IALP2016 in Taiwan

長岡技術科学大学
Nagaoka University of Technology

Japanese Orthographical Normalization Does Not Work for Statistical Machine Translation

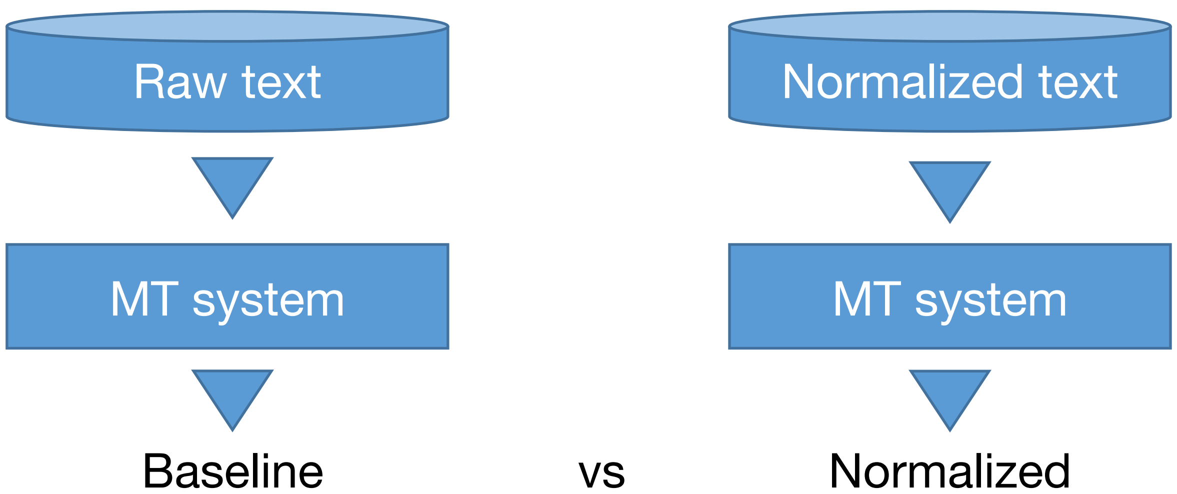
Natural Language Processing Lab
Kazuhide Yamamoto, Kanji Takahashi

Summary
Japanese orthographical normalization does not work for statistical machine translation.

Summary

10% of Japanese words have different notations.

Normalization reduces a vocabulary size.



Result shows normalization does not improve Statistical Machine Translation.

Agenda

1. Motivation
2. Japanese Orthographical Variants and Normalizing
3. The Effect on Language Model
4. The Effect on PBSMT

Agenda

1. Motivation
2. Japanese Orthographical Variants and Normalizing
3. The Effect on Language Model
4. The Effect on PBSMT

Motivation

The main problem of SMT is data sparseness(Callison-Burch et al., 2006).

Orthographic Processing for Persian-to-English improves SMT quality(Rassoli et al., 2013).

10 % of Japanese vocabulary have more than one orthographical variations(Sato, 2004;Ogura, 2009).

Our hypothesis

Normalizing orthographical variants improve a SMT quality.

Agenda

1. Motivation
2. Japanese Orthographical Variants and Normalizing
3. The Effect on Language Model
4. The Effect on PBSMT

Japanese Orthographical Variants

“center” and “centre” are the same word with a slight spelling difference.

Japanese writing system causes orthographical variants.
They have the same reading but spelling are different

Some examples

Chinese Character

- 附属、付属(attach)

