Construction of a Japanese Learner-Friendly Dictionary Interface

Slaven Bilac[†], **Timothy Baldwin**[‡] **and Hozumi Tanaka** [†]
[†]Tokyo Institute of Technology <{sbilac,tanaka}@cs.titech.ac.jp>
[‡]CSLI, Stanford University <tbaldwin@csli.stanford.edu>

Abstract

We propose a method which allows learners to look up words according to their expected, but not necessarily correct, reading. In preprocessing, we calculate the possible readings each kanji character can take and different types of phonological and conjugational changes that can occur, and associate a probability with each. Using these probabilities and corpus based frequencies we calculate a plausibility measure for each generated reading given a dictionary entry, based on the naive Bayes model. In response to a user-entered reading, we calculate the plausibility of each dictionary entry corresponding to the reading and display a list of candidates for the user to choose from. We have implemented our system in a web-based environment and are currently evaluating its usefulness to learners of Japanese.

1 Introduction

Unknown words are a major bottleneck for learners of any language, due to the high overhead involved in looking them up in a dictionary. 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. While the proposed method is directly transferrable to other language pairs, for the purposes of this paper, we will focus exclusively on a Japanese–English dictionary interface.

The Japanese writing system consists of the three orthographies of hiragana, katakana and kanji, which appear intermingled in modern-day texts (NLI, 1986). The hiragana and katakana syllabaries, collectively referred to as kana, are relatively small (46 characters each), and each character takes a unique and mutually exclusive reading which can easily be memorized. Thus they do not present a major difficulty for the learner. Kanji characters (ideograms) 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) 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. The kanji 発, for example, has readings including *hatu* and *ta(tu)*, whereas 表 has readings including *omote*, *hyou* and *arawa(reru)*. Based on simple combinatorics, therefore, the kanji compound 発表 *happyou* "announcement" can take at least 6 basic readings, and when one considers phonological and conjugational variation, this number becomes much greater. Learners presented with the string 発表 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 in, e.g., looking the word up in a conventional dictionary.

Japanese dictionary look-up typically occurs in two forms: (a) directly based on the reading of the entire word, or (b) indirectly via an individual component kanji and an index of words involving that kanji. Clearly in the first case, the reading of the word must be known in order to look it up, which is often not the case. In the second case, the complicated radical and stroke count systems make the kanji look-up process cumbersome. We attempt to resolve these failings by: (a) allowing access to the desired dictionary entry via an incorrect reading, and (b) supporting regular expression-based queries, such that it is possible to look up a novel kanji combination as long as at least one kanji is known (and can be input into the dictionary interface).

This paper describes a system that allows a learner to use his/her knowledge of kanji to the fullest extent in looking up unknown words according to their expected, but not necessarily correct, reading. Learners are exposed to certain kanji readings before others, and quickly develop a sense of the pervasiveness of different readings. We attempt to tap into this intuition, in predicting how Japanese learners will read an arbitrary kanji string based on the relative frequency of readings of the component kanji, and also the relative rates of application of phonological processes. An overall probability is attained for each candidate reading using the naive Bayes model over these component probabilities. Below, we describe how this is intended to mimic the cognitive ability of a learner, how the system interacts with a user and how it benefits a user.

In Proceedings of the Eighth Annual Meeting of the Association of Natural Language
Processing (Japan), Keihanna, Japan, pp. 439-42.

The remainder of this paper is structured as follows. Section 2 describes the preprocessing steps of reading generation and ranking. Section 3 describes the actual system as is currently visible on the internet. Finally, Section 4 provides an analysis and evaluation of the system.

2 Reading Generation and Grading

In order to generate a set of plausible readings we first extract all entries containing kanji and their respective readings and for each entry perform the following steps:

1. Segment the kanji string into minimal units and align each resulting unit with its corresponding reading. We have successfully adopted the TF-IDF algorithm to this purpose as is described in Baldwin and Tanaka (2000).

2. Perform conjugational, phonological and morphological analysis of each kanji–reading pair and standardize the reading to canonical form. In particular, we consider gemination (*rendaku*) and sequential voicing (*onbin*) phenomena as the most commonly occurring alternations in kanji compound formation (see Tsujimura (1996)).

