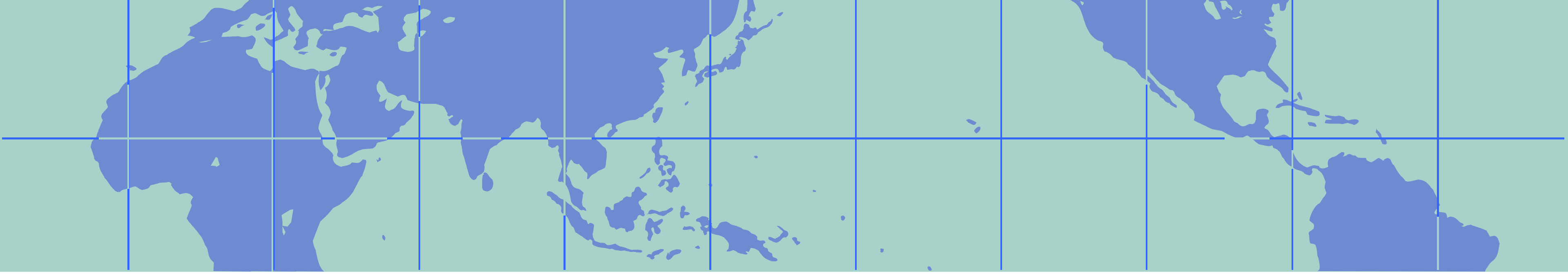
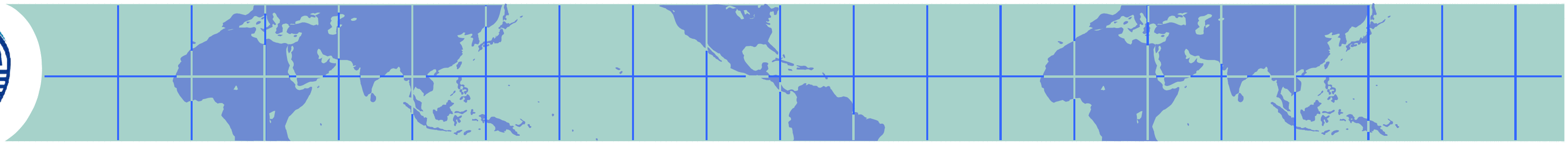


The ICT Statistical Machine Translation Systems for IWSLT 2007

Zhongjun He, Haitao Mi, Yang Liu, Devi Xiong, Weihua Luo, Yun Huang, Zhixiang Ren, Yajuan Lu, Qun Liu

Institute of Computing Technology
Chinese Academy of Sciences

2007.09.15– 2007.08.16

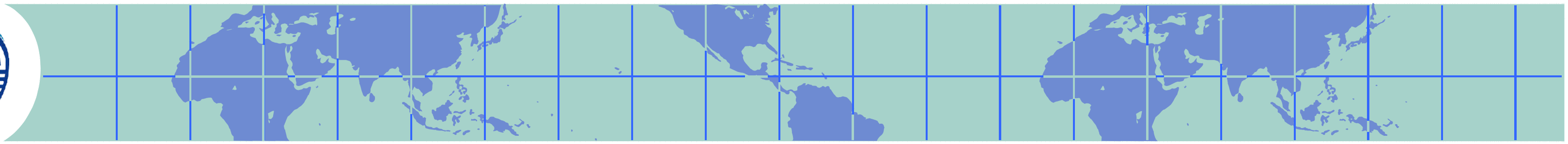


Outline

- Overview
- MT Systems
  - Bruin
  - Confucius
  - Lynx
- Official Evaluation
- Discussion
- Summary

Introduction of Our Group

- Multilingual Interaction Technology Laboratory, Institute of Computing Technology, Chinese Academy Sciences
- Long history for working on MT
  - Rule-based
  - Example-based
- Focus on SMT from 2004
- Website: http://mtgroup.ict.ac.cn/



People Working on SMT at ICT

- **Staffs**
  - Qun Liu (Researcher)
  - Yajuan Lu (Associate Researcher)
  - Yang Liu (Associate Researcher)
  - Weihua Luo (Assistant Researcher)
- **PhD Students**
  - Zhongjun He
  - Haitao Mi
  - Jinsong Su
  - Yang Feng
- **Master Students**
  - Yun Huang
  - Wenbin Jiang
  - Zhixiang Ren
  - …

IWSLT 2007 Evaluation

- Chinese-English transcript translation task

Systems for IWSLT 2007 Evaluation

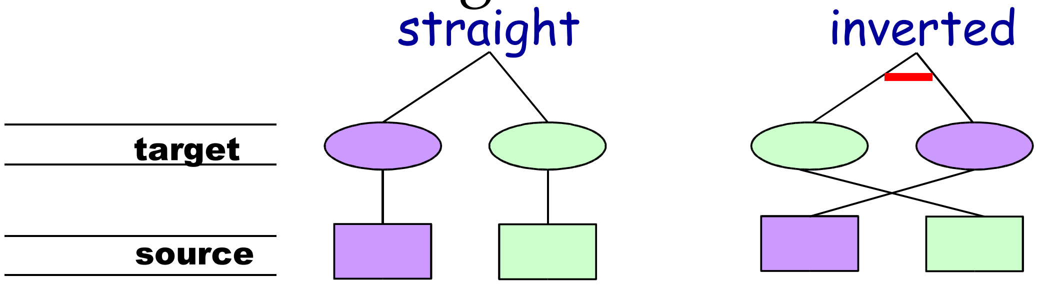
- MT Systems:
  - ***Bruin*** (formally syntax-based)
  - ***Confucius*** (extended phrase-based)
  - ***Lynx*** (linguistically syntax-based)

Outline

- Overview
- MT Systems
  - Bruin
  - Lynx
  - Confucius
- Official Evaluation
- Discussion
- Summary