Character

- りんご、リンゴ、林檎、苹果（an apple）

Abbreviation

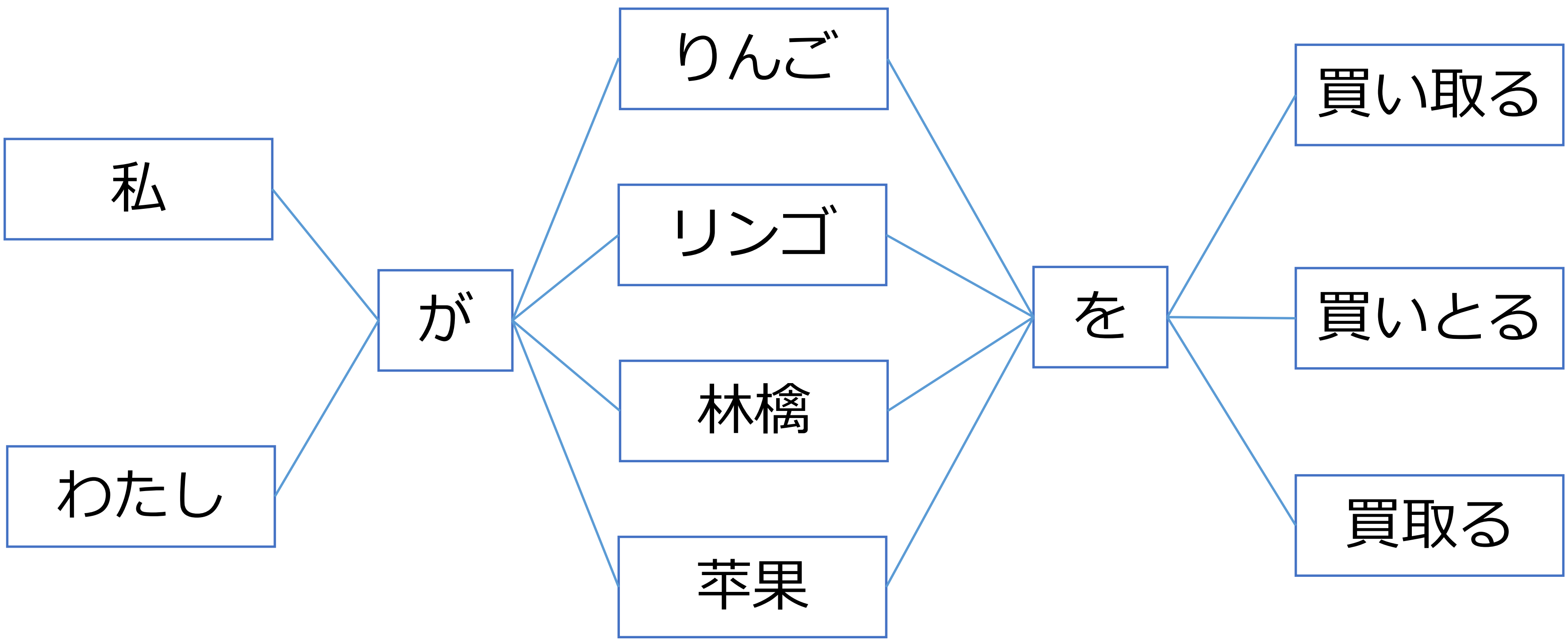
- 取説、取り扱い説明書(a manual)

Katakana(a phonographic writing system)

- コンピュータ、コンピューター(a computer)

Japanese Orthographical Variants

Ex: “I buy an apple. “ by 24 variation.



I (SUBJ) apple (OBJ) buy

How to Normalize?

SNOWMAN, our Japanese word analyzer

Word segmentation

Part-of-speech tagging

Normalizing orthographical variants(Abbreviations)

Many Features

Web-based system

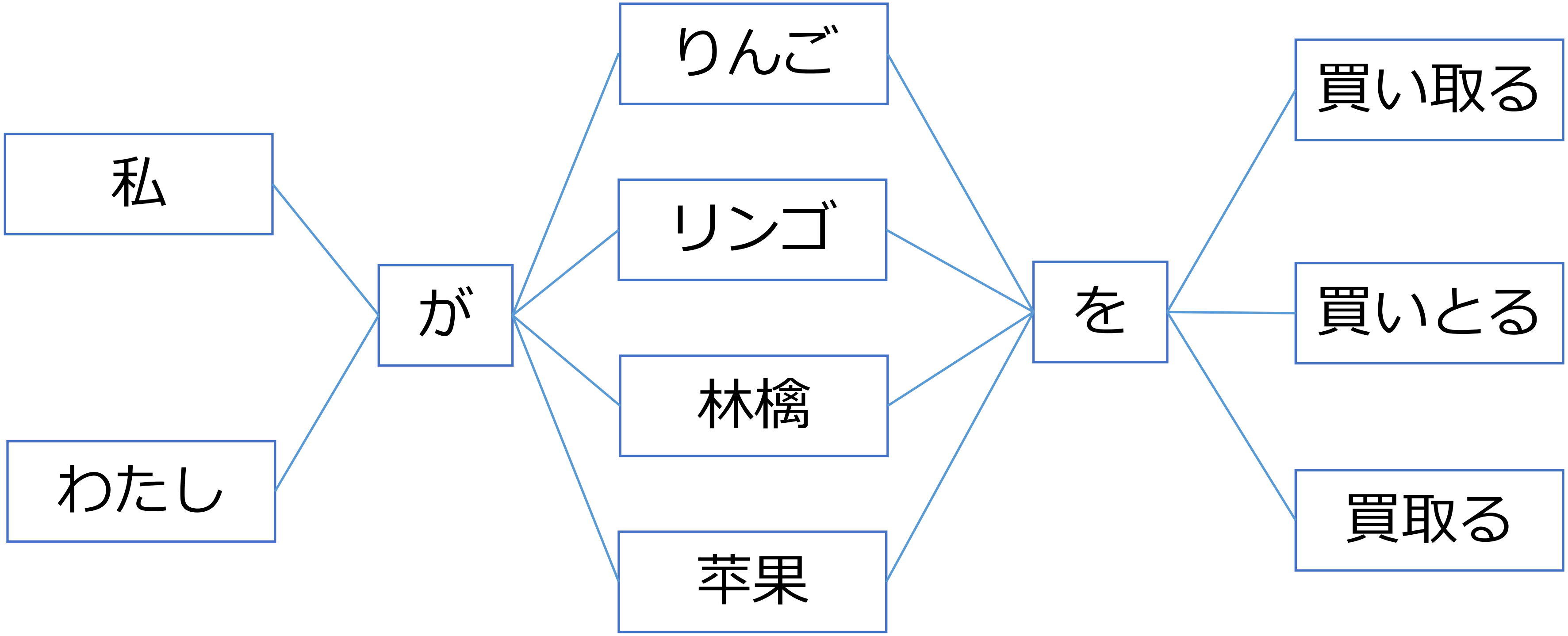
Identify idioms and functional expressions

