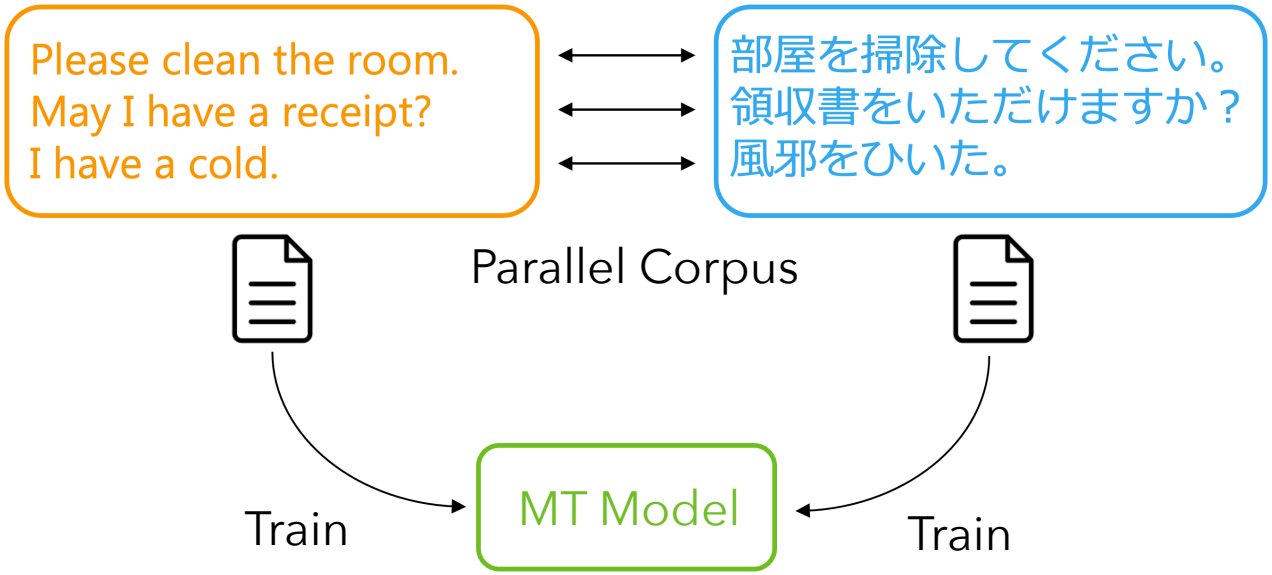
Domain Adaptation of Machine Translation with Crowdworkers