Bruin

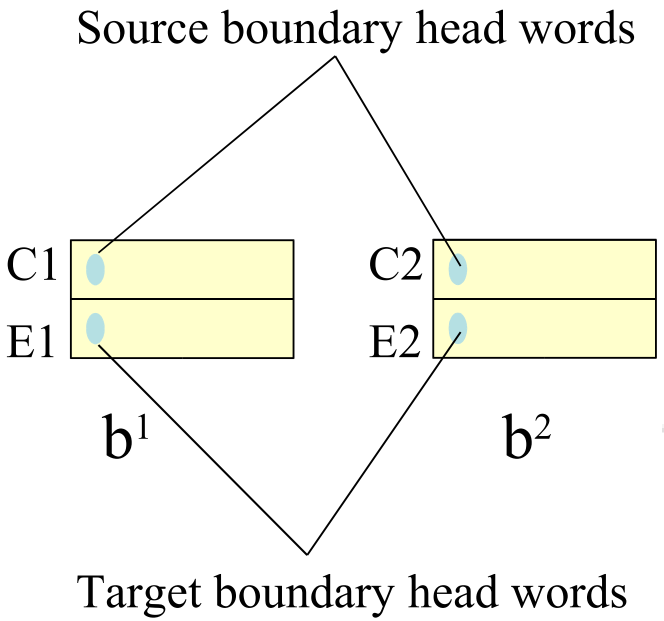
- Bruin is a formally syntax-based system
- MaxEnt Reordering Model build on BTG rules



- Regard reordering as a binary classification
  - Building a MaxEnt-based classifier
  - Using boundary words instead of whole phrases as features for the classifier

Features

- Source and target boundary words (lexical feature)
- Combinations of boundary words (collocation feature)



$$h_i(o,b^1,b^2) = \begin{cases} 1, & b^1.t_1 = E_1, o = O \\ 0, & otherwise \end{cases}$$

$$h_j(o,b^1,b^2) = \begin{cases} 1, & b^1.t_1 = E_1, b^2.t_1 = E_2, o = O \\ 0, & otherwise \end{cases}$$

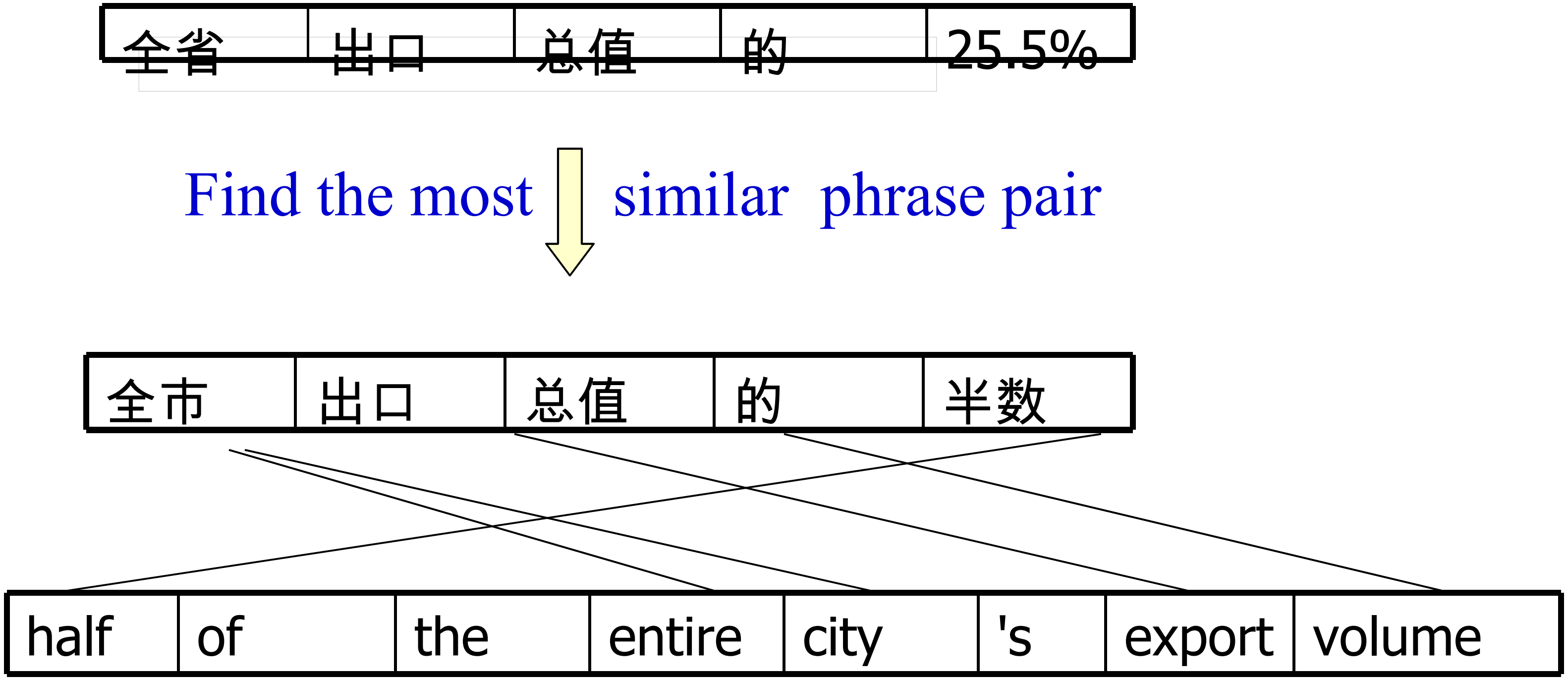
Training and Decoding

- Training the model
  - Learning reordering examples from bilingual word-aligned corpus
  - Generating features from reordering examples
  - Training a MaxEnt model on the features
- Decoding
  - CKY algorithm
- For details, see *Xiong et al., ACL2006*

Confucius

- An extended phrase-based system
- Log-linear model
- Monotone decoding
- We try a phrase-based similarity model, in which a translation for a certain source phrase can be applied for other similar phrases