Customized POS structure

etc.

http://snowman.jnlp.org/english

SNOWMAN Normalization



I (SUBJ) apple (OBJ) buy

24 paths into 1 path!

Agenda

1. Motivation
2. Japanese Orthographical Variants and Normalizing
3. $\boxed{\text{The Effect on Language Model}}$
4. The Effect on PBSMT

Impact of Normalization on Language Model

Language Model is a main part of SMT.

Our hypothesis in Japanese

If normalization reduce the size of LM, the SMT's quality will improve.

Compare

Baseline

Normalized corpus

Denormalized corpus

- contains a lot of orthographical variants

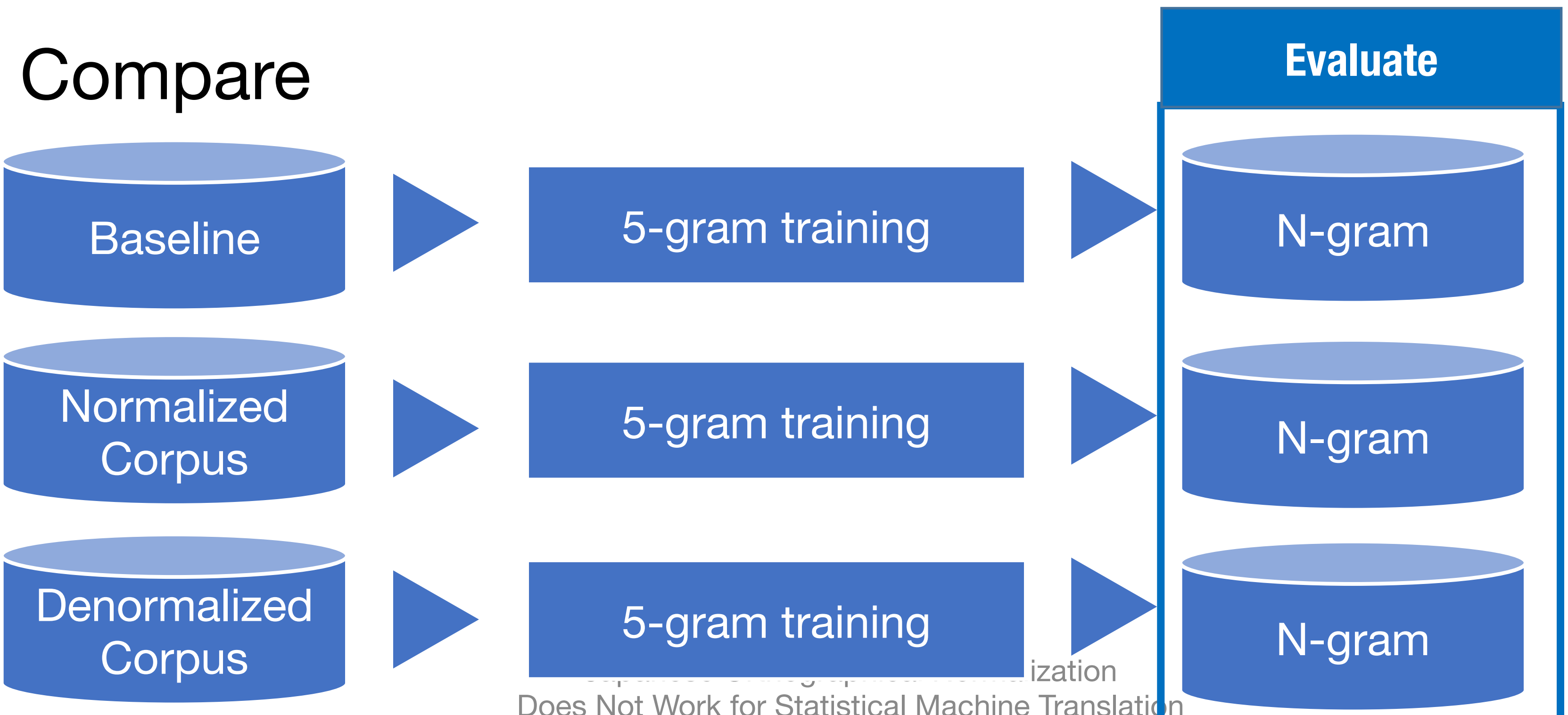
Impact of Normalization on Language Model

