

# Named Entity-Factored Transformer for Proper Noun Translation

Kohichi Takai Gen Hattori Akio Yoneyama  
Keiji Yasuda Katsuhito Sudoh Satoshi Nakamura

**KDDI Research Institute, Inc**  
**Nara Institute of Science and Technology**  
**Japan**

# Background

- ✓ Proper noun translation for Machine Translation(MT)
  - ❑ MT: useful in practical applications
  - ❑ Social implementation: importance of proper noun translation

*Watashi wa Akage no Anne no musical wo miniikitai*  
*(I want to watch a musical of Anne Green Gables)*

↓  
MT result : I want to see Anne's musical with red hair

Examples of proper noun (bleu)  
mistaken translation (red)

Problem: Proper noun translation for Machine Translation(MT)

Proper nouns are often processed as out-of-vocabulary (OOV) words in MT systems

Two major approach:

(1) Handcraft bilingual corpus or dictionary ⇒ High developing cost

(2) Subword ⇒ Doesn't always work well for proper noun translation

---

# Motivation

- ✓ Work well for proper noun translation by effectively **using training corpus**

## Proposed method based on subword approach

- ✓ Use **Named Entity (NE) features** for an enhance of NMT
- ✓ Also use distributions for an unclear NE of proper noun

*Akage no Anne*: Person, Work of Art, or Movie Title as NE

---

# Conventional Method

## Factored Transformer:

- ✓ García-Martínez et al use morphological and grammatical features of a word
- ✓ Jordi et al use grammatical features in low-resource NMT

Subword  
Embedding

Encoder

Decoder

Transformer

Subword  
Embedding

POS Feature  
Embedding

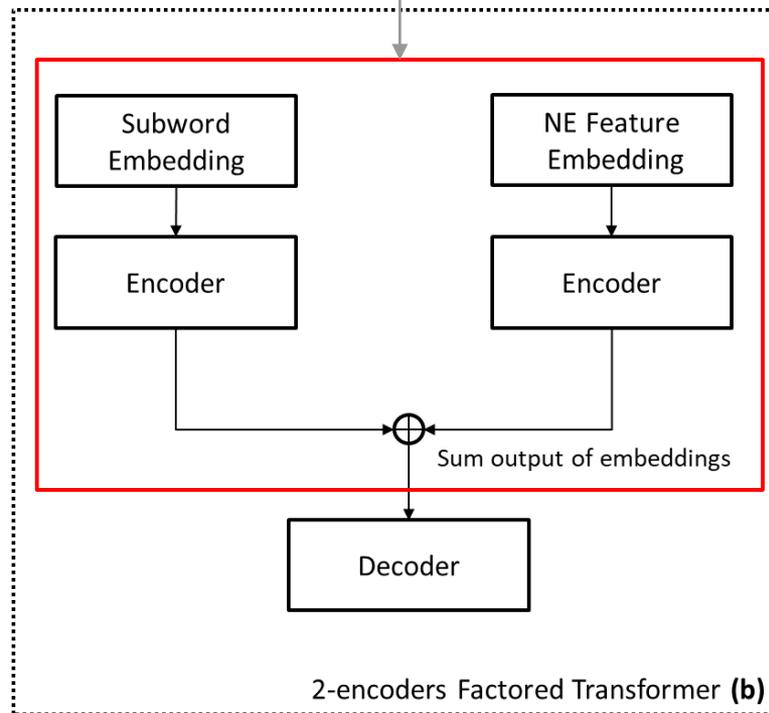
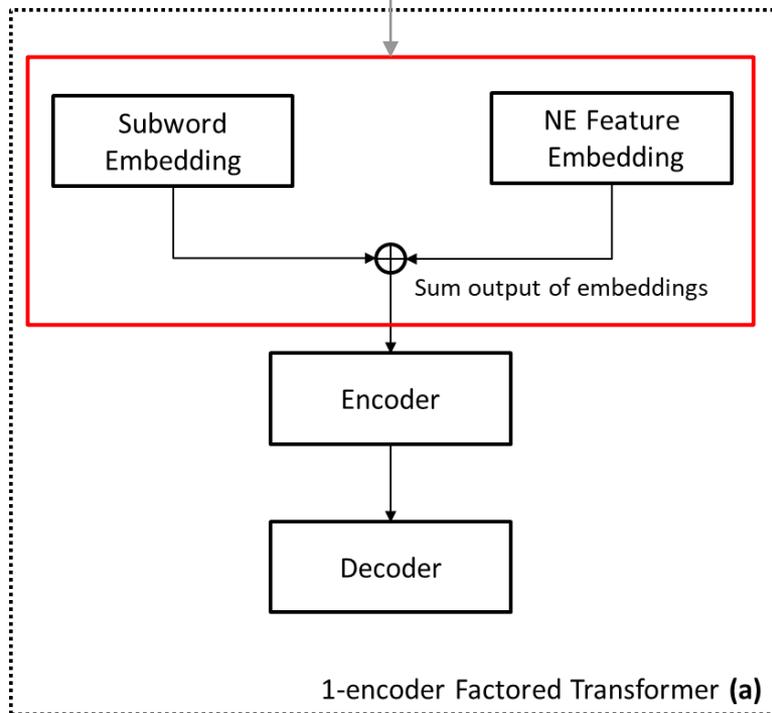
Encoder(s)

