Increasing the error coverage of the FOKS Japanese dictionary interface

Slaven Bilac†, Timothy Baldwin∗ and Hozumi Tanaka†

† Tokyo Institute of Technology 2-12-1 Ookayama, Meguro Tokyo, JAPAN
{sbilac,tanaka}@cl.cs.titech.ac.jp

∗ Center for the Study of Language and Information (CSLI)
210 Panama Street
Stanford, CA 94305-4115, USA
tbaldwin@csli.stanford.edu

Abstract

With the advent of electronic dictionaries, significant progress has been made in improving the accessibility of dictionary entries allowing for speedy and wide-ranging dictionary lookups. Nonetheless, very little work has been done in trying to accommodate user errors or supplement knowledge deficiencies in dictionary lookups. These two factors are common causes of failed searches and user frustration, especially so for learners of a foreign language trying to use a dictionary as a learning aid.

We have developed FOKS, a system aimed at learners of the Japanese language, which works on removing the requirement of knowing the prescriptively correct reading of the word in order to be able to look it up in a dictionary. Users can enter their best guess as to the reading of the word and have the system suggest likely corresponding dictionary entries. While calculating the candidate entries, FOKS considers common reading errors and sources of reading confusion, and displays dictionary entries that could give rise to the reading.

In this paper we describe how we increase the error coverage of the FOKS system by adding a mechanism to handle reading errors based on character co-occurrence and common substrings. We use a corpus of general Japanese texts to extract the character/substring confusion sets and assign a probability measure to each pair in the confusion set. Based on the extracted sets we can generate novel readings for each dictionary entry and assign a plausibility of such readings appearing in a dictionary query. The generation process is constrained with various thresholds to limit the overall number of readings generated. We add the resulting reading set to the existing FOKS system and evaluate the overall change in performance on the data obtained from the Japanese Proficiency Test collection as well as the effect of the increased number of readings on the overall number of candidates displayed to the user.

1 Introduction

Learning a foreign language is a time-consuming and painstaking process, and made all the more daunting by the existence of unknown words (Groot 2000). Without a fast, low-cost way of looking up unknown words in a dictionary, the learning process is impeded (Humble 2001). This is particularly true in non-alphabetic languages such as Japanese, as there is no easy way of looking up the component characters of new words. This research attempts to alleviate the dictionary look-up bottleneck by way of a comprehensive dictionary interface which allows Japanese learners to look up Japanese words in an efficient, robust manner.

The Japanese writing system consists of the three orthographies of hiragana, katakana and kanji, which appear intermingled in modern-day texts. The hiragana and katakana syllabaries, collectively referred to as kana, are relatively small (46 characters each), and most characters take a unique and mutually exclusive reading which can easily be memorized. Kana thus do not present a major difficulty for the learner. Kanji characters (ideograms), on the other hand, present a much bigger obstacle. The high number of these characters (1,945 prescribed by the government for daily use, and up to 3,000 appearing in newspapers and formal publications (NLI 1986)) in itself presents a challenge, but the matter is further complicated by the fact that each character can and often does take on several different and frequently unrelated readings, and that readings undergo morpho-phonological changes in the word formation process. The kanji 発, for example, has readings including *hatsu*[1] and *ta(tsu)*, whereas 表 has readings including *omote*, *hyou* and *arawa(reru)*. Learners presented with the string 発表 *happyou* "announcement"[2] for the first time will, therefore, have a possibly large number of potential readings (conditioned on the number of component character readings they know) to choose from. With paper dictionaries, look-up typically occurs in two forms: (a) directly based on the reading of the entire word, or (b) indirectly in a kanji dictionary via component kanji characters and an index of words involving those kanji. Clearly in the first case, the correct reading of the word must be known in order to look it up, which is often an unreasonable assumption. In the second case, the complicated radical and stroke count systems make the kanji look-up process cumbersome and time consuming.

