Joint Decoding with Multiple Translation Models

Yang Liu, Haitao Mi, Yang Feng, and Qun Liu

Institute of Computing Technology,

Chinese Academy of Sciences

{yliu,htmi,fengyang,liuqun}@ict.ac.cn

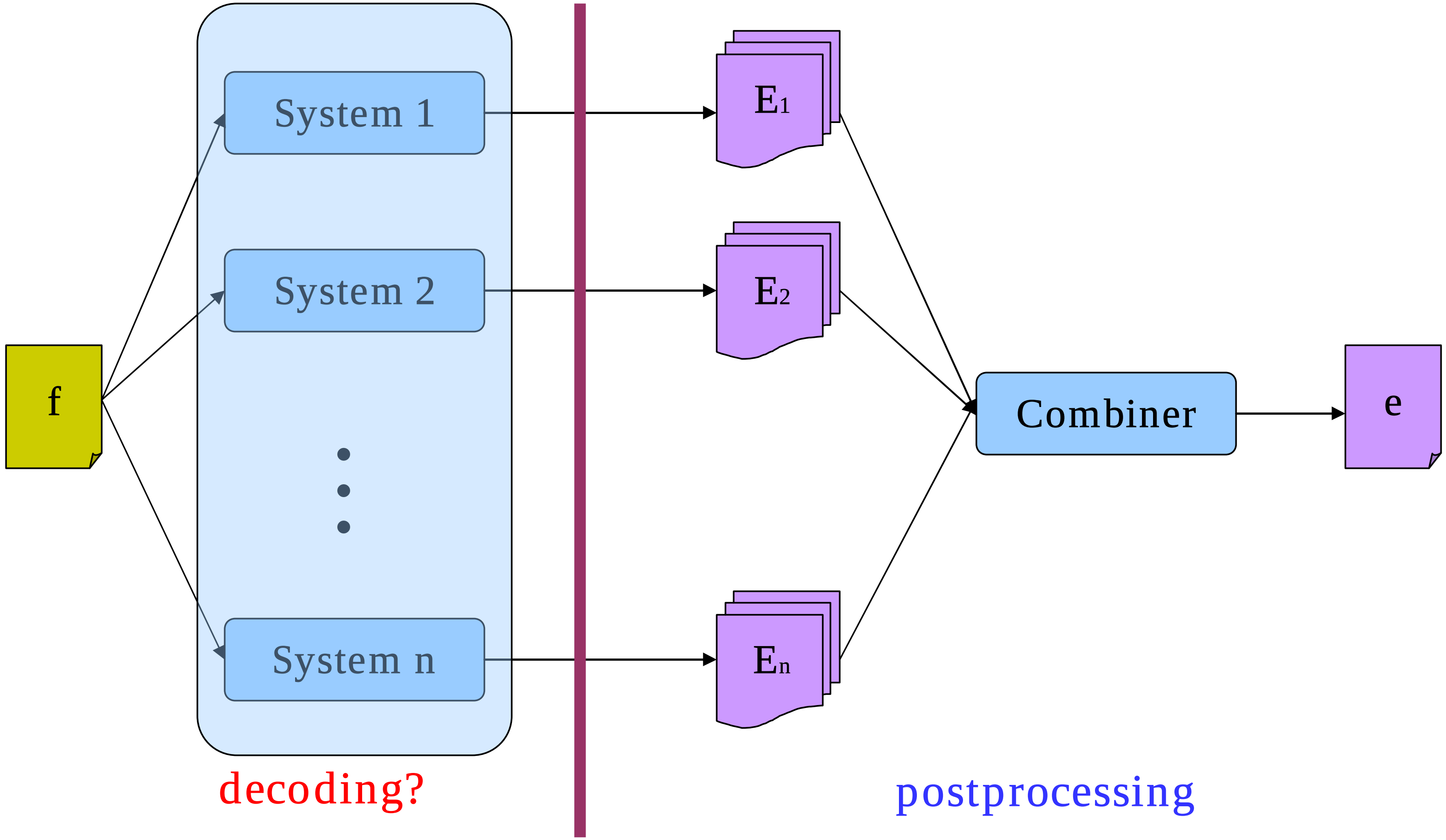
ICT

中科院计算所

INSTITUTE OF COMPUTING TECHNOLOGY

System Combination

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY



decoding?

