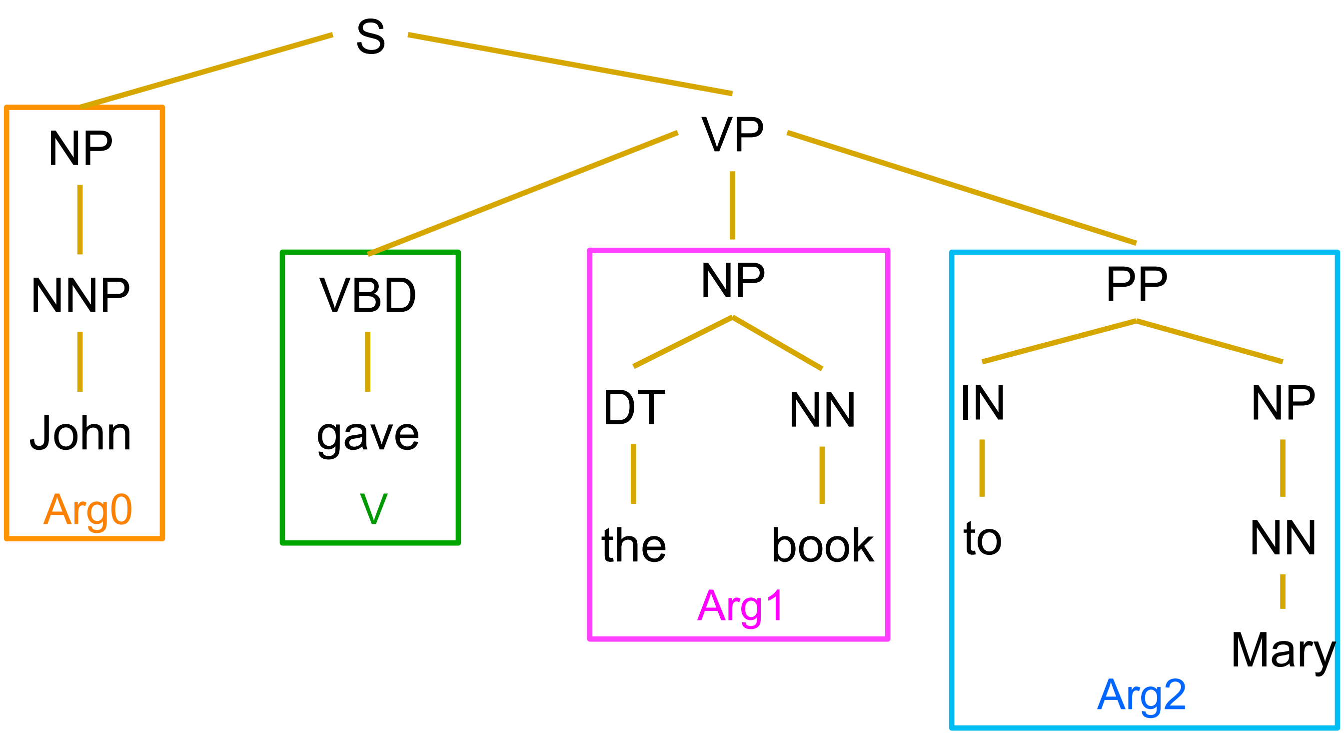
Improving Chinese-English PropBank Alignment

Shumin Wu, Martha Palmer
University of Colorado Boulder,
Hitachi Data Systems
June 4, 2015