Phrase-based Similarity Model



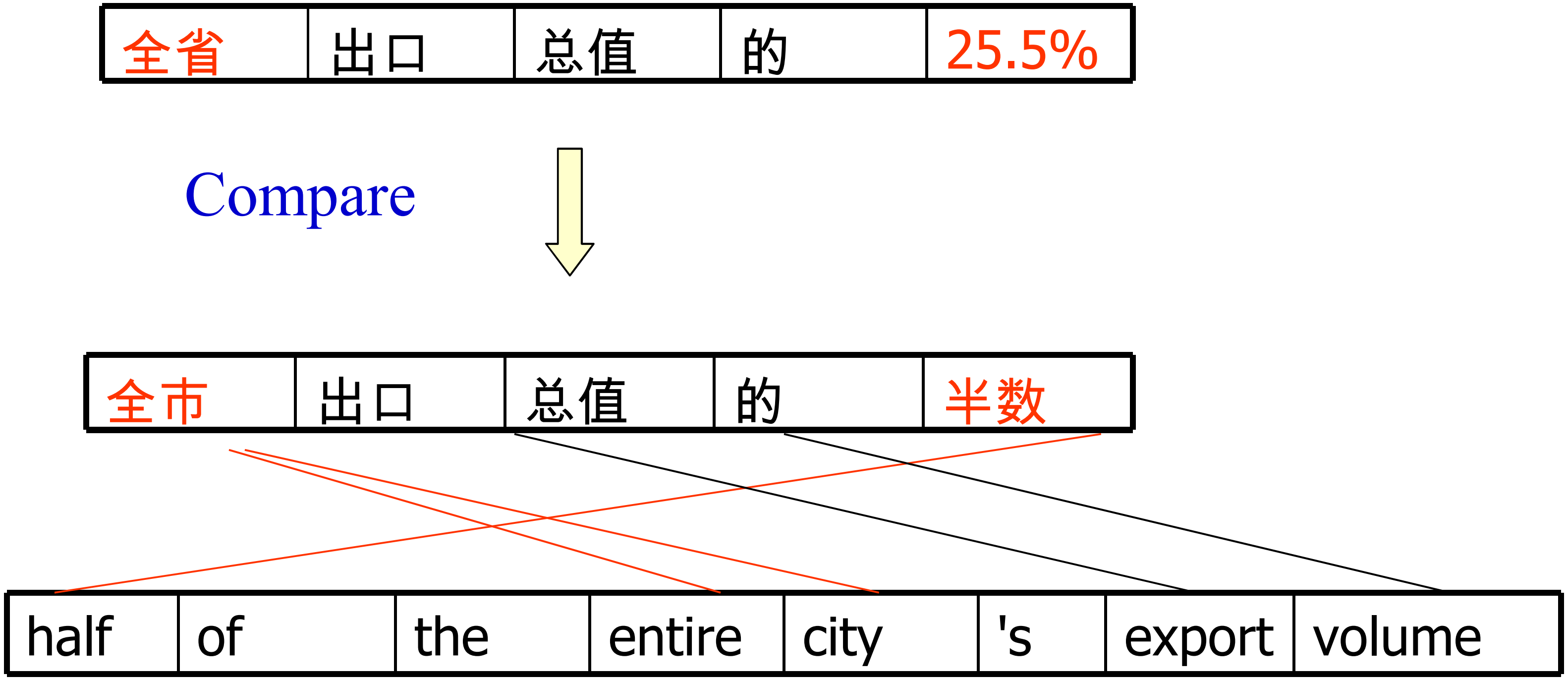
全省 出口 总值 的 25.5%

Find the most similar phrase pair

全市 出口 总值 的 半数

half of the entire city 's export volume

Phrase-based Similarity Model



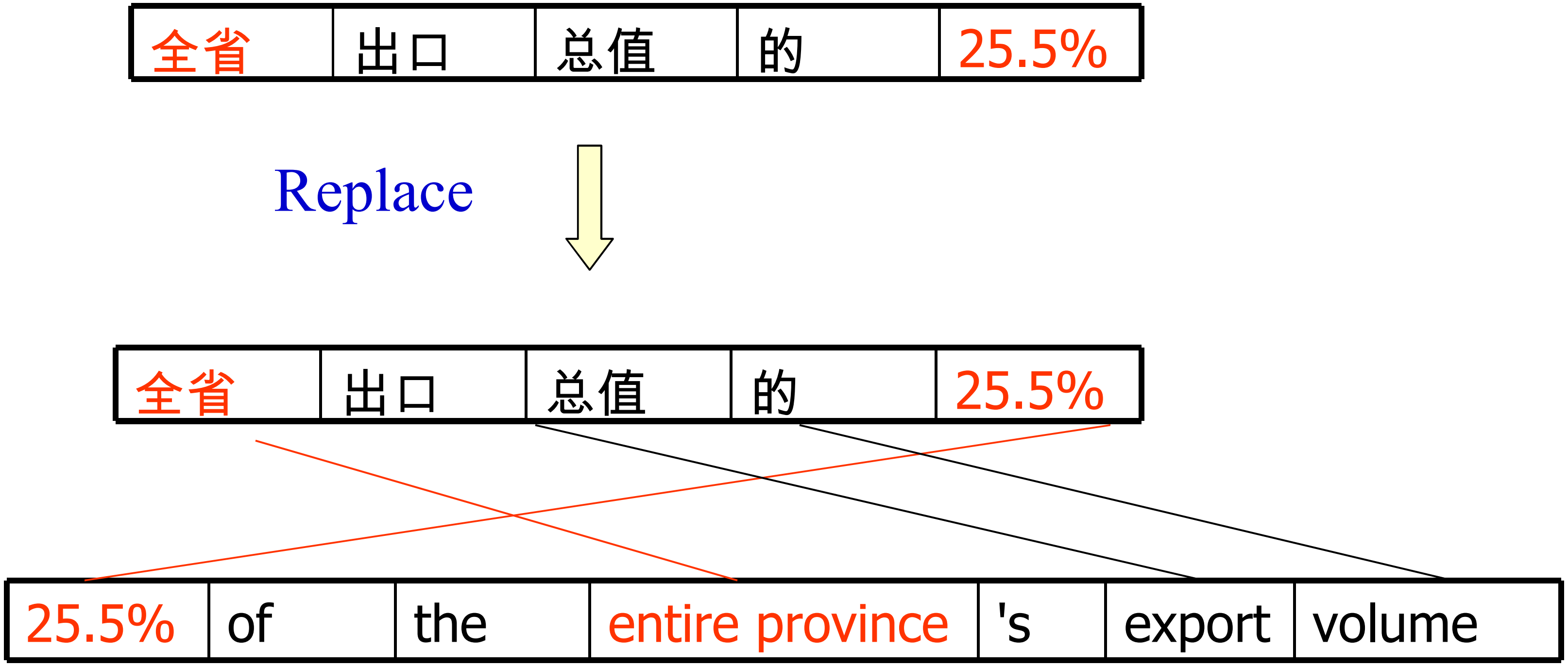
全省 出口 总值 的 25.5%

Compare

全市 出口 总值 的 半数

half of the entire city 's export volume

Phrase-based Similarity Model



全省 出口 总值 的 25.5%

Replace

全省 出口 总值 的 25.5%

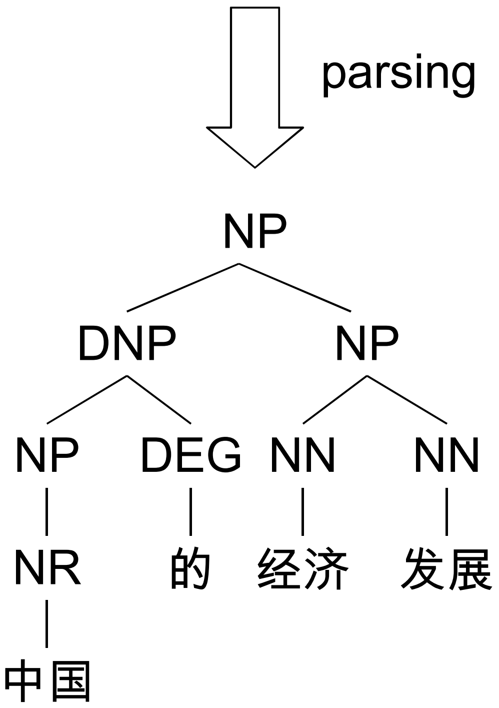
25.5% of the entire province 's export volume

Lynx

- A linguistically syntax-based system
- Based on tree-to-string alignment template (TAT), which map the source language tree to target language string
- Log-linear Model

