CASIA Phrase-Based SMT System for IWSLT'07

Yu ZHOU
Institute of Automation
Chinese Academy of Sciences (CASIA)
yzhou@nlpr.ia.ac.cn

2007-10-15

Outline

- Overview
- Data
- Word Alignments
- Phrase Extraction
- Language Model
- Decoder
- Other Process
- Experiments
- Conclusion & Future Work

Overview

- Using the Phrase-Based SMT
- Using a log-linear model

$$e^* = \arg\max_{e} \sum_{m=1}^{M} \lambda_m h_m(e, f)$$

Chinese-English translation task

Data

--Data Collection

- Downloading all the open resources from the web;
- Filtering these corpus which is highly correlative with the released train data by IWSLT 2007.

Data

--Data Preprocessing

- Chinese word segmenting using the software ICTCLAS3.0 (http://www.nlp.org.cn)
- Removing the noises words or characters in the training data
- Transforming the SBC case into DBC case in Chinese corpus
- Tokenizing the English words

Word Alignments

- Obtaining the initial word alignments by GIZA++ (http://www.fjoch.com/GIZA++.html)
- Using the method grow-diag-final method to modify the initial alignments
- Using our method to modify the word alignments by using the dictionary and jumping-distance

Word Alignments (Cont.)

- Dictionary:
  - The bilingual dictionary download from the open resources on the web (http://isslt07.itc.it/menu/resources.html)
  - From the bi-directional dictionaries generated by GIZA++, only using such word pairs with the highest probabilities.

Word Alignments (Cont.)

◆ If the word pair (f_i, e_j) which is inexistent in the bilingual dictionary but existent in the word alignments of the sentence pair ;

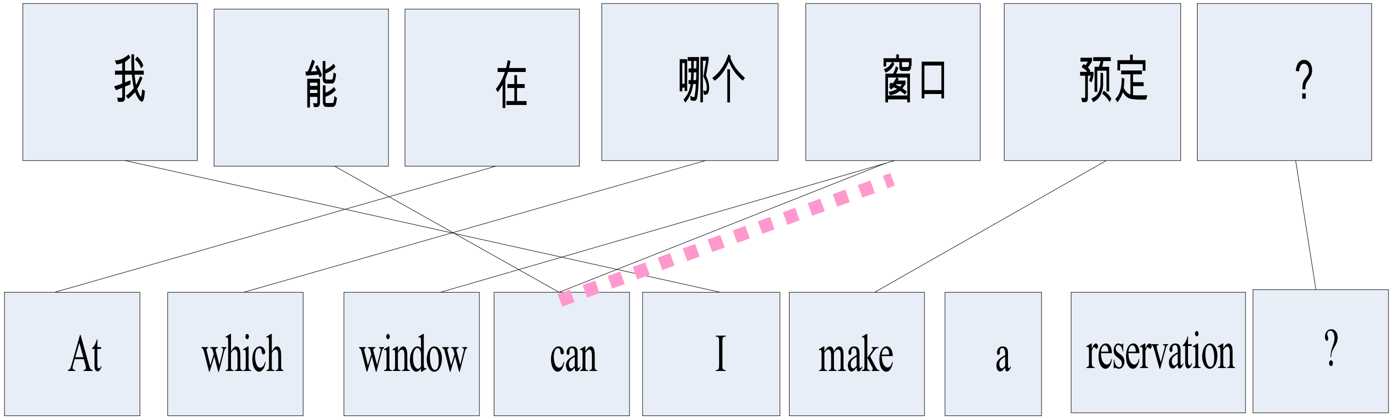
◆ If the word pair (f_i, e_j) is inexistent both in the bi-lingual dictionary and the word alignments, we will not deal with such case;

◆ If the word pair (f_i, e_j) is existent in the bilingual dictionary but inexistent in word alignments we will add the word pair alignment information;

◆ If the word pair (f_i, e_j) is existent both in the bi-lingual dictionary and in the word alignments, we will keep the word pair alignment information.

