The XMU Phrase-Based Statistical Machine Translation System for IWSLT 2006

Yidong Chen, Xiaodong Shi
Institute of Artificial Intelligence
Xiamen University
P. R. China

November 28, 2006 - Kyoto

XIAMEN UNIVERSITY

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- 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
保存
返回
Just agree to stop your horse?
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!
MP3

Overview (Cont.)

- IWSLT 2006
  - First participation
  - We implemented a simple **phrase-based** statistical machine translation system.
  - We participated in the **open data track** for **ASR lattice** and **Cleaned Transcripts** for the **Chinese-English translation direction**.

Outline

- Overview
- **Training**
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- 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.

Training (Cont.)

- Phrase Extraction
  - similar 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 $lex(\tilde{c} \mid \tilde{e})$
- $$p(\tilde{e} \mid \tilde{c}) = \frac{N(\tilde{e},\tilde{c})}{\sum_{\tilde{e}'} N(\tilde{e}',\tilde{c})}$$
- $$lex(\tilde{e} \mid \tilde{c}) = lex(e_1^I \mid c_1^J, a) = \prod_{i=1}^{I} \frac{1}{|\{j \mid (i,j) \in a\}|} \sum_{\forall (i,j) \in a} p(c_i \mid e_j)$$

Outline

- Overview
- Training
- System
  - **Translation Model**
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- 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.)

- Seven features
  - 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 $lex(\tilde{c} \mid \tilde{e})$
  - English language model $lm(e_1^I)$
  - English sentence length penalty I
  - Chinese phrase count penalty $-J'$
- We didn't use features on reordering.

Outline

- Overview
- Training
- System
  - Translation Model
  - **Parameters**
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- 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.)

