Technical Report of NEUNLPLab System for CWMT08

Click to edit Master subtitle style

Xiao Tong, Chen Rushan, Li Tianning, Ren Feiliang,
Zhang Zhuyu, Zhu Jingbo, Wang Huizhen
xiaotong@mail.neu.edu.cn
http://www.nlplab.com

Outline

- Overview
- System description
  - Basic MT system
  - Systems for CWMT08
- Data
- Experiment
- Summary

Outline

- Overview
- System description
  - Basic MT system
  - Systems for CWMT08
- Data
- Experiment
- Summary

Our group

- Natural Language Processing Laboratory, College of information science and engineering, Northeastern University
- Long history for working on a variety of problems related to machine translation including
  - Multi-language MT
  - Rule-based MT
  - Example-based MT
- Our research work on SMT started in late 2007
- Welcome to our homepage http://www.nlplab.com

People (SMT) at NEUNLPLab

- Faculty
  - Zhu Jingbo / 朱靖波 (Professor)
  - Ren Feiliang / 任飞亮 (Lecturer)
  - Wang Huizhen / 王会珍 (Lecturer)
- PhD Students
  - Xiao Tong / 肖桐
- Master Students
  - Li Tianning / 李天宁
  - Zhang Zhuyu / 张祝玉
  - Chen Rushan / 陈如山

Task of CWMT08

- Four sub-tasks
  - Chinese -> English News
  - English -> Chinese News
  - English -> Chinese Science and Technology
  - System combination of Chinese -> English News
- We participated in
  - Sub-task1 (2 systems)
  - Sub-task2 (2 systems)

Outline

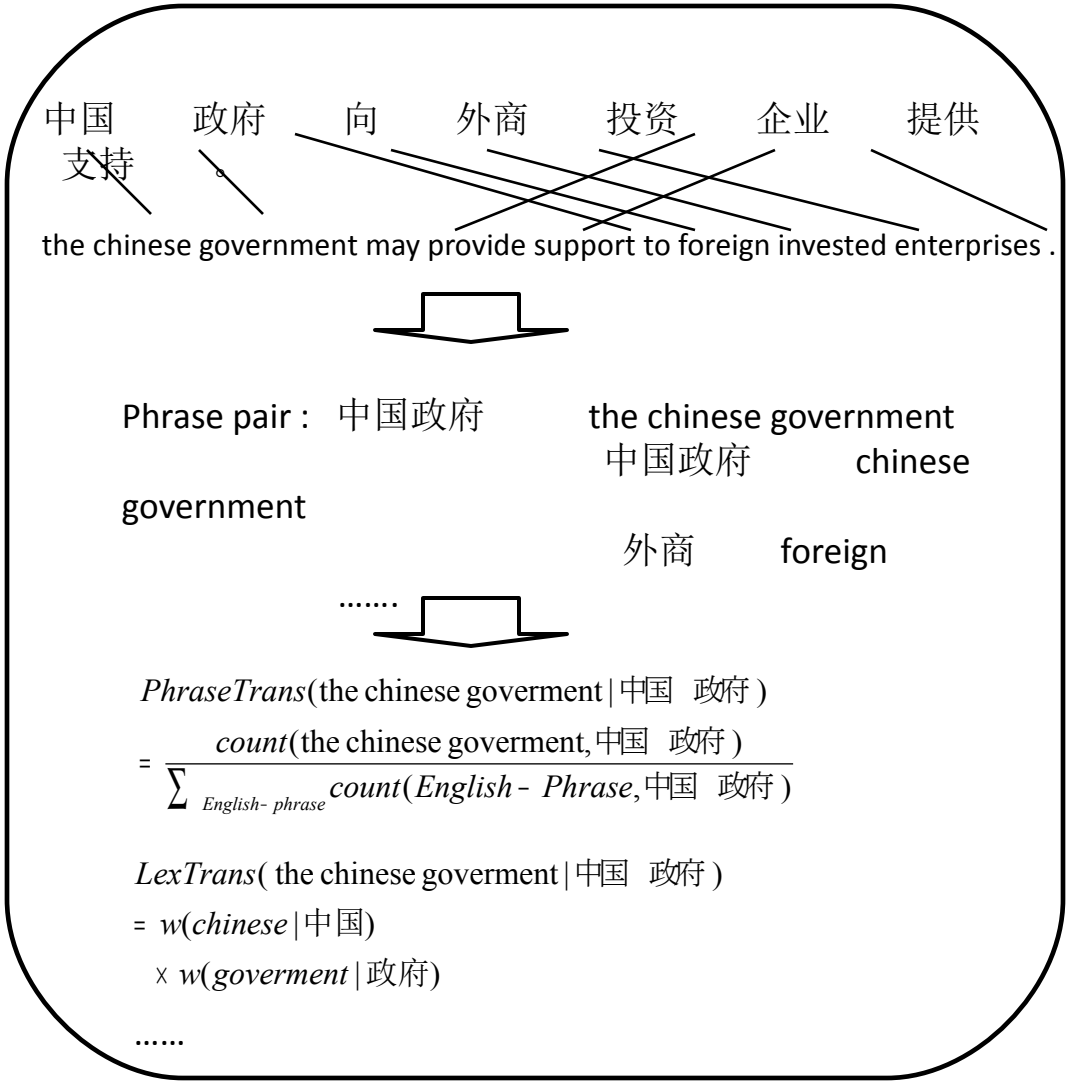
- Overview
- System description
  - Basic MT system
  - Systems for CWMT08
- Data
- Experiment
- Summary