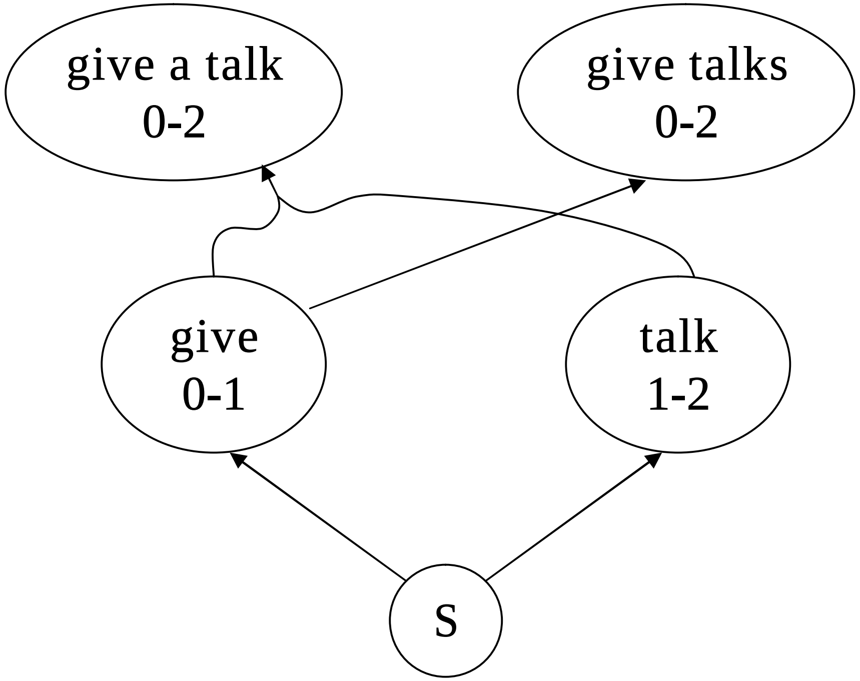
postprocessing

This Work

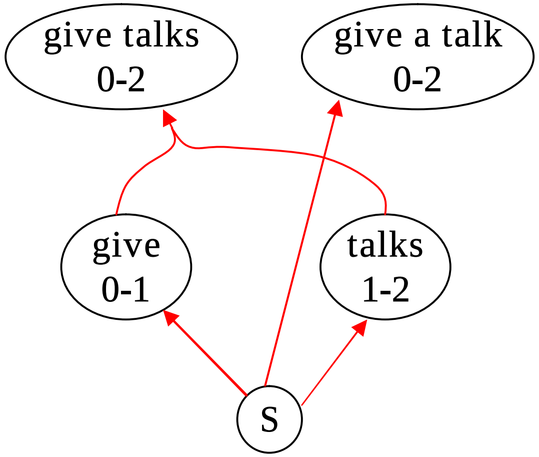
- We propose a technique called joint decoding to combine different models directly in the decoding phase.
- Our preliminary work shows promising results.

