Evaluating the FOKS Error Model*

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)
Stanford, CA 94305-4115, USA
tbaldwin@csli.stanford.edu

Abstract

Learners of Japanese face great difficulty when trying to lookup words containing kanji in a dictionary, due to the requirement of knowing the correct reading of the target word. We propose a system that imitates the cognitive process learners go through in generating readings for novel kanji strings, and provide direct access to the dictionary entries based on the generated readings. In doing so we remove the correct reading requirement. The system described here is implemented in a web-based environment and freely available for general use. In this paper we provide an analysis of query and error data collected by our server.

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. 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 (broadly divided into *on* readings of Chinese origin and *kun* readings of Japanese origin: Backhouse, 1994). These readings often undergo morpho-phonological changes, such as gemination and sequential voicing, in the process of word formation (Tsujimura, 1996). The kanji 発, for example, has several 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 Japanese paper dictionaries, look-up typically occurs in two forms: (a) directly based on the reading of the entire word, and/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. Quite often, this is an unreasonable assumption. In the second case, the complicated radical and stroke count systems make the kanji look-up process cumbersome and time consuming. For example, to look up 遷移 *seNi* "transition" without knowing the correct reading the user needs to look up each character individually (i.e. look up 遷 via its radical 辶 or stroke count of 15, and 移 via its radical 禾 or stroke count of 11).

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] and Asunaro[5]) 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

*See Bilac et al. (2004) for an extended version of this paper.

[1] In this paper, we follow the romanization style used in Bilac et al. (2003).

[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/

the kana-kanji conversion system. Again, this is complicated and time inefficient, hence 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 (Breen, 2000). However, such queries often result in a large number of responses, making it hard to locate the desired entry even when it is included in the system output.

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 he 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.

Once the candidate entries are displayed, the user can easily select the target word from the list 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 system is effective in guiding the user to the target word even when queried with an incorrect reading (Bilac et al., 2003).

In this paper we provide an analysis of query data collected by our dictionary server in an attempt to evaluate the adequacy of the error model used to predict (that is generate and score) erroneous readings and evaluate the effectiveness of the system in leading the user to the dictionary entry based on the incorrect reading.

The remainder of this paper is structured as follows. Section 2. describes common error types and their causes. Section 3. describes the current version of the FOKS system and the error types that it is able to handle. Finally, Section 4. provides an analysis and evaluation of the system performance.

2. Common reading errors

Previously Bilac et al., (2003) proposed a classification of common learning errors according to several basic types. While this classification was adequate for constructing the system, we felt that a more fine-grained classification was necessary to describe the errors actually appearing n the query data. Accordingly, we classify the error types as given in Table 1. As can be seen from this table, a larger number of causes can affect the derivation of an incorrect reading for a target entry. Quite commonly, several causes combine simultaneously, making the classification difficult. In Section 4., we look into the observed distribution of these error classes, but first, we introduce the FOKS system in greater detail.

3. 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 less frequent 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.

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 likelihood score) 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 score of the reading given the desired dictionary entry.

The system is built under the assumption that the cognitive process a user goes through in deriving a reading for a novel string is the following: (a) for each kanji in the word postulate a plausible (possibly erroneous) reading; (b) form an overall reading for the word by combining the individual readings of all components; and (c) when necessary, apply any phonological/morphological changes to the overall reading to get the final reading postulate. Depending on the proficiency of the learner, the number of choices available at each step varies.

The system imitates this cognitive process by first calculating the probabilities of each reading given a kanji character and the probabilities of various morpho-phonological changes affecting the overall reading based on dictionary data and a training corpus (Bilac et al., 2002). Then, for each dictionary entry we apply the extracted readings and their calculated probabilities to generate novel readings which we score with a likelihood measure based on the above probabilities and corpus frequencies. The score is assigned under an assumption of segment independence, thus failing to take into account the interaction of various readings and phonological changes. Although this deviates slightly from the observations of the researchers (Itô and Mester, 1995; Frellesvig, 1995) it simplifies the calculations significantly and should still be adequate for modelling learners of the language.

Currently, FOKS handles all the error types given in Section 2. except for types 7 and 12, although the handling of types 8 and 14 is limited.

