A Japanese Corpus of Many Specialized Domains for Word Segmentation and Part-of-Speech Tagging

Shohei Higashiyama[1], **Masao Ideuchi**[1,2], **Masao Utiyama**[1],
Yoshiaki Oida[2], **Eiichiro Sumita**[1]
[1]National Institute of Information and Communications Technology, Kyoto, Japan
[2]FUJITSU LIMITED, Tokyo, Japan
{shohei.higashiyama,masao.ideuchi,mutiyama,eiichiro.sumita}
@nict.go.jp,oida.yoshiaki@fujitsu.com

Abstract

We present a Japanese morphological corpus of sentences from 27 specialized domains for the two tasks of word segmentation and part-of-speech tagging. Experiments on the corpus demonstrated that recent neural models with domain adaptation techniques and pretrained language models achieved accurate performance for the two tasks for many specialized domains.

1 Introduction

Because the Japanese language has no explicit word delimiters, word segmentation (WS) and part-of-speech (POS) tagging are fundamental and important steps for downstream natural language processing (NLP) tasks, such as linguistic analysis and text mining. In previous studies, researchers devoted much effort to developing WS and POS tagging systems (Kudo et al., 2004; Neubig et al., 2011; Tolmachev et al., 2020), often as Japanese morphological analysis, which simultaneously performs WS, POS tagging, and lemmatization. However, the majority of existing systems were evaluated on general domains, such as news and the web.

Although researchers constructed morphologically annotated corpora of specialized domain text, the domains in publicly available corpora are limited, for example, Mori et al. (2014, 2016); Harashima and Hiramatsu (2020). Moreover, researchers proposed domain-specific or domain-independent adaptation methods (Tsuboi et al., 2008; Fujita et al., 2014; Sudoh et al., 2014; Kameko et al., 2015; Higashiyama et al., 2020); however, they evaluated their systems on one or a few specialized domains. Therefore, a benchmark corpus that includes text for many specialized domains is beneficial for conducting comprehensive system evaluation and developing robust adaptation methods for many domains.

In this paper, we present a Japanese Corpus of Many Specialized Domains (JCMS) for WS and POS tagging. The corpus consists of 32,310 sentences annotated with word boundary and POS tag information for 27 specialized domains. Using our corpus, we evaluated existing morphological analysis and WS systems, including popular non-neural systems and recent neural cross-domain systems. Our experiments demonstrated that (1) most systems trained with general source domain resources resulted in degraded performance for specialized target domains; however, (2) domain adaptation (DA) techniques and pretrained language models (PLMs) contributed to robust performance without annotated text for target domains.[1]

2 Construction of the JCMS

2.1 Data Sources and Domains

To construct a multi-domain corpus with public availability and domain diversity, we extracted raw sentences from several publicly available corpora with their sentence segmentation.

To include various science and engineering text (SCI) in our corpus, we used the ASPEC[2] (Nakazawa et al., 2016), NITCIR-9 PatentMT test collection[3] (Fujii et al., 2010), and NTCIR-11 MedNLP-2 test collection[4] (Aramaki et al., 2014). The ASPEC is a parallel corpus of paper abstracts in various scientific fields; we extracted 24K Japanese sentences for 20 domains (from AGR to TRA, as shown in Table 1). The PatentMT data form a parallel corpus of patent documents (PAT); we extracted 1K sentences. The MedNLP-2 data consist of pseudo electronic medical records (EMR); we used all 1.4K unique sentences.

To include other domain text, we used the BC-

[1] The JCMS will be available at https://github.com/shigashiyama/jcms.

[2] https://jipsti.jst.go.jp/aspec/

[3] http://research.nii.ac.jp/ntcir/permission/ntcir-9/perm-en-PatentMT.html

[4] https://research.nii.ac.jp/ntcir/permission/ntcir-11/perm-en-MedNLP.html

Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems, pages 1 - 10
November 20, 2022 ©2022 Association for Computational Linguistics

Group	Domain		Sent.	Word
SCI	AGR	agriculture, forestry, fisheries	900	19.3k
	BIO	biology	1,000	20.2k
	CHE-B	basic chemistry	1,700	38.3k
	CHE-E	chemical eng.	750	18.2k
	CHE-I	chemical industry	950	18.7k
	CON	construction eng.	1,700	39.0k
	ELC	electrical eng.	2,000	39.3k
	ENE	energy eng.	1,360	37.5k
	ENV	environmental eng.	870	19.0k
	ETH	earth and space science	800	19.2k
	INF	information eng.	900	18.9k
	MAN	eng. management	1,500	36.8k
	MEC	mechanical eng.	1,750	38.3k
	MED	medicine	1,300	20.0k
	MIN	mining eng.	640	19.1k
	NUC	nuclear eng.	800	18.6k
	PHY	physics	1,000	17.8k
	SYS	system control eng.	1,500	36.8k
	THM	thermal eng.	1,500	38.3k
	TRA	traffic and transportation eng.	1,430	37.7k
	PAT	patent	1,000	19.1k
	EMR	electronic medical record	1,362	28.7k
GOV	LAW	law	1,060	37.9k
	DIE	diet minute	650	36.3k
	PRM	PR magazine	1,238	19.1k
OTH	TBK	textbook	1,650	17.7k
	VRS	verse	1,000	15.9k
	Total		32,310	726k

Table 1: Statistics of the JCMS SUW data. Scientific (SCI), government document (GOV), and other (OTH) domains are grouped.