Translation Hypergraph

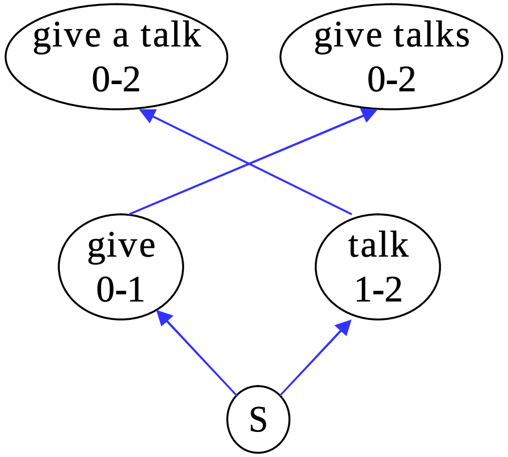
$_0$ fabiao $_1$ yanjiang $_2$



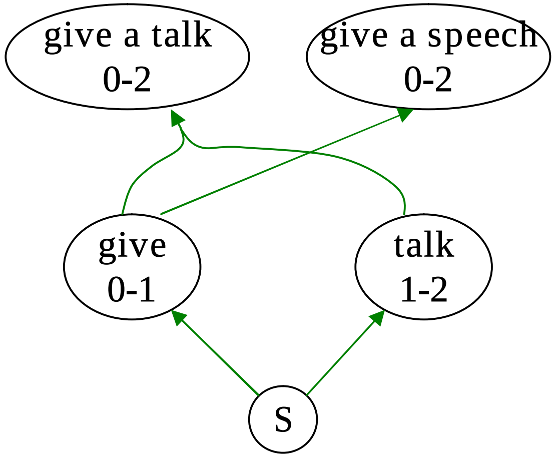
Consensus Translations



phrase-based

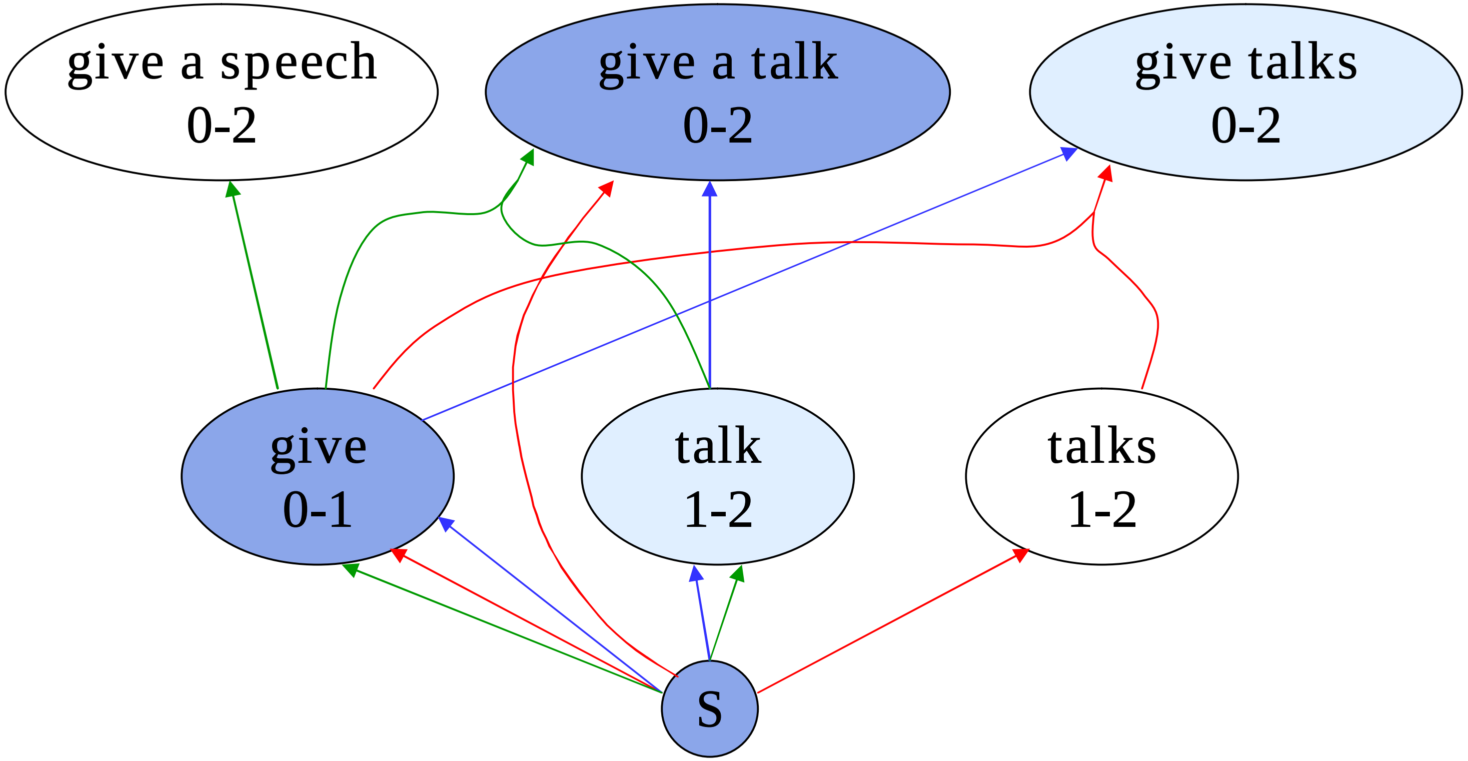


hierarchical phrase-based

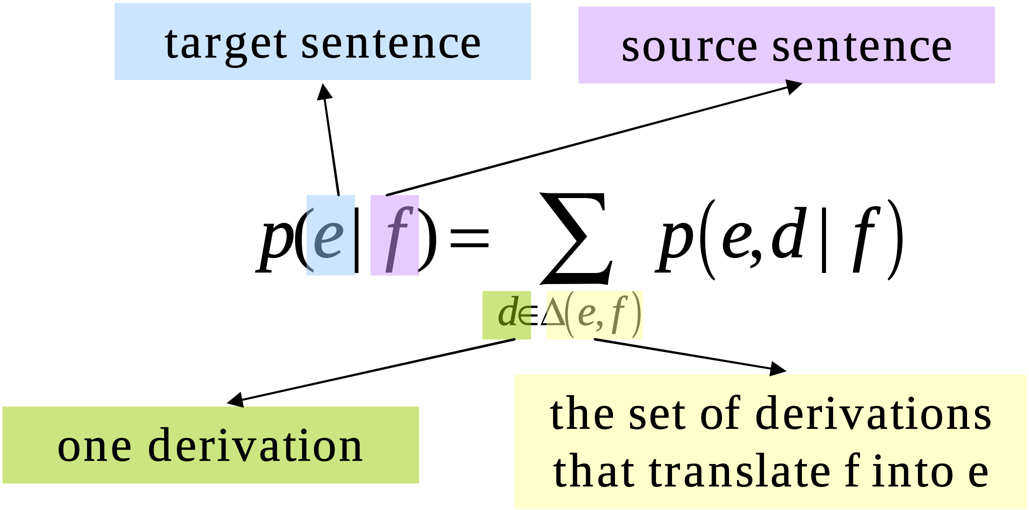


tree-to-string

Sharing Nodes



Scoring a Translation



target sentence

source sentence

$$p(e|f) = \sum_{d\in\Delta(e,f)} p(e,d|f)$$

one derivation

the set of derivations that translate f into e

a derivation can come from any model!

MDD and MTD

$$p(e \mid f) = \sum_{d \in \Delta(e,f)} \frac{\exp\left(\sum_i l_i h_i(d,e,f) \right)}{Z}$$

Blunsom et al. (2008)

$$\hat{e} = \arg\max_e \left\{ \sum_{d \in \Delta(e,f)} \exp\left(\sum_i l_i h_i(d,e,f) \right) \right\}$$

max-translation decoding

$$\approx \arg\max_{e,d} \left\{ \sum_i l_i h_i(d,e,f) \right\}$$

max-derivation decoding

An Example of MTD

model	feature		
	name	weight	value
hier	p(e|f)	1.0	0.1
	p(f|e)	1.0	0.2
	l(e|f)	1.0	0.1
	l(f|e)	1.0	0.3
	rc	1.0	3
t2s	p(e|f)	1.0	0.2
	p(f|e)	1.0	0.1
	l(e|f)	1.0	0.3
	l(f|e)	1.0	0.2
	rc	1.0	4
const	lm	1.0	0.7
	wc	1.0	3

exp(3.7)

