Using Language Models to Improve Rule-based Linguistic Annotation of Modern Historical Japanese Corpora

Jerry Bonnell and **Mitsunori Ogihara**
University of Miami
Coral Gables FL 33124, USA
{j.bonnell,m.ogihara}@miami.edu

Abstract

Annotation of unlabeled textual corpora with linguistic metadata is a fundamental technology in many scholarly workflows in the digital humanities (DH). Pretrained natural language processing pipelines offer tokenization, tagging, and dependency parsing of raw text simultaneously using an annotation scheme like Universal Dependencies (UD). However, the accuracy of these UD tools remains unknown for historical texts and current methods lack mechanisms that enable helpful evaluations by domain experts. To address both points for the case of Modern Historical Japanese text, this paper proposes the use of unsupervised domain adaptation methods to develop a domain-adapted language model (LM) that can flag instances of inaccurate UD output from a pretrained LM and the use of these instances to form rules that, when applied, improves pretrained annotation accuracy. To test the efficacy of the proposed approach, the paper evaluates the domain-adapted LM against three baselines that are not adapted to the historical domain. The experiments conducted demonstrate that the domain-adapted LM improves UD annotation in the Modern Historical Japanese domain and that rules produced using this LM are best indicative of characteristics of the domain in terms of out-of-vocabulary rate and candidate normalized form discovery for "difficult" bigram terms.

1 Introduction

Annotating unlabeled corpora with linguistic metadata is a fundamental task in many scholarly workflows in the digital humanities (DH) (Aurnhammer et al., 2019; Kirschenbaum, 2007). These can benefit from the application of "off-the-shelf", or pretrained, natural language processing (NLP) pipelines that supply tokenization, tagging, and dependency parsing simultaneously using an annotation scheme like Universal Dependencies (UD) (Nivre et al., 2020). For these tools to warrant any integration, the annotations produced must be accurate for the target corpus under study and enable helpful evaluation by domain experts. Current efforts have prioritized the former as the accuracy of pretrained UD tools remains unknown for historical texts sampled from domains that are different from pretraining domains (Suissa et al., 2022).

In applications to historical text in East Asian languages like Chinese and Japanese that can be written without specifying word boundaries, substantial errors in word segmentation can result in degraded accuracy of resulting dependency parsings (Yasuoka, 2020). The insufficiency of the vocabulary used by pretrained UD tools can be a major source of errors in word segmentation. Consequently, a straightforward application is likely not possible without extensive manual revision by domain experts and the process can be prohibitive when corpus size is large (Shirai et al., 2020).

Recent work in NLP has addressed the accuracy of sequence labeling tasks (e.g., word segmentation, part of speech annotation, named entity recognition) for historical materials through unsupervised finetuning of pretrained contextualized word embeddings using transformer-based language models (LMs) (Han and Eisenstein, 2019; Manjavacas and Fonteyn, 2022). These finetuned LMs are capable of capturing useful high-level features about the domain without requiring any labels from the target corpus. The improvements are encouraging, yet, the usefulness of a transformer-based method for a domain expert remains unknown. The lack of supervision in these methods means that direct mechanisms for finetuning the LM other than through the training data used are beyond the control of a domain expert. Put another way, the explicit representations of domain knowledge that occur during manual revision of pretrained output is not possible with current NLP for adapting pretrained LMs to historical corpora.

In prior work, Bonnell and Ogihara (2022) pro-

Proceedings of the 6th LaTeCH-CLfL Workshop, pages 30–39

posed a rule-based expert system for the case of Modern Historical Japanese corpora that generates accurate UD annotations using a set of handcrafted rules. The rules compose a two-parted workflow that (1) *normalizes* non-standard lexical variants to a more canonical form so that the normalized text can receive a more accurate parsing by a pretrained tool, and (2) an *assignment* step where the updated UD is then linked to word forms from the original text. Figure 1 shows an overview. The use of rules allows for immediate improvement in UD annotation for out-of-vocabulary terms by parsing dependencies using substituted terms that are present in the vocabulary, and then restoring the historical forms in the parsed sentence. Moreover, each rule forms a direct application of domain knowledge and supports human comprehension. However, because rule generation cannot proceed without manual review, the workflow can be a time intensive undertaking when the number of rules needed to achieve improved accuracy is not known.

