TerMT: A Dataset for Evaluating Terminology Consistency in Translation

(翻訳における訳語一貫性評価用データセット)

東北大学大学院　情報科学研究科
システム情報科学専攻
乾・鈴木研究室　修士課程2年

B8IM2001　阿部 香央莉

Contents

Background: Current NMT

- NMT: Neural Machine Translation
- Sentence-level NMT: high translation quality (e.g., Google Translation)
- Next Step: domain-specific translation
  - to consider **specific context** and **circumstances**

Promoting cross-language communication

Daily Conversation

Clever Dictionary
・Phrase bank

Domain-specific Translation

- Scientific Paper
- Corporation Document
- Specific Article
- Historical Text
- etc…

Background: Problem of Current NMT

- e.g., a corporation document

- Title: "**Our Company**'s Product: "
- In **our company**, we announced xxx …
- Strategy of **our company**: ……

Encoding → sentence-level NMT model → Decoding → **Outputs**

Background: Problem of Current NMT

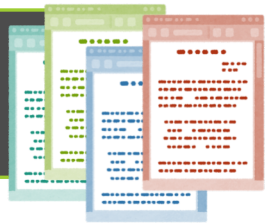
- e.g., a corporation document
  - Maybe, cause some terminology-inconsistent errors …

- Title: "**Our Company**'s Product: "
- In **our company**, we announced xxx …
- Strategy of **our company**: ……

- タイトル：「**弊社**の商品 … 」
- **我が社** では、xxx を発表した …
- **当社** の戦略：～

Encoding → sentence-level NMT model → Decoding → **Outputs**

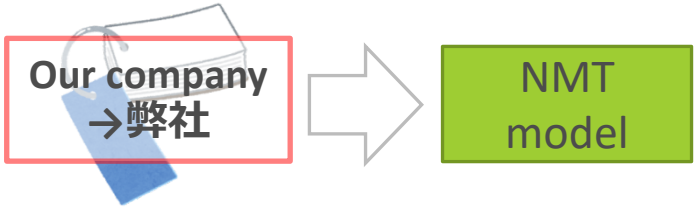
Current NMT models usually learned **a variety of articles** for fluency



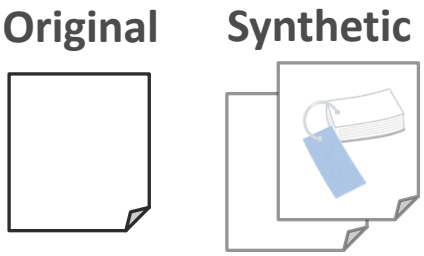
Background: NMT with Dictionaries

- One approach to terminology-inconsistent errors

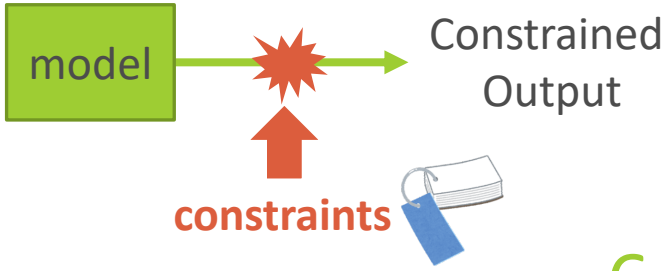
  → Using **phrase dictionaries**



- Previous studies of NMT with dictionaries
  - Synthetic Training Data [Song+, 2019]
    - Training with synthetic Code-Switching data generated by **the dictionary**



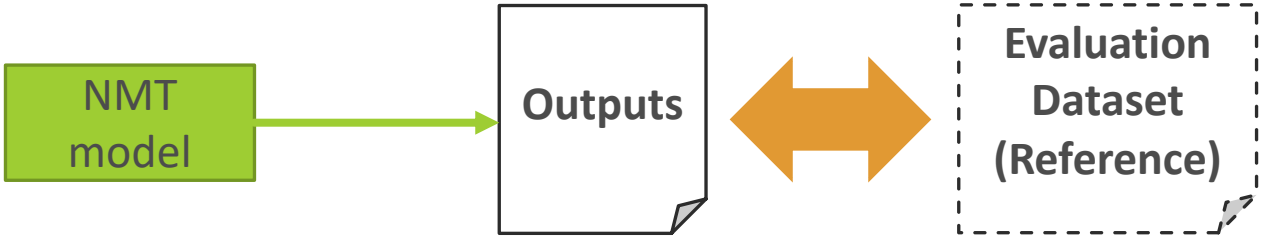
  - Constraint Decoding [Post and Vilar, 2018]
    - No change in training
    - constrained the output with **the dictionary** in decoding