With electronic dictionaries – both commercial and publicly available – the options are expanded somewhat. In addition to reading- and kanji-based look-up, for electronic texts, simply copying and pasting the desired string into the dictionary look-up window gives us direct access to the word.[3] Several reading-aid systems (e.g. Reading Tutor[4] (Kitamura & Kawamura 1998; Kawamura et al. 2000) and Asunaro[5] (Nishina et al. 2000; Nishina et al. 2002)) provide greater assistance by segmenting longer texts and outputting individual translations for each segment (word). While these dictionaries and reading aides are a welcome addition to the learner's repertoire, they provide little help to the user when the text is not available in electronic form. To deal with texts available only in hard copy the user still needs to input the word into the dictionary interface. It is often possible to use kana-kanji conversion to manually input component kanji, assuming that at least one reading or lexical instantiation of those kanji is known by the user. Essentially, this amounts to individually inputting the readings of words the desired kanji appear in, and searching through the candidates returned by the kana-kanji conversion system. Again, this is complicated and time inefficient so the need for a more user-friendly dictionary look-up method remains. Finally, many electronic dictionaries support the use of regular expressions (REGEXPs) in searches, enabling lookup of words when partial input is possible. However, such queries often result in a large number of alphabetically (or phonetically) ordered responses, making it hard to locate the desired entry even it when it is included as one of the responses.

In order to allow the user to maximize the use of available knowledge of kanji characters and their readings, and remove the requirement that the user possesses the correct reading knowledge of the word she is trying to lookup, we have implemented the FOKS (Forgiving Online Kanji Search) system. The system is a web-based facility that allows the user to enter the estimated reading of a novel word. Based on the input reading the system calculates the dictionary entries that could be perceived as taking that reading and displays the candidates for the user to choose from. Since the (minimum) input is a single field (i.e. the estimated reading) , the system can readily be used by small hand-held devices such as web-enabled cellular phones. Furthermore, the system can easily be integrated with a voice recognition/input system to completely remove the need to type the input.[6]

Once the candidate entries are displayed to the user she can easily select the target word from the list of candidates to obtain the translation of the word. For example, the user can search for the string 頭上 *zujou* "overhead" by inputting the reading *toujou* or *atamajou*, derived from more common readings of the characters 頭 and 上, *tou*/*atama* and *jou*, respectively. We have previously demonstrated that this

---

[1]In this paper, we loosely follow the Hepburn system of romanization, with the exception that we romanize long vowels as separate characters giving rise to *hyou* instead of *hyoo* or *hyō* for ひょう. The other notable divergence—taken from (Backhouse 1994)—is the use of the upper-case *N* for syllable-final nasals (corresponding to the kana ん) and lower-case *n* for syllable-initial nasals (as found in な, for example).

[2]Here, *hatsu* undergoes gemination and *hyou* sequential voicing to produce *happyou*.

[3]Although even here, life is complicated by Japanese being a non-segmenting language, putting the onus on the user to correctly identify word boundaries.

[4]`http://language.tiu.ac.jp/`

[5]`http://hinoki.ryu.titech.ac.jp/`

[6]The limiting factor here is the accuracy of the speech recognition system.