Computational Language and EducAtion Research
C L E A R

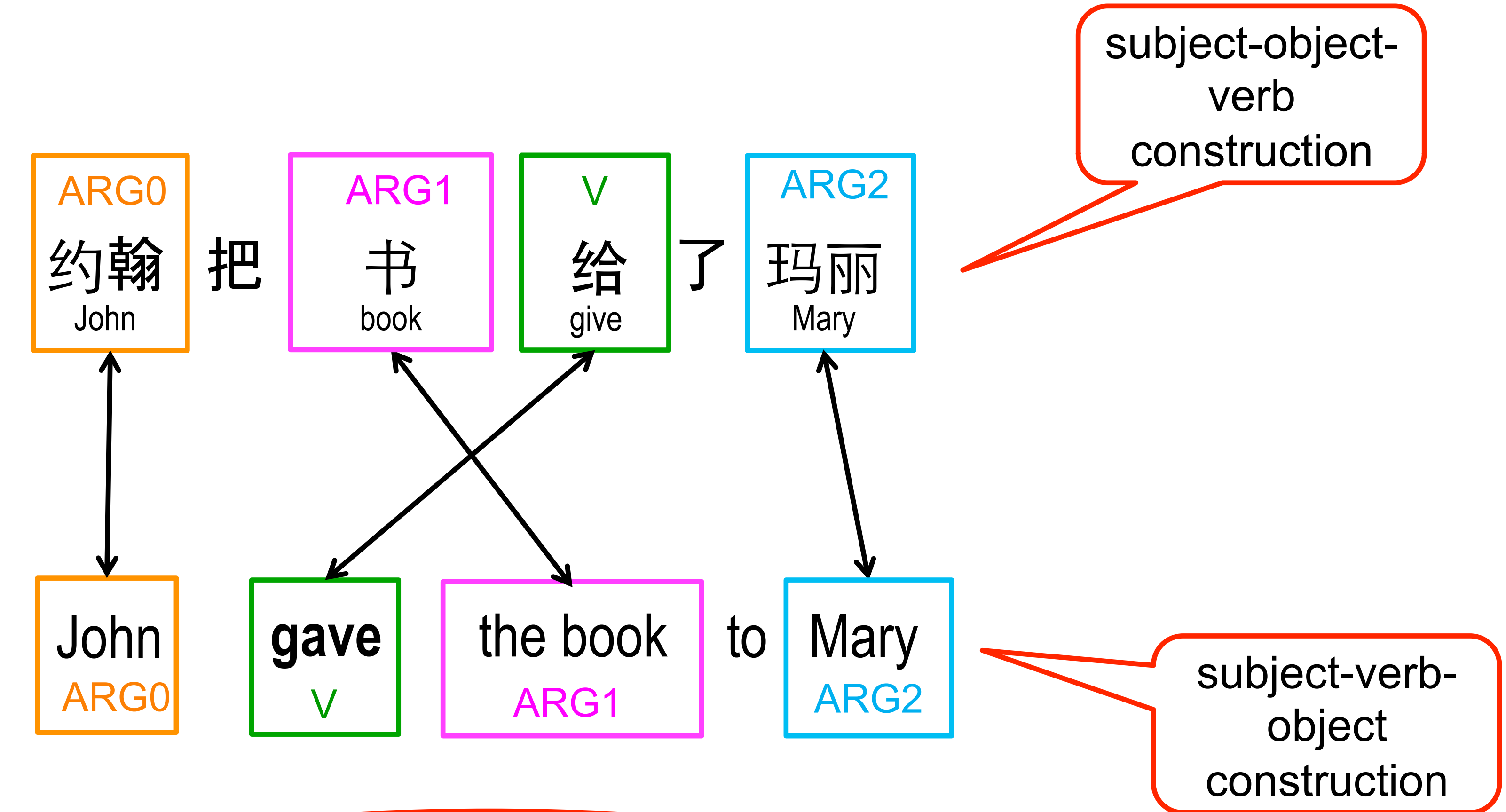
PropBank semantic annotation

- A corpus annotated with verbal (and more recently nominal and adjective) propositions and their arguments.
- Adds semantic information (semantic roles) to the phrase structures.
  - e.g. John gave the book to Mary



- Has a set of core (numbered) arguments and adjunct argument types
  - ARG0 is typically agent, ARG1 is typically patient or theme

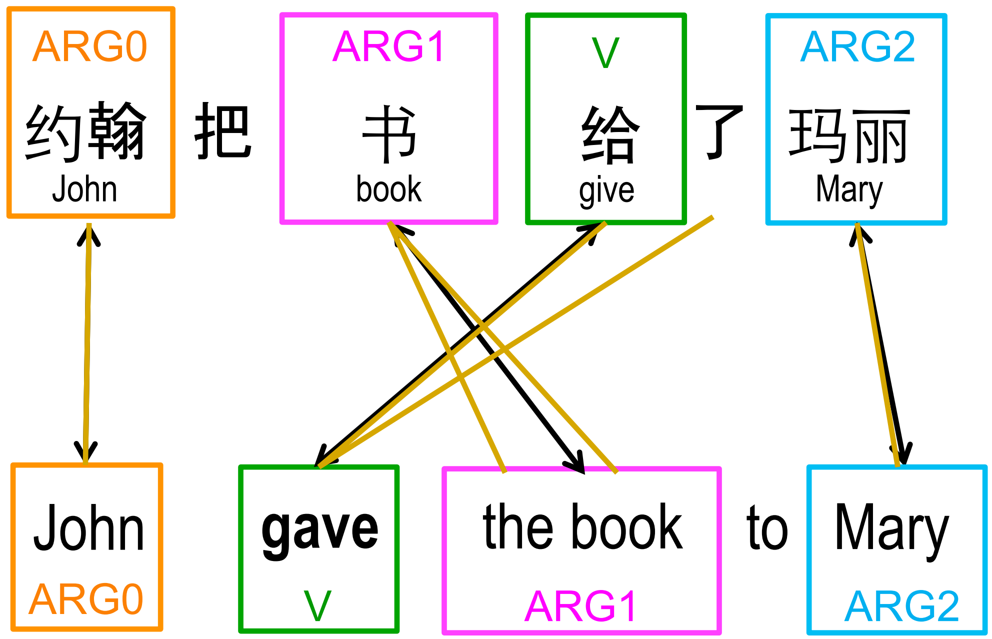
Why use cross-lingual PropBank alignment?



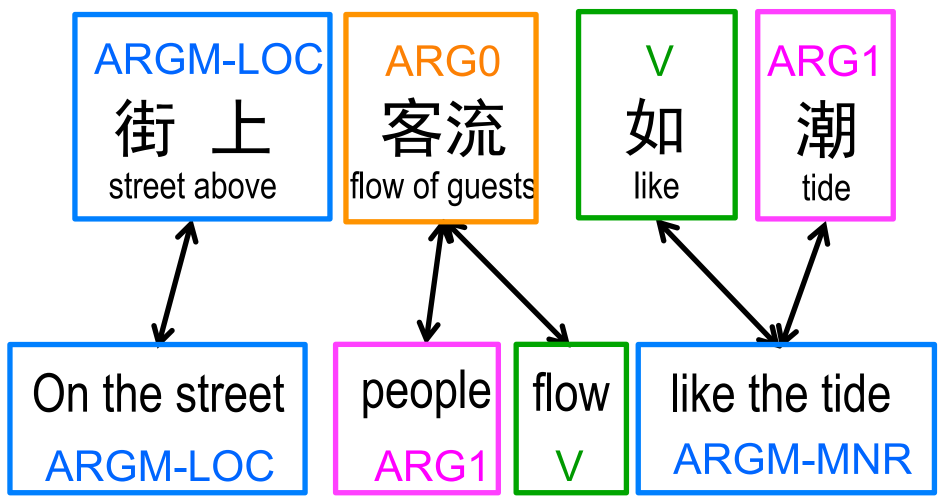
Potentially abstracts away syntactic differences

