The CASIA Phrase-Based SMT System

Wei Wei

Institute of Automation

Chinese Academy of Sciences (CASIA)

weiwei@hitic.ia.ac.cn

//www.ia.ac.cn

Outline

1. Introduction
2. CASIA Phrase-based SMT System
3. Experiments
4. Conclusion

1. Introduction

//www.ia.ac.cn

1.1 SMT Research in CASIA

- CASIA began the research on SLT in 1996, and as the partner of C-STAR, we has built firm foundation of SMT Research
  - Chinese word Segmentation and POS tagging
  - Syntactic parsing
  - Large scale corpora
  - Machine translation in some small domains:
    - based on interlingua (IF)
    - example-based system
    - word-based on IBM model1 and IBM model2
    - hybrid translation system

- Recently, we participate in kinds of MT evaluations and workshop
  - International: IWSLT’04, IWSLT’05,TC-STAR 2006
  - Internal: China’s HTRDP(“863”) Machine Translation Evaluation,2005,
  - Workshop: 1st SMT workshop in China,2005

1.2 1st SMT workshop in China,2005

- Purposes of the workshop:
  - To enhance the SMT research in China
  - Specifically as in beginning stage, algorithm and methods implementation and understanding
  - Planned from Oct. 2004
- Participants & founder-member:
  - Institute of Automation, CAS(CASIA)
  - Institute of Computing Technology, CAS (ICT)
  - Computer Department , Xiamen University

1.2 1st SMT workshop in China,2005

- Corpus Preparing– training
  - CASIA50K: 50,000 bilingual corpus in travel domain by CASIA
  - ICT150K:150,000 bilingual corpus of movie caption
  - XMU200K: 200,000 bilingual corpus of movie caption
  - All sentences are not very long, not very large because the purpose of the workshop and some copyright problem
- Corpus Preparing– testing
  - CASIA1500: 1500 test corpus with every sentence 5 translations
  - 863-03 and 863-04 standard dialogue test in previous 2 years
    - 863-03: 350 (4 translation)
    - 863-04: 400 (4 translation)

About Workshop

- Held from July 13-14
- Email discussion and result exchange before workshop
- Two-days workshop
  - On site evaluation
  - System technical report for every group
  - Discussing

1.3 China's HTRDP("863") Machine Translation Evaluation

- 863 MT Evaluation is part of the HTRDP Evaluation on Chinese Information Processing and Intelligent Human-Machine Interface Technology
  - supported by China's national High-Tech Research and Development Programme ("863" Programme)
- Technologies covered by the whole HTRDP Evaluation
  - Machine translation (MT)
  - Automatic speech recognition (ASR)
  - Speech to text (TTS)
  - Chinese character recognition (CR)
  - Information retrieval (IR)
  - Chinese word segmentation (CWS, includes part of speech tagging and named entity recognition)
  - Text classification (TC)
  - Text summarization (TS)
  - Human face detection and recognition (FR)
- History:
  - Begin from 1991, up to 2005 it was the 8th evaluation. In the meantime, MT evaluation has been held for 6 times from 1994 to 2005.

1.3 China's HTRDP("863") Machine Translation Evaluation

- Test Data (recent HTRDP MT evaluation 2003, 2004, and 2005)
  - The test data are mainly collected from real language
  - Both dialog data and text data are collected
  - Size: about 700-1000 sentences in each track
  - Domain: General and Olympic, Where Olympic-related domain covers: weather, sports, travel, traffic, hotel, restaurant, and etc.
- Training data
  - No training data were provided before 2004
  - Training data are provided for only E->C and C->E tracks in HTRDP MT evaluation 2005
  - Amount: 870,000 sentence pairs, which have been examined manually in 2005

2. CASIA Phrase-based SMT System

//www.ia.ac.cn

2.1 The CASIA SMT system

- Chinese-to-English
- in the domain of BTEC
- phrase-based SMT
- Some improvements:
  - using templates with variables
  - different tracing back algorithms in decoding
  - Improvement with word alignment

2.2 The phrase-based SMT model

$$p(e \mid c) = p_T(c \mid e)^{\lambda_t} \times p_L(e)^{\lambda_l} \times p_D(e,c)^{\lambda_d}$$

