Consensus Decoding Approaches to Statistical Machine Translation

12/14/2011

Nan DUAN

Advisor: Ming ZHOU and Mu LI
School of Computer Science and Technology
Tianjin University

Outline

- Consensus Decoding to SMT
- Collaborative Decoding
- Mixture Model-based MBR Decoding
- Hypothesis Mixture Decoding
- Conclusion

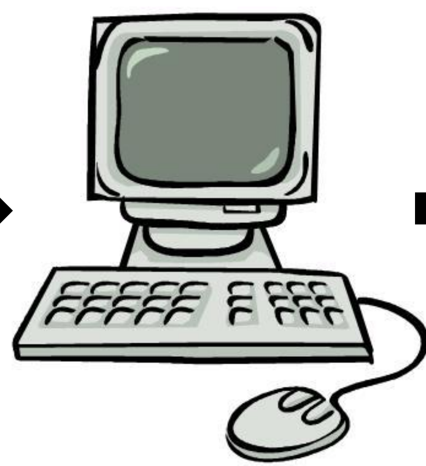
Outline

- Consensus Decoding to SMT
- Collaborative Decoding
- Mixture Model-based MBR Decoding
- Hypothesis Mixture Decoding
- Conclusion

“Statistical” Machine Translation

- **MT**: translation from one natural language into another natural language, using computer
- **SMT**: an MT paradigm based on *statistical model* and analysis of *bilingual data*

统计机器翻译



statistical machine translation

إحصائية الترجمة الآلية

統計的機械翻訳

सांख्यिकीय मशीनी अनुवाद

Modeling

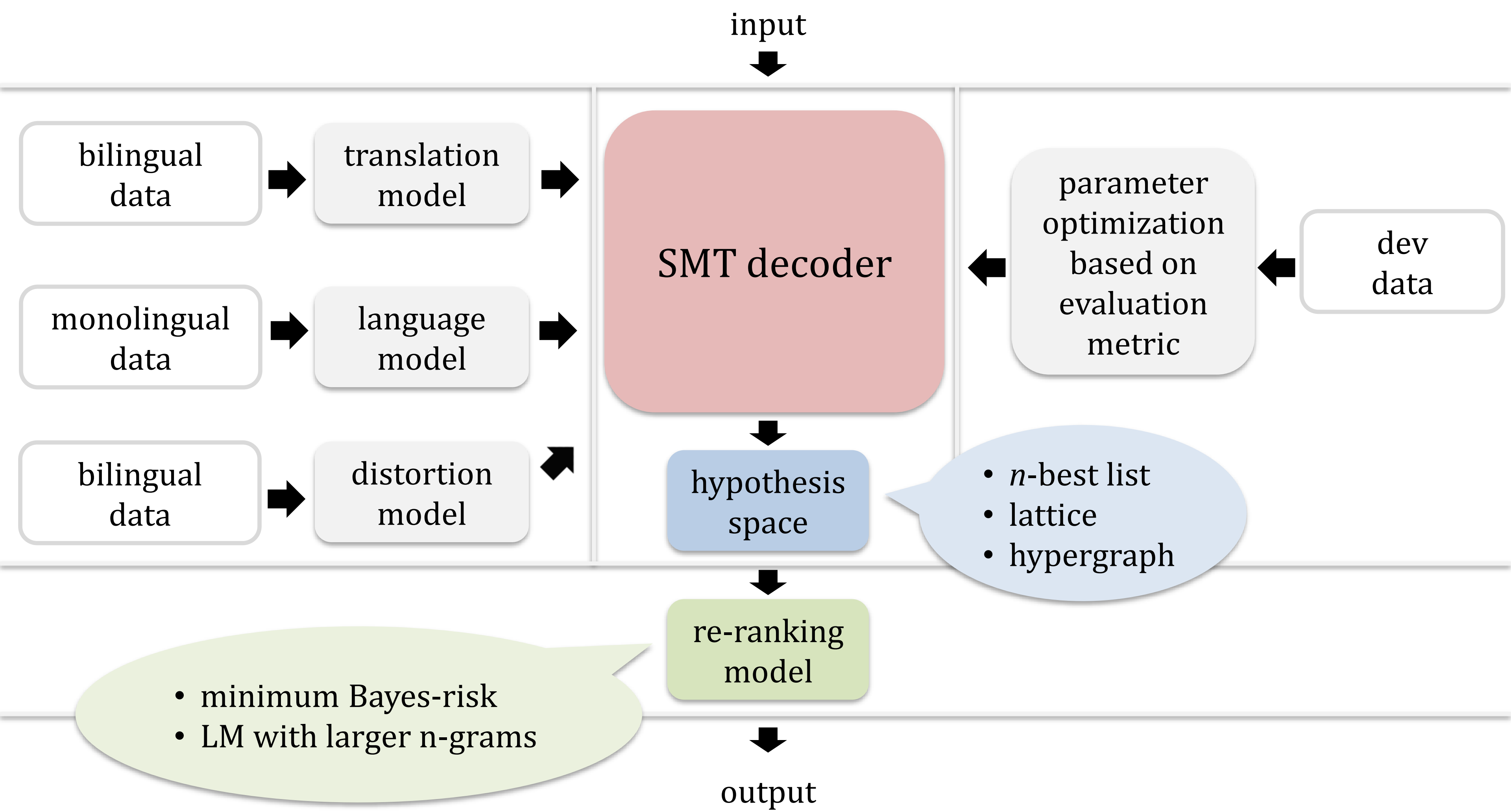
- Source-Channel Model

$$\hat{e} = argmax_e\{P(e|f)\}$$
$$= \underbrace{argmax_e}_{\text{Decoding Algorithm}}\{\underbrace{P(e)}_{\text{Language Model}} \cdot \underbrace{P(f|e)}_{\text{Translation Model}}\}$$

- Maximum Entropy Model

$$\hat{e} = argmax_e \frac{exp[\sum_{m=1}^{M} \lambda_m \cdot h_m(e,f)]}{\sum_{e'} exp[\sum_{m=1}^{M} \lambda_m \cdot h_m(e',f)]}$$
$$= argmax_e\{\sum_{m=1}^{M} \underbrace{\lambda_m}_{\text{Parameter Optimization}} \cdot \underbrace{h_m(e,f)}_{\text{Feature Selection}}\}$$

Generic System Overview



Existing Problems

- Single SMT model
  - 1-best translation is usually not the REAL-BEST one
  - better ones exist in the hypothesis space

How to select a better one?

- Multiple SMT models
  - different pros and cons
  - phrase-based models
    - good at handling local reorderings
    - bad at dealing with long-distance dependencies
  - syntax-based models
    - good at handling long-distance dependencies
    - bad at phrase coverage

How to combine merits of different SMT models?

Minimum Bayes-Risk Decoding

- Seek the translation with *the least expected loss* under a probability distribution

loss function

$$\hat{e} = \underset{e' \in H(f)}{argmin} \sum_{e \in H(f)} L(e,e')P(e|f)$$

$$= \underset{e' \in H(f)}{argmax} \sum_{e \in H(f)} G(e,e')P(e|f)$$

hypothesis space

gain function

probability distribution

MBR Decoding on Hypergraph
[Kumar et al., 2009]

MBR Decoding on Lattice
[Tromble et al., 2008]

Using BLEU as the loss function
[Ehling et al., 2007]

MBR Decoding on N-best
[Kumar and Byrne, 2004]

Related work

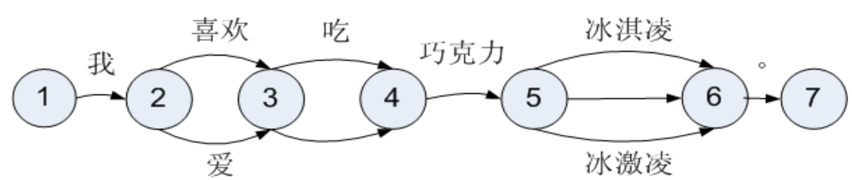
Word-Level System Combination

- Given outputs generated by multiple SMT systems
  1) align multiple hypotheses to build a confusion network
  2) output the path with the highest score as final translation

I like eating chocolate icecream.

我 喜欢 巧克力 冰激凌 。
我 喜欢 吃 巧克力 冰激凌 。
我 爱 吃 巧克力 冰激凌。
我 爱 巧克力 冰激凌。

我	喜欢	~	巧克力	冰激凌	。
我	喜欢	吃	巧克力	~	。
我	爱	吃	巧克力	冰淇凌	。
我	爱	~	巧克力	冰激凌	。



confidence score, null penalty, LM score, length penalty

$$\hat{e} = \underset{e \in H(f)}{argmax}\{\alpha P(e) + \beta N_{null}(e) + \chi logP_{LM}(e) + \sigma N_{words}(e)\}$$

我 爱 吃 巧克力 冰激凌 。

Consensus Decoding

- Techniques that can *re-rank* or *re-produce* better translations by making use of *consensus statistics* computed between hypotheses
  1. **system construction**
     - single system
     - multiple systems
  2. **first-pass decoding**
     - n-best list
     - hypergraph/lattice
  3. **search space re-construction**
     - simple union
     - more complicated structure
  4. **second-pass decoding**
     - re-rank existing translations
     - re-produce unseen translations

Outline

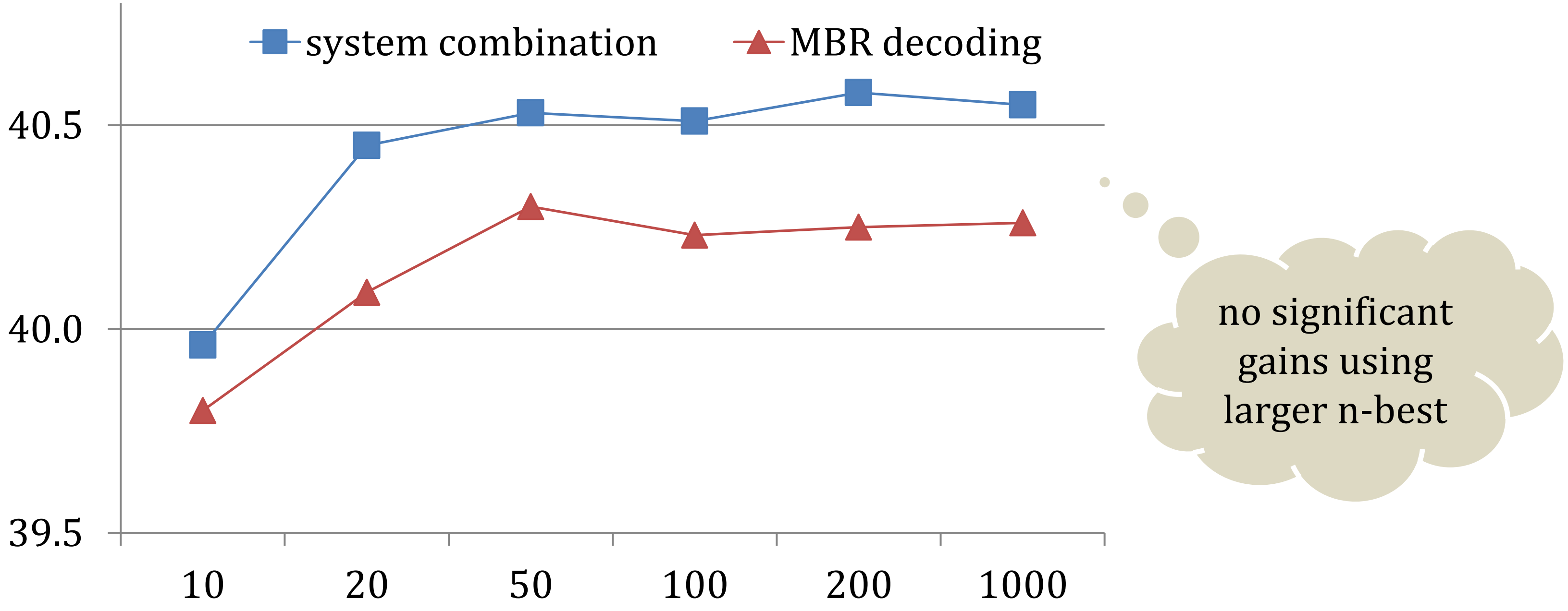
- Consensus Decoding to SMT
- Collaborative Decoding
- Mixture Model-based MBR Decoding
- Hypothesis Mixture Decoding
- Conclusion