Framework for PropBank alignment

1. PropBank SRL
2. Word alignment
3. Argument alignment



More PropBank alignment example



Many-to-many argument alignment
Different argument labels can align to each other

Alignment by argument label type

label	**A0**	**A1**	**A2**	**A3**	**A4**	**V**
A0	***1610***	79	25	-	-	9
A1	432	***2665***	128	**11**	-	142
A2	43	*310*	**140**	8	3	67
A3	2	14	*21*	7	-	4
A4	1	*37*	9	3	6	4
V	25	*28*	22	1	-	3278

Chinese argument type (column) to English argument type (row) alignment on Triple-Gold Xinhua

Alignment by argument type

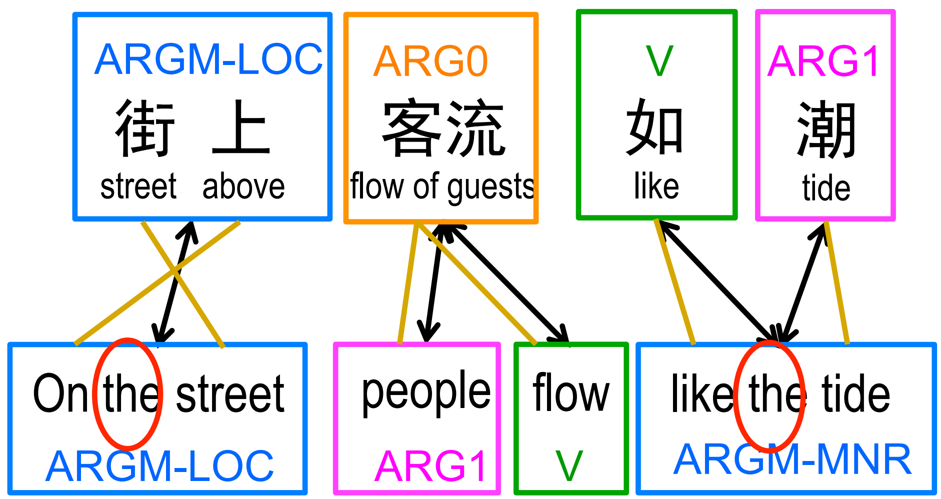
type	A0	A1	A2	A3	A4	ADV	BNF	DIR	DIS	EXT	LOC	MNR	PRP	TMP	TPC	V
A0	***1610***	79	25			28	1				8	5	1	11	1	9
A1	432	***2665***	128	**11**		83	**9**	**12**			29	12	5	21	**3**	142
A2	43	*310*	**140**	8	3	55	6	9		**2**	20	10	1	4	1	67
A3	2	14	*21*	7		2	4	2			1	2	1		1	4
A4	1	*37*	9	3	6						1		1			4
ADV	33	36	9	6		***307***	2	5	6		44	**121**	6	11	2	19
CAU	1					1							*16*			1
DIR	1	*13*	3	2		1		3			3					20
DIS	2					*69*			**40**		2	1	3	3		
EXT		4				*26*										2
LOC	23	65	13	1		3	1				***162***			5		4
MNR	9	9	5			*260*				1	3	34				25
MOD	1					*159*										84
NEG						*24*										5
PNC	3	23	11			1	6	1			1	2	***35***	2		8
PRD			*3*													1
TMP	14	21	2			235		3		1	8	16		***647***		6
V	25	28	22	1		211	1		1		2	12				***3278***

Chinese argument type (column) to English argument type alignment on Triple-Gold Xinhua

Many-to-many argument alignment

Objective:

choose source/target argument set to maximize total number of words in the set while minimizing number of unaligned words between arguments



100% precision, low recall

Many-to-many argument alignment

- Predicate-argument alignment score
  - a_i: argument, A_I: set of arguments
  - W_i: set of words in a_i
  - $map(a_i)$: word mapping (alignment) to set of words in target language
  - Source alignment precision/recall:

portion of word aligned words over all words in argument set

$$P_{I_C} = \frac{\left|(\bigcup_{i\in I} map_e(a_{i,c}))\bigcap(\bigcup_{j\in J} W_{j,e})\right|}{\left|(\bigcup_{i\in I} map_e(a_{i,c}))\right|}, R_{I_C} = \frac{\sum_{i\in I}\left|W_{i,c}\right|}{\sum_{\forall i}\left|W_{i,c}\right|}$$

portion of words in aligned argument set over words in all arguments

  - Choosing the argument set(s) to maximize alignment score:

Harmonic mean of precision and recall

$$A_{I,c} = \arg\max_{I} \frac{2\cdot P_{I,c}\cdot R_{I,c}}{P_{I,c}+R_{I,c}} = F_{I,c}$$

$$A_{I,c}, A_{J,e} = \arg\max_{I,J} \frac{2\cdot F_{I,c}\cdot F_{J,e}}{F_{I,c}+F_{J,e}} = F_{I,J}$$

F measure of source and target argument mapping score

Aligning multiple predicate-arguments

ARG0: 大 通道 建设 (big passage construction) | V: **搞活** (invigorated) | 了 | ARG1: 大 西南 的 物流、人流、信息流，(big southwest material flow people flow information flow)

V: **促进** (promote) | 了 | ARG1: 沿海 港口 城市 经济 的 发展。(coastal *port* city economy development)

