| | 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.

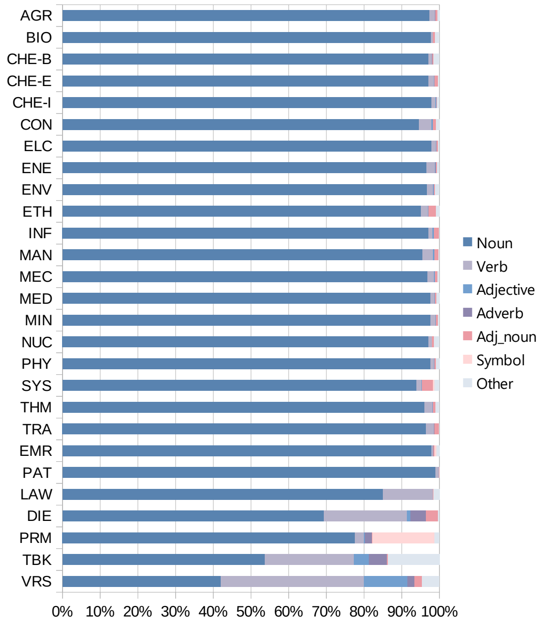

Figure 1: POS Proportions of Unknown Tokens in the SUW data

## 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 MA 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.

| Domain | MeCab | | BL-LWP | |
|---|---|---|---|---|
| | $D_h$ | | $D_s, D_m, U_m$ | |
| | Seg | POS | 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 | | | FP |
|---|---|---|---|---|---|
| | | Seg | POS | FPOS | |
| 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,[11] J-GLOBAL Mesh,[12] ComeJisyo,[13] Manbyo dictionary,[14] and Hyakuyaku dictionary.[15] We merged 400K sentences from the ASPEC medical domain and 137K sentences from the MedTxt[16] 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%.[17] 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 | | BL-LWP | | BERT | |
|---|---|---|---|---|---|---|---|
| | | $D_s$ | | $D_s, D_t, U_t$ | | – | |
| | | Seg | POS | Seg | POS | 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

## 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. 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[18] 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 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

---

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

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

[13] `https://ja.osdn.net/projects/comedic/`

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

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

[16] `https://sociocom.naist.jp/medtxt/`

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

[18] 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.

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

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.'

correctly segmented by the systems but were evaluated as errors because of the annotation errors (m–o).