Collaborative Decoding

ACL, 2009

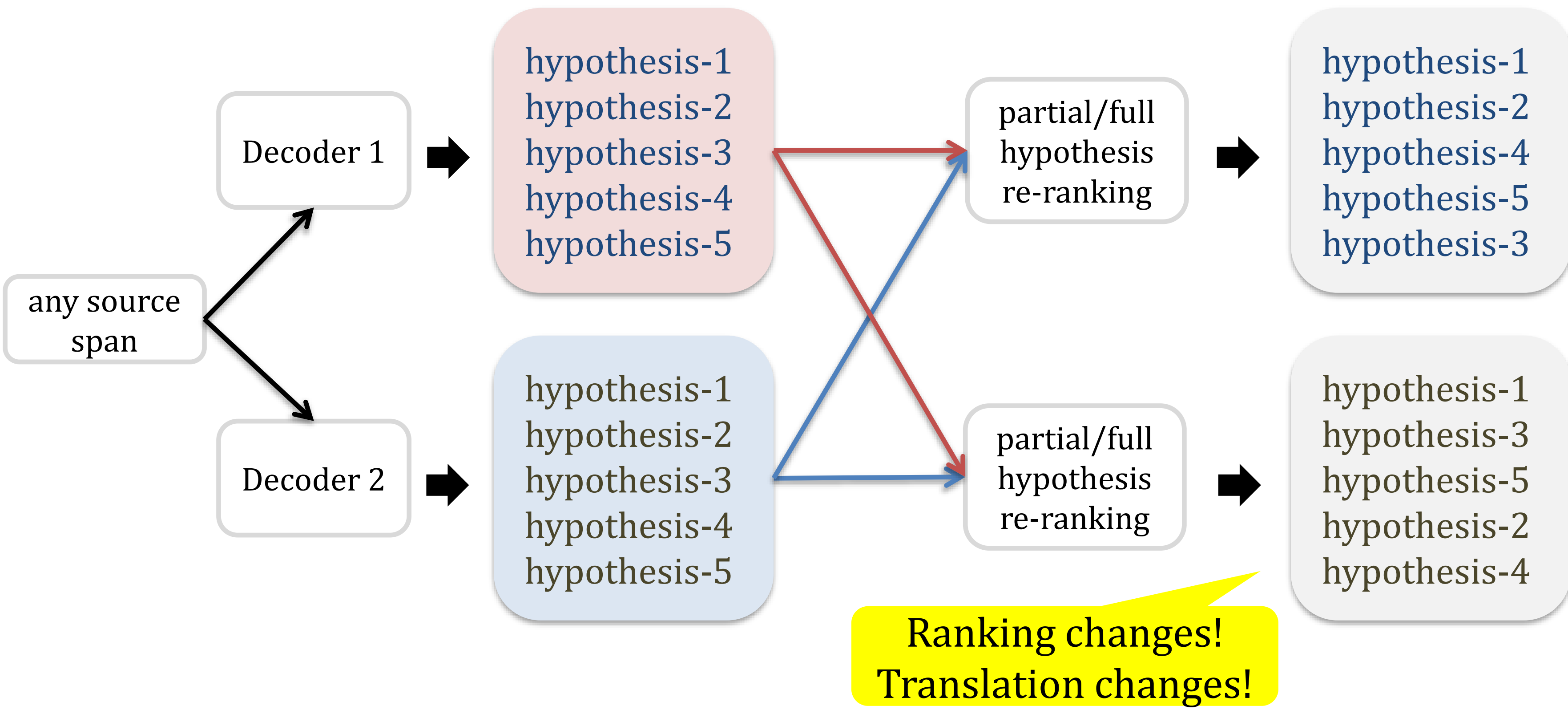
Motivation

- MBR decoding & system combination
  - take n-best translations as input
    - present *a small portion* of the whole search space
    - some useful partial translations are *pruned* during decoding



Collaborative Decoding

- Multiple SMT decoders work collaboratively
  - hypothesis scoring/re-ranking in decoding directly
  - explore translations beyond n-best lists of full translations



Co-decoding Model

- Baseline model
  - original SMT model

$$\Phi_m(f,e) = \sum_{i=1}^{N} \lambda_{m,i} \cdot \boxed{h_{m,i}(f,e)}$$

model specified features

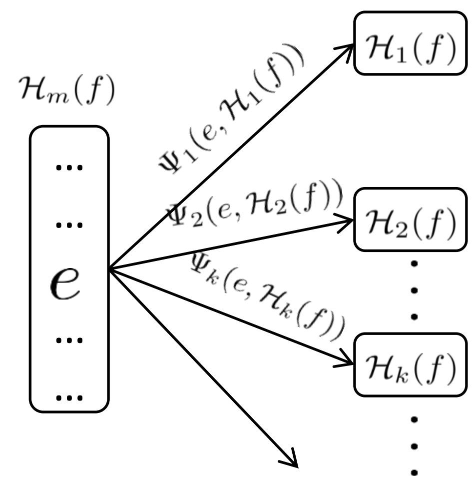
- Augmented model
  - consensus-based model

$$\sum_{k,k\neq m} \Psi_k(e,H_k(f)) = \sum_{k,k\neq m} \sum_{l} \lambda_{k,l} \cdot \boxed{h_{k,l}(e,H_k(f))}$$

n-gram -based features

- Co-decoding model
  - baseline model + augmented models

$$F_m(e) = \Phi_m(f,e) + \sum_{k,k\neq m} \Psi_k(e,H_k(f))$$



Co-decoding Feature

$$h_{k,l}(e, H_k(f)) = \sum_{e' \in H_k(f)} P(e' \mid H_k(f)) G_l(e, e')$$

Posterior Probability

- alpha is the scaling-factor that controls the shape of the probability distribution

$$P(e' \mid H_k(f)) = \frac{\exp\{\alpha \cdot F_k(e')\}}{\sum_{e'' \in H_k(f)} \exp\{\alpha \cdot F_k(e'')\}}$$

n-gram Measure Functions

- n-gram agreement measure

$$G_n^+(e, e') = \sum_{i=1}^{|e|-n+1} \delta(e_i^{i+n-1}, e')$$

- n-gram disagreement measure

$$G_n^-(e, e') = \sum_{i=1}^{|e|-n+1} (1 - \delta(e_i^{i+n-1}, e'))$$

Experiments

- Training
  - bilingual data (5.1M sentence pairs)
    - all data available for the NIST 2008 constrained track of C-to-E MT task
  - monolingual data
    - Xinhua portion of LDC English Gigaword V3.0
- Evaluation
  - development data
    - NIST 2003 (919 sentences)
  - test data
    - NIST 2005 (1,082 sentences) and NIST 2008 (1,357 sentences)
- Metric
  - case insensitive NIST BLEU

Experiments

- Baseline system
  - Hiero
    - (Chiang, 2005)
  - BTG
    - (Xiong et al., 2006)
  - DepHiero
    - (Shen et al., 2008)

- Combination system
  - more than one SMT model are used in co-decoding, so further combination is straightforward
    - word-level
      - (Rosti et al., 2007)
    - sentence-level
      - (Hildebrand and Vogel, 2008)

Overall Comparison

	NIST 2005	NIST 2008
	before co-decoding/**after co-decoding**	
Hiero	38.66%/**40.08%**	27.67%/**29.19%**
BTG	38.06%/**39.93%**	27.25%/**29.14%**
DepHiero	39.50%/**40.32%**	28.75%/**29.68%**
	Further Combination	
Word-Level Combination	40.45%/**40.85%**	29.52%/**30.35%**
Sentence-Level Combination	40.09%/**40.50%**	29.02%/**29.71%**

Summary

- Advantages
  - re-rank both partial and full translations
  - improve all involved SMT systems

- Future work
  - more systems included

Mixture Model-based MBR Decoding

COLING, 2010

Motivation

MBR decoding improves over max-derivation decoding

$$\hat{e} = \underset{e' \in H(f)}{argmin} \sum_{e \in H(f)} L(e,e') P(e|f)$$

Limitations of MBR decoding

- **It can only perform on single SMT system**
- **It can not adapt to multiple SMT systems**

Model distribution is the key point!

MBR Decoding for Hypergraph
[Kumar et al., 2009]

MBR Decoding for Lattice
[Tromble et al., 2008]

MBR Decoding for n-best
[Kumar and Byrne, 2004]

...

Different Search Spaces

Log-BLEU
[Tromble et al., 2008]

BLEU
[Ehling et al., 2007]

Word Error Rate (WER)
Position-independent WER
[Kumar and Byrne., 2004]

...

Different Loss Functions

In this work, we present Mixture Model-based MBR decoding (*MMMBR decoding*)

- ➢ **integrates multiple model distributions for Bayes-risk computation;**
- ➢ **integrates multiple search spaces for hypothesis selection.**

Overview of MMMBR Model

- **Generates multiple search spaces**

- **Integrates multiple search spaces as a combined search space**

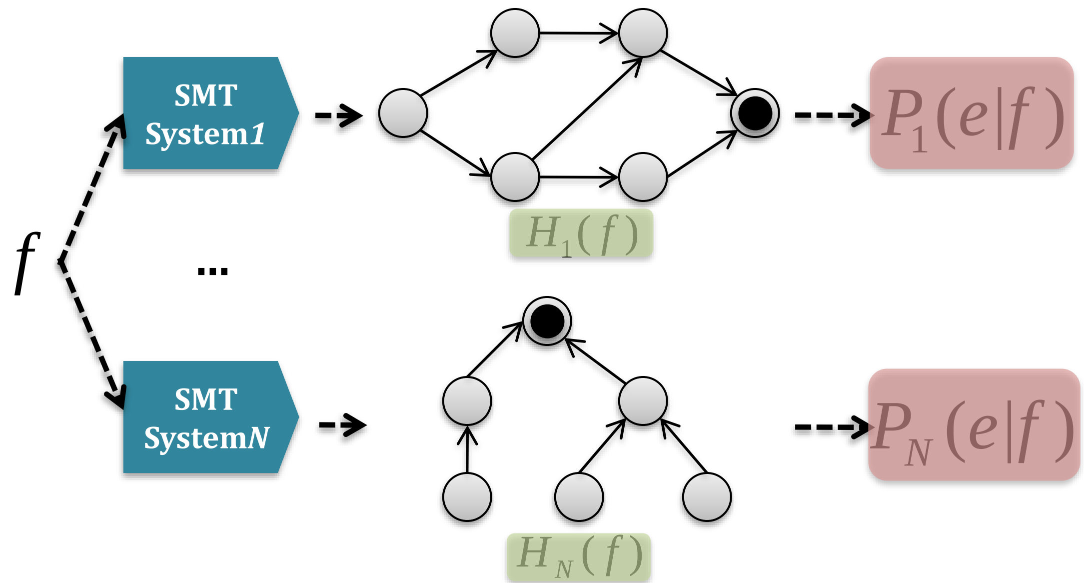
$$H(f) = \bigcup_{i=1}^{N} H_i(f)$$

- **Integrates multiple distributions by mixture modeling**

$$P(e|f) = \sum_{i=1}^{N} \lambda_i P_i(e|f)$$

$$\sum_{i=1}^{N} \lambda_i = 1 \quad 0 \le \lambda_i \le 1$$

- **Rewrites MBR to MMMBR**



$$\hat{e} = \underset{e' \in H(f)}{argmax} \sum_{e \in H(f)} G(e,e') \boxed{\sum_{i=1}^{N} \lambda_i P_i(e|f)}$$

Mixture Model

Mathematical Derivation

$$\hat{e} = \underset{e' \in H}{argmax} \sum_{e \in H} G(e,e') \sum_{i=1}^{N} \lambda_i P_i(e|f)$$

$$= \underset{e' \in H}{argmax} \sum_{e \in H} \{ \sum_{i=1}^{N} \lambda_i \sum_{k=1}^{N} \sum_{e \in H_k} \boxed{G(e,e')} P_i(e|f) \}$$

$$= \underset{e' \in H}{argmax} \sum_{e \in H} \{ \sum_{i=1}^{N} \lambda_i \sum_{k=1}^{N} \sum_{e \in H_k} \{ \theta_0 |e'| + \sum_{\omega} \theta_{|\omega|} \#_{\omega}(e') \delta_{\omega}(e) \} P_i(e|f) \}$$

$$= \underset{e' \in H}{argmax} \sum_{e \in H} \{ \sum_{i=1}^{N} \lambda_i \sum_{k=1}^{N} \theta_0 |e'| + \sum_{\omega} \theta_{|\omega|} \#_{\omega}(e') p_i(\omega \mid H_i) \}$$

MAP decoding score

$$= \underset{e' \in H}{argmax} \, \theta_0 |e'| + \sum_{\omega} \theta_{|\omega|} \#_{\omega}(e') \sum_{i} \lambda_i p_i(\omega \mid H_i) + \sum_{k} \theta_k \log C_{MAP}(e' \mid f, d_k)$$

- **decomposing Bayes-risk computation to each local search space**
- **using *log-BLEU* (Tromble et al., 2008) as the similarity function**
- **using *Algorithm 4* (Kumar et al., 2009) for *n*-gram posterior probability computation**

Model Training

- A two-pass training procedure

$$\hat{e} = \underset{e' \in H}{argmax}\, \theta_0 |e'| + \sum_{\omega} \theta_{|\omega|} \#_{\omega}(e') \sum_i \lambda_i p_i(\omega \mid H_i) + \sum_k \theta_k \log C_{MAP}(e' \mid f, d_k)$$

$$= \underset{e' \in H}{argmax} \sum_i \lambda_i \{\theta_0 |e'| + \sum_{\omega} \theta_{|\omega|} \#_{\omega}(e') p_i(\omega \mid H_i) + \sum_k \theta_k \log C_{MAP}(e' \mid f, d_k)\}$$

Step1: for MBR and MAP

- Fix system weights in mixture model
- Run Mert to tune MBR and MAP parameters

Step2: for Mixture Model

- Fix parameters in MBR and MAP models
- Run Mert to tune system weights

This procedure can be iteratively processed!

