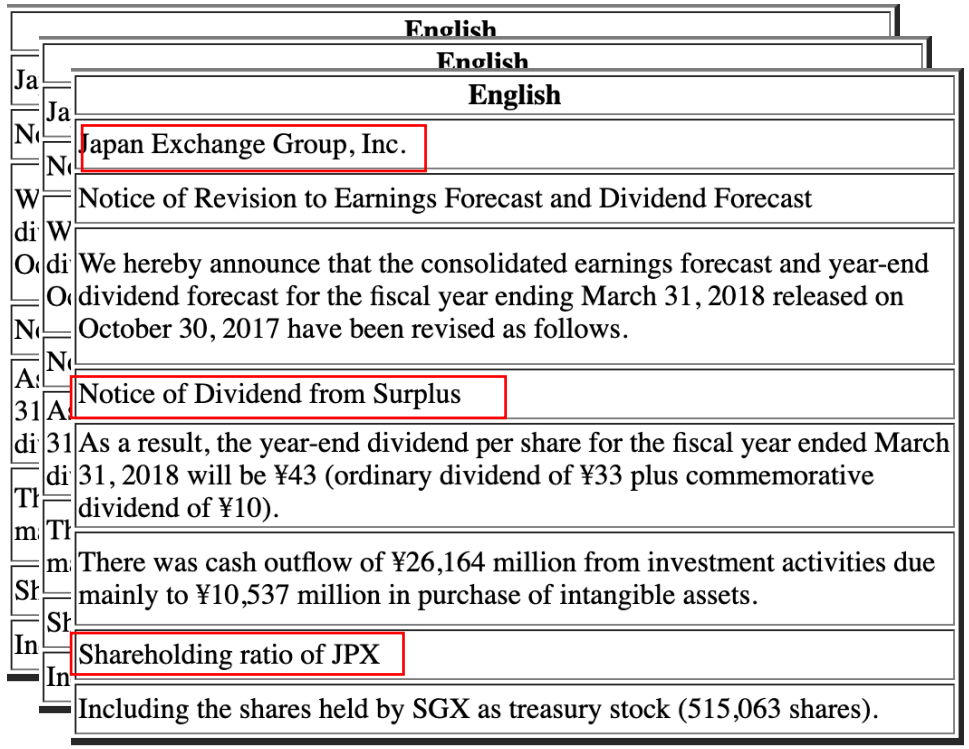
Combining Translation Memory with Neural Machine Translation

11/04/2019

Akiko Eriguchi, Spencer Rarrick, and Hitokazu Matsushita

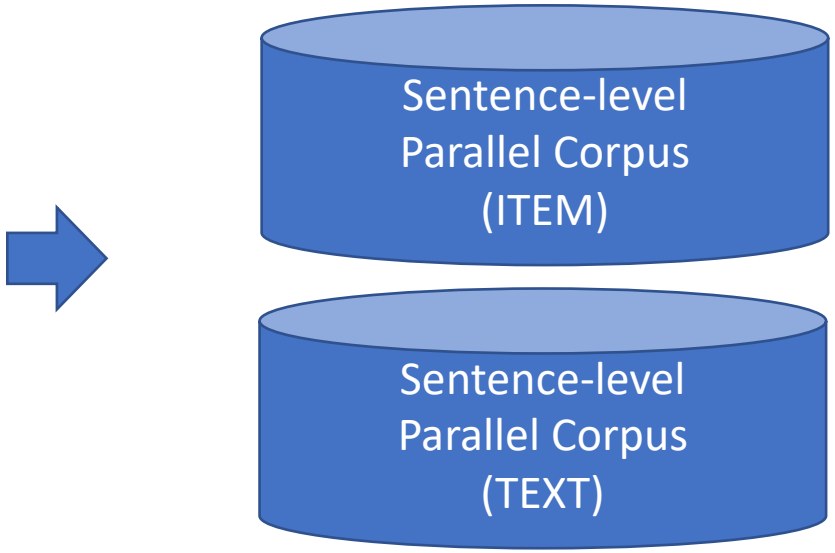
Microsoft

Timely Disclosure Documents Corpus (Ja->En)



Sample of original timely disclosure document

Time-series data
(2016-2018)



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

http://lotus.kuee.kyoto-u.ac.jp/WAT/Timely_Disclosure_Documents_Corpus/

Characteristics of Data sets

- Data sets contain duplicated src/tgt sentences or parallel pairs
- Such "duplicates" can be found between Train and Dev/Devtest/Test
  - Dev: 1,047 (26% duplicated sentence pairs of Train.)
  - DevTest: 1,117 (28% duplicated sentence pairs of Train.)

	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	----

General Translation Performance

- White-box MT systems (Transformer_base) show significantly better BLEU
- Lower performance in black-box MT systems, due to domain mismatch

	Dev	DevTest
Transformer A from scratch	48.6	50.6
Transformer B from scratch	40.9	42.2
Online A	24.8	24.8
Online B	24.5	24.4
Online C	24.5	24.5

White-box systems

Black-box systems

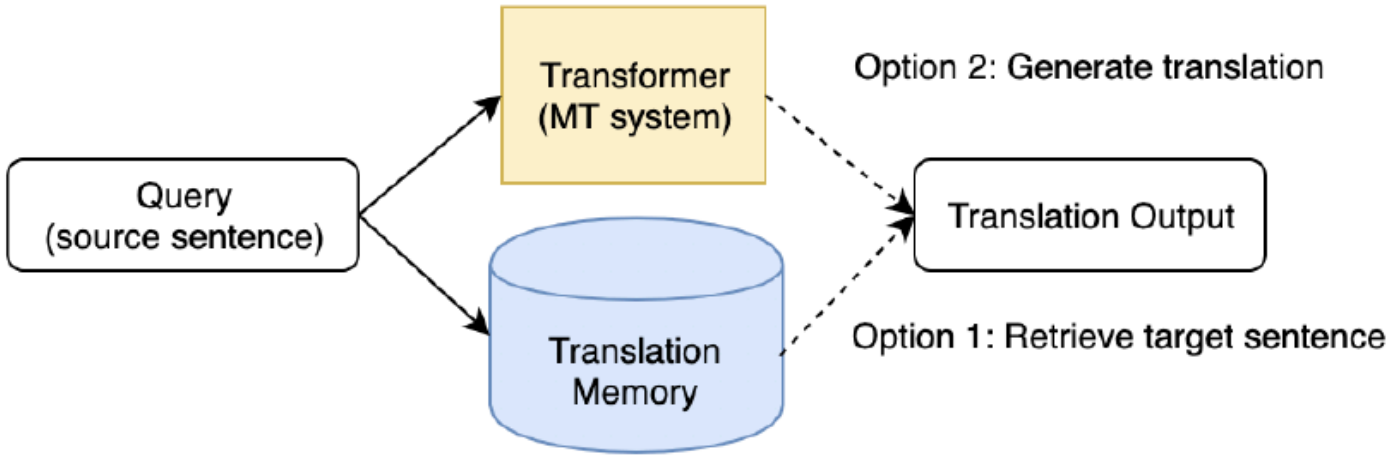
* Google Translate, Microsoft Translator, Mirai Translator as of July, 2019.
All systems are anonymized in random order

BLEU score results

Proposed Approach

- Combine Translation Memory with Neural Machine Translation system

  - Train **white-box MT model** from scratch on provided training data, or
  - Employ **black-box MT** systems (e.g. Microsoft Translator)



  - Keep provided training data as **Translation Memory**