Background: Lack of Evaluation Tools

- However, we have some problems …

1. **Evaluation dataset** considering terminology consistency is ~~nothing~~ -> **a little**

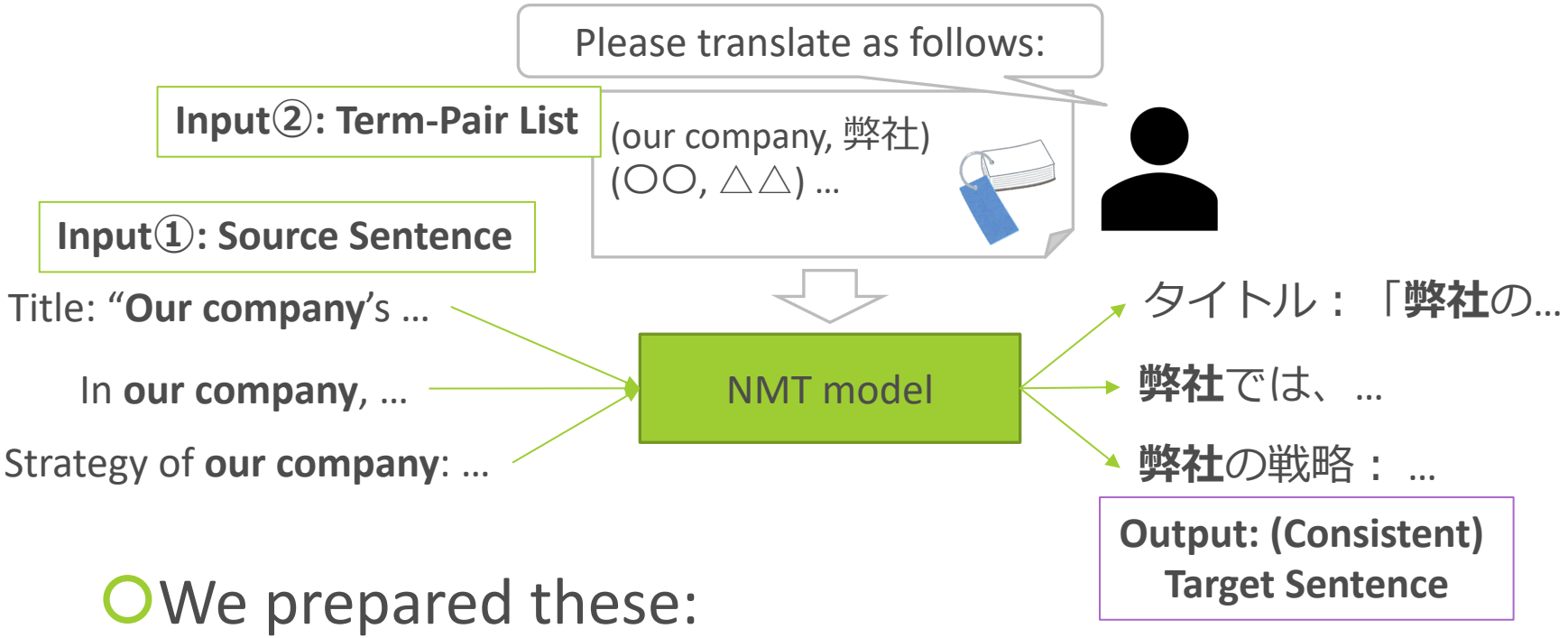


   - Very current study: HABLeX [Thompson+, 2019]
     - 2019/11/03 published

2. **Lack of automatic evaluation metric** for terminology consistency
   - [Thompson+, 2019] do not consider this strictly

Task: Terminology Consistency Evaluation

- Task definition:



- We prepared these:
  1. Evaluation dataset
  2. Automatic evaluation metric

Critical Problem on Terminology Consistency Evaluation

- The term in Term-Pair List  sometimes be inappropriate **because of contextual rightness**
  - e.g. the term in KFTT corpus (“公家”)
    - 公家 <-> nobles <-> aristocrats
    - 公家 (社会) <-> aristocratic (society) <-> aristocrats, nobles
    - 公家 (様) <-> aristocratic (style) <-> aristocrats, nobles

- How we evaluate this?
  - We need **sentence-wise correct term information**
    -> make datasets!