Makoto Morishita, Jun Suzuki, Masaaki Nagata

NTT Communication Science Laboratories, Tohoku University

EMNLP 2022 Industry Track

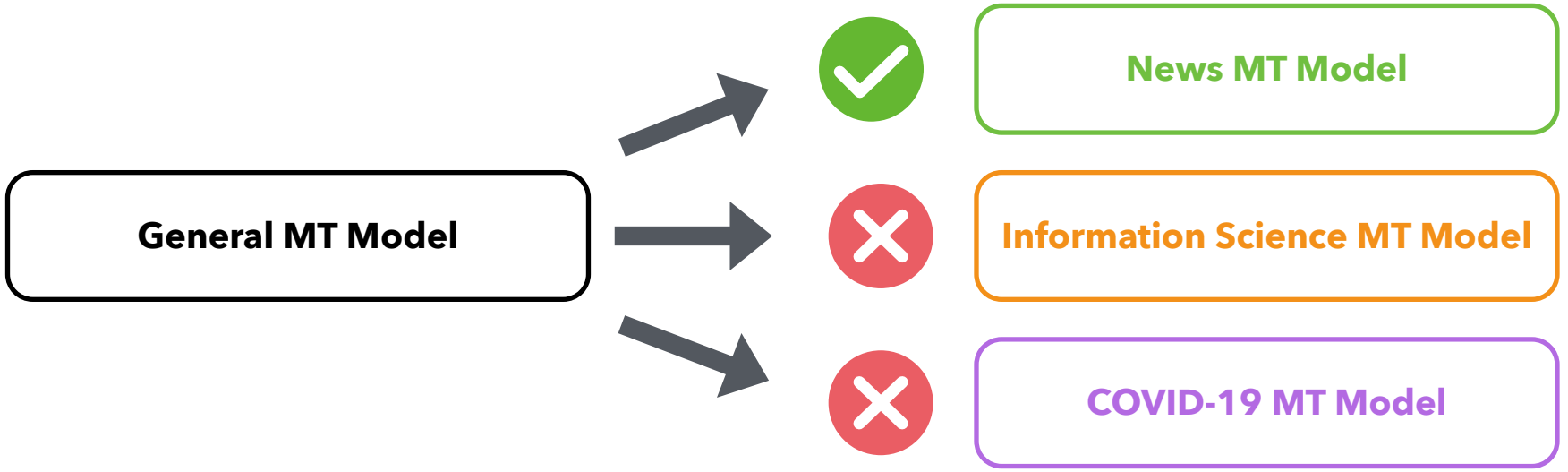
MT Model Training



- Current MT models are mainly trained with a parallel corpus
  - The model is specialized in the training domain
  - MT models are weak in domains where they have not been trained
    - If you want to translate these domains
      → Domain-adaptation

[Müller et al., AMTA 2020]

Limitations of domain-adaptation



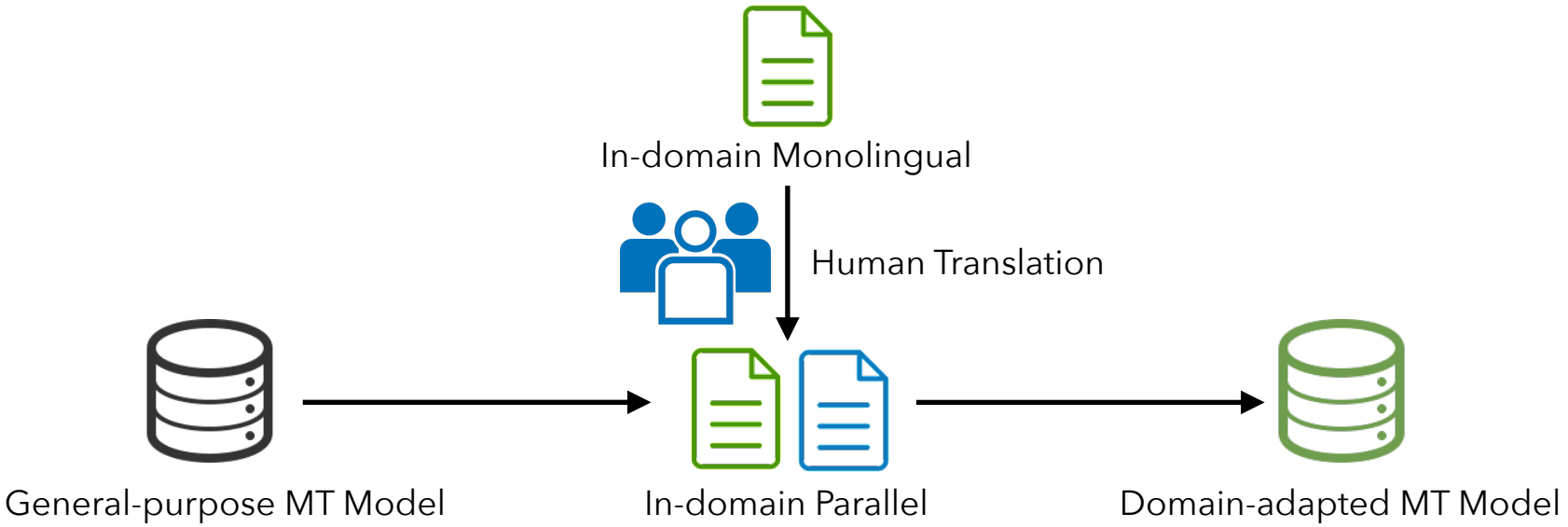
We want to adapt the model to various domains

