Description and Findings of OPPO's Machine Translation Systems for CCMT 2020

2020/10/11

Tingxun Shi
OPPO Research Institute
shitingxun@oppo.com

Contents

Data Processing - Text Conversion in General

- Simplify Chinese characters
- Normalize punctuations
  - Symbol unification. E.g. Convert different hyphens to the ASCII one（-, ASCII Code 45）
  - Convert all punctuations in Chinese corpus to half width form (except for full stops, commas, question marks and exclamation marks)
- Segment Chinese corpus (pkuseg)
- Tokenize (moses)
- True case

Data Processing - Special Conversions

- Multilingual translation
  - Convert all CJK characters in Japanese corpus to Kanji
  - Segment corpus using mecab
- Minority languages translation
  - More careful data processing: we listed all non-Chinese characters and design ad hoc rules
    - Unify non-alphabet symbols (e.g. Tibetan numbers)
    - Delete invalid/invisible symbols
  - Special process for Tibetan (see our report for more details)
  - Jointly use different segmentation tools for Chinese

Data Processing - Text Filtering

- Heuristic filtering: Remove sentence pairs that
  - Contain too many non-sense characters (e.g. Emoji)
  - Contain too long sentences (count of words > 160)
  - |count_number(src) - count_number(tgt)| >= 3
  - |count_punc(src) - count_punc(tgt)| >= 5
  - Len(en) / len(zh) < 0.7 or > 2.2
  - Deduplication
- Alignment-based filtering
  - Get alignment scores using fast_align
  - Remove pairs that sentence-level score < -16 or word-level score < -2.5
- Setting the threshold
  - Fixed threshold by experiences
  - Based on statistical information (0.1 or 99.9 percentile)

Results of Text Filtering

Task	# Pairs before Cleaning	# Pairs after Cleaning	Retention Rate
EnZh/ZhEn	28M	17M	60.71%
JaEn	JaZh: 3M EnZh: 3M	JaZh: 2.9M EnZh: 2.8M	JaZh: 96.67% EnZh: 93.33%
UgZh	169,525	163,762	96.60%
BoZh	162,096	147,440	90.96%
MnZh	261,454	228,225	96.18%

Applying Multiple Segmentation Tools

- Inspiration: multilingual translation
- Segement Chinese corpus using different segmentation tools (e.g. jointly use pkuseg and jieba. For Uighur we also use scws), and combine the results with character-based texts
  - Add symbol "<tag>" to mark how the sentence is segmented, for BOTH source and target corpus
  - Remove BPE suffices "@@" for Chinese corpus
    - Chinese has no explicit words boundaries, post-processing is not a problem
    - Component of a subword may have the exactly same meaning as the individual word, e.g. "国际" and "国际@@" in "国际@@ 贸易" (if "国际贸易" is separated by BPE)
  - Shatter low frequency words to characters

Applying Multiple Segmentation Tools

Method	Validation set BLEU	Improvement	Online test BLEU
Baseline model (Character-based)	44.2	-/-	54.74
+ pkuseg segmentation	45.4	+1.2/+1.2	(not tested)
+ Multiple segmentation w/o segmentation tag	45.7	+1.5/+0.3	(not tested)
+ Segmentation tag & keeping BPE symbol	46.1	+1.9/+0.4	(not tested)
+ Removing BPE symbol	46.2	+2.0/+0.1	55.90
+ selected by kenlm	46.7	+2.5/+0.5	56.69

Applying Multiple Segmentation Tools

Method	Validation set BLEU	Improvement	Online test BLEU
Baseline model (Character-based)	44.2	-/-	54.74
+ pkuseg segmentation	45.4	+1.2/+1.2	(not tested)
+ Multiple segmentation w/o segmentation tag	45.7	+1.5/+0.3	(not tested)
+ Segmentation tag & keeping BPE symbol	46.1	+1.9/+0.4	(not tested)
+ Removing BPE symbol	46.2	+2.0/+0.1	55.90
+ selected by kenlm	46.7	+2.5/+0.5	56.69