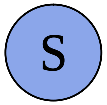
exp(4.8)

exp(3.7)

(exp(3.7) + exp(4.8)) × exp(3.7)

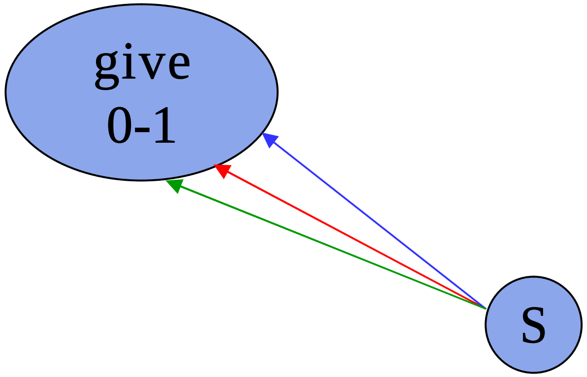
Joint Decoding

$_0$ fabiao $_1$ yanjiang $_2$



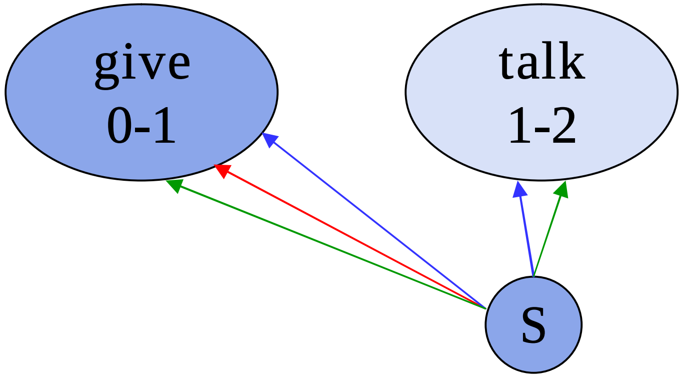
Joint Decoding

$_0$ fabiao $_1$ yanjiang $_2$



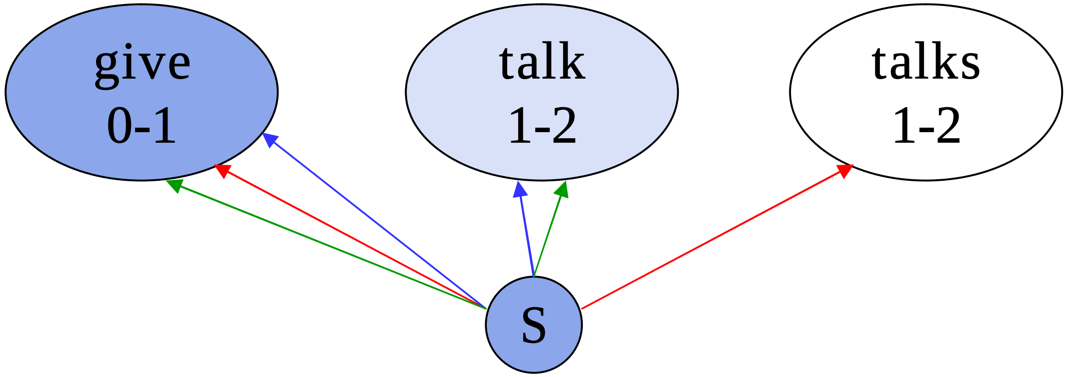
Joint Decoding

$_0$ fabiao $_1$ yanjiang $_2$



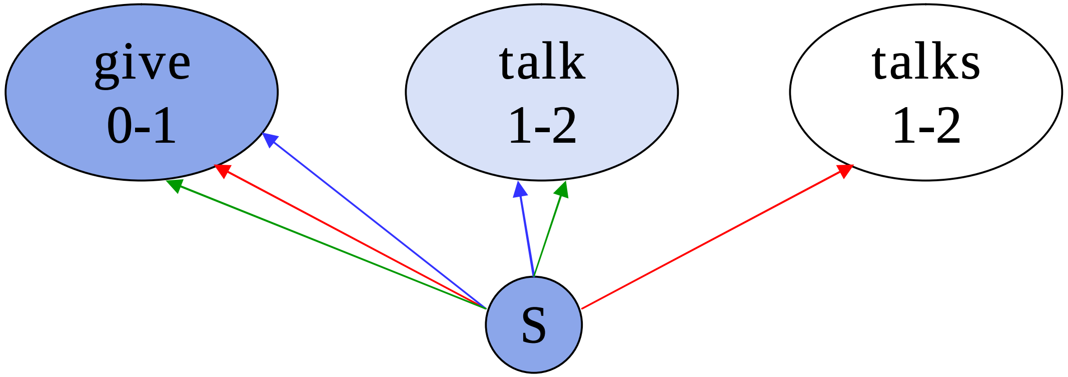
Joint Decoding

$_0$ fabiao $_1$ yanjiang $_2$



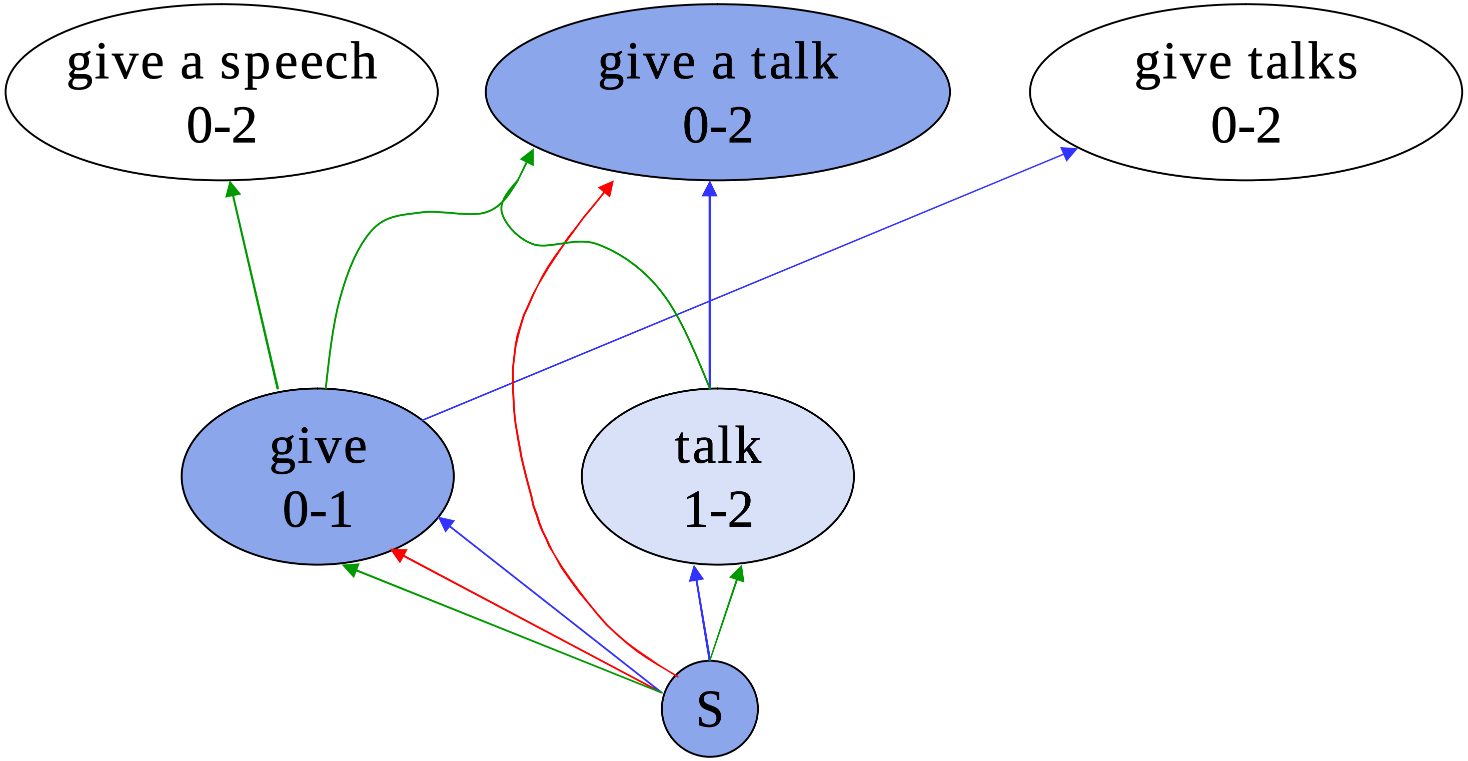
Joint Decoding

$_0$ fabiao $_1$ yanjiang $_2$



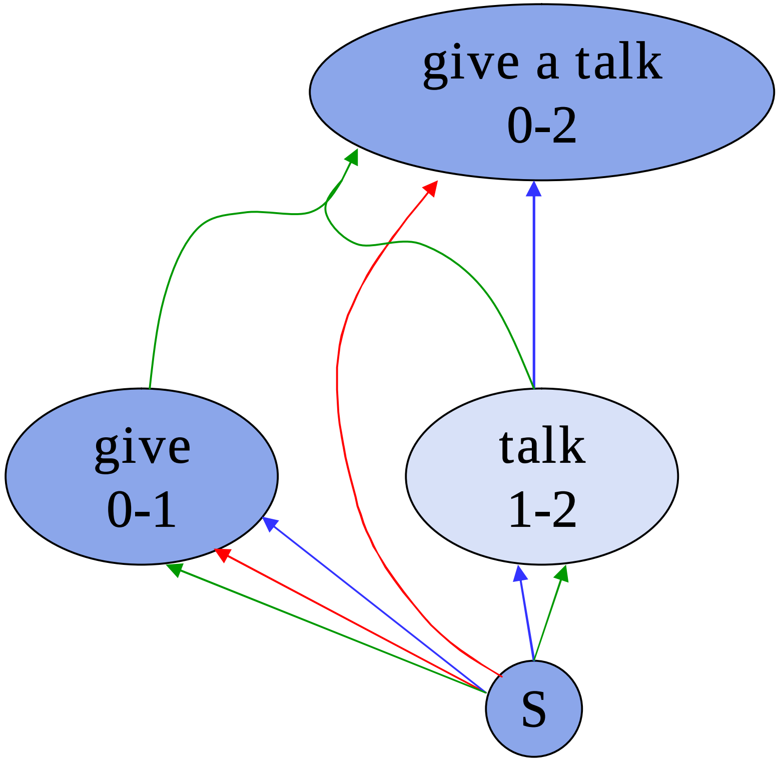
