Combining Translation Memory with Neural Machine Translation

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

1. Timely Disclosure Documents Corpus

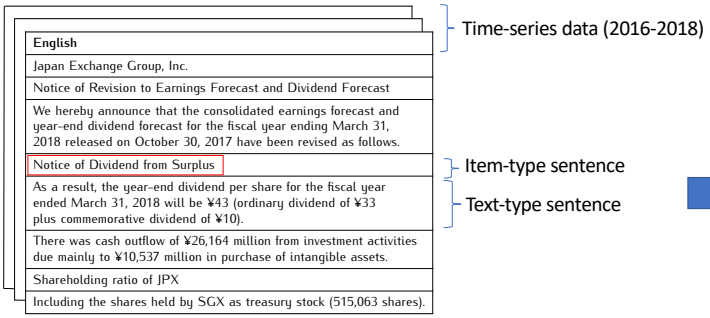
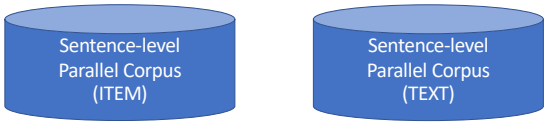


Figure 1. Samples of original timely disclosure document.



Each sampled sentence pair has a unique hash id, meaning that the same contents/sentence pairs may appear multiple times in a dataset.

	size	# Uniq. Ja	# Uniq. En
Train.	1.4M	583K	709K
Dev.	3.9K	3.5K	3.7K
Devtest	4.0K	3.7K	3.4K
Test	3.2K	2.8K	----

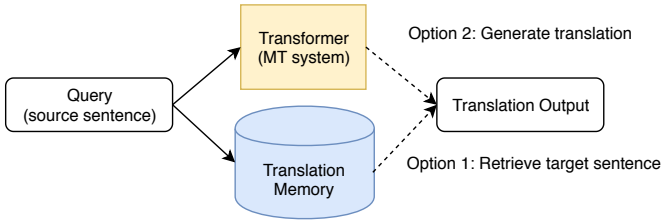
Table 1 Data statistics of Timely Disclosure Documents Corpus.

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/DevTest 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)

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)

$$\mathrm{Sim}_{idf}(S_1, S_2) = 2 \times \sum_{t \in (S_1 \bigcup S_2)} f_t - \sum_{t \in (S_1 \bigcap S_2)} f_t,$$

$$f_t = \log \frac{|C_{TM}|}{n_t}.$$

$|C_{TM}|$: Number of sentence pairs in TM
n_t: number of occurences of token t

3. Results and Discussion

Results

	Dev.	Devtest
Transformer A	48.6	50.6
Transformer B	40.9	42.2
Online A	24.8	24.8
Online B	24.5	24.4
Online C	24.5	24.5

Table2: General translation performance on the corpus (ITEM+TEXT).

+ TM

		threshold	Dev.	Devtest
ITEM	Transformer A	89	54.1	58.1
	Transformer B	89	53.3	57.0
	Online A	83	51.2	55.6
	Online B	80	51.8	55.8
	Online C	83	51.6	55.6
TEXT	Transformer A	18*	57.7	57.9
	Transformer B	14*	57.1	57.6
	Online A	15*	55.9	56.7
	Online B	10*	55.6	56.4
	Online C	80	55.7	56.8

Table3: Results of our proposed approach.
* Employed IDF-based retrieval.

- **Combining TM improves each NMT system dramatically on both data sets**
- Higher thresholds for retrieval suggest that white-box systems are more reliable
- Fill gap between white-box and black-box systems up to 1-2 BLEU by using TM

Experimental 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.

Human evaluation results and Evaluation example

	ITEM	TEXT	ALL
HREF	73.4	71.0	72.5
Transformer A	**73.2**	66.5	70.8
Transformer B	**73.0**	66.3	70.5
Online A	**71.9**	69.2	**70.9**
Online B	**71.4**	68.2	70.2
Online C	**71.7**	67.1	70.0

Table4: Human Evaluation results (score ∈ [0, 100]).
Bold indicates indistinguishable from HREF ($p < 0.05$).

- ITEM: all systems achieved human-parity (Hassan et al., 2018)
- TEXT: online systems are more highly evaluated than in-house systems due to more flexible translation capability

	score	ITEM
Source	—	依頼者提示資料に基づき査定
HREF	28	Based on materials provided by The Institute of Internal Auditors IIA
All systems	99	Assessed based on documents presented by the requester.

Table5: Translation example and its human evaluation score.

- Practically, consider which translation is preferred
- Sentence-level evaluation may lack context of a document

4. Conclusion

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

Microsoft