Full Sentence Verb Identification and Conjugation Classification in Japanese

Theodore Chen
Boston University, Boston, USA

Introduction

Japanese verbs, unlike English verbs, pack a significant amount of information inside its structure without much use of outside auxiliary words. This makes conjugating verbs in Japanese incredibly simple, as in order to express a certain tense, voice, or level of politeness, one only needs to remember a set of rules that has minimal exceptions.

However, the opposite is not true. In order to identify and distinguish between certain types of verbs, an entire sentence needs to be accurately understood. Two completely different conjugations can appear identical, and without knowing the context of the sentence or prior statements, it cannot be identified.

Motivation

When learning Japanese, intuitively, more common and easily understood verb forms like the past tense and polite form are taught first. These verb forms and their meanings are self-contained: If you see a verb conjugated into the past tense form, it means it is in the past tense. Additionally, unlike in English, nouns are marked by particles, each of which has seemingly clear rules as to what they indicate. For example, [に] (ni) can be thought of as a when/where indicator. [が] (ga) can be thought of as a subject marker. [を] (wo) can be thought of as direct object marker. While some exceptions are present in these rules at this stage, it's limited to only a few specific situations.

Once potential form, passive form, causative form, and causative passive form come into the picture, several confusing exceptions are introduced. First, some verbs are identical in potential and passive form, making the context of a sentence necessary to determine the difference. Second, the rules surrounding particles also change. In causative, passive, and causative passive form, [に] (ni) can now also mark the subject of the verb (the one doing the action), or the one causing the subject to do an action. [が] (ga) is still a subject marker, but can also mark the object (when thinking of a sentence intuitively in English).

As a long-time Japanese language learner, I still continue to struggle with quickly understanding sentences using these more complex verb forms, so my goal is to find a way to generate labeled corpora to practice reading.

お姉さん　は　りんご　が　食べられた　。
oneesan　wa　ringo　ga　taberareta

”My sister was able to eat the apple.” (potential)

お姉さん　に　りんご　を　食べられた
oneesan　ni　ringo　wo　taberareta

”The apple was eaten by my sister.” (passive)

Figure 1: 「食べられた」is used in both sentences, despite different meanings in the sentences.

Research Objectives

I will use one corpus for training: the "Tanaka Corpus".

- RQ_1: What is the baseline performance training and testing when using a Logistic Regression model?
- RQ_2: What is the performance training and testing when using a BiLSTM RNN Model?

Labeling Sentence Corpora

There is a lack of datasets that are labeled with the conjugation type of Japanese verbs, making it necessary to automate a way to label existing un-labeled datasets with a set of established rules.

To do this, I created a script that would automatically identify and classify verbs from a sentence using a set of rules. Building off an existing conjugation script, the process to do so is as follows: Tokenize a sentence using the spaCy library. Identify the verbs in the sentence, saving the lemma (root/dictionary) form, full form, its conjugation rule, the presence of the particle [に] (ni) in the prior sentence segment, and the presence of [を] (wo) or [が] (ga) as the noun marker directly before the verb. Conjugate/inflect the verb into all possible forms. Compare these forms with the full verb form, and classify the verb with the one that matches. Return the full list of verbs found and their types. This script was then applied to the Tanaka Corpus.