$P_T(c/e)$ is the translation model

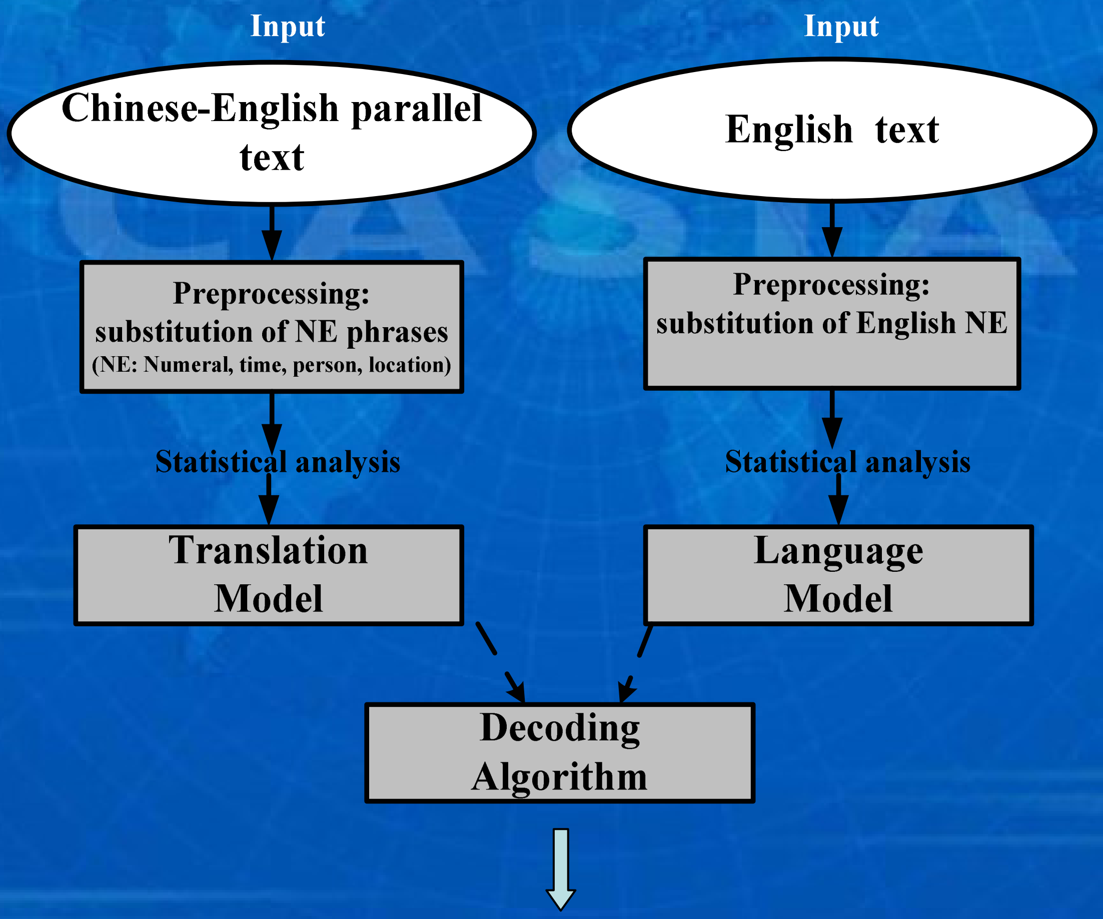
$P_L(e)$ is the target language model

$P_D(e,c)$ is the distortion model and

$P_D(e,c) = \lambda \mid a_i - b_{i-1} - 1 \mid$

2.2 The phrase-based SMT model

- Components: Preprocessing, Translation model, language model, decoder


Input
Input
Chinese-English parallel text
English text
Preprocessing: substitution of NE phrases (NE: Numeral, time, person, location)
Preprocessing: substitution of English NE
Statistical analysis
Statistical analysis
Translation Model
Language Model
Decoding Algorithm

//www.ia.ac.cn

Preprocessing model

- Pre-processing of numerals using rule-based approach

  - Arabic numbers: telephone No., room No. etc

  - Number written in Chinese, such as "一百 one hundred"

  - Ordinal numbers:

