Domain Adaptation of Machine Translation with Crowdworkers

Makoto Morishita[1], Jun Suzuki[2], Masaaki Nagata[1]
[1]NTT Communication Science Laboratories, [2]Tohoku University

Abstract

NMT models are weak in domains where they have not been trained
→ If you want to translate these domains, need domain-adaptation
→ However, currently, we can adapt to the limited domains due to training data scarcity
→ We need to collect in-domain data efficiently for domain-adaptation

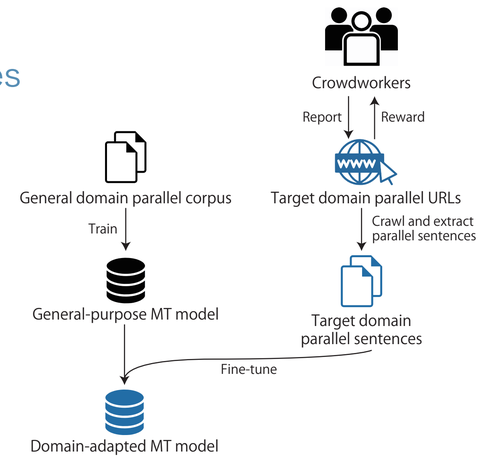
We proposed a method to collect in-domain parallel sentences rapidly from the web with crowdworkers
→ Our model drastically improved the accuracy on the target domain with the collected in-domain parallel sentences

Collecting In-domain Parallel Corpus

Co-operate with Crowdworkers

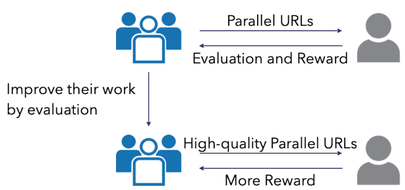
We asked crowdworkers to report URLs that contain target-domain parallel sentences
→ Hypothesis: People can quickly find target-domain parallel websites based on their experience

A previous method asked workers to translate in-domain monolingual sentences
→ Translation is a difficult task and requires substantial cost and time
→ Collecting parallel URLs is much easier, and many workers can do



Variable Reward

We varied the reward based on the quality of reported URLs to improve workers' performance
→ Our system returns evaluation results and rewards to the workers once we get parallel URLs
→ The workers can learn how to improve their performance



Evaluation Criterion

- Number of extracted parallel sentences
- Translation quality based on the sentence aligner
- Domain similarity based on the sentence embeddings

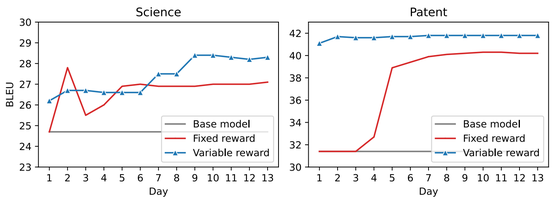
Experiments

Settings

Language: En-Ja
Domain: Science, Patent, COVID-19, News, Legal
General-purpose model: Transformer big trained with JParaCrawl v2.0 (10M sents)
Crowdsourcing: 97 workers, 13 days

Fixed or Variable Reward?

Variable rewards achieved better accuracy than fixed reward
→ Workers might get motivated to find good websites



Results on Other Domains

Our method drastically improved the BLEU scores on other domains as well
→ Up to +20 points compared to the baseline
→ Crowdsourcing costs around 2,000 USD for each domain. This is quite reasonable than asking workers to translate

Comparison with Moore-Lewis

Moore-Lewis: A method extracting similar sentences from corpora based on the language model
→ Moore-Lewis slightly improved the accuracy (green line)
→ However, our method significantly surpassed it

Ref: Intelligent Selection of Language Model Training Data, Robert C. Moore and William Lewis, ACL 2010