3.1. Implementation Details

The base dictionary for the FOKS system is the publicly-available EDICT Japanese-English electronic dictionary.[6] We extracted all entries containing at least one kanji character and created a set of 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 obtain the final plausibility measure. Once the complete set of readings is generated, it is stored in a relational database and queried through CGI

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

Type No.	Description	Example
1	*Inadequate choice of kanji reading*	大会 *taikai* “convention” misread as *ookai* or *daikai*
2	*Vowel length confusion*	主催 *shusai* “organization” misread as *shuusai*
3	*Inadequate palatalization*	亜流 *aryuu* “epigone” misread as *aruu*
4	*Incorrect voicing*	団塊 *dankai* “baby boom” misread as *dangai*
5	*Incorrect gemination*	脱出 *dasshutsu* “escape”, misread as *datsushutsu* or *dashutsu*
6	*Other phonological*	春雨 *harusame* “spring rain” misread as *haruame*
7	*Due to graphic similarity of characters*	墓地 *bochi* “graveyard” confused with 基地 *kichi* “base”
8	*Due to grapho-phonetic character similarity*	闇 *yami,in* “darkness” misread as *on* or *oto* due to 音 *oto,on* “sound”
9	*Due to character co-occurrence*	激しい *hageshii* “violent” misread as *kibishii* due to the common suffix
10	*Proper nouns (personal and place names)*	弘前 [*Hirosaki*], where 前 *mae,zen* “front” has an unconventional reading
11	*Idiomatic expressions*	珈琲 *koohii* “coffee” where reading and meaning do not correspond
12	*Character-level semantic similarity*	火事 *kaji* “fire” confused with 火災 *kasai* “(disastrous) fire”
13	*Inadequate kana content*	滑稽 *kokkei* “comical” misread as *suberukei* due to 滑る *suberu* “(to) slide”
14	*Other*	塵芥 *chiriakuta* “garbage” (inexplicably) misread as *chiNke*

Table 1: Common Error Types

scripts. Since the readings and scores are pre-calculated, there is no time overhead in response to a user query.

In addition to reading-based search, the system provides the means to limit the search space by various additional constraints (number of characters, prefix, etc.). Also, the users can search in “simple” mode whereby the candidates are selected based on direct match with only the correct readings, effectively reducing the search ability to that of a conventional dictionary system.

The system is available for public use and easily accessible through any Japanese language-enabled web browser at `http://www.foks.info`.

4. Evaluation

The FOKS website was initially made accessible to the public in November 2001. Since that time we have collected logs for all queries to our system (94,180 queries to date). Based on the sequence of user input, queries can be divided into two groups: **full queries** with input reading and target dictionary entry pairs recorded[7] and **partial queries** where it is not possible to determine the target entry.[8] The latter group is ignored for the purposes of evaluation since we only aim to analyze how the system models user errors at the reading-level. We thus use only the full queries in evaluation.

The complete set of full queries contains 5,820 input/target entry tokens. Of these, we analyzed 4,675 token pairs comprising 2,658 distinct types. In 2,076 cases (1,158 distinct types), or 44.4% of the data, the input reading is not the correct dictionary reading of the target entry. The high percentage of queries with an erroneous reading clearly shows that the ability to handle reading errors helps the user get to the target entry in a large number of cases.

[7]Here, the user entered a reading and subsequently selected the target entry from the list of candidates displayed.

[8]Here, the user either queried the system with a string containing kanji characters or a regular expression to obtain the translation directly, the user was only interested in the reading and did not click through to the translation, or the target entry was not available in the candidate listing.

Target Entry	Incorrect Input	Frequency
市井 *shisei* “town”	*ichii*	120
乱高下 *raNkouge* “fluctuations”	*raNkouka*	99
山車 *dashi* “festival car”	*yamasha*	78
幕間 *makuai* “intermission”	*makukaN*	58
⋮	⋮	⋮
返戻 *heNrei* “giving back”	*heNmodosi*	1
工作 *kousaku* “construction”	*kosaku*	1
地点 *chiteN* “spot”	*jiteN*	1
一息 *hitoiki* “one breath”	*ichii*	1

Table 2: Example full queries

Table 2 gives some examples of stored pairs with their occurrence frequency.

The remaining 2,599 pairs (1,594 distinct types) were cases where the input reading was the correct dictionary reading of the target entry. In such cases our system returned 12.9 candidates on average and the target entry mean rank was 3.2. This shows that the number of candidate entries is low enough that the correct ones are not obscured in the list and competent users are not unduly penalized.

As a second step, we looked at each unique input/target pair where there the input reading did not correspond to the correct reading of the target entry (1,158 pairs in all), and classified each entry as corresponding to one or more of the error types listed above. The error analysis was conducted by a teacher of Japanese as a Foreign Language with over 30 years of teaching experience. Table 3 gives the most common error types with representative examples. We can see that (uniquely) type 1 errors are the most common, accounting for 61.1% of the data, and an additional 18.8% when instances classified according to multiple error types are factored in. From this we can conclude that choosing the correct reading for a character based on the available context is the hardest problem the learner has to deal with. Such error distribution is adequately reflected in the reading set we generate as the majority of the probability mass is assigned to type 1 errors.