Decoder

Factored Transformer

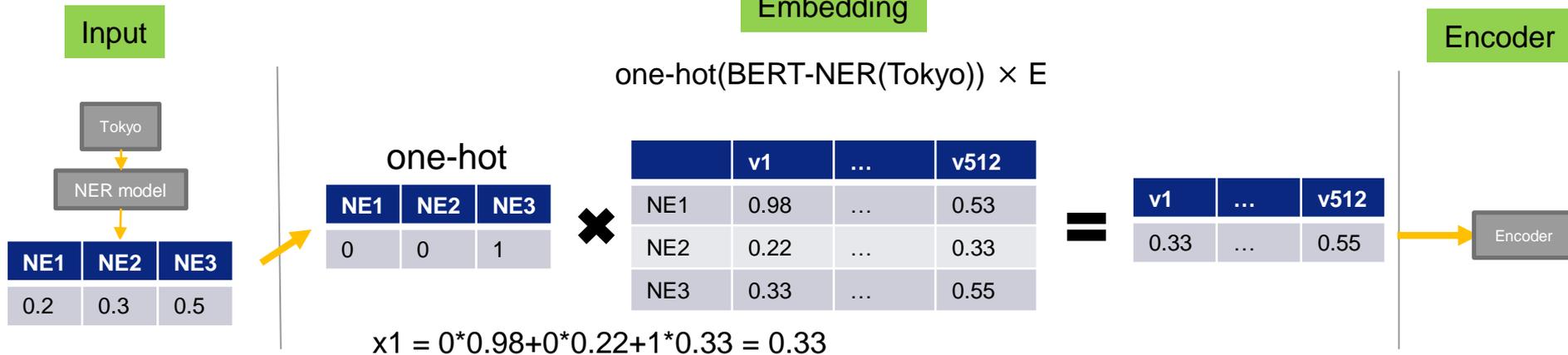
# Concatenation structures in Factored Transformer