3. Calculate the probability of a given kanji being realized with each reading $P(r|k)$ and phonological ($P_{phon}(r)$) or conjugational ($P_{conj}(r)$) alternation occurring. The set of reading probabilities is unique for each character but the latter two are calculated based on the reading only. See Baldwin et al. (2002) for a detailed explanation.

4. Create an exhaustive listing of reading candidates for each dictionary entry s and calculate the probability $P(r|s)$ for each, based on evidence from step 3 and the naive Bayes model (assuming independence between all parameters).

$$\begin{aligned} P(r|s) &= P(r_{1..n}|k_{1..n}) && (1) \\ P(r_{1..n}|k_{1..n}) &= \prod_{i=1}^{n} P(r_i \mid k_i) \times \\ & \quad \times P_{phon}(r_i) \times P_{conj}(r_i) && (2) \end{aligned}$$

5. Calculate the corpus based frequency $F(s)$ of each dictionary entry in the corpus and then its probability $P(s)$ according to equation (3). Notice that the term $\sum_i F(s_i)$ depends on the given corpus and is constant for all strings s in the corpus.

$$P(s) = \frac{F(s)}{\sum_i F(s_i)} \qquad (3)$$

6. Use Bayes rule to calculate the probability $P(s|r)$ of each resulting reading according to equation (4). Since both $P(r)$ and $\sum_i F(s_i)$ are constant[1], the final plausibility grade can be estimated as in equation (5).

$$\begin{aligned} \frac{P(s|r)}{P(s)} &= \frac{P(r|s)}{P(r)} && (4) \\ Grade(s|r) &= P(r|s) \times F(s) && (5) \end{aligned}$$

3 System Description

3.1 System Overview

The base dictionary for our system is the public-domain EDICT Japanese-English electronic dictionary.[2] We extracted all entries containing at least one kanji character and executed the steps described above for each. Corpus frequencies were calculated over the EDR Japanese corpus (EDR, 1995). During the generation step we ran into problems with extremely large numbers of generated readings, particularly for strings containing large numbers of kanji. Therefore, to reduce the size of generated data, we kept only generated readings satisfying $P(r|s) \geq 5 \times 10^{-5}$ for entries with less than 5 segments. Finally, to complete the set we inserted correct readings for all dictionary entries s_{kana} that did not contain any kanji characters (for which no readings were generated above), with plausibility grade calculated by equation (6).[3]

$$Grade(s_{kana}|r) = F(s_{kana}) \qquad (6)$$

The resulting data set is as follows:

Total Entries:	97,399
Entries containing kanji:	82,961
Average number of segments:	2.30
Total readings:	2,646,137
Unique readings:	2,194,159
Average readings per entry:	27.24
Average entries per reading:	1.21
Maximum readings per entry:	471
Maximum entries per reading:	112

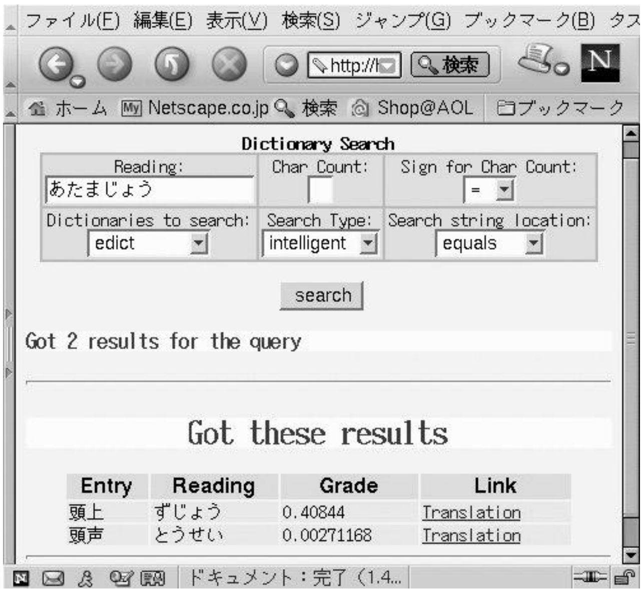
The above set is stored in a MySQL relational database and queried through a CGI script. Figure 1 depicts the system output for the query *atamajou*.[4] The system is easily accessible by any Japanese language-enabled web browser. Currently we include only a Japanese-English dictionary but it would be a trivial task to add links to translations in additional languages.