Language Model is a main part of SMT.

Our hypothesis in Japanese

If normalization reduce the size of LM, the SMT's quality will improve.

Compare



Building Denormalized Corpus

Artificially denormalized corpus is built for investigating the effect of a lot of orthographical variants in a corpus.

Word	meaning	Orthographical variants	Output
私	I	わたくし,ワタクシ,私	私
が	(SUBJ)	が, ヶ	が
りんご	apple	りんご,リンゴ,林檎,苹果	リンゴ
を	(OBJ)	を,ヲ	を
買い取る	to buy	買い取る,買いとる, 買取る	買いとる

Building Denormalized Corpus

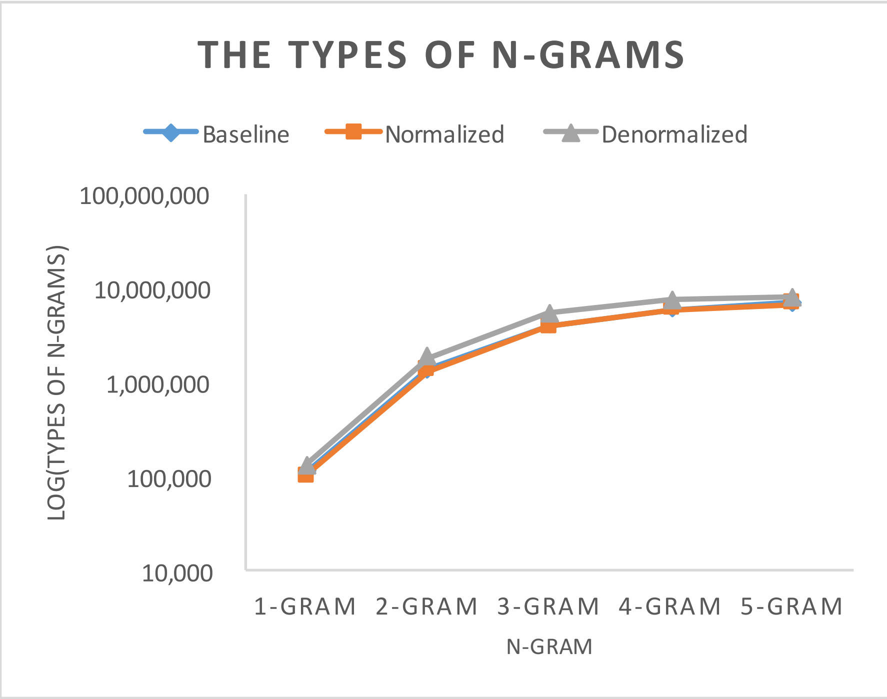
Artificially denormalized corpus is built for investigating the effect of a lot of orthographical variants in a corpus.