→ However, currently, we can adapt to the limited domains

→ Due to training data scarcity

→ Requires a method to collect in-domain efficiently

A Typical Method for Parallel Sentence Collection

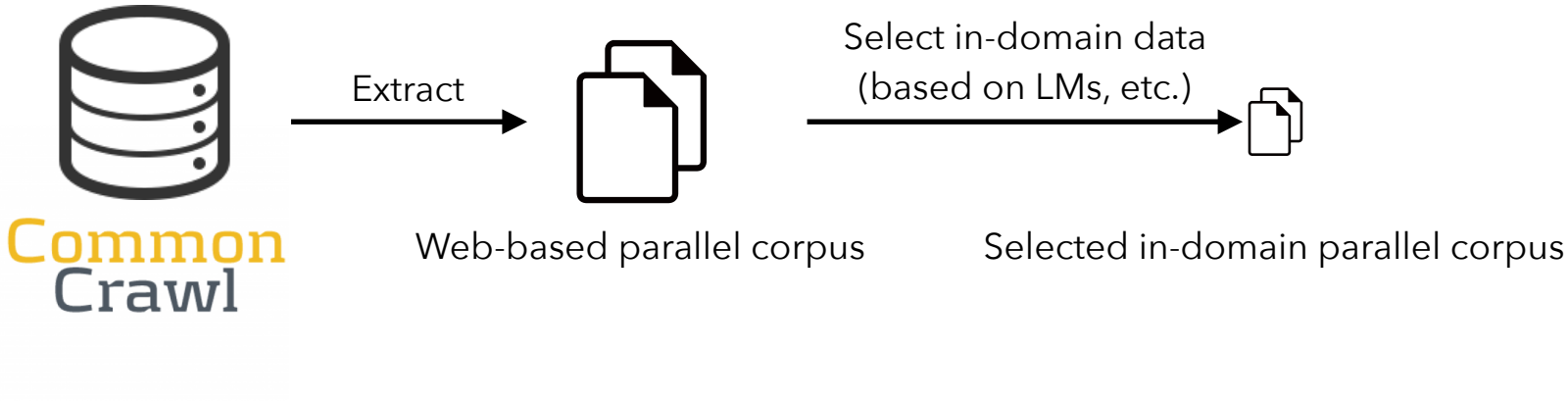


Create parallel data by human translators

→ A limited number of workers who can translate

→ Require substantial cost, and it is time-consuming

