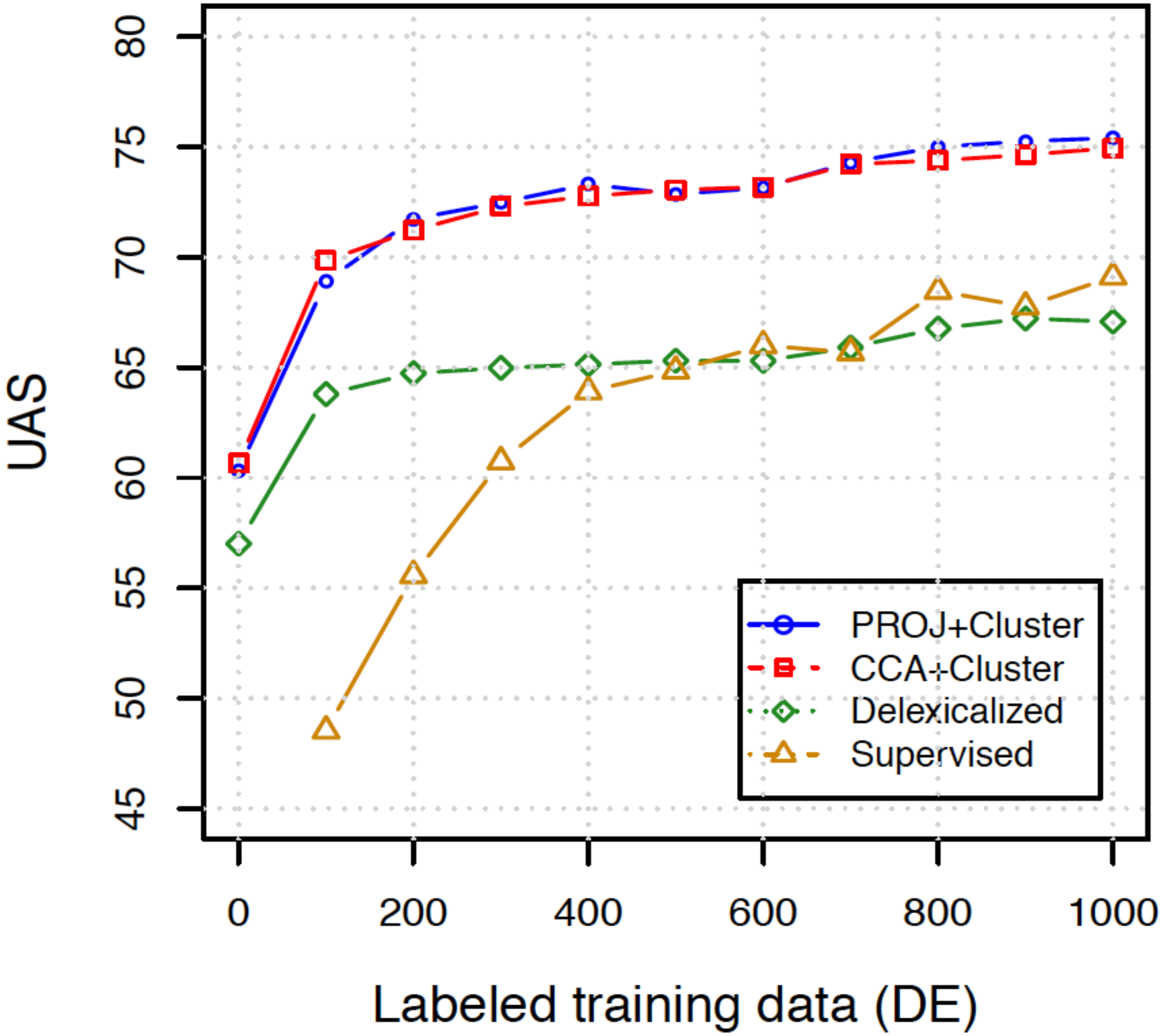
Relation: *dobj*; Language: EN *vs.* DE			
	$dobj_{\curvearrowright}$	$dobj_{\curvearrowleft}$	ratio
EN	38,395	764	50.3 : 1
DE	4,277	3,457	1.2 : 1

Results of Minimal Supervision

Solution: Use small labeled dependency trees (100 → 1000) from the target language to fine-tune the parsing model



Results Zoom-in



Effect of Minimal Supervision (100 sent)

- Case studies
  - *dobj* (EN *vs.* DE)
  - *amod* (EN *vs.* ES, FR)

Relation: *dobj*; Language: DE		
	P	R
PROJ+Cluster	41.45	31.09
+100	41.90	51.40
Δ	↑ 0.45	↑ **20.31**
CCA+Cluster	39.47	31.74
+100	43.59	57.57
Δ	↑ 4.12	↑ **25.83**

Relation: *amod*; Language: ES, FR				
	ES		FR	
	P	R	P	R
PROJ+Cluster	94.97	80.05	92.94	81.70
+100	91.60	92.52	93.61	95.75
Δ	↓ 3.37	↑ **12.47**	↑ 0.67	↑ **14.05**
CCA+Cluster	93.37	77.31	92.08	72.22
+100	91.85	92.77	92.77	96.41
Δ	↓ 1.52	↑ **15.46**	↑ 0.69	↑ **24.19**

Conclusion

- A novel cross-lingual dependency parsing based on distributed feature representation
- Two methods of cross-lingual word representations
  - Robust projection and CCA
- Achieve significant improvements by combining with word clusters
- Further boosted by minimal supervision from target language

Thanks Q&A

http://ir.hit.edu.cn/~car/