This paper aims to enrich the rule-based expert system by incorporating it into a workflow that is usefully guided by a pretrained LM adapted to the domain of Modern Historical Japanese. The contributions of the proposed workflow are as follows:

- Use of domain adaptation methods to train a LM that can flag instances of incorrect UD output from a pretrained tool not adapted to the historical domain.
- Automatic generation of a rule set from flagged differences in UD output using a domain-adapted LM trained with the masked language modeling objective. The developed rule set can be directly used to enhance a rule-based system for linguistic UD annotation.
- The domain-adapted LM improves pretrained UD annotation in the historical domain and rules suggested using this LM are best indicative of unique characteristics of the domain when compared to baseline methods, in terms of out-of-vocabulary rate and candidate rule discovery for "difficult" bigram terms.

The proposed method offers improved UD accuracy for Modern Historical Japanese corpora while also enabling useful evaluations by a domain expert. This can serve as welcome news to DH scholars who would like to make more frequent use of transformer-based NLP methods in their scholarship.

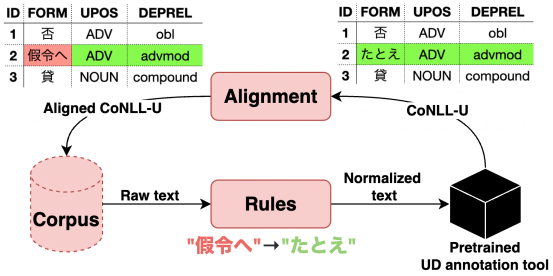

Figure 1: Overview of the rule-based expert system. In this example, the rule '假令へ'→ 'たとえ' is applied to the sentence '否假令へ貸倒れとならずとするも、'. After string replacement of the left-hand side with the right-hand side, the normalized sentence is submitted to a pretrained UD tool for annotation. The updated UD returned is then aligned with historical word forms (i.e., '假令へ') from the source text.

2 Related Work

Gururangan et al. (2020) has shown that continued pretraining of large pretrained language models using unlabeled data from the task domain yields improvements in task performance for both small and large corpus sizes. The literature is rich with proposals applying pretraining strategies in different domains (Gururangan et al., 2020; Desai et al., 2020). These methods have also been successfully applied in the case of historical texts (Manjavacas and Fonteyn, 2022; Han and Eisenstein, 2019).

Universal Dependencies (UD) is a community effort for cross-linguistic annotation of grammar (parts of speech, morphological information, and syntactic dependencies) (Nivre et al., 2020). Recent methods have been put forward for generating UD annotations using transformer-based language models, e.g., BERT and ELECTRA, as a backbone (Kondratyuk and Straka, 2019). These are also available for generating UD from Japanese text input. (Yasuoka, 2022; Matsuda et al., 2019). However, the transfer of adaptive BERT models when using these methods for the case of historical text – and specifically historical Japanese text – remains to be evaluated thoroughly.

The need for helpful evaluations of NLP-based tools is gaining traction in the DH community (McGillivray et al., 2020; Suissa et al., 2022). In the case of historical Japanese, Shirai et al. (2020) trains a combination of UDPipe and CRF++ for sequence labeling using training data made available through the labor-intensive corrections done by domain experts. To the best of our knowledge, this paper is the first to address the applicability of

already pretrained tools through domain adaptation methods and the use of a pretrained neural architecture as a mechanism for enhancing an expert system that generates UD annotations for the case of Modern Historical Japanese corpora.

3 Method

3.1 Data and Model Selection

We adopt the Taiyo (太陽) magazine as a historical corpus of written Japanese published by Hakubunkan and maintained by the National Institute for Japanese Language and Linguistics (NINJAL) as part of its Corpus of Modern Japanese. Taiyo was the best-selling general interest magazine during the Meiji (明治) and Taisho (大正) periods (1895-1925) and contains 3400 documents written by 1000 different writers. The magazine saw significant changes in literary and colloquial writing during this period where both styles can coexist within the same article (Maekawa, 2006). Taiyo is made available without any linguistic metadata and, therefore, obtaining ground truth UD annotations for the corpus is not possible.[1]

