Sentence Boundary Detection on Line Breaks in Japanese

Yuta Hayashibe*
hayashibe@megagon.ai
Megagon Labs, Tokyo, Japan, Recruit Co., Ltd.
7-3-5 Ginza Chuo-ku, Tokyo,
104-8227, Japan

Kensuke Mitsuzawa*
kensuke-mi@megagon.ai
Freelance
7-3-5 Ginza Chuo-ku, Tokyo,
104-8227, Japan

Abstract

For NLP, sentence boundary detection (SBD) is an essential task to decompose a text into sentences. Most of the previous studies have used a simple rule that uses only typical characters as sentence boundaries. However, some characters may or may not be sentence boundaries depending on the context. We focused on line breaks in them. We newly constructed annotated corpora, implemented sentence boundary detectors, and analyzed performance of SBD in several settings.

1 Introduction

Many NLP tasks treat a sentence as a unit of processing. The task for decomposing a text into sentences is called sentence boundary detection (SBD). In Japanese, periods (e.g. "。", "."), exclamation marks, and question marks are delimiters to segment sentences in most cases. For this reason, the SBD in most studies takes only the positions of these typical delimiters as sentence boundaries. For example, in the construction of the "Web Japanese N-gram database[1]" provided by Google, Inc., they extracted sentences by segmenting on their positions.

However, line breaks can also indicate sentence boundaries without periods as the following text[2].

> オアシズの大久保さんが最近気になります⏎</s>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 ⏎</s>What TV shows does she often appear on?) (1)

Many line breaks do not follow the typical delimiters. For example, 33.4% of line breaks in the balanced corpus of contemporary written Japanese (BCCWJ) (Maekawa, 2008) were not followed by them. On the other hand, line breaks may be placed in the middle of a sentence. Therefore, we can not simply treat the positions of line breaks as sentence boundaries.

> 最近の映画で⏎ゲイリー・オールドマンが出演している映画ってありますか？ (Among recent movies ⏎ are there any with Gary Oldman?) (2)

This type of line break is used to make long sentences easy to read. Shinmori et al. (2003) performed a structural analysis of Japanese patent documents. They reported that 48.5% of the first claim in the 59,968 patent documents contain line breaks in the sentence. They explain it is common that claims written in Japanese are described in one sentence and the use of line breaks is intended to improve readability.

There can be more sentence boundaries than these. Nishimura (2003) showed that there are more than six variations of Japanese sentence boundaries in an online forum: Description of actions (e.g. "(照)": embarrassment, "(涙)": tears), "Smiley" Icons (e.g. "(*^∇^*)", "§^。^§"), and so on. Sakai (2013) conducted a linguistic analysis of Japanese emails written by young people on their mobile phones and found that about 63% of the emails used emoticons instead of punctuation marks for sentence boundaries.

In this paper, we focus on SBD on line breaks in Japanese. We newly construct annotated corpora to answer the following three research questions:

1. Is it possible to train a sentence boundary detector on line breaks with annotated corpora? (Section 4.2)

*Both authors contributed equally.

[1] https://www.gsk.or.jp/files/catalog/GSK2007-C/GSK2007C_README.utf8.txt

[2] In this paper, we use "⏎" to show line breaks and "</s>" to show sentence boundaries.

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 71–75

2. Is a trained model accurate enough to work with data in another writing style? (Section 4.3)

3. Is it possible to train a sentence boundary detector with unannotated corpora? (Section 4.4)

2 Related Work

Zhu et al. (2007) removed noise in English email text by removing extra lines and spaces and restoring wrong cases of characters. They show that 49.5% of the noise in about 5,000 texts is due to line breaks. Three labels for line breaks were trained and predicted by the Conditional Random Fields (CRF) algorithm (Lafferty et al., 2001): PRV (Preserve line break), RPA (Replace line break by space), DEL (Delete line break). The accuracy is reported as F-measure 93.75.

Huang and Chen (2011) insist that the concept of "sentences" is fuzzier and less-defined in Chinese, and Native Chinese writers seldom follow the usage guidelines of punctuation marks. They listed the symbols used as sentence boundaries, such as whitespaces, commas, periods, line breaks. They reported F1 of manual SBD is 81.18 and one of CRF is 77.48.

