



Rule-based Adornment of Modern Historical Japanese Corpora using Accurate Universal Dependencies

Jerry Bonnell <j_dot_bonnell_at_miami_dot_edu>, Department of Computer Science, University of Miami 
<https://orcid.org/0000-0002-7404-9160>

Mitsunori Ogihara <m_dot_ogihara_at_miami_dot_edu>, Department of Computer Science, University of Miami 
<https://orcid.org/0000-0002-5690-7854>

Abstract

Historical materials are an indispensable resource for many scholarly workflows in the Digital Humanities. These workflows can benefit from the application of natural language processing (NLP) pipelines that offer support for tokenization, tagging, lemmatization, and dependency parsing. However, the application of these tools is not trivial as “off-the-shelf,” or pre-trained, tools are prone to error when given historical text as input and training data development can be expensive to carry out in terms of time and expertise needed. This paper introduces a rule-based workflow that can produce improved annotations encoded in Universal Dependencies (UD) targeted for modern historical Japanese corpora using only a pre-trained UD tool as a starting point. The proposed workflow reduces the amount of manual review time needed for training data development and brings improvements over pre-trained tools on a word segmentation task. Moreover, the workflow has the potential to pave the path toward adapting advanced NLP technologies to historical corpora under study.

1. Introduction

Scholarly workflows in the Digital Humanities can benefit from the application of natural language processing (NLP) tools [Argamon and Olsen 2009]. To serve any practical use to DH scholars, these tools must produce accurate and reliable results for the corpus under study, especially when the intent is to provide linguistic annotations that enable syntactic analyses. For historical texts, such an application is non-trivial because pre-trained tools are usually trained over contemporary materials (e.g., “newswire” articles, microblog text, and general books), making any direct application prone to error [McGillivray et al. 2020]. For trainable language-agnostic pipelines like UDPipe that provide annotations for tokenization, tagging, lemmatization, and dependency parsing simultaneously using Universal Dependencies (UD), it becomes necessary to train UDPipe from scratch using model training data derived from the historical literary materials [Shirai et al. 2020] [Straka and Straková 2017]. Developing these UD annotations can be done through manual effort alone or by applying a pre-trained UDPipe model to the raw text and then later post-editing its output manually for errors [Scannell 2020]. However, the structure of nodes in the dependency tree may also require changing and, for East Asian languages like Chinese and Japanese that can be written without specifying word boundaries other than insertion of punctuation marks, tokenization is often difficult, and thus errors in the tree can be severe when tokenization is inaccurate. Consequently, manual revision can be challenging to carry out in terms of time and expertise needed in working with UD.

Motivated by this issue, the present paper purposes to address the following questions with respect to modern historical Japanese corpora: (1) can accurate UD annotations be developed from scratch using pre-trained tools while also minimizing the amount of manual effort needed for correction, (2) can said generated annotations be used as model training data to achieve improved accuracy on a fundamental NLP task, e.g., word segmentation, (3) does the trained model adapted to historical materials have a substantial effect on the output parsings produced when compared to pre-

trained tools, and (4) can the proposed workflow be carried out by non-experts in UD? The answers can be encouraging to DH scholars working with historical corpora who are not subject experts in UD, but would like to make more frequent use of linguistic metadata in their scholarship.

To attempt an answer, this paper introduces a rule-based workflow for modern historical Japanese corpora that produces more accurate UD annotations directly from the raw text in the corpus using the output of pre-trained tools as a starting point. It consists of:

- A *text normalization* step using handcrafted rules to normalize historical lexical variants to a more canonical form so that the input text may receive a more accurate parsing by a pre-trained model, and
- An *assignment* step in which these annotations are linked with word forms in the original text.

In this way, any parser that learns their model from this data also learns how to deal with historical text.

2. Related Work

[Shirai et al. 2020] proposes a related workflow to reduce the amount of manual effort needed for correcting sentence boundaries in modern historical Japanese materials maintained by the National Institute for Japanese Language and Linguistics (NINJAL) [NINJAL 2021]. Their method trains a combination of UDPipe and CRF++ using “core data” manually corrected by experts to predict sentence boundaries on unlabeled data. Our approach also corrects UD annotations, but aims to do so automatically and does not use any “gold-standard” or already corrected data as a starting point.

Text normalization is a well-known preprocessing problem in NLP with many proposed solutions. For Japanese text normalization, [Ikeda et al. 2016] presents a deep learning encoder-decoder model for normalizing Japanese social media and microblog text. Our approach uses text normalization in the context of historical text, which has received recent attention in Bollmann’s survey though Japanese is not covered specifically [Bollman 2019]. Moreover, the goal of our approach is not historical text normalization and involves it only as an intermediary step as part of a larger workflow that aims to provide accurate UD annotations to modern historical Japanese text.

3. Methods

3.1. Data Source and Model Selection

We follow [Shirai et al. 2020] and adopt the Taiyo (太陽) magazine published by Hakubunkan and maintained by NINJAL as a historical corpus of written Japanese [Maekawa 2006] [NINJAL 2021]. Taiyo was the best-selling general-interest magazine of the time, consisting of 3400 articles written by about 1000 writers published during the Meiji (明治) and Taisho (大正) periods between the years 1895 and 1925. Chief among the reasons for its selection in this study is the challenge the corpus presents computationally because it reflects the dramatic changes in literary and colloquial writing styles that were occurring at that time with both styles sometimes coexisting within the same article. Therefore, Taiyo is representative of the kinds of historical corpora humanities scholars regularly deal with.