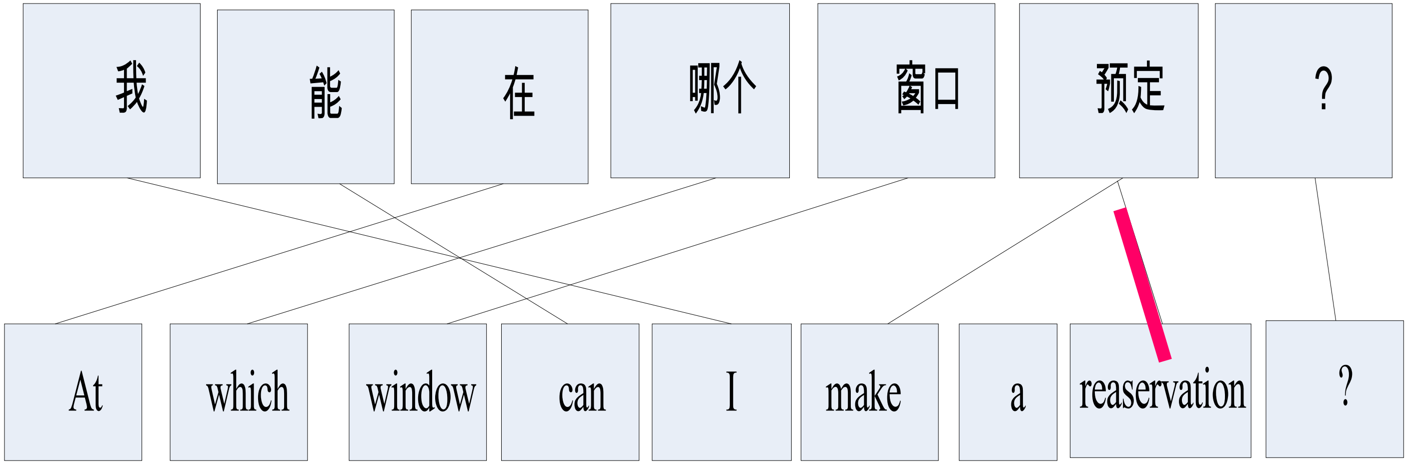
Word Alignments (Cont.)

- For example:



我 能 在 哪个 窗口 预定 ?

At which window can I make a reservation ?



我 能 在 哪个 窗口 预定 ?

At which window can I make a reaservation ?

Phrase-Extraction

- Och’s method:
  - Simple and easy to realized
  - Totally consistent with word alignments
- Extend Och’s method:
  - Using a flexible scale to extract the phrase pairs

Phrase-Extraction (Cont.)

$$(\tilde{f}, \tilde{e}) \in BP <=>$$

$$\forall f_i \in \tilde{f} : (f_i, e_j) \in A \rightarrow e_j \in \tilde{e}$$

$$AND \quad \forall e_j \in \tilde{e} : (f_i, e_j) \in A \rightarrow f_i \in \tilde{f}$$

Phrase-Extraction

$$(\tilde{f}, \tilde{e}) \in BP <=>$$

$$\forall f_i \in \tilde{f} : (f_i, e_j) \in A \rightarrow e_j \in \tilde{e}$$

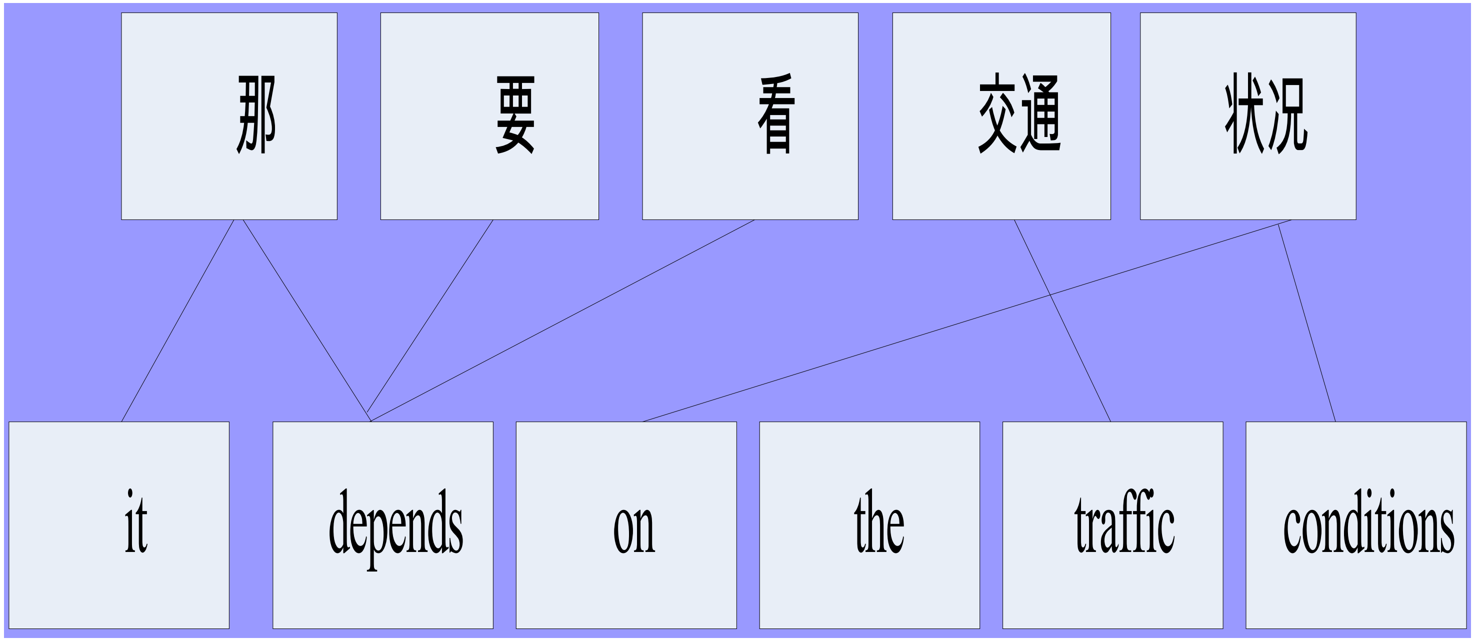
$$AND \quad \begin{cases} \forall e_j \in \tilde{e} : (f_i, e_j) \in A \rightarrow f_i \in \tilde{f} \\ OR \quad \dfrac{\{e_j \mid (e_j, f_i) \in A\}}{\{e_j \mid e_j \in \tilde{e}\}} \geq Threshold \;, \\ e_j \; is \; not \; a \; functional \; word \;, \\ \qquad \rightarrow \underset{\{f_i \mid (f_i, e_j) \in A\}}{\arg\max} \, p(f_i \mid e_j) \in \tilde{f} \end{cases}$$