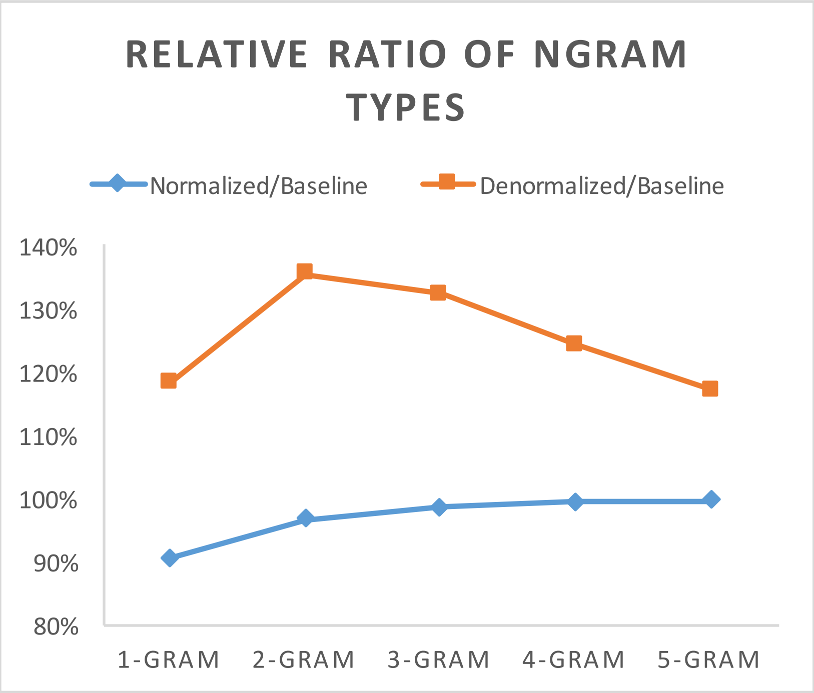
Randomly selected

Word	meaning	Orthographical variant	Output
私	I	わたくし,ワタクシ,私	私
が	(SUBJ)	が,ヶ	が
りんご	apple	りんご,リンゴ,林檎,苹果	リンゴ
を	(OBJ)	を,ヲ	を
買い取る	to buy	買い取る,買いとる,買取る	買いとる

N-gram Types

The types of n-grams with normalization slightly decreases.





Reduction ratio of phrase table:2%

Orig:23,446,800 -> Normalized:23,033,827

Agenda

1. Motivation
2. Japanese Orthographical Variants and Normalizing
3. The Effect on Language Model
4. The Effect on PBSMT

SMT Experiments Setup

SMT system (standard baseline)

