The XMU SMT System for IWSLT 2007

Yidong Chen, Xiaodong Shi, Changle Zhou
Department of Cognitive Science
School of Information Sciences and Technologies
Xiamen University, P. R. China
{ydchen, mandel, dozero}@xmu.edu.cn

16 October 2007 - Trento

XIAMEN UNIVERSITY

Outline

- **Overview**
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Reordering of the Source Sentences
  - Dealing with the Unknown Words
- Experiments
- Conclusions

Overview

- Who we are?
  - NLP group at Institute of Artificial Intelligence, Xiamen University
  - Begin research on SMT since 2004
  - Have worked on rule-based MT for more than 15 years
  - First web MT in China (1999)
  - First mobile phone MT in China (2006)
  - Website: http://ai.xmu.edu.cn/
    http://mt.xmu.edu.cn
    http://nlp.xmu.edu.cn


NOKIA
Neon
我们用蓝牙传输文件
We transmit a file with
blue tooth
汉语互译
退出
选择
取消
Neon
6:38
Neon
欢迎来到厦门大学
英汉互译
Welcome to Xiamen university
编辑 打开
iPAQ
KONKA
翻译结果
this is the first tr
anslation system on
a mobile phone
这是手机上的第一个翻
译系统
保存
返回
Just agree to stop your horse?
It is we reply without method only.
You can have seen desert blossoms.
You can have seen glacier melts.
Had not you seen?
Had not seen!
So wandering traveller,
He never also can stop!
malloc: 16945, free: 16945
MP3
SHIFT

Overview (Cont.)

- IWSLT 2007
  - We implemented a **phrase-based** statistical machine translation system.
  - We incorporated a **reordering** model based on chunking and reordering of source language sentences.
  - We participated in the **open data track** for **Cleaned Transcripts** for the **Chinese-English translation direction**.

Outline

- Overview
- **Training**
- System
  - Translation Model
  - Parameters
  - Decoder
  - Reordering of the Source Sentences
  - Dealing with the Unknown Words
- Experiments
- Conclusions

Training

- Preprocessing (Chinese part)
  - Segmentation
  - Mixed (DBC/SBC) case to SBC case
- Preprocessing (English part)
  - Tokenization
  - Truecasing of the first word of an English sentence

Training (Cont.)

- Word Alignment
  - Firstly, we ran **GIZA++** up to IBM model 4 in **both translation directions** to get an initial word alignment.
  - Then, We applied "**grow-diag-final**" method (Koehn, 2003) to refine it and achieve n-to-n word alignment.

XIAMEN UNIVERSITY

Training (Cont.)

- Reordering of the Training Set (Chinese Part)
  - We used an algorithm similar to **selection sort algorithm** to perform the reordering.
  - We regard the chunk reordering problem as a problem of finding a permutation of the chunks that is the best one according to the target language order, and thus is similar to the problem of **sorting**, whose aim is finding a permutation of a given integer sequence so that the integers are in ascending or descending order.
  - The word alignment matrix is used as a clue for how a Chinese chunk sequence should be reordered.

Training (Cont.)

- Phrase Extraction
  - A similar way to (Och, 2002).
  - We limited the length of phrases from 1 word to 6 words.
  - For a Chinese phrase, only 20-best corresponding bilingual phrases were kept. $\sum_{i=1}^{N}\lambda_i \cdot h_i(\tilde{e},\tilde{c})$ is used to evaluate and rank the bilingual phrases with the same Chinese phrase.

Training (Cont.)