CWJ[5] (Maekawa et al., 2014) non-core data and extracted 3K sentences from three government documents (GOV): letter of the law (LAW), minutes of the national diet (DIE), and public relations magazines of local governments (PRM). Additionally, we extracted 2.7K sentences from two other domains (OTH): textbooks (TBK)[6] and Japanese verse (VRS).

As shown in Table 1, the JCMS included 27 domains and 16–40K words per domain. We regard PAT and TBK data as single domains, although they include text in multiple academic or industry fields.

2.2 Segmentation Criteria and POS Tag Sets

Regarding the word boundary and POS tag annotation, we adopted two WS criteria (and corresponding POS tag sets). One is the short unit word (SUW). The SUW was designed by the National Institute for Japanese Language and Linguistics (NINJAL) to achieve consistent WS and has been adopted in various NINJAL corpora (Oka et al., 2020). Additionally, we defined a new criterion, SUW-SC, by separating conjugate words (verb, adjective, verbal/adjectival suffix, and auxiliary verb) into stems and conjugation endings, similar to EDR (2001) and Mori et al. (2014).[7] This criterion has the advantage that different conjugation forms of (regular) conjugate words (e.g., 読-む *yo-mu* ‘read’, 読-ま *yo-ma*, and 読-み *yo-mi*) can be treated as the same stem token (e.g., 読 *yo*) without an additional lemmatization process. The SUW-SC POS tags that differ from the SUW POS tags are shown in Appendix A.

2.3 Annotation and Checking Process

Using auto-analyzed sentences with SUW-SC information, five experienced annotators at an annotation company annotated sentences with word boundaries and POS tags, following the SUW-SC criterion and the BCCWJ annotation guidelines (Ogura et al., 2011a,b).[8] After the annotation, the annotators performed (1) unknown word checks to detect erroneous out-of-dictionary words and (2) full-sentence checks to detect any erroneous words, and then fixed annotation errors. Finally, we automatically converted SUW-SC information to SUW information by merging adjacent conjugate word stems and conjugation endings.[9] As a result, we obtained 32,310 sentences with 726k SUW tokens (771k SUW-SC tokens), as shown in Table 1, of which 10,520 sentences included one or more words modified by the annotators.

Through the annotation process, we also found approximately 350 character errors in the original sentences, which may have been caused by, for example, OCR and typographic errors,[10] and replaced them with the correct strings, while retaining the original strings as meta information.

To assess the quality of SUW-SC annotation, the first author randomly sampled and checked 200 annotated sentences comprising 4,928 words. The author found 15 erroneous (multi-) word spans. The F1 scores of the annotators’ annotation were 99.75 (WS), 99.64 (top-level POS), and 99.56 (full POS)[11] when the annotation refined by the author

[5] https://clrd.ninjal.ac.jp/bccwj/en/index.html

[6] The BCCWJ compiles textbooks on ten subjects for elementary, middle, and high schools. The JCMS used sentences from Japanese textbooks for elementary schools.

[7] We did not separate words with irregular conjugations, such as する *suru* ‘do;’ we treated them as single words.

[8] We ignored word attributes, such as the base forms of conjugate words because of the high annotation cost.

[9] If conjugation type and form information are available, SUW annotation can also be converted to SUW-SC annotation using several simple rules. We will publish the conversion script together with the JCMS data.

[10] For example, we found ヌクレチド and フログラム but correct forms were assumed to be ヌクレオチド ‘nucleotide’ and プログラム ‘program.’

[11] This check was done on the manually annotated sentences. This means that the reported F1 scores were not inter-annotator agreement on the auto-analyzed sentences.

System		GEN		SCI Avg.		GOV Avg.	
		Seg	POS	Seg	POS	Seg	POS
MeCab	D_s	**99.6**	99.0	97.9	97.2	98.0	**97.6**
KyTea	D_s	99.1	98.4	98.5	96.8	97.5	96.6
	D_s, D_t	99.1	98.4	98.6	97.1	97.5	96.7
BiLSTM	–	98.7	98.1	98.0	97.2	97.6	96.9
BiLSTM-LF	D_s	99.4	98.8	98.1	97.3	97.9	97.3
	D_s, D_t	99.4	98.8	98.1	97.3	97.9	97.3
BiLSTM-LWP	D_s, D_t, U_t	98.9	98.3	98.9	98.1	97.7	97.1
BERT	–	99.4	**99.1**	**99.3**	**98.7**	**98.1**	**97.6**
BERT-WM	–	99.4	–	**99.3**	–	98.0	–
	U_t	99.4	–	**99.3**	–	98.0	–

Table 2: System performance on the BCCWJ test (GEN), and the JCMS SCI and GOV domain data.

was regarded as the gold standard.

3 Experiments

3.1 Systems and Language Resources

In this section, we report the experimental results for the JCMS data with the SUW annotation. See Appendix F for the results for the SUW-SC data.

