Making Use of Furigana

Gary Kacmarcik
Microsoft Research
Redmond, WA
garykac@microsoft.com

Abstract

An interesting aspect of written Japanese that has not been well studied is the use of furigana, or reading cues, to assist linguistic processing of text. Difficulties in processing this material have led to the situation where it is sometimes considered more convenient to simply remove the parenthetical material rather than to process it. This paper describes a system that makes use of the furigana to assist with various tasks, including segmentation, word sense disambiguation and support for OOV items. The system reports an F-measure score of 93.3% on the task of matching the base text with its furigana.

1 Introduction

Furigana are the pronunciation guide characters that are used to assist the reader when difficult, ambiguous or rare Kanji characters (typically those outside the Joyo Kanji List of common characters) are presented in text. They usually take the form of hiragana (phonetic) characters, and in typeset text, they are printed above or to the side of the base text being described using a smaller font, as in the following example:

こ し た ん た ん
虎 視 顧 々

Since this richly structured arrangement for assigning furigana is not an option in many documents, various other methods for encoding furigana are used, as summarized in Table 1.

Table 1: Some examples\(^2\) of furigana markup: (A) are examples of the standard parenthesized form; (B) are in an interchange format suggested in [JIS-4052]; (C) are in the format proposed in the most recent W3C document [W3C_Ruby01].

<table>
<thead>
<tr>
<th></th>
<th>A 質（ただ）した</th>
<th>B ん（たど）って</th>
<th>C &lt;rubygroup&gt;黄&lt;ruby&gt;とら&lt;/ruby&gt;&lt;/rubygroup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ぬ &quot;千尋／ちひろ&quot;</td>
<td>走 &lt;rubygroup&gt;黒&lt;/ruby&gt;&lt;rubygroup&gt;赤&lt;/ruby&gt;&lt;rubygroup&gt;黄&lt;/ruby&gt;</td>
<td>&lt;rubygroup&gt;赤&lt;/ruby&gt;&lt;rubygroup&gt;黒&lt;/ruby&gt;</td>
</tr>
</tbody>
</table>

In this paper, we deal exclusively with the cases of furigana in parentheses as shown in (A) of Table 1. While this format is not the most complete (since it does not identify the start of the furigana block) it is by far the most common and the other formats are trivially converted into it.

Though furigana (and parenthetical materials in general) have not been well studied in the area of Japanese text processing, we show that they provide information that (when available) can be quite valuable. We continue with a discussion of the benefits of processing furigana in assisting various tasks, including segmentation disambiguation, word sense disambiguation and the recognition of out-of-vocabulary (OOV) items. This paper finishes with a discussion of how furigana is handled by our system, along with an evaluation.

\(^1\) koshitantan - eyeing covetously

\(^2\) (A) tadahita = verified; koukou = mimetic for the shining of the sun; (B) Chihiro = NAME; tadotte = following; tora = tiger, the third sign of the Chinese zodiac; (C) aru = some
2 Types of Parenthetical Material

Parentheses are used for a broad range of applications in natural language, but this paper will be focusing on how they are used for furigana.

For our purposes, it is useful to divide parenthetical material into two broad categories based on how they are generally handled by a parsing system: TOKENIZER parentheses and GRAMMAR parentheses.

TOKENIZER parentheses are those that are typically handled during the tokenization process. For example, in English: “Camping with your dog(s).” or “Please briefly explain expense(s) and attach proof/receipt”.

GRAMMAR parentheses, on the other hand, are those that are typically handled by the grammar and do not present a problem for tokenization (for example, this sentence).

In English, the vast majority of parenthetical material is of the GRAMMAR type; the TOKENIZER parentheses are somewhat rare and majority of the instances are simply adding a final “(s)” to identify an ambiguously singular/plural noun.

Like the English examples given above, the parenthetical material in Japanese can also be divided into tokenizer and grammar types. However, there are a few crucial differences between how parentheses are used in English and Japanese.

First, the lack of spaces in Japanese text means that it is more difficult to identify a priori parentheses that need to be handled by the tokenizer versus those to be handled by the grammar.

Second, the TOKENIZER class of parenthetical material is far larger in Japanese than in English because this is a common method for providing furigana. As shown in detail in Section 3, furigana can be inserted word-externally to provide a reading for an inflectional stem.

This difficulty in processing furigana has prompted some to simply remove the furigana and other parenthetical materials prior to text processing. However, parenthetical material, both the TOKENIZER and GRAMMAR types, provide important information for text processing that can be used to improve performance.

3 Furigana

As mentioned in the introduction, furigana characters provide pronunciation cues for rare or difficult kanji. This section provides additional details about how furigana is used in Japanese text and proposes a classification scheme for furigana when applied to kanji.