Our Dataset Usage

- Purpose: Terminology-consistent test set **corresponding to the Term-Pair list (TP-List)**
  - Necessary condition
    - Terminology consistency
    - Then, appropriate wording in the context

源氏 は …
武家と**公家**の間では …
公家 …

NMT model

Genji is …
… between ***warriors*** and ***nobles*** …
Aristcrats …

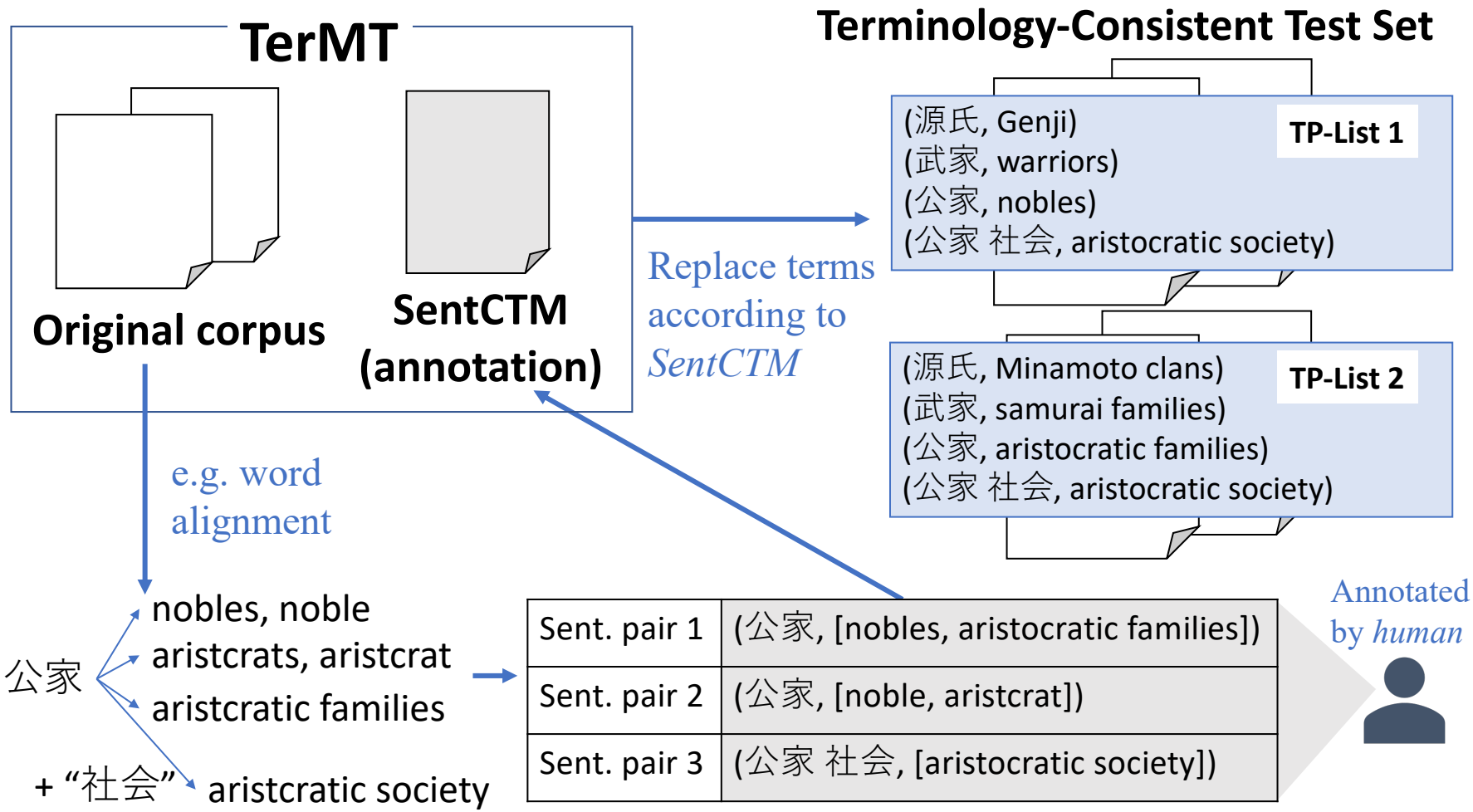
Evaluation with some metrics

Terminology-Consistent Test Set

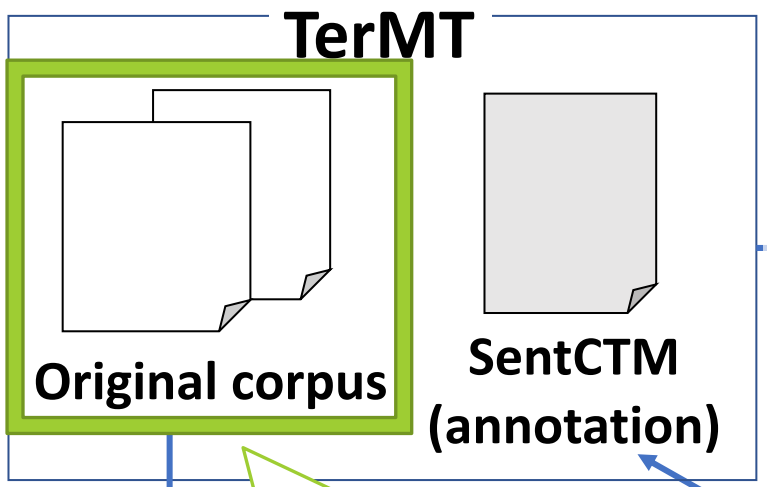
TP-List 1

(源氏, Genji)
(武家, warriors)
(公家, nobles)
(公家 社会, aristocratic society)

Procedure of Dataset Construction

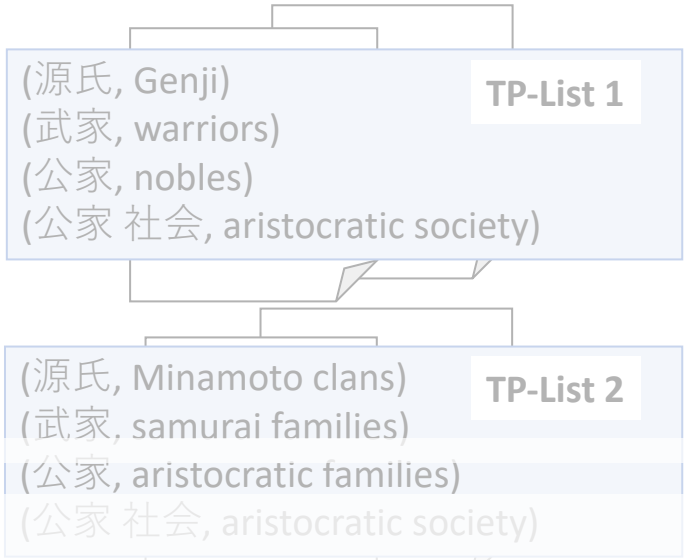


(I) Select Original Dataset



Replace terms according to *SentCTM*

Terminology-Consistent Test Set



First,

- Selected the **original corpus** to add annotations

Two Original Dataset

- Chose two original datasets for Ja-En/En-Ja translation directions