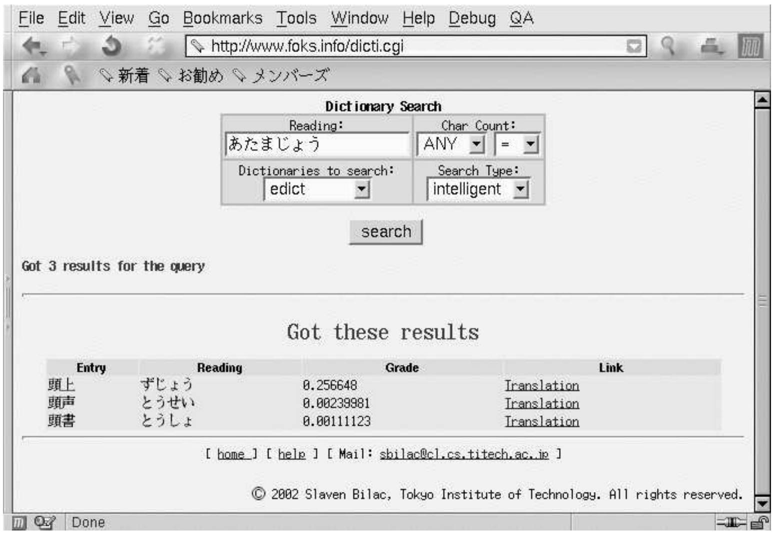

Figure 1: Results displayed by FOKS system as a response to query *atamajou*

system is effective in guiding the user to the target dictionary entry even when queried with an incorrect reading (Bilac et al. 2003).

In this paper we describe how we expand the error-handling ability of the FOKS system by adding awareness of errors induced by character/substring co-occurrence. This is the type of error where learners replace the reading of an unknown kanji character by the reading of a known character due to common kana suffix or common context. For example, a user might try to read 慰める *nagusameru* "console, comfort" as *osameru* due to knowledge of string 納める *osameru* "dedicate, offer". In a similar fashion, a learner familiar with the commonly-occurring 訴訟 *soshou* "lawsuit" may construct the erroneous reading *kishou* for 起訴 *kiso* "(to) charge". We start out with a corpus and extract confusion sets of character/substring pairs. Then we create novel (erroneous) readings for all dictionary entries containing any members of the confusion set and integrate it with the set generated for the previous version of the FOKS system.

The remainder of this paper is structured as follows. Section 2 describes the previous version of the FOKS system and the error types that it is able to handle, and also points out some remaining problems. Section 3 describes how we extend the error coverage through the creation of confusion sets and generation of novel readings to account for previously unhandled error types. Finally, Section 4 provides an analysis and evaluation of the system.

2 System description

The FOKS system was implemented at the Tokyo Institute of Technology as a means of improving dictionary accessibility for learners of the Japanese language. It is based on the notion that learners acquire Japanese character readings gradually, starting with the most common characters and readings and then moving on to rarer ones. Due to such ordering of the learning process they might be unable to construct the prescriptively correct reading for a novel string, even though familiar with some (or all) of the characters contained in the string.

The majority of electronic dictionaries require accurate reading knowledge in order to look up a desired dictionary entry, since they employ direct matching sometimes augmented with REGEXPs (see above) as the lookup criteria. Unlike most other dictionary interfaces, the FOKS system does not assume correct reading knowledge of the target string, but instead tries to estimate what string the user is looking for based on the input reading. The system judges the plausibility (in the form of a compositional probability) of each reading-dictionary pair based on the probability of each kanji character taking a particular reading and the overall reading undergoing further morpho-phonological changes. The corpus frequency is then combined with the calculated probability to produce the overall plausibility measure of the reading given the desired dictionary entry.

The base dictionary for the FOKS system is the publicly-available EDICT Japanese-English electronic dictionary.[7] We extracted all entries containing at least one kanji character and created novel (potentially erroneous) readings, which we scored for plausibility as described above. Corpus frequencies calculated over the complete set of 200,000+ sentences in the EDR Japanese corpus (EDR 1995) were used to produce the final plausibility measure. Once the complete set of readings is generated, it is stored in a relational database and queried through CGI scripts. Since the readings and scores are pre-calculated, there is no time overhead in response to a user query. Figure 1 depicts the system output for the query *atamajou*.[8] We can see that the target string 頭上 is the highest ranking candidate, meaning that the user can easily access its translation even though the initial query was based on an incorrect reading. Notice that conventional electronic dictionaries employing a direct match search criterion would not return any candidates since *atamajou* is not a valid word in Japanese. The system is available for public use and easily accessible through any Japanese language-enabled web browser. Currently, only a Japanese-English dictionary is included but it would be a trivial task to add links to translations in alternative languages once the dictionaries are available. Furthermore, since the whole generation and scoring process is fully automatic it would be straightforward to apply the same process to a different Japanese dictionary.[9]

