Non-Standard Word and Homograph Resolution for Asian Language Text Analysis

Craig Olinsky and Alan W Black

Language Technologies Institute
Carnegie Mellon University
colinsky@cs.cmu.edu, awb@cs.cmu.edu

ABSTRACT

In this paper we present a general model for text analysis of Asian languages (Chinese and Japanese). That is a method for mapping strings of characters to strings of identified trivially pronounceable words. This work is based on the English Non-Standard Word analysis model suitably augmented to deal with both the lack of spaces between words in Japanese and Chinese and addressing the issues of homographs. Results are present for the sub-components of the process.

1. INTRODUCTION

This paper addresses issues of text analysis for Asian languages, particularly Chinese and Japanese. Text analysis in most languages typically consists of a set of hand-written, often hacky rules that expand text tokens to identifiable words. Text analysis is often dismissed as an afterthought to language processing as many aspects have been ignored. With the increase in applications using technologies like speech synthesis, language modeling for speech recognition and information retrieval there is a growing demand for good, well-designed text analysis systems that can be properly gauged and ported between domains.

A project at the Johns Hopkins University Summer Workshop 1999 addressed this issue and devised a general trainable model for analyzing English text [5]. The project concentrated on the expansion of so-called “non-standard words” (NSWs) which at first approximation is basically all tokens that do not appear directly in the lexicon. NSWs consist of numbers, symbols, abbreviations, etc as well as genuinely out-of-vocabulary words. These must be reliably identified, properly classified, and expanded to conventional words.

The JHU project addressed issues in text normalization for English alone, though it did investigate four different text genre: news articles, recipes, email and classified ads. This paper reports the application of that basic model to Asian languages and our identification of what changes and enhancements are required for success in these different languages.

The basic NSW expansion model consists of 5 basic modules, as shown in *Figure* 1. Strings of characters are first tokenized into white-space separated tokens. A splitter then further splits these tokens as required based on punctuation, case/number distributions etc. These *split tokens* are then classified in to one of around 23 types. The NSW identifies how a token should be expanded. For example, a digit string “23” may be assigned the tag NUM if it is used as a standard number quantity with pronunciation “twenty-three”, while in other contexts (e.g. when preceded by the name of a month) it would be tagged with NDAY identifying it as an ordinal with the pronunciation “twenty-third”.

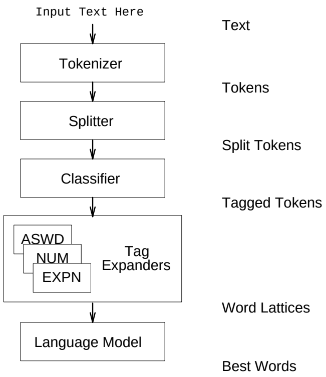

Figure 1: *Basic NSW Expansion Model*

After classification tokens are then expanded to full word forms. For most token/tag pairs this is an deterministic algorithmic process but there could also be some ambiguity. One particular token/tag expansion is interesting. Identifying something as an abbreviation is one thing but you also need to know what it expands too. In many cases the expansion will be defined by a lexicon of abbreviations, but abbreviations are productive and new ones may appear, particularly in things like email and classified ads. Although the JHU Workshop project includes a mechanism for prediction of unknown abbreviations we have not yet addressed this issue within our Asian language text analyzer.

The final stage of text analysis is the use of a language model to choose between possible multiple expansions of some tokens.

Given this basic model we had to modify this slightly to accommodate Chinese and Japanese. This first important observation is that unlike most European languages, Chinese and Japanese do not use any form of white space between words. Although both use punctuation in a similar way to English it is common to even insert newlines within what would normally be called words to allow appropriate paragraph formatting.

Thus we had to replace the initial tokenizer with a statistically trained model to provide a basic token splitter. Although this part of the English NSW model is considered trivial, as is often the case in dealing with multiple languages what is trivial is one can be much harder in another. The following section describes the technique we used.

The second key difference between Chinese, Japanese and English is the distribution of homographs, that is words written the same way but pronounced differently. In English they are relatively rare in alphabetic tokens, apart from a predictable class of stress position in verb/nouns like ``segment", ``project" etc, there are probably less than 100 words in the class, (e.g. ``bass", ``Dr", ``wind" etc). Though digit strings too can be ambiguous. In Japanese however, and to a lesser extent in Chinese, homographs are much more common than in English.