To evaluate our methods, we also incorporate three other magazines made available by NINJAL in its corpora of Modern Japanese: Josei (女性) (1894 - 1925), Meiroku Zasshi (鳴鹿) (1874 - 1875), and Kokumin (國民) (1887 - 1888) [Tanaka et al. 2012] [NINJAL 2021]. Table 1 shows the distribution of punctuation and sentences across the four collections according to the NINJAL markup.^[1] The Meiroku and Kokumin collections are especially valuable to this study as NINJAL has further adorned the text with morphological information in the form of short-word units (SUWs).

Corpus	% sentences ending with period marker (。)	% sentences ending with comma marker (、)	% sentences ending with mixed markers	% sentences ending without markers	Total # sentences
Taiyo	18%	65%	0.05%	17%	361K
Josei	22%	66%	0.1%	12%	68K
Kokumin	2%	58%	0.5%	40%	46K
Meiroku	0.01%	4%	0%	96%	9K

Table 1. Punctuation and sentence distribution in the four NINJAL collections.

GiNZA is a recent open-source NLP framework that advertises as an easy “one-stop” solution for providing tokenization, part of speech tagging, and dependency analysis on Japanese text simultaneously and enjoys popularity among NLP researchers [GiNZA 2021]. It can also supply output in CoNLL-U format which can be used to train pipelines like UDPipe to fulfill fundamental NLP tasks, e.g., word segmentation [Nivre et al. 2016] [Straka and Straková 2017]. However, its parsing model is trained on a portion of UD Japanese BCCWJ – an annotated corpus of contemporary Japanese – and, therefore, is unsuitable for historical Japanese corpora [Maekawa et al. 2014]. This makes GiNZA a candidate for use as a pre-trained tool in our study.

3.2. Workflow

We present the workflow used by our research to produce improved UD metadata directly from the raw sentences of the Taiyo corpus. Three steps organize the work: (1) a development phase where a set of handcrafted rules is generated to normalize portions of the historical text, (2) a text normalization phase that then applies the rule set to fulfill the needed transformation, and (3) application of GiNZA to the normalized text followed by an alignment step that assigns the UD metadata generated to word forms from the historical text. These steps are realized by means of a Python script, which we have made available through GitHub at <https://github.com/jerrybonnell/Rules2UD>. We have also provided a Binder link through this repository that launches a live Jupyter notebook in an executable environment so that users can interact with the tool without needing to install any packages on their machine.

3.2.1. Overview

Before proceeding to describe our workflow in detail, we offer a brief overview of the dichotomy between symbolic and non-symbolic approaches, and how both are combined into a single workflow in the proposed work. The developed collection of rules used for producing normalized text can be viewed as an “expert system” that generates direct applications of domain knowledge for decision-making tasks. A salient aspect of this system is that rule application is a symbolic transformation: a sequence of symbols is substituted with another sequence where the symbols compose written language as either ASCII (e.g., “a”, “z”, “+”), UTF-8 unicode (e.g., “あ”, “夕”, “勉”), or numerals (e.g., “0”, “2”, “3”), and the representations before and after the substitution support human comprehension. In contrast, non-symbolic systems in the form of pretrained language models like GiNZA transform raw text into an internal numerical representation (e.g., word embeddings, contextual word embeddings, etc.) through means of deep learning that, while necessary for its computation and effective for achieving state-of-the-art across NLP benchmarks, obstructs any meaningful interpretation as per current methods [Qiu et al. 2020]. A system, then, that were to make use of both symbolic and non-symbolic approaches simultaneously would be difficult to achieve because the representations are incompatible.

However, if the expert system is allowed to carry out its work as a separate preprocessing step and the language model follows with its own independent computation, then the language model can take advantage of any symbolic transformations made by the expert system when receiving its input, thereby guiding its own computation. While the representations used during that computation are no longer symbolic, the output returns to a representation that is, which can again be used by the expert system for further postprocessing to complete the required annotation. By making these interactions indirect, the systems can inform one another effectively. This approach forms the basis for the

workflow presented here.

3.2.2. Phase I: Rule Development

We define a rule as some mapping between two word forms, a historical usage and a normalized usage. A collection of rules is a set of these mappings which, in implementation, is a Python dictionary of key-value pairs. We generate rules by manual evaluation of the GiNZA output to identify errors in the parsing that occur primarily because of historical usages. For instance, a significant portion of the rule set are character substitutions, mapping historical kanjis presently rare in use (called Kyukanji 旧漢字), e.g., 黨, to contemporary usages (called Shinkanji 新漢字), e.g., 党. Also included are substitutions of 「わ行」 Hiragana and 「ワ行」 Katakana to their modern forms, e.g, changing 「ゐ」 to 「い」. A key feature of this evaluation is that only the FORM field is considered for correction and no review time is given to the HEAD and DEPREL fields that form the dependency tree, reducing the overall amount of manual effort needed.^[2] We have crafted approximately 600 rules under this approach using only the Taiyo corpus as a data source. Figure 1 shows an example output from the GiNZA tool with its associated dependency structure. The transliteration of the sentence is as follows:

社 會 の 發達 に 從ふ て、

sha kai no hattatsu ni shitagou te

The fourth component consists of two characters. The first character of the pair is ordinarily pronounced “hatsu”, but when combined with the next character, which is pronounced “tatsu”, the pronunciation changes to its shortened version “ha-”, thereby yielding “hattatsu” instead of “hatsutatsu”. The sixth component also contains the modification in the pronunciation. The first character is normally pronounced “shitaga”, but when “ふ” is attached, the pronunciation changes to “shitago” and the pronunciation of “ふ” changes from “hu” to “u.” The encoded dependency tree given by the HEAD and DEPREL fields is shown using the CoNLL-U viewer tool [Straka and Sedlák 2021]. The fields FEATS and DEPS specified by the CoNLL-U format are not supplied by GiNZA and the MISC field is omitted for sake of presentation. A curious result is the treatment of the first two characters “社會”, which means “society.”