Construction of the main passage (ARG0) has **activated** (V)

the flow of materials, the flow of people and the flow of information (ARG1) in the great southwest, (ARGM-LOC)

and has **promoted** (V) development (ARG1) in the coastal port cities' economies (ARGM-LOC)

Aligning multiple predicate-arguments

- Find the best one-to-one alignment (linear assignment problem) using Kuhn-Munkres method:

$$g^* = \arg\max_x \sum_{i \in C} \sum_{j \in E} F_{ij} \cdot x_{ij} \qquad x_{ij} \in \{0,1\}, \sum_i x_{ij} \le 1, \sum_j x_{ij} \le 1$$

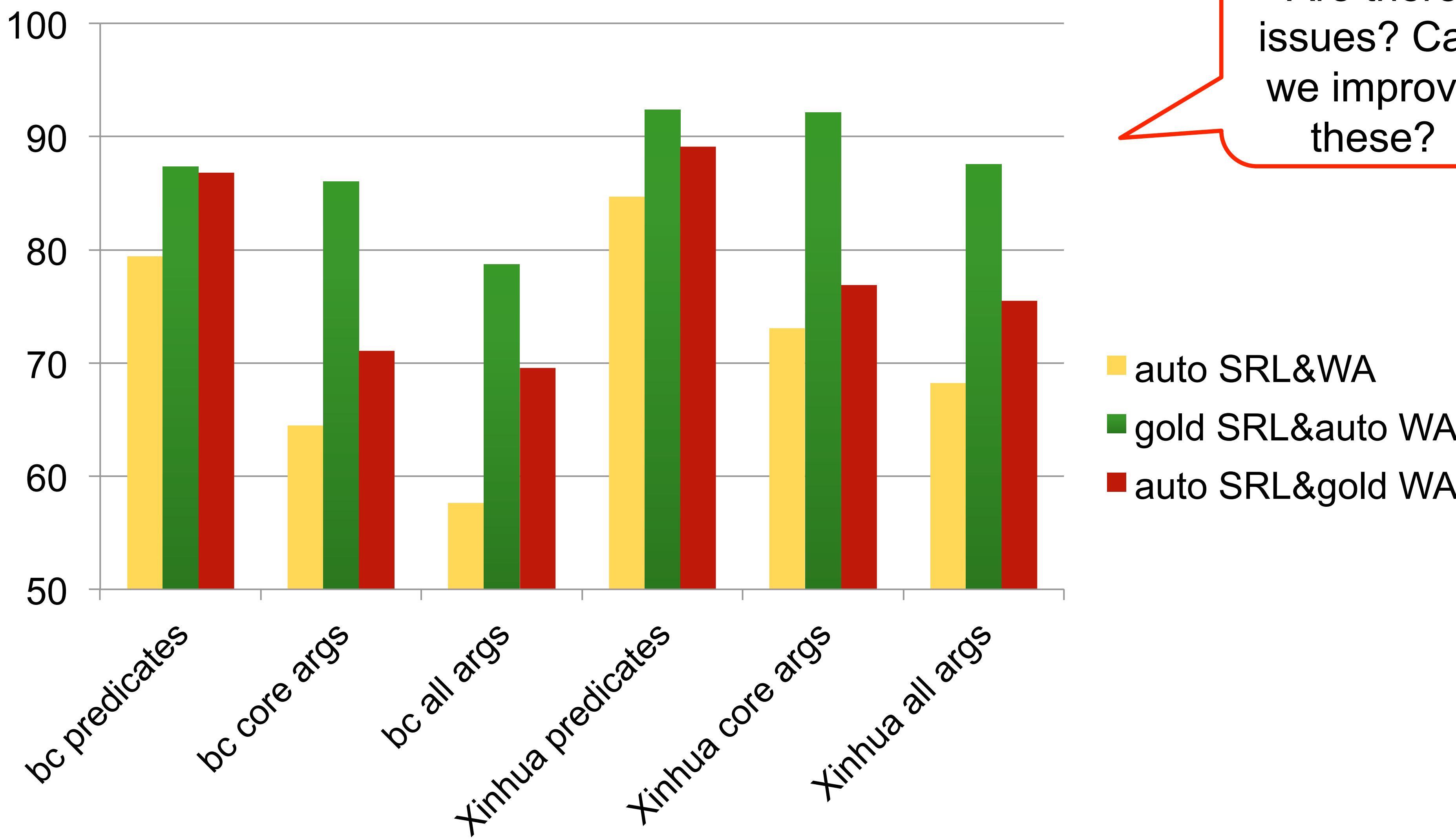
- Argument alignment score:

	activate.01	*promote*.01
搞活. 01 (invigorate)	**0.49**	0.25
促进. 01 (promote)	0.23	**0.77**

- Alignments:

