

Kanbun-LM: Lire et traduire du chinois classique en japonais grâce à des modèles de langues

Hao WANG, Hirofumi SHIMIZY, Daisuke KAWAHARA

Amaury GAU

Florian JACQUOT

Agathe WALLET

Introduction

Cet article nous promet :

- Création du premier corpus CC pour le Kanbun
- Présentation d'une méthodologie de prétraitements pour transformer le chinois classique en japonais :
 - réarrangement de l'ordre des caractères
 - traduction automatique
- Résultats à l'état de l'art pour les deux tâches ci-dessus
- Comparaison avec des traductions humaines

Plan

Introduction

I. Présentation des données

- A. Qu'est-ce que le Kanbun ?
- B. Les données
- C. Prétraitements

II. Présentation des expériences

- A. Modèles
- B. Métriques
- C. Annotations manuelles

III. Les résultats

- A. Réarrangement des caractères
- B. Traduction automatique
- C. La pipeline

Conclusion

I. Présentation des données

A. Qu'est-ce que le Kanbun ?

Chinois
classique

- SVO
- langue isolante



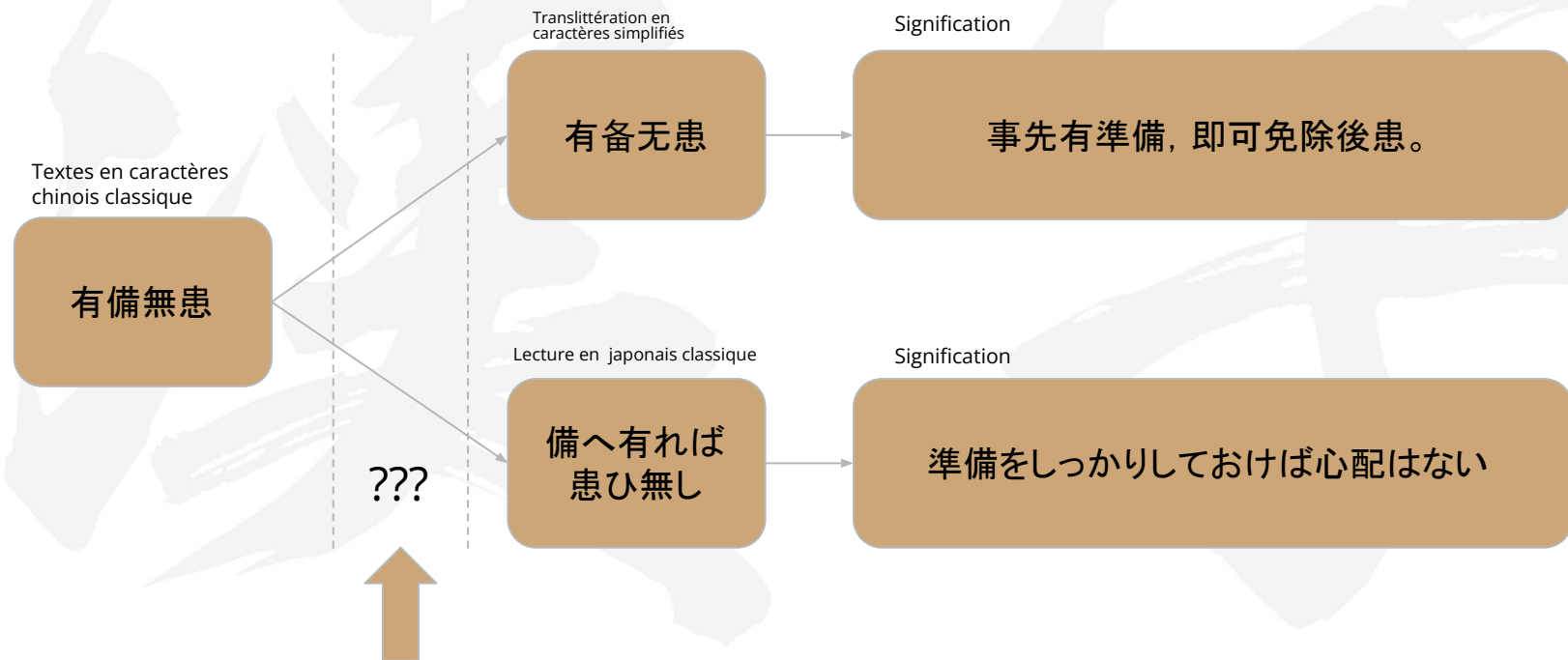
KANBUN

Japonais

- SOV
- langue agglutinante

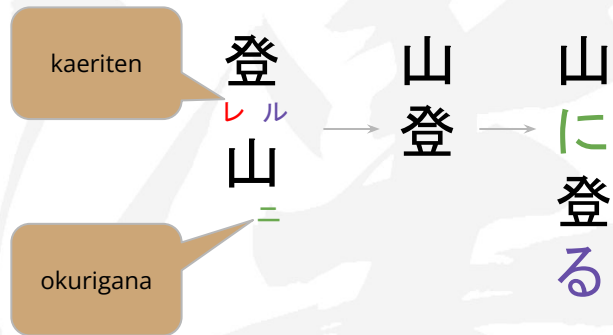
I. Présentation des données

A. Qu'est-ce que le Kanbun ?



I. Présentation des données

A. Qu'est-ce que le Kanbun ?



有備無患

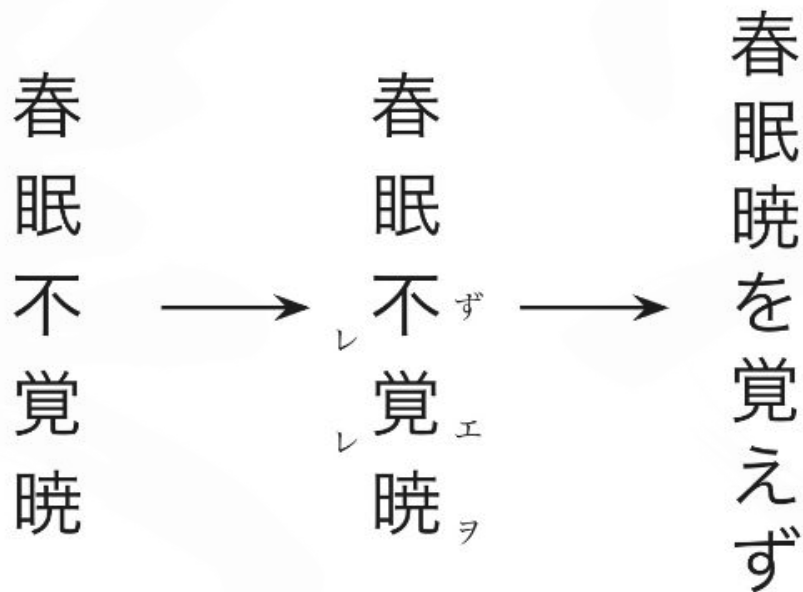
有
レ
備
無
レ
患

備有患無

備へ有れば患ひ無し

I. Présentation des données

A. Objectif du travail



I. Présentation des données

B. Les données

- 465 poèmes Tang annotés :

	poèmes	vers	caractères
train	372	2731	16411
dev	46	320	2038
test	47	370	2254

- Format :

id	id_poème	ver	lecture jp	ordre
0	0	中原還逐鹿	中原還た鹿を逐い	12354

I. Présentation des données

C. Prétraitements

- Extraction de l'ordre des caractères via un script à base de règles
- Annotation manuelle pour les tokens “spéciaux” :
 - qui ne doivent pas apparaître dans la lecture japonaise
 - qui doivent être lus deux fois
- Conversion des caractères classiques en caractères japonais :
 - À base de dictionnaires
 - ex :