Note that furigana can sometimes be used to provide readings for katakana, for example in texts geared toward younger audiences, or even archaic hiragana (as in やあ（やえ）). We do not discuss these variations in this paper because they do not typically occur in the types of corpora that we’re focusing on and because a katakana-hiragana match/detection algorithm is trivial to implement.

3.1 Types of furigana

Furigana can be separated into three distinct classes: PARTIAL WORD, FULL WORD and MULTI-WORD.

PARTIAL WORD furigana have the interesting (and confounding) property that they can occur in the middle of word units (as in 妖（あや）しく or 踊（こ）線橋 or at the end (as in 研鑽(さん))). They typically provide a reading for a single kanji character, but may apply to multiple characters. A single word may have multiple PARTIAL WORD furigana blocks.

FULL WORD furigana give a reading for an entire word, which may be one or more kanji characters.

MULTI-WORD furigana span word boundaries and typically identify phrasal units like proper names or book titles. MULTI-WORD furigana can also make use of the nakaten (・) character within the furigana string to identify word boundaries, as in 山根俊英(やまね・としひで) where the nakaten identifies the boundary between the family name and the given name.

It is worth noting that the PARTIAL WORD class is larger than it might initially seem because of the

---

3 For example, Kyoto University Text Corpus ([Kurohashi97] and [Kurohashi99]) consists of sentences after all parenthetical material has been removed.

4 Yae = NAME
5 ayashiku = dubious
6 kosenkyo = an overpass
7 kensan = a study
8 Yamane Toshihide = NAME
convention of placing the furigana immediately after the kanji. For inflected verbs and adjectives that end with a hiragana inflectional ending, the desire to place the furigana next to the kanji forces the furigana to be placed word internally.

3.2 Frequency

Depending on the corpus being analyzed, the frequency of furigana can range from being fairly frequent (e.g.: in the Shincho corpus of novels where roughly half of all sentences have furigana) to practically non-existent (as in spoken dialog or chat-room transcripts).

<table>
<thead>
<tr>
<th></th>
<th>Mainichi</th>
<th>Shincho</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of clean sentences</td>
<td>893,693</td>
<td>126,145</td>
</tr>
<tr>
<td># of sentences with (...)s</td>
<td>145,297</td>
<td>64,507</td>
</tr>
<tr>
<td>Percentage</td>
<td>16.3%</td>
<td>51.1%</td>
</tr>
<tr>
<td>Total # of (...)s</td>
<td>195,709</td>
<td>211,913</td>
</tr>
<tr>
<td># of all-hiragana (...)s</td>
<td>16,368</td>
<td>196,103</td>
</tr>
<tr>
<td>Percentage</td>
<td>8.4%</td>
<td>92.5%</td>
</tr>
</tbody>
</table>

Table 2: Summary of parenthetical material in Mainichi 1995 newspaper and Shincho novel corpora.

An analysis of the Mainichi 1995 newspaper corpus reveals that roughly 16% of clean sentences contain parenthetical material and that more than 8% of all parentheses contain furigana readings for the preceding kanji character(s).

While this is not an overwhelming percentage, it is also not insignificant. In addition, when furigana does occur, it is typically added to resolve ambiguity or to identify difficult or rare words. These are cases where additional information can be quite useful.

3.3 Identifying furigana

Given a parenthetical expression in a sentence, it is relatively straightforward to determine whether or not the expression is furigana for the preceding kanji characters.

The simple heuristic of tagging any parenthetical that contains only hiragana (and nakaten) characters achieves 98.3% precision (see Table 3) with 100% recall (F-measure = 99.1%).

This can be improved to 99.9% precision by adding the additional constraint that forces the character immediately preceding the left parenthesis to be a kanji character. Not surprisingly, increasing the precision with the kanji restriction has a negative effect on recall with 22 examples of furigana being lost. This reduces the recall value to 99.9% (F-measure = 99.9%).

4 Using furigana to Improve Analysis

The main purpose of using the furigana characters is to improve the performance of our parsing system. The following sections describe ways in which our system can benefit from the additional information that the furigana provides.

4.1 Assisting Segmentation

There are a few ways in which segmentation can be assisted by furigana.

The most obvious is with respect to word internal furigana (for example in 着 (か) ける). Without furigana analysis, the word (着ける in this case) will not be identified as such and will result in serious segmentation problems.

In addition, because of the manner in which our system performs segmentation, we can use the furigana to improve our segmentation precision. Our segmentation phase provides a maximal-recall word lattice to our parser, which is then responsible for determining the correct path through the lattice (see [Suzuki00]). Anything that we can confidently remove from this lattice improves our system overall. For example, if our segmenter encounters 独楽 we will also produce individual words for 独 and 楽 (which the parser will later eliminate). If we instead encounter 独楽 (こま), then we can eliminate these subwords immediately.

