The NICT/ATR Speech Translation System for IWSLT 2008

Masao Utiyama, Hailong Cao (NICT)
Andrew Finch, Hideo Okuma, Michael Paul, Hirofumi Yamamoto, Keiji Yasuda, Eiichiro Sumita (NICT/ATR)

Tasks

- English-Chinese Challenge Task
- Chinese-English Challenge Task
- Pivot task

CleopATRa (Inhouse decoder)

- Phrase-based SMT system
- Log-linear model whose features are the same as those of the MOSES decoder
- Dynamic Interpolation

English-Chinese Challenge Task

Factors affecting English-Chinese SMT were examined

system	devset	devset3	factor
org	0.4282	0.4301	Original BTEC
dict	0.4462	0.4363	Chinese word segmentation
cldc	0.4399	0.3834	CLDC-2004-863-0009
all	0.4963	0.4710	BTEC+CLDC
all+dict+cldc	0.4966	0.4691	Clustering
all+questions+declarations	0.5055	0.4743	Clustering
all+dict+cldc+q.+d.	0.5070	0.4745	Clustering

These BLEU scores were obtained without MERT.

Chinese word segmentation (CWS)

Comparison of the original CWS in the supplied BTEC training corpus with a re-segmentation of the same corpus +1.8% BLEU, +0.6 % BLEU

- Dictionary-based CWS system
- Viterbi-segmetaion according to a language model
- Dictionary was augmented by the words in the BTEC corpus

CWS is important.

Additional Corpus

- MODEL = devset, devset3
- BTEC = 0.4462, 0.4363
- CLDC = 0.4399, 0.3834
- BTEC+CLDC = 0.4963, 0.4710

+5.01% BLEU, +3.47 % BLEU

- BTEC was more suitable than CLDC-2004-863-0009
- Using BTEC and CLDC was very effective

Dynamic Interpolation

- Our decoder, **CleopATRa**, can linearly interpolate all the models from all the sub-systems according to a vector of interpolation weights that are supplied for each sentence to be decoded
- phrase-table, reordering-table, language model can be combined

Clusters made from the training data were used to make models

Clustering by corpora (1/2)

- BTEC and CLDC were regarded as classes
- Pr(BTEC|sentence) was the weight for BTEC
- Probabilities were learned by an ME classifier

Class	Features
BTEC	please input your pin number
BTEC	we want to have a table near the window
CLDC	yes please
CLDC	thank you sir

Clustering by corpora (2/2)

- MODEL = devset, devset3
- BTEC = 0.4462, 0.4363
- CLDC = 0.4399, 0.3834
- BTEC+CLDC = 0.4963, 0.4710
- BTEC, CLDC, BTEC+CLDC = 0.4966, 0.4691

The weight of "BTEC+CLDC" was fixed.

+0.03% BLEU, −0.19 % BLEU

Clustering by corpora was not effective

Clustering by sentence type (1/2)

- Question sentences and non-question sentences were regarded as clusters.
- Pr(Question|sentence) was the weight for the question model
- Probabilities were learned by an ME classifier

Class	Features
Q	`<s>_where <s>_where_do where where_do where_do_i ...`
Q	`<s>_how <s>_how_long how how_long how_long_is ...`
NQ	`<s>_the <s>_the_light the the_light the_light_was ...`
NQ	`<s>_i <s>_i_have i i_have i_have_a have_a ...`

Clustering by sentence type (2/2)

- MODEL = devset, devset3
- BTEC+CLDC = 0.4963, 0.4710
- Questions, Non-questions, BTEC+CLDC = 0.5055, 0.4743

The weight of "BTEC+CLDC" was fixed.

+0.92% BLEU, +0.33 % BLEU

Clustering by sentence type was slightly effective

Combination of all models

- MODEL = devset, devset3
- BTEC+CLDC = 0.4963, 0.4710
- BTEC, CLDC, BTEC+CLDC = 0.4966, 0.4691
- Questions, Non-questions, BTEC+CLDC = 0.5055, 0.4743
- BTEC, CLDC, Questions, Non-questions, BTEC+CLDC = 0.5070, 0.4745

+0.15% BLEU, +0.02 % BLEU

Combination of all models was slightly effective.

Pivot Task

Strategies examined

- Cascade
- Pseudo corpus
- Phrase table composition

Cascade strategy (Baseline)

- SMT-1: Chinese sentence $\rightarrow$ English sentence
- SMT-2: English sentence $\rightarrow$ Spanish sentence
- SMT-1 + SMT-2: Chinese sentence $\rightarrow$ English sentence $\rightarrow$ Spanish sentence

Pseudo Corpus

- English–Chinese training data → EC-SMT system
- Spanish–English training data → English part → EC-SMT system → Translated Chinese (100-best) → Spanish–Translated Chinese training data → SC-SMT system
- English–Spanish training data → ES-SMT system
- Chinese–English training data → English part → ES-SMT system → Translated Spanish (100-best) → Chinese–Translated Spanish training data → SC-SMT system

Phrase table composition

$$\phi(\bar{s}|\bar{c}) = \sum_{\bar{e} \in T_{SE} \cap T_{EC}} \phi(\bar{s}|\bar{e})\phi(\bar{e}|\bar{c})$$

- $\bar{s}$, $\bar{c}$, $\bar{e}$: Spanish, Chinese, and English phrases
- T_{SE}, T_{EC}: Spanish-English, English-Chinese phrase-tables
- $\phi(\bar{s}|\bar{e})$, $\phi(\bar{e}|\bar{c})$: Phrase translation probability

Lexicalized reordering models were also induced.

Comparison of BLEU

1. Cascade = 0.2529
2. Pseudo Corpus (EC-SMT) = 0.2860
3. Pseudo Corpus (ES-SMT) = 0.2740
4. Phrase-table induction = 0.2703
5. Linear interpolation (2+3+4) = 0.3050

Summary

- English–Chinese translation Challenge Task:

  Chinese word segmentation and external resources had a significant impact on the translation results

- Chinese–English translation Challenge Task:

  We used a novel clustering method based on WER

- PIVOT Task:

  We integrated two strategies for pivot translations by linear interpolation.