- Moses
- GIZA++
- KenLM toolkit 5-gram
- MERT tuning

Japanese-English Corpus

- KFTT : Wikipedia's Kyoto articles
- NTCIR-7 : Patents

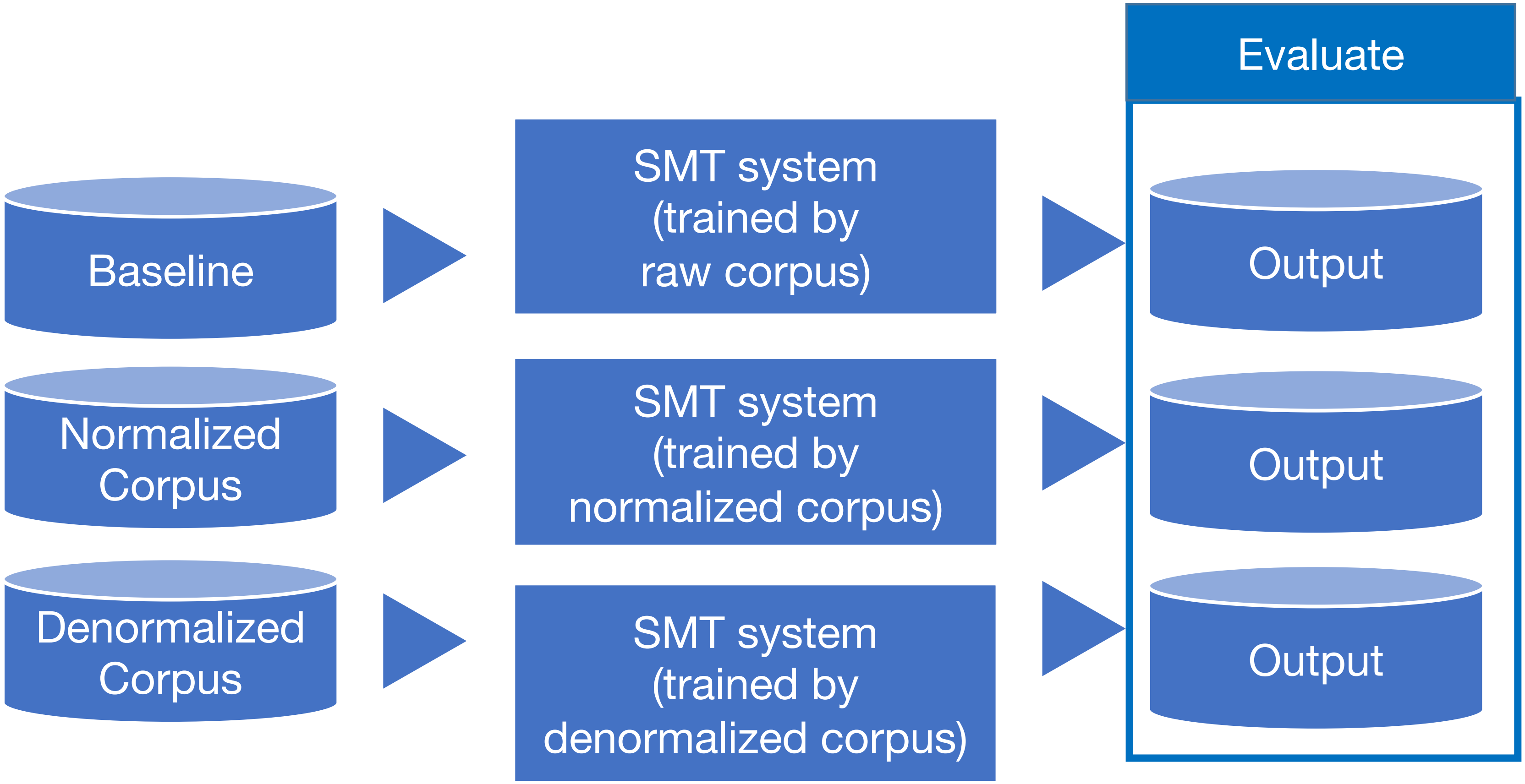
Corpus preprocessing

- English : TreeTagger tokenization and lowercasing
- Japanese : Word segmentation and some preprocessing
- Delete ignore ratio sentences for GIZA++

* Experimental scripts are available on https://github.com/kanjirz50/mt-ialp2016

SMT Experiments Setup

Experimental Flow



Test-set Statistics

Corpus	Token	Vocabulary	OOV	Perplexity
KFTT-Baseline	27,761	4,637	152	74.0
KFTT-Normalized		4,558	134	71.2
KFTT-Denormalized		5,274	133	152.3
NTCIR7-Baseline	33,565	3,505	65	34.5
NTCIR7-Normlized		3,424	64	33.9
NTCIR7-Denormalized		4,490	482	82.6

Test-set Statistics

Corpus	Token	Vocabulary	OOV	Perplexity
KFTT-Baseline	27,761	4,637	152	74.0
KFTT-Normalized		4,558	134	71.2
KFTT-Denormalized		5,274	133	152.3
NTCIR7-Baseline	33,565	3,505	65	34.5
NTCIR7-Normlized		3,424	64	33.9
NTCIR7-Denormalized		4,490	482	82.6

Result

There is no improvement on both evaluation metrics.
EN to JP, it's difficult to compare exactly because the surface forms are changed by normalizing

	Japanese to English		English to Japanese	
Condition	***BLEU***	***RIBES***	***BLEU***	***RIBES***
KFTT-Baseline	19.3	66.4	21.3	68.5
KFTT-Normalized	19.7	66.2	22.0	69.2
KFTT-Denormalized	17.3	63.6	9.7	61.0
NTCIR7-Baseline	26.2	65.8	29.1	67.6
NTCIR7-Normalized	26.0	65.6	29.7	67.4
NICIR7-Denormalized	23.3	64.0	10.0	58.5

* No statistical significance was found

Result

There is no improvement on both evaluation metrics.
EN to JA, it's difficult to compare exactly because the surface forms are changed by normalizing

	Japanese to English		English to Japanese	
Condition	***BLEU***	***RIBES***	***BLEU***	***RIBES***
KFTT-Baseline	19.3	66.4	21.3	68.5
KFTT-Normalized	19.7	66.2	22.0	69.2
KFTT-Denormalized	17.3	63.6	9.7	61.0
NTCIR7-Baseline	26.2	65.8	29.1	67.6
NTCIR7-Normalized	26.0	65.6	29.7	67.4
NICIR7-Denormalized	23.3	64.0	10.0	58.5