Joint Decoding

$_0$ fabiao $_1$ yanjiang $_2$



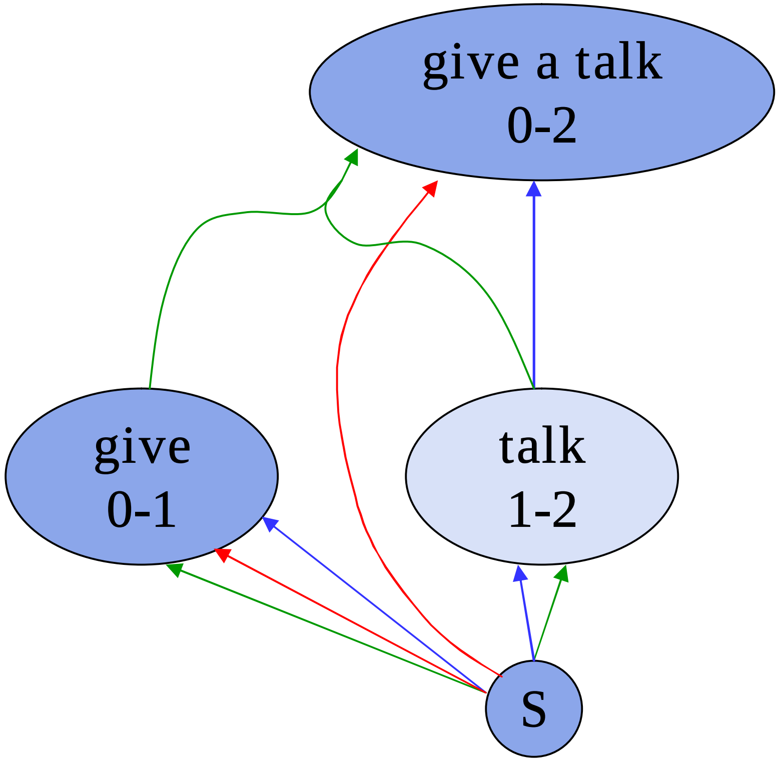
Joint Decoding

$_0$ fabiao $_1$ yanjiang $_2$

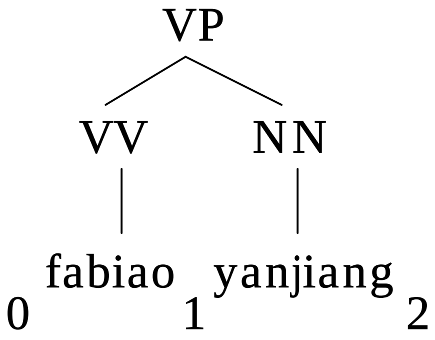


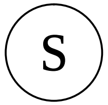
a pruned packed hypergraph

Sharing Hyperedges

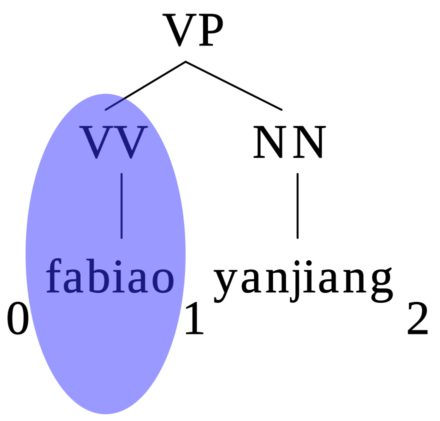


Joint Decoding with Different Rules

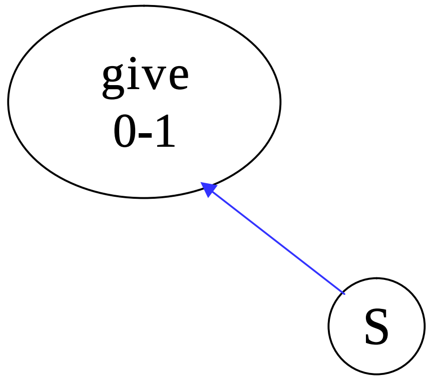




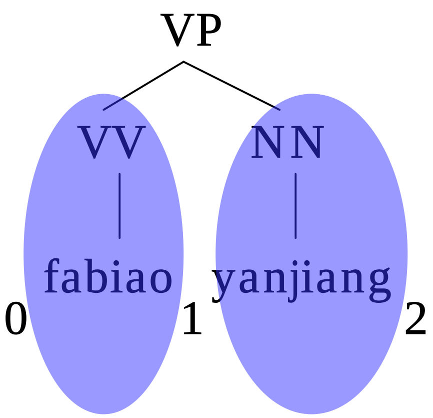
Joint Decoding with Different Rules



X -> <fabiao, give>

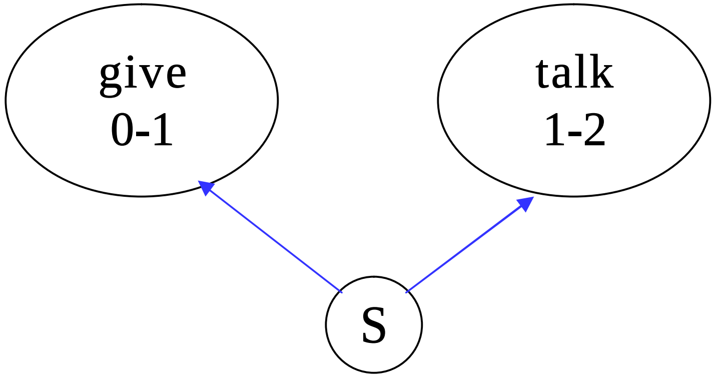