Translation Process: Parsing

中国 的 经济 发展



parsing

NP

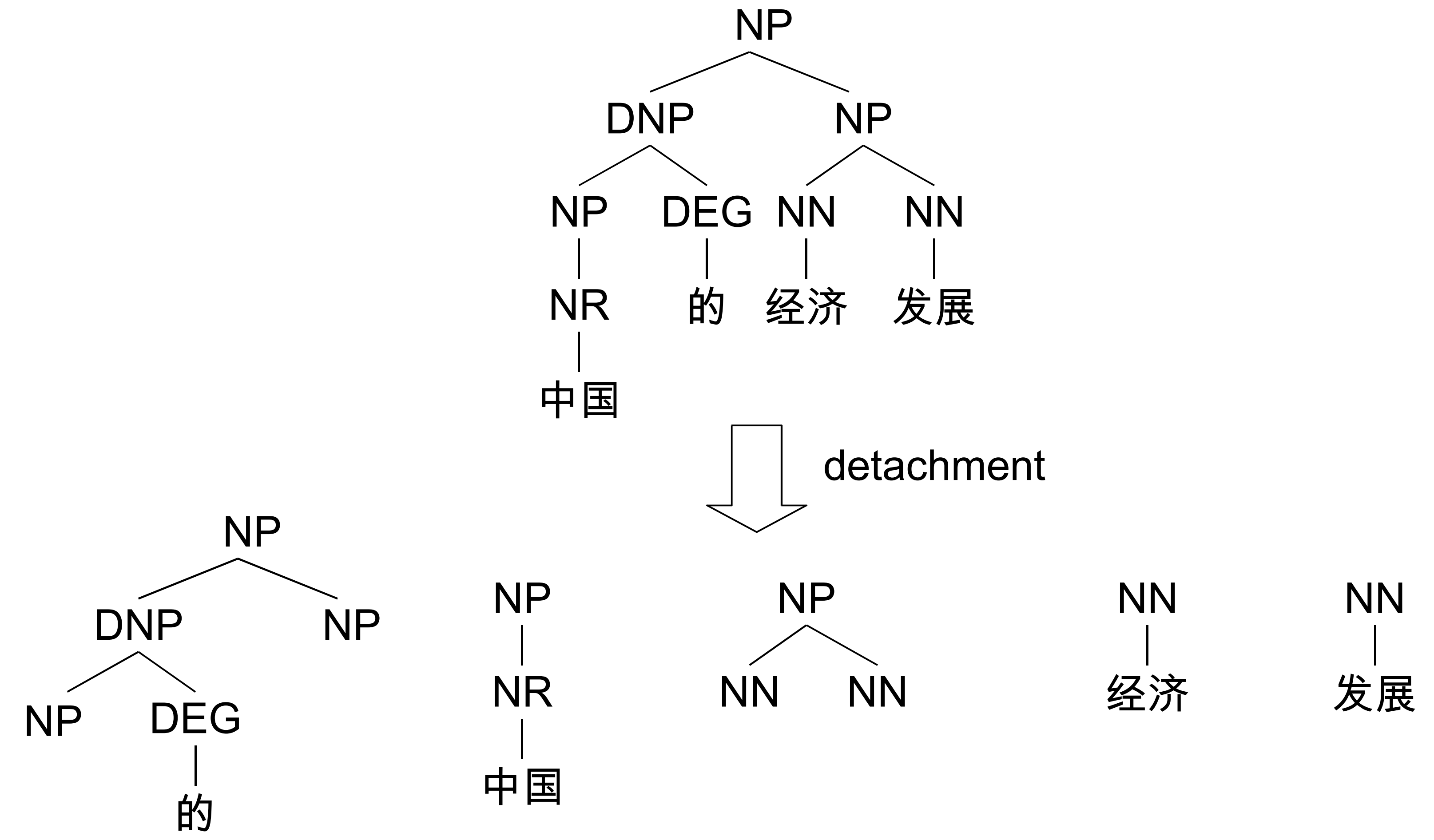
DNP NP

NP DEG NN NN

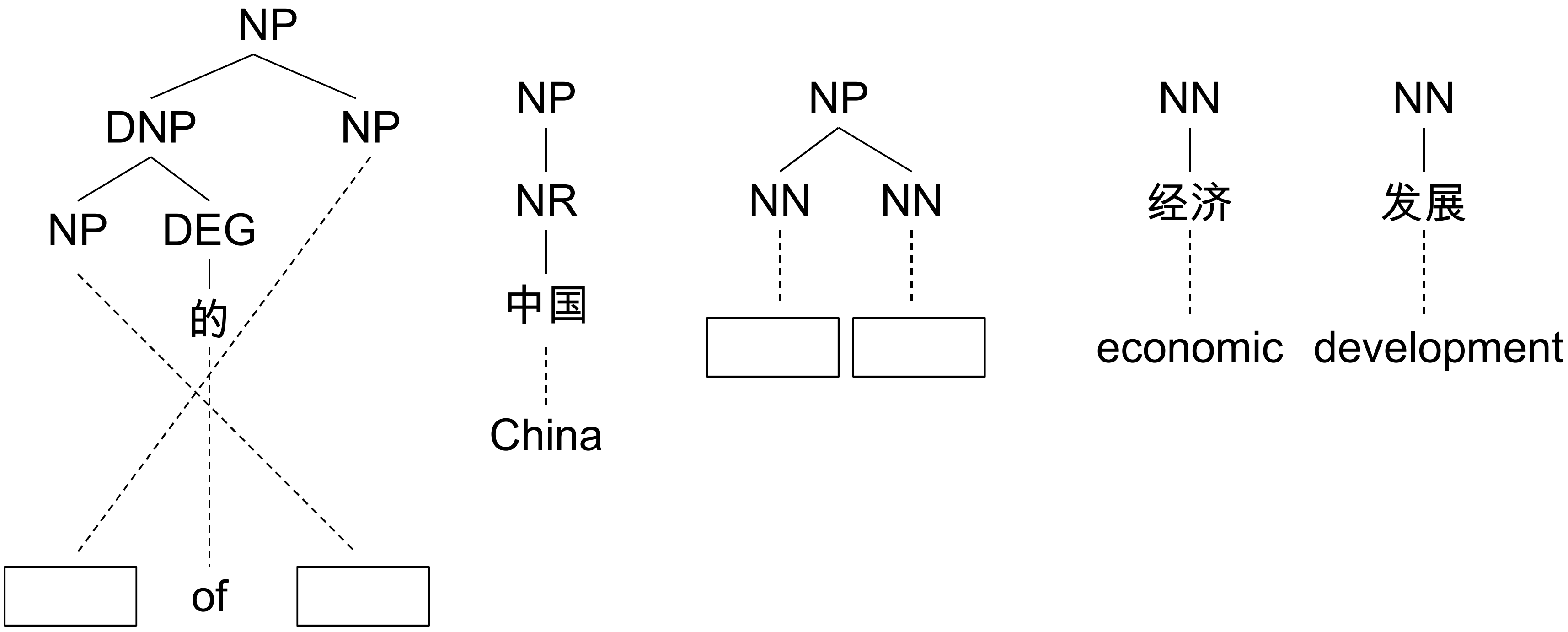
NR 的 经济 发展

中国

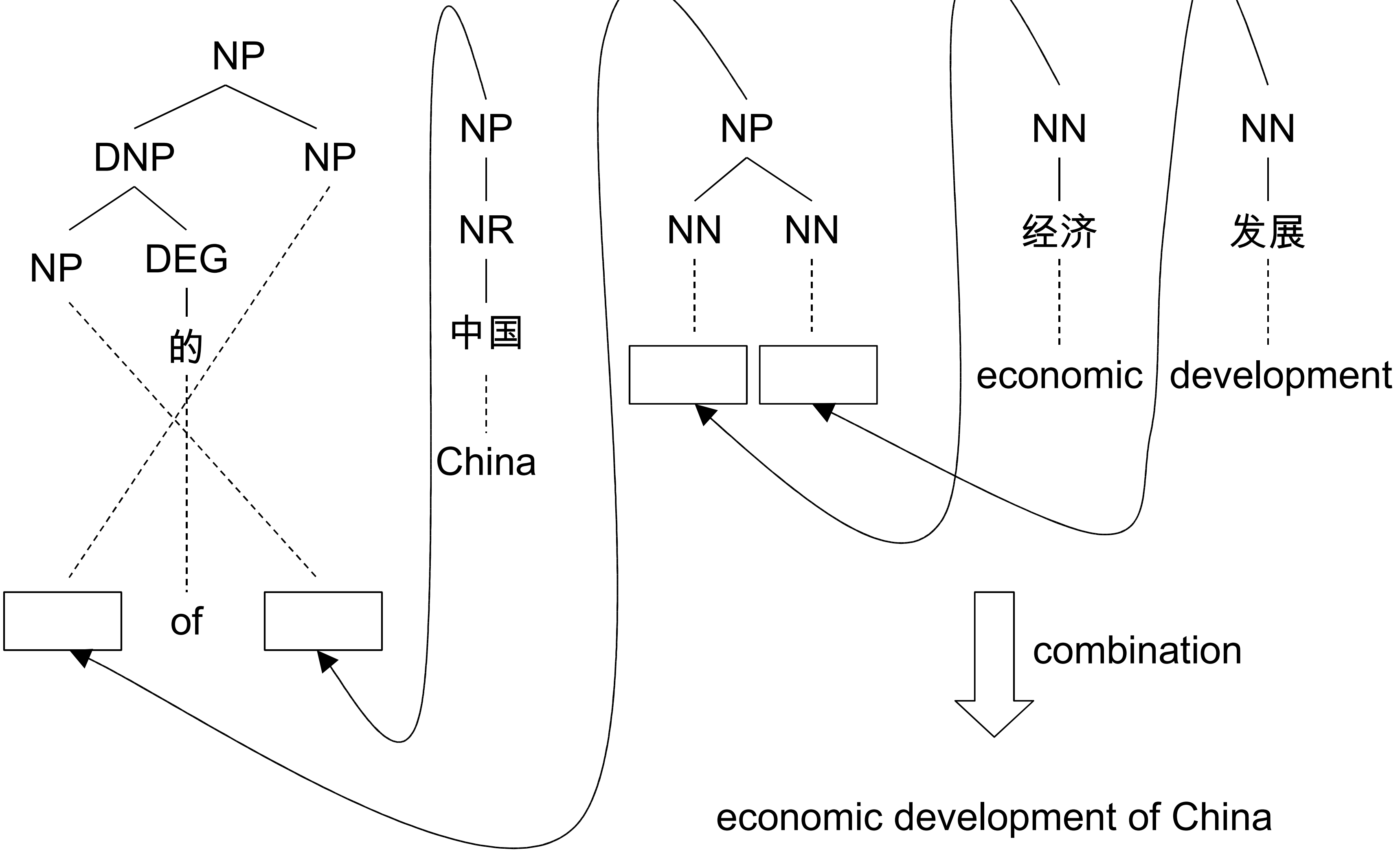
Translation Process: Detachment



Translation Process : Production



Translation Process : Combination



economic development of China

Training and Decoding

- Training
  - Extract TATs from word-aligned, source side parsed bilingual corpus using bottom-up strategy
  - Impose several restrictions to decrease the magnitude
- Decoding
  - bottom-up beam search
- For details, see *Liu et al., ACL2006*

