Toshiba MT System Description for the WAT2014 Workshop

Satoshi Sonoh, Satoshi Kinoshita, Hiroyuki Tanaka and Satoshi Kamatani
Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation.
satoshi.sonoo@toshiba.co.jp

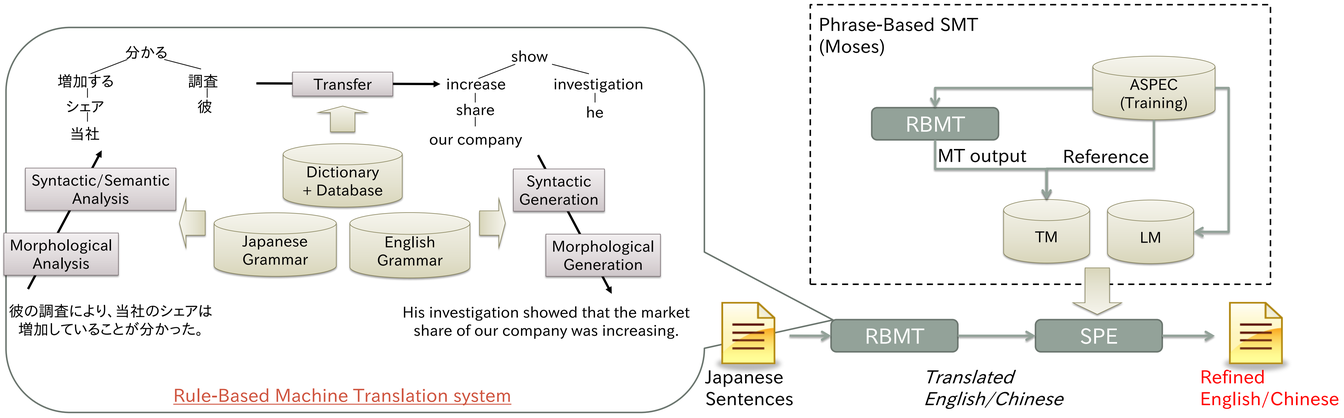
Abstract
We introduce a system description of Toshiba Machine Translation System for WAT2014. We participated in two tasks, namely Japanese-English (JE) translation and Japanese-Chinese (JC) translation. In each task, we submitted two results; one is a result of a rule-based translation system, and the other is a result which is an output of statistical post editing trained with the ASPEC training corpora. In both tasks, output by statistical post editing shows improvement in machine evaluation, but we obtained different results from human evaluation.

Toshiba Machine Translation System

We have been developing a Rule-Based Machine Translation (RBMT) system. The core functions can realize both high performance and flexibility of customization by using a large volume of dictionaries (rules) including translation knowledge.
Although Statistical-based Machine Translation (SMT) has practical translation performance in the target domain, it is extremely high cost to develop parallel corpora in wide-domain for commercial use.

Purpose: We applied statistical approach to RBMT system in order to **improve its performance** and **analyze its availability**.

System: **A combination of RBMT and Statistical Post Editing (SPE)** with selected technical term dictionaries .



Selecting Technical Term Dictionaries

Available Technical Term Dictionaries
- JE: 32 dictionaries, 3M words
- JC: 14 dictionaries, 0.46M words

Selecting best dictionary set by BLEU

Selected Technical Term Dictionaries
- JE: 4 dictionaries, **0.5M words** (Chemistry, Medicine 1&2, Science)
- JC: 2 categories, **48K words** (Chemistry, Biology)

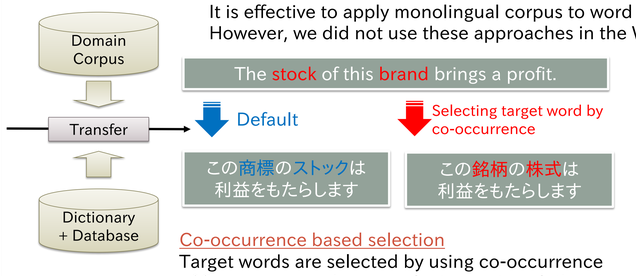
Using technical term dictionaries (BLEU for dev-set)

	JE	JC
Baseline Dict.	15.47	18.73
Add Selected Dict.	**16.21**	**18.92**

Baseline Dictionaries
- JE: **1M words** (Common, Proper noun, Internet, Natural/Social Science)
- JC: **0.4M words** (Common, Proper noun, Internet)

Alternative approach using a monolingual corpus

We have been used statistical approach in our commercial systems.
It is effective to apply monolingual corpus to word selection in the specific domain.
However, we did not use these approaches in the WAT 2014 because ASPC includes multi-domain contents.