Experiments

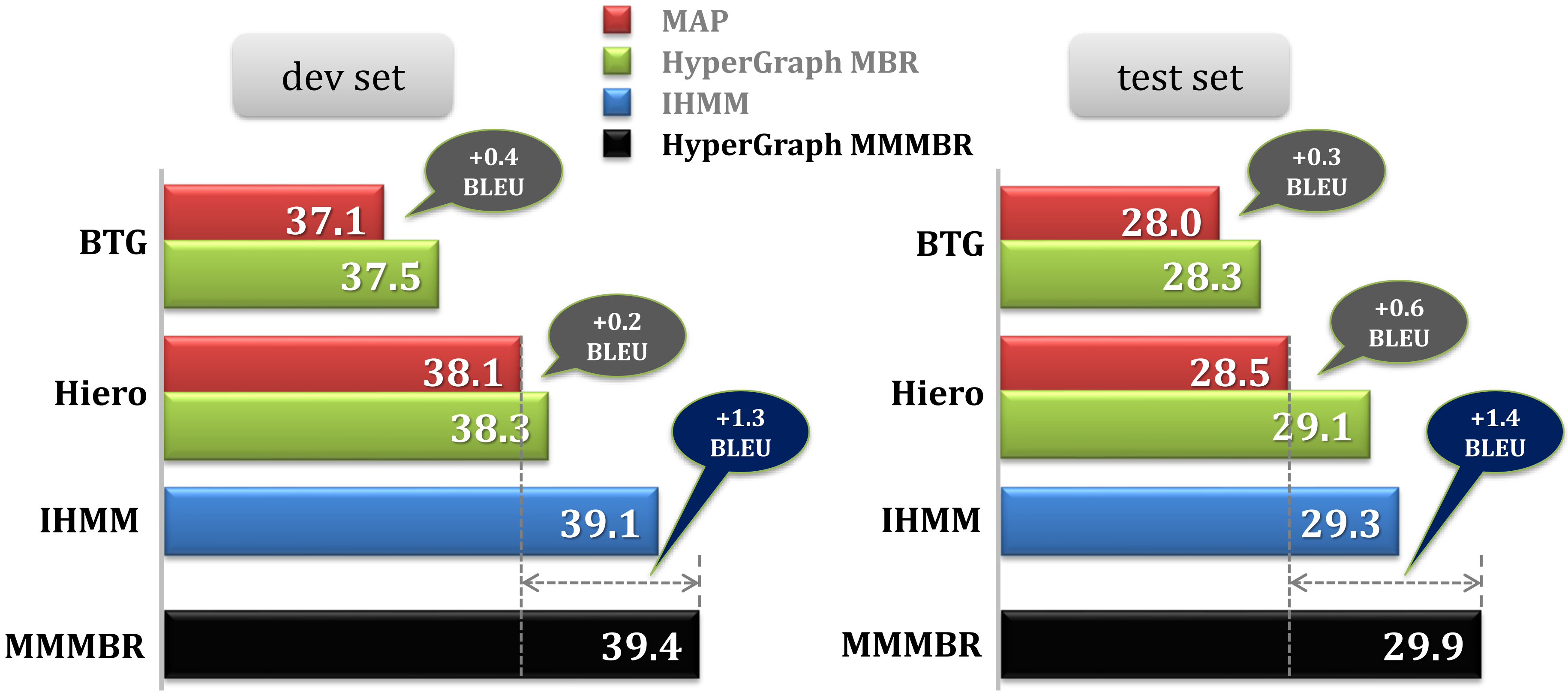
- Training
  - bilingual data (5.1M sentence pairs)
    - all data available for the NIST 2008 constrained track of C-to-E MT task
  - monolingual data
    - Xinhua portion of LDC English Gigaword V3.0
- Evaluation
  - development data
    - newswire portion of the NIST 2006 (616 sentences)
  - test data
    - NIST 2008 (1,357 sentences)
- Metric
  - case insensitive NIST BLEU

Experiments

- Baseline system
  - Hiero
    - (Chiang, 2005)
  - BTG
    - (Xiong et al., 2006)

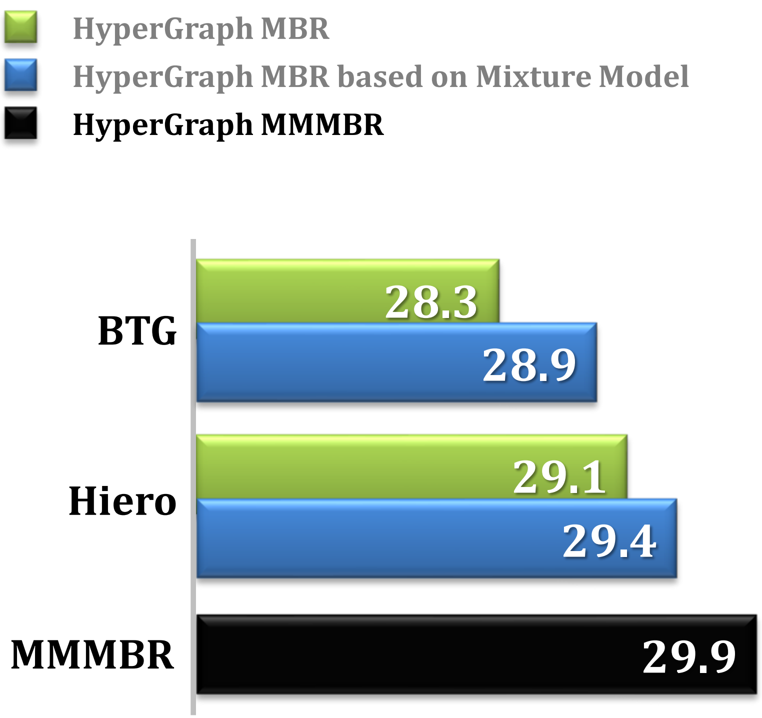
- Comparison technique
  - word-level system combination
    - (Li et al., 2009)

Overall Comparison



- MBR based on single system improves, but few (+0.2~+0.6 BLEU points)
- MMMBR based on multiple systems improves significantly (+1.3~+1.4 BLEU points)
- MMMBR is also comparable to word-level system combination (+0.3~+0.6 BLEU points)

Impacts of P(...) and H(...)



$$\hat{e} = \underset{e' \in H_i}{argmax} \sum_{e \in H_i} G(e,e') \boxed{P_i(e|f)}$$

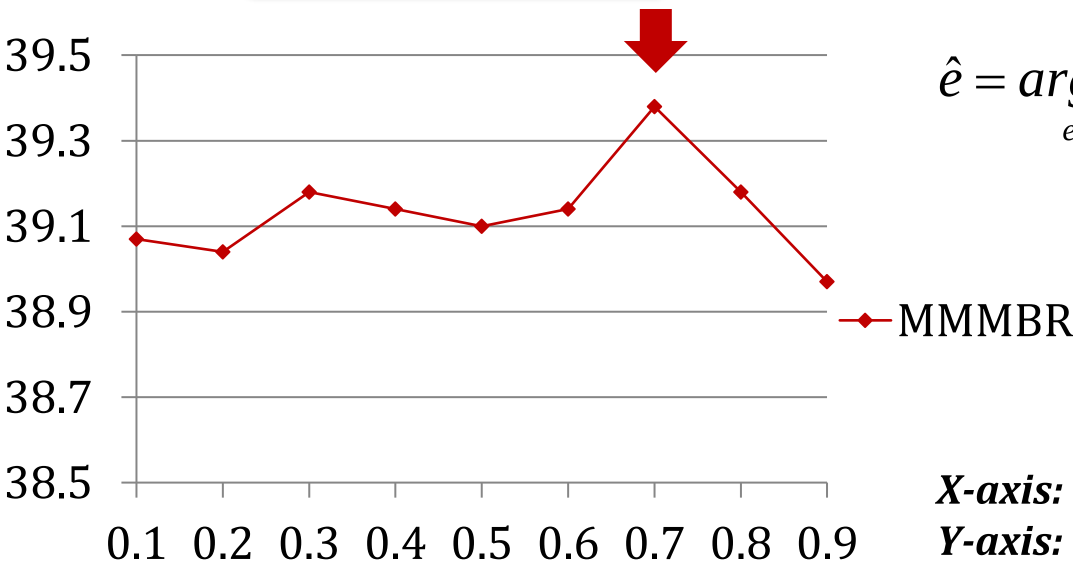
$$\hat{e} = \underset{e' \in H_i}{argmax} \sum_{e \in H_i} G(e,e') \boxed{\sum_{i=1}^{N} \lambda_i P_i(e|f)}$$

$$\hat{e} = \underset{e' \in H}{argmax} \sum_{e \in H} G(e,e') \sum_{i=1}^{N} \lambda_i P_i(e|f)$$

Using mixture model helps! + **Enlarging search space helps!** = best

Impacts of System Weight

NIST 2006



$$\hat{e} = \underset{e' \in H_i}{argmax} \sum_{e \in H_i} G(e,e') \sum_{i=1}^{N} \boxed{\lambda_i} P_i(e|f)$$

X-axis: system weight for Hiero
Y-axis: BLEU score

The performance is stable after the 1st round of two-pass training has been finished.

Summary

- Advantages
  - significant gains achieved
    - better than MBR decoding
    - comparable to word-level system combination
  - large hypothesis spaces explored
  - flexible and tunable system weights

- Future work
  - more systems included
    - syntax-based models

Hypothesis Mixture Decoding

ACL, 2011

Motivation

Various SMT Models

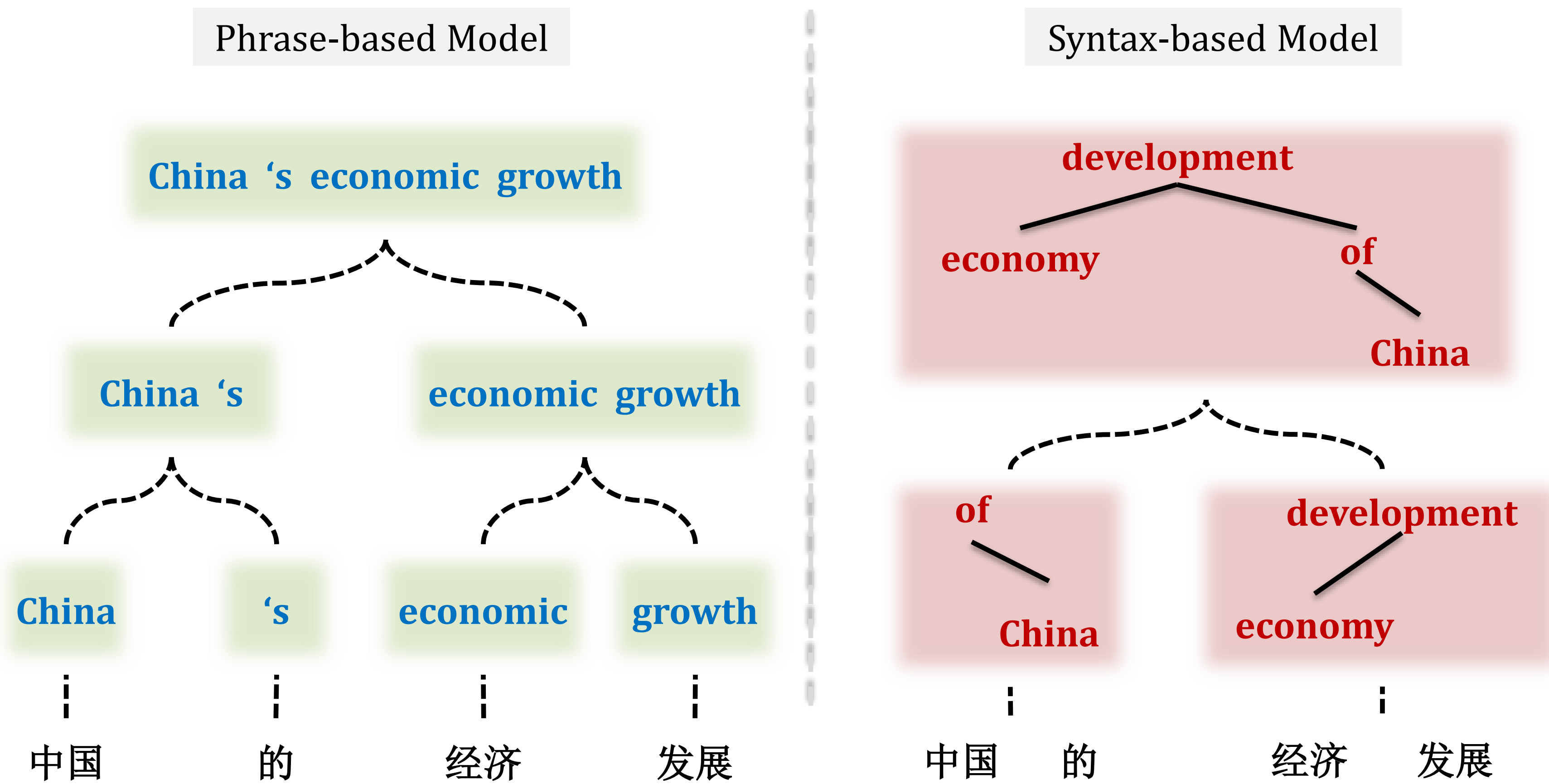
phrase-based, hierarchical phrase-based, syntax-based…

System Combination	**MBR Decoding**
n-best input	n-best/hypergraph input
confusion-network decoding	re-ranking
output	output
Pros new hypotheses can be generated	**Pros** more candidates can be leveraged
Cons few candidates can be leveraged	**Cons** no new hypothesis can be generated

How to combine these two merits?

Hypothesis Mixture Decoding

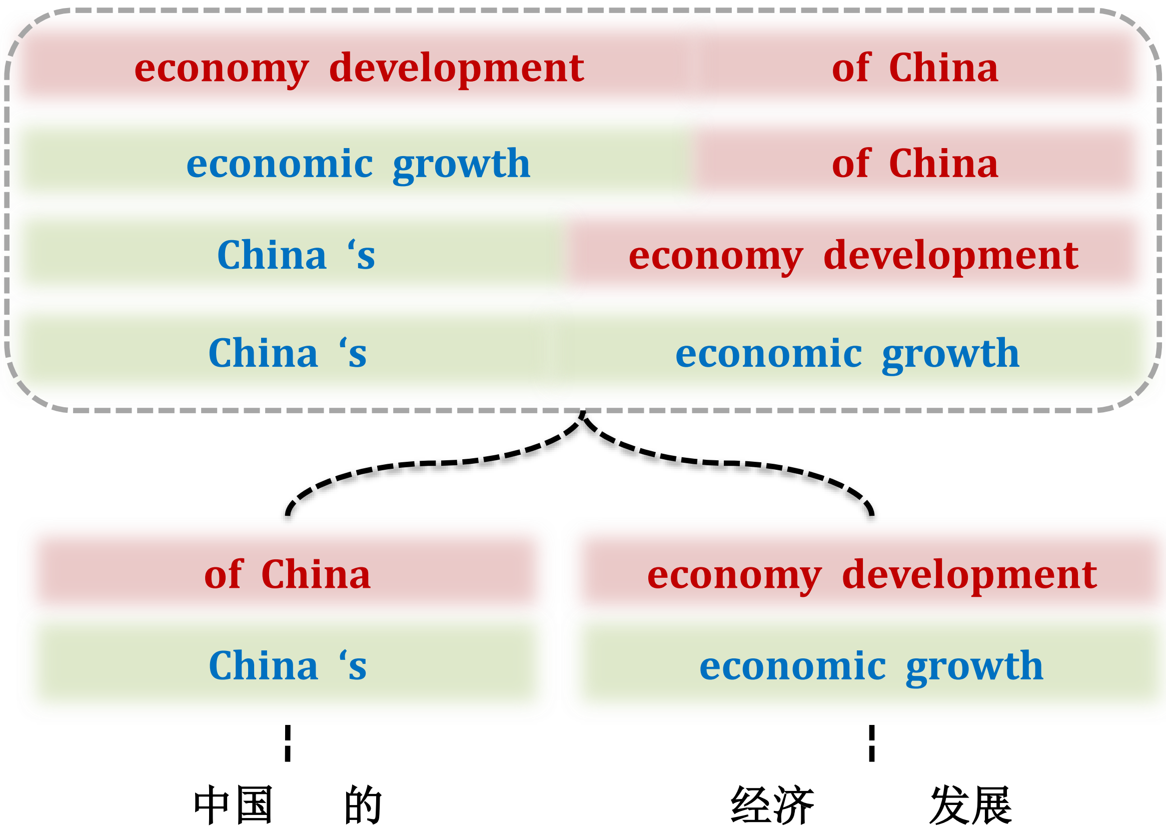
- Step-1: independent decoding



Hypothesis Mixture Decoding

- Step-2: hypothesis mixture decoding (HMD)