Stanza[3] (Qi et al., 2020) is a language-agnostic fully neural pipeline for text analysis, including tokenization, multiword token expansion, lemmatization, part-of-speech and morphological feature tagging, dependency parsing, and named entity recognition. Unlike most existing toolkits, it does tokenization and SBD at the same time by using a bidirectional long short-term memory network (Graves and Schmidhuber, 2005) (Bi-LSTM) for characters in texts. It provides models for 66 languages including Japanese. The Japanese model is trained with with UD Japanese GSD[4]. Its architecture enables SBD on any characters, including line breaks. However the training corpus does not contain line breaks. Therefore the model can not perform SBD on line breaks.

3 Corpus Preparation

3.1 BCCWJ

The balanced corpus of contemporary written Japanese (BCCWJ) is a corpus annotated with sentence boundaries (Konishi et al., 2015) and morphological information. It covers a wide range of genres such as books, magazines, newspapers, business reports, blogs, internet forums, and textbooks. Some of them contain line breaks. Table 1 shows the statistics of the corpus[5,6] It consists 44,770 sentences in 2,918 documents. They contain 1,721 line breaks that do not segment sentences out of 26,056 line breaks. Such line breaks are contained only limited domains: QA texts and blog texts on the web.

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

Table 1: Statistics of corpus. LB means "line break," and SB means "sentence boundary." Each two letters for BCCWJ reperesents 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 setcion for a question section is in a different document in the question.

3.2 Jalan Corpora

We create two kinds of Japanese corpora with sentence boundary annotation: Jalan-F and Jalan-A, in order to perform experiments in another domain and another writing style. Both of them are composed of a part of hotel reviews posted on Jalan[7], which is a popular travel information web site. Table 1 shows the statistics of the corpora. All annotations are performed by one worker and confirmed by another worker.

Jalan-F[8] comprises 500 reviews. We fully anno-

[3] https://stanfordnlp.github.io/stanza/

[4] https://universaldependencies.org/treebanks/ja_gsd/

[5] In this paper, we removed all line breaks at the end of documents because they are obvious sentence boundaries. Additionally, if there is a series of line breaks or a space before or after a line break, we replaced it with a single line break.

[6] We only used the "core" in BCCWJ. Its annotation is manual while no manual correction is performed for "non-core."

[7] https://www.jalan.net

[8] The "F" is an abbreviation for "full annotation."

tated sentence boundaries for all texts. As a result, we found 3,290 sentences. It contains 1,484 line breaks. Out of them, 170 line breaks do not segment sentences.

Jalan-A[9] comprises 298 reviews in an atypical writing style. They do not contain typical Japanese periods (“。”). This is an example.

個室を利用させていただきました⏎
</s>清潔感があり⏎お部屋も広く⏎
またお邪魔させていただきますね⏎
</s>スタッフの対応も最高でした
(We stayed in a private room ⏎</s>
The room was clean ⏎ and spacious,
⏎ so we'll be back again ⏎</s> The
staffs were great) (3)

Some line breaks segment sentences and some do not. We only annotated sentence boundaries on line breaks. While the number of boundaries is 1,374, there may be more sentences. It contains 1,983 line breaks. Out of them, 153 line breaks do not segment sentences.

3.3 Pseudo Annotation Corpora

To answer the third research question, we created two pseudo annotation corpora: P-BCCWJ and P-Jalan. First, we removed all line breaks from BCCWJ and 10,000 reviews additionally extracted from Jalan. Then, we replaced typical Japanese sentence boundaries “。” into line breaks and regard all of them as sentence boundaries. Finally, we replaced ideographic commas “、” into line breaks with 50% probability. This is an example.

Original: 眺めのよいところで、遠くを
見ることですよ。(It is to look into the
distance from a good view.)
Pseudo annotation: 眺めのよいところ
で⏎遠くを見ることですよ⏎</s> (4)

4 Experiments

4.1 Experiment Settings

We create sentence boundary detectors by fine-turning the BERT (Devlin et al., 2019) model[10] pre-trained on Japanese Wikipedia by Tohoku University. Texts are first tokenized with MeCab[11] morphological parser and then spitted into subwords by WordPiece. Its vocabulary size is 32,000. We exploit implementations of sequence labeling in “transformers”[12] by Hugging Face with three labels[13]: “Sentence boundary” (SB) and “Not sentence boundary” (NSB) for line breaks, and “Others” (O) for tokens that are not line breaks. We only use predictions for line breaks. Table 2 shows an example of input, output, and evaluation for detectors. In training, all tokens are labeled “O” except for line breaks. Whatever predictions are output for them, we do not consider them in the evaluation. Line breaks are labeled “SB” or “NSB” for training. We recognize sentence boundaries only on the tokens whose predictions are “SB.”