- Phrase Probabilities
  - Phrase translation probability $p(\tilde{e} \mid \tilde{c})$
  - Inversed phrase translation probability $p(\tilde{c} \mid \tilde{e})$
  - Phrase lexical weight $lex(\tilde{e} \mid \tilde{c})$
  - Inversed phrase lexical weight $lex(\tilde{c} \mid \tilde{e})$
- $$p(\tilde{e} \mid \tilde{c}) = \frac{N(\tilde{e}, \tilde{c})}{\sum_{\tilde{e}'} N(\tilde{e}', \tilde{c})}$$
- $$lex(\tilde{e} \mid \tilde{c}) = lex(e_1^I \mid c_1^J, a) = \prod_{i=1}^{I} \frac{1}{|\{j \mid (i,j) \in a\}|} \sum_{\forall (i,j) \in a} p(c_i \mid e_j)$$

XIAMEN UNIVERSITY

Outline

- Overview
- Training
- System
  - **Translation Model**
  - Parameters
  - Decoder
  - Reordering of the Source Sentences
  - Dealing with the Unknown Words
- Experiments
- Conclusions

Translation Model

- We use a log-linear modeling (Och, 2002):

$$\Pr(e_1^I \mid c_1^J) = \frac{\exp[\sum_{m=1}^{M} \lambda_m \cdot h_m(e_1^I, c_1^J)]}{\sum_{e_1^I} \exp[\sum_{m=1}^{M} \lambda_m \cdot h_m(e_1^I, c_1^J)]}$$

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m \cdot h_m(e_1^I, c_1^J) \right\}$$

Translation Model (Cont.)

- Six features
  - Phrase translation probability $p(\tilde{e} | \tilde{c})$
  - Inversed phrase translation probability $p(\tilde{c} | \tilde{e})$
  - Phrase lexical weight $lex(\tilde{e} | \tilde{c})$
  - Inversed phrase lexical weight $lex(\tilde{c} | \tilde{e})$
  - English language model $lm(e_1^I)$
  - English sentence length penalty I

Outline

- Overview
- Training
- System
  - Translation Model
  - **Parameters**
  - Decoder
  - Reordering of the Source Sentences
  - Dealing with the Unknown Words
- Experiments
- Conclusions

Parameters

- We didn’t used discriminative training method to train the parameters. We adjust the parameters by hand.
- We didn’t readjust the parameters according to the develop sets provided in this evaluation. we simply used an empirical setting, with which our decoder achieved a good performance in translating the test set from the *2005 China’s National 863 MT Evaluation*.

Parameters (Cont.)