ID	FORM	LEMMA	UPOS	XPOS	HEAD	DEPREL
1	社	社	NOUN	名詞-普通名詞-助数詞可能	2	compound
2	會	會	NOUN	名詞-普通名詞-一般	4	nmod
3	の	の	ADP	助詞-格助詞	2	case
4	發達	發達	NOUN	名詞-普通名詞-サ変可能	6	obl
5	に	に	ADP	助詞-格助詞	4	case
6	從ふ	從ふ	VERB	動詞-一般	0	root
7	て	て	SCONJ	助詞-接続助詞	6	mark
8	、	、	PUNCT	補助記号-読点	6	punct

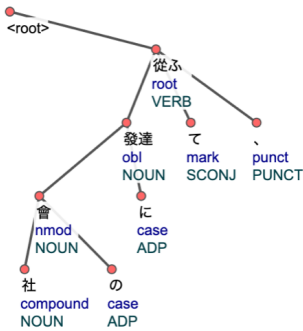


Figure 1. An example CoNLL-U table returned by the GiNZA tool and its corresponding dependency structure for the sentence “社會の發達に従ふて、”.

Because some rules can be general, a global pattern match-and-replace could be too aggressive and prone to error. To overcome this, the rule set is partitioned into disjoint sets so that some rules may apply only after a condition is met. For instance, “つた” → “つた”. may trigger only when the word form “つた” appears after some kanji. Table 3 shows the different rule sets and distribution.

Set / Condition	Proportion	Example
Global rules	93%	假令 → たと, 舊 → 旧
KANJI + *	5%	はら → わら, へば → えば
HIRAGANA + * + {さ, し, す, せ, そ}	0.8%	照し → 照らし
* + ¬HIRAGANA	0.8%	假令い → たとえ
HIRAGANA + *	0.3%	がよい → が良い
¬KANJI + *	0.3%	大なる → 大いなる
* + KANJI	0.2%	亦た → また

Table 2. Top seven rule sets and proportions with respect to the rule collection. Rule sets are expressed as conditions needed for application, e.g., rules in KANJI + * may only fire if the historical word form (denoted by wildcard *) is preceded by some kanji.

3.2.3. Phase II: Text Normalization

This step receives as input a single sentence from the target corpus and returns the sentence after normalization. Each rule set is visited in turn for possible applications. If a match is found and meets the condition of the group, the word form is replaced by the value in the rule's key-value pair. The procedure builds state about the match in a list of “start” and “end” index pairs; this information is stored in a dictionary and is needed for successful alignment of UD annotations to the historical word forms in the proceeding step. Following is a breakdown showing an example sentence, two matches, and the corresponding normalized output:

16

"其	れ	で	寧	ろ	小	黨*	分	立	で	行
<	所	ま	で	行	<	が*	よ*	い*"		
so	re	de	mushi	ro	shou	tou*	bun	ritsu	de	i
ku	tokoro	ma	de	i	ku	ga*	yo*	i*		

黨 → 党 (global rule), がよい → が良い (hiragana + * rule)
{'黨': [(6, 6)], 'がよい': [(17, 19)]}

"其	れ	で	寧	ろ	小	党*	分	立	で	行
<	所	ま	で	行	<	が*	良*	い*"		
so	re	de	mushi	ro	shou	tou*	bun	ritsu	de	i
ku	tokoro	ma	de	i	ku	ga*	yo*	i*		

This sentence contains one historical word form, 黨, and an ambiguous character sequence がよい. The former is normalized to the form 党. The latter in this context can be spelled as が良い with the use of one kanji. The alternate form is 「通い」, which is normally pronounced as 「かよい」, but in the case where it is preceded by a general noun representing a commercial location, pronounced as 「がよい」. The former is a global rule while the latter may only trigger if the word form is preceded by a hiragana. This normalized sentence is ready for submission to GiNZA for parsing, which returns UD metadata in CoNLL-U format.

17

3.2.4. Phase III: Aligning UD Metadata

The fundamental problem with the CoNLL-U output returned by GiNZA in the previous step is that the UD annotations supplied are for the *normalized* text and not the *historical* text. Meaning, any model trained using the normalized output loses information about lexical variation and, consequently, is ill-suited for direct application on the historical text. Therefore, we desire a procedure that aligns *normalized* UD annotations with the *historical* word forms present in the source sentence.^[3] The state built up during the match is brought forward to this step to determine the start-end locations where the alignment should occur in the word forms given in the FORM field. Once found, the historical form is

18

referenced from the state dictionary to string replace the normalized form identified by the start-end pair. This aligns the historical word form with the normalized UD annotation.

However, the alignment is complicated by nature of tokenization: the normalized word form that needs to be replaced may not be contained within a single row of the FORM field. This yields two scenarios when doing the alignment:

19

1. the normalized word form is contained within a single row, and
2. the normalized word form spans multiple rows in the FORM field

We envisage each row as containing two parts: a part that is not influenced by the normalization (part A) and a part that is (part B). The two scenarios are demonstrated using the GiNZA output from the example sentence in the previous step.^[4]

20

1. 其れ
2. で
3. 寧ろ
4. 小党
5. 分立
6. で
7. 行く
8. 所
9. まで
10. 行く
11. ガ
12. 良い

Scenario #1: the normalized form 党. The start-end pair (6,6) identifies 小 as the part not influenced by normalization (part A) and 党 as the part that is (part B). The normalized form is fully contained by the row and, therefore, string replacing 党 with 黨 completes the alignment.