Joint Decoding with Different Rules

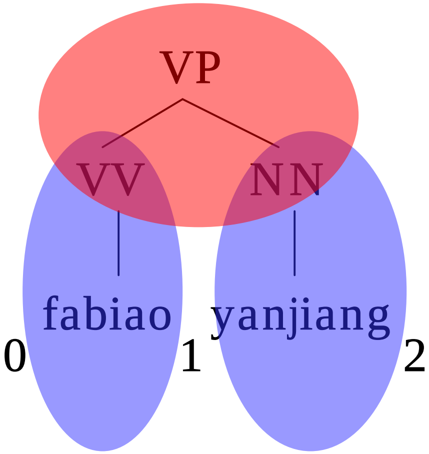


X -> <fabiao, give>

X -> <yanjiang, talk>



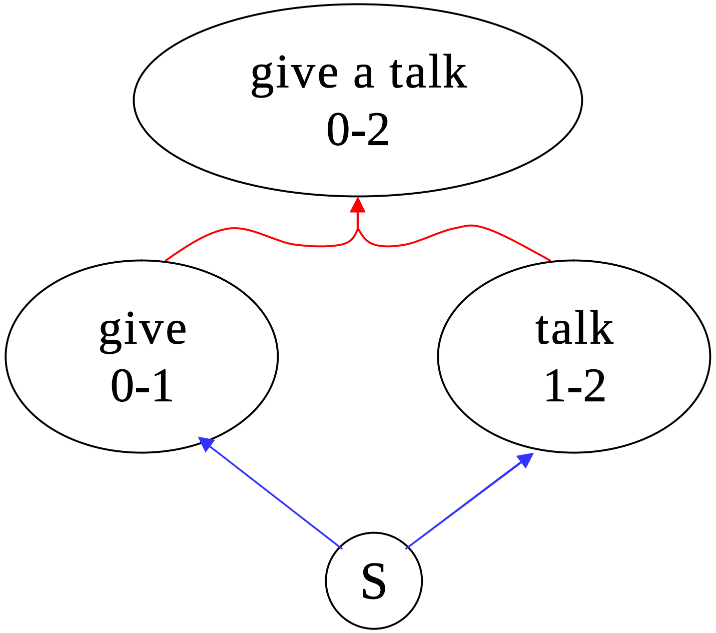
Joint Decoding with Different Rules



(VP (VV:x_1) (NN:x_2)) -> x_1 a x_2

X -> <fabiao, give>

X -> <yanjiang, talk>



An Example of MDD

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

model	feature		
	name	weight	value
hier	p(e|f)	1.0	0.1
	p(f|e)	1.0	0.2
	l(e|f)	1.0	0.1
	l(f|e)	1.0	0.3
	rc	1.0	2
t2s	p(e|f)	1.0	0.2
	p(f|e)	1.0	0.1
	l(e|f)	1.0	0.3
	l(f|e)	1.0	0.2
	rc	1.0	1
const	lm	1.0	0.7
	wc	1.0	3

