Two methods for stabilizing MERT: NICT at IWSLT 2009

Masao Utiyama, Hirofumi Yamamoto, and Eiichiro Sumita

NICT

Our system

- Participated in the Challenge Task
- A baseline phrase-based SMT system

Outline

- Language resources
- Combination of Chinese segmentations
- Two methods for stabilizing MERT
- Official results
- What we tried but didn't work

Language resources

- Training data: `IWSLT09_BTEC.train.*`, `IWSLT09.devset*`, `IWSLT09_CT.train.*`
- Expanded the devset data
- Two language models from the CT portion and rest
- Development data: Sample sentences from the training data. These sentences were excluded from the training data.
- Development test data:
  `IWSLT09_CT.devset.*.with_interpreter.txt`
- In-house tokenizers for Chinese and English
- Lowercased English sentences in EC
- Truecase in CE

Combination of Chinese segmentations

Method1

Word segmentation with our in-house tokenizer

Method2

1. Segmentation into characters with '$\langle w \rangle$' tags inserted between words.
2. Insertion of '$\langle w \rangle$' into English texts
3. Made a phrase (reordering) table from this data.

Combination of the two tables

1. Phrases in the first phrase (reordering) table were segmented into characters.
2. Removal of "$\langle w \rangle$" from the output

Two methods for stabilizing MERT

- Devset sampling
- Averaged mert

Causes of instability in MERT

Mismatch between development and test data

- *Devset sampling* tries to sample sentences that are similar to the input data
- Best parameters on devset $\neq$ Best parameters on test

  Averaged MERT avoids over-fitting by using averaged parameters

Devset sampling

- Sampling similar sentences to the input texts
- Sampled sentences were excluded from the training
- For test each sentence in the 500 test sentences
  - Extracted the most similar 2 sentences
  - Average of BLEU1, ... BLEU4 scores as the similarity score
  - Input sentence was regarded as the reference
  - Used the most similar 1000 sentence development data
- After the tuning, the development data were added to the training data again to make the final model

Results using devset sampling

	with sampling	w/o sampling
EC	32.16	30.34
CE	28.66	26.12

Similar sentences (testset // devset)

- hotel royal plaza may i help you // holiday inn crowne plaza may i help you
- yes can you tell me the number of people type of room and approximate budget please // yes would you tell me the address and phone number of the hotel please
- yes let me check for vacancies // yes let me check hold on a moment please
- sorry to keep you waiting // sorry to keep you waiting
- we have two types of rooms available in your budget range // we have two types of dressing japanese or french

Averaged MERT

- Run MERT several times on a development data
- Average tuned parameters to get final parameters

Why this works?

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

- Average of parameters (weights) $\rightarrow$ average of scores
- A kind of a system combination method

Results of Averaged Mert

	average	max
EC	32.61	31.93
CE	29.24	28.49

Additional experiments on IWSLT-2007 Japanese-English translation task

- Bootstrap method
- Run MERT 100 times on devset1
- Obtained 100 parameter sets
- Calculated the averages and standard deviations of BLEU scores for 1, 2, 3, 5, 7, 10, 20, 30, 50, 70, and 100 parameter sets by sampling these parameter sets
- Sampled 100 parameter sets for each number of parameter sets.

Two methods

- Averaged parameters
- Parameters that obtained the maximum BLEU score on devset1

Results

method	average	maximum
No.	av. (std.)	av. (std.)
1	62.22 (0.54)	62.22 (0.54)
2	62.59 (0.41)	62.32 (0.42)
3	62.63 (0.37)	62.08 (0.59)
5	62.72 (0.38)	62.18 (0.53)
7	62.72 (0.29)	62.14 (0.56)
10	62.73 (0.27)	62.14 (0.54)
20	62.71 (0.21)	62.27 (0.52)
30	62.73 (0.21)	62.16 (0.55)
50	62.69 (0.19)	62.36 (0.45)
70	62.70 (0.16)	62.42 (0.41)
100	62.71 (0.15)	62.50 (0.33)

BLEU scores for our official submissions

	EC		CE	
	c+p	nc+np	c+p	nc+np
ASR	35.83	35.44	26.67	25.80
CSR	38.42	38.15	29.70	28.72

What we tried but didn't work

- Increasing the size of the CT corpus
- Alignment with lowercased prefixes
- Replacing numbers with a special symbol

Increasing the size of the CT corpus

- Adding several replications of each sentence of the CT corpus when we added them to the BTEC corpus

	with	w/o
EC	30.89	31.25
CE	25.12	26.11

Alignment with lowercased prefixes

- Using lowercased 4-letter prefixes of English words in word alignment

	with	w/o
EC	29.58	32.22
CE	26.73	26.91

Replacing numbers with a special symbol

	with	w/o
EC	29.56	32.22
CE	24.17	26.91

Examples failed

- a Chinese word sequence "0 0 0" was translated into "triple o"

Conclusions

- Participated in the Challenge Task
- Two methods for stabilizing MERT to reduce mismatch between development and test data
  - *Devset sampling* tries to sample sentences that are similar to the input data
  - Best parameters on devset $\neq$ Best parameters on test

    Averaged MERT avoids over-fitting by using averaged parameters