21

Scenario #2: the normalized form ガ 良い. The start-end pair (17, 19) identifies part A empty and part B to be split across two rows, rows 11 and 12. To help make informed decisions, we use the normalized GiNZA output to guide the character lengths to maintain for each row. Meaning, the lengths of each normalized row should remain unchanged after the alignment is completed.^[5] For row 11 in the example, this means string replacing ガ with just the first character from the historical word form, which happens to also be ガ. Proceeding on to the next row, the remaining characters to be aligned are now contained within a single row – a Scenario #1 case. In other examples where this is not true, the operation of Scenario #2 is repeated.

22

3.3. Corner Cases

There is a possibility for conflicts to arise during processing. Two main issues are addressed here: (1) overlapping rules, and (2) rows with “blank” forms.

23

3.3.1. Overlapping Rules

There are scenarios where multiple rules can fire on the same historical word form or portions of it. These are usually due to the large number of kanji substitution rules in the rule set. This introduces undefined behavior as the alignment step is unable to determine which rule should have precedence and be applied first. The following gives an example of such a case with two rules that overlap:

24

將*	た	又	軍	港	な	り
ha*	ta	mata	gun	kou	na	ri

candidate rule #1: 將*た又 → 果て又
 candidate rule #2: 將* → 將

25

The second rule is a simple substitution of a traditional kanji character with a modern character. The first rule is a bit of a “hack” where the second character 「た」, pronounced “ta,” is substituted with 「て」, pronounced “te.” The application of the first rule thus changes it to a more modern way of communicating the same meaning at the cost of changing the pronunciation.

26

We define two or more rules to be “overlapping” when the ranges of the indices covered by the historical forms to be substituted overlap. When a rule is fully covered by another rule, that is, its historical form is fully contained by the historical form of another rule, the covered rule is jettisoned from application as it is assumed that longer rules are more specific – and, hence, more useful – than “general” short rules like kanji substitution rules. The above is an example of such a scenario where candidate rule #2 is a kanji substitution rule and is removed from processing.

27

Some scenarios can be more complex when the historical form is not covered by another rule, as in the following example. The seventh character of this sentence 「か」 is normally pronounced “ka,” but the traditional (early modern period) spelling, with the succeeding character 「う」 (“u”), forces the pronunciation 「こ」 instead.

働	く	べ	き	に	働	か	う	と	す	る
hatara	ku	be	ki	ni	hatara	ko*	u	to	su	ru

candidate rule #1: か*う → こう
 candidate rule #2: 働か* → 働

28

When the simple deletion technique above is no longer applicable, we form non-conflicting combinations of rules that also maximize the number of rules to include. Only two exist for this example: {かう} and {働か}. To determine which to use for processing, each combination is scored along four axes:

- Number of rows in the CoNLL-U table returned by GiNZA.
- Number of rows with a “bad reading,” that is, the reading given in the MISC field does not contain a katakana pronunciation (e.g., 日本 instead of ニホン).
- Number of rules present in the combination.
- A normalized value giving the “agreement percentage” in the BLEX, CLAS, and MLAS metrics by comparing the aligned accuracy parsing from a combination against that with no rule application [Zeman et al. 2018].^[6] Lower values denote more difference which shows that the rule combination exhibits a larger effect.

29

The combination with a minimum score is selected for processing. If multiple minima exist, the instance is flagged for inspection. However, this has not occurred during our experiments with the rule set applied as of this writing.

3.3.2. Rows with Empty Forms

30

Situations can arise where there are not enough characters in the historical word form to “fill” the rows spanned by the normalized word form. In the following example sentence, the normalized form “而かして” (or, alternatively, “しこうして”) spans the first four rows of the FORM field in the CoNLL-U output returned by GiNZA.^[7]

而*		て*	移	轉	せ	ら	れ	た	る	繪
畫	は									
shikoushi*		te*	i	ten	se	ra	re	ta	ru	
kai	ga	ha								

```
rule: 而て → 而かして
FORM field from CoNLL-U output:
1   而
2   て
3
4
```

Only two characters from the historical form are available to distribute among four rows, which results in the alignment step after completion leaving the third and fourth rows in the FORM field empty. While the issue seems like implementation error, it points to a problem with the rule itself: the normalized form is not helpful in guiding GiNZA to a more accurate parse that sees 而かして as a single word form, hence the tokenization into multiple rows. The solution is an adjustment of the normalized form in the rule, e.g., changing 而かして to 而して. This yields a parse where the normalized form spans a single row, is more accurate, and allows the alignment step to proceed without error.

31

4. Results

This section evaluates the rule set introduced by this research along three criteria:

32

1. generalization of rules crafted from the Taiyo corpus to other similar corpora,
2. the effect of the rules on the CoNLL-U output when compared to a parsing without rule application, and
3. evaluating the performance of a word segmentation task when training a NLP pipeline using the proposed rule set.

4.1. Generalization to Other Corpora

Figure 2 shows the top 10 most frequently applied rules in the Taiyo corpus, and the frequency of application for said rules across the other three corpora.