Mixture Hypothesis Space



translations beyond existing hypothesis spaces can be generated

partial hypotheses generated by arbitrary SMT models can be used

Models and Features

$$\hat{e} = \underset{e \in H(f)}{arg\,max} \sum_i \lambda_i \cdot h_i(e, f)$$

- Linear combination of two sets of features
  - **consensus-based features**
    - n-gram posteriors
    - length posteriors
  - **general features**
    - LM score
    - word penalty
    - count of lexicon pair
    - reordering score
    - count of new generated n-grams

Two Decoding Algorithms

HM Decoding Algorithm

for each component model M_n **do**
    output the hypothesis space $H_n(f)$ for the input
end for
for $l = 1$ to $|f| - 1$ **do**
    for all i, j s.t. $j - i = l$ **do**
        $\mathrm{H}(f_i^j) = \{nil\}$
        Hypothesis Re-Construction
        for each hypothesis $e \in \bigcup_{n=1}^{N} H_n(f_i^j)$ **do**
            compute HMD features for e
            add e to $H(f_i^j)$
        end for
    end for
end for
return $\hat{e} \in H(f)$ with the maximum score

(I)BTG-based Hypothesis Re-Construction

for all k s.t. $i \le k < j$ **do**
    for $e_1 \in H(f_i^k)$ and $e_2 \in H(f_{k+1}^j)$ **do**
        add $e = Comb_{[\bullet]}(e_1, e_2)$ to $H(f_i^j)$
        add $e = Comb_{\langle\bullet\rangle}(e_1, e_2)$ to $H(f_i^j)$
    end for
end for

(II)SCFG-based Hypothesis Re-Construction

for each rule $r \in R$ that matches f_i^j **do**
    for $e_1 \in H(r_{\#1})$ and $e_2 \in H(r_{\#2})$ **do**
        add $e = Comb_r(e_1, e_2)$ to $H(f_i^j)$
    end for
end for

Experiments

- Training
  - bilingual data (5.1M sentence pairs)
    - all data available for the NIST 2008 constrained track of C-to-E MT task
  - monolingual data
    - Xinhua portion of LDC English Gigaword V3.0
- Evaluation
  - development data
    - NIST 2004 (1,788 sentences)
  - test data
    - NIST 2005 (1,082 sentences)
    - newswire portion of NIST 2006 (616 sentences) / NIST 2008 (691 sentences)
- Metric
  - case insensitive NIST BLEU

Experiments

- Baseline System
  - DHPB: string-to-dependency system
    - (Shen et al., 2008)
  - PB: phrase-based system
    - (Xiong et al., 2006)

- Comparison Technique
  - Comb: word-level system combination
    - (Li et al., 2009)
  - CD: collaborative decoding
    - (Li et al., 2009)
  - MMMBR: mixture model-based MBR decoding
    - (Duan et al., 2010)

Overall Comparison

		NIST 2004	NIST 2005	NIST 2006	NIST 2008
Baseline	DHPB	39.90%	39.76%	35.00%	30.43%
	PB	38.93%	38.21%	33.59%	29.62%
System Combination	Comb	41.14%	40.70%	36.04%	31.16%
Collaborative Decoding	CD-DHPB	40.81%	40.56%	35.73%	30.87%
	CD-PB	40.39%	40.34%	35.20%	30.39%
	CD-Comb	41.27%	41.02%	36.37%	31.54%
MMMBR Decoding	MMMBR	41.19%	40.96%	36.30%	31.43%
HM Decoding	HMD-BTG	41.24%	41.26%	36.76%	31.69%
	HMD-SCFG	41.31%	41.19%	36.63%	31.52%
	HMD-Comb	**41.74%**	**41.53%**	**37.11%**	**32.06%**

Summary

- Advantages
  - HM decoding combines merits of word-level system combination and MBR decoding
    - large hypothesis spaces
    - capability to produce unseen translations
- Future work
  - explore richer consensus information
  - explore better reordering models for HM decoding

Outline

- Consensus Decoding to SMT
- Collaborative Decoding
- Mixture Model-based MBR Decoding
- Hypothesis Mixture Decoding
- Conclusion

Conclusion

- Consensus decoding helps SMT, by
  - *re-ranking* existing translations
    - MBR decoding
    - mixture model-based MBR decoding
    - model combination
    - collaborative decoding
    - …
  - *re-producing* new translations
    - word-level system combination
    - hypothesis mixture decoding
    - …
- Future work
  - explore richer consensus information
  - adapt to other NLP tasks

My Research at MSRA

(since 2006.12-now)

- C-E and J-E SMT Projects
  - EngKoo
  - Bing Translator
  - Prototype SMT systems for research
- MT Competitions
  - NIST 2008 (international)
  - CWMT 2008 (domestic)
  - CWMT 2009 (domestic)
  - IWSLT 2010 (international)
    - joint work with MSR Redmond
- Publications
  - 10 international conference papers
    - 3 ACL, 2 COLING, 1 EMNLP and etc.

APPENDIX

- More contents/details after…

Feature Subspace-based System Combination

EMNLP, 2009

Motivation

- Two ways to build multiple SMT systems for system combination
  - using models based on different paradigms
    - phrase-based
    - hierarchical phrase-based
    - syntax-based
    - …
  - using models based on different training methods
    - different word-aligners
    - different LMs
    - different word segmentations
    - …

Challenging,
even for "old birds"

Time-consuming,
sometimes not applicable

- Is there any *light-weight* method for system building?

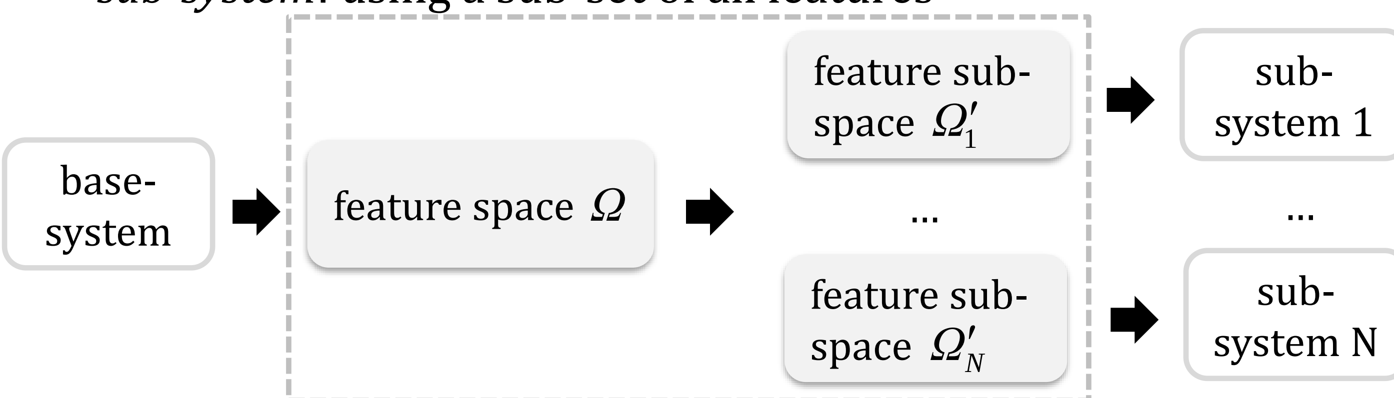
Feature Subspace-based System Combination (1)

- Ensemble Generation using Single SMT System
  - most SMT systems are based on linear models
    - feature space: $\Omega = \{h_m(e,f) \mid m = 1,\dots,M\}$
    - feature subspace: $\Omega' \subset \Omega$

$$\hat{e} = argmax_{e \in H(f)} \sum_{m=1}^{M} \lambda_m \cdot \boxed{h_m(e,f)}$$

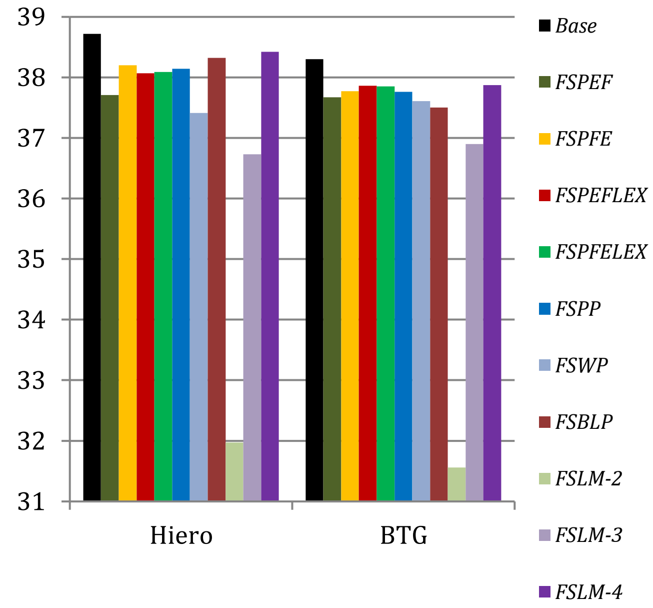
a set of features

  - each feature subspace maps to a new SMT system
    - *base-system*: using all features
    - *sub-system*: using a sub-set of all features



Principle of Feature Selection

- Not all feature subspaces are helpful
  - too many candidates: $|\Omega'| = 2^{|\Omega|}$
  - poor performances for most of them



- Our Feature Selection Principle
  - for non-LM features
    - remove one feature each time
    - use the remainders for system construction
  - for LM feature
    - lower its n-gram order
    - use it with the remainders for system construction

Now's SMT systems CANNOT live without LM!

Feature Subspace-based System Combination (2)

- Sentence-Level System Combination
  - re-ranking the union of hypotheses generated by all systems

$$H(f) = \bigcup_{i=1}^{N} H_i(f)$$

$$\hat{e} = \underset{e \in H(f)}{argmax}\{\lambda_{LM} h_{LM}(e) + \lambda_l h_l(e) + \psi(e,H(f))\}$$

simply union all hypotheses generated by base-system and sub-systems without bias

A weighted combination of n-gram consensus features

- n-gram *agreement* feature: $h_n^+(e,H(f)) = \sum_{e' \in H(f)} G_n(e,e')$
- n-gram *disagreement* feature: $h_n^-(e,H(f)) = \sum_{e' \in H(f)} (|e| - n + 1 - G_n(e,e'))$
- n-gram consensus: $G_n(e,e') = \sum_{i=1}^{|e|-n+1} \delta(e_i^{i+n-1}, e')$

Experiments

- Training
  - bilingual data (5.1M sentence pairs)
    - all data available for the NIST 2008 constrained track of C-to-E MT task
  - monolingual data
    - Xinhua portion of LDC English Gigaword V3.0
- Evaluation
  - development data
    - NIST 2003 (919 sentences)
  - test data
    - NIST 2004 (1,788 sentences) and NIST 2005 (1,082 sentences)
- Metric
  - case insensitive NIST BLEU

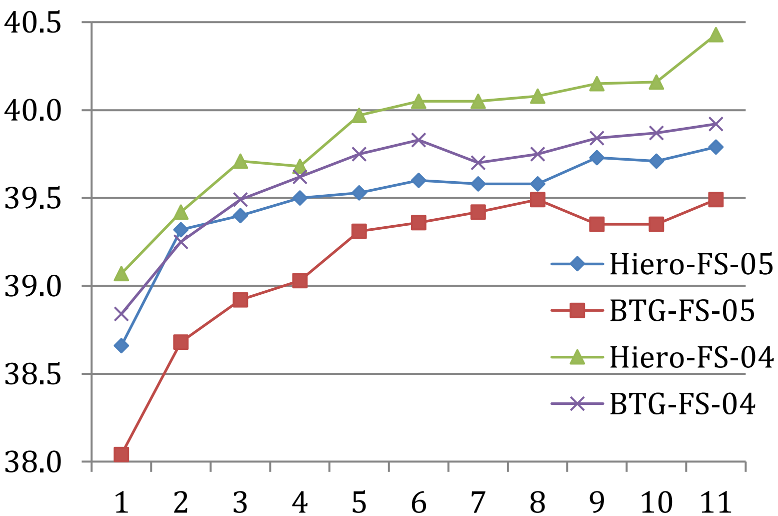
Experiments

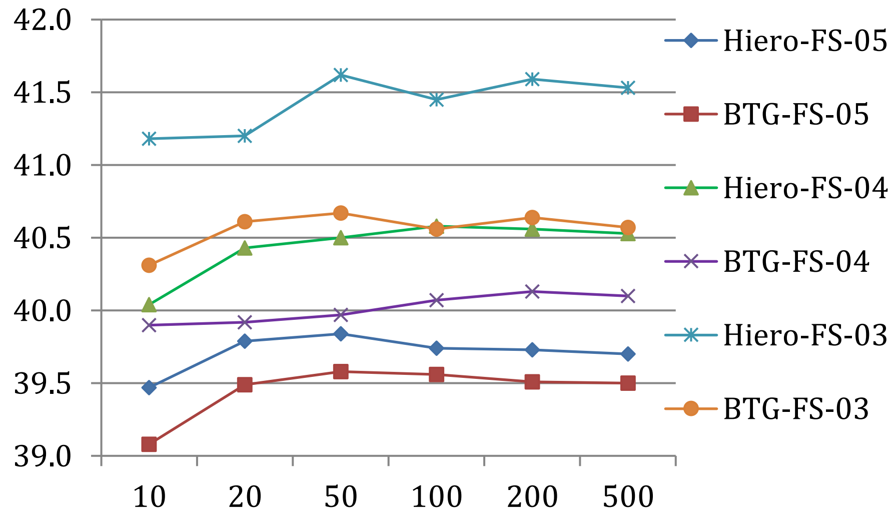
- Base-System
  - Hiero (Chiang, 2005)
  - BTG (Xiong et al., 2006)
- Sub-Systems
  - removing one of the non-LM features
    - *PEF*/*PFE*: source-to-target/target-to-source translation probability
    - *PEFLEX*/*PFELEX*: source-to-target/target-to-source lexical weight
    - *PP*: phrase penalty
    - *WP*: word penalty
    - *BLP*: bi-lexicon counting
  - lowering the order of an 5-gram LM
    - *LM-4*/*LM-3*/*LM-2*
  - (7+3)=10 sub-systems for each base-system