System description

- Our basic system is a state-of-the-art Phrase-based statistical machine translation system.
- Characters of our system
  - Phrase-based SMT (Philipp et al, 2003)
  - Log-linear model (Och and Ney, 2002)
  - 7 features
- The major part (decoder) of our system is Moses (http://www.statmt.org/moses/)

Features – part1

- Feature
  - phrase translation probability
  - inverse phrase translation probability
  - lexical translation probability
  - inverse lexical translation probability
  - language model
  - sentence length penalty
  - MSD reordering model



Features – part2

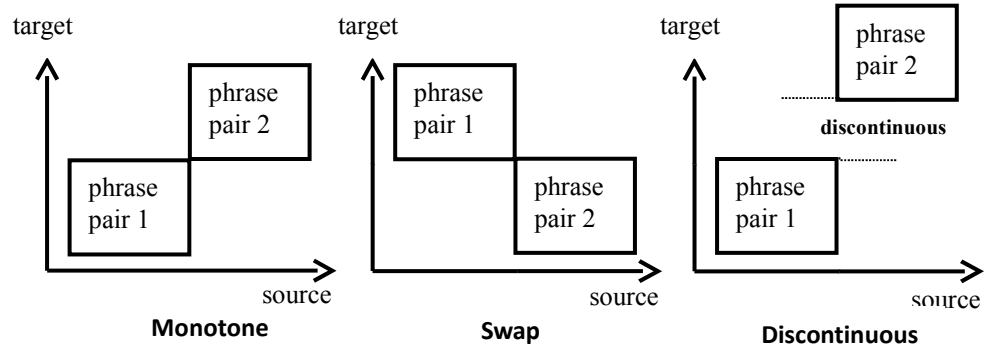
- Feature
  - phrase translation probability
  - inverse phrase translation probability
  - lexical translation probability
  - inverse lexical translation probability
  - language model
  - sentence length penalty
  - MSD reordering model

Pr(the chinese government may provide support …)
=Pr(the) ×
Pr(chinese | the) ×
Pr(government | chinese the) ×
… Pr(wi | wi-n+1 wi-2 … wi-1) …

- 5-gram (For both Chinese->English and English->Chinese sub-tasks)
- smoothing

Features – part3

- Feature
  - phrase translation probability
  - inverse phrase translation probability
  - lexical translation probability
  - inverse lexical translation probability
  - language model
  - sentence length penalty
  - MSD reordering model



- Reordering model
  - Three ways (Monotone, Swap and Discontinuous)
  - 6 sub-features

System

- Pre-processing
  - Chinese segmentation (our lab)
  - English tokenization (tokenizeE.perl.tmpl)
  - NE recognition of time, date and person (a rule-based system of our lab)
  - Remove case
- Word alignment
  - GIZA++ (http://code.google.com/p/giza-pp/)
  - Alignment symmetrization (Philipp et al, 2003)
- Phrase extraction and scoring (Philipp et al, 2003)
- Language model
  - SRILM (http://www.speech.sri.com/projects/srilm/)
- Decoding
  - Moses decoder (http://www.statmt.org/moses/)
  - NE translation (a rule-based system of our lab)
- Post-processing

Systems for CWMT08

- System 1 for Chinese->English News
  - Basic system
  - Limited data condition
- System 2 for Chinese->English News
  - Basic system
  - Large Language model (trained on NIST data)
- System 1 for English->Chinese News
  - Basic system
  - Limited data condition
- System 2 for English->Chinese News
  - Modified pro-processing strategy
  - Limited data condition

Outline

- Overview
- System description
  - Basic MT system
  - Systems for CWMT08
- Data
- Experiment
- Summary

Data

- Data provided within the CWMT08 evaluation tasks

	Chinese	English
Sentence	849930	
Word	9977500	10997208
Vocabulary	105239	112160

- Data used to train LM (C->E system2)
  - The English side of the LDC parallel corpus + gigaword-xinhua

    LDC2003E14, LDC2004T08, LDC2005T06, LDC2004T07, LDC2003E07, LDC2004E12, LDC2007T07

Outline

- Overview
- System description
  - Basic MT system
  - Systems for CWMT08
- Data
- Experiment
- Summary

Experimental results

- Environment

  Xeon(TM) 3.00GHz + 32GB memory + Linux

- CWMT08 Chinese->English News

system	time	BLEU4	NIST5	GTM	mWER	mPER	ICT
System1	2:49:15	0.2033	7.2819	0.6836	0.7262	0.5274	0.3220
System2	3:30:01	0.2331	7.6770	0.6968	0.7159	0.5178	0.3367

- CWMT08 English->Chinese News

system	time	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
System1	2:54:03	0.2408	0.1838	7.5465	7.5504	0.7101	0.6851	0.4566	0.3564
System2	2:55:32	0.2417	0.1847	7.6279	7.6319	0.7074	0.6866	0.4567	0.3482

Outline

- Overview
- System description
  - Basic MT system
  - Systems for CWMT08
- Data
- Experiment
- Summary

Summary

- We built
  - Two systems for Chinese->English News
  - Two systems for English->Chinese News
- Problems of our system
  - Word alignment
  - Reordering
- In future
  - Syntax-based SMT
  - System combination

Thank you