2. WORD-SPLITTING

In English and other Western languages, word boundary segmentation involves a simple algorithm: boundaries are marked by white space or punctuation. While there is still limited ambiguity in this case, such parsing is not generally considered a hindrance to or major stumbling block for subsequent text processing.

In Chinese, various characters can function as independent words, as part of character compounds, as part of proper names, or even completely abstracted from their meaning in phonetic spellings of foreign words. The character 到 *dao4*， for instance can appear as part of a complex verb (e.g. 看到 *kan2dao4*), or by itself as a preposition. As another example, most everyone is familiar with the fact the many Chinese names are not just comprised of arbitrary phonetic characters but are often sets of words or phrases favorably describing or claiming skill or luck for the name-bearer.

Thus, a simple lexicon search and breakdown, while finding a number of possible interpretations of a sentence, will not serve to disambiguate between these interpretations without taking into account some degree of usage information (whether through full parsing, rule application, or statistical methods).

Because Japanese uses a combination of writing systems, with Chinese characters used primarily for root forms but most inflectional endings and grammatical devices in *hiragana* (a syllabic script), word boundary discrimination can generally be simplified by paying attention to the transitions between the different scripts. A third script, *katakana* (a syllabic script), is used primarily for phonetic renderings of foreign words, further reducing ambiguity. Thus, in some ways, Chinese can be considered the "harder" case for word segmentation

A number of different approaches have been attempted for the word boundary disambiguation task, ranging from hand-crafted rules to full neural network systems. The most common techniques employed involve variations of the *Maximum Matching Algorithm*, which involves aligning word boundaries according to the longest possible matching character compounds in the given lexicon, backtracking only when an end-of-sentence (or other punctuation) is reached without having achieved full coverage of the text segment in question.

The system developed for this study, following the Bell Labs GETEX text analysis system [4] uses trained Weighted Finite-State Transducers (WFSTs) for the disambiguation task. A lexicon of Chinese words was weighted on word-frequency computations over the corpus, and arranged into a WFST such that each path through the transducer comprised one complete word ending with its estimated cost based upon the word frequency determined from a test set pres-segmented by a native speaker. An iterative process was used to learn word frequencies of the training set to quiescence, starting from an initial naïve segmentation based on assignment of equal weights to all words in the lexicon.

Testing was performed against a corpus extracted from 1996 issues of the *People's Daily* taking from the Linguistic Data Consortium mandarin Chinese News Text Corpus (LDC95T13). All text markup was removed, and a selection of 500 paragraphs was randomly chosen from the text, of which the first 100 (13000 characters) were used for testing and provided to a native Chinese speaker for hand-marking; and the remaining 400 (47,000 characters) were used for iterative training of word frequencies.

As a lexicon, the GB2312-encoded version of the freely available CEDICT online dictionary [1] (15,000 words) was chosen. The AT&T Finite State Machine Library (FSM) toolkit [2] was used for WFST construction. All manipulation of the corpus and lexicon was done with Java.

Precision, recall, and accuracy were tabulated for three separate text segmentations: (1) a naïve segmentation of the test set assuming all words to be composed of single characters, used as a baseline measurement; (2) segmentation of the test set using the initial evenly weighted lexicon; and (3) the final segmentation using converged word frequency weightings based upon multiple iterations through the training set.

Word frequencies obtained from the training set stabilized after three iterations.

To calculate the following figures, the space between every character in the document (for a total of 12,885 unique data points) was considered to have one of two values: "segment here" (positive value) or "don't segment here" (negative value). Performance was then scored relative to a similarly marked hand-labeled test set.

Results were thus calculated here in reference to the existence and position of segmentation breaks: (*tp* = true positive, *fn* = false negative, etc.)

$$\text{precision} = \frac{tp}{(tp + fp)} \qquad (1)$$

$$\text{recall} = \frac{tp}{(tp + fn)} \qquad (2)$$

$$\text{accuracy} = \frac{tp + tn}{(tp + tn) + (fp + fn)} \qquad (3)$$

Text	Precision	Recall	Accuracy
Single-character segmentation (baseline)	65.63 %	100 %	65.63 %
WFST: Evenly weighted lexicon (Maximal Matching)	86.46 %	95.23 %	87.09 %
WFST: Converged word frequency weightings	86.57 %	95.38 %	87.26%

3. NSW TAGGING

The basic set of NSW tags originally devised is somewhat English-specific. They capture the basic differences but many of the distinctions, such as differentiating years from standard numbers, are language-specific. Although we would like a set of language independent NSW tags, this ultimately is not necessarily practical and is actually not necessary. As the NSW tags will always be used for a particular language and the predictive models are unlikely ever to be useful cross-language, having a modified NSW tag set per language is most likely the more practical solution. On the other hand, a large number of the NSW categories are multilingually common, and thus we feel that identical techniques for their automatic labeling should be applicable.