Co-occurrence based selection
Target words are selected by using co-occurrence of fixed words, which are selected by translation rules, in the domain corpus.

N-gram based selection (for Japanese and Korean)
Target words are selected to maximize log-likelihood for a sequence of target words in the domain corpus.

Results

RBMT is better than SPE.

Japanese-English evaluation results

System	BLUE	RIBES	HUMAN	RANK(*)
RBMT	15.69	0.69	20.25	**1.38**
+SPE	**20.61**	**0.71**	**23.25**	1.45
SMT	18.45	0.65	-	2.24

Japanese-Chinese evaluation results

System	BLUE	RIBES	HUMAN	RANK(*)
RBMT	19.28	0.76	-5.25	2.13
+SPE	27.42	**0.80**	**0.75**	1.76
SMT	**27.96**	0.79	-	**1.63**

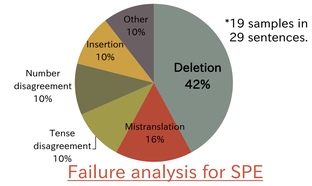
*RANK: the average of relative ranks from 1(BEST) to 3(WORST) for three systems by a bilingual evaluator.

SPE achieved **improvements of 31.4% for Japanese-English and 42.2% for Japanese-Chinese** in BLUE.
On the other hand, in RANK of Japanese-English, **RBMT showed better translation than SPE**.
We found improvements and worse related to the difference as follows:

- Improvements
  Some phrases including noun, verb and adjective are post-edited to better phrases. Because a number of vocabulary in Chinese dictionaries are significantly smaller than that of English dictionaries, the effect of SPE for Chinese may be large.

SRC	そこで, 流体の性質や条件の違いにより適切なセンサを選択することが必要である。
REF	Then, it is necessary to choose the appropriate sensor in accordance with a difference in the property and condition of the fluid.
RBMT	Then, it is required to choose a suitable sensor by the character of a fluid or the difference among conditions.
+SPE	Then, it is necessary to choose the appropriate sensor by the properties of the fluid and the difference between the conditions.

- Mistranslation
  However, deletion, that translated phrases by RBMT get worse by post-editing (e.g., “interface mold” -> “mold”), have high proportion of failures of SPE.
  Furthermore, tense disagreement (e.g., RBMT generated past form but SPE modified present form) and number disagreement occurred.



Failure analysis for SPE

SRC	りん酸基をもつ界面鋳型樹脂によるCu(II)の吸着量はpHの増大に伴い増大した。
REF	The amount of Cu(II) adsorption by the interface template resin with phosphate group increases with an increase in pH value.
RBMT	The amount of adsorption of Cu(II) by interface mold resin with a phosphoric acid group increased with increase of pH.
+SPE	The adsorption amount of Cu (II) by the mold resin with phosphoric acid group increased with increasing pH.

SRC	標題光スイッチングと光記録画像を形成し, その安定性を調べた。
REF	The optical switching of the title and its optical recording image were formed, and the stability was examined.
RBMT	Title optical switching and an optical record picture were formed, and the stability was investigated.
+SPE	Titled optical switching and the optical recording images were formed, and the stability is investigated.

Context-aware Machine Translation

Our RBMT system has following functions for context-aware translation. But we *can not confirm effectiveness* of these function in WAT2014 task.

U.S. Supreme Court will move quickly to review a new law regulating indecent materials on the Internet.…

“Law” is estimated as domain

Two panels of judges have summarized important issues on the **act**.

Default
行為 (action of a particular kind)
法律 (law made by a parliament)
“Law” domain (芝居の)幕 (division of a stage play)

Target word selection based on domain estimation

John Snow (15 March 1813 – 16 June 1858) was an English physician and a leader in the adoption of anesthesia and medical hygiene. …

If “John Snow” was found in Proper noun dictionary, a memorized word “Snow” will be preferably analyzed as a proper noun.

Snow was born 15 March 1813 in York, England.
(source: Wikipedia)
スノーは1813年3月15日にヨーク(英国)で生まれました。

Part-of-speech disambiguation and the reuse of translations by using preceding context

Conclusion
A combination of RBMT and SPE achieved improvements of BLUE score in both Japanese-English and Japanese-Chinese translation. In contrast, in a part of the human evaluation, RBMT showed better performance than SPE for Japanese-English translation.