Summed vectors are different

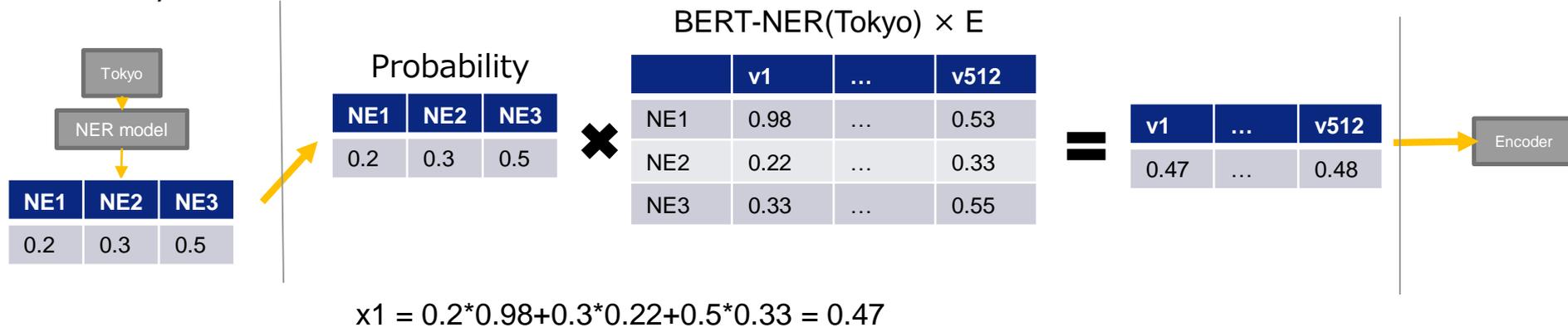


# NE features

- one-hot vector of NE



- Probability Distribution of NE



# Experiments ~Corpus~

- JParaCrawl: the training of the NMT models (Morishita et al., 2020)  
In Japanese-English chose 160,000 sentence pairs that contain proper nouns  
Sentence pair scores higher than 0.786, and shorter than 250 subwords
- Our field experiments with taxis in Japan:  
For Japanese-English task, we used an evaluation dataset of 271 sentences containing a single proper noun.
- WMT2020: For English-Japanese task

	Corpus	direction	# of sentence	# of subwords	# of uniq subwords
Training	JParaCrawl	JA-EN	159,888	5,318,140	10,073
Test	Field experiments	JA-EN	271	4,258	646
Training	JParaCrawl	EN-JA	10,116,570	332,520,888	47,087
Test	WMT2020	EN-JA	1,000	32,696	5,171

# Compared Methods

- **Transformer: baseline**
- Proposed method with combination  
1-encoder / 2-encoder  
NE one-hot vector / NE probability distribution vector

Baseline		
Transformer	PRPacc1	BLEU1
Proposed method	1-encoder	
	2-encoders	
Factored-Transformer with NE one-hot vector	PRPacc2	BLEU2
	PRPacc3	BLEU3
Factored-Transformer + NE probability Distribution vector	PRPacc4	BLEU4
	PRPacc5	BLEU5

# Compared Methods

- **Proposed method with combination**  
**1-encoder / 2-encoder**  
**NE one-hot vector / NE probability distribution vector**
- NER model  
Japanese: BERT-NER(Devlin et al., 2018)  
English: Stanza

Baseline		
Transformer	PRPacc1	BLEU1
Proposed method	1-encoder	
	2-encoders	
Factored-Transformer with NE one-hot vector	PRPacc2	BLEU2
	PRPacc3	BLEU3
Factored-Transformer + NE probability Distribution vector	PRPacc4	BLEU4
	PRPacc5	BLEU5

# Compared Methods

- **Proposed method with combination**

**1-encoder / 2-encoder**

NE one-hot vector / **NE probability distribution vector**

- NER model

Japanese: BERT-NER(Devlin et al., 2018)

English: Stanza\*

\*<http://nlp.stanford.edu/software/stanza/1.2.2/en/ner/ontonotes.pt>

Baseline		
Transformer	PRPacc1	BLEU1
Proposed method	1-encoder	
	2-encoders	
Factored-Transformer with NE one-hot vector	PRPacc2	BLEU2
	PRPacc3	BLEU3
Factored-Transformer + NE probability Distribution vector	PRPacc4	BLEU4
	PRPacc5	BLEU5

# Evaluation Metrix

- proper noun translation accuracy (**PRPacc**):
- **BLEU** (Papineni et al., 2002)

Baseline		
Transformer	<b>PRPacc1</b>	<b>BLEU1</b>
Proposed method	1-encoder	
	2-encoders	
Factored-Transformer with NE one-hot vector	<b>PRPacc2</b>	<b>BLEU2</b>
	<b>PRPacc3</b>	<b>BLEU3</b>
Factored-Transformer + NE probability Distribution vector	<b>PRPacc4</b>	<b>BLEU4</b>
	<b>PRPacc5</b>	<b>BLEU5</b>

# Experiments Results: PRPacc and BLEU

## ■ Japanese-English

Baseline		
Transformer	56.1	11.37
Proposed method	1-encoder	
	2-encoders	
Factored-Transformer with NE one-hot vector	43.2	10.1
	63.4	13.8
Factored-Transformer + NE probability Distribution vector	53.5	10.9
	<b>65.7</b>	<b>13.8</b>

Training set: 160k sentences of JParaCrawl  
Test set: 271 sentences of Gifu field experiments  
NER model: BERT-NER