For application of domain adaptation methods, we reference the Balanced Corpus of Contemporary Japanese (BCCWJ), a large corpus of contemporary written Japanese and Japan's first 100 million words balanced corpus (Maekawa et al., 2014).[2] UD-Japanese-BCCWJ r2.8 is a UD resource curated from BCCWJ and contains UD annotations from the core (edited) portion of BCCWJ (1980 documents; 57K sentences) (Asahara et al., 2018). To augment the labeled data available, we incorporate UD-Japanese-GSD, another UD Japanese resource from Google Universal Dependency Treebanks v2.0 (Asahara et al., 2018). Finally, for evaluation of our methods against the Modern Historical Japanese domain, we appeal to the UD-Japanese-Modern treebank, a small UD annotation corpus based on samples from the Meiroku Zasshi corpus in NINJAL's Corpus of Modern Japanese where Taiyo is also sourced from (822 labeled sentences) (Asahara et al., 2018).[3]

Figure 2: Heatmap matrix showing similarity between data sampled from target and pretraining domains. Similarity is defined as the percentage overlap in the top 10K words gathered from each sample. Texts are tokenized with MeCab initialized using UniDic for Modern Literary Japanese (Ogiso et al., 2013). '1895', '1901', '1917', and '1925' refer to publication year of texts sourced from Taiyo. The other samples compared with here: 'meiroku' (from Meiroku Zasshi), 'google' (from UD Japanese GSD), 'bccwj' (from UD Japanese BCCWJ), and 'jawiki' (from Japanese Wikipedia).

Pretrained models that generate Japanese UD annotations are selected based on the pretraining data used and if a significant difference in the vocabulary is observed between the pretraining domain and the Modern Historical Japanese domain.[4] This is also done in expectation of workflows in production where only a pretrained tool trained strictly on contemporary text is available. These models: (1) GiNZA (5.1.0), an ELECTRA-based model pretrained on large web crawl of Japanese text and UD Japanese BCCWJ r2.8 (Matsuda et al., 2019), and (2) esupar (1.1.5), a BERT-based model for UPOS prediction that also trains a BiLSTM with deep biaffine attention for dependency parsing (Yasuoka, 2022). esupar can be initialized using different models and we select KoichiYasuoka/bert-base-japanese-char-extended that is pretrained on Japanese Wikipedia. Figure 2 shows a heatmap quantifying vocabulary similarity in the selected domains by comparing overlap in the top 10K words from samples collected in each domain. Low similarity observed between Taiyo and pretraining do-

[1] Due to the presence of copyrighted text, the Taiyo corpus is not publicly available and obtaining the entire contents is possible only through DVD (CD-ROM) purchase through NINJAL.

[2] The OW, OB, OM, and OL registers have the longest coverage in the corpus, spanning 30 years from 1976-2005. DVD (CD-ROM) purchase is also required for access due to copyrighted articles.

[3] More specifically, the Meiroku Zasshin samples in the UD-Japanese-Modern resource are sourced from *CHJ Meiji / Taishō Era Series I: Magazines.*

[4] For the case of Japanese, several pretrained BERT models exist where the pretraining data used is similar to the historical domain where Taiyo is sourced form. However, for the purposes of the research questions forwarded here, we do not include said models here.

Corpus	UD Labels?	Domain Tuning	Task-specific Training	Rule Set Generation
Taiyo		✓		✓
UD-Japanese-BCCWJ	✓	✓	✓	
UD-Japanese-GSD	✓	✓	✓	
UD-Japanese-Modern	✓			

Table 1: Overview of the corpora used at each phase in the workflow.

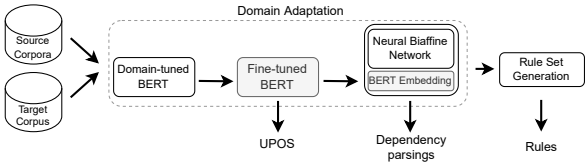

Figure 3: Component diagram showing different steps and outputs in the workflow.

mains offer credence to domain adaptation methods for this kind of data.

3.2 Workflow

We develop a domain-adapted LM that can predict accurate UD in the Modern Historical Japanese domain by employing pretraining methods that can enable transfer between disparate domains. We base our approach on the training steps proposed in Han and Eisenstein (2019) and treat Taiyo as a *target* corpus of unlabeled data, and UD-Japanese-BCCWJ and UD-Japanese-GSD as labeled *source* corpora. Then, following Yasuoka (2022), we train a combination of BERT and a BiLSTM-based neural biaffine network for UPOS and semantic dependency parsing prediction, respectively. The resulting domain-adapted LM is subsequently used for generating a rule set that brings improved pretrained UD annotation in the historical domain. Figure 3 presents an overview of the workflow and Table 1 shows the role of the corpora along each phase. Our code for realizing these steps are publicly available.[5]