This information can also be used to provide hints to the parser about the boundaries of the structures that should be created. For example, in 正月元日歌 (むときつきたちのみようた) the

---

9 51.1% of all sentences have parenthetical hiragana; of which we estimate more than 99% are actually furigana.
10 Our simple definition of a “clean” sentence is that it must end with a “...” character.
11 8.36% of parentheses are all-hiragana, and 98.3% of all-hiragana parentheses are furigana for kanji (see Table 3).
12 Only 18 examples out of the 16,368 were not furigana, and an additional 4 contained furigana but did not properly match the immediately preceding characters.
13 4 hiragana, 4 katakana, 7 number and 7 roman.
14 kakeru = to wager
15 koma = top. The two subwords are doku = Germany and raku = ease.
16 mutsukitsukita no yomiuta = New Year's Day song of joy.
parsable can be told to prefer structures that coincide with the given furigana boundaries.

4.2 Sense Disambiguation

When there are multiple senses associated with a word, the furigana can be used to determine the author's intended sense. In the sentence:

心願成就のお札（ふだ）を買った。  

the 札 character can be read either satsu (= bank note) or fuda (= card). If the furigana were not present, then this ambiguity would need to be preserved and resolved via further analysis. However, with the furigana, the sense can be correctly determined during the segmentation phase.

4.3 Handle OOV

Out of Vocabulary (OOV) items are a serious issue for Japanese because the problem is exacerbated by the lack of spaces in the text - it is hard to know the extent of the OOV word and so neighboring words tend to get disrupted along with the OOV word.

For example, in the sentence:

喪主は妻か志子（かじこ）さん。  

the boundary of the proper name Kajiko would be very difficult to determine heuristically. Using the information contained within the furigana allows the correct name boundary to be determined in a straightforward manner.

![Figure 1](image.png)  

Figure 1: Graphical representation of [長:なが][!居:い], the orthography lattice for the noun 長居 (nagai = "a long visit"). This allows the following forms: 長居, なが居 and ながい, but disallows 長い (which is confusable with much more common adjective 長い (nagai = "long").

5 Method

To enable our segmenter to handle parenthetical furigana, we made use of orthography lattices (as described in [Kacmarcik00]) that we initially employed to handle spelling variations.

5.1 Orthography Lattices

These lattices concisely encode all possible surface forms for each lexical entry. Each component of the lattice consists of a BASE part (typically kanji) and a READING part (always hiragana). An example orthography lattice [長:なが][!居:い] is depicted graphically in Figure 1. More complex lattice examples that support a wider range of phenomenon (like okurigana) are described in [Kacmarcik00].

The standard method for using this lattice involves working from right to left and following the connections as long as the either the BASE or the READING matches the input text stream. When the left side is reached, a new token is created for the word lattice that will later be passed to the parser for further analysis.

5.2 Handling Partial/Full Word furigana

To handle PARTIAL and FULL WORD furigana, the model was adapted to verify the consistency between the text and the furigana characters.

Our system handles this by accumulating furigana characters (working from right to left) until the matching open parenthesis is encountered. At that point, it walks the two strings in parallel and ensures that the furigana string follows the READING while the text matches either the BASE or the READING. A valid match is one where the furigana boundary coincides with an orthography lattice boundary. In this way we can ensure that the text is consistent.

To support word internal furigana, we simply allow these parenthetical expressions to start at any lattice boundary.

In this fashion, our system can support a wide variety of furigana forms. Using the lattice given above as an example, support is provided for the following forms: 長居(ながい), 長(なが)居(い), 長(なが)居, 長居(い) in addition to other (unlikely) forms like なが居(い). For each of these strings, an orthographically normalized word (長居) is passed to the grammar component for further analysis.
5.3 Handling MULTI-WORD furigana

For MULTI-WORD furigana the basic algorithm needed to be extended to apply across multiple words. In essence, this is simply stringing the lattices from multiple words together, but there are two aspects of MULTI-WORD furigana that make them more interesting to work with.

First of all, it is possible for MULTI-WORD furigana to contain readings for hiragana characters that occur between kanji. These characters can act as anchors to facilitate identifying how the furigana maps to the text being described. For example, the の [no] in 「三日の餅（みかのもち）」, serves as an anchor between the reading for the first part and the second part. These anchors can be of great use when attempting to match OOV items.

Additionally, there are cases where extra characters (typically の [no], which roughly corresponds to 'of' in English) are inserted into the furigana that are not present in the base text. In 正月元日慶歌（むつききたちのよみうた） and 安倍晴明（あべのせいまい） the の does not directly correspond to any of the kanji characters and is inserted to facilitate readability.

5.4 Handling OOV

The method described above works when furigana is applied entries that occur in our lexicon, but it does not handle OOV items. To properly handle this class of furigana, a different approach was needed.