33

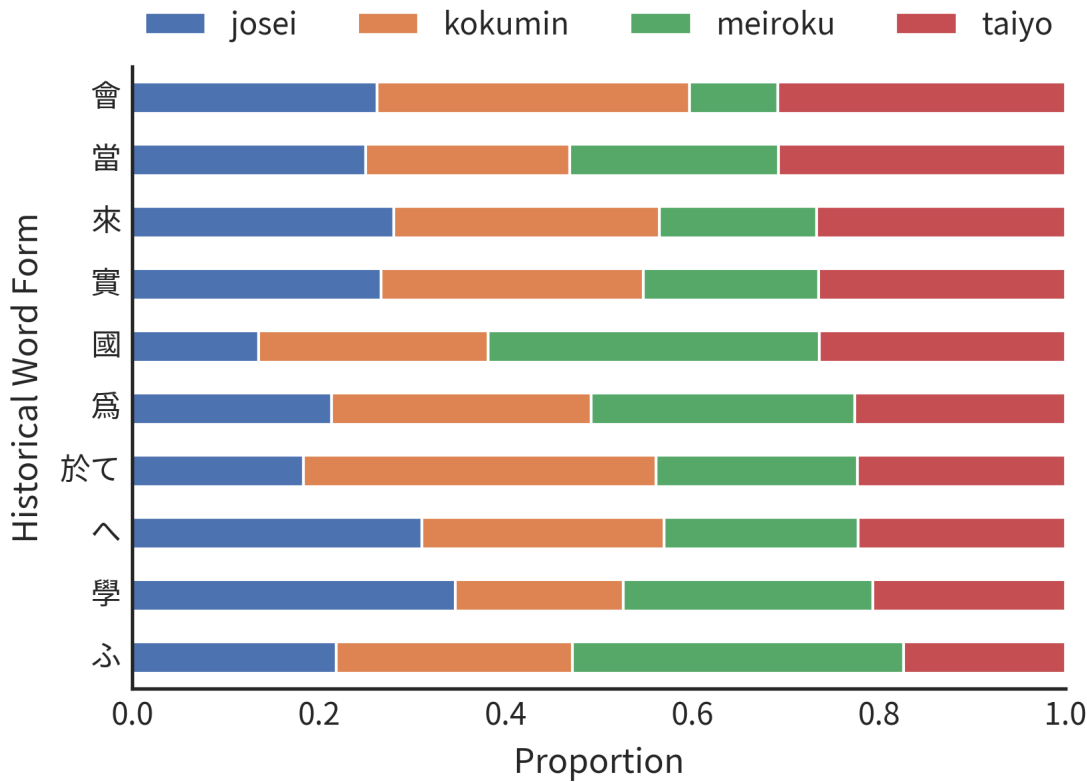


Figure 2. Top 10 most frequently applied rules in the Taiyo corpus and application of these rules across the three other corpora in the collection. Frequency is measured in terms of proportion with respect to other corpora. The historical word form of the rule is shown.

Overall, we observe fair representation of the Taiyo rules in the other corpora. These rules make up roughly, on average, a quarter of applications with respect to the other corpora, with the other three contributing about evenly to the remaining 75% of application. Indeed, some rules saw disproportionately more application in Taiyo than in other corpora, e.g., ‘會’ saw over 30% application in Taiyo while only 10% in Meiroku. The character as a noun means “group” and as a verb means “to meet”. The use of the character is used mostly in the former sense. The difference in the frequency is due to the fact that Taiyo speaks more frequently about groups (specifically, political parties and groups) than Meiroku.

34

4.2. Effect on CoNLL-U Output

If the proposed approach is to be successful in bringing improvement to a fundamental NLP task like word segmentation, it must first have an observable effect on the resulting parsings generated by GiNZA. This is especially critical when evaluating the method against unlabeled corpora like Taiyo where it is not possible to compare predictions made with any ground truth labels. In the absence of ground truth, visualizing disagreements in CoNLL-U output between the proposed approach and what a pre-trained tool would normally generate can showcase whether any effect can be seen in the output and the quantity of that difference. If the answer is in the positive, then this raises the possibility for improved performance on the historical materials.

35

To evaluate this, CoNLL-U output with and without rule application is compared across the 4 corpora using the BLEU, CLAS, and MLAS metrics as defined in [Zeman et al. 2018]. The proposed rules are gradually introduced into the collection before application, and a test is done at every 10% interval. Rules are selected according to frequency, with most frequently occurring rules introduced last. Because of the large corpus size and the repetition of this test at multiple intervals, a random sample of 1,000 sentences is selected to avoid incurring high computation costs. This process is repeated 10 times and the mean disagreement percentage over the 10 runs, derived from the aligned accuracy score given by [Zeman et al. 2018], is reported. Figure 3 shows these results.