3.2.1 Domain tuning

In this phase we fine-tune the BERT contextualized embeddings through continued pretraining on the dynamic masked language modeling (MLM) objective in the target domain. Ten random maskings are generated for each sample and, in each masking, 15% of the tokens are randomly masked as per Han and Eisenstein (2019). Three epochs are done over the masked data. We use all unlabeled training samples from the Taiyo corpus and add unlabeled samples from the training splits in UD-Japanese-BCCWJ and UD-Japanese-GSD. The maximum sequence length is set to 50 and we segment the text into chunks of this size; we find empirically that using smaller segments for this data helps drive down the training loss when compared to larger sequence lengths.

[5] https://github.com/jerrybonnell/adapt-esupar

3.2.2 Task-specific training

We develop two models for simultaneous prediction of UPOS and dependency parsings using the domain-tuned BERT. To learn the former we fine-tune the contextualized embeddings on the UPOS labeling objective using only labeled samples from the training splits in the source corpora; no labels from the target corpus are used during learning of the prediction model.

For learning the latter, the contextualized embeddings from the fine-tuned LM are then used as an embedding layer in a BiLSTM-based neural biaffine network for learning dependency parsings. The BiLSTM model is also trained using only labeled dependencies from the source corpora. An implicit assumption of this step is that the additional transfer of the fine-tuned contextualized embeddings to the BiLSTM-based model is useful for generating improved UD annotations in the target domain.

3.2.3 Generating a rule set using the domain-adapted LM

A domain-adapted LM that can produce accurate UD in the Modern Historical Japanese domain can be used to flag instances of inaccurate UD output generated by a pretrained LM. These flagged discrepancies provide critical regions where potential rules can be generated. Consistent with Bonnell and Ogihara (2022), we restrict our attention to differences only in the FORM field and define a rule as a mapping from a historical word form to a normalized usage such that, after rule application and submission of the normalized sentence to a pretrained system, the output FORM, UPOS, and DEPREL fields of the domain-adapted and pretrained LM become identical. Figure 4 provides an overview of this module.

We focus specifically on prediction of two-character bigrams consisting of at least one kanji character that are misclassified by a pretrained LM. Instances of this form are raised using sentences from the testing split of the Taiyo corpus. The mis-

Masked Contexts	Masked Predictions	Rules
世に立つ上に於ては何處までも**[MASK]**の點を避ける様に努めねばならぬと思ふ。	そ, こ, 此, 別, ...	之の → その*, 之の → この*, 之の → 此の, 之の → 別の, ...
世に立つ上に於ては何處までも之**[MASK]**點を避ける様に努めねばならぬと思ふ。	丶, 是, は, 二, ...	之の → 之丶*, 之の → 之是, 之の → 之は, 之の → 之二, ...

Table 2: Example flow showing candidate normalized form discovery for the flagged bigram '之の' in the context sentence '世に立つ上に於ては何處までも之の點を避ける様に努めねばならぬと思ふ。" where predicted UPOS and DEPREL for '之の' is "DET" and "det", respectively, according to ADAPT-ESUPAR. Normalized forms are candidates when, after substitution in the respective context sentences, pretrained annotation consistently aligns with the FORM, UPOS, and DEPREL fields given by ADAPT-ESUPAR for this term. Candidate normalized forms here are marked with asterisks.

classified bigram directly composes the left-hand side of the rule.

For discovery of candidate normalized forms (that then compose the right-hand side), we apply the domain-tuned BERT trained with a masked language modeling (MLM) head. We collect all contexts where the bigram appears in the Taiyo testing set and, with respect to each context, separately mask each of the two characters in the bigram and generate the top 15 predictions for the masked token.

The masked predictions are used to substitute the misclassified bigram term in its respective contexts, and are then submitted to a pretrained LM for UD annotation. Predicted terms are ranked best if, after substitution, UD supplied for the bigram in the parsed sentence consistently aligns with the domain-adapted LM annotation in terms of FORM, UPOS, and DEPREL fields. Table 2 demonstrates an example flow. The best ranked predicted terms compose a candidate normalized form list that can be further curated by a domain expert, and used to supplement and expand the rule set used to guide the rule-based expert system.