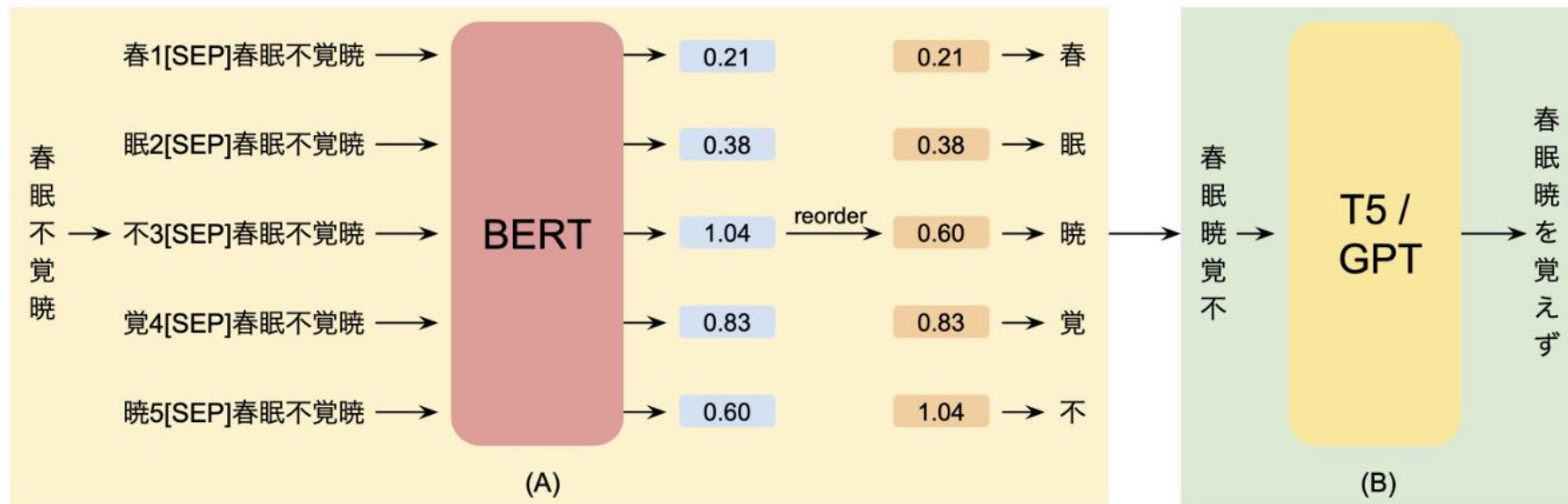
II. Présentation des expériences

A. Modèles (transformers)

- 5 modèles pour réordonner :
 - Taille : 12 couches, 768 dimensions cachées, 12 points mécanismes d'attention
 - Pré-entraînements : 2 sur du japonais, 2 sur du chinois, 1 sur du chinois classique
- 2 modèles pour traduire :
 - Pas de modèles japonais (voc trop restreint ; produit trop de UNK)
 - mT5 : entraîné sur le corpus mC4
 - mGPT : entraîné sur le corpus mC4 et wiki

II. Présentation des expériences

A. Modèles (fonctionnement)



II. Présentation des expériences

B. Mesures utilisées

1. Réarrangement des caractères

- Tau de Kendall : évalue le réarrangement au niveau du caractère
- Perfect Match Ration (PMR) : évalue le réarrangement au niveau de la séquence

2. Traduction automatique

- BLEU
- RIBES
- ROUGE-L
- BERTScore

II. Présentation des expériences

C. Annotation manuelle

3 annotateurs bilingues japonais-chinois ayant eu la note maximale à l'épreuve de Kanbun à l'examen d'entrée à l'université

Tâches :

- Réarrangements des caractères à partir des connaissances
- Evaluation de la traduction automatique selon 3 critères :
 - Relevance
 - Accuracy
 - Fluency

III. Présentation des résultats

A. Réarrangement des caractères

Model Setup	τ	PMR
UD-Kundoku	0.770	0.402
Human	0.844	0.606
BERT-japanese-char	0.898	0.637
RoBERTa-japanese-char-wwm	0.894	0.600
BERT-chinese	0.917	0.689
RoBERTa-chinese-wwm-ext	0.920	0.718
RoBERTa-classical-chinese-char	0.944	0.783

Table 3: Kendall's Tau (τ) and PMR scores of character reordering. UD-Kundoku is the baseline, and human scores are the average of the three annotators' results.