## ■ English-Japanese

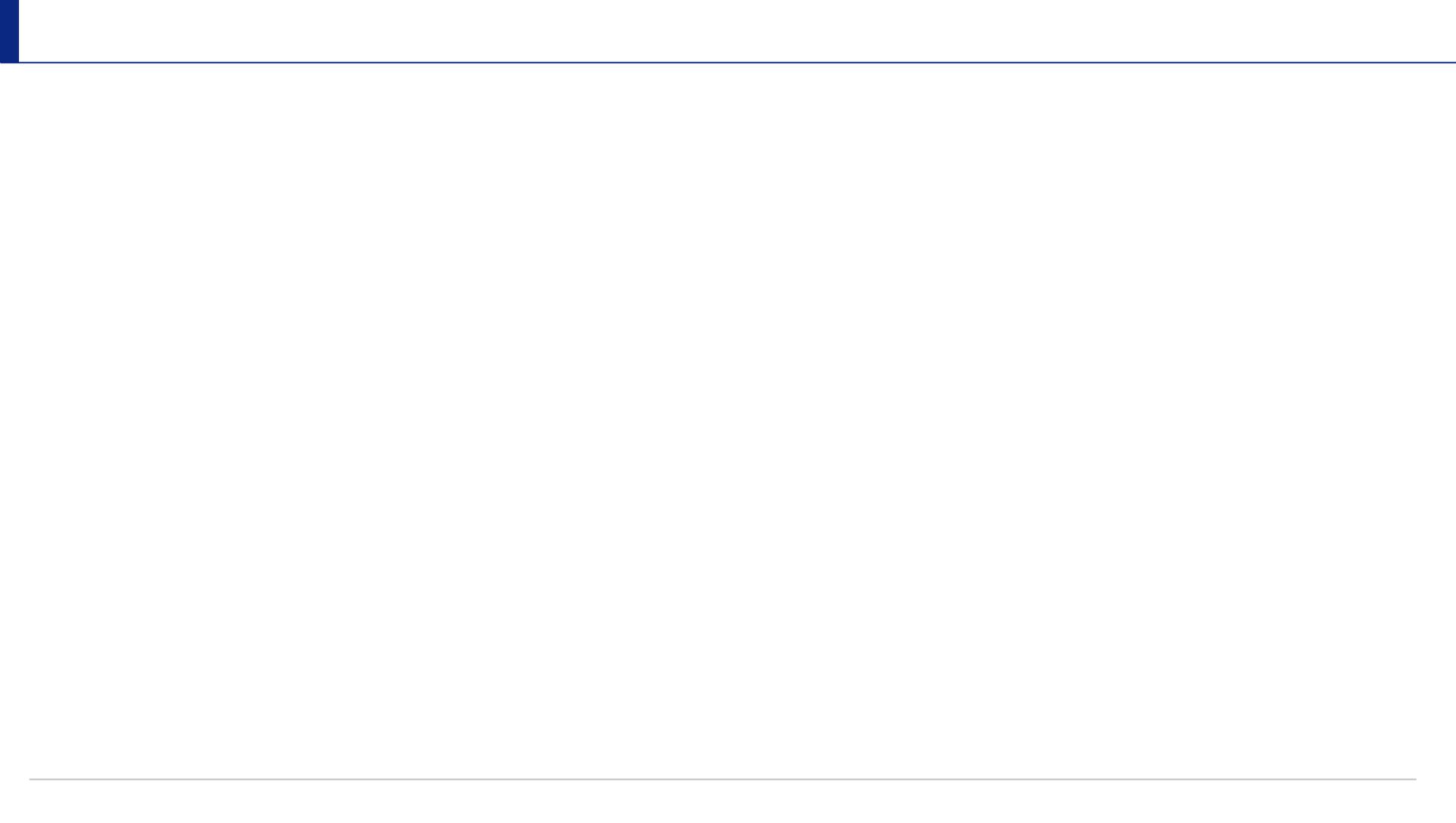
Baseline		
Transformer	46.5	17.54
Proposed method	1-encoder	
	2-encoders	
Factored-Transformer with NE one-hot vector	<b>50.1</b>	<b>18.8</b>
	47.5	17.8
Factored-Transformer + NE probability Distribution vector	49.5	18.4
	46.7	17.6

Training set: 10m sentences of JParaCrawl  
Test set: 1,000 sentences of WMT 2020 News  
NER model: Stanza in Stanford NLP toolkit

# Conclusion

The NE feature vectors are injected into Factored Transformer model as factors.