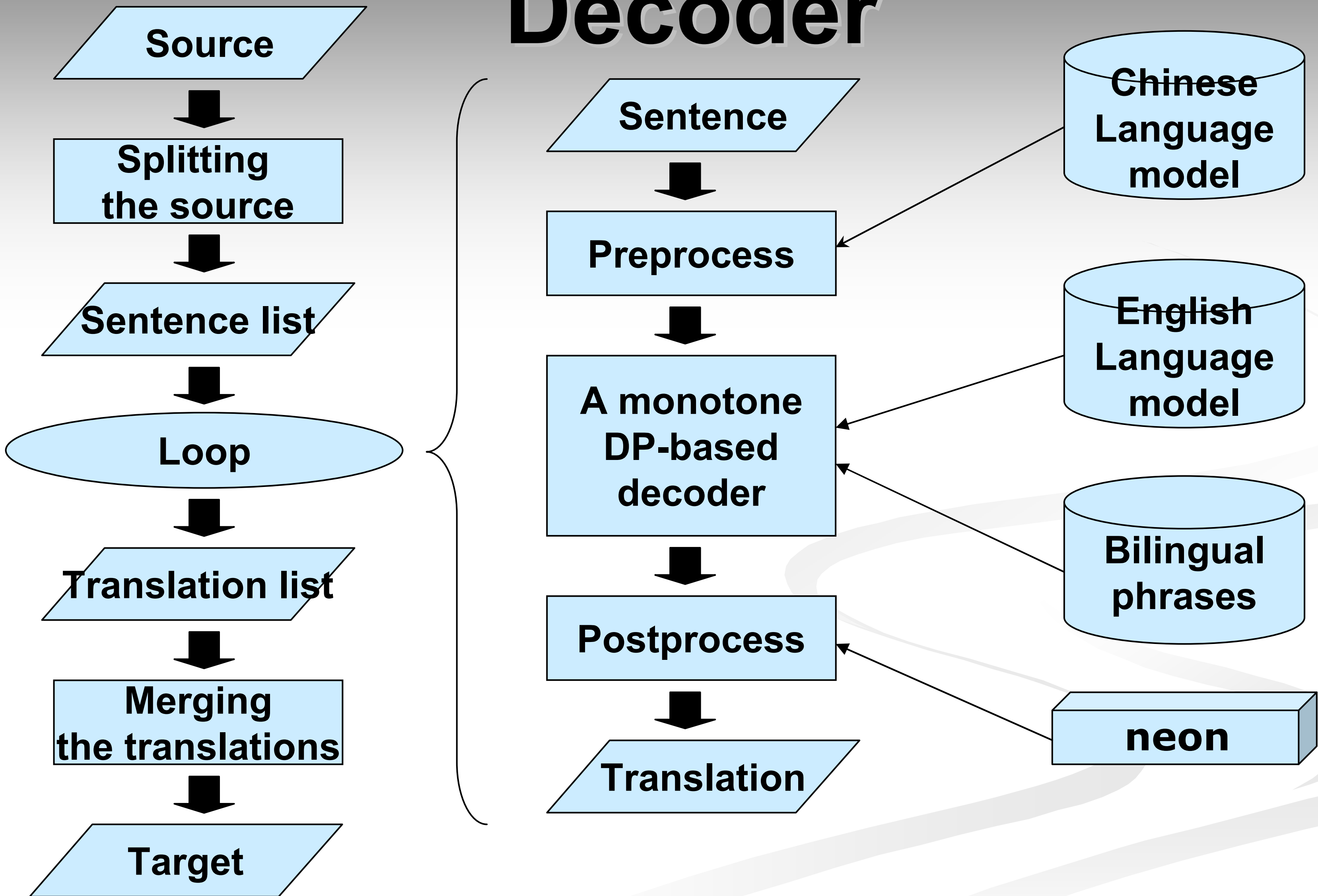
- The parameter settings for our system

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
λ_7	$-J$	0.48

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - **Decoder**
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions

Decoder



Source
Splitting the source
Sentence list
Loop
Translation list
Merging the translations
Target
Sentence
Preprocess
A monotone DP-based decoder
Postprocess
Translation
Chinese Language model
English Language model
Bilingual phrases
neon

Decoder (Cont.)

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

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

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

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

XIAMEN UNIVERSITY

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - **Dealing with the Unknown Words**
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- 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.
- All the 63 unknown words in the test data for the Cleaned Transcripts task in this evaluation are translated into English.

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - **Recovering the Missing Punctuations**
  - Translating the ASR Lattice
- Experiments
- Conclusions

Recovering the Missing Punctuations

- There are no punctuations in the Chinese sentences.
- The missing of punctuations can have an adverse effect on the translation quality, so we developed a preprocessing model to recover the missing punctuations.

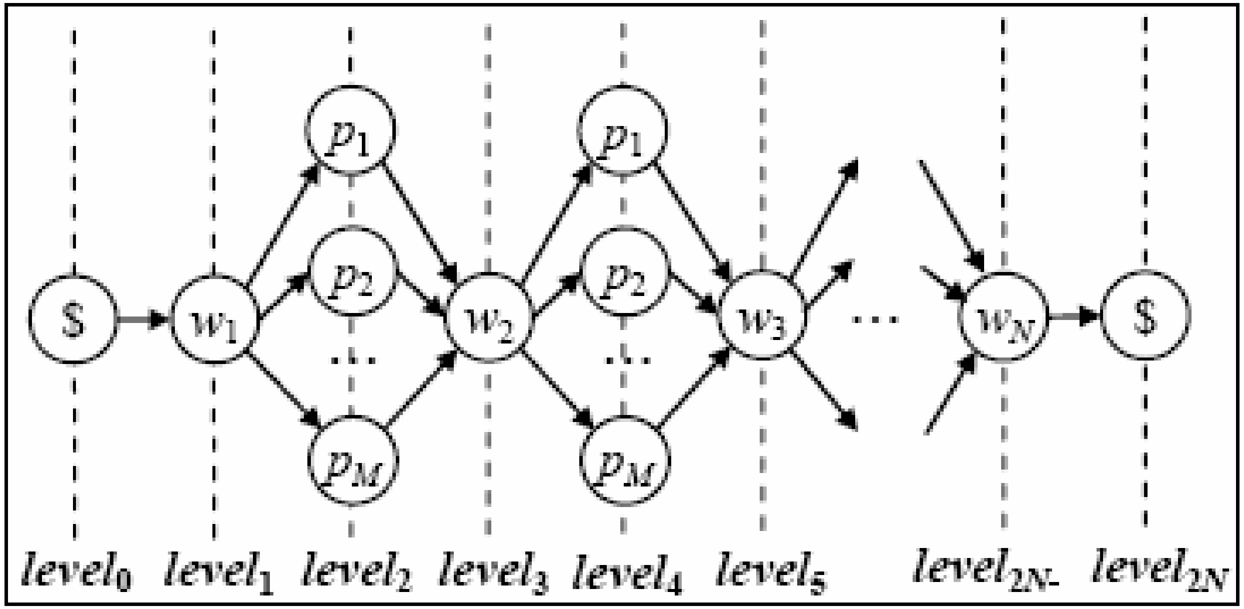
Recovering the Missing Punctuations (Cont.)