We evaluated popular morphological analysis systems and recent neural WS models: MeCab version 0.996[12] (Kudo et al., 2004), KyTea version 0.4.7[13] (Neubig et al., 2011), BiLSTM, BiLSTM with Lexicon Features (LF) (Higashiyama et al., 2020), and BERT (Devlin et al., 2019). Additionally, we evaluated two domain-adaptable neural models proposed for Japanese and Chinese WS: BiLSTM with Lexicon Word Prediction (LWP) (Higashiyama et al., 2020) and BERT with Wordhood Memory (WM)[14] (Tian et al., 2020). We used the off-the-shelf MeCab model based on UniDic,[15] "unidic-cwj-3.1.0" (Den, 2009), and trained the other systems on the corpora and lexicons described later. We used a pretrained Japanese BERT model[16] with character-level tokenization for the BERT-based models. The detailed settings are described in Appendix B.

As source domain labeled data, we split the BCCWJ core data into 51K/6K/3K sentences and used them as training, development, and test data, respectively, for the above systems. As target domain test data, we used all the sentences in each JCMS domain.

For lexicon-enhanced models, we used entries in UniDic as the source domain lexicon D_s and entries in the MeCab-IPADIC user dictionaries for science and technology terms[17] as the target domain lexicon D_t.[18] As target domain unlabeled data for BiLSTM-LWP and BERT-WM, we used 0.98M Japanese sentences in the ASPEC extracted from 20+ domains as single unlabeled data U_t shared for scientific target domains. Using these resources, we trained single domain-adapted model instances for SCI domains. We used no additional resources for the GOV and OTH domains.

3.2 Overall Results

Table 2 shows the WS and POS tagging (top-level POS) F1 scores for each system on the BCCWJ test data (GEN), and the JCMS SCI and GOV domain data; the scores in the SCI and GOV rows are the macro average F1 scores for 22 SCI domains and three GOV domains, respectively. The neural model scores are the mean F1 scores of three runs with random initialization.

For the GEN domain, MeCab, BiLSTM-LF, and BERT-based models achieved high performance: ≥99.4% and ≥98.8% F1 scores for WS and POS tagging, respectively.[19] For the SCI domains, for the two tasks, the systems with only source domain resources (except BERT) had a 0.6–1.8 F1 point degradation from the scores for the GEN domains. Training with target domain resources contributed to robust performance; for example, BiLSTM-LWP achieved a 0.9 F1 point improvement over BiLSTM for each task. BERT achieved the best performance

[12] https://taku910.github.io/mecab/

[13] http://www.phontron.com/kytea/

[14] https://github.com/SVAIGBA/WMSeg

[15] https://clrd.ninjal.ac.jp/unidic/back_number.html

[16] https://huggingface.co/cl-tohoku/bert-base-japanese-char-v2

[17] https://dbarchive.biosciencedbc.jp/en/mecab/download.html

[18] Because the dictionaries included many compound words, we split the original entries into substrings at the positions before and after continuous Japanese characters, continuous Latin characters, continuous Arabic numerals, and each symbol character, as preprocessing.

[19] Notably, the MeCab model was trained on the BCCWJ core data and other corpora (Den, 2009; Oka, 2017), which may have included the GEN test sentences.

Dom.	Unknown Tok/Type Ratio	MeCab D_s		BL-LWP D_s, D_t, U_t		BERT –	
		Seg	POS	Seg	POS	Seg	POS
GEN	2.7 / 16.1	99.6	99.0	98.9	98.3	99.4	99.1
ENE	2.5 / 15.4	99.3	98.9	99.6	99.2	99.7	99.4
TRA	3.0 / 18.2	98.8	98.4	99.4	98.9	99.6	99.2
ENV	3.2 / 15.1	98.8	98.1	99.3	98.7	99.5	99.2
MAN	3.3 / 19.5	98.6	98.2	99.4	99.0	99.6	99.3
CON	3.5 / 19.5	98.9	98.1	99.2	98.6	99.5	99.1
AGR	4.5 / 21.0	98.5	98.0	99.0	98.4	99.4	99.0
THM	4.5 / 24.0	98.4	97.7	99.1	98.3	99.4	98.8
INF	4.7 / 22.6	97.9	97.5	99.1	98.5	99.5	99.1
MEC	5.0 / 25.3	98.4	97.8	99.3	98.7	99.5	99.1
NUC	5.3 / 20.2	98.1	97.3	98.9	98.0	99.4	98.9
CHE-I	5.5 / 23.7	97.9	97.3	99.0	98.3	99.5	99.0
ETH	5.5 / 24.5	98.5	97.8	99.3	98.4	99.4	98.8
MED	5.6 / 27.0	97.1	96.6	99.1	98.6	99.5	99.1
SYS	5.6 / 24.8	98.4	97.7	98.9	98.0	99.4	98.7
ELC	5.8 / 29.4	97.4	97.0	99.0	98.5	99.5	99.1
PAT	6.0 / 26.8	97.0	96.8	99.1	98.7	99.4	99.2
CHE-E	6.1 / 23.7	97.9	97.0	99.0	98.0	99.2	98.7
MIN	6.6 / 22.6	98.0	97.4	98.8	98.1	99.0	98.6
BIO	6.7 / 30.2	96.7	96.0	98.8	98.0	99.3	98.7
PHY	7.5 / 29.6	97.1	96.4	98.5	97.7	99.2	98.8
CHE-B	8.1 / 35.4	97.0	96.1	98.5	97.4	99.1	98.4
EMR	11.2 / 30.2	95.4	91.9	95.6	92.5	97.1	94.0
DIE	0.9 / 7.5	98.0	97.6	97.7	97.0	97.9	97.4
LAW	2.1 / 11.0	97.4	97.0	97.6	97.4	97.9	97.8
PRM	2.8 / 11.1	98.7	98.1	97.7	96.8	98.3	97.8
TBK	4.4 / 19.0	99.0	97.0	97.7	95.5	98.7	96.8
VRS	19.7 / 47.4	87.3	82.3	81.8	75.1	87.1	83.0

