

Combining Translation Memory with Neural Machine Translation

11/04/2019

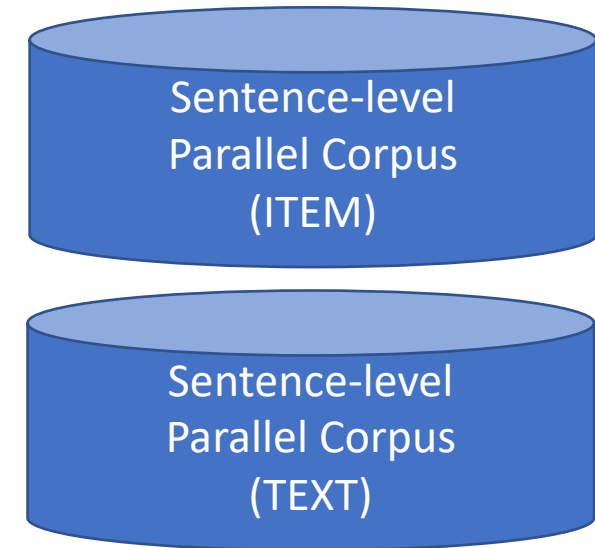
Akiko Eriguchi, Spencer Rarrick, and Hitokazu Matsushita

Microsoft

Timely Disclosure Documents Corpus (Ja->En)

	English
Ja	English
Ja	English
Ne	Japan Exchange Group, Inc.
Ne	Notice of Revision to Earnings Forecast and Dividend Forecast
W	Notice of Revision to Earnings Forecast and Dividend Forecast
di	We hereby announce that the consolidated earnings forecast and year-end dividend forecast for the fiscal year ending March 31, 2018 released on October 30, 2017 have been revised as follows.
W	Notice of Dividend from Surplus
di	As a result, the year-end dividend per share for the fiscal year ended March 31, 2018 will be ¥43 (ordinary dividend of ¥33 plus commemorative dividend of ¥10).
Th	There was cash outflow of ¥26,164 million from investment activities due mainly to ¥10,537 million in purchase of intangible assets.
m	Shareholding ratio of JPX
St	Including the shares held by SGX as treasury stock (515,063 shares).
In	
In	

} Time-series data
(2016-2018)



Sample 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.

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

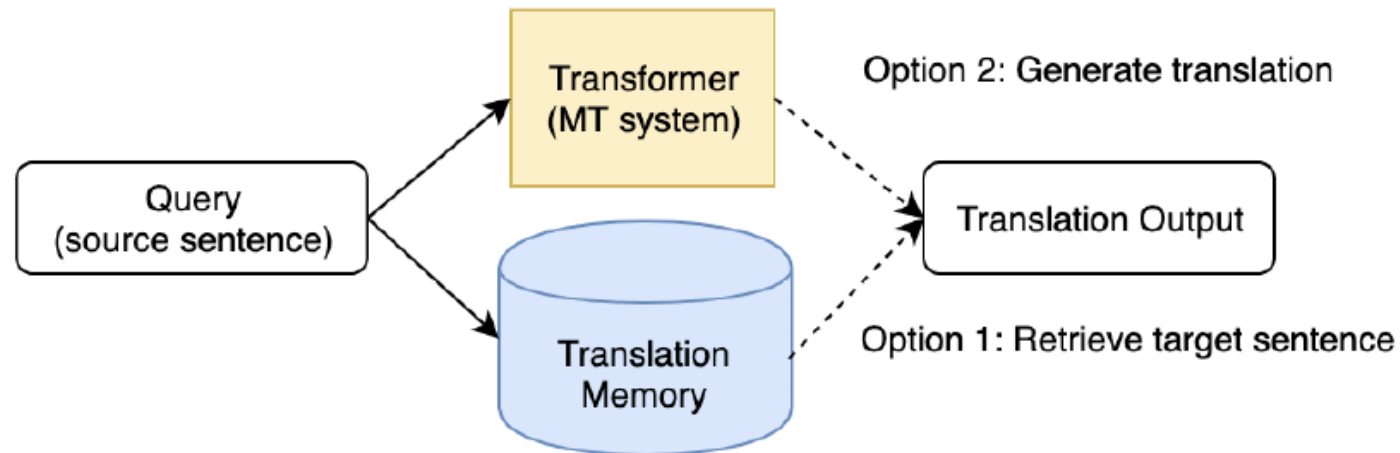
* Google Translate, Microsoft Translator, Mirai Translator as of July, 2019.

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**