Select In-domain Data from Web-based Corpora



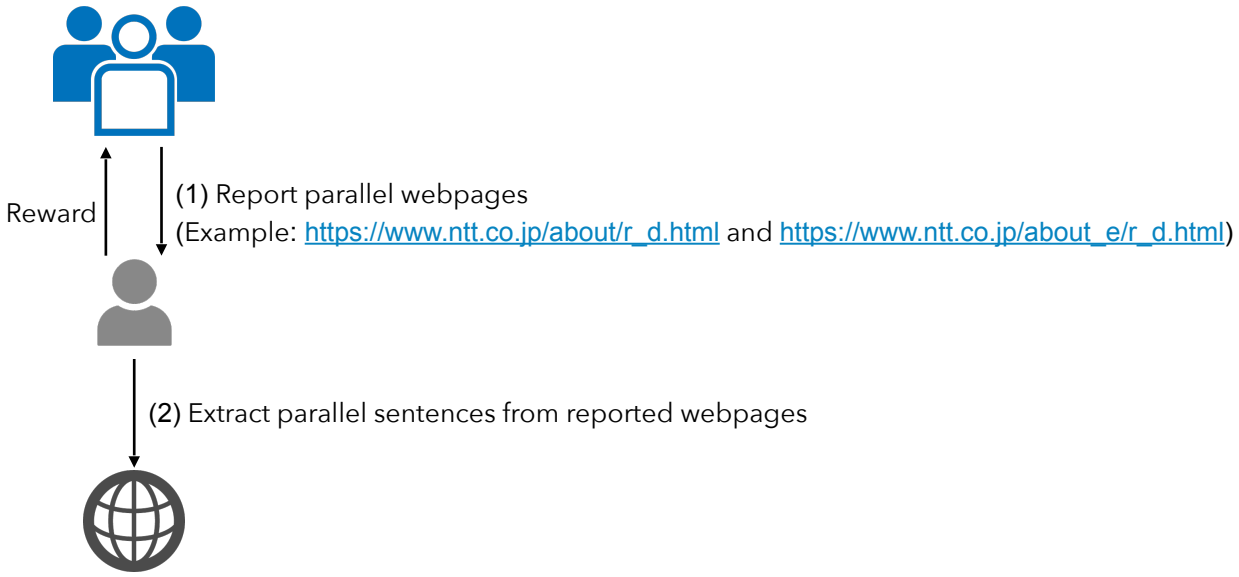
- Extracting in-domain sentences from the web-based corpus
- Pros
  - Easy
- Cons
  - It does not include enough data for some domains
    - CommonCrawl does not cover the whole web

Method: Co-operate with Crowdworkers

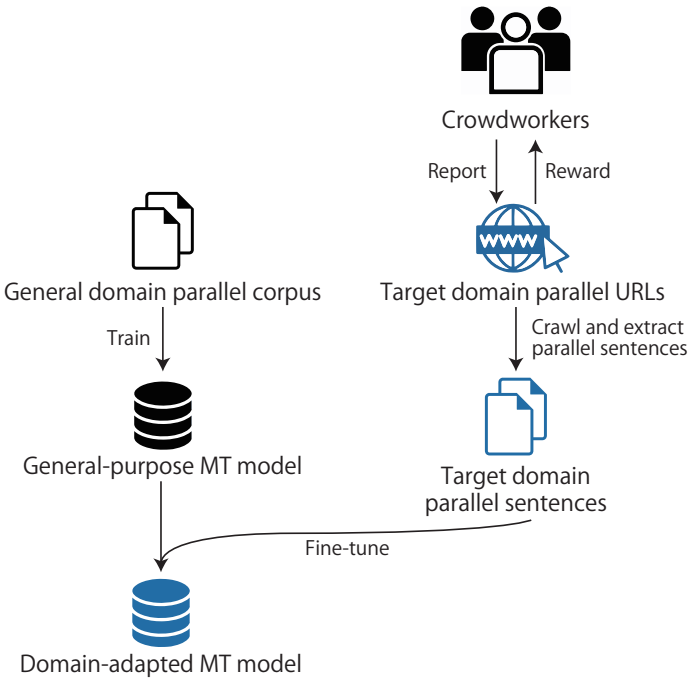
Hypothesis: People can quickly find parallel websites based on their experience

