Towards Better Translation Performance on Spoken Language

Chao Bei and Hao Zong
{beichao,zonghao}@gtcom.com.cn
1 Task description

Table 1 Number of sentences summary for in-domain training and development data for bilingual task in small data condition

| NMT direction | training data | development data | monolingual data (target) |
|---------------|---------------|------------------|---------------------------|
|               | 231K          | 1,372 1,297 1,205| 520K                      |
| en-zh         | 231K          | 1,372 1,297 1,205| 234K                      |
| zh-en         |               |                  |                           |
1 Task description

Table 2 Number of sentences summary for in-domain training and development data for zero-shot multilingual task

| language          | de-en | de-it | de-ro | en-it | en-nl | en-ro | it-nl | nl-ro |
|-------------------|-------|-------|-------|-------|-------|-------|-------|-------|
| training data     | 204K  | 203K  | 200K  | 230K  | 236K  | 219K  | 232K  | 205K  |
| development set   | 1,138 | 1,133 | 1,121 | 1,147 | 1,181 | 1,129 | 1,183 | 1,123 |
Contents

1 Task description
2 Bilingual task
3 Multilingual task
4 Summary
5 Q&A
2 Bilingual task
Data Preprocessing

- For English:
  - Tokenizer and Truecase (Moses)

- For Chinese:
  - Apply Jieba segmentation without recognizing new words.

- BPE for English and Chinese respectively:
  - \( N_{operation} = \text{number of words(\text{word frequency} > 10)} \)
2 Bilingual task

Model architecture

Encoder Decoder with Attention
Table 3 Model configuration for bilingual task in small data condition

| Type                                | value   |
|-------------------------------------|---------|
| English vocabulary size             | 19623   |
| Chinese vocabulary size             | 25377   |
| word embedding                      | 512     |
| hidden units                        | 1024    |
| embedding dropout                   | 0.2     |
| hidden dropout                      | 0.2     |
| source dropout                      | 0.1     |
| target dropout                      | 0.1     |
| layer normalization                 | True    |
| maximum sentence length             | 100     |
2 Bilingual task

Result Analysis

Table 4 Case-insensitive BLEU score in development set of Chinese-to-English in small data condition. WN means weight normalization and SD means synthetic data.

|                      | tst2013 | tst2014 | tst2015 | average |
|----------------------|---------|---------|---------|---------|
| 2 layers             | 20.32   | 18.07   | 21.48   | 20.03   |
| + annealing Adam     | 20.85   | 18.39   | 22.04   | 20.47   |
| 4 layers             | 20.89   | 17.91   | 21.87   | 20.33   |
| + annealing Adam     | 20.81   | 17.91   | 22.24   | 20.33   |
| 4 layers with WN     | 20.95   | 17.99   | 21.98   | 20.43   |
| + annealing Adam     | 21.24   | 18.1    | 21.81   | 20.48   |
| 4 layers with SD     | 21.05   | 18.4    | 21.94   | 20.49   |
| + annealing Adam     | 20.94   | 18.57   | 22.41   | 20.65   |
| 4 layers with SD and WN | 21.34   | 18.72   | 22.5    | 20.91   |
| + annealing Adam     | 21.53   | 18.72   | 22.46   | 20.98   |
| Deep transition      | 20.68   | 17.56   | 21.49   | 19.97   |
| + annealing Adam     | 21.11   | 17.66   | 21.64   | 20.28   |
| Deep transition with WN | 20.71   | 17.98   | 21.96   | 20.78   |
| + annealing Adam     | 21.40   | 18.33   | 22.30   | 20.80   |
| Deep transition with SD | 21.49   | 18.1    | 22.40   | 20.73   |
| + annealing Adam     | 21.75   | 18.83   | 22.77   | 21.16   |
| Deep transition with SD and WN | 21.31   | 18.78   | 22.07   | 20.78   |
| + annealing Adam     | 21.86   | 18.64   | 22.23   | 20.97   |
| ensemble             | 22.83   | 19.72   | 23.73   | 22.13   |
| + r2l reranking      | 23.02   | 19.94   | 24.26   | 22.43   |
Table 5 Case-insensitive BLEU score in development set of English-to-Chinese in small data condition. WN means weight normalization and SD means synthetic data.