4 Evaluation

A prerequisite to using a domain-adapted LM to enrich the rule-based expert system is that it must first produce accurate UD in the target domain. Because Taiyo is unlabeled, we evaluate against ground truth labels from similar corpora in the target domain using the UD-Japanese-Modern treebank. Moreover, no canonical train-test split exists for Taiyo so we randomly partition the documents in a 75%/25% scheme (2121/707 documents) where the training set is used for domain tuning and the testing set for development of the rule set.

BERT and BiLSTM-based systems are implemented in PyTorch using the HuggingFace transformers (4.16.2) and esupar (1.1.5) libraries, respectively. Furthermore, we use the parallel shell tool for achieving speed-ups during UD annotation work (Tange, 2011). All experiments are performed on an NVIDIA Tesla P100 GPU.

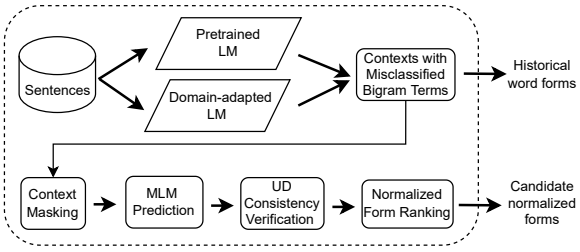

Figure 4: Flow diagram illustrating steps for rule generation using the domain-adapted LM. The goal of the module is to develop candidate normalized forms that can be used as substitutions for historical word forms. The mapping from a historical word form to some candidate normalized form composes one potential rule.

4.1 Systems

We evaluate the following four systems:

- **ADAPT-ESUPAR.** This system fine-tunes the contextualized embeddings on unlabeled data from the source and target domains, and then applies task-specific training using source domain labeled data from UD-Japanese-BCCWJ and UD-Japanese-GSD.

- **FINETUNED-ESUPAR.** This baseline omits the domain tuning step and only applies task-specific training using labeled data from the source domain.

- **OMIT-ESUPAR.** This baseline *omits* any samples from the target domain. It applies domain tuning using only unlabeled data from the source domain, and task-specific training also using source domain labeled data.

- **SUB-ESUPAR.** This baseline *substitutes* all target domain samples used during domain tuning with an equal amount of unlabeled samples from the source domain collected from the non-core portion of BCCWJ. Task-specific training is then applied using source domain labeled data.

All above systems use the pretrained BERT model KoichiYasuoka/bert-base-japanese-char-extended as a starting point, the same base used for fine-tuning the default configuration used by esupar. For evaluation, we draw comparisons with the following two pre-trained UD annotation tools:

- **ESUPAR**. This pretrained tool is applied under the default setting (KoichiYasuoka/bert-base-japanese-upos), as described in Yasuoka (2022). The systems are compared against its output for flagging instances of inaccurate UD output and generating bigram rules.

- **GINZA**. This pretrained tool, as described in Matsuda et al. (2019), is used when evaluating performance on ground truth labels from the UD-Japanese-Modern treebank.

5 Results

The results we report in this section are designed to answer the questions: (1) does the use of domain adaptation bring an improvement in UD annotation for the Modern Historical Japanese domain, (2) when compared to baseline systems, do the flagged instances raised by ADAPT-ESUPAR suggest unique characteristics about the target corpus, and (3) does the application of ADAPT-ESUPAR bring the best discovery rate for potential rules in the target domain?

5.1 Improving UD annotation in the target domain

We evaluate each system against the UD-Japanese-Modern treebank using the UPOS, UAS, MLAS, LAS, and BLEX metrics defined in Zeman et al. (2018), and incorporate performance from GINZA into our results.[6] Table 3 reports F1 scores for each system.

[6]We do not report results from the baseline ESUPAR as the specification of its training data for task-specific training are not made clear and we are unable to confirm whether this treebank is used as part of any step during its training procedure.

Model	UPOS	UAS	MLAS	LAS	BLEX
ADAPT-ESUPAR	**74.04**	**71.63**	**32.76**	**51.89**	**42.95**
FINETUNED-ESUPAR	70.28	64.44	29.41	47.42	38.47
OMIT-ESUPAR	69.77	65.15	28.56	47.49	38.12
SUB-ESUPAR	68.87	63.97	28.62	47.39	38.17
GINZA	69.51	58.88	24.26	44.05	31.84

Table 3: F1 score performance on the UD-Japanese-Modern treebank. Best score on a metric is bolded.