Table 3: System performance for each domain. BL represents BiLSTM.

without explicit DA steps and demonstrated the strong effectiveness of PLMs. This may be because the BERT representations were pretrained on Wikipedia text, including articles on scientific topics. BERT-WM did not show salient improvements over BERT, even when we used U_t unlabeled data. For the GOV domains, MeCab and BERT achieved the best WS and POS tagging performance. Most systems achieved lower performance than that for the GEN and SCI domains, which may be because of the high proportions of unknown non-noun tokens, such as verbs, in the GOV domains, as shown in Appendix C.

3.3 Results for Each Domain

Table 3 shows the performance (F1 scores) of the three accurate systems for the GEN and each JCMS domain. For each domain group, the domains are shown in descending order of the unknown token ratio (UTR).[20] The performance of the systems for the two tasks tended to decrease as the UTR increased. However, BiLSTM-LWP and BERT achieved robust performance for SCI domains with higher UTR (scores $\geq$98% and $\geq$99% are shown with the light blue and blue background). As indicated by the high UTR and low system performance, EMR and VRS were two difficult domains.

In Appendices D, E, and G, we present additional experiments for domain-specific enhanced models for EMR and VRS, the evaluation of the differences between the JCMS annotation and original annotation for GOV and OTH, and segmentation examples output by the systems, respectively.

3.4 Discussion

The JCMS comprises well-formed written text from, for example, scientific papers and government documents. Because of this characteristic, systems trained only on source domain resources achieved reasonable performance (WS and POS tagging F1 scores of 96.6–98.5%, on average), and more sophisticated systems enhanced with DA techniques or PLMs, that is, BiLSTM-LWP and BERT, achieved more accurate performance (F1 scores of 97.1–99.3%), as shown in Table 2. Straightforward extensions include the introduction of POS tagging-oriented DA techniques and the integration of DA techniques into PLM-based models.

Furthermore, possible research directions include WS and POS tagging on more challenging text registers, such as speech and social media text on specialized topics. Another important text analysis task is chunking or recognizing multi-word terms because NLP applications in specialized domains can require term-level processing.

4 Related Work

Japanese Morphology Corpora The representative Japanese morphology corpora used in the 1990s and early 2000s include the EDR Japanese Corpus (Miyoshi et al., 1996) and RWCP Text Database (Toyoura et al., 1998), and those used from the 2000s to the present include the Kyoto University Text Corpus (Kurohashi and Nagao, 2003) and BCCWJ (Maekawa et al., 2014). These corpora mainly comprise newspaper articles and other written language text, such as magazines, books, and dictionary example sentences. These corpora have played a significant role in the development of many Japanese morphological analysis and WS systems (Takeuchi and Matsumoto, 1995; Asahara and Matsumoto, 2000; Kudo et al., 2004; Neubig et al., 2011; Tolmachev et al., 2020). Additionally, web corpora (Hashimoto et al., 2011; Hangyo et al., 2012) and transcribed speech corpora (Maekawa, 2003; Koiso et al., 2022) annotated with morphology information have been con-

[20] The unknown token (type) ratio is the percentage of word tokens (types) that did not occurr in the BCCWJ training sentences among all test word tokens (types).

structed and released. Efforts have also been made to construct and publish corpora of other specialized domain text: patent (Mori et al., 2014), shogi (Japanese chess) commentary (Mori et al., 2016), and recipes (Harashima and Hiramatsu, 2020).

Domain Adaptation Methods To improve Japanese morphological analysis and WS performance on target domains, domain-specific or domain-independent adaptation methods have been proposed. Fujita et al. (2014) explored data augmentation techniques to improve morphological analysis performance on picture book text. Kameko et al. (2015) enhanced a WS model for shogi commentary text using shogi game state information. Partially labeled data have been used to fine-tune general WS models to target domains; Tsuboi et al. (2008) adapted a CRF model to a medical domain and Neubig et al. (2011) adapted a pointwise prediction model to a web domain. Higashiyama et al. (2020) enhanced a BiLSTM-based WS model by introducing an auxiliary word prediction task and adapted the model to several Japanese and Chinese target domains.

5 Conclusion

We presented the JCMS, which is a Japanese corpus of 27 specialized domains annotated with word boundaries and POS tags. The experiments on the corpus demonstrated the robust WS and POS tagging performance of recent neural models on many out-of-domain datasets. Our corpus could be a useful benchmark for developing and evaluating cross-domain systems for WS and POS tagging.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. We used the Asian Scientific Paper Excerpt Corpus, NITCIR-9 PatentMT test collection, NTCIR-11 MedNLP-2 test collection, and Balanced Corpus of Contemporary Written Japanese to construct our corpus.

References

Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. 2014. Overview of the NTCIR-11 MedNLP-2 task. In *Proceedings of the 11th NTCIR Conference*, pages 147–155, Tokyo, Japan.