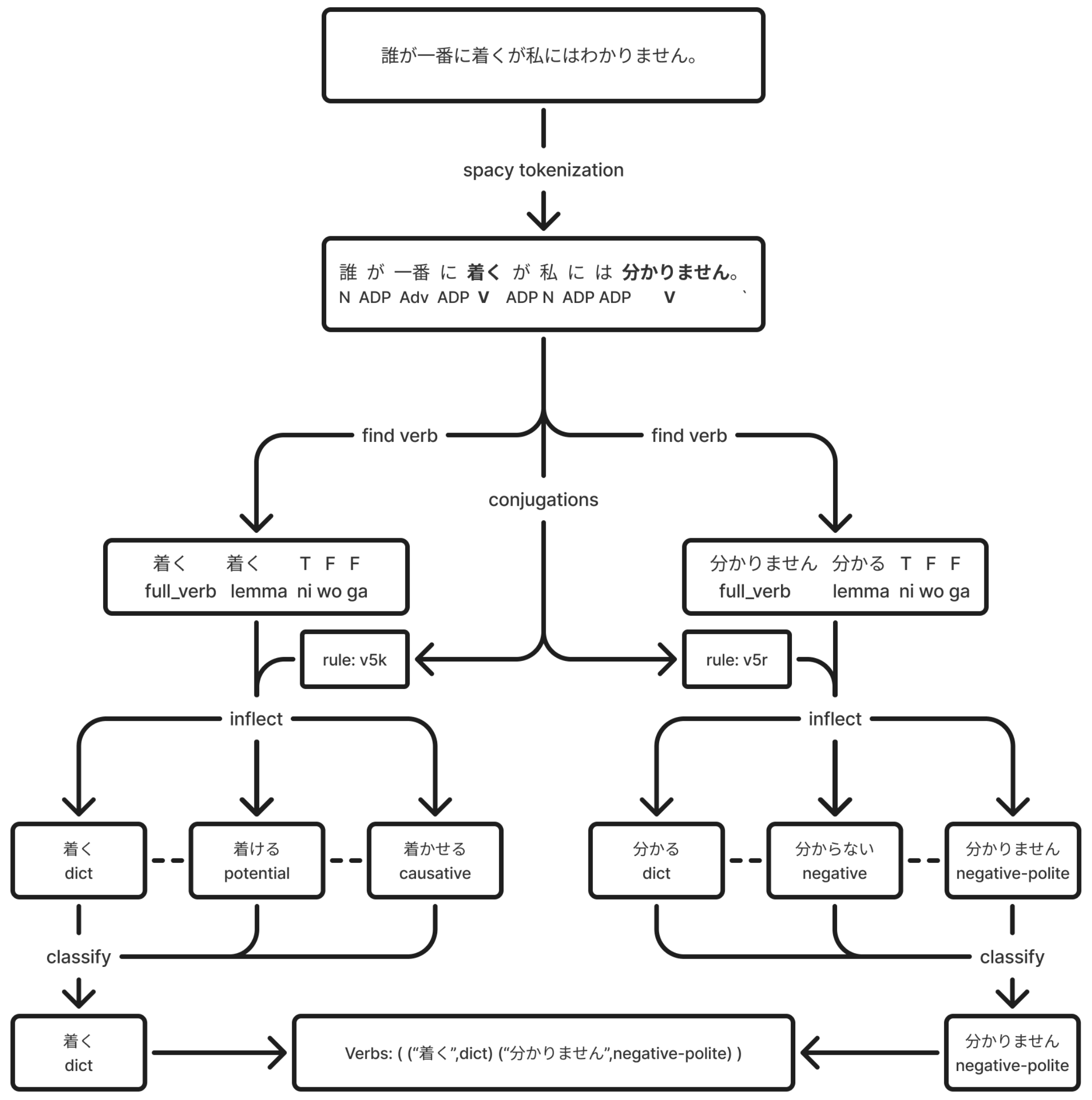
誰が一番に着くが私にはわかりません。
spacy tokenization
誰 が 一番 に **着く** が 私 に は **分かりません**。
N ADP Adv ADP **V** ADP N ADP ADP **V**
find verb
find verb
conjugations
着く 着く T F F
full_verb lemma ni wo ga
分かりません 分かる T F F
full_verb lemma ni wo ga
rule: v5k
rule: v5r
inflect
inflect
着く dict
着ける potential
着かせる causative
分かる dict
分からない negative
分かりません negative-polite
classify
classify
着く dict
Verbs: ((“着く”,dict) (“分かりません”,negative-polite))
分かりません negative-polite

Figure 2: A flow chart showing the process taken by the script to classify verbs given a sentence.

When comparing all possible verb forms, the information of the presence of [に] (ni), [を] (wo), and [が] ga are used to distinguish between the forms found in Figure 1.

Dataset Details

The dataset post-script contains 107,980 successfully labeled sentences out of a total of 147,865, with 171,363 individual verbs defining sentence segments that could be used for training. The models were trained with a training/test split of 80/20.

Model Details

Logistic Regression model:

Input: each segment s is mapped via character n-grams (1–3) to a feature vector

$$\mathbf{x} = \phi(\mathbf{s}) = \mathrm{CountVectorizer}_{\textbf{char,1-3}}(\mathbf{s}) \in \mathbb{R}^d.$$

Training objective: maximize the L_2–regularized log-likelihood with regularization $C = 1.0$:

$$\mathcal{L}(\mathbf{W}, \mathbf{b}) = \sum_i \log P\big(y^{(i)} \mid \mathbf{x}^{(i)}\big) \; - \; \frac{1}{2C}\|\mathbf{W}\|_F^2.$$

BiLSTM model:

Input: each segment $\mathbf{s} = (c_1, \ldots, c_T)$ of up to $T = 50$ characters is mapped to indices $i_t = \text{char\_to\_idx}(c_t)$

Embedding:

$$\mathbf{x}_t = \mathbf{E}\big(i_t\big) \in \mathbb{R}^{d_e}, \quad d_e = 128.$$

Training objective: minimize the multiclass cross-entropy loss

$$\mathcal{L} = -\sum_{i=1}^{N} \log P\big(y^{(i)} \mid \mathbf{s}^{(i)}\big),$$

with learning rate $1\mathrm{e}{-3}$, batch size 64, for 10 epochs.