  - Criteria: has a variety of translation

1. **KFTT [Neubig+, 2011] (Ja-En)**
   - Wikipedia articles about “Kyoto”

e.g., “武家”

“Samurai” (Phonetic)
“Samurai familes”

“Warrior”(Semantic)
“Warrior familes”

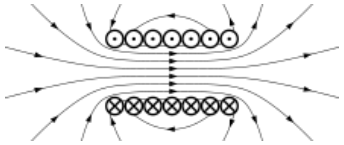
…

2. **ASPEC [Nakazawa+, 2016] (En-Ja)**
   - Scientific papers (multi-domains)

e.g., “Field”

“分野” (most scientific paper domain)

“電界” (electronic domain)



“圃場” (agricultural domain)

(II) Obtain One-to-Many Term Mappings

TerMT

Terminology-Consistent Test Set

Second,

- Extracted our focused term pairs with word alignment
- Merged term pairs to ***one-to-many term mappings***

(公家, aristocratic families)
(公家 社会, aristocratic society)

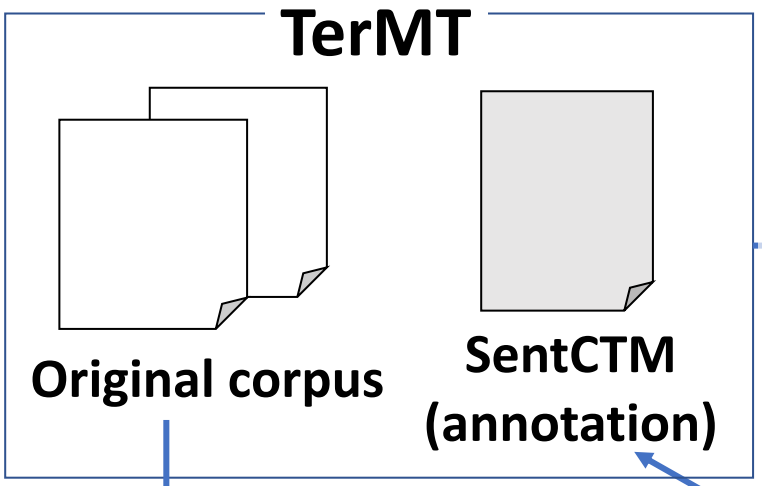
e.g. word alignment

公家 → nobles, noble
→ aristcrats, aristcrat
→ aristcratic families
+ “社会” → aristcratic society