For our initial study of the international applicability of NSW tagging, we chose to apply the NSW framework to a Japanese rather than a Chinese corpus. While Japanese morphology simplifies the Japanese word segmentation task over that in Chinese, the combination of native Japanese vocabulary and Chinese loan-words, combined with the higher proportion of Romanized words and numbers in Japanese texts suggests that NSW tagging and pronunciation discrimination to be a more complex task.

In English text, the NSW tags break down into three major classes: alphabetic, numeric and other (including punctuation, etc) Romanized alphabetics are common in Japanese text and mostly offer the same problems as the do in English so in this case we have chosen the same set of tags – although the same sequence of letters or numbers sometimes belongs to a different class multilingually, depending on regional preferences and differences in the phoneme sets between languages.

Alphabetic non-standard word classes include:

EXPN. (Expansion) An abbreviation or contraction which, when pronounced, is expanded into a full word (or words) rather than pronounced as is. Examples include rd. for road, and etc. for et cetera. This type of expansions are fairly rare in Japanese text, and are almost exclusively used for non-native words.

LSEQ. (Letter Sequence) An abbreviation or contraction which is read out letter-by-letter rather than expanding to a full word. These are commonly abbreviations of a number of words or a phrase, rather than a single word – common examples include IBM, NEC, and the WTO.

ASWD. (As a Word). Some multi-word abbreviations, generally those which contain a sufficient number of vowels, are pronounceable as single words and are thus not spelled out. NATO and SCUBA are generic examples of ASWDs.

Numbers (western digits) in Japanese have more varied pronunciation than English but there are usually stronger cues to which pronunciation is required. As with Chinese, counter particles may follow numbers characterizing what type of object that is being counted. Although this phenomena is somewhat reduced in modern Japanese a number of these very common, often the pronunciation of the number is directly affected. For example a counter for people (人) is affects the pronunciation of the numbers 1 and 2. A generic counter (つ) affects numbers 1 to 10 (though its rarely used for any other numbers). Other counters affect the pronunciation less though often they affect phonetic boundaries through assimilation.

We have preserved the original numeric NSW tags as they are mostly still useful for Japanese and Chinese, though (as with English) mapping of a digit string to distinct tags does not necessarily give rise to separate pronunciations – another language-dependent factor. There is, however, a major distinction in that both Japanese and Chinese can express numbers (in all of these categories) using either western digits or native Chinese characters.

Numeric NSW tags include (the following examples include some additional non-numeric characters to clarify usage, but keep in mind that such characters are not part of the actual tag):

ひとり ひとり
NUM. A cardinal number: 1, 2, 3; １人/一人 *hitori*

NDAY. An date. E.g., 1 日, 二日

NTEL A telephone number (or partial number, such as an extension): (123) 555-5555; x2837; +1 (03) 4545-4545

NDIG. A number that is read out as digitals (with no other specific meaning).

NIDE. A number used as an identifier (such as a Student ID #.).

NZIP. A Zip Code or PO box.

NTIME. A compound time. 9:30, 9 時 30 分, 九時三十分

NYER. A year. 1999 年, 二千円。

NZIP. A Zip code, PO Box, or numeric block identifier. e.g., 東京１―３―５

We have so far only tagged a small amount of Japanese news text to look at the distribution of these tags as well as predictive models.

From a sampling of two Japanese Business News texts, the *Nikkei* (日本経済新聞) and *Kyodo News Service*, and the Chinese *People's Daily* from the LDC collection, we first cataloged the distribution of (Roman) alphabetic and numeric characters:

	Kyodo (J)	Nikkei (J)	P. Daily (C)
Sample (chars)	1,062,053	2,279,623	6,737,0631
Alphabetic	1.3 %	0.7 %	0.3 %
Numeric	5.6%	0.8 %	3.9 %

From the Kyodo corpus, a selection of 5000 alphabetic and numeric strings (evenly divided) were extracted and hand-labeled with the appropriate NSW class, 10% of which was set aside for test data and the rest used for classification training.. These strings were then run through a set of feature detectors, to catalog a set of lexical features about the string and surrounding characters for automatic training. These features and the training tags were then composed into a stepwise classification tree, and NSW tag selection performed on the test data.