2.1 Error handling ability

The version of the FOKS system as described above is able to handle a wide range of common errors by learners of the Japanese language. Among these is the substitution of one kanji character reading for another, resulting in an incorrect overall reading. For example, the FOKS system allows the user to access 外科 *geka* "surgery" via the more salient (but incorrect) reading of *gaika*. Here the learner has applied one of the standard readings for 外 *gai, soto* "outside,outer" to obtain the overall reading and queried the system accordingly to obtain the translation. While querying a conventional system would result in an unsuccessful search, the FOKS system guides the user to the translation directly based on this incorrect reading. Additionally, the system can handle queries where the user has incorrectly predicted phonological alternation (resulting in queries like *hahhyou* or *hatsuhyou* when searching for 発表 *happyou* "announcement") or where the learner is confused as to the correct vowel length of the character reading (resulting in queries such as *ryoukou* for 旅行 *ryokou* "travel").

However, in evaluation of the Japanese Proficiency Test data (see below) we were able to pinpoint systematic error types that we are currently unable to handle. Namely, we were able to isolate three error types:

1. Errors due to character co-occurrence. The learner applies a reading from a kanji character occurring in the same context. E.g., a user unfamiliar with the character 激 *geki, hage(shii)* "violent, fierce" might apply a known reading of the character 厳 *geN, kibi(shii)* "strict,severe" due to the common kana suffix (i.e. 激しい *hageshii* "violent" vs. 厳しい *kibishii* "strict").

2. Errors due to character-level semantic similarity. Characters like 右 *migi* "right" and 左 *hidari* "left" have a similar meaning and as such are easily confused by learners, resulting in an erroneous reading. Semantic confusion sometimes occurs at the word level, too, such as between 火事 *kaji* "fire" and 火災 *kasai* "(disastrous) fire".

[7] `http://www.csse.monash.edu.au/~jwb/edict.html`

[8] This is a screenshot of the system as it is visible at `http://www.foks.info/`.

[9] For example, we have employed the same methodology to allow searching ENAMDICT, a proper name dictionary in EDICT distribution, on the same principles.

3. Errors due to graphic similarity of characters. This is a common error type in learners from non-kanji language backgrounds. For example, 墓 *bo, haka* "grave" and 基 *ki, moto* "base" are graphically very similar, resulting in possible substitution of readings among words containing them (e.g. 墓地 *bochi* "graveyard" and 基地 *kichi* "base").

3 Extending the error coverage

In this paper we concentrate on the first of the three error types outlined above, since it accounts for a significant portion of the failed queries in our evaluation data. The main cause of this error is users associating characters with the context in which they commonly occur, and consequently being unable to distinguish the context from the reading knowledge. In this section we describe how we extract possible candidates for this type of error and create additional (erroneous) readings for dictionary entries to enable the system to handle errors due to character co-occurrence.

3.1 Extracting the character confusion sets

We capture the context similarity of characters by looking at their corpus occurrence in different word contexts, in order to model character co-occurrence tendencies as perceived by a learner dealing with Japanese texts. Learners often associate character reading knowledge with their context. In other words, they find it easy to identify readings in a fixed context, but become confused when the context changes. We aim to capture such contextual effects and use them to extend the error coverage of our system.

We employ Mutual Information (MI) as the measure of correlation between two characters. It is calculated according to equation 1, where the numerator and denominator represent the probability of two characters occurring together and independently, respectively.

$$MI(a,b) = log_2 \frac{P(a,b)}{P(a)P(b)} \tag{1}$$

Roughly speaking, it is a measure of how much one character tells us about the other. The probabilities of each event in equation 1 are calculated based on Maximum Likelihood Estimation, as calculated over the EDR Japanese corpus.