Sent. pair 1	(公家, [nobles, aristocratic families])
Sent. pair 2	(公家, [noble, aristcrat])
Sent. pair 3	(公家 社会, [aristocratic society])

Annotated by *human*

(III) Annotation of SentCTM



Terminology-Consistent Test Set

Third,

- Attached mappings that consists only of ***interchangeable*** terms ***regard to context***

e.g. word alignment

公家 → nobles, noble / aristcrats, aristcrat / aristcratic families

+ "社会" → aristcratic society

Sent. pair 1	(公家, [nobles, aristocratic families])
Sent. pair 2	(公家, [noble, aristcrat])
Sent. pair 3	(公家 社会, [aristocratic society])

Annotated by *human*

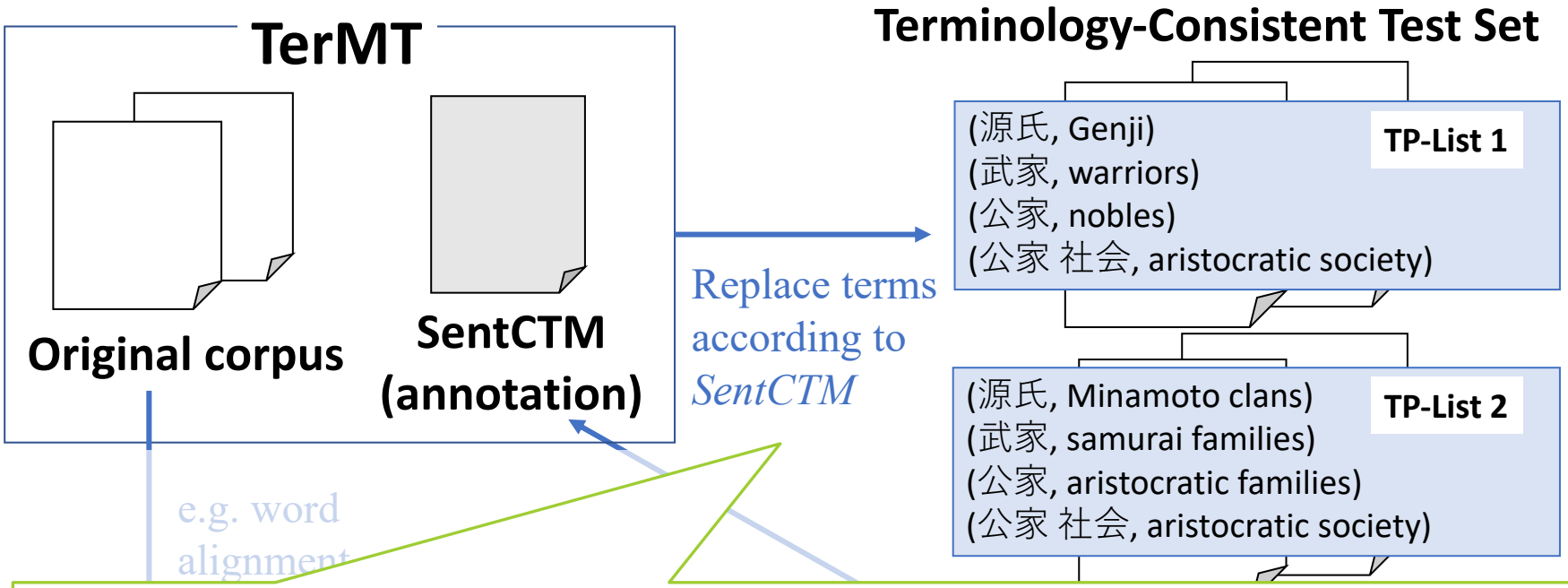
The Detail of Our Dataset

Index	Src-sent	Tgt-sent	Src-Term	Tgt-Term
705	この時代の文化を、武家様・**公家**様・唐様（禅宗様）が融合した北山文化と呼ぶことも多い。	The culture of this period is called the Kitayama Culture , where the samurai style , **aristocratic** style and the Tang style (Zen Buddhism style) were merged .	公家	**aristocratic**
1209	形式的に言えば、朝廷が正規の政府で幕府は地方における臨時の政府であると**公家**の間では認識していた。	Formally speaking , among the **nobles** it was recognized that the Court was the true government and the Shogunate was a temporary government in the provinces .	公家	**nobles**,aristcrats
1217	これらの事から、征夷大将軍になるのは源氏でも平氏でも、さらには**公家**の藤原氏でもなんら支障は無いと解釈できる。	All this can be interpreted to mean that someone from the Minamoto or Taira clans , or even from the **noble** Fujiwara clan , could become Seii Taishogun .	公家	**noble**,aristcrat

Original corpus

SentCTM (annotation)

Final Test Set



Finally,

- We can obtain terminology-consistent test set from our dataset TerMT

+ “社会” aristcratic society

Dataset Statistics

	KFTT（Ja-En）	ASPEC（En-Ja）
# of our focused source terms	84	196
A: # of sent. pairs w/ SentCTM	457	407
B: # of original sent. pairs	1245	1812
Ratio (A/B)	(36.7%)	(22.5%)
avg. # of target candidates	2.27	2.77