Overall Comparison

		NIST 2004	NIST 2005
Hiero	Base-system 220-best	39.07%	38.72%
	Combination (1 base-system+ 10 sub-systems) 20-best for each system	40.43% (+1.36%)	39.79% (+1.07%)
BTG	Base-system 220-best	38.84%	38.30%
	Combination (1 base-system+ 10 sub-systems) 20-best for each system	39.92% (+1.08%)	39.49% (+1.19%)
Hiero + BTG	Combination (Hiero + BTG) 220-best for each system	39.98%	39.43%
	Combination (2 base-systems+ 20 sub-systems) 20-best for each system	40.96% (+0.98%)	39.49% (+1.06%)

Impacts of Different Settings





Impacts of different numbers of sub-systems for system combination

- X-axis: number of sub-systems
- Y-axis: BLEU score

Impacts of different sizes of n-best lists for system combination

- X-axis: n-best size for each system
- Y-axis: BLEU score

Contributions of Sub-Systems

	NIST 2004	NIST 2005
Hiero	69.71%	69.69%
BTG	59.07%	58.54%

Ratio of unique hypotheses from sub-systems

	NIST 2004	NIST 2005
Hiero	44.63%	46.12%
BTG	47.54%	44.73%

Ratio of final translations from sub-systems

Oracle Performance		Hiero	BTG
		BLEU/TER	BLEU/TER
NIST 2004	Base-System 220-best	49.68%/0.6411	49.50%/0.6349
	1 Base-System + 10 Sub-Systems 20-best for each system	51.05%/0.6089	50.53%/0.6056
NIST 2005	Base-System 220-best	48.89%/0.5946	48.37%/0.5944
	1 Base-System + 10 Sub-Systems 20-best for each system	50.69%/0.5695	49.81%/0.5684

Summary

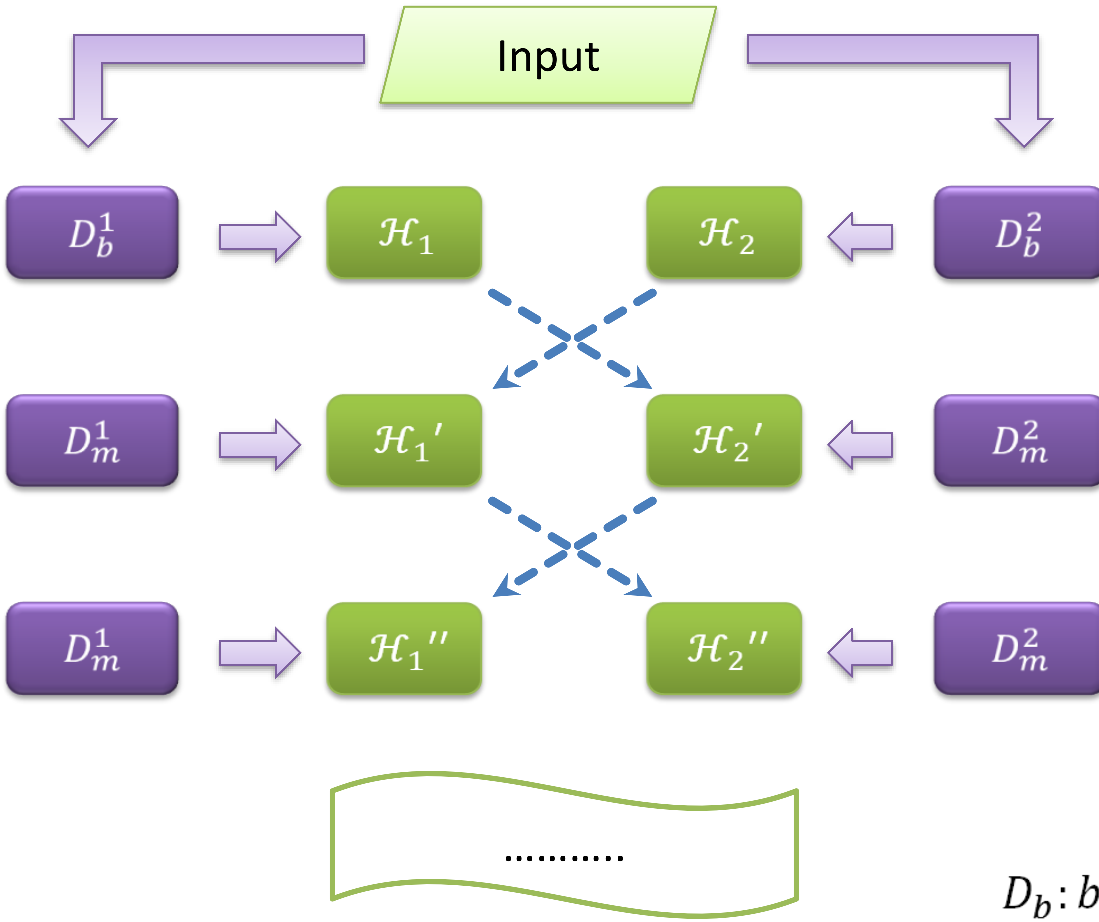
- Pros
  - simple and effective way for multiple system construction
  - useful to all SMT systems based on the linear model

- Cons
  - feature selection principle is simple

Collaborative Decoding

ACL, 2009

Decoder coordination: Iterative Decoding



D_b: *baseline decoder*
D_m: *member decoder*

Impacts of Different Settings

NIST 2005	Hiero	BTG
Baseline	38.66%	38.04%
+agree -disagree	39.36%	39.02%
-agree +disagree	39.12%	38.67%
+agree +disagree	39.68%	39.61%

TER before CD **TER after CD**	**NIST 2005**	**NIST 2008**
Hiero-BTG	0.3190 **0.2204**	0.4016 **0.2686**
Hiero-DepHiero	0.3252 **0.1840**	0.4176 **0.2469**
BTG-DepHiero	0.3498 **0.2171**	0.4238 **0.2665**

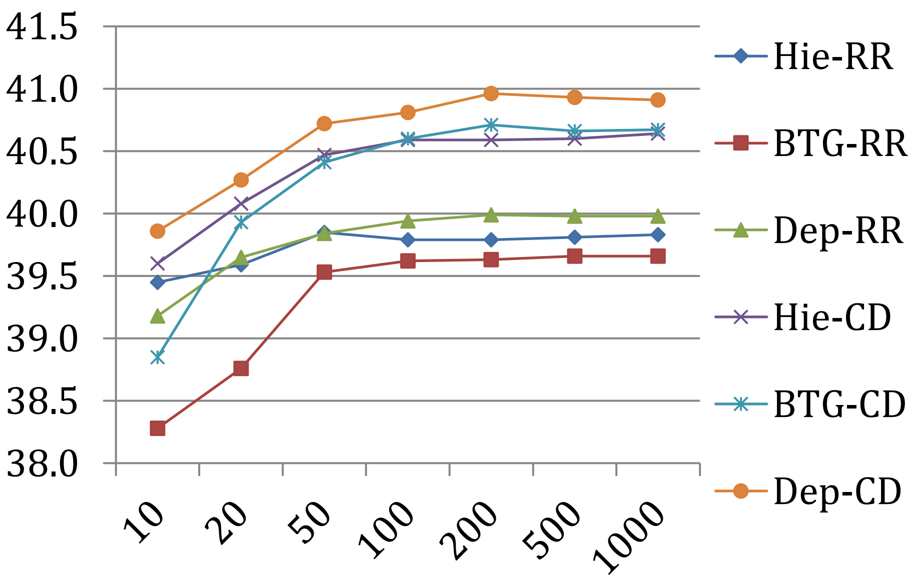
Impact of different consensus features

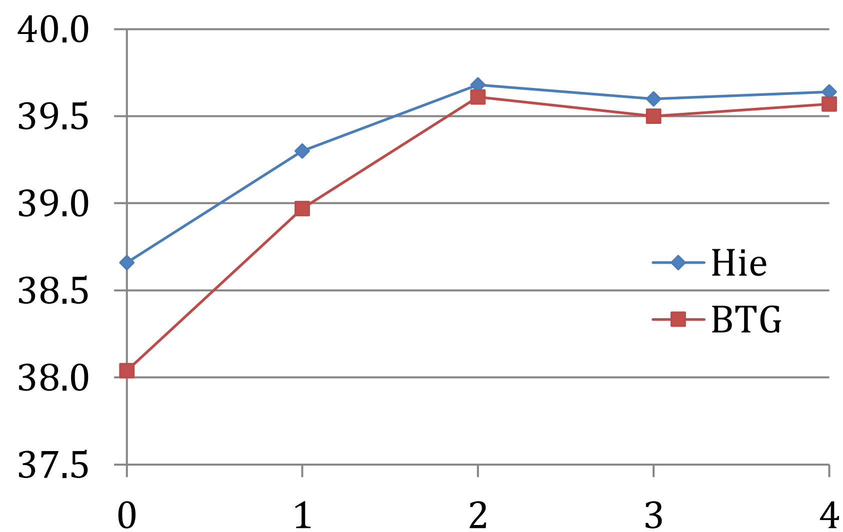
- Both n-gram agreement features and disagreement features are useful

TER scores between translations

- Outputs become more similar due to the use of consensus information

Impacts of Different Settings





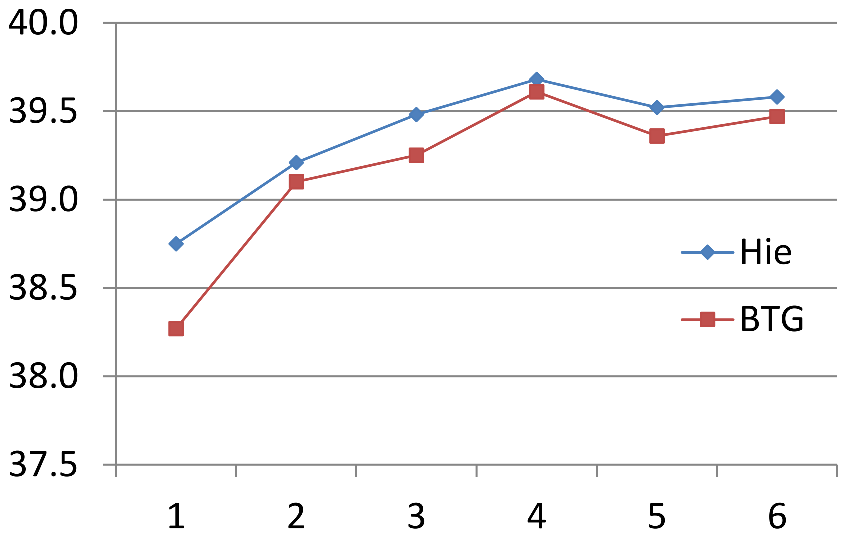
Co-decoding (**CD**) vs. Re-ranking(**RR**)

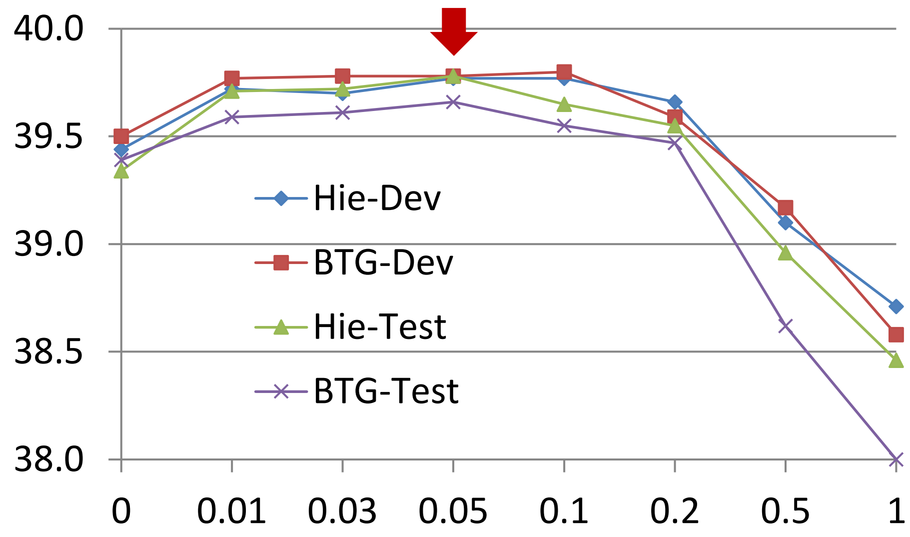
- X-axis: n-best size for each system
- Y-axis: BLEU score
- test data: NIST 2005