- Japanese-to-English experiments: small bilingual training corpus
    - 2-encoders and NE probability distribution vector is best
    - 9.6 points in proper noun accuracy and 2.5 points in the BLEU
  - English-to-Japanese experiments:
    - 1-encoder and NE one-hot vector is best
    - 3.6 points in proper noun accuracy and 1.3 points in the BLEU
  - In English-to-Japanese the improvement in PRPacc and BLEU was smaller:
    - One of reasons is the difference in the training data sizes and NER
    - Another is the difference in the degrees of difficulty in these domains
-



# Appendix

## Approaches for proper noun translation

- ✓ Subword-based approach: Sennrich et al 2016  
Subwords decompose a word into shorter units but does not always work on a proper noun translation.
- ✓ Class-based approach: Okuma et al 2008, Li et al 2016  
Hand-crafted bilingual lexicon as external knowledge uses a bilingual proper noun dictionary.
- ✓ LCD(lexically constrained decoder) and LeCA(Lexical-Constraint-Aware NMT) approach:  
Hokamp et al 2017, Chen et al 2020, Chousa et al 2021  
Select target language sentence constrained by a bilingual dictionary  
Extends the beam search algorithm to find the hypothesis that contains all of the proper nouns

	Subword-based	Class-based	LCD	LeCA	Proposed method
Source side proper noun	✓	✓ + NE	✓	✓	✓
Target side proper noun	Nan	✓	✓	✓	Nan
Real-time processing	✓	✓	Not enough	✓	✓