Masayuki Asahara and Yuji Matsumoto. 2000. Extended models and tools for high-performance part-of-speech. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Yasuharu Den. 2009. A multi-purpose electronic dictionary for morphological analyzers [in Japanese]. *Journal of the Japanese Society for Artificial Inteligence*, 34(5):640–646.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

EDR. 2001. EDR denshika jisho shiyō setsumēsho (The EDR electronic dictionary specification manual) [in Japanese].

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehiro Utsuro, Terumasa Ehara, and Sayori Shimohata. 2010. Overview of the patent translation task at the NTCIR-8 workshop. In *Proceedings of the 8th NTCIR Conference*, pages 371–376, Tokyo, Japan.

Sanae Fujita, Hirotoshi Taira, Tessei Kobayashi, and Takaaki Tanaka. 2014. Japanese morphological analysis of picture books. *Journal of Natural Language Processing*, 21(3):515–539.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 535–544, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.

Jun Harashima and Makoto Hiramatsu. 2020. Cookpad parsed corpus: Linguistic annotations of Japanese recipes. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 87–92, Barcelona, Spain. Association for Computational Linguistics.

Chikara Hashimoto, Sadao Kurohashi, Daisuke Kawahara, Keiji Shinzato, and Masaaki Nagata. 2011. Construction of a blog corpus with syntactic, anaphoric, and sentiment annotations. *Journal of Natural Language Processing*, 18(2):175–201.

Shohei Higashiyama, Masao Utiyama, Yuji Matsumoto, Taro Watanabe, and Eiichiro Sumita. 2020. Auxiliary lexicon word prediction for cross-domain word segmentation. *Journal of Natural Language Processing*, 27(3):573–598.

Hirotaka Kameko, Shinsuke Mori, and Yoshimasa Tsuruoka. 2015. Can symbol grounding improve low-level NLP? word segmentation as a case study. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2298–2303, Lisbon, Portugal. Association for Computational Linguistics.

Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. 2022. Design and evaluation of the corpus of everyday Japanese conversation. In *Proceedings of the Language Resources and Evaluation Conference*, pages 5587–5594, Marseille, France. European Language Resources Association.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Sadao Kurohashi and Makoto Nagao. 2003. Building a Japanese parsed corpus. In *Treebanks*, pages 249–260. Springer.

Kikuo Maekawa. 2003. Corpus of spontaneous Japanese: Its design and evaluation. In *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Hideo Miyoshi, Kenji Sugiyama, Masahiro Kobayashi, and Takano Ogino. 1996. An overview of the EDR electronic dictionary and the current status of its utilization. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Shinsuke Mori, Hideki Ogura, and Tetsuro Sasada. 2014. A Japanese word dependency corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 753–758, Reykjavik, Iceland. European Language Resources Association.

Shinsuke Mori, John Richardson, Atsushi Ushiku, Tetsuro Sasada, Hirotaka Kameko, and Yoshimasa Tsuruoka. 2016. A Japanese chess commentary corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1415–1420, Portorož, Slovenia. European Language Resources Association.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533, Portland, Oregon, USA. Association for Computational Linguistics.

Toshinobu Ogiso, Mamoru Komachi, and Yuji Matsumoto. 2013. Morphological analysis of historical Japanese text [in Japanese]. *Journal of Natural Language Processing*, 20(5):727–748.

Hideki Ogura, Hanae Koiso, Yumi Fujiike, Sayaka Miyauchi, Hikari Konishi, and Yutaka Hara. 2011a. Gendai nihongo kakikotoba kinkō corpus kētairon kitēshū dai 4 ban ge (Regulations of morphological information for balanced corpus of contemporary written Japanese 4th edition volume 2) [in Japanese]. *NINJAL Internal Reports*.

Hideki Ogura, Hanae Koiso, Yumi Fujiike, Sayaka Miyauchi, Hikari Konishi, and Yutaka Hara. 2011b. Gendai nihongo kakikotoba kinkō corpus kētairon kitēshū dai 4 ban jō (Regulations of morphological information for balanced corpus of contemporary written Japanese 4th edition volume 1) [in Japanese]. *NINJAL Internal Reports*.

Teruaki Oka. 2017. UniDic for morphological analysis with reduced model size by review of CRF feature templates [in Japanese]. In *Language Resources Workshop*, pages 144–153. National Institute for Japanese Language and Linguistics.

Teruaki Oka, Yuichi Ishimoto, Yutaka Yagi, Takenori Nakamura, Masayuki Asahara, Kikuo Maekawa, Toshinobu Ogiso, Hanae Koiso, Kumiko Sakoda, and Nobuko Kibe. 2020. KOTONOHA: A corpus concordance system for skewer-searching NINJAL corpora. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7077–7083.

Katsuhito Sudoh, Masaaki Nagata, Shinsuke Mori, and Tatsuya Kawahara. 2014. Japanese-to-English patent translation system based on domain-adapted word segmentation and post-ordering. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 234–248, Vancouver, Canada. Association for Machine Translation in the Americas.

Koichi Takeuchi and Yuji Matsumoto. 1995. HMM parameter learning for Japanese morphological analyzer. In *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*, pages 163–172, Hong Kong. City University of Hong Kong.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020. Improving Chinese word segmentation with wordhood memory networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285, Online. Association for Computational Linguistics.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2020. Design and structure of the Juman++

morphological analyzer toolkit. *Journal of Natural Language Processing*, 27(1):89–132.