Max # of target candidates	12	10

Problem of Evaluation Metric

- **BLEU**: strong benchmark of evaluation metric in NMT
  - compare model outputs <-> references with n-gram
  - is unable to properly capture the terminology consistency

Model Output 1

It was customarily **agreed** by the **warrior** families … .

It was customarily **accepted** by the **warrior** families … .

Model Output 2

It was customarily **accepted** by the **samurai** families… .

It was customarily **accepted** by the **warrior** families … .

We cannot distinct **terminology error** or not

- So, We consider F-score using TER[Snover+, 2006] alignment
  - use **word alignment** model outputs <-> references

Proposed Metric: F-score of Terms

- Utilized **word alignment** (model output <-> reference)
  - focus on the **translation of terminology itself**
  - impose penalties for **over/under generation** of the terms

Reference	it fell precisely on the 100th day after the death of takauji **ashikaga**
Model 1	it was the 100th day after takauji **ashikaga** 's death .

Reference	**it** fell precisely on the 100th day after the death of takauji **ashikaga**
Model 2	**ashikaga** was the 100th day after takauji **ashikaga** 's death

- [Thompson+, 2019] used simple Recall of the terminology
  - cannot consider over/under generation

How to Calculate "F-score of Terms"

References (Recall)	Outputs (Precision)	Alignment (for focused term)
Ashikaga	it	Not agree
…	…	
Takauji	Takauji	
Ashikaga	**Ashikaga**	**agreed**

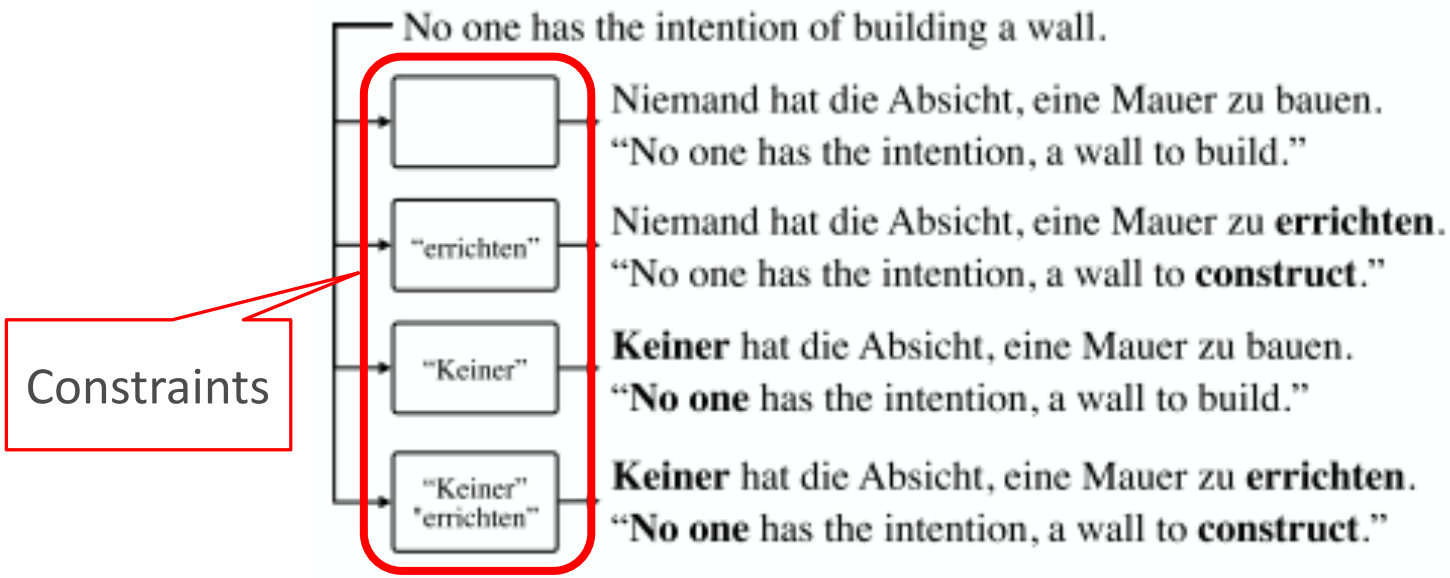
- Precision
  - (# of **agreed alignments**) / (# of the term appearance in the **outputs**)
- Recall
  - (# of **agreed alignments**) / (# of the term appearance in the **references**)
- F-score
  - 2 * Precision* Recall / (Precision + Recall)