Outline

- Overview
- MT Systems
  - Bruin
  - Lynx
  - Confucius
- Official Evaluation
- Discussion
- Summary

Toolkits Used

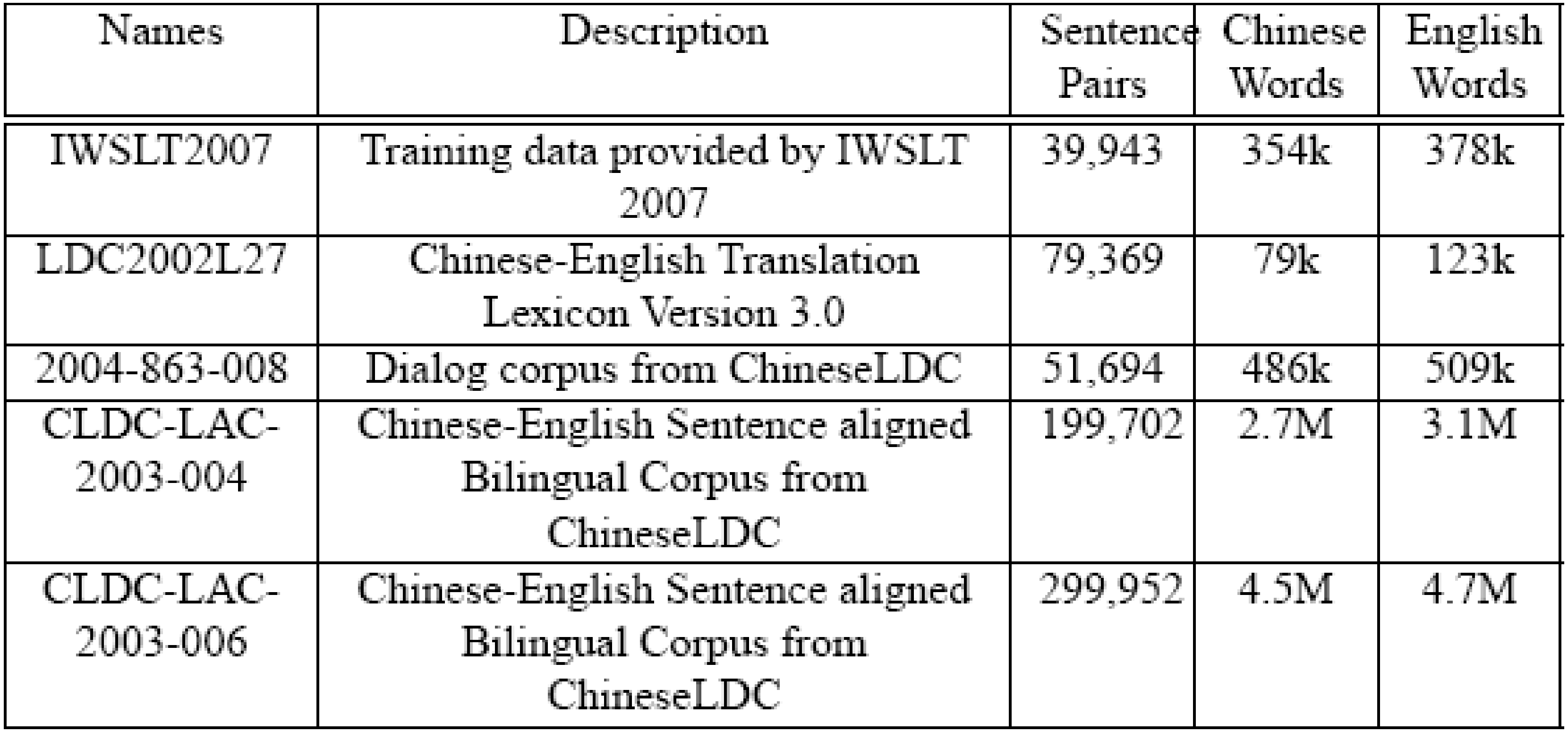
- Word alignment
  - GIZA++ plus “grow-diag-final” refinement method
- Language model
  - SRILM
- Chinese parser
  - Deyi Xiong’s

    A lexicalized PCFG model trained on PennTree bank
- Chinese word segmentation
  - ICTCLAS

Preprocessing and Postprocessing

- Preprocessing
  - Chinese word segmentation
  - Rule-based translations of numbers, dates and Chinese names
  - Chinese sentences Parsing (for Lynx only)
- Postprocessing
  - Remove unknown words
  - Capitalize the first word of each sentence