Jun Toyoura, Hitoshi Isahara, Shiho Ogino, Wakako Kuwahata, Hironobu Takahashi, Takenobu Tokunaga, Koichi Hashida, Minako Hashimoto, and Fumio Motoyoshi. 1998. RWCP niokeru kenkyūyō text database no kaihatsu (Development of the text database for research at RWCP) [in Japanese]. pages 454–455.

Yuta Tsuboi, Hisashi Kashima, Shinsuke Mori, Hiroki Oda, and Yuji Matsumoto. 2008. Training conditional random fields using incomplete annotations. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 897–904, Manchester, UK. Coling 2008 Organizing Committee.

	SUW		SUW-SC		
	POS tag	Example	POS tag (stem)	POS tag (ending)	Example
V1	動詞-一般	ある	動詞-語幹-一般	活用語尾-動詞型	あ|る
V2	動詞-非自立可能	すぎる	動詞-語幹-非自立可能	活用語尾-動詞型	すぎ|る
V3	動詞-一般	有し	動詞-特殊型-一般	–	有し
V4	動詞-非自立可能	する	動詞-特殊型-非自立可能	–	する
A1	形容詞-一般	高い	形容詞-語幹-一般	活用語尾-形容詞型	高|い
A2	形容詞-非自立可能	欲しい	形容詞-語幹-非自立可能	活用語尾-形容詞型	欲し|い
A3	形容詞-非自立可能	ねえ	形容詞-特殊型	–	ねえ
S1	接尾辞-動詞的	(悪)ぶる	接尾辞-動詞型語幹	活用語尾-動詞型	(悪)ぶ|る
S2	接尾辞-形容詞的	っぽい	接尾辞-形容詞型語幹	活用語尾-形容詞型	っぽ|い
AV1	助動詞	させる	助動詞-動詞型語幹	活用語尾-動詞型	させ|る
AV2	助動詞	(行か)ない	助動詞-形容詞型語幹	活用語尾-形容詞型	(行か)な|い
AV3	助動詞	だろう	助動詞-特殊型	–	だろう

Table 4: POS tags and example words of the SUW and SUW-SC criteria

A SUW-SC POS Tags

Table 4 shows the SUW-SC POS tags that differ from the SUW POS tags. Characters in "()" indicate the preceding context and the symbol "|" presents a word boundary.

B Details for the Evaluated Systems

We used the default hyperparameters of KyTea. We used similar model architectures, hyperparameters, and training settings to Higashiyama et al. (2020) for BiLSTM, BiLSTM-LF, and BiLSTM-LWP, except we introduced an additional multi-layer perceptron with one hidden layer (300 hidden units) for POS tagging for each model. We used Tian et al. (2020)'s code for BERT and BERT-WM models with their hyperparameters and training settings for the MSR data, except we used softmax inference similarly to BiLSTM-based models and decreased the mini-batch size to 4 or 8 because of the memory limitation. The BERT model predicted joint segmentation and POS tags, such as B-名詞 (noun), using a single inference layer.

C POS Proportions of Unknown Tokens

Figure 1 shows the proportions of POS tags of unknown tokens for each domain in the JCMS SUW data. Nouns accounted for 95–99% of all unknown tokens for the SCI (AGR to PAT) domains, whereas non-noun tokens, such as verbs and symbols, accounted for 15–60% for the GOV and OTH domains.

D Performance of domain-specific models

The VRS data consisted of Japanese verse sentences written in historical literary styles. The EMR data consisted of medical history summaries of imaginary patients. We additionally evaluated two domain-specific models for the VRS and EMR domains of the SUW data. One is the off-the-shelf MeCab model with the morphological analysis dictionary for historical literary style text: "UniDic-202203_65_novel" D_h (Ogiso et al., 2013). The other is a BiLSTM-LWP model trained with medical domain-specific lexicon D_m and unlabeled data U_m, which we describe later. As shown in Table 5, the improved performance of the MeCab model on the VRS domain indicates the alleviation of domain mismatch. The BiLSTM-LWP model adapted for the EMR domain achieved 1.2–1.3 F1 point improvement for WS and POS tagging over the model adapted for all scientific domains, and achieved competitive scores to BERT.

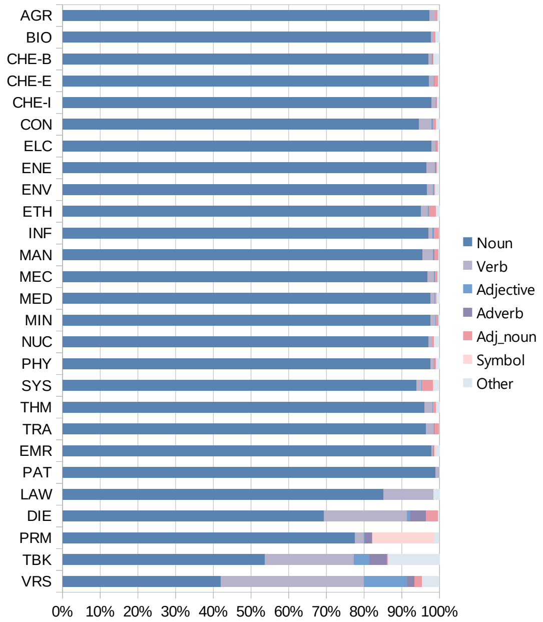

Figure 1: POS Proportions of Unknown Tokens in the SUW data

Domain	MeCab D_h Seg	 POS	BiLSTM-LWP D_s, D_m, U_m Seg	 POS
EMR	–	–	96.9	93.7
VRS	94.1	91.3	–	–