Experimental Setting

- Term Pair List
  - randomly select one from multiple target candidates
- Dataset (Training, Development)
  - KFTT (Ja-En), ASPEC (En-Ja)
- Metrics
  - BLEU, F-score of Terms
- Model (https://github.com/awslabs/sockeye)
  - **BASE**: baseline (Transformer [Vaswani+, 2017] model)
    - Transformer model: generally SOTA model in NMT
  - **CONST**: BASE + dynamic beam allocation [Post and Vilar, 2018]
    - Decoding methods for integrating the dictionary
  - **POST**: BASE + simple post-editing

Model: CONST (Beam-Allocation) [Post and Vilar, 2018]

- A promising method for tackling terminology consistency
  - **always outputs given constraint words or phrases**, utilizing beam search



※図は[Post and Vilar, 2018]より引用

How to Use Term-Pair List in CONST?

- CONST needs a target phrase list to constrain
  - Note that, this can only use the term-pair list
    = without other information like SentCTM
- We selected this setting:
  - If the source term in Term-Pair List (simply) **appear** in the Src sent., use the target one as the constraint

1. … **examination** …
2. ……
3. …**flow range**…
   examination …..

Src sent. (Input①)

(**examination**, 調査)
(**flow range**, 流速 範囲)
(flow, 流量)
…

Term-Pair List (Input②)

→

1. [“調査”]
2. (None)
3. [“流速 範囲”, “調査”]
…..

Constraint List for CONST

Results: Quantitative Evaluation

	BLEU			F-score		
	BASE	CONST	POST	BASE	CONST	POST
Ja-En	**13.47**	13.45	13.24	54.45	**78.89**	66.72
En-Ja	**41.63**	41.03	33.91	71.31	**77.64**	45.25

- Overall,
  - BLEU : CONST < BASE
  - F-score : BASE < CONST
  - POST (simple method) is worser in both metrics

Constraint Decoding hurt translation performance of baseline model

Results: Qualitative Evaluation

Example 1 (ASPEC, En-Ja)	
Input	Macroscopic heat ***transfer*** characteristics of fluid in the vicinity of a critical point was clarified by *experiments* and numerical analysis .
correct terms	[**(transfer, “伝熱”)**, (experiments, “実験”)]
Reference	臨界点近傍の流体の巨視的な**伝熱**特性を，実験及び数値解析によって明らかにした。
BASE	臨界点近傍の流体の巨視的な**熱伝達**特性を実験と数値解析により明らかにした。
(constraints)	[“**移植**”, “実験”]
CONST	臨界点近傍の流体の巨視的**熱伝達**特性を実験と数値解析により明らかにした。**移植実験**の結果を報告した。
POST （伝達→移植）	臨界点近傍における流体の巨視的**熱移植**特性を実験と数値分析により明らかにした。

