

Say-as Classification for Alphabetic Words in Japanese Texts

Hisako Asano, Masaaki Nagata, Masanobu Abe

NTT Cyber Space Laboratories, NTT Corporation

{asano.hisako,nagata.masaaki,abe.masanobu}@lab.ntt.co.jp

Abstract

Modern Japanese texts often include Western sourced words written in Roman alphabet. For example, a shopping directory in a web portal, which lists more than 8,000 shops, includes a total of 6,400 alphabetic words. As most of them are very new and idiosyncratic proper nouns, it is impractical to assume all those alphabetic words can be registered in the word dictionary of a text-to-speech synthesis system; their pronunciations must be derived automatically. Our solution consists of two steps. Step 1 classifies each unknown alphabetic word into a *say-as* class (English, Japanese, French, Italian or English spell-out), which indicates how it is to be read, and Step 2 derives the pronunciation using the grapheme-to-phoneme conversion rules for the classified *say-as* class. This paper proposes a method of *say-as* classification (i.e. Step 1) that uses the Support Vector Machine. After some trial and error, we achieved 89.2% accuracy for web shop data, which we think sufficient for practical use.

1. Introduction

Japanese is a non-Roman alphabet language, but modern Japanese texts often include Western sourced words written in Roman alphabet (e.g., “OUTLET横浜”), rather than in Japanese scripts for loan words (i.e. Katakana). Even Japanese sourced words are sometimes transcribed in Roman alphabet for style reasons, such as “イタリア料理店 Ristrante Tokyo” (means “the Italian restaurant, Ristrante Tokyo”) to give it an Italian taste. For example, the category “buy” (apparel, cosmetic, book and sport shops etc.) of a shopping directory in a popular regional web portal covering major cities in Japan, lists more than 8,000 shops, and includes a total of 6,400 alphabetic words¹ (2,900 distinctive words). As most of them are very new and idiosyncratic proper nouns, it is impractical to assume all those alphabetic words can be registered in the word dictionary of a text-to-speech synthesis system, so their pronunciations must be derived automatically.

A *say-as* element in Speech Synthesis Markup Language² indicates how the text in the element is to be read by its *interpret-as* attribute. For example, “123” in a *say-as* element is read in English as “one two three” if the *interpret-as* attribute has *digits* value, and “one hundred and twenty third” if the *interpret-as* attribute has *ordinal* value. We define *say-as* classes for alphabetic words in Japanese texts as an extension of values for *interpret-as* attribute in *say-as* element; different *language* classes (e.g., *English*: “shop”, *Italian*: “ristrante”, *Japanese*³: “Tokyo”), as well as each

language spell-out class (e.g., *English spell-out*: “USA”) and each *language abbreviation* class (e.g., *English abbreviation*: “NY (= New York)”).

Say-as classification differs from language identification [1] [2]. The former classifies how a word is pronounced in the text, while the latter classifies the identity of the text.

There are a few grapheme-to-phoneme conversion methods for alphabetic words in Japanese texts. [3] uses letter bigrams and the target *say-as* class is only English (single language). It states that most of the errors are personal names (probably derived from other nationalities) and initials, that is, many errors occur in words that are other than English class. [4] uses the longest match method. It notes that using language dependent data is better than using language independent data.

Our target texts are shop names and publicity material on the web, and include alphabetic words having a variety of *say-as* classes such as English, Japanese, and French. Considering the characteristics of the target texts, our approach for grapheme-to-phoneme conversion of unknown alphabetic words consists of two steps; the first step classifies an unknown word into a *say-as* class, and the second step derives the pronunciation using each grapheme-to-phoneme conversion rule set for the classified *say-as* class (e.g., *English* rule: application [3], *English spell-out* rule: table of alphabetic characters and pronunciation). This paper proposes a new method for the first step, i.e. *say-as* classification using the Support Vector Machine (SVM) [5]. The main focus of the paper is to investigate what kind of feature is effective for the task. The reason for using SVM is it can handle a large number of feature sets without suffering the sparse data problem. SVM was recently applied to a variety of NLP tasks and was shown to be very effective. It is necessary to extend SVM to handle multi-class classification because SVM is a binary classifier. We use YamCha 0.2⁴, which is a general-purpose SVM classifier extended to support multi-class classification by the pair-wise method and the one-versus-rest method.