Table 5: Performance of domain-specific models for the EMR or VRS domain of the SUW data

Domain		F1 Seg	 POS	 FPOS	FP
GOV	DIE	98.2	98.1	97.9	544
	LAW	98.3	98.3	98.1	501
	PRM	98.6	98.1	96.8	637
OTH	TBK	99.7	99.6	99.3	100
	VRS	95.3	92.9	91.7	1,380

Table 6: Accuracy of original annotation in the BCCWJ non-core data evaluated on the JCMS SUW data

Regarding the resources for the EMR domain, we preprocessed and merged five medical dictionaries into a single lexicon D_m: MEDIS hyojun byomei master,[21] J-GLOBAL Mesh,[22] ComeJisyo,[23] Manbyo dictionary,[24] and Hyakuyaku dictionary.[25] We merged 400K sentences from the ASPEC medical domain and 137K sentences from the MedTxt[26] case report and radiography report corpus into a single unlabeled dataset U_m.

E Accuracy of the Original BCCWJ annotation

The original annotation of the BCCWJ non-core data was performed semi-automatically; hence, the average annotation accuracy was 98%.[27] We regarded the original annotation of the GOV and OTH domain data as system prediction and evaluated it using the SUW annotated sentences in the JCMS as the gold standard. Table 6 shows the WS and POS tagging (top-level POS as "POS" and full POS as "FPOS") F1 scores and the numbers of false positives (FP) based on the FPOS errors. All domain data contained annotation errors, which corresponded to 100–1380 FPs; however, the original annotation achieved higher F1 scores than the evaluated systems in §3.3 because of manual correction efforts by NINJAL.

Dom.	Unknown Tok/Type Ratio	MeCab D_s Seg	 POS	BL-LWP D_s, D_t, U_t Seg	 POS	BERT – Seg	 POS
GEN	3.7 / 21.1	99.6	99.1	98.8	98.3	99.3	99.1
SCI Avg.		98.0	97.3	98.9	98.2	99.3	98.8
GOV Avg.		98.0	97.6	97.5	97.0	98.0	97.7
ENE	3.1 / 18.1	99.3	98.9	99.5	99.2	99.7	99.4
TRA	3.6 / 20.9	98.8	98.4	99.4	98.9	99.6	99.2
ENV	3.8 / 17.4	98.8	98.2	99.3	98.8	99.6	99.3
MAN	3.9 / 22.0	98.6	98.3	99.4	99.0	99.6	99.3
CON	4.0 / 22.2	98.9	98.2	99.3	98.7	99.5	99.1
THM	4.9 / 26.7	98.4	97.8	99.1	98.4	99.4	98.9
AGR	5.1 / 23.5	98.5	98.1	99.0	98.5	99.4	99.1
INF	5.1 / 25.2	98.0	97.6	99.1	98.6	99.5	99.1
MEC	5.5 / 27.8	98.4	97.9	99.2	98.7	99.5	99.2
NUC	5.7 / 22.6	98.2	97.5	98.9	98.1	99.4	99.0
CHE-I	5.9 / 26.2	98.0	97.4	99.0	98.4	99.5	99.2
ETH	6.0 / 27.1	98.6	97.9	99.4	98.5	99.4	98.9
MED	6.0 / 29.3	97.2	96.8	99.1	98.6	99.6	99.2
SYS	6.1 / 27.7	98.4	97.8	98.9	98.1	99.4	98.8
ELC	6.2 / 31.8	97.5	97.1	99.0	98.5	99.5	99.1
PAT	6.4 / 29.9	97.1	96.9	99.0	98.6	99.4	99.3
CHE-E	6.5 /26.5	97.9	97.1	98.9	98.1	99.3	98.8
MIN	6.9 / 24.9	98.0	97.5	98.8	98.1	99.1	98.7
BIO	7.2 / 32.6	96.8	96.2	98.8	98.1	99.3	98.8
PHY	8.0 / 32.2	97.2	96.6	98.7	97.9	99.2	98.8
CHE-B	8.6 / 38.2	97.1	96.3	98.6	97.5	99.2	98.6
EMR	11.1 / 32.4	95.5	92.1	95.9	92.5	97.3	94.3
LAW	2.7 / 12.4	97.4	97.0	97.6	97.3	98.1	97.9
DIE	3.4 / 12.0	98.1	97.8	97.7	97.1	98.0	97.5
PRM	3.7 / 14.3	98.5	97.9	97.3	96.6	98.1	97.7
TBK	5.5 / 23.6	98.9	97.2	97.6	95.7	98.6	97.0
VRS	18.1 / 47.6	88.6	81.4	80.0	72.9	85.0	81.1

Table 7: Performance of the three systems on the JCMS SUW-SC data. BL represents BiLSTM.

F Results for the SUW-SC POS Tag Set

Table 7 shows the performance of the three systems trained and evaluated on the SUW-SC annotation data. For MeCab, we applied the conversion rules mentioned in §2.3 to SUW results and obtained SUW-SC results. For BiLSTM-LWP and BERT, we trained new model instances with SUW-SC training data. Similar to the results of the SUW experiments, we observed that system performance tended to decrease as the UTR increased.