36

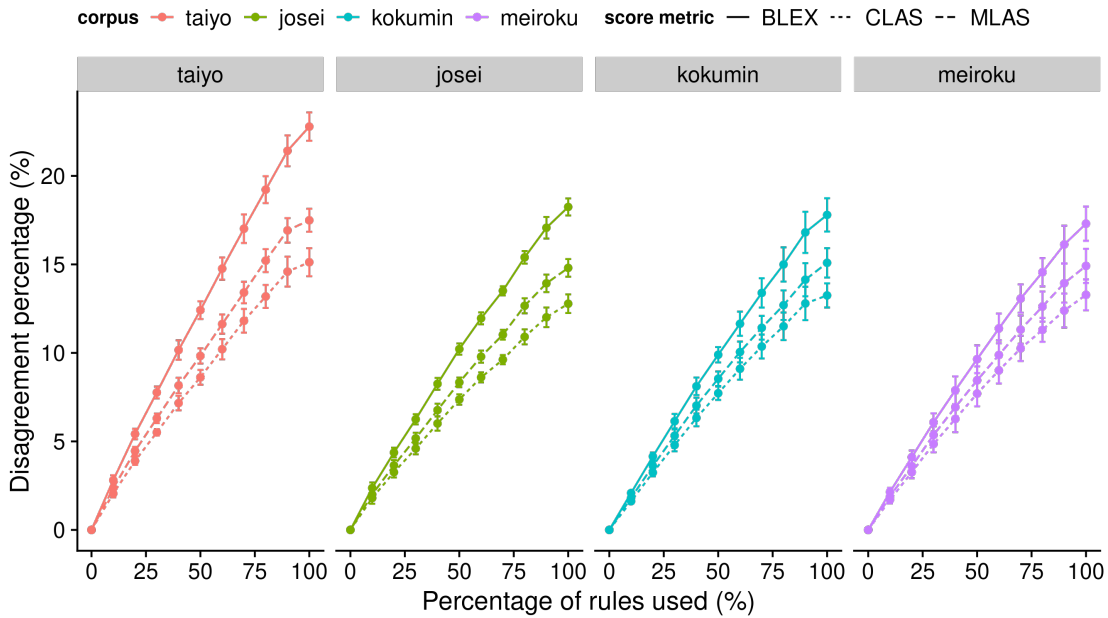


Figure 3. Disagreement percentage in BLEX, CLAS, and MLAS metrics between CoNLL-U output with and without rule application across the 4 corpora in the collection. X-axis shows percentage of rules from the rule collection introduced during application, and tests are done at every 10% interval. Each test selects a random sample of 1,000 sentences for comparison and the test is repeated 10 times. The mean score is reported and error bars are given at 99% significance.

We observe a gradual increase in disagreement with the original parsing as more rules are introduced, with the maximum disagreement at full rule usage reaching 22.8% in the BLEX metric from the Taiyo corpus and the minimum 12.8% in the CLAS metric from Josei. BLEX results yield the highest disagreements because of the strict conditions it places on the two CoNLL-U files being compared, relative to the other two metrics. Despite frequent rules being introduced last, the amount of disagreement begins to plateau when rule usage approaches the complete rule set; this could be an artefact of character substitution rules that, while frequently applied, have minimal effect on the dependency structure.

37

4.3. Improving Word Segmentation

We evaluate whether the observed effect on the dependency structure of the CoNLL-U output can bring an improvement in a basic NLP task in word segmentation using UDPipe. The test is performed by comparing three experimental set-ups:

38

1. a pre-trained UDPipe model trained on UD Japanese-GSD, a UD resource curated from Japanese Wikipedia [Asahara et al. 2018];
2. a UDPipe model trained over GiNZA CoNLL-U output using Taiyo documents without any rule application; and
3. a UDPipe model trained identically but with the addition of rule application.

The training data is prepared using five-fold cross-validation over the documents in the Taiyo corpus where the testing fold is not used due to lack of word-level metadata. Instead, the trained models are tested against documents from the Kokumin and Meiroku collections which supply short word unit (SUW) annotations and can be used for ground truth. Figure 4 shows an example of the SUW tag for two tokens in the Kokumin corpus, “陛下” and “及び.” The experiment is repeated 10 times for the two setups with and without rule application. In total, 101 different models are evaluated.

```

<SUW end="210" form="ヘイカ" kanaToken="ヘイカ" lForm="ヘイカ" lemma="陛下"
orderID="120" orth="陛下" orthToken="陛下" pos="名詞-普通名詞-一般"
pronToken="ヘーカ" start="190" wType="漢">
陛下
</SUW>
<SUW end="230" form="オヨビ" kanaToken="オヨビ" lForm="オヨビ" lemma="及び"
orderID="130" orth="及び" orthToken="及び" pos="接続詞"
pronToken="オヨビ" start="210" wType="和">
及び
</SUW>

```

Figure 4. An example of two SUW tags corresponding to the tokens “陛下” and “及び” as they appear in the XML file “k188701.xml” in the Kokumin corpus made available by NINJAL.

In keeping with the methods proposed in [Shirai et al. 2020], our evaluation scheme measures correctness of “B” label estimation using the metrics precision, recall, and F1. However, we adjust the meaning of the “B” label to mean “start of token” and “I” label as “rest of token”. Predictions using the tokenizer option in UDpipe are compared with the “B” and “I” truth labels derived from the SUW tags in Meiroku and Kokumin. Figure 5 reports the results.