搞活. 01 ↔ activate.01 促进. 01 ↔ promote.01

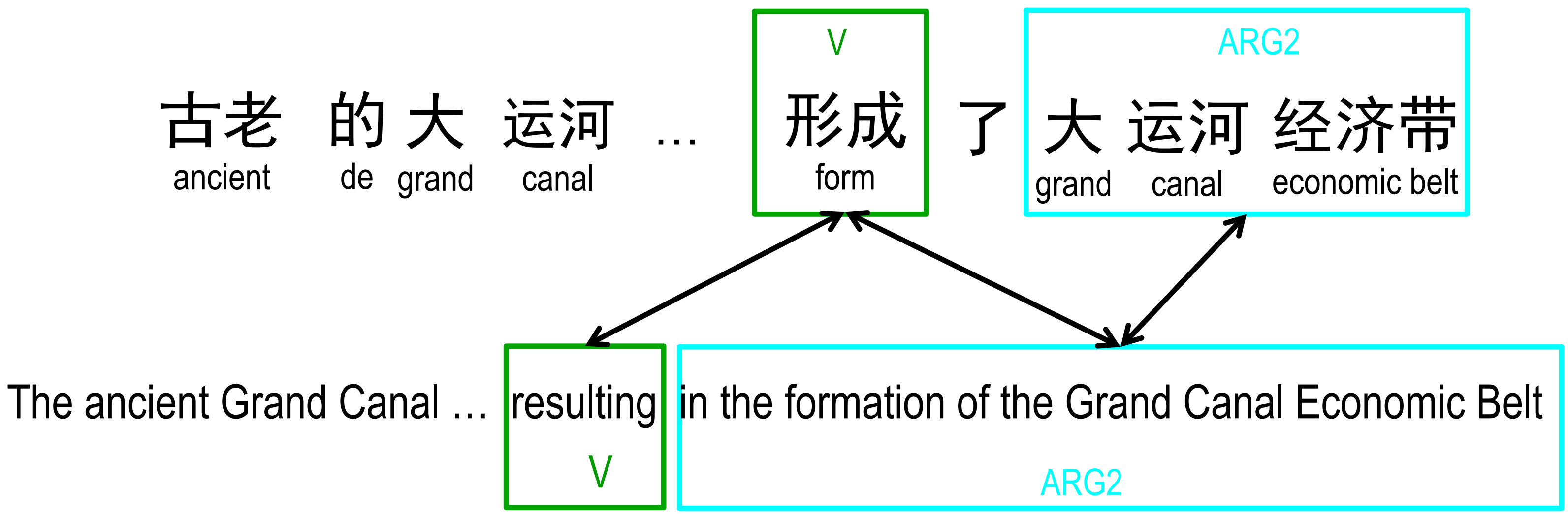
PropBank alignment results



Are there issues? Can we improve these?

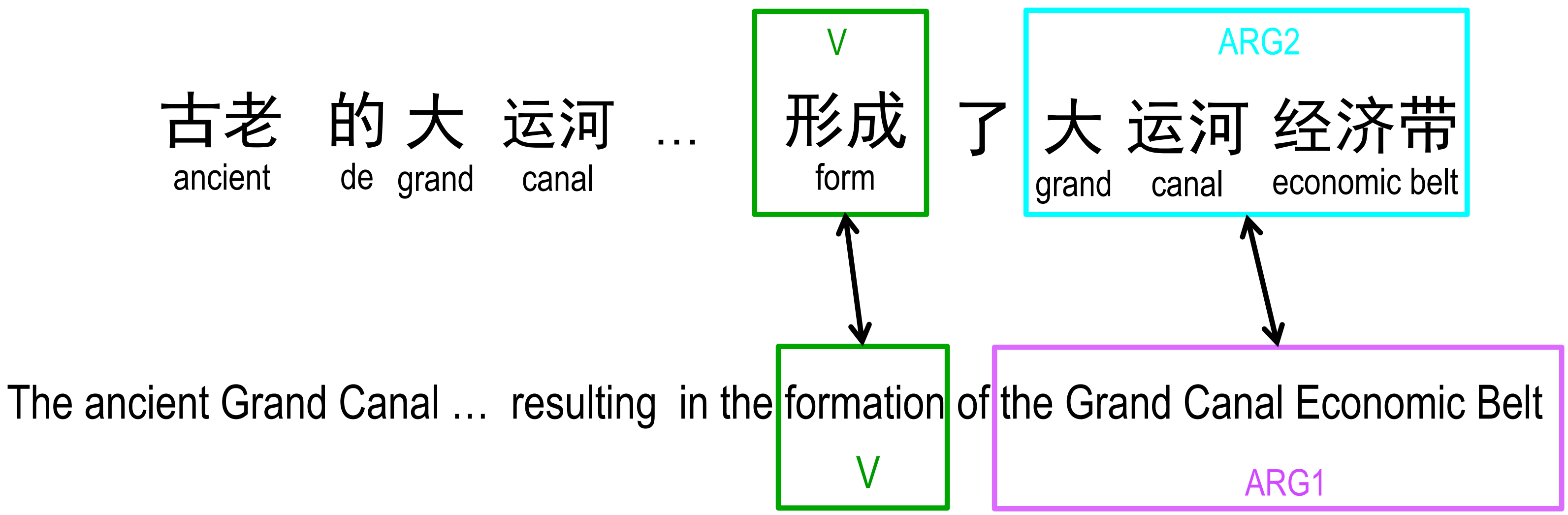
Semantic alignment results on broadcast conversation (bc) & Xinhua News