Phrase-Extraction (Cont.)

- Computing the percentage of consistent target words in $\tilde{e}$;
- Judging if these non-consistent target words are functional words;
- Checking if the source words that the non-consistent and non-functional target word is aligned to are all outside $\tilde{f}$;
- Extracting the phrases.

Phrase-Extraction (Cont.)

- For example:



那 要 看 交通 状况

it depends on the traffic conditions

Phrase-Extraction (Cont.)

Och’s
phrase
table

那 要 看 ||| it depends
那 要 看 交通 状况 |||
it depends on the traffic conditions
交通 ||| traffic
交通 ||| the traffic
交通 状况 ||| on the traffic conditions

那 要 看 ||| it depends
那 要 看 交通 ||| it depends on the traffic
那 要 看 交通 状况 ||| it depends on the traffic conditions
要 看 交通 ||| depends on the traffic
要 看 交通 状况 ||| depends on the traffic conditions
看 交通 ||| depends on the traffic
看 交通 状况 ||| depends on the traffic conditions
交通 ||| traffic
交通 ||| the traffic
交通 状况 ||| on the traffic conditions

Our
phrase
table

Phrase-Extraction (Cont.)

$$\phi(\tilde{f} \mid \tilde{e}) = \frac{N(\tilde{f},\tilde{e})}{\sum_{\tilde{f}'} N(\tilde{f}',\tilde{e})}$$

$$\phi(\tilde{e} \mid \tilde{f}) = \frac{N(\tilde{f},\tilde{e})}{\sum_{\tilde{e}'} N(\tilde{f},\tilde{e}')}$$

$$lex(\tilde{f} \mid \tilde{e},a) = \prod_{i=i_1}^{i_2} \frac{1}{|\{j \mid (i,j) \in a\}|} \sum_{\forall (i,j) \in a} p(f_i \mid e_j)$$

$$lex(\tilde{e} \mid \tilde{f},a) = \prod_{j=j_1}^{j_2} \frac{1}{|\{i \mid (i,j) \in a\}|} \sum_{\forall (i,j) \in a} p(e_j \mid f_i)$$

Language Model

- Using the ngram-count tool in the open SRILM toolkit (http://www.speech.sri.com/projects/srilm)
- Training on the English part of the training data in GIZA++
- Using the 4-gram language model with Kneser-Ney smoothing method

Decoder

- Using the beam search algorithm which is similar to the Pharaoh decoder
- Adding the ‘expanding F-zerowords’ model
- Using a new tracing back method;
- Using the monotone search without any distortion and reordering model

Other Process

--Name Entities

- For the person name and location name, we translate them only by looking up its translations in the common phrase pair table
- For the organization name, we translate them using the model based on a synchronous CFG grammar
- For the number and date, we adopt the method based on the man-written rules to translate

Other Process

--Post-processing

- Transforming the lowercase of the first character of the English words into uppercase
- Recombination the separated punctuations with its left closest English words

Experiments

Data	Chinese	English
CE_train	39,950	39,950
CE_sent_filtered	188,282	188,282
CE_dict_filtered	31,132	31,132
CE_newdev1	24,192	24,192
CE_newdev2	10,423	10,423
CE_test	489	-

Experiments (Cont.)

DEV_train	Chinese	English
Sentences	283,556	283,556
Words	1,754,932	1,900,216
Vocalbulary	11,424	10,507
Average Length	6.2	6.7

Experiments (Cont.)

TST_train	Chinese	English
Sentences	293,979	293,979
Words	1,890,984	2,051,619
Vocalbulary	11,661	11,273
Average Length	6.4	7.0

Experiments (Cont.)

System	BLEU4
Baseline	0.2730
CASIA	0.3648

Baseline means the system with the base methods on word alignments and phrase extraction. The baseline system is only looking the name entities as the common words. CASIA means the system with the new methods described in our paper.

Conclusion & Future Work

- Using several new approaches in this our system: word alignments, phrase extraction, name entity identification and translation
- Adding the semantic information into our model
- Using non-consecutive phrase pair into our decoder
- Adding the reorder model into our decoder
- Re-ranking the N-best of the decoder
- Combining with other translation systems

Thanks！

谢谢

！