2. Target say-as class

We investigated the *say-as* class distribution of alphabetic words in the above-mentioned category “buy” web shop data to select suitable *say-as* classes for target texts. The result is shown in Table 1. This paper considers the top five *say-as* classes: English, English spell-out (called “spell-out” hereafter), Japanese, French, and Italian, are treated as the target classes. We ignore the abbreviation class (e.g., “NY” = New York) because for grapheme-to-phoneme conversion it

¹ An alphabetic word is a word that consists of only alphabetic letters and apostrophe.

² <http://www.w3.org/TR/speech-synthesis>

³ *Japanese* here means alphabetic transliteration of Japanese.

Any Japanese word can be written in alphabetic

transliteration, but this is unusual. Proper nouns are sometimes written in alphabetic transliteration depending on the context.

⁴ <http://cl.aist-nara.ac.jp/~taku-ku/software/yamcha>

is very difficult to automatically restore the original word sequence from the abbreviation without the help of a dictionary.

Table 1: Say-as class distribution in web shop data: shop category “buy”

Say-as class	Total number of words (ratio %)	Number of distinct words (ratio %)
English	3883 (60.5%)	2023 (69.5%)
(Eng) spell-out	1366 (21.3%)	291 (10.0%)
Japanese	486 (7.6%)	278 (9.5%)
French	333 (5.2%)	199 (6.8%)
Italian	97 (1.5%)	56 (1.9%)
Abbreviation ⁵	222 (3.5%)	37 (1.3%)
Others	30 (0.5%)	28 (1.0%)
Total	6416	2911

3. Accents

French and Italian have accented letters such as “é”, “ù” and “ê”. In Table 1 data, however, such letters are replaced by non accented letters (e.g., “é” to “e”) or decomposed into a non accented letter and an apostrophe (') or a grave (`) (e.g., “é” to “e'”, “ù” to “u'” or “u`”)⁶ because the Japanese character set does not include accented letters; no accented letters were present in the data used below.

4. Word boundaries

Basically, say-as class is assigned on a word-by-word basis because there are some cases where adjacent words have different say-as classes (e.g., “Ristrante Tokyo”: “Ristrante” = *Italian* class, “Tokyo” = *Japanese* class). However, defining what a word is not trivial, it affects the overall accuracy. It is necessary to consider where word boundaries for unknown alphabetic words are because the say-as class boundary should match the word boundary.

An apostrophe can indicate a say-as class boundary if it separates the English “’s” from other say-as class (e.g., “Hanako’s”; “Hanako”: Japanese female name, “’s”: English). The apostrophe is connected to the next word in the possessive case, but it is connected to the previous word in the decomposition form of an accented letter as described in Section 3 (e.g., “é” to “e'”). It is necessary to confirm where the best word boundary is for apostrophes: pre-boundary, post-boundary or both boundaries. This is confirmed by the experiments in Section 6.5.1.

The boundary indicated by a transition from a lower-case letter to an upper-case letter can also be a say-as class boundary (e.g., “CafeHanako”; “cafe”: English, “Hanako”: Japanese). It is noted capitalization could be changed for stylistic reasons such as “DeeP”. There is some possibility that treating all lower to upper case transitions as boundaries will degrade the accuracy. This is examined in Section 6.5.2.