Forced alignment between verb predicates



Alignment of non-synonymous predicates, and predicate to core arguments

Add nominal predicates to alignment



Using nominal predicates can produce more semantically similar alignments

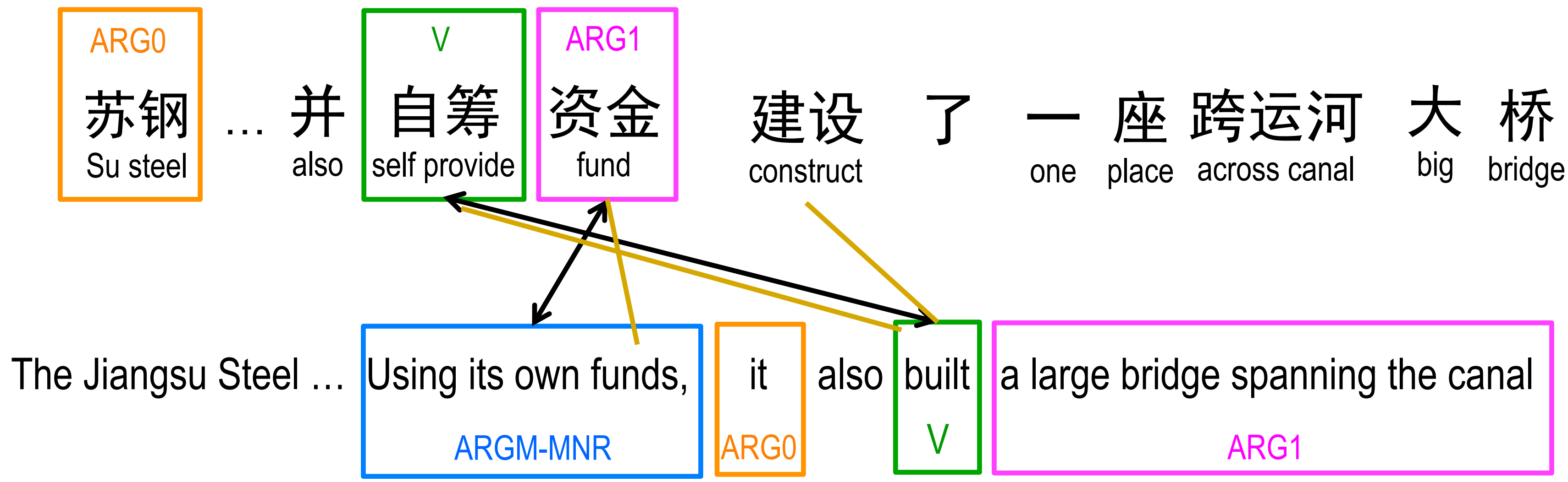
Alignment counts w/ nominal predicates

pred. type	V_c-V_e	N_c-V_e	V_c-N_e	N_c-N_e	total	Δ
verb only	4879	-	-	-	4879	-
+Ch nom.	4762	274	-	-	5036	+3.2%
+En nom.	4849	-	384	-	5233	+7.3%
all pred.	4759	239	314	760	6072	+24.5%

Predicate-argument mappings counts on Xinhua News with inclusion of nominal predicates

Using nominal predicates also increases alignment coverage

Alignment issue w/ automatic WA



	use.01	*build*.01
自筹. 01 (self provide)	0.50	**0.60**
建设. 01 (construct)	0.22	0.50

But 自筹⇔*build*, Arg1⇔AM-MNR
are very unlikely!!

Building alignment probability

- Collect predicate-to-predicate & argument-to-argument mapping counts from corpus to build:

$$p(pred_{j,e} \mid pred_{i,c})$$

$$p(a_{l,e} \mid a_{k,c}, pred_{i,c}, pred_{j,e})$$

- Counts can be really sparse, so apply smoothing
  - Simple Good-Turing smoothing for predicate-to-predicate probability

$$p_0 = \frac{N_1}{N}$$

$$p_c = (c+1)\frac{N_{c+1}}{N_c N}$$

Building alignment probability (cont)

- Smooth argument-to-argument probability with absolute discounting

$$p(a_{l_E} | a_{k_C}, pred_{i_C}, pred_{j_E}) = \frac{\max(freq(a_{l_E} | a_{k_C}, pred_{i_C}, pred_{j_E}) - d, 0)}{\sum_l (a_{l_E} | a_{k_C}, pred_{i_C}, pred_{j_E})}$$

$$+ (1-\lambda) \cdot p_{backoff}(a_{l_E})$$

$$p_{backoff}(a_{l_E}) = p(a_{l_E} | a_{k_C}, pred_{i_C}) \quad \text{or} \quad p(a_{l_E} | a_{k_C}, pred_{j_E})$$

- Intermediary back-off probabilities can also be smoothed:

$$p(a_{l_E} | a_{k_C}, pred_{i_C}) = \frac{\max(freq(a_{l_E} | a_{k_C}, pred_{i_C}) - d, 0)}{\sum_l (a_{l_E} | a_{k_C}, pred_{i_C})} + (1-\lambda) \cdot p(a_{l_E} | a_{k_C})$$

Alignment w/ probability model

- Update previously defined argument precision & recall

$$P'_{kl} = (1-\beta+\beta\cdot w(a_{l,e} \mid pred_{i,c}, pred_{j,e}, a_{k,c}))P_{kl}$$

$$R'_{kl} = (1-\beta+\beta\cdot w(a_{k,c} \mid pred_{i,c}, pred_{j,e}, a_{l,e}))R_{kl}$$

Portion (β) of alignment score weighted by argument label alignment probability

where $w(a_k) = \dfrac{p(a_k)}{\sum_k p(a_k)\cdot p(a_k)}$, $P_{kl} = \dfrac{|map_e(a_{k,c}) \cup W_{l,e}|}{|map_e(a_{k,c})|}$, $R_{kl} = \dfrac{|map_c(a_{l,e}) \cup W_{k,c}|}{|map_c(a_{l,e})|}$,