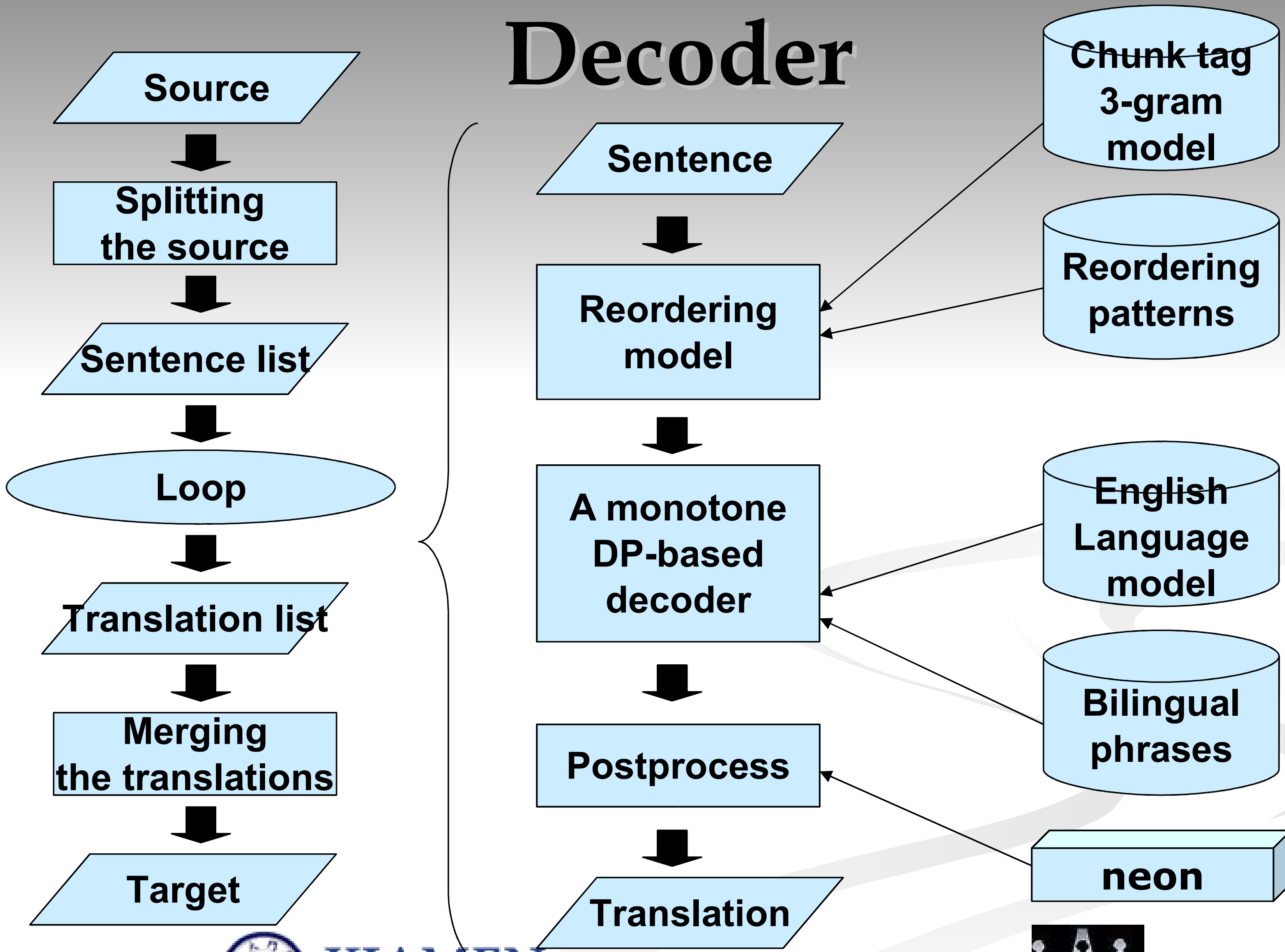
- The parameter settings for our system

Parameters	Corresponding Features	Values
λ_1	$p(\tilde{e} \mid \tilde{c})$	0.15
λ_2	$p(\tilde{c} \mid \tilde{e})$	0.03
λ_3	$lex(\tilde{e} \mid \tilde{c})$	0.16
λ_4	$lex(\tilde{c} \mid \tilde{e})$	0.03
λ_5	$lm(e_1^I)$	0.13
λ_6	I	0.48

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - **Decoder**
  - Reordering of the Source Sentences
  - Dealing with the Unknown Words
- Experiments
- Conclusions

Decoder


Source
Splitting the source
Sentence list
Loop
Translation list
Merging the translations
Target
Sentence
Reordering model
A monotone DP-based decoder
Postprocess
Translation
Chunk tag 3-gram model
Reordering patterns
English Language model
Bilingual phrases
neon

Decoder (Cont.)

- We used the monotone search in the decoding, similar to (Zens, 2002).
- Dynamic programming recursion:

$$Q(0,\$) = 1$$

$$Q(j,e) = \max_{\substack{0 \le j' < j \\ e', \tilde{e}}} \left\{ Q(j',e') + \sum_{m=1}^{M} \lambda_m \cdot h_m(\tilde{e}, c_{j'+1}^{j}) \right\}$$

$$Q(J+1,\$) = \max_{e'} \{ Q(J,e') + p(\$ \mid e') \}$$

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - **Reordering of the Source Sentences**
  - Dealing with the Unknown Words
- Experiments
- Conclusions

Reordering of the Source Sentences

- Reordering of the source sentences is a translation problem.
- We use a way similar to the monotone decoding of phrase-based SMT to performing the reordering. A dynamic programming recursion is used.

Reordering of the Source Sentences (Cont.)

- Two kinds of data are required:
  - Reordering Patterns, which is a set of triple <*CST*, *Perm*, *Prob*>. Here, *CST* is a chunk tag sequence, *Perm* is a permutation, and *Prob* is the corresponding probability.
  - Chunk tag 3-gram.
- These two types of data could both be trained used the training bitexts, with the Chinese part reordered at the training time.