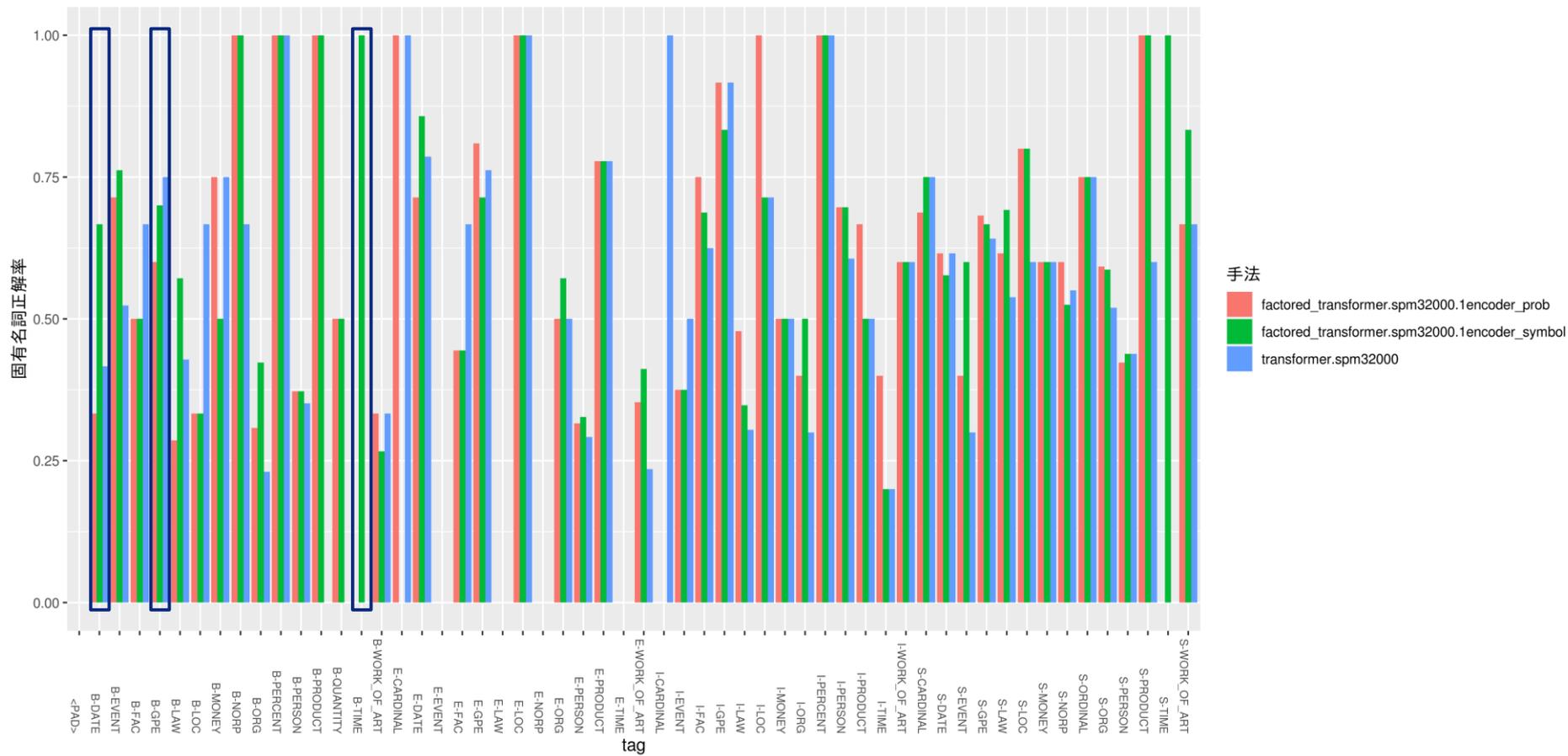
# Appendix

(1) Input sentence: 飛騨牛を使った料理です。		
baseline	-	it is a cooking cuisine using hida flavored cuisine .
1-encoder	symbols	i tried to use the hida dish of the cuisine .
1-encoder	Probability Distributions	i used cuisine as a dish of hidafuruga .
2-encoder	symbols	it's a dish that was used with <b>hida beefs</b> .
2-encoder	Probability Distributions	it is a dish used with <b>hida beef</b> .
(2) Input sentence: 戦国武将の豊田秀吉が三日で作ったと言われています。		
baseline	-	the minister of battlefield was created by the third day in yoshino , Japan
1-encoder	symbols	thirdly , the minister of foreign affairs said that japan had been pursuing three days .
1-encoder	Probability Distributions	the minister's commendation was created in toyotomi , where he said , "it was created in 3 days"
2-encoder	symbols	<b>toyotomi hideyoshi's</b> president of the battleship is said to be made in three days .
2-encoder	Probability Distributions	the minister of battlefield headquartered in mie , japan , said it was created on third day .
(3) Input sentence: このお城は豊田秀吉が作りました。		
baseline	-	this castle was created by an excellent Japanese castle .
1-encoder	symbols	this castle was created by yoshino <b>hideyoshino hideyoshinori</b> .
1-encoder	Probability Distributions	this castle was created by minister <b>toyotomi hideyoshi</b> .
2-encoder	symbols	the castle of this castle was created by <b>toyotomi hideyoshi</b> .
2-encoder	Probability Distributions	this castle was created by <b>toyotomi hideyoshi</b> .
(4) Input sentence: 岐阜は初めてですか？		
baseline	-	what is <b>gifu</b> ?
1-encoder	symbols	what is the first battery ?
1-encoder	Probability Distributions	what is the first time ?
2-encoder	symbols	is <b>gifu</b> is the first time ?
2-encoder	Probability Distributions	<b>gifu</b> is the first time ?
(5) Input sentence: 島根県ですここから六時間ほどかかります。		
baseline	-	It takes about 6 hours from shimane here .
1-encoder	symbols	you can take about 6 hours from shimane island .
1-encoder	Probability Distributions	it takes about 6 hours to get here from <b>shimane prefecture</b> here .
2-encoder	symbols	it takes about six hours from here to the island .
2-encoder	Probability Distributions	it takes about six hours from here to <b>shimane prefecture</b> .



# Appendix ~ 2. proper noun acrcy each category ~

カテゴリ別固有名詞正解率



## Appendix ~2. Token's difference of between subword and NE inputs~

		NE token											
		買い物	も	最近	は	名古屋	へ	出かけ	方	が	多い	です	。
Subword token		0	0	0	0	0	0	0	0	0	0	0	0
	買い物	3	0	0	0	0	0	0	0	0	1	0	0
	も	3	4	3	3	3	3	3	3	3	3	3	3
	最近	3	4	6	4	4	4	4	4	4	4	4	4
	は	3	4	6	7	6	6	6	6	6	6	6	6
	名古屋	3	4	6	7	10	7	7	7	7	7	7	7
	へ	3	4	6	7	10	11	10	10	10	10	10	10
	出	3	4	6	7	10	11	12	11	11	11	11	11
	かけ	3	4	6	7	10	11	14	12	12	12	12	12
	る	3	4	6	7	10	11	15	14	14	14	14	14
	方	3	4	6	7	10	11	15	16	15	15	15	15
	が多い	4	4	6	7	10	11	15	16	17	18	16	16
	です	4	4	6	7	10	11	15	16	17	18	20	18
。	4	4	6	7	10	11	15	16	17	18	20	21	

- How to synchronize two tokens of NE and subword whose sometimes has multiple POS  
⇒ In this research, after Japanese tagger, divide token as subword