III. Présentation des résultats

B. Traduction automatique

Model Setup	BLEU	RIBES	ROUGE-L	BERTScore	Relevance	Accuracy	Fluency
UD-Kundoku	0.097	0.309	0.546	0.884	-	-	-
reference	-	-	-	-	4.958	4.951	4.949
mT5-small	0.317	0.428	0.659	0.914	3.219	3.002	3.153
mT5-base	0.462	0.520	0.735	0.930	-	-	-
mT5-large	0.514	0.583	0.747	0.934	3.948	3.884	3.904
mGPT	0.303	0.476	0.606	0.898	2.548	2.270	2.236

Table 4: Results of machine translation, containing the automatic and manual evaluation metrics. UD-Kundoku is the baseline, and reference is the Kanbun target of translation.



III. Présentation des résultats

C. La pipeline

Model Setup	BLEU	RIBES	ROUGE-L	BERTScore
mT5-small	0.317	0.428	0.659	0.914
+ reorder	0.328	0.420	0.701	0.916
+ reorder (gold)	0.359	0.451	0.727	0.919
mT5-base	0.462	0.520	0.735	0.930
+ reorder	0.413	0.486	0.735	0.926
+ reorder (gold)	0.461	0.529	0.770	0.932
mT5-large	0.514	0.583	0.747	0.934
+ reorder	0.479	0.551	0.748	0.931
+ reorder (gold)	0.502	0.573	0.774	0.935
mGPT	0.303	0.476	0.606	0.898
+ reorder	0.303	0.467	0.612	0.894
+ reorder (gold)	0.340	0.508	0.642	0.900

Conclusion

Succès

- Premier corpus Chinois Classique
→ Kanbun
- État de l'art atteint pour :
 - Réarrangement 
 - Traduction automatique 
- Code et dataset disponibles
- Possibilité de tester l'outil sur Hugging Face

Limites

- Corpus très restreint
- Évaluation qualitative non-vérifiée
- Manière de collecter le corpus non détaillée
- Pas (encore) pip installable



Nous vous remercions de votre attention

Bibliographie non exhaustive

- Hao Wang, Hirofumi Shimizu, and Daisuke Kawahara. 2023. [Kanbun-LM: Reading and Translating Classical Chinese in Japanese Methods by Language Models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8589–8601, Toronto, Canada. Association for Computational Linguistics.
- Yasuoka, Koichi. « 漢文の依存文法解析と返り点の関係について ». 日本漢字学会第1回研究大会予稿集 1 décembre 2018, 33-48.
- Yasuoka, Koichi. « Universal Dependencies Treebank of the Four Books in Classical Chinese ». DADH2019: 10th International Conference of Digital Archives and Digital Humanities, décembre 2019, 20-28.
- Cui, Baiyun, Yingming Li, et Zhongfei Zhang. 2020. « BERT-enhanced Relational Sentence Ordering Network ». In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, édité par Bonnie Webber, Trevor Cohn, Yulan He, et Yang Liu, 6310-20.
- Isozaki, Hideki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, et Hajime Tsukada. 2010. « Automatic Evaluation of Translation Quality for Distant Language Pairs ». In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, édité par Hang Li et Lluís Màrquez, 944-52. Cambridge, MA: Association for Computational Linguistics. <https://aclanthology.org/D10-1092>.
- Zhang*, Tianyi, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, et Yoav Artzi. 2019. « BERTScore: Evaluating Text Generation with BERT ». In . <https://openreview.net/forum?id=SkeHuCVFDr>.