  - Dates

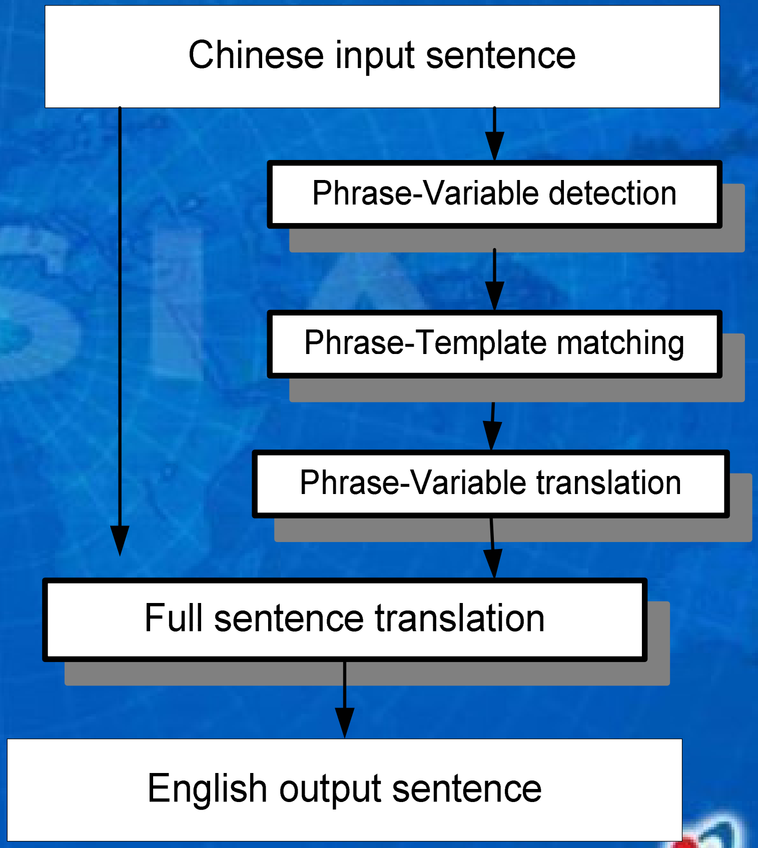
  - Combination of the different expressions

Preprocessing model

- After pre-processing, the numerals are replaced with the specific marks (variables).

- The phrase translations become the templates with variables, e.g.

X个单人间 -> X single rooms



Chinese input sentence

Phrase-Variable detection

Phrase-Template matching

Phrase-Variable translation

Full sentence translation

English output sentence

Preprocessing model

- In our experiment, about 5% extracted phrases contain variable.
- The performance of the system has been improved about 3.2% using pre-processing of the numerals.

Phrase Translation Model

- Refined model from bi-direction Word-Based Alignment (Och 2002,Koehn et al.,2003).
- Phrase translation for “我 要 买”

English Phrase	Probability
I want to buy	0.386
I would like to buy	0.234
I will buy	0.119
I wanna buy	0.108
I wan to get	0.101
……	……

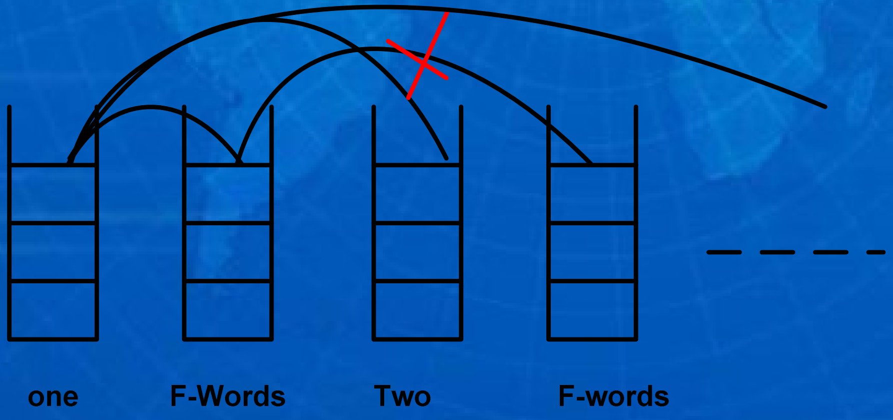
Language Model

- Standard Technique: Trigram Model
  - multiplication of trigram probabilities
- Using the available SRI language modeling toolkit.

http://www.speech.sri.com/projects/srilm/

Beam-Search Decoding

- The basic algorithm is Beam-Search, same as Pharaoh. (Och & Ney, 2002; Zens et al, 2002)
- Different Stack Organization:
  - The hypotheses are stored in different stacks
    - bin label: number of Chinese words covered
    - insert the Functional Words(of , the, for…..) bins.