| Method                                      | tst2013 | tst2014 | tst2015 | average |
|---------------------------------------------|---------|---------|---------|---------|
| 2 layers                                   | 23.71   | 21.03   | 26.80   | 23.83   |
| + annealing Adam                            | 24.3    | 21.45   | 26.69   | 24.14   |
| 4 layers                                    | 23.94   | 21.63   | 27.34   | 24.30   |
| + annealing Adam                            | 24.05   | 21.90   | 27.26   | 24.37   |
| 4 layers with WN                            | 24.27   | 21.61   | 27.64   | 24.54   |
| + annealing Adam                            | 24.46   | 21.8    | 27.42   | 24.54   |
| 4 layers with SD                            | 24.43   | 21.89   | 28.00   | 24.74   |
| + annealing Adam                            | 24.73   | 21.73   | 28.14   | 24.85   |
| 4 layers with SD and WN                    | 24.39   | 21.47   | 27.61   | 24.47   |
| + annealing Adam                            | 24.69   | 21.69   | 28.04   | 24.79   |
| Deep transition                             | 23.83   | 21.51   | 27.15   | 24.13   |
| + annealing Adam                            | 23.75   | 21.37   | 27.06   | 24.03   |
| Deep transition with WN                     | 23.85   | 21.77   | 27.66   | 23.74   |
| + annealing Adam                            | 24.21   | 21.92   | 27.43   | 24.49   |
| Deep transition with SD                     | 24.04   | 21.53   | 27.43   | 24.31   |
| + annealing Adam                            | 24.47   | 22.1    | 27.98   | 24.82   |
| Deep transition with SD and WN              | 23.7    | 21.7    | 26.5    | 23.74   |
| + annealing Adam                            | 24.41   | 21.64   | 27.65   | 24.55   |
| ensemble                                    | 25.86   | 23.21   | 29.41   | 26.13   |
| + r21 reranking                             | 26.21   | 23.61   | 30.35   | 26.68   |
2 Bilingual task

Result Analysis

Annealing Adam:

Halving learning rate after early stop and trained from the previous best model.

Result:

0 to 0.81 BLEU score improvement.

M. J. Denkowski and G. Neubig. Stronger baselines for trustable results in neural machine translation.
2 Bilingual task

Result Analysis

Back translation:

Monolingual data was translated by a shallow model trained with parallel data from target to source. And training with a mix of parallel and synthetic data.

Result:

0 to 0.88 BLEU score improvement.

-R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data.
2 Bilingual task

Result Analysis

Weight normalization:

A reparameterization of the weight vectors in a neural network that decouples the length of those weight vectors from their direction.

Result:

A fluctuation of -0.57 to 0.81 BLEU score.
2 Bilingual task

Result Analysis

Ensemble decoding:

Ensembling of the independent left-to-right models:

Result:

0.97 to 1.28 BLEU score improvement.
2 Bilingual task

Result Analysis

Right-to-left reranking:

Training right-to-left models and re-scoring the n-best lists that are produced by the left-to-right models.

Result:

0.97 to 1.28 BLEU score improvement.