- Update predicate-to-predicate score:

$$F'_{i,c} = (1-\alpha+\alpha\cdot w(pred_{j,e} \mid pred_{i,c}))F_{i,c}$$

$$F'_{j,e} = (1-\alpha+\alpha\cdot w(pred_{i,c} \mid pred_{j,e}))F_{j,e}$$

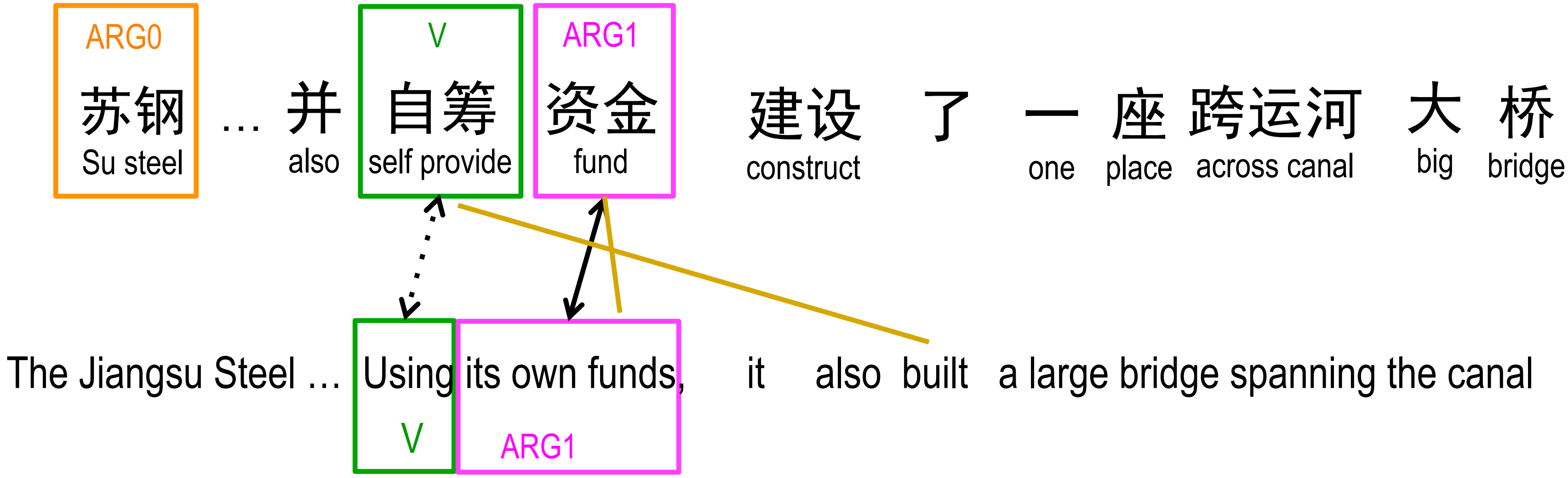
Portion (α) of alignment score weighted by predicate alignment probability

- Need to find "good" *α* and *β* values that balances the word alignment score and predicate/argument alignment probabilities

Alignment w/ probability model (cont)

- Find the best α and β using expectation maximization (EM)
  1. Perform semantic alignment on corpus
  2. Build probability models from mapping (E step)
  3. Find α and β that maximizes the sum of the mapping score of all predicate pairs in the corpus using grid search (M step)
  4. Repeat steps 2-3 until the score sum plateaus
- Data
  - 1.6M parallel sentence pairs from multiple LDC corpora
  - Stanford Chinese word segmenter, Berkeley parser, ClearSRL
- Need only 2 iterations to converge as the mapping output doesn't change drastically
- Optimal $\alpha=0.15$, $\beta=0.1$
  - α value (predicate-to-predicate probability) has a larger affect on score sum than β value (argument-to-argument probability)

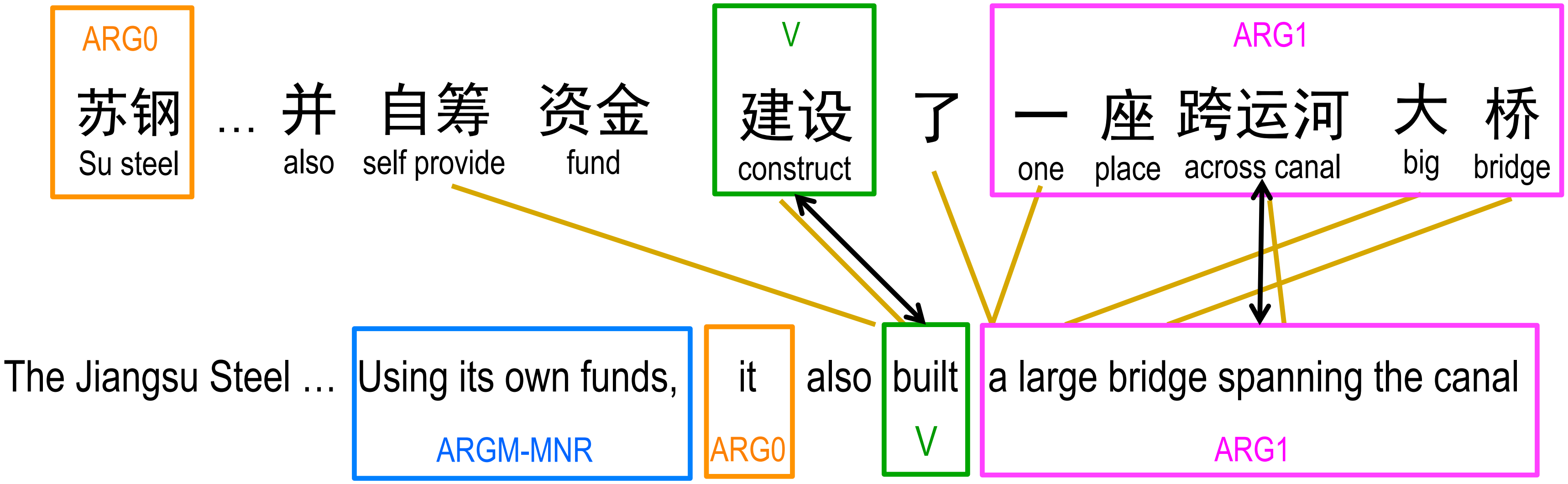
Alignment w/probability example



	use.01	*build*.01
自筹. 01 (self provide)	**0.65**	0.40
建设. 01 (construct)	0.10	**0.85**