Hyperparameters: Vocabulary size $|\mathcal{V}| = |\text{char\_to\_idx}| + 1$, embedding dim 128, hidden dim 256, layers $L = 2$, dropout 0.5, max length 50.

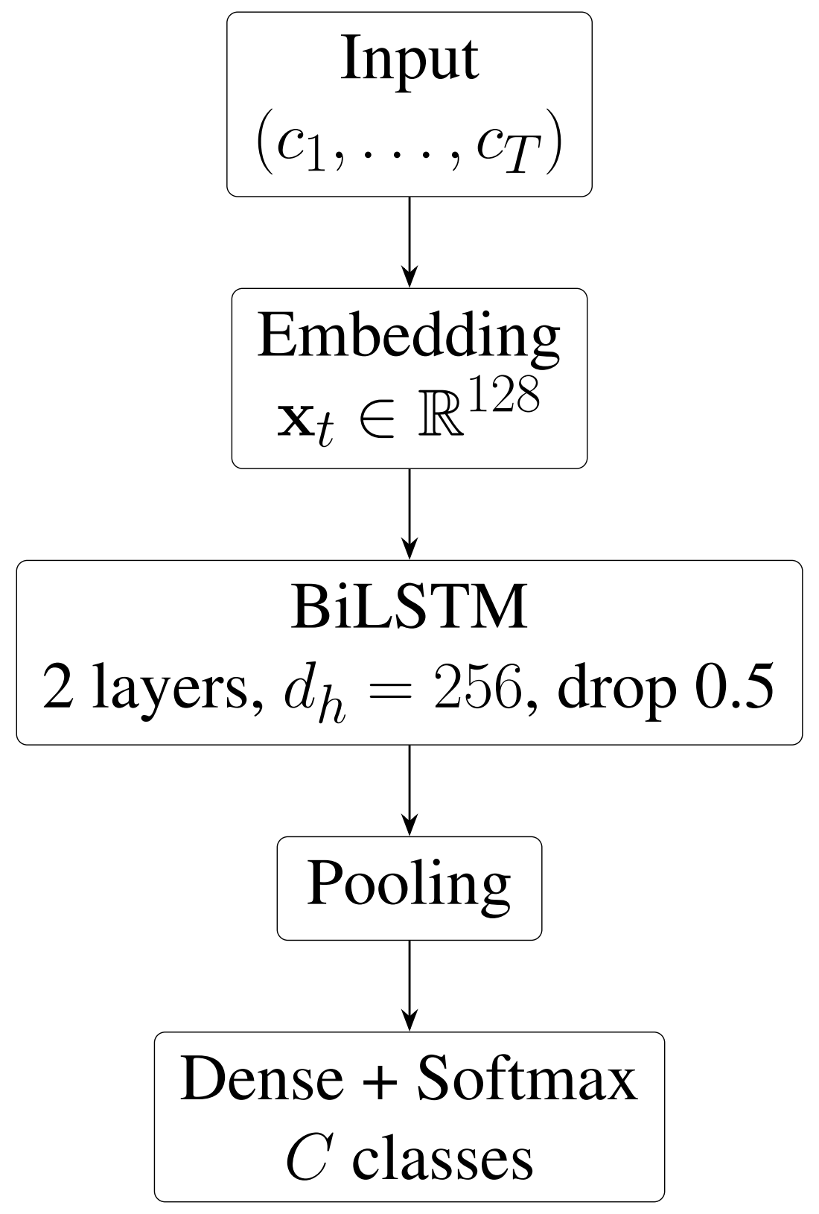
Input
$(c_1, \ldots, c_T)$
Embedding
$\mathbf{x}_t \in \mathbb{R}^{128}$
BiLSTM
2 layers, $d_h = 256$, drop 0.5
Pooling
Dense + Softmax
C classes

Figure 3: BiLSTM structure

Experiments & Results

The difficulty with classifying Japanese verbs lies in verb forms that require context, so these results will be focused on the performance of the Logistic Regression Model and the BiLSTM Model on the causative, passive, and potential forms.

