Combining Translation Memory with Neural Machine Translation

Aikko Eriguchi, Spencer Rarrick, and Hitokazu Matsushita (Microsoft Corporation)

1. Timely Disclosure Documents Corpus

- Time-series data (2016-2018)

Figure 1. Samples of original timely disclosure document.

Characteristics of Timely Disclosure Documents Corpus
- Timely disclosure documents provided by companies to the public every year
- Financial statements, corporate actions, governance policies
- Duplicated src/tgt sentences or sentence pairs found between Train. and Eval.
- Dev/Test respectively contain 26%/28% duplicated sentence pairs of Train.

2. Proposed Method

Combining Translation Memory with Neural Machine Translation (NMT)
- Keep provided training data as Translation Memory (TM)
- Train white-box MT model from scratch on provided training data, or
- Employ black-box MT systems (e.g. Microsoft Translator)

3. Results and Discussion

Results

Table 2: General translation performance on the corpus (ITEM+TEXT).

Table 3: Results of our proposed approach. * Employed IDF-based retrieval.

Retrieval Approaches over Translation Memory

1. Edit distance
   - Expected to capture similarity well on shorter sentences (ITEM)

2. IDF-based (Bapna and Firat, 2019)
   - Expected to work well on longer sentences (TEXT)
   \[ \text{Sim}_{\text{IDF}}(S_1, S_2) = \frac{1}{|C_{TM}|} \sum_{t \in S_1 \cup S_2} \left( f_t - \sum_{t \in S_1 \cap S_2} f_t \right), \]
   \[ f_t = \log \frac{|C_{TM}|}{n_t}, \]
   
   \( C_{TM} \): Number of sentence pairs in TM
   \( n_t \): number of occurrences of token \( t \)

Human evaluation results and Evaluation example

Table 4: Human Evaluation results (score \( \in [0, 100] \)). Bold indicates indistinguishable from HREF (p < 0.05).

Experiment Setup

- Train white-box MT systems: Transformer (base) (Vaswani et al., 2017)
- Transformer A, B trained respectively with 200k/80k updates
- Transformer B trained with 80k updates
- Employ online MT services as black-box MT systems
- Google Translate
- Microsoft Translator
- Mirai Translate
- * All the systems as of July, 2019.
- * Anonymized to Online (A, B, C) in random order in results.

4. Conclusion

- Proposed a simple method of combining TM with NMT
- Improved variety of vanilla NMT system performance
- Human evaluation results suggest drawbacks of sentence-level translation systems/human evaluation