→ Ask crowdworkers to report parallel webpages

→ It is easier than asking for translations and is reasonable

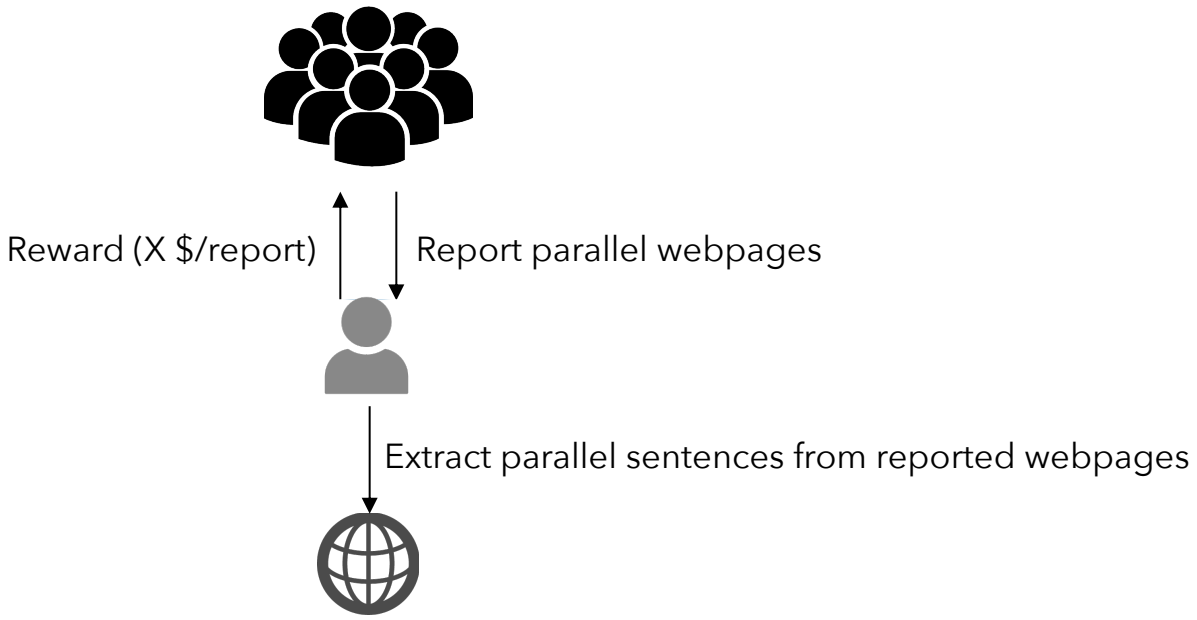


Domain-adaptation with Crowdworkers



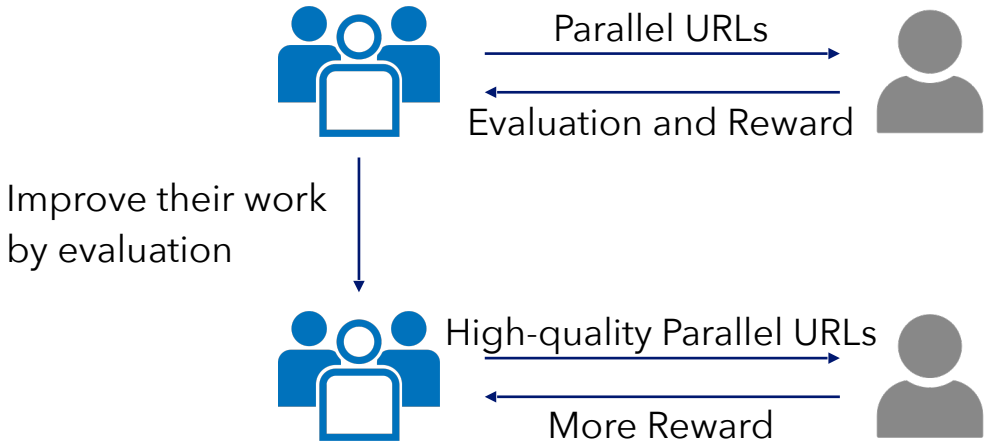
- Receive parallel URLs reported from crowdworkers
  → Extract parallel sentences from the reported URLs
  → Adapt the model to the target domain

Fixed Reward



- Pay a fixed reward per report
  → Workers may report easy-to-find webpages, that may not be useful for domain-adaptation.