5. Contexts and features

We consider that context is informative in addition to the

⁵ The total of all languages

⁶ In other accented letters such as “û” or “ö”, all accents are omitted, not decomposed.

target alphabetic word itself in determining the say-as class. For example, the say-as class of “Take”, which is ambiguous as to whether it is English or Japanese, depends on the context; “Take” in “Take it easy” is English (the pronunciation is “teik”), and “Takeさん” (Mr. Take; a person name) is Japanese (the pronunciation is “take”). Therefore, we use the feature sets for SVM classifier consisting of word features taken from the target words, its two prior and two following words.

The following word features are considered. The effectiveness of each feature is confirmed in Section 6.4.

- Spelling information: Previous studies on language identification suggest that spelling information is effective in determining the origin of words. We use letter or syllable representations of the first, the second, the second from the last and the last positions in a word as spelling information⁷. It is difficult to guess correct syllable boundaries in each language automatically, so we approximated the syllable boundaries with the boundaries from a vowel (i.e. the letters “AIUEO”) to a consonant (i.e. except “AIUEO”) in this paper (e.g., “sho/p”, “Yo/ko/ha/ma”, “Ri/stra/n-te”). The alphabetic words have various notations due to the combinations possible with upper-case and lower-case forms, but they are converted to upper-case in the representation. Non-alphabetic words do not use this feature (a dummy value is added). For example, the spelling information of “Ristrante”, “Yokohama” and “イタリア”⁸ are “R, I, T, E”, “Y, O, M, A” and “*, *, *, *” as letters and “RI, STRA, *, NTE”, “YO, KO, HA, MA” and “*, *, *, *” as syllables (“*” is a dummy value).
- Word length: The number of characters in the word.
- Word type: Alphabetic words are classified as all upper-case (e.g., “SHOP”), all lower-case (e.g., “shop”), capitalized (e.g., “Shop”), or others (e.g., “shoP”). Hiragana, Katakana, Kanji or Numeric words is assigned the kind of script as word type (e.g., “店”: Kanji, “100”: Number). Other words, i.e. Symbol words, are assigned their symbol (e.g., “%”: %, “\$”: \$).
- Part of speech: 33 kinds of part of speech such as verb and noun.

6. Experiments

We performed several experiments. In all experiments, the SVM parameters for learning were second degree of polynomial kernel and 1 slack variable, and the pair-wise method was used for multi-class extension.

6.1. Corpus

We prepared two training corpora. One was the web shop data of category “buy”, which was manually created, and the other was supplemental data from electronic dictionaries,

⁷ The order of the position in words less than four syllables is the first, the last and the second, and a dummy value is added in the empty features.

⁸ This word is not a say-as classification target because it is a non-alphabetic word, but it can appear as the context of a target word.

which was relatively easy to obtain but does not necessarily suit the usage in Japanese text.

The web shop training corpus consisted of 2,142 shop names⁹ and 1,492 shop publicity sentences extracted from the Table 1 web shop data. All sentences had at least one alphabetic word, and the total number of alphabetic words belonging to target says-as class was 6,164.

The dictionaries for English, French, and Italian had about 10,000 – 12,000 words extracted from electronic language dictionaries¹⁰. There is no dictionary for alphabetic transliteration of Japanese, so we made one using about 12,000 Japanese family and first names transliterated automatically. We also created a spell-out dictionary consisting of about 900 words collected manually from mainly various web sites. The total number of dictionary entries was about 46,000. Most entries were single words, but some were compounds. In learning by using dictionary entries, each entry was treated as one sentence, that is, there was no Japanese context information.

For the evaluation, 429 sentences including alphabetic words were randomly extracted from category “buy” (none overlapped the training corpus) and 423 sentences were extracted from category “eat” (restaurant, coffee shop and bar etc.) in the web shop directory. The targets of say-as classification are unknown alphabetic words, but the number of the unknown words depends on how many alphabetic words exist in the dictionary of a text-to-speech synthesis system. Since we assumed infrequent words to be unknown, words whose frequencies were less than five in the web shop training corpus were treated as unknown words in the evaluation corpus. If the word has say-as class ambiguity in the training corpus, the word is also treated as an unknown word, irrespective of word frequency in the training corpus. The evaluation excluded words belonging to a say-as class that is not one of the five target classes. Table 2 shows the number of all alphabetic words and the number treated as unknown words in the evaluation corpus. Accuracy was evaluated by precision rate (number of correctly classed words / total number of unknown words [%]).