Group	**Classes**	**Support**	**Macro F1**	**Weighted F1**
Causative	19	301	0.643	0.849
Potential	9	36	0.464	0.595
Passive	21	1173	0.590	0.933
Other	24	32770	0.915	0.983

Table 1: Baseline Group Analysis

Group	**Classes**	**Support**	**Macro F1**	**Weighted F1**
Causative	19	301	0.990	0.987
Potential	9	36	0.926	0.906
Passive	21	1173	0.900	0.983
Other	24	32770	0.993	0.999

Table 2: BiLSTM Group Analysis

In addition to an improvement on the average F1 score across all classes, classes that were previously unclassified due to a low support value were still able to be classified.

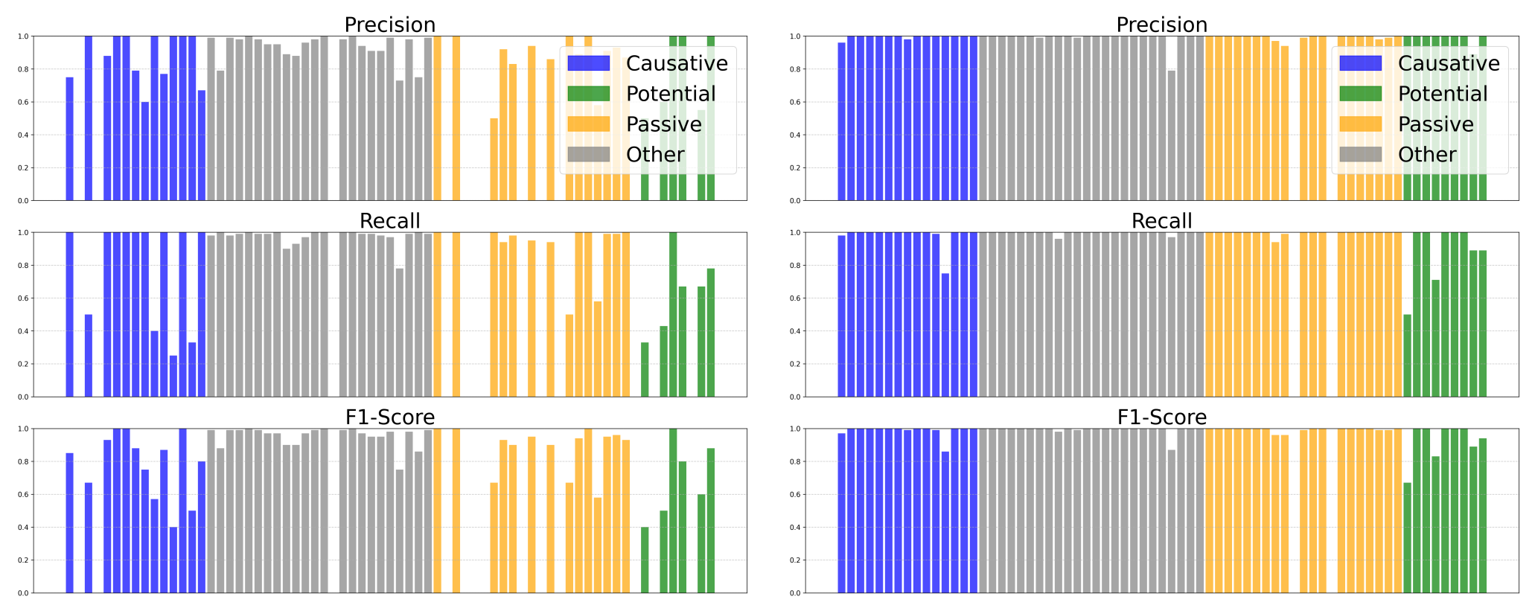

Figure 4: Charts of Precision, Recall, and F1 Score for all verb classes for RQ_1 (left) and RQ_2 (right).

Conclusion

I found that my initial intuition was correct, the use of a model that is able to take in the full context of a sentence performs better than a model that only observes the verb and a few surrounding characters.

In the future, I aim to fine-tune the hyperparameters of the BiLSTM Model (they were chosen arbitrarily), improve the existing classification script in Figure 2 , and add classification for [です] (desu) ”to be” and compound verbs.

One potential alternative to rectify inconsistencies between potential and passive form classification when labeling the dataset is instead of using particle information, translate the sentence to English first, then label based off of that.