- Two ways to do with the missing punctuations
  - Method 1: To remove punctuations from the Chinese part of the training set, then train the model using the training set, and then translate the sentences without punctuations directly.
  - Method 2: To recover the punctuations from the input sentences, then translate the result sentences using a model trained from normal training set.
- Experiments and results on develop set 4

	bleu-4
Method 1	0.1936
Method 2	0.2139

Recovering the Missing Punctuations (Cont.)

- Given a Chinese sentence with N words, w_1 w_2 … w_N, we may construct a directed graph with $2N+1$ levels



XIAMEN UNIVERSITY

Recovering the Missing Punctuations (Cont.)

- Given such a graph, the problem of punctuation recovering could be looked on as a problem of searching the optimal path from the node in $level_0$ to the node in $level_{2N}$.
- In this problem, a path is said to be better than the other one if the **language model score** for it is larger than that for the latter.
- We then used the Viterbi algorithm to solve the search problem.

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - **Translating the ASR Lattice**
- Experiments
- Conclusions

Translating the ASR Lattice

- In the task of translating the ASR lattice, three types of test data were given:
  - word lattice
  - the 20-best results generated from ASR lattice
  - the 1-best result generated form ASR lattice
- A possibly better way is to regenerate the 1-best result based on Chinese language model from word lattice and then to translate it.

Translating the ASR Lattice (Cont.)

- We used a simpler approximate way
  - We first used our system to translate all the 20-best results and got 20 translations for each corresponding sentence.
  - Then we used the English language model to choose the best translation for each sentence.

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- **Experiments**
- Conclusions

Experiments

- The data we used

Purposes	Corpus	
	genre	statistics
Bilingual Phrase	Training set from IWSLT 2006	39,952 sentence pairs
	Training set from the 2005 China's National 863 MT Evaluation	152,049 sentence pairs
English Language Model	English part of the training set from the 2005 China's National 863 MT Evaluation	7.4M words
Chinese Language Model	Chinese part of the training set from IWSLT 2006	350K Chinese words
	Chinese Reader (Duzhe) Corpus	7.9M Chinese words

Experiments (Cont.)

- The use of additional data did help improving the performance of our system on the develop sets.
- Influence of the additional bitexts (bleu-4)

	Training without additional bitexts	Training with additional bitexts
develop set 1	0.3305	0.3922
develop set 2	0.3652	0.4349
develop set 3	0.4319	0.4823
develop set 4	0.1869	0.2139

Experiments (Cont.)

- Scores of our system in IWSLT 2006

	official (with case + punctuation)	additional (without case + punctuation)
CE spontaneous speech ASR output	0.1505	0.1623
CE read speech ASR output	0.1579	0.1718
Correct Recognition Result	0.1976	0.2162

Experiments (Cont.)

- Some lessons
  - The scores on Correct Recognition Result are significantly **higher** than those on ASR output. This may result from the influence of the ASR errors. And the other reason may be the simple method we used to translate ASR lattice.

Experiments (Cont.)

- Some lessons (Cont.)
  - The scores on CE read speech ASR output are slightly **higher** than those on CE spontaneous speech ASR output. This indicates that the ASR system used to give the ASR output may be cleverer at the read speech data than at the spontaneous speech data.

Experiments (Cont.)

- Some lessons (Cont.)
  - The additional scores are **higher** than the official scores. This indicates that post-editing models such as truecasing or punctuation correction may help improving the translation quality. We will integrate such models in the future.

Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- **Conclusions**

Conclusions

- We describe the system which participated in the 2006 IWSLT Speech Translation Evaluation of Institute of Artificial Intelligence, Xiamen University.
- It is a rather crude phrase-based SMT baseline, for example, without even considering phrase reordering.
- 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 and Grant No, 60373080), National 863 High-tech Program (Grant No. 2004AA117010-06) and Key Research Project of Fujian Province (Grant No. 2006H0038)