ADAPT-ESUPAR brings the most improvement across all metrics when compared to GINZA, with a 4.53% improvement in UPOS prediction, 12.75% in UAS, 8.5% in MLAS, 7.84% in LAS, and 11.11% in BLEX. Moreover, ADAPT-ESUPAR also improves over the best performing baseline (FINETUNED-ESUPAR) with a 3.76% improvement in UPOS, 7.18% in UAS, 3.35% in MLAS, 4.47% in LAS, and 4.48% in BLEX.

We also find that smaller fine-tuned LMs can outperform larger generally-trained LMs on this treebank. We compare the systems tested against the generally trained UDify (Kondratyuk and Straka, 2019). All four systems improve on the BLEX benchmark (35.47%; 7.48% improvement from ADAPT-ESUPAR) and remain competitive across other metrics.

5.2 Flagged FORM differences in bigram prediction

We evaluate each system against ESUPAR by collecting instances where bigrams predicted by the system in the FORM field are split into denominations by ESUPAR, e.g., “夫れ” is parsed as “夫” and “れ”. To quantify the importance of these discrepancies, we compare the degree of overlap in the flagged instances found for each system. Figure 5 shows the overlap in bigram parsing differences in the FORM field using a Venn diagram.

We observe heavy agreement in parsing differences among the four systems (33%). The similarity drops off considerably for any other pairing, however, there is notable agreement in the differences found by SUB-ESUPAR, FINETUNED-ESUPAR, and OMIT-ESUPAR (8%). Moreover, the diagram indicates regions that are unique to each system. SUB-ESUPAR has the largest share of these differences (9%), ADAPT-ESUPAR the second-largest (8%), and FINETUNED-ESUPAR the lowest (5%).

We explore these differences further by looking at the out-of-vocabulary (OOV) rate for bigram terms in the unique regions. We define an OOV

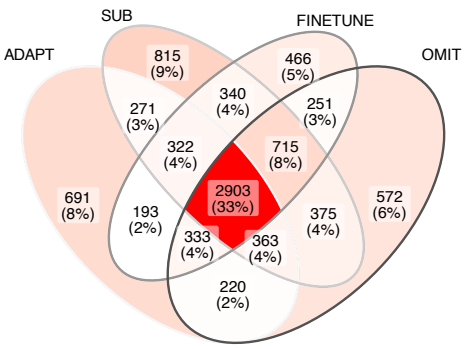

Figure 5: Venn diagram of differences in bigram parsing among the 4 tested systems. System names are abbreviated, e.g., "OMIT-ESUPAR" is shortened to "OMIT".

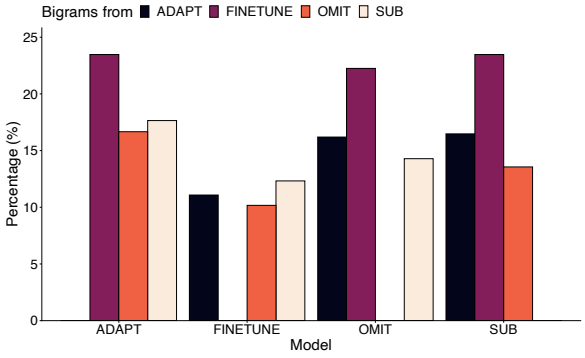

Figure 6: Bar plot showing candidate prediction success rate for "difficult" bigram terms. "Difficult" bigrams are defined as bigram terms identified by some system that do not contain any normalized forms that, according to ADAPT-ESUPAR, bring an improvement in UD annotation for that term when applied. X-axis shows the system being tested and Y-axis shows the percentage of difficult bigrams (with respect to another system) that have at least one candidate normalized form predicted by the tested system. System names are abbreviated.

term as any bigram that does not appear in the source domain training set: the concatenation of UD-Japanese-BCCWJ, UD-Japanese-GSD, and the samples collected from the non-core portion of BCCWJ. We find that for bigrams from these regions ADAPT-ESUPAR incurs the highest OOV rate (75.3%). For baseline systems, SUB-ESUPAR yields a 71.7% OOV rate, OMIT-ESUPAR 71.7%, and FINETUNED-ESUPAR 69.7%. Moreover, bigrams from the region consistent among all four systems incur the lowest OOV rate (67.8%).