We calculate MI for each pair of units found in a given word (e.g. 厳 *kibi* and しい *shii* in 厳しい *kibishii* "strict"), and retain all pairs of units with an MI value over an experimentally-determined threshold value of θ. After repeating this process for all words, we merge the confusion pairs into confusion sets by taking each kana unit occurring in a confusion pair as an index and merging all kanji character units it is confusable with. Confusion pairs consisting of kanji only are merged into confusion sets indexed on each kanji. Within each confusion set, we normalize the probabilities of the members to sum to 1, based on the corpus frequencies of the source words.

3.2 Generating novel readings

Having extracted the confusion sets, we proceed to generate and score the various readings for plausibility. We first segment each dictionary entry up into single kanji characters and kana character sequences. Then, for each confusion pair contained in the word, we replace each member with alternate readings of the characters in the confusion set indexed on its counterpart. For example, given a confusion pair (納 *osa*, める *meru*) derived from the word 納める *osameru* "obtain" and dictionary entry 慰める *nagusameru* "comfort" segmented into units as 慰 *nagusa* and める *meru*, we would create a novel reading of *osameru* for 慰める due to a common pivot element (i.e. める *meru*).[10] The overall probability of each generated reading is obtained under the assumption of segment independence by calculating the product of the probabilities assigned to each member of the confusion set and the original probability of the reading.

[10] Since (慰 *nagusa*, める *meru*) is not a pair in our sample confusion set we would not create a *nagusameru* reading for 納める.

	Conventional	FOKS	COOC	FOKS + COOC
Number of Queries	1189	1189	1189	1189
Average Number of Results	2.3	11.0	4.5	14.1
Successful Queries	18	547	56	581
Error Reduction (%)	0	45.2	3.5	48.1
Mean Rank	1.6	1.8	10.1	2.9

Table 1: Performance comparison on the Japanese Proficiency Test data

We exhaustively generate readings for all dictionary entries in this manner using the confusion data. The resulting set is then stored in the relational database to be queried by the user.

4 Evaluation

Finding an appropriate test set was a significant problem. Many publications on Japanese language education mention common learner errors (MEIJI 1997), but to our knowledge there is no readily available database containing real-life examples of learner reading errors. We thus opted to use the Japanese Proficiency Test data (Suzukawa & Katori 1996; Matsuoka 1995) for evaluation. The Japanese Proficiency Test is offered by the Japanese government as a means for learners of Japanese to evaluate their language ability. In the vocabulary component of the test, subjects are asked to select the correct reading for a kanji-containing word out of four candidates (only one of which is correct). The incorrect reading candidates are carefully crafted to exploit reading errors common in learners of the language. Therefore, we feel this set is appropriate for evaluating the effectiveness of the FOKS system.

We divide evaluation into two parts. In the first part we evaluate the character co-occurrence module (COOC) independently to select the best threshold parameters. For the second part, we combine the best version of the COOC data with the original FOKS system data and evaluate the overall change in performance. We are particularly interested in the relative change in the number of dictionary entries correctly retrieved when queried with an incorrect reading (given as successful queries in Table 1), the average number of results returned for each query, and the rank of the entry in the candidate listing.

We extracted a set of 1189 incorrect readings from the level 2 Japanese Proficiency Test sample data and ran queries over different implementations of the system. The results are given in Table 1. Here, the conventional system represents direct match over the EDICT dictionary. We can see that, based on a threshold value of $\theta = 3.6$, the COOC module increases the number of incorrect readings handled by 34[11] without incurring an excessively large increase in the number of results generated (+3.1).[12] The mean rank of the resulting system is 2.9 showing that when the target is in the candidate listing it is high enough to be easily located. As such, the addition of the COOC module should increase the usefulness of the FOKS system to learners of Japanese.