8.2

Sharing Matrix

	Phrase	Hiero	T2S	S2T	T2T
Phrase	node, edge	node, edge	node, edge	node	node
Hiero	node, edge	node, edge	node, edge	node	node
T2S	node, edge	node, edge	node, edge	node	node
S2T	node	node	node	node, edge	node, edge
T2T	node	node	node	node, edge	node, edge

How to Tune Feature Weights for MTD?

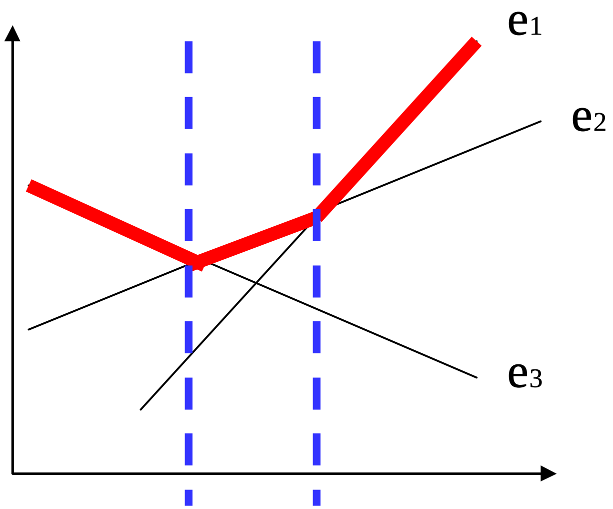
model	feature		
	name	weight	value
hier	p(e|f)	1.0	0.1
	p(f|e)	1.0	0.2
	l(e|f)	1.0	0.1
	l(f|e)	1.0	0.3
	rc	1.0	3
t2s	p(e|f)	1.0	0.2
	p(f|e)	1.0	0.1
	l(e|f)	1.0	0.3
	l(f|e)	1.0	0.2
	rc	1.0	4
const	lm	1.0	0.7
	wc	1.0	3