G Segmentation Examples

Table 8 shows the gold standard annotation and segmentation results of several JCMS sentence fragments[28] output by three systems: MeCab, BiLSTM-LWP, and BERT. Incorrect segmentation (including incorrect manual annotation) is highlighted in the gray background. System errors include oversegmentation of Latin characters (a–c), oversegmentation of English loanwords written

[21] http://www2.medis.or.jp/stdcd/byomei/index.html

[22] https://dbarchive.biosciencedbc.jp/en/mecab/data-2.html

[23] https://ja.osdn.net/projects/comedic/

[24] https://sociocom.naist.jp/manbyou-dic/

[25] https://sociocom.naist.jp/hyakuyaku-dic/

[26] https://sociocom.naist.jp/medtxt/

[27] https://clrd.ninjal.ac.jp/bccwj/doc/manual/BCCWJ_Manual_01.pdf

[28] The Japanese writing system uses multiple script types, including *kanji* (e.g., '漢字'), *hiragana* (e.g., 'ひらがな'), *katanaka* (e.g., 'カタカナ'), Arabic numerals (e.g., '012' or '０１２'), Latin characters (e.g., 'ABC' or 'ＡＢＣ'), and punctuation and auxiliary symbols.

<table>
<tr><th></th><th>Domain</th><th>Gold</th><th>MeCab</th><th>BiLSTM-LWP</th><th>BERT</th></tr>
<tr><td>(a)</td><td>PHY</td><td>ＮａＣｌ (型)</td><td>Ｎａ|Ｃｌ</td><td>Ｎａ|Ｃｌ</td><td>ＮａＣｌ</td></tr>
<tr><td>(b)</td><td>INF</td><td>Ｂｌｕｅｔｏｏｔｈ</td><td>Ｂｌｕｅ|ｔｏｏｔｈ</td><td>Ｂｌｕｅ|ｔｏｏｔｈ</td><td>Ｂｌｕｅｔｏｏｔｈ</td></tr>
<tr><td>(c)</td><td>BIO</td><td>ＨＥＶ (の感染)</td><td>Ｈ|Ｅ|Ｖ</td><td>ＨＥＶ</td><td>ＨＥＶ</td></tr>
<tr><td>(d)</td><td>INF</td><td>サブルーチン (の効率)</td><td>サブルーチン</td><td>サブ|ルーチン</td><td>サブルーチン</td></tr>
<tr><td>(e)</td><td>INF</td><td>(ＴＣＰ) スループット</td><td>スルー|プット</td><td>スループット</td><td>スループット</td></tr>
<tr><td>(f)</td><td>CHE-B</td><td>クロマトグラフィー</td><td>クロマトグラフィー</td><td>クロマト|グラフィー</td><td>クロマト|グラフィー</td></tr>
<tr><td>(g)</td><td>LAW</td><td>(関係) 市町村長</td><td>市町村長</td><td>市|町村長</td><td>市|町村長</td></tr>
<tr><td>(h)</td><td>PHY</td><td>(Ｂ) 中間|子 (物理)</td><td>中間|子</td><td>中間子</td><td>中間|子</td></tr>
<tr><td>(i)</td><td>PHY</td><td>希|土類|金属</td><td>希|土類|金属</td><td>希土|類|金属</td><td>希土|類|金属</td></tr>
<tr><td>(j)</td><td>LAW</td><td>ただし書 (又は)</td><td>ただし書</td><td>ただし|書</td><td>ただし書</td></tr>
<tr><td>(k)</td><td>PHY</td><td>撹はん (する)</td><td>撹|はん</td><td>撹|はん</td><td>撹はん</td></tr>
<tr><td>(l)</td><td>PHY</td><td>り患 (年数)</td><td>り患</td><td>り患</td><td>り|患</td></tr>
<tr><td>(m)</td><td>PHY</td><td>(パルス) 静電|場</td><td>静電|場</td><td>静|電場</td><td>静電|場</td></tr>
<tr><td>(n)</td><td>EMR</td><td>右下|腹部|痛</td><td>右下|腹部|痛</td><td>右|下腹|部|痛</td><td>右下|腹部|痛</td></tr>
<tr><td>(o)</td><td>EMR</td><td>両下|肢</td><td>両|下肢</td><td>両|下肢</td><td>両|下肢</td></tr>
</table>

Table 8: Segmentation results of the JCMS sentence examples using the three systems. Characters in "()" indicate the surrounding context. The meanings of the examples are as follows: (a) 'NaCl (-type),' (b) 'Bluetooth,' (c) 'HEV (infection),' (d) '(efficiency of) the subroutine,' (e) '(TCP) throughput,' (f) 'chromatography,' (g) '(the relevant) municipal mayors,' (h) 'B-meson physics,' (i) 'rare earth metal,' (j) 'proviso (or),' (k) 'stir,' (l) '(duration years of) the disorder,' (m) '(pulse) electrostatic field,' (n) 'right lower quadrant pain,' and (o) 'both lower extremities.'

with katakana (often into English morphemes) (d–f), incorrect segmentation of kanji sequences (g–i), and incorrect segmentation of hiragana and kanji mixed sequences (j–l). We found words that were correctly segmented by the systems but were evaluated as errors because of the annotation errors (m–o).