39

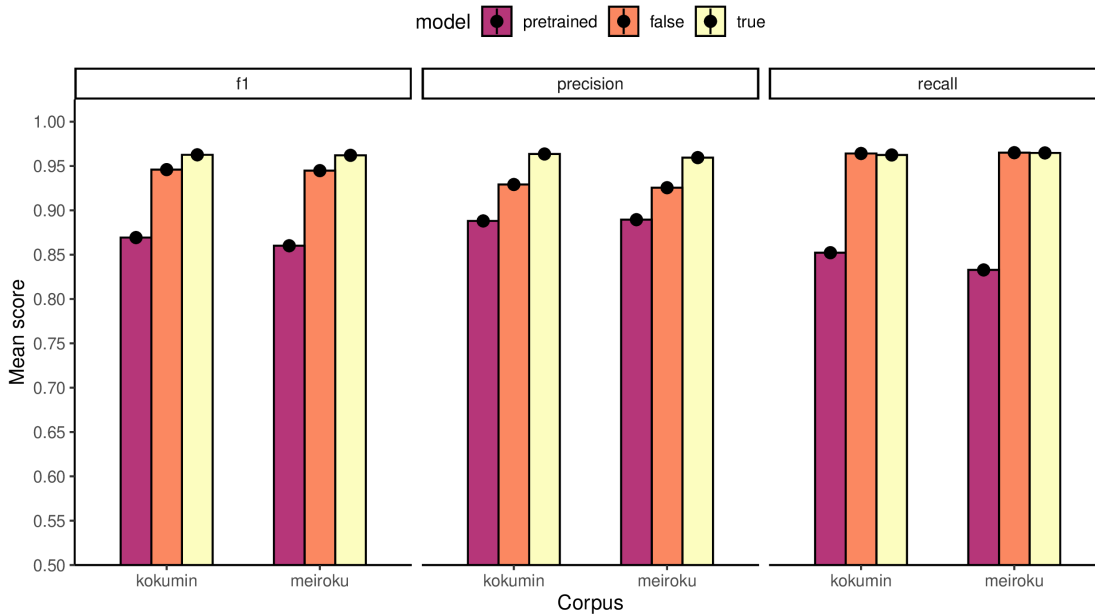


Figure 5. Mean precision, recall, and F1 scores on a word segmentation task on the Meiroku and Kokumin corpora using the tokenizer option in UDpipe and comparing against SUW annotations supplied by NINJAL. Three UDpipe set-ups are evaluated: a pretrained model (PRETRAINED), a model trained over 5-fold cross-validated GiNZA CoNLL-U output from Taiyo documents (FALSE), and a model trained identically but with addition of the proposed rule application approach (TRUE). Error bars are reported with 99% significance.

Significant improvements obtained using rule application are in precision and F1. For Kokumin, the “true” model gives a 3.4% improvement in precision and a 1.7% improvement in F1 over the “false” model, and a 7.5% improvement in precision and a 9.3% improvement in F1 over the pretrained model. For Meiroku, the “true” model gives a 3.4% improvement in precision and a 1.7% improvement in F1 over the “false” model, and a 7.0% improvement in precision and a 10.2% improvement in F1 over the pretrained model. We emphasize the improvements brought by precision as being most significant as tokenizations that are inaccurate often produce many tokens (i.e., “B” labels) that result in high recall but low precision.

40

5. Discussion

Indeed, some management of the rule set is needed to achieve an observable improvement. The proposed workflow is not totally automatic and care is needed to ensure rules that are introduced into the collection do in fact lead to more accurate parsings produced by GiNZA (or a pretrained tool of choice) and that overlapping rules do not produce a condition where multiple minima exist. Problems with the former usually present as rows with empty forms that can be detected with ease. Moreover, the amount of manual time needed for review is still reduced as the reviewer need only to concentrate review on the FORM field to obtain improved dependency parsings on historical materials.

41

While the proposed workflow has an effect on the CoNLL-U output produced by GiNZA and said effects bring a significant improvement in precision and F1 for “B”-label word estimation, the results also point toward a need for developing mechanisms that facilitate expansion of the rule collection that, in turn, furthers the changes made to the dependency structure in the CoNLL-U output; this has the potential to bring more improvement in the performance of trainable NLP pipelines like UDPipe on fundamental NLP tasks for historical materials.

42

Perhaps one step in this direction are NLP methods that can flag instances of pretrained output with inaccurate parsings that need review, thereby allowing easier rule introduction. Alternatively, another approach is to orient the research towards automatic rule inferencing that pave the path for rapid expansion of the current rule collection. One possibility is to allow for “rule chaining,” that is, the application of one rule that triggers the application of one or more new rules that were not directly applicable on the original sentence. Furthermore, methods from deep learning like the encoder-decoder model shown in [Ikeda et al. 2016] offer rich potential for automatic inferencing. These research directions are appealing, however, caution must be exercised to avoid developing methods that are exciting computationally but offer little to the DH scholars that make use of them [McGillivray et al. 2020]. Therefore, any NLP tool developed for the purpose of aiding analysis with historical materials must first work for the DH corpus at hand.

43

6. Conclusion

In this work we introduced a rule-based workflow for providing improved UD annotations to historical Japanese corpora. The principal advantage of our approach is that no “gold standard” data is required for training data development, only the availability of a pre-trained model in the target language. Moreover, the amount of time needed for post-editing pre-trained model output is significantly reduced as the reviewer need only to develop rules that address problems in the FORM field and the review does not require deep expertise in UD. We showed that this “cheaper” review strategy exhibits an effect on the dependency structure in the CoNLL-U output and, furthermore, brings an improvement in the performance of trainable language-agnostic NLP pipelines like UDPipe on word segmentation tasks.

44

These results are encouraging to DH scholars who would like to enhance their scholarship by annotating historical materials with linguistic metadata that is customizable and more reliable than what would be possible by the straightforward application of “off-the-shelf” tools. Future work will do well to further expedite the manual review needed to achieve good results on the target corpus by incorporating methods that flag pretrained output that is inaccurate and allow automatic rule inferencing. We also caution against the development of techniques that could be interesting for a venue in NLP but offers little for the scholar that would use said techniques on a target DH corpus.

45

Acknowledgments

We would like to thank the Department of Computer Science at the University of Miami for providing the computational resources necessary for running the experiments in this research. The work is in part supported by the National Science Foundation Grant CNS P2145800.