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)

Table 2: An example of input, output, and evaluation for sentence boundary detectors

We set the 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. We perform the Unicode NFKC normalization for all inputs.

For training and evaluation, we exclude 663 documents from BCCWJ and 164 documents from Jalan-F that do not contain line breaks. Each corpus of BCCWJ, Jalan-F, and Jalan-A is divided into 8:2 for learning and training. We built four models by using the three training sources and the data from the combination of Jalan-F and Jalan-A.

4.2 Experiments 1: Impact of Domains

First, we investigate the impact of domains. As shown in Table 3, In BCCWJ test data, the F_1 score of the model Jalan-F+A (96.8) is not very

[9] The “A” is an abbreviation for “atypical.”

[10] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

[11] https://github.com/taku910/mecab

[12] https://github.com/huggingface/transformers

[13] We did a preliminary experiment with binary labels “Sentence boundary” (SB) and “Not sentence boundary” (NSB), but it was low performance.

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

Table 3: SBD Performance on line breaks by models traied with annotated copora

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

Table 4: SBD Performance on line breaks in atypical writing style by models traied with annotated copora

Test	Train	TP	TN	FP	FN	F_1
P-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

Table 5: SBD Performance on line breaks by models traied with pseudo copora

bad compared to one of the model BCCWJ (98.2). This shows that we can make reasonably accurate models using training data even from different domains. On the other hand, F_1 scores for Jalan-F test data are close to 100 for all models. Therefore, we consider Jalan-F only contains simple cases.

4.3 Experiments 2: Impact of Writing Styles

Second, we investigate the impact of writing styles. As shown in Table 4, the F_1 score of the model BCCWJ is the best (97.2) among the four models. This shows that models trained on a large amount of data are more accurate, even if the writing styles are different.

4.4 Experiments 3: Effect of Pseudo Corpora

Third, we investigate the effect of pseudo corpora. Table 5 shows the result. The F_1 scores of the model P-BCCWJ for BCCWJ is 78.8. It is much worse than one of the model BCCWJ (98.2). This is an example of a false negative (FN) by the model P-BCCWJ.

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, ...) (5)

They were often wrong even in the almost obvious cases where periods "。" were just before line breaks.

The F_1 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.

5 Conclusion

We implemented sentence boundary detectors by using BERT and revealed the following facts:

- It is possible to train a sentence boundary detector on line breaks with annotated corpora.
- Training with much annotation data is effective even for texts in another writing style.
- Although it is possible to train a sentence boundary detector even with pseudo-corpus to some extent, more performance will be gained by training with annotated corpora.

There are two main issues that we need to address in the future. The first issue is to do is to use active learning to increase the number of learning examples and improve accuracy. The second issue is to perform SBD for other atypical sentence boundary expressions other than line breaks.

Acknowledgements

We recognize Dr. Yuki Arase at Osaka University for the many discussions and insightful comments. Furthermore, we thank the anonymous reviewers for their careful reading and valuable comments.

References

Jacob Devlin, Ming-Wei Chang, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks*, 18(5):602–610.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Pause and Stop Labeling for Chinese Sentence Boundary Detection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 146–153.

Hikari Konishi, Takenori Nakamura, et al. 2015. Correction of Sentence Boundaries in the Balanced Corpus of Contemporary Written Japanese DVD Version 1.0. *NINJAL Research Papers*, (9):81–100.

John Lafferty, Andrew McCallum, et al. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

Kikuo Maekawa. 2008. Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Yukiko Nishimura. 2003. Linguistic Innovations and Interactional Features of Casual Online Communication in Japanese. *Journal of Computer-Mediated Communication*, 9(1).

Peng Qi, Yuhao Zhang, et al. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Noboru Sakai. 2013. The role of sentence closing as an emotional marker: A case of Japanese mobile phone e-mail. *Discourse, Context & Media*, 2(3):149–155.

Akihiro Shinmori, Manabu Okumura, et al. 2003. Patent Claim Processing for Readability: Structure Analysis and Term Explanation. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*, pages 56–65.

Conghui Zhu, Jie Tang, et al. 2007. A Unified Tagging Approach to Text Normalization. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 688–695.