自筹⇔*use*, Arg1⇔Arg1
are much more likely

Alignment w/probability example (cont)

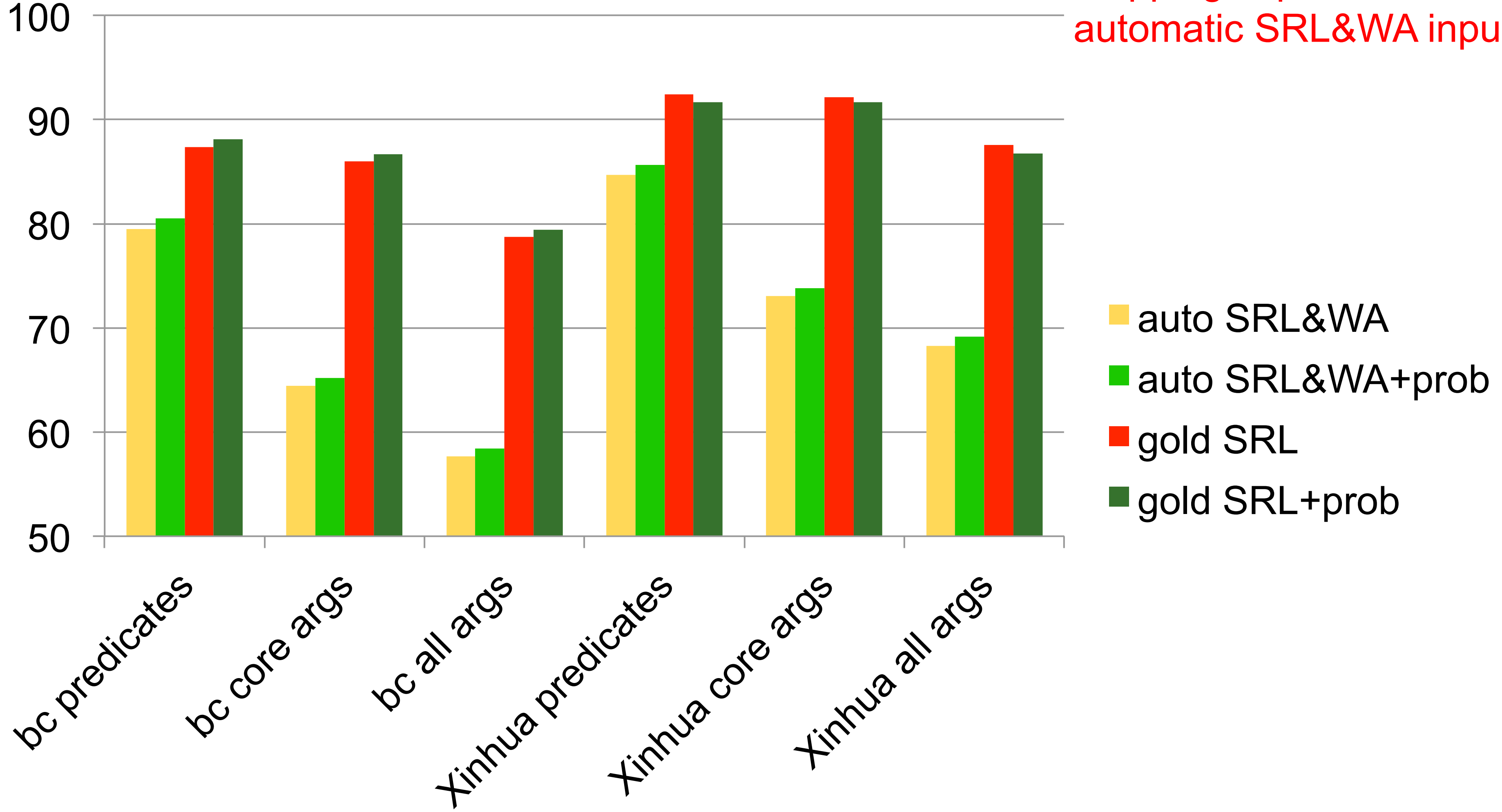


	use.01	*build*.01
自筹. 01 (self provide)	**0.65**	0.40
建设. 01 (construct)	0.10	**0.85**

Also allowed correct
建设⇔*build* mapping

Probability model improvements

- About 1 F point predicate mapping improvement on automatic SRL&WA input



Semantic alignment results on broadcast conversation (BC) & Xinhua News

Summary

- PropBank alignment with nominal predicates
  - Improves alignment quality and coverage
- Predicate-argument alignment with probability models
  - Generate predicate-to-predicate and argument-to-argument alignment probabilities on large corpora
  - Iteratively improve model with EM
  - Provides ~1 F-point improvement w/ automatic SRL input

Future work

- Build alignment probability models based on verb classes

$$p(pred_{j,c} \mid pred_{V_i,e})$$

- Use alignment probability model to perform joint Chinese, English SRL
  - Can improve both SRL accuracy and alignment model
- PropBank alignment of other language pairs
  - Arabic-English

Questions?