$$\hat{e} = \arg\max_e \left\{ \sum_{d\in\Delta(e,f)} \exp\left(\sum_i l_i h_i(d,e,f) \right) \right\}$$

$$(\exp(3.7) + \exp(4.8)) \times \exp(3.7)$$

MERT for MDD

- f
  - e_1
    - d_1 0.1 0.2 0.3 0.1
  - e_2
    - d_1 0.2 0.1 0.3 0.1
  - e_3
    - d_1 0.1 0.3 0.1 0.2



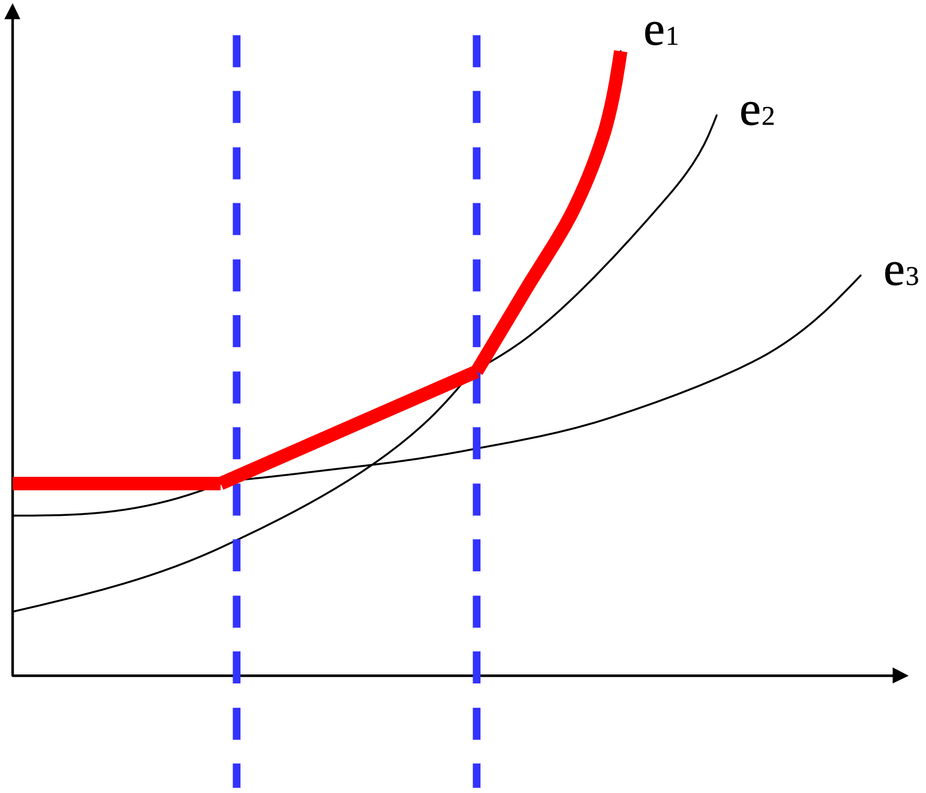
MERT for MTD

- f

	model 1				model 2			
e_1								
d_1	0.1	0.2	0.3	0.1	0.0	0.0	0.0	0.0
d_2	0.0	0.0	0.0	0.0	0.2	0.3	0.4	0.1
e_2								
d_1	0.0	0.0	0.0	0.0	0.3	0.1	0.2	0.1
d_2	0.1	0.2	0.3	0.1	0.0	0.0	0.0	0.0
e_3								
d_1	0.1	0.2	0.1	0.2	0.0	0.0	0.0	0.0
d_2	0.0	0.0	0.0	0.0	0.1	0.3	0.2	0.1

Curves

$$f(x)=\sum_{k=1}^{K} e^{a_k \times x + b_k}$$



Setup

- Models
  - Hierarchical phrase-based (Chiang, 2005)
  - Tree-to-string (Liu et al., 2006)
- Training set: FBIS (6.9M + 8.9M)
- Language model: 4-gram trained on GIGAWORD Xinhua portion
- Development set: NIST 2002 C2E
- Test set: NIST 2005 C2E

Individual Decoding Vs. Joint Decoding

Model	Sharing	Max-derivation		Max-translation	
		Time	BLEU	Time	BLEU
Hiero	-	40.53	30.11	44.87	29.82
T2S	-	6.13	27.23	6.69	27.11
both	node	-	-	55.89	30.79
	node & edge	48.45	31.63	54.91	31.49

Compared with System Combination

Method	Model	BLEU
individual	Hiero	30.11
	T2S	27.23
system comb.	both	31.50
joint	both	31.63

Individual Training Vs. Joint Training

Training	Max-derivation	Max-translation
individual	30.70	29.95
joint	31.63	30.79

Conclusion and Future Work

- We have presented a framework for combining different translation models in the decoding phrase.
- Future work
  - Including more models
  - Forced decoding
  - Hypergraph-based MERT (Kumar et al., 2009)

Thanks!