Co-decoding with different iterations

- X-axis: times of iteration
- Y-axis: BLEU score
- test data: NIST 2005

Impacts of Different Settings





Impact of n-gram consensus measure

- X-axis: the order of n-grams
- Y-axis: BLEU score
- NIST 2005

Impact of scaling factor

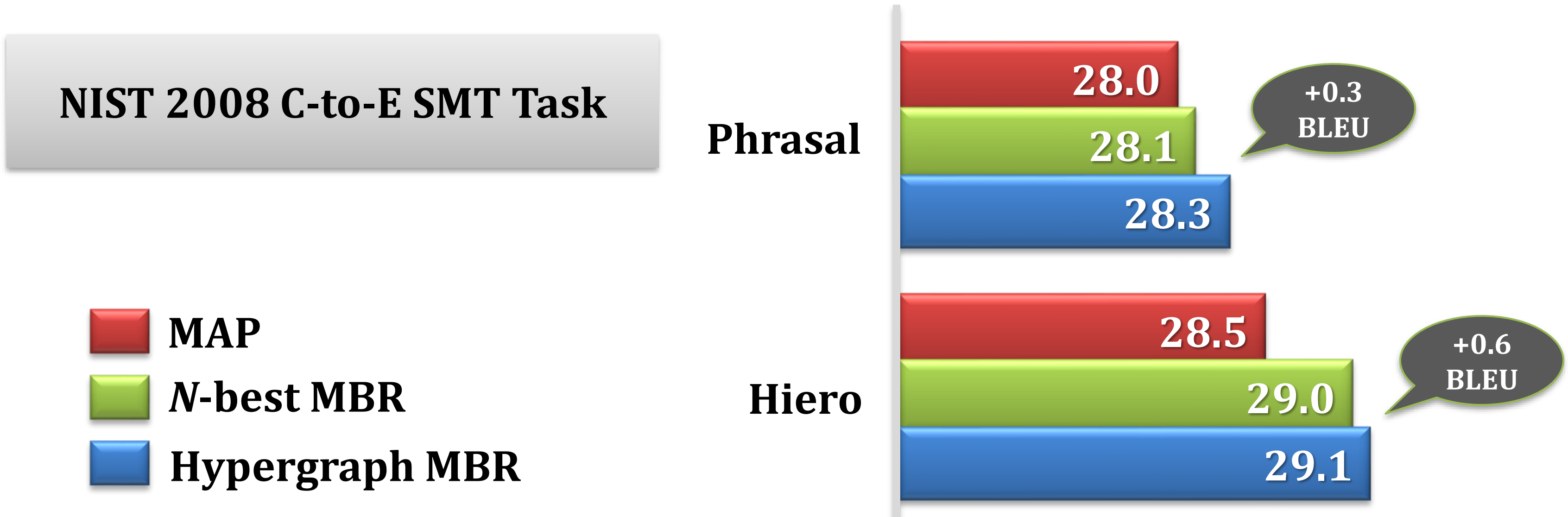
- X-axis: value of scaling factor
- Y-axis: BLEU score
- NIST 2003 (Dev) and NIST 2005 (Test)

Mixture Model-based MBR Decoding

COLING, 2010

Limitations of MBR Decoding

- Single system –based
- High correlation between hypotheses
- Relative few improvements achieved



Extending MBR decoding to multiple SMT systems is straight-forward!

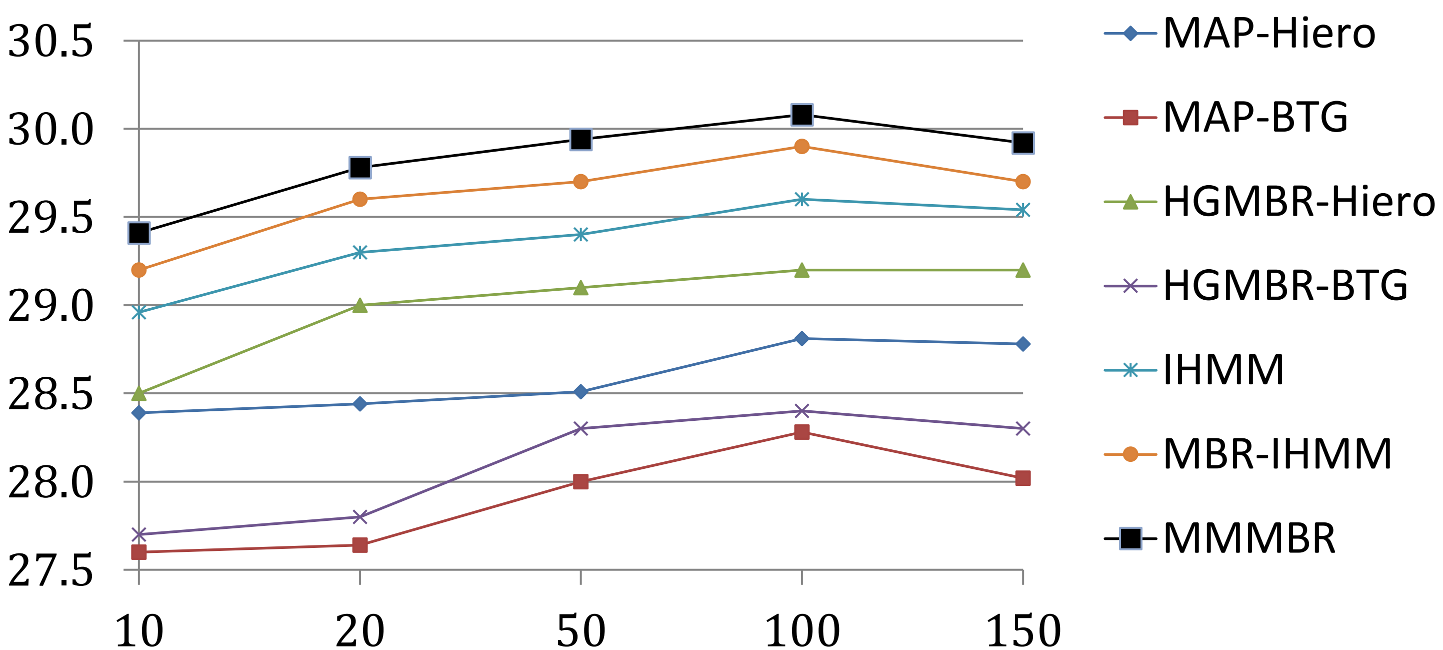
MMMBR Decoding Algorithm

MMMBR decoding on multiple MT systems

for each component MT system d_k **do**
  generate the output search space H_k
  compute the n-gram posterior probability set $\{p_k(w|H_k)\}$ for H_k
end for
compute the mixture n-gram posterior probability for each n-gram:
for each unique n-gram w appeared in $\cup_k H_k$ **do**
  for each search space H_i **do**
    $p(w) \mathrel{+}= \lambda_i p_i(w|H_i)$
  end for
end for
for each hyperedge e in $\cup_k H_k$ **do**
  assign $p(w)$ to the edge e for all w contained in e
end for
return the best path according to the MMMBR decision rule

Impacts of Beam Size

NIST 2008



X-axis: beam size
Y-axis: BLEU score

MBR Decoding on Single Model

- **Decision Rule**

$$\hat{E} = \underset{E' \in H}{\operatorname{argmax}} \sum_{E \in H} G(E, E')\, P(E \mid F)$$

$$= \underset{E' \in H}{\operatorname{argmax}} HGMBR(H, E' \mid P)$$

- **Loss Function**
  - $|E'|$ is hypothesis length
  - $\#_\omega(E')$ is the number of times ω occurs in E'
  - $Cost(E')$ is the model cost of the first-pass decoding
  - θ_i are feature weights

$$\hat{E} = \underset{E' \in H}{\operatorname{argmax}} \sum_{E \in H} \boxed{\{\theta_0 |E'| + \sum_\omega \theta_{|\omega|} \#_\omega(E')\, \delta_\omega(E') + \theta_m Cost(E')\}} P(E \mid F)$$

$$= \underset{E' \in H}{\operatorname{argmax}} \theta_0 |E'| + \sum_\omega \theta_{|\omega|} \#_\omega(E')\, p(\omega \mid H) + \theta_m Cost(E')$$

$$p(\omega \mid H) = \sum_{E \in H} \delta_\omega(E)\, P(E \mid F)$$

Algorithm for HGMBR

Step-1
- **Sort the hypernodes topologically**

Step-2
- **Compute inside-outside probabilities of each hypernode**

Step-3
- **Compute posterior probabilities of each hyperedge**

Step-4
- **Compute posterior probabilities of each *n*-gram***

Step-5
- **Find the path with the highest score in hypergraph**

Inside-Outside Probability

- **Inside Probability**
  - $w(h)$ is the weight of hyperedge h, given by $\exp(\alpha x^T \varphi(h))$, α is the scaling factor
  - $I(u)$ and $O(u)$ refer to the incoming and outgoing hyperedges of hypernode u

$$I(u) = \sum_{h \in In(u)} w(h)[\prod_{v \in tail(h)} I(v)]$$

- **Outside Probability**
  - For the root node of the hypergraph, $O(root) = 1$
  - $head(h)$ is the hypernode which h pointed to

$$O(u) = \sum_{h \in Out(u)} w(h)[O(head(h)) \prod_{\substack{v \in tail(h) \\ v \neq u}} I(v)]$$

Posterior Probability

- **Hyperedge Posterior Probability**
  - $Z(f) = I(root)$ is the inside probability of the root of the hypergraph

$$p(h \mid H) = \frac{1}{Z(f)} w(h) O(head(h)) \prod_{v \in tail(h)} I(v)$$

- ***n*-gram Posterior Probability**
  - Approximate this quantity by the ***sum*** of posterior probabilities of edges which contribute the *n*-gram ω and have the highest edge posterior probability relative to their predecessors on each path

$$p(\omega \mid H) = \sum_{e \in H} 1_{\omega \in e} f^{*}(e, \omega, H)\, p(e \mid H)$$

Hypothesis Mixture Decoding

ACL, 2011

Consensus-based Features

- n-gram posteriors based on existing hypothesis space

$$h_{\mathrm{H}_n(f)}(e, f) = \sum_{\omega \in e} \#_{\omega}(e) p(\omega \mid \mathrm{H}_n(f))$$

- n-gram posteriors based on stemmed hypothesis space

$$h_{\mathrm{H}_n^S(f)}(e^S, f) = \sum_{\omega \in e^S} \#_{\omega}(e^S) p(\omega \mid \mathrm{H}_n^S(f))$$

- n-gram posteriors based on mixture hypothesis space

$$h_{\mathrm{H}(f)}(e, f) = \sum_{\omega \in e} \#_{\omega}(e) p(\omega \mid \mathrm{H}(f))$$

- length posterior based on mixture hypothesis space

$$h_I(e, f) = \sum_{e' \in \mathrm{H}(f), |e'| = |e| = I} p(e' \mid f)$$

General Features

- LM score + word penalty
- count of lexicon pairs
- reorder features for (BTG/SCFG)-HMD respectively

$$h_{[\bullet]}(e,f)=\sum_{r\in D(e)}\delta_r([\bullet])$$

$$h_{\langle\bullet\rangle}(e,f)=\sum_{r\in D(e)}\delta_r(\langle\bullet\rangle)$$

$$h_{SCFG-Rule}(e,f)=\sum_{r\in D(e)}\delta_r(R)$$

$$h_{Glue-Rule}(e,f)=\sum_{r\in D(e)}\delta_r([\bullet])$$

- count of new generated n-grams

$$h_{New}(e,f)=\sum_{\omega\in e}\#_{\omega}(e)\bar{\delta}_{\omega}(\cup_{n=1}^{N}\mathrm{H}_n(f))$$

Background

On-line SMT Engines

- Microsoft
  - http://dict.bing.com.cn

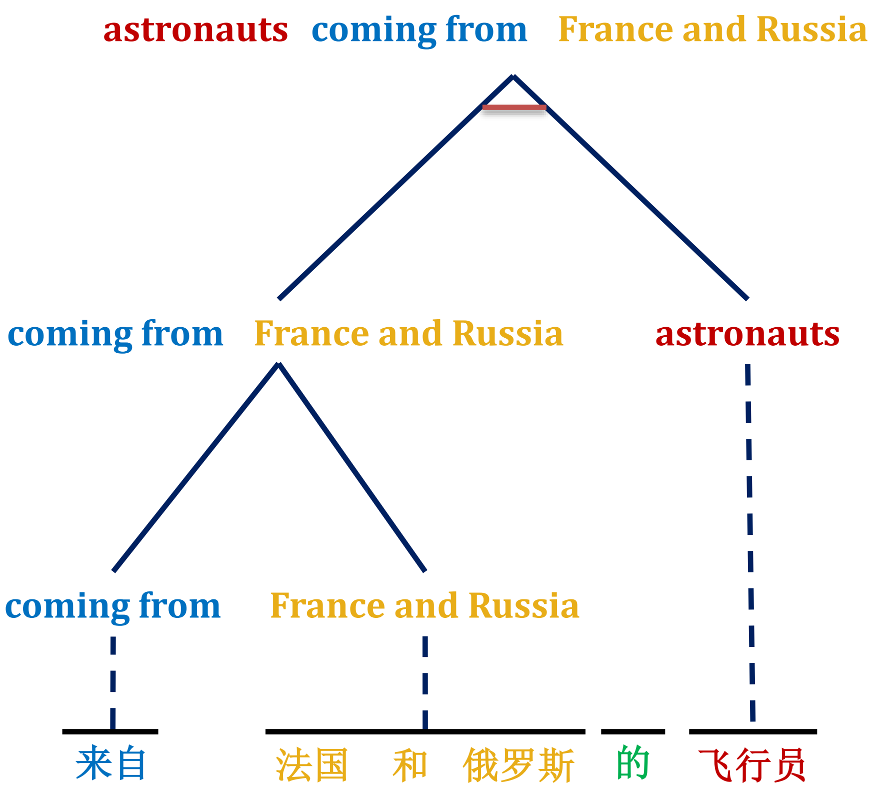
- Google
  - http://translate.google.com