46

Notes

[1] The four collections make available metadata that classifies articles as either “colloquial” or “formal.” Because the goal of this work is to verify whether a rule-based approach can bring any improvement in UD annotations for historical text, the incorporation of colloquial works makes assessing the efficacy of the approach more difficult to gauge. Therefore, we make exclusive use of formal materials during all experiments presented here. Nevertheless, the rules constructed here (e.g., character substitution rules) are applicable to the colloquial portion and new rules can always be generated by manual analysis of the colloquial texts; these would not be in opposition to the current ruleset developed from

the formal texts.

[2] The underlying assumption in our review criteria is that, by concentrating our efforts on correction of the FORM field to the exclusion of all other fields, an updated parsing by GiNZA that yields an improved tokenization will also necessarily yield an improved dependency parse as well.

[3] The text header must also be updated during this procedure. However, this work is trivial to complete as we need only to replace the normalized sentence with the original source sentence

[4] Note that other fields from the returned CoNLL-U table are omitted for sake of presentation.

[5] Using normalized GiNZA output to inform the upper bound on row lengths is an underlying assumption of this step and there are scenarios where this can result in incorrect parsings. These are reviewed in the next section.

[6] The labeled attachment score (or LAS) is the percentage of nodes with correctly assigned reference to the parent node in the dependency tree. The morphologically-aware labeled attachment score (or MLAS) aims at cross-linguistic comparability of the scores. Finally, the bilexical dependency score (or BLEX) is similar to MLAS but also incorporates lemmatization into the analysis and aims to evaluate both dependencies and lexemes [Zeman et al. 2018].

[7] The remaining rows of the FORM field, as well as the other fields present in the CoNLL-U output, are omitted in the interest of brevity.

Works Cited

- Argamon and Olsen 2009** Argamon, S. and Olsen, M. “Words, Patterns and Documents: Experiments in Machine Learning and Text Analysis.” *DHQ: Digital Humanities Quarterly* 3.2 (2009).
- Asahara et al. 2018** Asahara, M., Kanayama, H., Tanaka, T., Miyao, Y., Uematsu, S., Mori, S., et al. “Universal Dependencies Version 2 for Japanese.” *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. Miyazaki, Japan. (2018).
- Bollman 2019** Bollmann, M. “A Large-Scale Comparison of Historical Text Normalization Systems.” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. (2019): 3885–3898.
- GiNZA 2021** GiNZA - Japanese NLP Library. Megagon Labs. <https://megagonlabs.github.io/ginza/>.
- Ikeda et al. 2016** Ikeda, T., Shindo, H., and Matsumoto, Y. “Japanese Text Normalization with Encoder-Decoder Model.” *Proceedings of the 2nd Workshop on Noisy User-generated Text*. Osaka, Japan. (2016): 129–137.
- Maekawa 2006** Maekawa, K. “Kotonoha. The Corpus Development Project of the National Institute for Japanese Language.” *Proceedings of the 13th NIJL International Symposium: Language Corpora: Their Compilation and Application*. (2006): 55–62.
- Maekawa et al. 2014** Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., et al. “Balanced corpus of contemporary written Japanese.” *Language Resources and Evaluation*, 48, 345–371, doi: 10.1007/s10579-013-9261-0 (2014).
- McGillivray et al. 2020** McGillivray, B., Poibeau, T., Fabo, PR. “Digital Humanities and Natural Language Processing: 'Je t'aime... Moi non plus'.” *DHQ: Digital Humanities Quarterly* 14.2 (2020).
- NINJAL 2021** NINJAL. National Institute for Japanese Language and Linguistics. <https://www.ninjal.ac.jp/english/>.
- Nivre et al. 2016** Nivre, J., de Marneffe, M-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, CD., et al. “Universal Dependencies v1: A Multilingual Treebank Collection.” *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia. (2016): 1659–1666.
- Qiu et al. 2020** Qiu, X., Sun, T., Xu, Y., et al. “Pre-trained models for natural language processing: A survey.” *Science China Technological Sciences*, 63, 1872-1897, doi: 10.1007/s11431-020-1647-3(2020).
- Scannell 2020** Scannell, K. “Universal Dependencies for Manx Gaelic.” *Proceedings of the Fourth Workshop on Universal Dependencies*. Barcelona, Spain. (2020): 152–157.
- Shirai et al. 2020** Shirai, R., Matsumura, Y., Ogiso, T., Komachi, M. (2020). “Machine Learning-based Sentence Boundary Detection for Modern Japanese Texts.” *Information Processing Society of Japan*, 61(2), 152–161.
- Straka and Sedlák 2021** Straka, M., and Sedlák, M. CoNLL-U Viewer. Universal Dependencies.

https://universaldependencies.org/conllu_viewer.html

Straka and Straková 2017 Straka, M. and Straková, J. "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe." *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada. (2017): 88-99.

Tanaka et al. 2012 Tanaka, M., Okajima, A., Ogiso, T., Ono, M., Kojima, S., Shimada, Y., et al. "Study on Documents and Meta-languages for Designing a Corpus of Modern Japanese." *Academic Repository of the National Institute for Japanese Language and Linguistics*, doi: <https://doi.org/10.15084/00002759> (2012).

Tange 2018 Tange, O. "GNU Parallel 2018." GNU Parallel 2018, doi: 10.5281/zenodo.1146014 (2018).

Zeman et al. 2018 Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., et al. "CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies." *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium. (2018):1–21.



This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.