W-NUT 2020

Sentence Boundary Detection on Line Breaks in Japanese

Yuta Hayashibe
(Megagon Labs, Tokyo, Japan, Recruit Co., Ltd.)

Kensuke Mitsuzawa
(Freelance)

Sentence Boundary (SB)

- Sentences are basic units of NLP
- In Japanese, periods (。), exclamation marks (!), and question marks (?) are delimiters to segment sentences in most cases.
- Some characters may or may not be sentence boundaries depending on the context

Line break as SB in Japanese

Line breaks in web texts can be SBs.

オアシズの大久保さんが最近気になります⏎
テレビはどの番組によく出るんですか?

Ms. Okubo of "Oasiz" has been on my mind lately ⏎
What TV shows does she often appear on?

Corpus Preparation

Corpus	Documents	Sentences	LBs	LBs w/o SB
BCCWJ	2,918	44,760	23,099	1,702
(PN)	340	8,747	3,069	0
(PB)	83	8,956	3,290	0
(PM)	86	9,424	3,890	0
(OW)	62	3,751	2,223	0
(OC)	*1,876	6,413	4,055	818
(OY)	471	7,469	6,572	884
Jalan-F	500	3,290	1,484	170
Jalan-A	298	?	1,193	153

Each two letters for BCCWJ represents newspaper articles (PN), books (PB), magazines (PM), white papers (OW), QA texts in the Internet (OC) and blog texts (OY). *In OC, we regarded an answer section for a question section is in a different document in the question.

BCCWJ
The balanced corpus of contemporary written Japanese

Jalan-F
Full SB annotation on reviews posted on Jalan, which is a popular travel information web site

Jalan-A
Partial annotation on reviews in an atypical writing style where they do not contain typical Japanese periods ("。")

P-BCCWJ
Pseudo corpus of BCCWJ

P-Jalan
Pseudo corpus consisted of 10k reviews extracted from Jalan

*Pseudo annotation: Replaced typical Japanese sentence boundaries "。" into line breaks and regard all of them as sentence boundaries, and ten replaced ideographic commas "、" into line breaks with 50% probability.

SB Detector Setup

- Fine-tuned BERT model pre-trained on Japanese Wikipedia by Tohoku University
- Texts are first tokenized with MeCab morphological parser and then spitted into subwords by WordPiece (Size=32k)
- "Sentence boundary" (SB) and "Not sentence boundary" (NSB) for line breaks, and "Others" (O) for tokens that are not line breaks
- Maximum sequence length 320, the training batch size 32, and the number of epochs five (If the maximum number of input tokens is exceeded, we divide them into multiple inputs)

Token	Gold	Prediction	Evaluation
ます (is)	O	SB	(ignored)
⏎	SB	SB	TP
テレビ (TV)	O	O	(ignored)
⏎	NSB	SB	FP
は (is)	O	NSB	(ignored)

Exp1: Impact of Domain

- We can make reasonably accurate models using training data even from different domains
- On the other hand, F1 scores for Jalan-F test data are close to 100 for all models
- Therefore, we consider Jalan-F only contains simple cases.

Test	Train	TP	TN	FP	FN	F_1
BCCWJ	BCCWJ	4,029	568	50	96	**98.2**
	Jalan-F	3,749	520	98	376	94.1
	Jalan-A	4,014	325	293	111	95.2
	Jalan-F+A	3,921	559	59	204	96.8
Jalan-F	BCCWJ	258	18	0	0	**100.0**
	Jalan-F	258	15	3	0	99.4
	Jalan-A	258	7	11	0	97.9
	Jalan-F+A	258	17	1	0	99.8

Exp2: Impact of Writing Styles

- Models trained on a large amount of data are more accurate, even if the writing styles are different.

Test	Train	TP	TN	FP	FN	F_1
Jalan-A	BCCWJ	210	46	1	11	**97.2**
	Jalan-F	188	39	8	33	90.2
	Jalan-A	204	27	20	17	91.7
	Jalan-F+A	202	45	2	19	95.1

Exp3: Effect of Pseudo Corpora

防災対策を構築する必要がある。⏎ </s>
消防庁においては、. . .
(It is necessary to build disaster prevention measures. ⏎</s>
In the fire and disaster management agency, . . .)

Example of a false negative by the model P-BCCWJ

Test	Train	TP	TN	FP	FN	F_1
BCCWJ	P-BCCWJ	2,715	570	48	1,410	78.8
	P-Jalan	1,868	575	43	2,257	61.9
Jalan-A	P-BCCWJ	200	46	1	21	94.8
	P-Jalan	192	46	1	29	92.8

- They were often wrong even in the almost obvious cases where periods "。" were just before line breaks.
- The F1 scores of the models P-BCCWJ and P-Jalan are respectively 94.8 and 92.8. Though they are better than one of the model Jalan-F (90.2), worse than one of the model Jalan-F+A (95.1).
- These results suggest that although a sentence boundary detector with pseudo-corpus could achieve moderate performance, we can obtain better detectors by training with annotated corpora.