- Yahoo!
  - http://babelfish.yahoo.com/

- Baidu
  - http://fanyi.baidu.com/

State-of-the-Art SMT Models

- Phrase-based Model



X_1 X_2 ||| X_2 X_1

飞行员 ||| astronauts

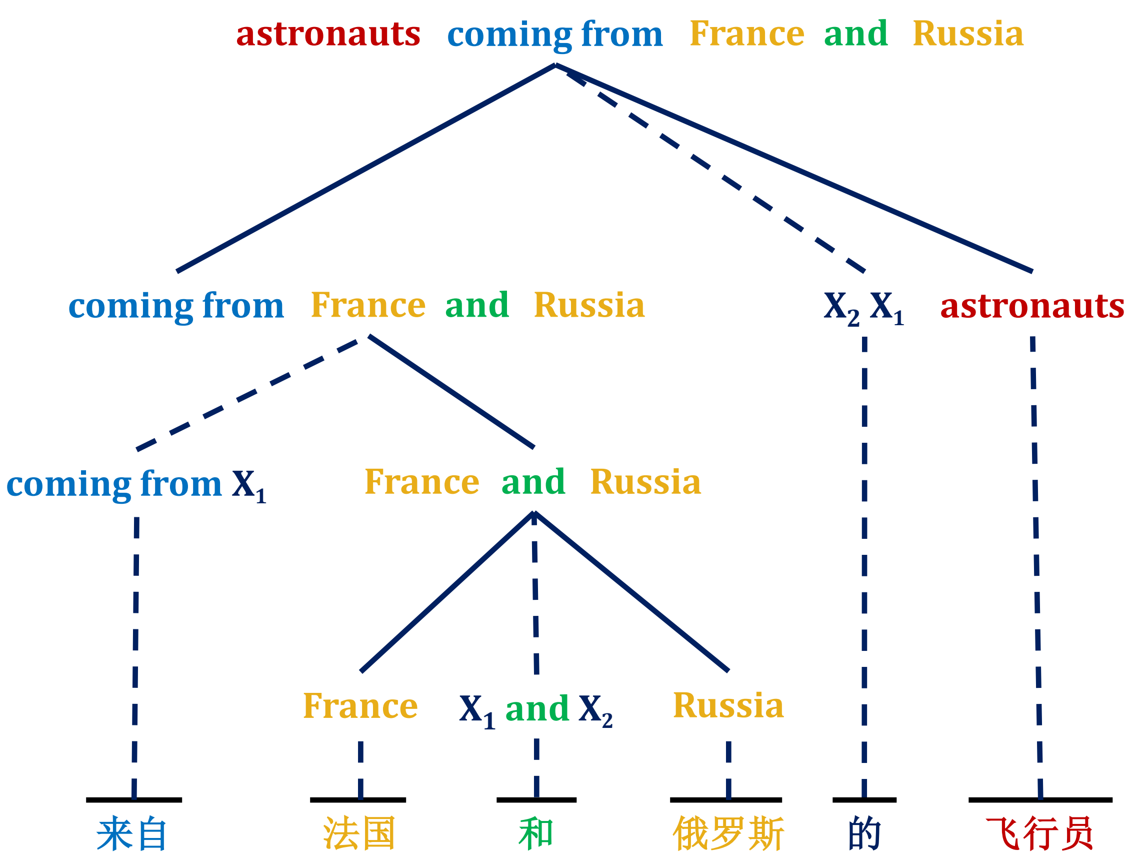
X_1 X_2 ||| X_1 X_2

法国 和 俄罗斯 ||| France and Russia

来自 ||| coming from

State-of-the-Art SMT Models

- Hierarchical Phrase-based Model



X_1 的 X_2 ||| X_2 X_1

飞行员 ||| **astronauts**

来自 X_1 ||| **coming from** X_1

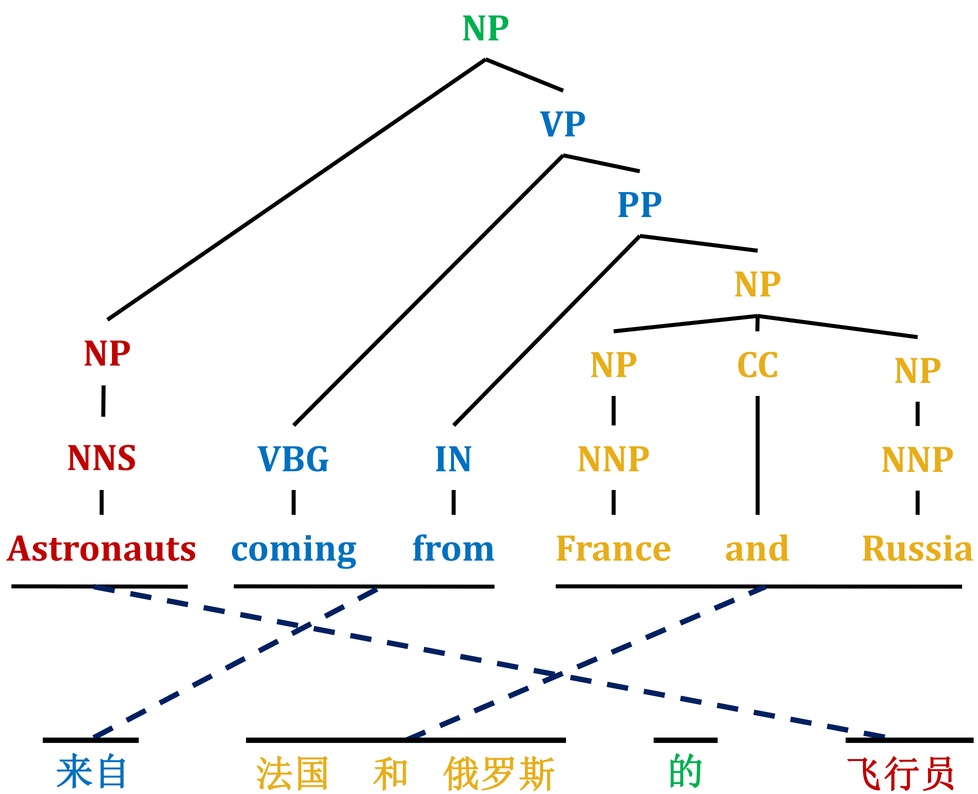
X_1 和 X_2 ||| X_1 **and** X_2

俄罗斯 ||| **Russia**

法国 ||| **France**

State-of-the-Art SMT Models

- Syntax-based Model



法国 和 俄罗斯 ||| NP{NNP{France}} CC{and} NP{NNP{Russia}} ||| NP

来自 X_1 ||| VBG{coming} IN{from} NP:X_1 ||| VP

飞行员 ||| NNS{astronauts} ||| NP

X_1 的 X_2 ||| NP:X_2 VP:X_1 ||| NP

n-best List & Hypergraph

n-best list

a cat on the mat

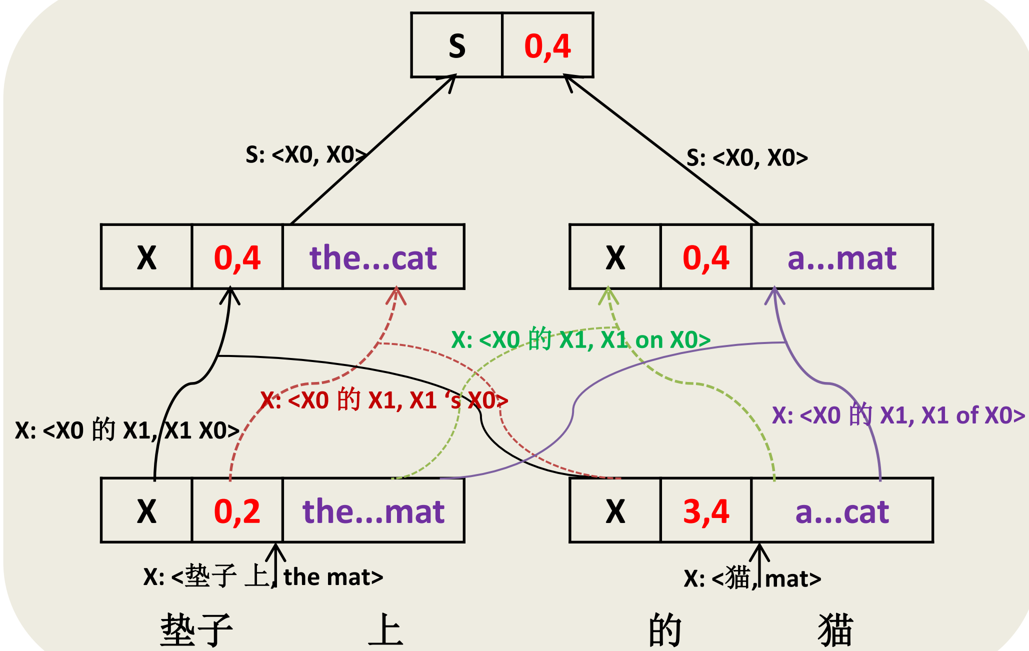
the mat a cat

a cat of the mat

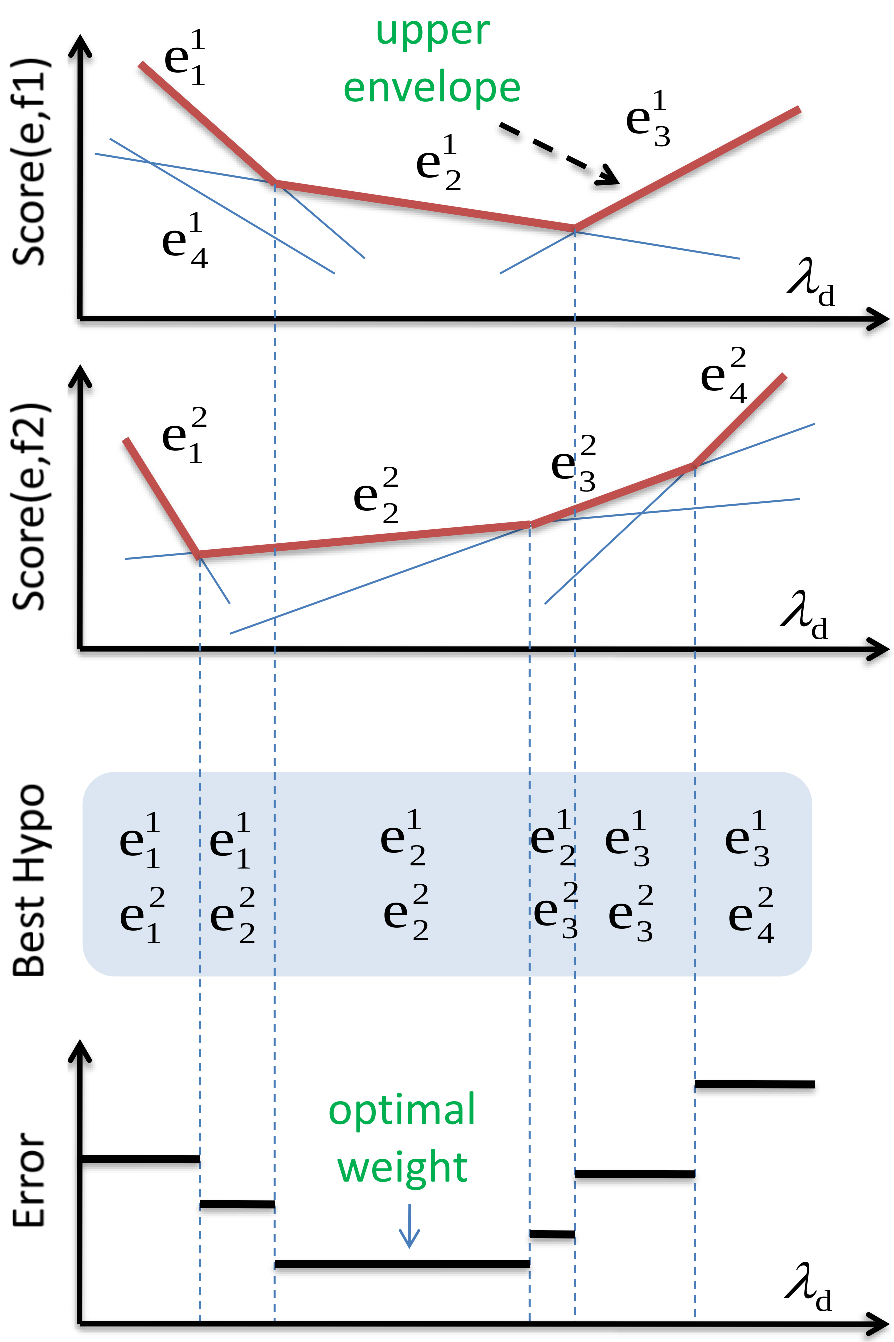
the mat 's a cat

垫子 上 的 猫

Hypergraph



垫子 上 的 猫



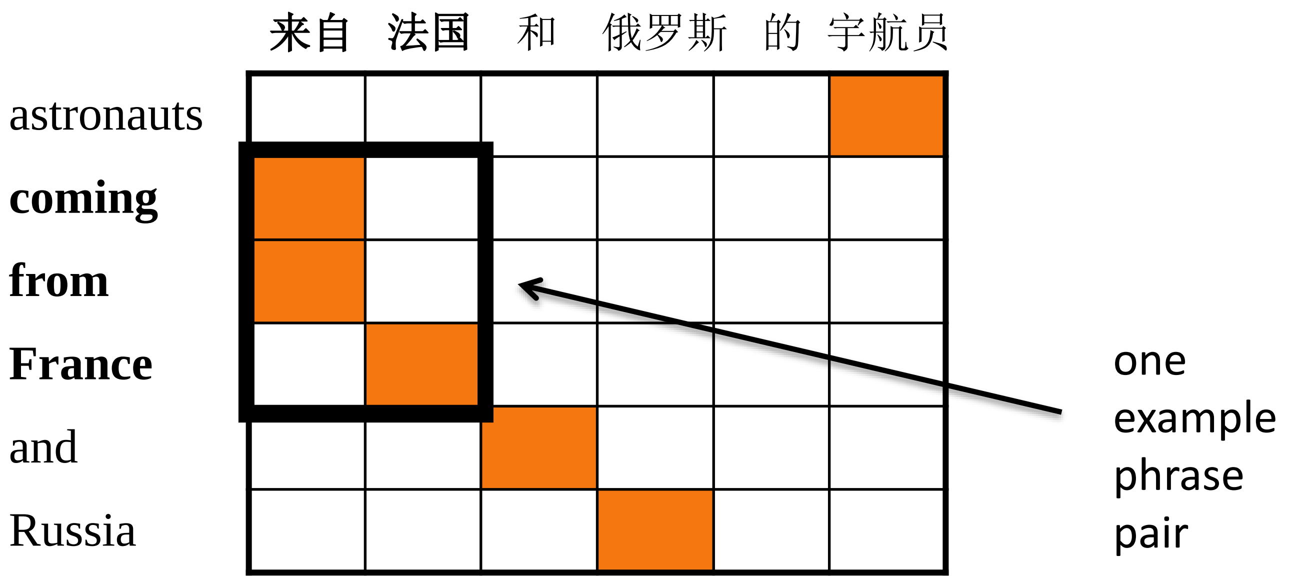
Score(e,f1)
upper envelope
e^1_1
e^1_2
e^1_3
e^1_4
λ_d
Score(e,f2)
e^2_1
e^2_2
e^2_3
e^2_4
λ_d
Best Hypo
e^1_1 e^2_1
e^1_1 e^2_2
e^1_2 e^2_2
e^1_2 e^2_3
e^1_3 e^2_3
e^1_3 e^2_4
Error
optimal weight
λ_d