Results: Qualitative Evaluation

Example 1 (ASPEC, En-Ja)	
Input	Macroscopic heat ***transfer*** characteristics of fluid in the vicinity of a critical p nalysis .
correct terms	[(transfe
Reference	臨界 点 及び 数値 解析 によ
BASE	臨界 点 近傍 の 流体 の 巨視 的 な **熱** 性 を 実験 と 数値 解析 により 明らかに した。
(constraints)	[“**移植**”, “実験”]
CONST	臨界 点 近傍 の 流体 の 巨視 的 **熱 伝達** 特性 を 実験 と 数値 解析 により 明らかに した。 **移植 実験** の 結果 を 報告 した。
POST （伝達→移植）	臨界 点 近傍 に お け る 流体 の 巨視 的 **熱 移植** 特性 を 実験と 数値 分析 に よ り 明らかに した。

This verbose phrases might cause by the compulsory constraints

Conclusions

- Two contributes:
  - constructed **evaluation dataset** for terminology consistency
  - proposed more rigorous **evaluation metric** for terminology consistency than accuracy
- Future Work:
  - propose the methodology that can balance terminology consistency and satisfying the context

References

[Song+, 2019] Song, K., Zhang, Y., Yu, H., Luo, W., Wang, K., and Zhang, M. (2019). Code-Switching for Enhancing NMT with Pre-Specified Translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 449–459.

[Post and Vilar, 2018] Post, M. and Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1314–1324. Association for Computational Linguistics, June.

[Thompson+, 2019] Thompson, B., Knowles, R., Zhang, X., Khayrallah, H., Duh, K., and Koehn, P. (2019). HABLex: Human annotated bilingual lexicons for experiments in machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1382–1387, Hong Kong, China, November. Association for Computational Linguistics.

[Neubig+, 2011] Neubig, G. (2011). The Kyoto free translation task. http://www.phontron.com/kftt.

[Nakazawa+, 2016] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). Aspec: Asian scientific paper excerpt corpus. In Proceedings of the Ninth International Conference on Language Resources and Evaluation, pages 2204–2208. European Language Resources Association, may.

[Snover+, 2006] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In Proceedings of Association for Machine Translation in the Americas, pages 223–231.

[Vaswani+, 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. CoRR, abs/1706.03762.

Appendix

Comparison between HABLex and TerMT

- Same
  - Dataset was made by supplying new annotation for original dataset
  - Motivation of dataset creation
    - want to integrate bilingual lexicon into NMT
      → adapt some unknown words or domain-specific terms
- Difference in HABLex [Thompson+, 2019]
  - Multilingual Dataset （{Ru, Ch, Ko} -> En）
  - Annotation schema: 1 sentence <-> 1 term-pair
  - use Recall for evaluation of terminology consistency

Two Original Dataset: KFTT (Ja-En)

- Chose two original datasets for Ja-En/En-Ja translation directions
  - **Original dataset 1. KFTT [Neubig+, 2011] → Ja-En**
    - has many kind of **Japanese-specific culture names or concepts** → a variety of translation in other languages

E.g., “侍”

“Samurai” (Phonetic)

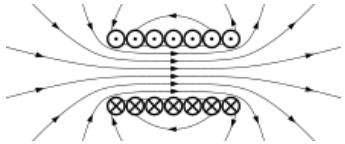
“Warrior”(Semantic)
“Warrior familes”
…

Two Original Dataset: ASPEC (En-Ja)

- Chose two original datasets for Ja-En/En-Ja translation directions
  - **Original dataset 2. ASPEC [Nakazawa+, 2016] → En-Ja**
    - has many kind of **scientific terms** which has different meaning from daily conversation
      → a variety of translation depends on their domain, translator, time, …

e.g., “Field”

- “分野” (most scientific paper domain)
- “電界” (electronic domain)



- “圃場” (agricultural domain)

Model: POST

- Very simple batch replacement
  - Concept: replacing other candidates to the correct term
    - ”other chandidates”?
    - Originally, **references do not tell us ”What terms should be replaced to the correct term”**
    - We obtained one-to-many term mapping in procedure of the dataset construction

    → So, use this as a list for replacing!

Correct term

  - e.g. source: “弟子” -> target: [apprentice, disciple, apprentices, disciples]

→ `$ sed –E “s/(apprentices/disciple/disciples)/apprentice/g” [model_output]`

Results : Quality Evaluation (KFTT)

Example 2	
Input	**本尊** は 阿弥陀 如来 一 仏 で あ る。
SentCTM	[(本尊, Honzon), (如来, Nyorai)]
Reference	The **Honzon** is only Amida Nyorai .
BASE	The **principal image** is Amida Nyorai (Amitabha Tathagata) .
CONST	**Honzon** (**principal image** of Buddha) is Amida Nyorai (Amitabha Tathagata) .
POST	The **Honzon** is Amida Nyorai (Amitabha Tathagata) .