Table 2: Word number in evaluation corpus

Alphabetic word	Category “Buy”	Category “Eat”
All words	706	771
Unknown words	526	629

6.2. Initial setting

A boundary from lower-case to upper case was not treated as a word boundary, and both apostrophe boundaries were treated as word boundaries. A word has seven features; the first, the second, the second from the last and the last syllable representations, word length, word type and part of speech.

6.3. Training data source and amount

To confirm the effectiveness of the training data source and the amount of training data in the initial setting shown in

⁹ Each shop name is treated as a sentence in learning.

¹⁰Accented letters are replaced by “decomposition” and “omission”.

Section 6.2, four types of training data were evaluated; (1): dictionaries only, (2): web shop data only, (3): dictionaries and web shop data and (4): one half of (2). The result is shown in Table 3.

(1), dictionaries only, is not effective even though there is a large amount of data. (2), web shop data only, offers high precision rates with a small amount of data. (3), the mix of dictionaries and web shop data, the precision rate rises in category “buy”, but hardly changes in category “eat”, its efficiency is low given that it uses eight times the training data compared to (2). General vocabularies without Japanese context information are not so useful for this particular task. (4) reduces the amount of training data and the precision rate also reduces. Accordingly, there is some possibility of increasing precision rate by adding more web shop training data.

The following experiments used training corpus (3) because it offers the maximum precision.

Table 3: Precision rate for training data source & amount

Training data source & amount	Category “Buy”	Category “Eat”
(1): Only Dictionaries	37.6%	34.8%
(2): Only shop data	87.1%	86.6%
(3): (1)+(2)	89.0%	86.8%
(4): Half of (2)	85.9%	82.4%

6.4. Features

6.4.1. Initial word features

We investigated what word features are effective by removing the feature one by one. Table 4 shows the change in precision rate for all features. Smaller values indicate higher effectiveness. The order is sorted according to the effectiveness for category “buy”.

Table 4: Precision difference with one feature dropped

Dropped feature	Category “Buy”	Category “Eat”
(1): The last syllable	-7.8%	-4.6%
(2): The first syllable	-6.7%	-4.0%
(3): Word length	-3.4%	+0.5%
(4): The second syllable	-2.3%	-1.9%
(5): Word type	-2.1%	-0.5%
(6): The last-1 syllable	-0.8%	-1.1%
(7): part of speech	+0.2%	+0.5%

Table 4 indicates that the most effective features are syllables, which directly indicate spelling. The second and the second from the last syllables are not as effective as the first and the last syllables because these features are empty in the words with fewer than 4 syllables; such words are in the majority. Part of speech lowers precision slightly, so in the following experiments we dropped part of speech from the word features used.

6.4.2. Comparison between syllable and letter

Table 5 shows precision rates where syllables are used as spelling information (= Table 4 (7)) and letters are used

instead of syllables. Syllables are extremely more effective than letters as spelling information because syllables better express the characteristic of each say-as class.

Table 5: Precision rate of syllable and letter

Spelling information	Category "Buy"	Category "Eat"
Syllable	89.2%	87.3%
Letter	83.7%	82.2%

6.5. Word boundaries

6.5.1. Apostrophe

Table 6 shows the precision rate achieved by using apostrophes as word boundaries (both prior and post word boundaries). The precision varies slightly and the level of effectiveness depends on the category because the evaluation corpus contained few apostrophes.