Case by case, different across tasks

Applying Multiple Segmentation Tools

- Why?
  - Data augmentation?
  - A sentence can be segmented into different forms, thus improve the model's robustness?
  - Interaction between character-based models and multiple word-character mixture models?
- Has no impact on high-resource tasks

Model Training

- Architecture: Transformer-Big
  - For EnZh, Dimension of the FFN is 15,000
- Framework: fairseq
- BPE
  - EnZh/ZhEn/EnJa: 32K joint BPE, separated vocabulary
  - UgZh/BoZh: 32K separate BPE
  - MnZh: 16K separate BPE

Main Techniques

- Back-translation
  - ZhEn/EnZh tasks: argmax-based back-translation performed better than noisy back-translation
  - Minority languages tasks: add tag <bt>
  - ZhEn/EnZh tasks: also benefitted from forward-translation
- Domain adaptation
  - "Translationese problem" —— fine-tune using original parallel corpus
  - Domain mismatch
    - fine-tune using validation set
    - fine-tune using corpus that similar to the test set (selected by FDA algorithm), try different corpus sizes (10K, 100K, 1M ...)
- Model Ensemble
- Reranking according to multiple features: K-Batched MIRA or Noisy Channel
- Multilingual training (for JaEn task only)

Final Results - EnZh Task

1st in the leaderboard

Method	Validation set BLEU	Absolute Improvement	Relative Improvement
Baseline (trained by parallel corpus only)	38.6	-	-
+ back-translation	39.1	+0.5	+0.5
+ fine-tuned by parallel corpus	40.6	+2.0	+1.5
+ fine-tuned by newstest2017	41.3	+2.7	+0.7
+ forward-translation	41.9	+3.3	+2.8
+ ensemble	42.7	+4.1	+0.8
+ reranking	43.2	+4.6	+0.5

Final Results - ZhEn Task

1st in the leaderboard

Method	Validation set BLEU	Absolute Improvement	Relative Improvement
Baseline (trained by parallel corpus only)	28.8	-	-
+ back-translation	29.8	+1.0	+1.0
+ forward-translation	34.5	+5.7	+4.7
+ fine-tuned by newstest2017	36.7	+7.9	+2.2
+ ensemble & reranking	38.3	+9.5	+1.6

Using the two models trained in these two tasks as scorers, we also ranked 1st in the Corpus Filtering task (500M English words subset)

Final Results - JaEn Task

1st in the leaderboard

Method	Validation set BLEU	Absolute Improvement	Relative Improvement
Baseline (trained by parallel corpus only)	37.8	-	-
+ forward-translation	39.5	+1.7	+1.7
+ multi-lingual processing	40.5	+2.7	+1.0
+ ensemble	41.1	+3.3	+0.6
+ reranking	41.5	+3.7	+0.4

Final Results - Minority Languages Task

Method	Uighur	Tibetan	Mongolian
Baseline (trained by parallel corpus only)	38.6	46.7	61.4
+ back-translation & ensemble kd	48.6 (+10.0)	47.9 (+1.2)	63.9 (+2.5)
+ fine-tune on the original parallel corpus	49.0 (+0.4)	50.0 (+2.1)	66.9 (+3.0)
+ model ensemble	49.4 (+0.4)	53.0 (+3.0)	69.5 (+2.6)
+ reranking	49.5 (+0.1)	53.0 (+0.0)	73.0 (+3.5)
	1st	1st	2nd

Conclusions

- Applying multiple segmentation tools helps on the low-resource translation tasks
- Forward-translation could bring gains as well as back-translation
- Fine-tune could contribute a lot if there is a domain mismatch
- Impact brought by back-translation also depends on domain
- Model performance generally benefits from model ensemble and reranking
- What one loses on the swings, he gets back on the roundabouts

Thank you

OPPO