* No statistical significance was found

Result

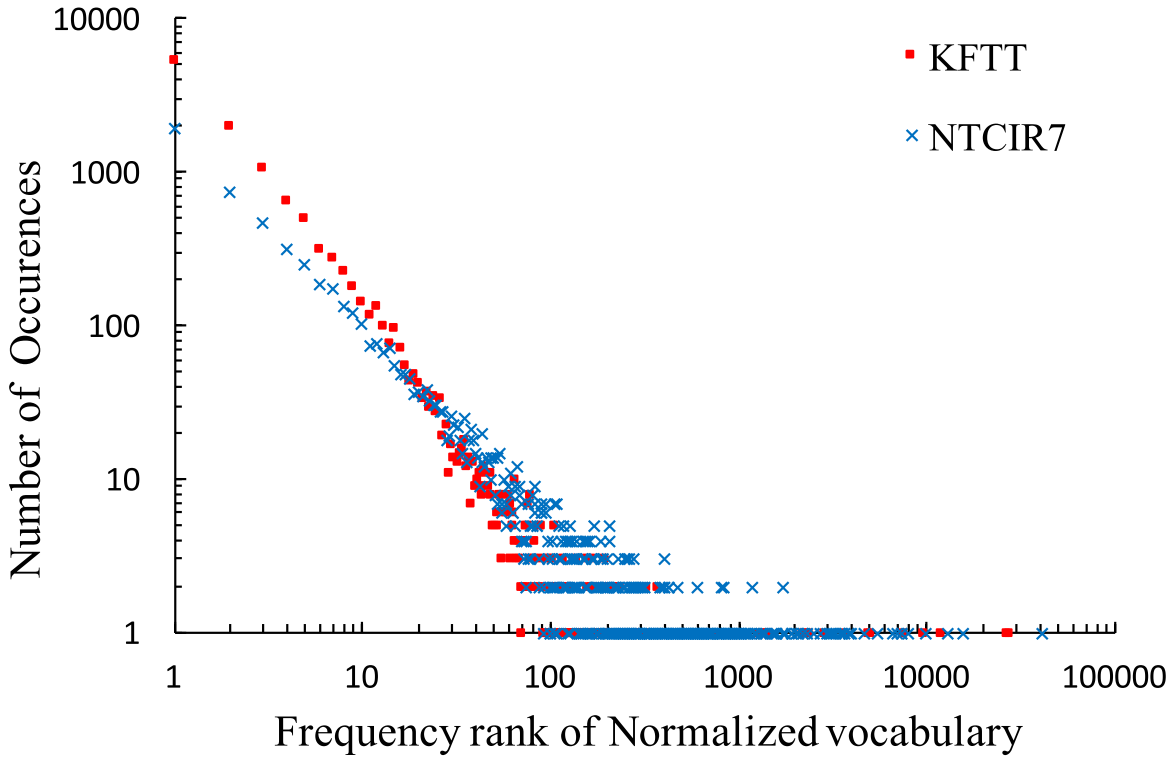
There is no improvement on both evaluation metrics.
EN to JA, it's difficult to compare exactly because the surface forms are changed by normalizing

	Japanese to English		English to Japanese	
Condition	*BLEU*	*RIBES*	*BLEU*	*RIBES*
KFTT-Baseline	19.3	66.4	21.3	68.5
KFTT-Normalized	19.7	66.2	22.0	69.2
KFTT-Denormalized	17.3	63.6	9.7	61.0
NTCIR7-Baseline	26.2	65.8	29.1	67.6
NTCIR7-Normalized	26.0	65.6	29.7	67.4
NICIR7-Denormalized	23.3	64.0	10.0	58.5