IBM Model-4

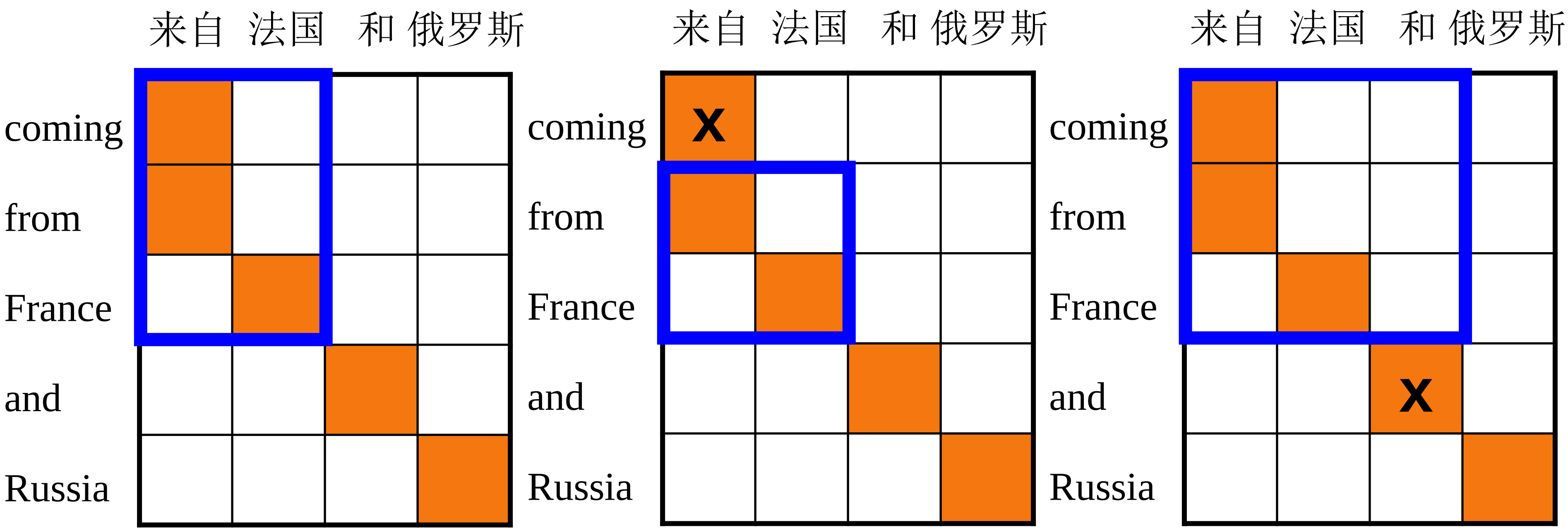
这 7 人 包括 来自 法国 和 俄罗斯 的 宇航员
这 7 人 包括 来自 来自 法国 和 俄罗斯 宇航员
这 7 人 包括 来自 来自 法国 和 俄罗斯 NULL 宇航员
These 7 people include coming from France and Russia the astronauts
These 7 people include the astronauts coming from France and Russia
n(2|来自)
p-null
p(astronauts|宇航员)
d(5-4|class(the), class(astronauts))

How to Learn the Phrase Translation Table?

- Collect all phrase pairs *that are consistent with the word alignment*



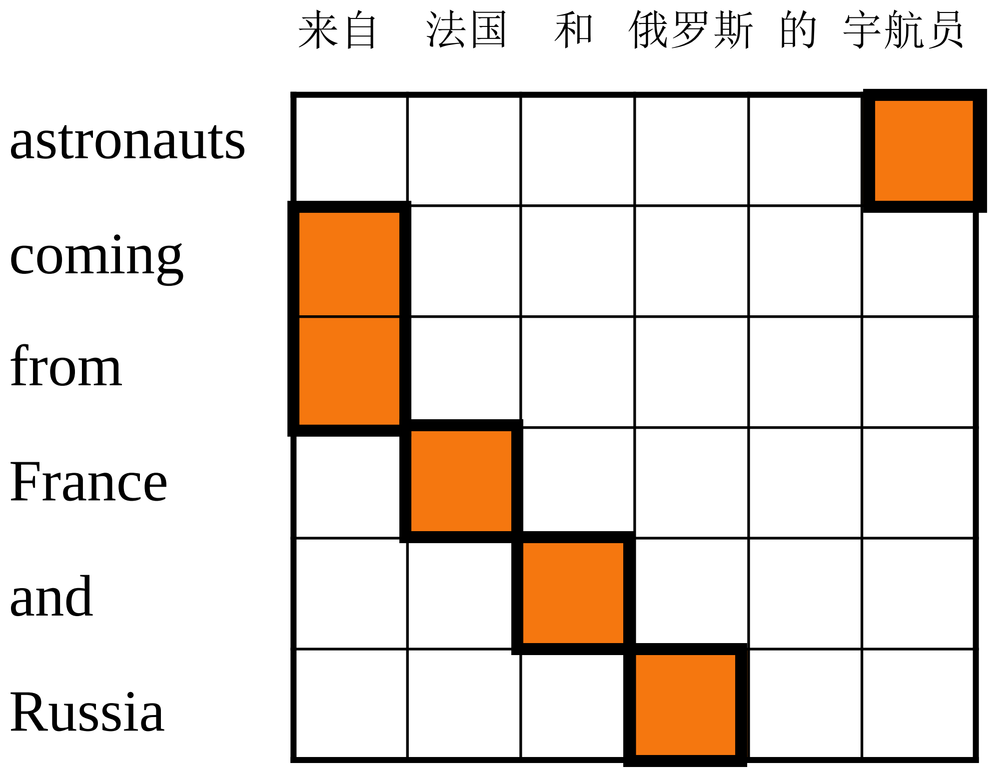
Consistent with Word Alignment



consistent inconsistent inconsistent

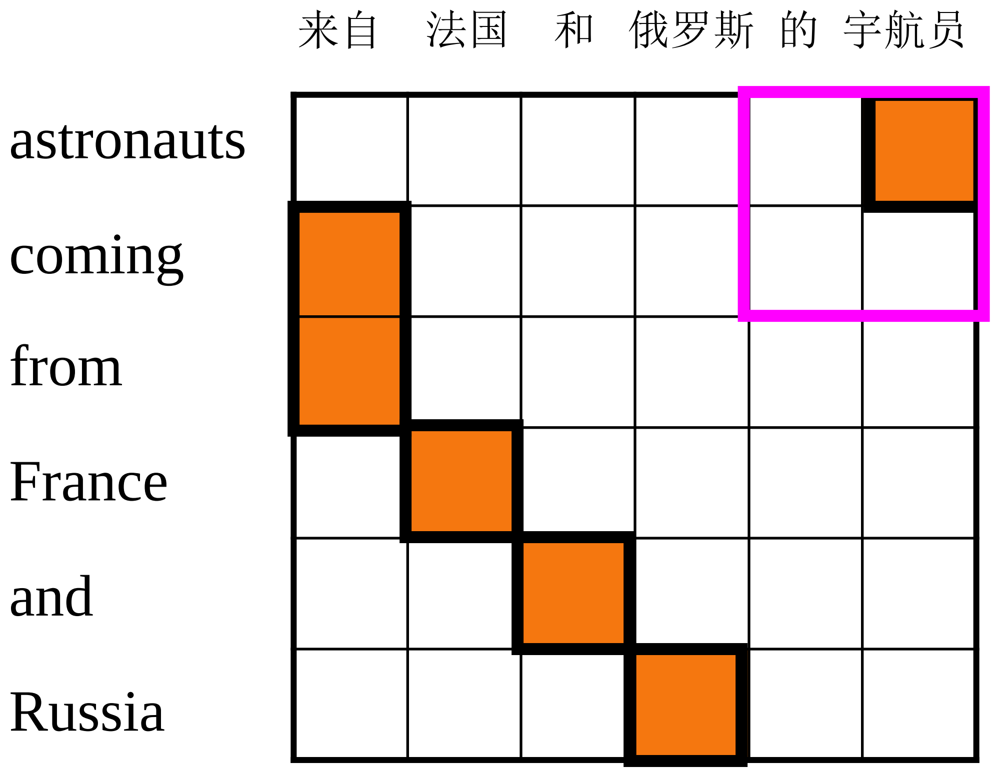
Phrase alignment must contain all alignment points for all the words in both phrases!

Word Alignment Induced Phrases



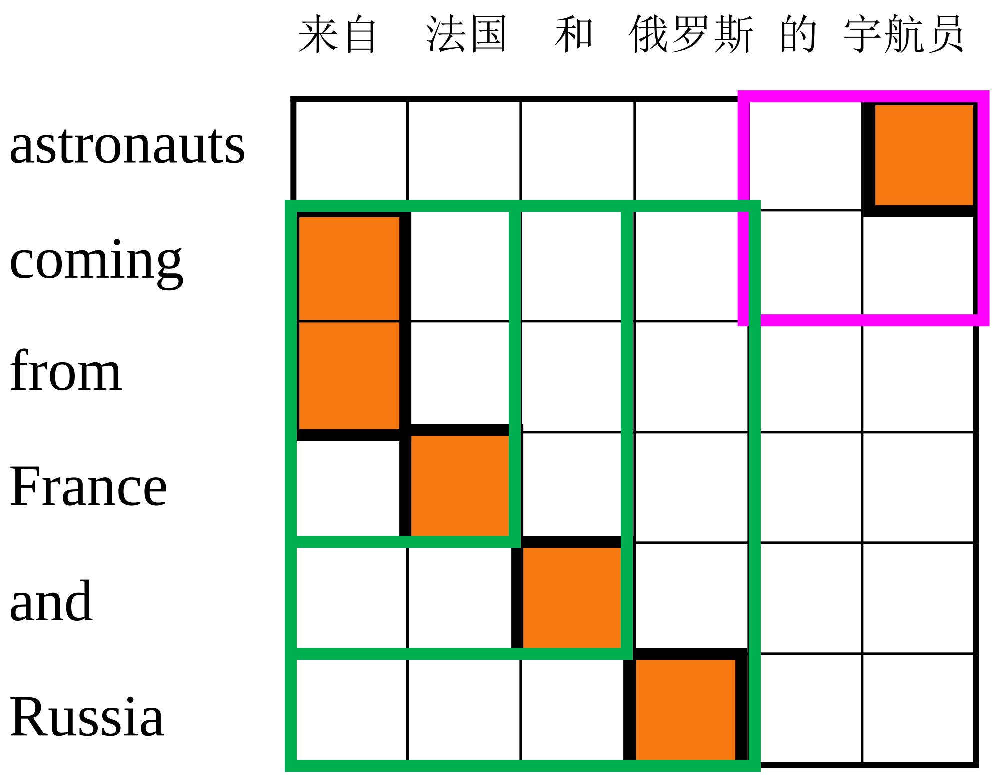
(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

Word Alignment Induced Phrases



(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)
(的 宇航员, astronauts) ...

Word Alignment Induced Phrases



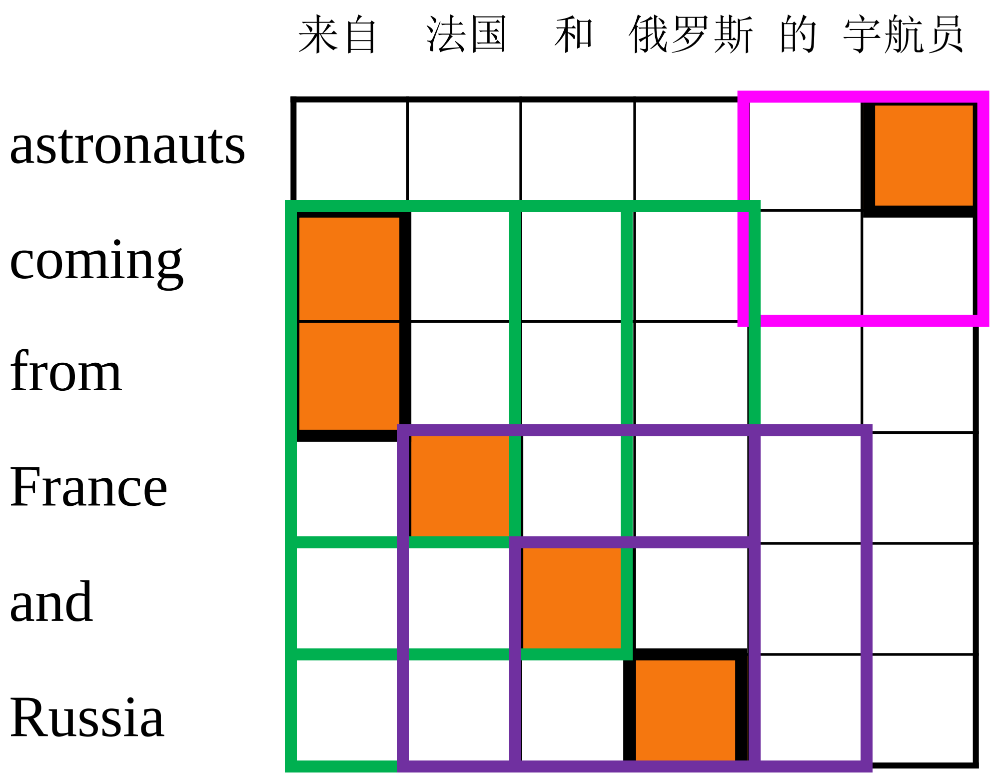
(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

(的 宇航员, astronauts) …

(来自 法国, coming from France) (来自 法国 和, coming from France and)

(来自 法国 和 俄罗斯, coming from France and Russia) …

Word Alignment Induced Phrases



(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

(的 宇航员, astronauts) …

(来自 法国, coming from France) (来自 法国 和, coming from France and)

(来自 法国 和 俄罗斯, coming from France and Russia) …

(和 俄罗斯, and Russia) (法国 和 俄罗斯, France and Russia)

(法国 和 俄罗斯 的, France and Russia) …