From Table 3 we can see that instances of multiple error

Error Type	%	Example Query and Target String
1	61.1	*yasai*→ 家裁 *kasai* “family court”
1,5	4.3	*shukkatsu*→ 祝勝 *shukushou* “victory celebration”
2	3.8	*soushi*→ 阻止 *soshi* “obstruct”
5	3.8	*dakkai*→ 大会 *taikai* “convention, congress”
1,11	3.3	*mizunaitsuki*→ 水無月 *minazuki* “June”
14	3.2	*chiNke*→ 塵芥 *chiriakuta* “garbage”
1,4	3.2	*kijiN*→ 着信 *chakushiN* “arrival”
1,2	3.0	*shoiji*→ 笑い事 *waraigoto* “laughing matter”
4	2.6	*taNji*→ 端子 *taNshi* “terminal”

Table 3: Most common error combinations in input queries

Error Type	Frequency	%
1	926	79.9
5	101	8.7
2	87	7.5
4	84	7.3
11	48	4.1
14	37	3.2
10	16	1.4
3	15	1.3
6	14	1.2
13	13	1.1
7	9	0.8
8,9,12	3	0.3

Table 4: Individual error types by frequency

classification are relatively frequent (19.5% of all types), underlining the effectiveness of our model at modeling the effects of compound errors. To ascertain the relative impact of the individual error types, we calculated the proportion of queries for which a given error type was evident, as given in Table 4.[9] From this, we can once again see that error type 1 (reading choice) is the most prevalent source of reading confusion, but also that phonology (error types 4, 5 and 6) leads to errors in 17.2% of the cases. Note the appearance of error types 7 and 12 in Table 4 despite there being no explicit handling of them in the reading generation process. This is due to other error types conspiring to produce readings which happen to coincide with the effects of graphic and semantic similarity, respectively.

For 3.2% of input readings, our judge could not determine the source of reading error (and hence assigned error type 14). There are two possible explanations for this. One is that the error model used in our system allows for the application of various sources of confusion in a layered fashion, ultimately masking the individual error types. The other possible explanation is that foreign language learners sometimes obtain and store reading knowledge based on its context and self-derived rules which native speakers (even teachers of the language) cannot readily identify. However, this hypothesis would need further exploration on larger data sets.

The most significant shortcoming of our evaluation is that we only have full queries in cases where our system offered the target entry as a candidate. The failure of the system to return the target entry can be ascribed to: (a) the target entry not being in the dictionary, (b) the system not handling the error type present in the query, or (c) the provided input being too ill-motivated to make the connection with the target entry. Currently, we have no way of differentiating these three effects, and moreover, we are unable to determine which of the partial queries were successful (but the translation was not accessed) and which were not.

[9] Percentages do not add up to 100% since a single input can involve more than one error type.

5. Conclusion and future work

The FOKS system is a Japanese dictionary interface aimed at removing the presupposition of complete and correct reading knowledge in the word look-up process. By allowing access to dictionary entries based on (predictably) incorrect readings, FOKS encourages the user to maximally use available knowledge. In this paper we analyzed the error distribution in actual input data obtained from system query logs. The system logs contain initial input and target entry pairs which were classified according to the error type (if any) appearing in the initial reading input. Using this data we demonstrated the effectiveness of the FOKS probabilistic model in reading-error modeling.

Finally, we identified two major types of error common in learners of Japanese but currently not handled adequately: errors due to graphic or semantic similarity of kanji. In the future, we would like to expand our model to incorporate handling of these factors.

Acknowledgements

We wish to thank Prof. Kikuko Nishina of Tokyo Institute of Technology International Student Center for classifying the query data, and Michael Zock for help in writing this paper.

6. References

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

Bilac, S., T. Baldwin, and H. Tanaka, 2002. Bringing the dictionary to the user: the FOKS system. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*. Taipei, Taiwan.

Bilac, S., T. Baldwin, and H. Tanaka, 2003. Improving dictionary accessibility by maximizing use of available knowledge. *Traitement automatique des langues*, 44:2.

Bilac, S., T. Baldwin, and H. Tanaka, 2004. Modeling learners’ cognitive processes for improved dictionary accessibility. Technical Report TR04-001, Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan.

Breen, J., 2000. A WWW Japanese Dictionary. *Japanese Studies*, 20:313–317.

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

Frellesvig, B., 1995. *A Case Study In Diachronic Phonology, The Japanese Onbin Sound Changes*. Aarhus University Press.

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

Itô, J. and R. A. Mester, 1995. Japanese phonology. In J.A. Goldsmith (ed.), *The Handbook of Phonological Theory*, chapter 29. Blackwell, pages 817–838.

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

Tsujimura, N., 1996. *An Introduction to Japanese Linguistics*. Cambridge, Massachusetts: Blackwell, 1st edition.