* No statistical significance was found

Analysis

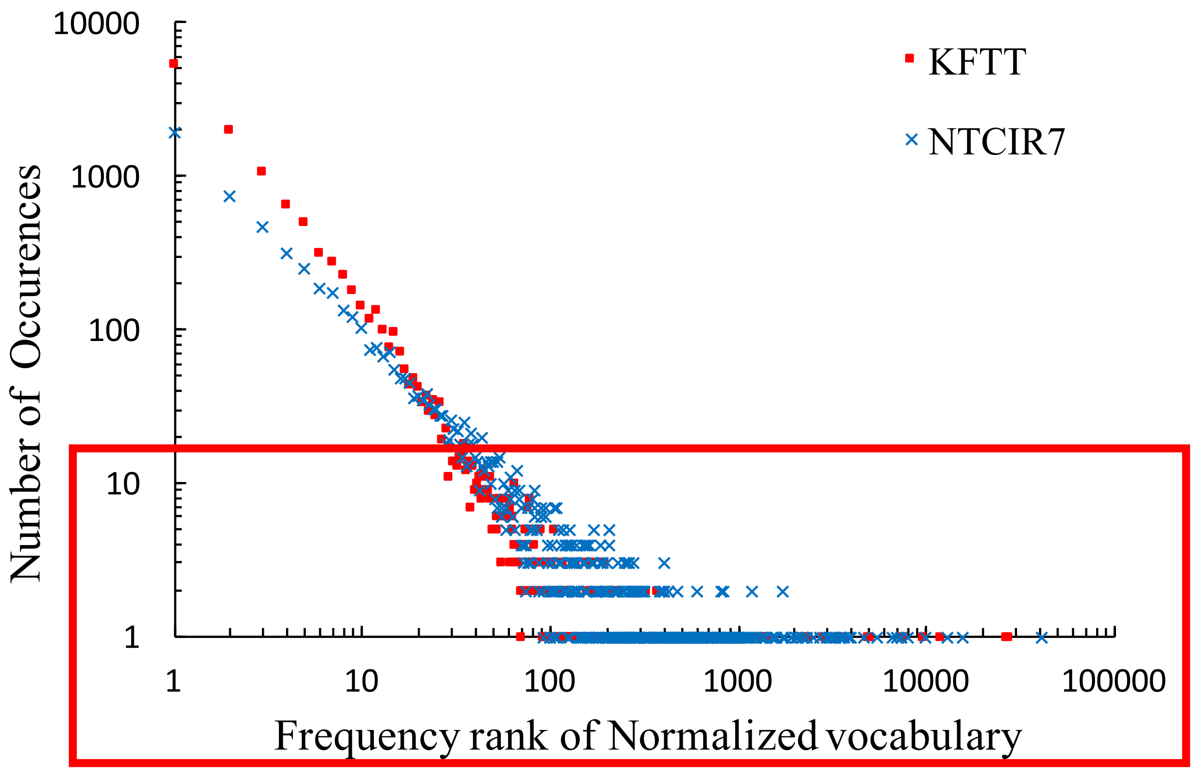
Real corpus contains low frequency orthographical variants.



Ex. lemma:freq
引っ越し:120 <- (引越:40, 引越し:3)

Analysis

Real corpus contains low frequency orthographical variants.



Ex. lemma:freq
引っ越し:120 <- (引越:40, 引越し:3)

Conclusion

Orthographical normalization of Japanese language does not improve SMT.

Real corpus contains low frequency orthographical variants.

Normalization slightly decreases

- Vocabulary size
- Perplexity
- Out-of-vocabulary

Summary

Japanese orthographical normalization does not work for statistical machine translation.

RIBES: Rank-based Intuitive Bilingual Evaluation Score

An automatic evaluation metric for MT, developed in NTT Communication Science Labs.

Automatic Evaluation of Translation Quality for Distant Language Pairs

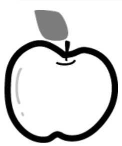
			BLEU	RIBES
Original	彼は雨に濡れたので、風邪を引いた。			
Reference	He caught a cold because he got soaked in the rain.			
RBMT	He caught a cold because he had gotten wet in the rain.	○	0.53	0.93
SMT	He got soaked in the rain because he caught a cold.	×	0.74	0.38

http://aamtjapio.com/kenkyu/files/discussion01/AAMT_Japio_discus(20120907)-02.pdf

Japanese Orthographical Normalization Does Not Work for Statistical Machine Translation

Investigating the effect of normalizing Japanese orthographical variants on SMT.

Japanese Orthographical Variants



An apple : “りんご”, “リンゴ”, ”林檎”, “苹果” → “りんご”

SMT with normalization is equivalent to that without normalization by both BLEU and RIBES.

* Experimental scripts are available on https://github.com/kanjirz50/mt-ialp2016

Refference

K. Yamamoto, Y. Miyanishi, K. Takahashi, Y. Inomata, Y. Mikami, and Y. Sudo, “What We Need is Word, Not Morpheme; Constructing Word Analyzer for Japanese,” *Proceedings of the International Conference on Asian Language Processing*, pp. 49– 52, 2015.

C. Callison-Burch, M. Osborne, P. Koehn, and M. Osborne, “Improved Statistical Machine Translation Using Paraphrases,” *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 17–24, 2006.

M. S. Rasooli, A. El Kholy, and N. Habash, “Orthographic and Morphological Processing for Persian-to-English Statistical Machine Translation,” *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 1047–1051, 2013.

S. Sato, “Identifying spelling variations of Japanese words,” *The Special Interest Group Technical Reports of IPSJ* , vol. 2004, no. 47, pp. 97–104, 2004, (in Japanese).

H. Ogura, “Corpus-Based Survey of the Orthographic Variation in Contemporary Japanese: Analysis of the BCCWJ-Core,” *JCLWorkshop 2012*, pp. 321–328, 2009, (in Japanese).