[1] Strictly speaking, $P(r)$ is not constant, but for the purposes of generating $P(s|r)$ the relative ranking of each $P(\cdot|r)$ is all that is important.

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

[3] Here, $P(r|s_{kana})$ is assumed to be 1, as there is only one possible reading (i.e. r).

[4] In this section we describe the system as it is visible at http://hinoki.ryu.titech.ac.jp/dicti/.



Dictionary Search

Reading:	Char Count:	Sign for Char Count:
あたまじょう		=
Dictionaries to search:	Search Type:	Search string location:
edict	intelligent	equals

search

Got 2 results for the query

Got these results

Entry	Reading	Grade	Link
頭上	ずじょう	0.40844	Translation
頭声	とうせい	0.00271168	Translation

Figure 1: Example of system display

3.2 Search Facility

The system supports two major search modes: **simple** and **intelligent**. **Simple** search emulates a conventional electronic dictionary search (see, e.g., Breen (2000)) taking both kanji and kana as query strings and displaying the resulting entries with their reading and translation. It also supports wild character and specified character length searches.

Intelligent search accepts only kana query strings[5] and proceeds in two steps. Initially, the user is provided with a list of candidates corresponding to the query, displayed in descending order of the score calculated from equation (5). The user then must click on the appropriate entry to get the full translation.

3.3 Example Search

Let us explain the benefit of the system to the Japanese learner through an example. Suppose the user is interested in looking up word 頭上 *zujou* "overhead" in the dictionary but does not know the correct reading. Both 頭 and 上 are quite common characters but frequently realized with different readings, namely *atama, tou*, etc. and *ue, jou*, etc., respectively. As a result, the user could interpret the string 頭上 as being read as *atamajou* or *toujou* and query the system accordingly. Tables 1 and 2 show the results of these two queries.[6]

From Table 1 we see that only two results are returned for *atamajou*, and that the highest ranking candidate corresponds to the desired string 頭上. Note that *atamajou* is not a valid word in Japanese, and that a conventional dictionary search would yield no results.

Entry	Reading	Grade	Translation
頭上	*zujou*	0.40844	overhead
頭声	*tousei*	0.00271168	head voice

Table 1: Results of search for *atamajou*

Entry	Reading	Grade	Translation
登場	***toujou***	73.2344	appearance
頭上	*zujou*	1.51498	overhead
搭乗	***toujou***	1.05935	embarkation
筒状	***toujou***	0.563065	cylindrical
道場	*doujou*	0.201829	dojo
東上	***toujou***	0.126941	going to Tokyo
凍傷	*shimoyake*	0.0296326	frostbite
凍傷	*toushou*	0.0296326	frostbite
刀匠	*toushou*	0.0144911	swordsmith
頭声	*tousei*	0.0100581	head voice
刀傷	*toushou*	0.00858729	sword wound
痘瘡	*tousou*	0.00341006	smallpox
凍瘡	*tousou*	0.0012154	frostbite
東清	*toushin*	0.000638839	Eastern China

Table 2: Results of search for *toujou*

Things get somewhat more complicated for the reading *toujou*, as can be seen from Table 2. A total of 14 entries is returned, for four of which *toujou* is the correct reading (as indicated in bold). The string 頭上 is second in rank, scored higher than three entries for which *toujou* is the correct reading, due to the calculation procedure not considering whether the generated readings are correct or not.

4 Evaluation

4.1 Experiment

To get a preliminary idea of our system's effectiveness we ran the following experiment. We used a collection of 139 entries taken from a web site describing reading errors made by native speakers of Japanese[7] and for each entry we queried our system with the erroneous reading to see whether the intended entry was contained in the system output. To transform this collection of items into a form suitable for dictionary querying we converted all readings into hiragana and in some cases removed context words. Table 3 gives a comparison of results returned in **simple** (conventional) and **intelligent** (proposed system) search modes. 61 entries, mostly proper names and 4-character proverbs, were not contained in the dictionary and have been excluded from evaluation.