F-Words must expand to Non F-Words Next

Decoding: search the best translation

- Considering there are many auxiliary words and mood words in the Chinese sentence, and these words sometimes don’t have the corresponding words in the English sentence. So, the system does not require the all words in the Chinese sentence to be translated.

Select the candidate sentences generated after *L*- *a* Chinese words have been translated.

L is the length of the input Chinese sentence

a is an integer.

2.3 Improvement with word alignment

- Statistical incorporation of dictionary cues
  - The dissimilarity between Chinese and English need the lexical cues to improve pure statistical word-alignment.
  - Size of dictionary ,including 1-N,N-1phrase pairs (After filtering through the training data)
    - 1-N(N>=1) 31,200
    - N-1(N>1) 73,811

2.3 Improvement with word alignment

- The approach of dictionary incorporation
  - Lower-case: when find proper English word in the dictionary, we convert the E-word in the sentence into lower-case.
  - Stem-word: in order to solve the morphosyntactic word in English ,and only stemming some first characters is better than using the porter-stem tool.
  - Word disambiguation: Integrate some restrictions of context when the word appears more than one time in the sentence (especially some functional word, such as of, the……..)

3. Experiments

- **Experiment-1**
  - **Comparison of the different searching algorithms using 100,000 sentence pairs ,Marks as follow:**

M means word-based translation model;

+NF0 means the Functional-zero words are not applied;

+F0 means the Functional-zero words are applied;

+BACK1 stands for our decoder;

+BACK2 stands for Koehn's decoder;

+NUM means the numerals are pre-processed

3. Experiments

Methods	Bleu (4-gram)
M+NF0+BACK2	0.1833
M+NF0+BACK1	0.1919
M+F0+BACK2	0.2372
M+F0+BACK1	0.2663
PBT +NF0+BACK2	0.2730
PBT +NF0+BACK1	0.2864
PBT +F0+BACK2	0.2763
PBT +F0+BACK1	0.2882
PBT +F0+BACK1+NUM	0.3177

0.0086

0.0291

//www.ia.ac.cn

3. Experiments

- Experiment-2
  - Comparison of the number of translation options in each stack and decoding time using 900,000 sentence pairs

Methods	Bleu (4-gram)	Decoding time
G+F0+BACK1	0.3418	2H6Min
G+F0+BACK1_Top100	0.3452	40Min
G+F0+BACK1_Top150	0.3446	54Min
G+F0+BACK1_Top200	0.3423	64Min
G+F0+BACK1_Top50	0.3366	23Min

3. Experiments

- Experiment-3
  - Comparison of the different methods for improving statistical word alignment.

	Baseline	Dic_stem	Dic_stem_lower	Dic_stem_lower_ disambiguation
Precision	0.8939	0.8177	0.8211	0.8725
Recall	0.5744	0.6312	0.6554	0.6888
F-Measure	0.7053	0.7125	0.7290	0.7698
ASR	0.2866	0.2828	0.2668	0.2254

Test data: 500 sentence pair from 2005 HTRDP WA evaluation

Training data: 870,000 sentence pairs from 2005 HTRDP MT evaluation

3. Experiments

- IWSLT2005 evaluation
- Training Corpus
  - 1,000,000 sentence pairs in the specific domain of C-Star, including BTEC corpus and CJK corpus and CASIA corpus
  - 500,000 sentence pairs in the general domain (news) from Chinese LDC

3. Experiments

- Perplexity of source language (Chinese)
  - Use SRILM tool
  - Results:
    - counting all input tokens:41.2084
    - excluding end-of-sentence tags:69.3387
- Results from IWSTL'2005

Track (C-E)	Data condition	Bleu4	NIST	Meteor	WER	PER
Manual transcription	unrestricted	0.5279	10.2499	0.7214	0.4160	0.3366
ASR Output	unrestricted	0.3845	8.0406	0.5802	0.5788	0.4770

4. Conclusion

- Statistical MT has been initially investigated in China that preliminary result is comparable with the state-of-art rule-based or hybrid system
- Phrase-based has shown to be superior to other system by now in view of implementation and accuracy
- Phrase with variables or phrase template has been initially tried to have some improvement in accuracy

Future direction

- Way of merging EBMT and SMT
  - Phrase template
  - What kind of words or parameter could be variables
    - Besides time\number, name entity like \name\location …..
  - Need to integrate more advanced preprocessing (shallow paring to medium-depth parsing)
- But Systematical integration of structure knowledge-morphological, syntax and so on ….

Thanks !

//www.ia.ac.cn