From the orthography lattices, a separate yomi (reading) table was created with an entry for each recognized kanji that listed all possible readings for that character. If the initial furigana match fails to find a match against the headwords in our lexicon, this yomi table is used to identify matches between the BASE and READING strings.

Where even this expanded yomi table was insufficient, we introduced guessing based on the surrounding characters. We differentiated between STRONG and WEAK guesses based on the how much supporting context the guess had.

A STRONG guess is one that has an anchor character that establishes a known boundary. These anchors can be characters for which we've already identified a reading (as in 彩挙（さいえん）, where the yomi table entry [彩.さい] provides the anchor for the [挙:えん] guess) or they can be external to the match (for example, in 語（まう）づ, where the hiragana に [ni] establishes the start of the base text being described).

In contrast, a WEAK guess is one without at least one supporting anchor. An example of this is 中村朗生（はるお） where there is no clear indication whether the furigana applies to the final one, two or three preceding kanji. Surrounding word context can sometimes help, but in this case each of the four individual kanji can be a valid word on their own.

6 Evaluation

We tested the improvements on the 16,368 occurrences of parenthetical hiragana extracted from the Mainichi 1995 newspaper corpus. Since we were interested primarily in assigning readings to kanji, we essentially ignored the 287 occurrences where the preceding character was not kanji. Note that our system normalizes some special characters like 々 and 々 into the appropriate kanji character before processing, so instances of these characters (67 and 1 respectively) are reflected in the kanji totals. In addition, the 34 instances of the 々 were also included in the kanji totals.

Our first version used the readings extracted from the orthography lattices in our primary lexicon. We were able to match 80.4% of the furigana using the information from this lexicon. Merging in the information from our proper noun lexicon (which contains primarily low frequency

---

19 三日 [mikka = third day] 餅 [mochi = rice cake]. Note the archaic spellings in the furigana.
20 matsukitsukitachi no yomiuha = New Year's Day song of joy
21 Abe Seimei = (NAME) a famous Heian period astronomer and fortuneteller
22 saien = an artistic decoration style
23 moudzu = archaic form of 語（もう）でる = mouderu = to go to worship at a shrine
24 Nakamura Haruo = NAME
25 々 is replaced with a copy of the preceding kanji and 々 is replaced with 綿.
26 getaji - used as a placeholder for unprintable kanji
27 Containing roughly 70,000 headwords that provide 13,697 unique kanji readings.
28 Providing an additional 10,673 kanji readings.
entries and is thus not used for general parsing) raised to this 87.5%.

Introducing guessing allowed us to significantly improve the number of matches with the caveat that precision was reduced by incorrect guesses.

An examination of a sampling of our STRONG guesses reveals that they are reasonable about 85.5% of the time and produce the correct word segmentation result (even with an incorrect guess for an individual character) an additional 4.8% of the time, resulting in roughly 90% guessing precision.

Using F-measure as an evaluation metric for these models, the non-guessing model scores at 93.3% and the anchored guessing model comes in slightly lower at 91.5%.

7 Conclusions

The information contained in furigana can be quite useful when processing Japanese and should not be ignored or removed. When presented within parentheses, it is straightforward to detect furigana and distinguish it from other parenthetical material with high accuracy.

We have also shown that an orthography lattice representation of lexical items can be quite useful for handling furigana. Our system makes use of the information extracted from the furigana to improve it’s handling of proper nouns and other OOV items.

8 Future Work

We briefly experimented with feeding the STRONG guesses back into our yomi tables to see if there was any improvement. We found that while we matched more furigana, we were not satisfied with the accuracy of the guesses. However, additional experimentation may allow us to improve this.

In addition, we would also like to experiment more with passing the boundary hints of multi-word furigana to the grammar to see if the system can benefit from this additional information.

9 Notes

All Japanese examples in this paper were taken directly from the Mainichi 1995 newspaper corpus. The examples in Table 1 were also originally from Mainichi but were modified to demonstrate the various furigana markup formats.

Table 3: Cumulative recall (%) for matching furigana

<table>
<thead>
<tr>
<th>Description</th>
<th>Items</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of hiragana (...)'s</td>
<td>16,368</td>
<td></td>
</tr>
<tr>
<td>(...)'s following non-kanji</td>
<td>287</td>
<td>1.7%</td>
</tr>
<tr>
<td>Matching main + PN readings</td>
<td>13,160</td>
<td>80.4%</td>
</tr>
<tr>
<td>Matching main readings</td>
<td>14,316</td>
<td>87.5%</td>
</tr>
<tr>
<td>With Anchored (Strong) Guesses</td>
<td>15,233</td>
<td>93.1%</td>
</tr>
<tr>
<td>With Unanchored (Weak) Guesses</td>
<td>15,999</td>
<td>97.7%</td>
</tr>
</tbody>
</table>

References