Table 6: Precision rate by the apostrophe word boundary

Word boundary of apostrophe	Category "Buy"	Category "Eat"
Both-boundary	89.2%	87.3%
Pre-boundary	86.9%	87.4%
Post-boundary	87.3%	87.1%

6.5.2. Boundary by lower-case to upper-case

Table 7 shows the precision rate for the lower-to-upper-case boundary. There is little difference in precision because category "buy" had just one entry and category "eat" had no entry whose boundary was a say-as class boundary. Even in the training corpus, only ten boundaries became say-as class boundaries. Due to this very low frequency, the treatment of these boundaries is not important in the web shop data.

Table 7: Precision rate for lower to upper case boundary

Lower-to-upper-case boundary	Category "Buy"	Category "Eat"
Word boundary	89.2%	87.3%
No word boundary	89.2%	87.6%

6.5.3. The best set and its error analysis

It is difficult to rigorously select the best set of word features because feature effectiveness depends on the category. However, based on the above results, the set yielding the highest precision consists of (1) the initial word features (omitting "part of speech"), and (2) pre-and-post apostrophes as word boundaries. This set has a precision of 89.2%.

Table 8 provides examples of typical errors in this best set. Most frequent error is the confusion between English and French such as (1). These errors account 29 of the total of 59 errors in category "buy" (51%) and 34 of the total of 80 errors in category "eat" (43%). The writing style shown in (2) is common in the web shop domain and is the cause of a few errors. Some pre-processing may be needed. Some classification errors do not indicate problems with grapheme-to-phoneme conversion because some words have the same

pronunciation in different say-as classes such as (3). "MARIO" in Japanese has the same pronunciation in Italian.

Table 9 shows each precision rate for the five say-as classes. For example, "buy" evaluation corpus has 343 unknown words of English class, and 316 of them were classified correctly as English class. The precision rate is low in classes with low frequency. We believe that the precision would increase if web shop training data with low frequency class examples were to be added.

In all experiments, the accuracy of category "buy" is better than category "eat". The main reason is that the frequency of Italian class words in category "eat" is higher than the frequency in category "buy". As shown in Table 9, the accuracy of Italian class is low.

Table 8: Examples of typical errors

No	Error class (Correct class)	Errors underlined
(1)	English (French)	La Boutique Montres <u>a</u> Paris
(2)	Spell-out (English)	<u>B-A-R</u>
(3)	Japanese (Italian)	<u>MARIO</u> GELATERIA

Table 9: Precision rate for each class with best set

Class	Category "Buy"	Category "Eat"
English	316/343 = 92.1%	393/419 = 93.8%
Spell-out	86/90 = 95.6%	78/84 = 92.9%
Japanese	32/39 = 82.1%	39/53 = 73.6%
French	31/43 = 72.1%	22/43 = 51.2%
Italian	4/10 = 40.0%	17/30 = 56.7%
Mix ¹¹	0/1 = 0%	0/0

7. Conclusions

We proposed a method that can assign say-as classes to unknown alphabetic words in Japanese, and confirmed its effectiveness. We intend to apply each grapheme-to-phoneme conversion rule set for the classified say-as class and evaluate the subsequent reading accuracy.

8. References

- [1] G. Kikui, "Identifying the Coding System and Language of On-line Documents on the Internet", *Proc. COLING-96*, 1996
- [2] W. B. Cavnar and J. M. Trenkle, "N-Gram-Based Text Categorization", *Proc. SDAIR-94*, 1994
- [3] K. Saito, A. Shinohara, M. Nagata and H. Ohara, "A transliteration algorithm for adapting a Japanese voice controlled browser to English", *Trans. of the Japanese Society for Artificial Intelligence*, Vol. 17, No.3, 2002, (In Japanese)
- [4] K. Masuda and K. Umemura, "An algorithm that extracts alphabet-kana rules from name database", *Journal of Information Processing Society of Japan*, Vol.40, No.7, 1997, (In Japanese)
- [5] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995

¹¹ A word consisting of more than one say-as class (e.g., "CafeHanako": English + Japanese)