5 Conclusion and future work

The FOKS system is a Japanese dictionary interface aimed at removing the presupposition of infallible reading knowledge in looking up words and encouraging the user to maximally use available knowledge. In this paper we have explained how we expand the error handling ability of the system to include the effects of character co-occurrence. We extract potential confusion pairs from corpus data based on mutual information and generate novel readings for dictionary entries containing such potentially confusing characters, scoring the readings with a plausibility measure in the process. The generated

[11] Notice however that the number is higher for the COOC module alone, showing that we get some overlap in generated readings. In such cases we combine the plausibility scores.

[12] It is possible to extend coverage by reducing the value of the θ threshold, but the mean rank tends to increase appreciably with relatively little gain in coverage.

readings have been added to the FOKS system to extend its coverage. Initial evaluation shows that the number of successful searches based on incorrect readings increases when using the improved system.

In Section 4, we saw that while we increased the number of successfully-handled queries, the system still failed to return the desired dictionary entry in a large number of cases. In order to further improve our system, we need to evaluate user data and analyze real-world patterns of reading errors. To this effect, we are collecting query data from our web server and intend to use it to perform extensive analysis. The reading generation and scoring procedure can be adjusted by adding and modifying various weight parameters to alter the calculation of probabilities, and thus tweak the system output to maximize dictionary accessibility.

Above, we identified two other major types of error commonly appearing in query data that our system currently does not handle, namely errors due to graphic or semantic similarity of kanji. In the future, we would like to expand our model to incorporate these factors.

References

Backhouse, A. E.: 1994, *The Japanese Language: An Introduction*, Oxford University Press.

Bilac, S., T. Baldwin & H. Tanaka: 2003, 'Improving dictionary acessibility by maximizing use of available knowledge', *Traitement automatique des langues*, **44:2**.

EDR: 1995, *EDR Electronic Dictionary Technical Guide*, Japan Electronic Dictionary Research Institute, Ltd., (In Japanese).

Groot, P. J. M.: 2000, 'Computer assisted second language vocabulary acquisition', *Language Learning & Technology*, **4**(1): 60–81.

Humble, Ph.: 2001, *Dictionaries and Language Learners*, Haag + Herchen.

Kawamura, Y., K. Tatsuya & R. Hobara: 2000, 'Development of a reading tutorial system for JSL and JFL learners using the EDR Japanese-English dictionary', *Japan Journal of Educational Technology*, **24**: 7–12, (In Japanese).

Kitamura, T. & Y. Kawamura: 1998, 'Construction of Japanese reading support environment enabling independent study', (In Japanese).

Matsuoka, T.: 1995, *Problems from Japanese Proficiency Test, characters and vocabulary (Levels 1 and 2)*, Kokusyo Kankoukai.

MEIJI, Meiji Publishing Planning/Editing Group: 1997, *Analysis of misuse of Japanese Language*, Meiji Publishing, (In Japanese).

Nishina, K., M. Okumura, S. Sugimoto, Y. Yagi, T. Abekawa, N. Totsugi & F. Ryang: 2000, 'Development research on multilingual Japanese reading aid for foreign students with scientific background', *Research Report of Telecommunications Advancement Foundation*, **15**: 151–159, (In Japanese).

Nishina, K., M. Okumura, Y. Yagi, N. Totsugi, F. Ryang, S. Sugimoto & T. Abekawa: 2002, 'Development of Japanese reading aid with a multilingual interface and syntax tree analysis', in *Proc. of the Eight Annual Meeting of The Association for Natural Language Processing (NLP2002)*, pp. 228–231, (In Japanese).

NLI: 1986, *Character and Writing system Education*, vol. 14 of *Japanese Language Education Reference*, National Language Institute, (in Japanese).

Suzukawa, K. & F. Katori, eds.: 1996, *Japanese Proficiency Test Preparation Measure, characters and vocabulary (Level 2)*, Kokusyo Kankoukai.