Change Workers' Behavior by Reward



For each reported URL, we return its evaluation result and reward

→ high-quality URLs:

a large number of sentences, high translation quality, similar domain

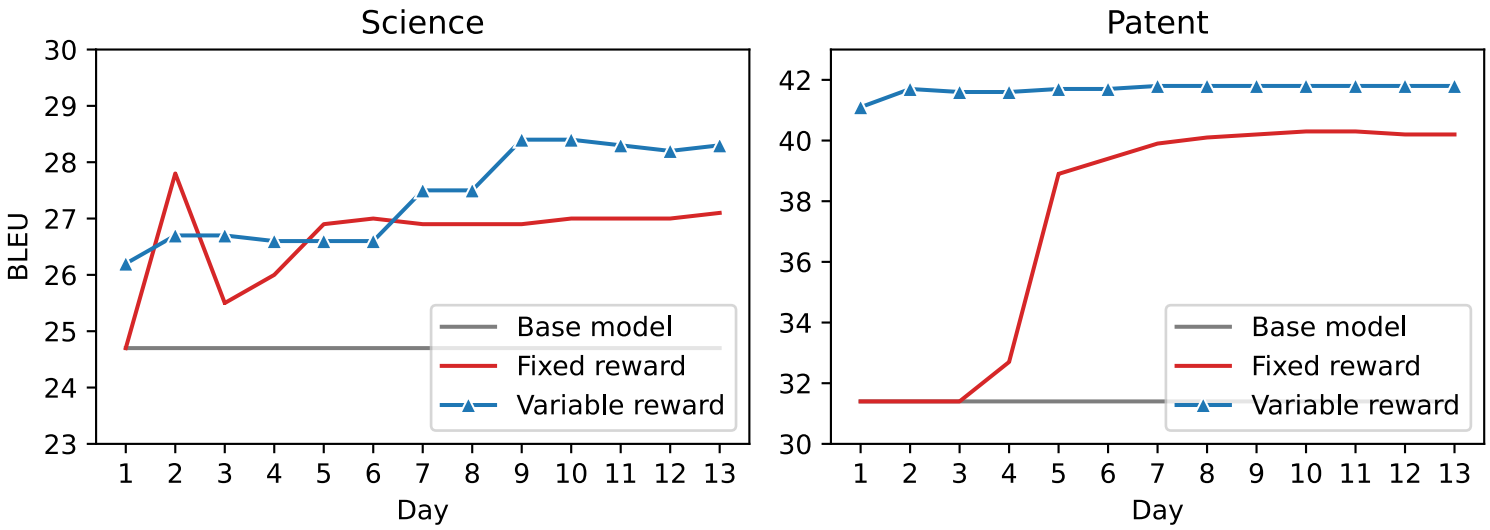
→ The worker will change their behavior to earn higher rewards efficiently

Experiments

Experimental Settings

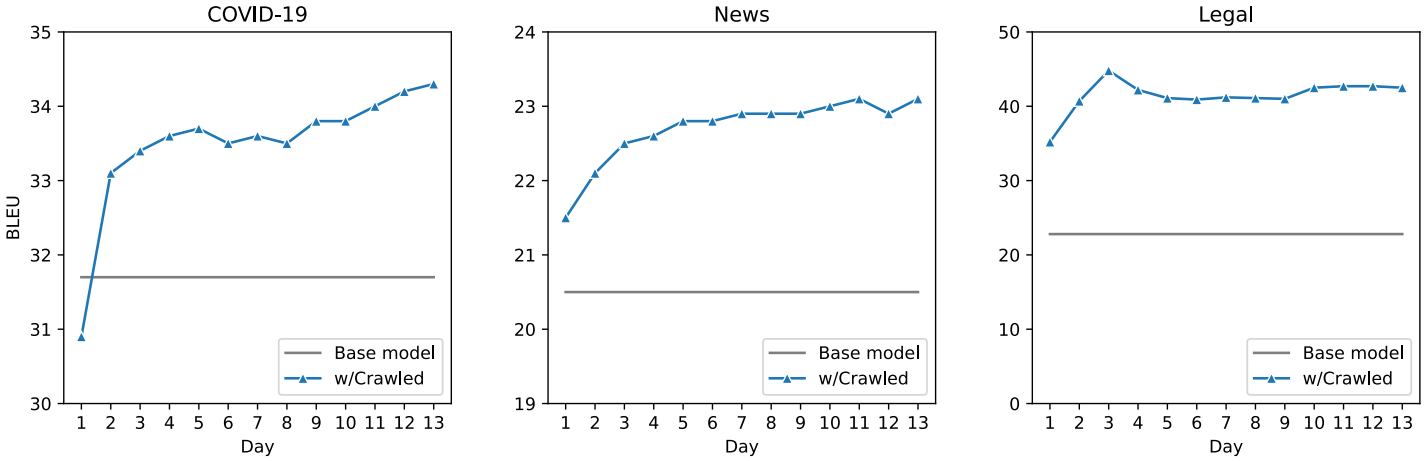
- Language: En-Ja
- Base model: Transformer trained with JParaCrawl (web-based parallel corpus)
- Target domain: Science, Patent, COVID-19, News, Legal
- Crowdsourcing: 97 workers, 13 days
- Rewards: Fixed or Variable