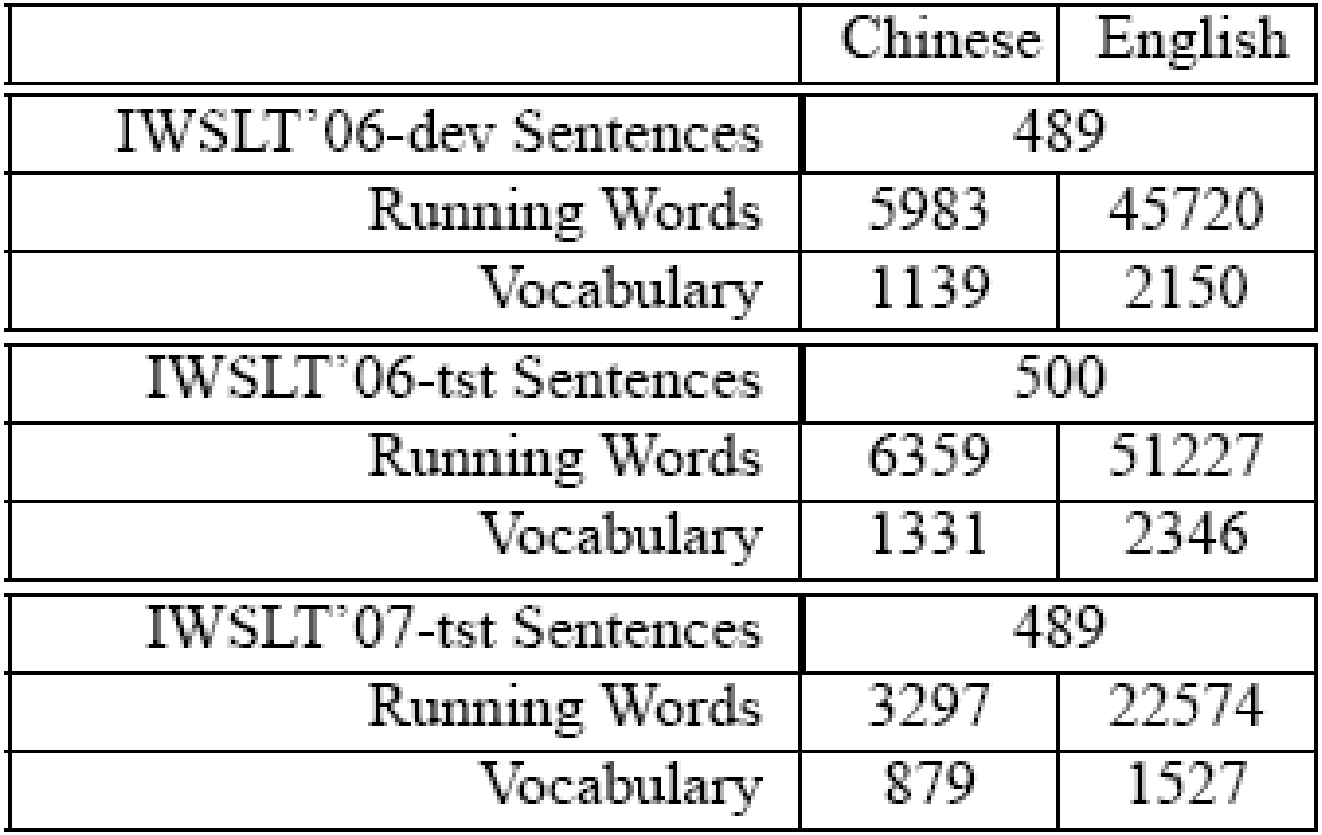
Training data



Names	Description	Sentence Pairs	Chinese Words	English Words
IWSLT2007	Training data provided by IWSLT 2007	39,943	354k	378k
LDC2002L27	Chinese-English Translation Lexicon Version 3.0	79,369	79k	123k
2004-863-008	Dialog corpus from ChineseLDC	51,694	486k	509k
CLDC-LAC-2003-004	Chinese-English Sentence aligned Bilingual Corpus from ChineseLDC	199,702	2.7M	3.1M
CLDC-LAC-2003-006	Chinese-English Sentence aligned Bilingual Corpus from ChineseLDC	299,952	4.5M	4.7M

Training Data List

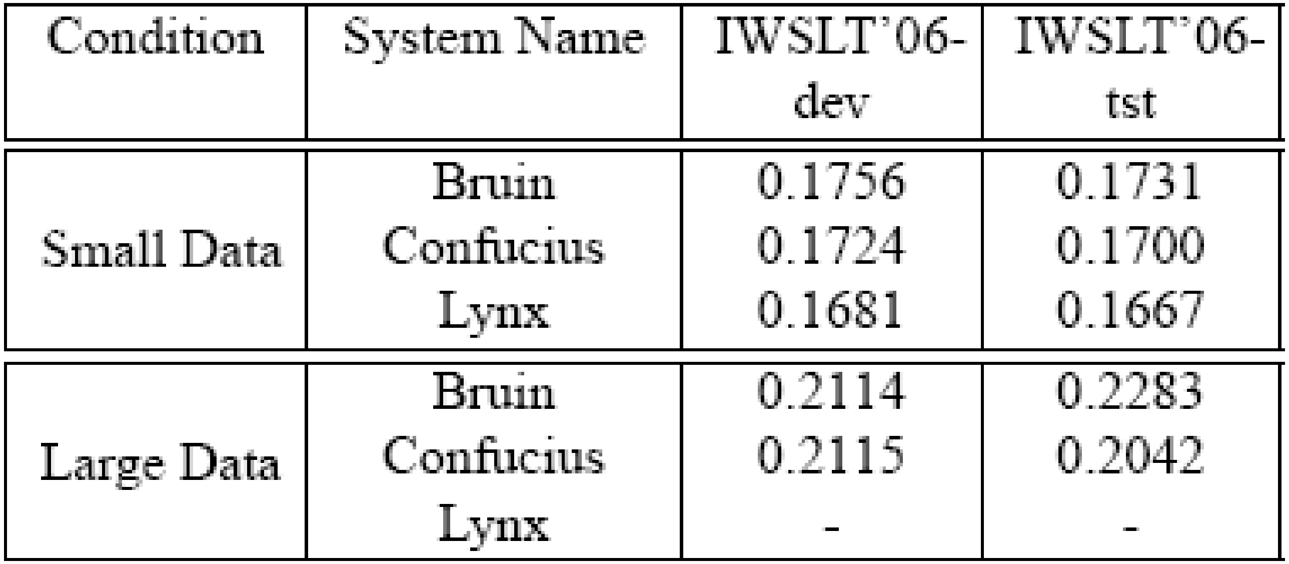
Development and test set



	Chinese	English
IWSLT'06-dev Sentences	489	
Running Words	5983	45720
Vocabulary	1139	2150
IWSLT'06-tst Sentences	500	
Running Words	6359	51227
Vocabulary	1331	2346
IWSLT'07-tst Sentences	489	
Running Words	3297	22574
Vocabulary	879	1527

Corpus statistics of the IWSLT 2006 and 2007 development and test set

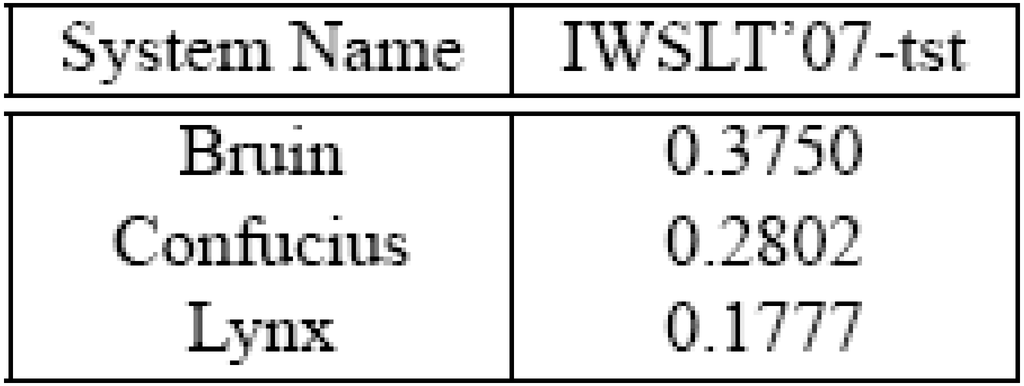
Results on IWSLT 2006 development set and test set