5.3 Candidate normalized form discovery for difficult bigram terms

A prerequisite to rule generation is the discovery of candidate normalized forms that can serve as substitutions for bigram terms misclassified by ESUPAR. To determine whether ADAPT-ESUPAR offers the best potential for generating these candidates when compared to baseline systems, we focus specifically on bigram terms that are "difficult." Meaning, assuming that ADAPT-ESUPAR parsing is accurate, bigrams where a system is unable to produce any candidate normalized forms that, after substitution and submission to ESUPAR for annotation, align with the parsing given by ADAPT-ESUPAR for this term with respect to FORM, UPOS, and DEPREL fields. We test each of the other 3 systems on a given system's difficult bigram terms and collect the percentage of those that contain at least one candidate normalized form. In terms of the Venn diagram in Figure 5, we apply this procedure to bigrams from the consistent portion among the 4 systems. Figure 6 reports the results from this experiment.

In terms of bigrams found to be difficult, we observe systems are most successful when predicting bigrams from FINETUNED-ESUPAR (purple bars). ADAPT-ESUPAR and SUB-ESUPAR are both able to suggest candidates for 23.5% of its terms. Conversely, systems appear to have the most difficulty predicting terms from OMIT-ESUPAR (orange bars). ADAPT-ESUPAR can suggest candidates for 16.7% of its bigrams, FINETUNED-ESUPAR 10.2%, and SUB-ESUPAR 13.6%.

If performance is to be measured as a model's ability to maintain a high candidate suggestion rate while also minimizing the rate for other models to suggest candidates for its own bigrams, then ADAPT-ESUPAR exhibits strong results on this task. We find that, with respect to each model, ADAPT-ESUPAR either ties for or has the highest candidate prediction rate on the respective bigrams. Moreover, candidate prediction rate on bigrams from ADAPT-ESUPAR (black bars) is lower than the rate ADAPT-ESUPAR delivers on the bigrams from any other model. This observation does not hold for other setups.

6 Discussion

Use of domain adaptation methods. A crucial component of this research is determining whether the application of domain adaptation methods is required for discovery of candidate normalized forms that can be used for refinement of a rule-based expert system. While a large portion of the flagged differences in bigram parsing are consistent among

the 4 systems tested here, we find these bigram terms incur the lowest OOV rate and, therefore, may not be best indicative of the lexical variants that are unique to the target domain. When putting this in context of rule generation, we also find that ADAPT-ESUPAR is the system that exhibits the best candidate discovery rate for difficult bigrams in this region.

It is important to highlight contributions made by our baselines. The domain tuned SUB-ESUPAR is able to flag more unique instances in bigram parsing than ADAPT-ESUPAR while only trading off a 3.6% reduction in the OOV rate incurred when compared to ADAPT-ESUPAR. It is only when domain tuning is excluded that deterioration in performance becomes apparent. FINETUNED-ESUPAR exhibits the lowest number of unique flagged instances and candidate discovery rate, and every other system finds most success when predicting its difficult bigram terms. These results are indicative of the effect of domain tuning and that any domain tuning, notwithstanding the use of target domain data in this step, is still able to bring improved performance on these tasks.

However, this result is strongly dependent on the source domain data used and the degree of overlap that exists between source and target domains. While the overlap in vocabularies between Taiyo and BCCWJ is relatively small (≈50%), the overlap that does exist may present BERT an opportunity to learn useful representations about the target domain from the unlabeled source domain data. Nevertheless, the degree of difference in our results is maximized when unlabeled samples from the target domain are incorporated into domain tuning.

Enabling helpful evaluations by domain experts. The true test of the proposed workflow is the value it creates for the domain expert. The flagged instances direct attention to errors in pretrained output that may be most egregious and the suggested rules provide a mechanism for improving that output in a manner that supports both human comprehension and further manual revision. The workflow, then, offers a more principled strategy for manual labeling campaigns than by rote manual revision of pretrained output.

Indeed, the potential rules that can be suggested currently are limited by the single masked character predictions that are possible under the constraint of differences in two-character bigram tokenization. The number of flagged instances can be increased by relaxing the constraint to include predicted bigrams that should be split into denominations and formulations that involve more than two characters. In the case of the former, the current masking strategy has a direct extension.

LM refinement using the rule-based expert system. The current proposed workflow has made use of a rule-based expert system that is usefully guided by a domain-adapted LM. However, an intriguing implication of this work for the DH community is the possibility to apply the steps in the reverse: using the expert system as a means to inform the training of a pretrained LM. Because the output of the expert system is capable of bringing improved UD annotation in the target domain, its output can serve as a means for obtaining accurate labeled target domain data that can then be used for supervised task-specific training. This can serve as an additional means for obtaining improved UD annotation in the target domain while enabling the process to be driven by the domain expert.[7]

7 Conclusion

This paper demonstrates the use of domain adaptation methods for bringing improved UD annotation in the Modern Historical Japanese domain. It incorporates a domain-adapted LM into a workflow designed to enable evaluation by domain experts. Features salient to this workflow: the domain-adapted LM is deployed to flag instances of incorrect pretrained UD output and these are then used to form rules that, when applied, improve annotation accuracy in these contexts. The rules can be used to supplement and enrich a rule-based expert system.