Fixed or Variable Reward?



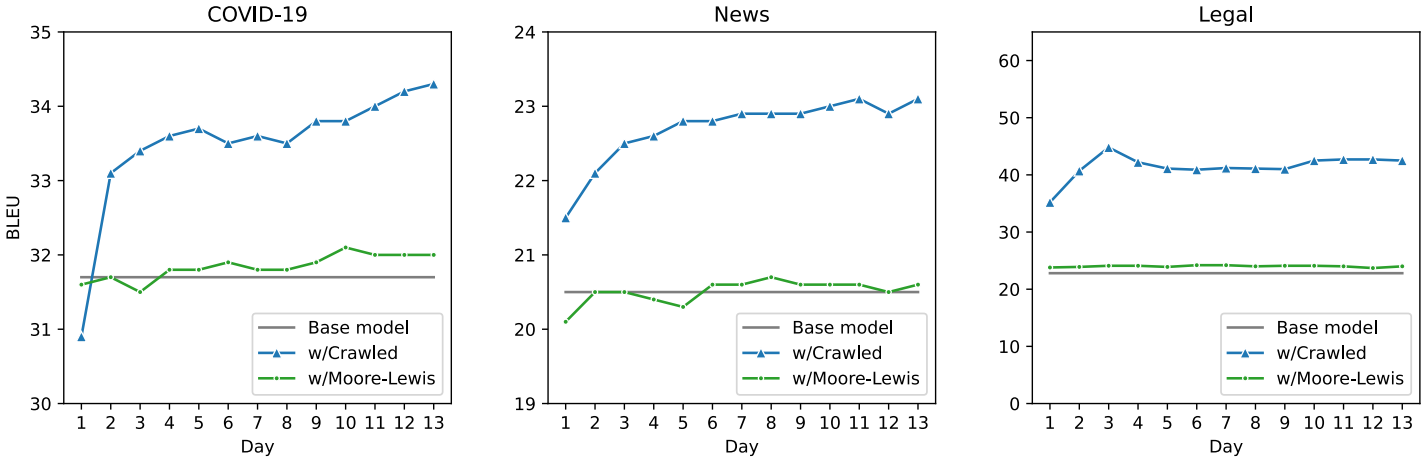
- Variable rewards achieved better accuracy than fixed
  - Workers might be motivated to find good websites
  - After 13 days of collection,
    science domain improved by +3 points
    patent domain improved by +10 points

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, which is quite reasonable than asking workers to translate

Compared to the Previous Method



- Moore-Lewis: Select in-domain data from web-based corpora based on the language model scores
- Moore-Lewis's method slightly improved the scores, but our method clearly surpassed it

Conclusion

- We proposed a method to collect parallel sentences in the target domain rapidly
  - It is reasonable and faster than asking translation
  - Two types of rewards: Fixed and variable rewards
    - Variable rewards changed workers' behavior to collect high-quality data
  - We could train a domain-specific MT model with collected target-domain parallel sentences in a few days

END