R. Sennrich, B. Haddow, and A. Birch. Edinburgh neural machine translation systems for WMT 16.
2 Bilingual task

Result Analysis

Table 6 Results on Official Test Sets for bilingual task

| direction | tst2016 | tst2017 |
|-----------|---------|---------|
| en-zh     | 28.13   | 28.30   |
| zh-en     | 21.35   | 22.16   |
Contents

1. Task description
2. Bilingual task
3. Multilingual task
4. Summary
5. Q&A
3 Multilingual task

Data Preprocessing

- Tokenizer and Truecase (Moses).
- Joint BPE for all corpus.
- Add a label at the start of each source sentence
  - consists of source language label and target language label
3 Multilingual task

Model architecture

Encoder Decoder with Attention
### 3 Multilingual task

#### Model configuration

Table 7 Model configuration for multilingual task in zero-shot condition

| Type                                | value  |
|-------------------------------------|--------|
| Source vocabulary size              | 40000  |
| target vocabulary size              | 40000  |
| word embedding                      | 512    |
| hidden units                        | 1024   |
| embedding dropout                   | 0.2    |
| hidden dropout                      | 0.2    |
| source dropout                      | 0.1    |
| target dropout                      | 0.1    |
| layer normalization                 | True   |
| maximum sentence length             | 80     |
### 3 Multilingual task

#### Result Analysis

Table 8 Case-insensitive BLEU score in development set of the zero-shot condition. WN means weight normalization.

|                         | en-de | en-nl | en-it | en-ro | de-en | de-it | de-ro | nl-en | nl-it |
|-------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| **shallow model**       | 28.29 | 32.22 | 29.67 | 27.56 | 34.43 | 20.60 | 19.47 | 38.01 | 22.42 |
| + annealing Adam        | 28.79 | 32.70 | 30.13 | 28.03 | 34.46 | 20.9  | 19.76 | 38.27 | 22.43 |
| **shallow model with WN** | 27.68 | 32.63 | 29.82 | 27.32 | 34.15 | 20.50 | 19.36 | 37.78 | 21.90 |
| + annealing Adam        | 27.79 | 32.56 | 30.15 | 27.72 | 34.42 | 20.82 | 19.81 | 38.03 | 22.05 |
| **deep transition**    | 29.43 | 32.79 | 30.86 | 28.96 | 35.33 | 21.93 | 20.54 | 39.45 | 23.48 |
| + annealing Adam        | 29.9  | 32.85 | 31.56 | 28.78 | 35.72 | 22.18 | 20.91 | 39.79 | 23.67 |
| **deep transition with WN** | 28.85 | 33.19 | 30.98 | 28.37 | 34.83 | 22.07 | 20.28 | 38.96 | 23.06 |
| **ensemble**            | 29.82 | 34.22 | 31.98 | 29.39 | 36.50 | 22.8  | 21.32 | 40.31 | 23.84 |
| + r21 reranking         | 29.60 | 32.70 | 31.58 | 28.77 | 35.76 | 22.48 | 21.45 | 39.50 | 24.22 |

|                         | nl-ro | it-de | it-en | it-nl | ro-de | ro-en | ro-nl | average |
|-------------------------|-------|-------|-------|-------|-------|-------|-------|---------|
| **shallow model**       | 20.79 | 20.75 | 34.22 | 22.1  | 22.05 | 35.81 | 23.15 | 27.28   |
| + annealing Adam        | 21.31 | 20.85 | 34.61 | 22.22 | 22.26 | 36.06 | 23.34 | 27.56   |
| **shallow model with WN** | 21.15 | 20.64 | 34.25 | 21.87 | 22.09 | 35.62 | 22.58 | 27.3    |
| + annealing Adam        | 20.78 | 20.29 | 33.71 | 22.04 | 21.63 | 35.31 | 22.48 | 27.05   |
| **deep transition**    | 22.13 | 21.51 | 35.25 | 22.99 | 22.84 | 37.06 | 23.29 | 28.3    |
| + annealing Adam        | 22.16 | 22.20 | 35.99 | 23.29 | 23.16 | 37.71 | 23.53 | 28.66   |
| **deep transition with WN** | 21.83 | 21.55 | 35.13 | 22.86 | 22.73 | 37.09 | 23.63 | 28.17   |
| **ensemble**            | 22.93 | 22.56 | 36.15 | 23.93 | 23.35 | 38.05 | 24.49 | 29.21   |
| + r21 reranking         | 22.74 | 24.41 | 35.74 | 23.76 | 23.68 | 37.47 | 24.61 | 28.99   |
3 Multilingual task

Result Analysis

Annealing Adam:

Halving learning rate after early stop and trained from the previous best model.