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Reordering of the Source Sentences
  - **Dealing with the Unknown Words**
- Experiments
- Conclusions

Dealing with the Unknown Words

- No special translation models for named entities are used. Named entities are translated in the same way as other unknown words.
- Unknown words were translated in two steps:
  - Firstly, we will look up a dictionary containing more than 100,000 Chinese words for the word.
  - If no translations are found in the first step, the word will then be translated using a rule-based Chinese-English translation system.

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Reordering of the Source Sentences
- **Experiments**
- Conclusions

Experiments

- The data we used

Purposes	Corpus	
	Names	Amounts
Bilingual Phrases and Reordering Patterns	Training set from IWSLT 2007	177,535 sentence pairs
	Three parts from CLDC-LAC-2003-004: oral.xml, n_train.txt and life_2.xml	
English Language Model	English part of the training set from the 2005 China's National 863 MT Evaluation	7.4M words
Chinese Chunker	LDC2005T01	18,782 trees

XIAMEN UNIVERSITY

Experiments (Cont.)

- Scores of our system in IWSLT 2007

	BLEU-4
Baseline + Reordering	**0.2888**
Baseline	**0.2742**

- After incorporating the chunk-based reordering model, the phrase-based SMT system could outperform the baseline system.

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Reordering of the Source Sentences
- Experiments
- **Conclusions**

Conclusions

- We describe the system which participated in the 2007 IWSLT Speech Translation Evaluation of Department of Cognitive Science, Xiamen University.
- The result shows that after incorporating a chunk-based reordering model, the baseline system may achieve great improvements.
- More improvements are underway.

References

- Koehn, Philipp, Och, Fraz Josef and Marcu Danie, "Statistical phrase-based translation", *Proceeding of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Edmonton, Canada, 2003, pp. 127-133.
- Och, Franz Josef, "Statistical Machine Translation: From Single Word Models to Alignment Templates", *Ph.D. thesis*, RWTH Adchen, Germany, 2002.
- Och, Fraz Josef and Ney, Hermann, "Discriminative training and maximum entropy models for statistical machine translation", *Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, 2002, pp. 295-302.
- Och, Franz Josef, "Minimum error rate training in statistical machine translation", *Proceeding of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, 2003, pp. 160-167.
- Zens, Richard, Och, Franz Josef and Ney, Hermann, "Phrase-Based Statistical Machine Translation", *Proceeding of the 25th German Conference on Artificial Intelligence (KI2002)*, ser. *Lecture Notes in Artificial Intelligence (LNAI)*, M. Jarke, J. Koehler, and G. Lakemeyer, Eds., Vol. 2479. Aachen, Germany: Springer Verlag, September 2002, pp. 18–32.

References (Cont.)

- Koehn, Philipp, Axelrod, Amittai, Mayne, Alexandra Birch, Callison-Burch, Chris, Osborne, Miles and Talbot, David, "Edinburgh system description for the 2005 iwslt speech translation evaluation", *Proceeding of International Workshop on Spoken Language Translation*, Pittsburgh, PA, 2005
- He, Zhongjun, Liu, Yang, Xiong, Deyi, Hou, Hongxu and Liu, Qun, "ICT System Description for the 2006 TCSTAR Run #2 SLT Evaluation", *Proceeding of the TCSTAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006, pp. 63-68.
- Forney, G. D., "The Viterbi algorithm", *Proceeding of IEEE*, 61(2): 268-278, 1973
- Stolcke, Andreas, "Srilm – an extensible language modeling toolkit", *Proceedings of the International Conference on Spoken language Processing*, 2002, volume 2, pp. 901–904.
- Chen, Stanley F. and Goodman, Joshua, "An empirical study of smoothing techniques for language modeling", *Technical Report TR-10-98*, Harvard University Center for Research in Computing Technology, 1998.

Thanks

This work was supported by the National Natural Science Foundation of China (Grant No. 60573189), National 863 High-tech Program (Grant No. 2006AA01Z139), Natural Science Foundation of Fujian Province (Grant No.2006J0043) and the Fund of Key Research Project of Fujian Province (Grant No. 2006H0038).