Condition	System Name	IWSLT'06-dev	IWSLT'06-tst
Small Data	Bruin	0.1756	0.1731
	Confucius	0.1724	0.1700
	Lynx	0.1681	0.1667
Large Data	Bruin	0.2114	0.2283
	Confucius	0.2115	0.2042
	Lynx	-	-

Small data: The training data released by the IWSLT 2007
Large data: All the training data

Results on IWSLT 2007 test set



System Name	IWSLT'07-tst
Bruin	0.3750
Confucius	0.2802
Lynx	0.1777

Outline

- Overview
- MT Systems
  - Bruin
  - Lynx
  - Confucius
- Official Evaluation
- Discussion
- Summary

Discussion

- **Lynx(0.1777)**
  - **Training Corpus:**
    - Training data:
      - About 39k sentence pairs **dialogs** data
        - Provided by IWSLT 2007
      - About 5M sentence pairs **newswire** data
        - Released by LDC
    - Domain is quite different
      - **Newswire** vs. **Dialogs**
    - Newswire data is too large

Discussion

- **Lynx (0.1777)**
  - **Parser :**
    - Trained on Penn Chinese Treebank
    - Domain is quite different too
      - **Newswire** vs. **Dialogs**
    - Parsing error (low performance of parser)
    - Lynx decoder
      - Only depends on the 1-best parsing tree

Discussion

- **Models:**
  - **Bruin (0.3750)**
  - **Confucius (0.2802)**

Discussion

- **Models:**
  - **Bruin (0.3750)**
    - **MaxEnt based reordering model**
    - **Long distance word reordering**
  - **Confucius (0.2802)**
    - **Monotone search**

Discussion

	2006 tst	2007 tst
Bruin	0.2283	0.3750
Confucius	0.2042	0.2802

- **Models:**
  - **Bruin (0.3750)**
    - **MaxEnt based reordering model**
    - **Long distance word reordering**
  - **Confucius (0.2802)**
    - **Monotone search**
  - **2007 test set (2006 test set)**

Discussion

	2006 tst	2007 tst
Bruin	0.2283	0.3750
Confucius	0.2042	0.2802

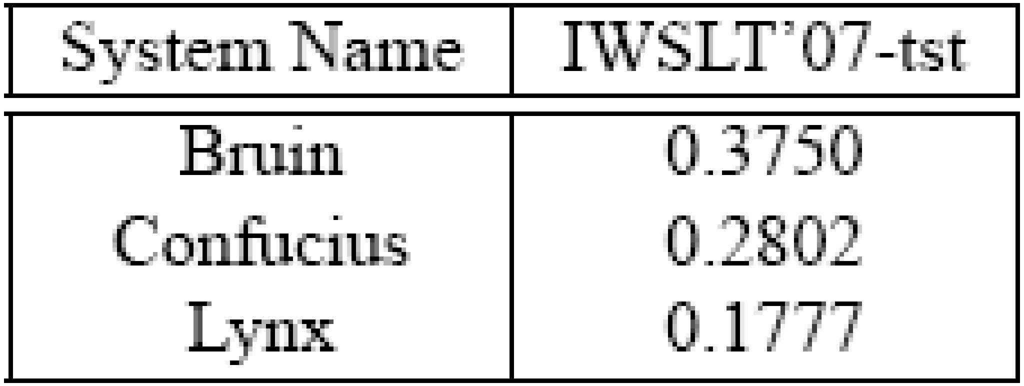
- **Models:**
  - **Bruin (0.3750)**
    - **MaxEnt based reordering model**
    - **Long distance word reordering**
  - **Confucius (0.2802)**
    - **Monotone search**
  - **2007 test set (2006 test set)**
    - **6.7words/sent (12.7words/sent)**
      - **Bruin will do better**

Discussion

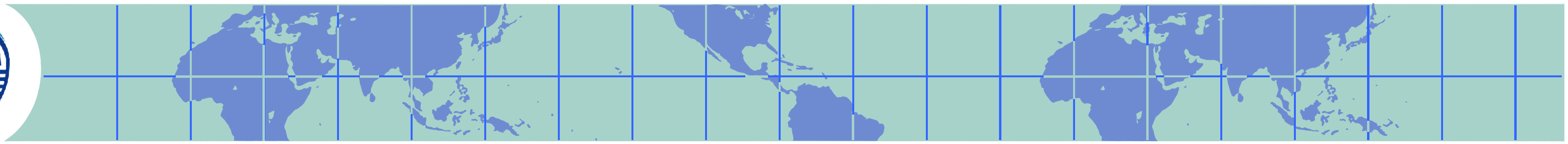
	2006 tst	2007 tst
Bruin	0.2283	0.3750
Confucius	0.2042	0.2802

- **Models:**
  - **Bruin (0.3750)**
    - **MaxEnt based reordering model**
    - **Long distance word reordering**
  - **Confucius (0.2802)**
    - **Monotone search**
  - **2007 test set (2006 test set)**
    - **6.7words/sent (12.7words/sent)**
      - **Bruin will do better**
    - **Punctuation marks (no)**
      - **More positive reordering information**
      - **Bruin will do better**

Results on IWSLT 2007 test set



System Name	IWSLT'07-tst
Bruin	0.3750
Confucius	0.2802
Lynx	0.1777



Outline

- Overview
- Systems
  - Bruin
  - Lynx
  - Confucius
- Official Evaluation
- Discussion
- Summary

Summary

- MT
  - 3 systems based on different translation models:
    - MaxEnt BTG Model
    - TAT model
    - Phrase-based Similarity Model
- Future Work
  - More new model
  - System combination

References

- Yang Liu, Qun Liu, and Shouxun Lin. 2006. **Tree-to-String Alignment Template for Statistical Machine Translation**. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 609-616, Sydney, Australia, July.
- Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. **Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation** . In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 521-528, Sydney, Australia, July.

Thanks!