Result:

0.28 to 0.36 BLEU score improvement.

M. J. Denkowski and G. Neubig. Stronger baselines for trustworthy results in neural machine translation.
3 Multilingual task

Result Analysis

Weight normalization:

A reparameterization of the weight vectors in a neural network that decouples the length of those weight vectors from their direction.

Result:

Get worse performance.

T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks.
3 Multilingual task

Result Analysis

Ensemble decoding:

Ensembling of the independent left-to-right models.

Result:

1.93 BLEU score improvement comparing shallow model.
3 Multilingual task

Result Analysis

Right-to-left reranking:

Training right-to-left models and re-scoring the n-best lists produced by the left-to-right models.

Result:

right-to-left re-ranking didn’t improve the performance of model.

R. Sennrich, B. Haddow, and A. Birch. Edinburgh neural machine translation systems for WMT 16.
3 Multilingual task

Result Analysis

Table 9: Results on Official Test Sets for multilingual task.

| direction | en-de | en-nl | en-it | en-ro | de-en | de-it | de-ro | de-nl | nl-en | nl-it |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| BLEU      | 23.08 | 29.08 | 32.84 | 23.89 | 28.04 | 18.56 | 16.23 | 19.59 | 32.78 | 21.21 |
| Nist      | 5.86  | 6.81  | 7.22  | 5.91  | 6.85  | 5.36  | 4.69  | 5.57  | 7.42  | 5.72  |
| Ter       | 60.63 | 51.46 | 47.63 | 58.81 | 51.41 | 63.43 | 69.04 | 61.26 | 47.34 | 60.83 |

| direction | nl-ro | nl-de | it-de | it-en | it-nl | it-ro | ro-de | ro-en | ro-nl | ro-it |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| BLEU      | 18.11 | 17.95 | 18.09 | 37.84 | 21.80 | 18.62 | 17.95 | 31.79 | 20.02 | 20.39 |
| Nist      | 4.97  | 5.06  | 5.09  | 8.10  | 5.78  | 5.03  | 5.06  | 5.59  | 5.59  | 5.57  |
| Ter       | 66.55 | 67.02 | 67.28 | 41.05 | 60.09 | 65.53 | 67.02 | 41.22 | 67.81 | 61.11 |
Contents

1. Task description
2. Bilingual task
3. Multilingual task
4. Summary
5. Q&A
4 Summary

- Annealing Adam training trick
- Deep model
- Weight normalization
- Right-to-left re-ranking
  - For multilingual task
  - For bilingual task
## 4 Summary

| Team      | System on 4 languages | BLEU  |
|-----------|-----------------------|-------|
| KYOTO     | ML ZS                 | 21.13 |
| GTCOM     | ML ZS                 | 19.40 |
| FBK       | ML ZS                 | 17.26 |
| UDSDEFKI  | ML ZS                 | 17.10 |

| Team      | System on other 16 languages | BLEU  |
|-----------|-------------------------------|-------|
| GTCOM     | ML ZS                         | 24.46 |
| KYOTO     | ML ZS                         | 24.10 |
| FBK       | ML ZS                         | 21.89 |
| UDSDEFKI  | ML ZS                         | 21.63 |
4 Summary

- Transformer > RNN
- Using RNN ML SD > ML ZS
- Using transformer ML ZS ≈ ML SD
Cooperation:

• Neural machine translation
• Our system is available on [http://translateport.yeekit.com:4305/index.html](http://translateport.yeekit.com:4305/index.html)
• Other language pair contribution
• Corpus
• Speech recognition
Opportunity:
• Internship
• Full time job on machine translation

GTCOM:
Shijingshan district, Beijing, China.
Thank you!