[5] In order to retain the functionality offered by the simple interface, we automatically default all queries containing kanji characters and/or wild characters into simple search.

[6] Note that the readings listed are always the correct readings for the corresponding Japanese dictionary entry, and not the reading in the original query. Also, readings here are given in romanized form whereas they appear only in kana in the actual interface. See Figure 1.

[7] `http://www.sutv.zaq.ne.jp/shirokuma/godoku.html`

	Conventional	Our system
In dictionary	78	78
Avg. # results	1.55	5.49
Successful	12	36
Mean rank	1.33	4.52

Table 3: Comparison between a conventional dictionary and our system

We can see that our system is able to handle 3 times more erroneous readings then the conventional system, representing an error rate reduction of 36.4%. However, the average number of results returned (5.49) and mean rank of the desired entry (4.52 – calculated only for successful queries) are still sufficiently small to make the system practically useful.

The fact that the conventional system covers any erroneous readings at all is due to the fact that those readings are appropriate in alternative contexts, and as such both readings appear in the dictionary. Out of 42 entries that our system did not handle, the majority of misreadings were due to graphical similarity-induced error (16) and the usage of incorrect character readings in compounds (16). Another 5 errors were a result of substituting the reading of a semantically-similar word, and the remaining 5 a result of interpreting words as personal names.

4.2 Discussion

In order to emulate the limited cognitive abilities of a language learner, we have opted for a simplistic view of how individual kanji characters combine. In step 4 of preprocessing, we use the naive Bayes model to generate an overall probability for each reading, and in doing so assume that component readings are independent of each other, and that phonological and conjugational alternation in readings does not depend on lexical context. Clearly this is not the case. For example, kanji readings deriving from Chinese and native Japanese sources (*on* and *kun* readings, respectively) tend not to co-occur in compounds. Furthermore, phonological and conjugational alternations interact in subtle ways and are subject to a number of constraints (see Vance (1987)).

However, depending on the proficiency level of the learner, s/he may not be aware of these rules, and thus may try to derive compound readings in a more straightforward fashion which is adequately modeled through a simplistic independence model. As can be seen from our preliminary experiments our model is effective in handling a large number of reading errors but can be improved further. We intend to modify it to incorporate further constraints as necessary after observing the correlation between the search inputs and selected dictionary entries.

Furthermore, we are working under the assumption that the target string is contained in the original dictionary and thus base all reading generation on the existing entries, assuming that the user will only attempt to look up words we have knowledge of. We also provide no solution for random reading errors.

4.3 Future Work

So far we have conducted only limited tests on the correlation between search results and target words. In order to truly evaluate the effectiveness of our system we need to perform experiments with a larger data set. The reading generation and scoring procedure can be adjusted by adding various weight parameters to modify the reading candidate scores and thus affect the results displayed.

The current cognitive model does not include any notion of errors due to graphic or semantic similarity of the different kanji. We intend to expand it to consider these error types, too.

5 Conclusion

In this paper we have proposed a method for constructing a system capable of handling motivated reading errors. Our method takes dictionary entries containing kanji characters and generates reading candidates for each. Dictionary entries plausibly associated with different reading inputs are precomputed and stored in the system database accessible through a web interface. In response to a user query, the system displays dictionary entries potentially corresponding to the reading entered. Evaluation indicates that the proposed system significantly enhances error-resilience in dictionary searches.

References

Timothy Baldwin and Hozumi Tanaka. 2000. A comparative study of unsupervised grapheme-phoneme alignment methods. In *Proc. of the 22nd Annual Meeting of the Cognitive Science Society (CogSci 2000)*, pages 597–602, Philadelphia.

Timothy Baldwin, Slaven Bilac, Ryo Okumura, Takenobu Tokunaga, and Hozumi Tanaka. 2002. Enhanced Japanese electronic dictionary look-up. In *Proc. of LREC*. to appear.

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

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

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

Natsuko Tsujimura. 1996. *An Introduction to Japanese Linguistics*. Blackwell, Cambridge, Massachusetts, first edition.

Timothy J. Vance. 1987. *Introduction to Japanese Phonology*. SUNY Press, New York.