Our experiments indicate that domain adaptation is a necessary step to enable flagging of incorrect UD output and generation of candidate normalized forms that can be used to build a rule set. However, we find that the choice of source domain data used for domain adaptation is significant, especially when there exists considerable similarity between data sampled from the source and target domains. To best maximize this transfer, the source data sampled should minimize similarity with the target domain. We are interested in exploiting associations between degree of dissimilarity in domains and

[7] We recognize that fine-tuning LMs on labeled data from the target domain can cause catastrophic forgetting in labeling accuracy in the source domain (Han and Eisenstein, 2019). Because the principal concern of this research is obtaining improvement exclusively in the target domain, we consider this side effect beyond the scope of this work.

margin of improvement brought by transfer learning in the context of historical texts. Future work will also do well to explore metrics other than vocabulary overlap for quantifying this similarity.

We hope to have the proposed workflow reviewed and evaluated by domain experts in the future, and that this work can help pave the path toward greater adoption of pretrained neural architectures into scholarly workflows in DH.

8 Acknowledgements

We would like to thank the Department of Computer Science at the University of Miami for providing computational resources necessary for running the experiments in this research. We would also like to thank the reviewers for their constructive comments and feedback.

References

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. 2018. Universal Dependencies version 2 for Japanese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christoph Aurnhammer, Iris Cuppen, Inge van de Ven, and Menno van Zaanen. 2019. Manual annotation of unsupervised models: Close and distant reading of politics on reddit. *Digital Humanities Quarterly*, 13(3).

Jerry Bonnell and Mitsunori Ogihara. 2022. Rule-based adornment of modern historical japanese corpora using accurate universal dependencies. *Digital Humanities Quarterly*, 16(4).

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Matthew Kirschenbaum. 2007. The remaking of reading: Data mining and the digital humanities. *The National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*, 134.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Kikuo Maekawa. 2006. Kotonoha, the Corpus Development Project of the National Institute for Japanese Language. In *Proceedings of the 13th NIJL International Symposium: Language Corpora: Their Compilation and Application*, pages 55–62.

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.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. Pre-training Language Models for Historical Languages. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Hiroshi Matsuda, Mai Ohmura, and Masayuki Asahara. 2019. tan tani hinshi no youhou aimaisei kaiketsu to izon kankei raberingu no douji gakushyu (simultaneous learning of ambiguity resolution and dependency labeling). *gengo shori gakkai dai 25 kai nenji taikai (The 25th Annual Meeting of the Association for Natural Language Processing)*.

Barbara McGillivray, Thierry Poibeau, and Pablo Ruiz Fabo. 2020. Digital humanities and natural language processing: "je t'aime... moi non plus". *Digital Humanities Quarterly*, 14(2).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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

Ryosuke Shirai, Yukio Matsumura, Toshinobu Ogiso, and Mamoru Komachi. 2020. Machine Learning-based Sentence Boundary Detection for Modern Japanese Texts. *jouhoushori gakkai ronbunshi (Information Processing Society of Japan Journal)*, 61(2):152–161.

Omri Suissa, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet. 2022. Text analysis using deep neural networks in digital humanities and information science. *Journal of the Association for Information Science and Technology*, 73(2):268–287.

O. Tange. 2011. Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47.

Koichi Yasuoka. 2020. keitaiso kaisekibu no tsukegae niyoru kindai nihongo (kyuu ji kyuu kamei) no kakari uke kaiseki (dependency analysis of modern japanese (old characters and old kana) by replacing the morphological analysis department). Technical Report 3, jouhoushori gakkai (Information Processing Society of Japan).

Koichi Yasuoka. 2022. Transformers to kokugokenchou tani niyoru nihongo kakari uke kaiseki moderu no seisaku (Production of Japanese dependency analysis model by Transformers and National Institute for Japanese Language and Linguistics). *IPSJ SIG Technical Report*, 2022-CH-128(7):1–8.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.