These three tables show the distribution of alphabetic and numeric NSW Tags in the test and training data, as well as the percentage of correct tag classifications for alphabetic and numeric NSWs across the test and training data.

On analysis of the hand labeled data we find alphabetic tags can be assigned through simple rules, without the need for a statistically learned process, if is contains lower case vowels or is in a predefined set of common pronounceable acronyms (e.g. "NATO") it is ASWD.

NUMERIC NSW Tag	Distribution
NUM	38.1 %
NDAY	22.7 %
MONEY	12.3 %
NYER	9.8%
NDIG	3.7%
PRCT	1.1%
NZIP	0.6 %
NTIME	0.2%
NTEL	< 0.1 %
URL	< 0.1 %

ALPHA NSW Tag	Distribution
LSEQ	86.6 %
ASWD	10.9 %
EXPN	3.5 %

If it is in a small set of common expansions, notably NY, and money symbols (USD, HKD etc) then it is an EXPN, otherwise it is a LSEQ. To apply this rule to our training, we used a small lexicon of such common ASWDs and EXPNs as a feature detector when training our classifications trees.

For numeric tags we trained a CART tree using features such as number of digits, punctuation in context, surrounding kana/kanji. The following character is a strong dictator of the type.

Final Results:

Data Set	ALPHA NSWs	NUMERIC NSWs
Training Data	94.7 %	77.7 %
Test Data	93.6 %	72.9 %

4. OTHER WORK

Another important phenomenon in Japanese (and to a lesser extent Chinese) less common in English is that of homographs. For example some Kanji characters have different pronunciation depending on the verb type they are being used for. This is especially evidenced in transitive/intransitive pairs, such as 集る（あつまる） *atsumaru* to gather (trans) and 集る（あつめる） *atsumeru* to gather (intrans).

Though sometimes the pronunciations (and meanings) are not even close, for instance the pair 行（い）く（く） *iku* to go, and 行う（おこなう） *okonau* to do, to carry out

Often these distinctions can be made by the kana context, but not always. Kanji characters typically have at least two pronunciations: one, an *on* reading derived from its original Chinese reading and a *kun* reading which is typically Japanese rooted. To a native the choice is usually obvious, and compound (multi-character) words generally share the same linguistic origin, although there are exceptions. One triplet of variant pronunciations are the is the character in 生（う）まれる *umareru* to be born, 生（しょう）じる *shoujiru* to bring about, and 生（なま）ビール *nama biru* draft beer.

In English homograph disambiguation has been based on the techniques described in [6]. There, each occurrence of the homograph is labeled (originally by hand, although automatic techniques were also investigated), and a collection for features such as nearby content words, part of speech, etc., were used to train decision tree that could classify the types.

We are developing a similar context dependent system to choose between expansions of such ambiguous characters.

5. CONCLUSION

The basic NSW model seems to work with Chinese and Japanese though substantial new work was required, especially for splitting, because of the significantly different writing conventions. Homograph disambiguation is also a major part of these languages though it is still left as an external part to the basic English treatment.

From the experience that was gained in building the English NSW it was obvious that accurate measures of results are very difficult without a very large amount of data as many phenomena are relatively rare. Though of course this does not mean they can be ignored, because as there so many rare phenomena it is likely that a relatively small example of data will include at least one rare example.

We are continuing to label data and improve our models so as to produce reliable structure text expansion suitable for general use as well as our speech synthesis research.

6. ACKNOWLEDGEMENT

This work was supported in part by an NSF Combined Research and Curriculum Development grant. The opinions expressed in this report do not necessarily represent the opinions of the NSF.

7. REFERENCES

[1] Denisowski, Paul. CEDICT (online Chinese machine-readable dictionary); available at www.mindpsring.com/~paul_denisowski/cedict.html

[2] Mohri, Mehryar, Fernando C. N. Pereira, and Michael D. Riley, "AT&T Finite State Machine Library (FSM)," available at http://www.research.att.com/sw/tools/fsm/

[3] Shih C., Sproat R. (1996), "Issues in Text-to-Speech Conversion for Mandarin." *Computational Linguistics and Chinese Language Processing* 1 (1), 37-86.

[4] Sproat R., Shih C., Gale W., and Chang N. (1996). "A Stochastic Finite-State Word Segmentation Algorithm for Chinese." *Computational Linguistics*, 22 (3).

[5] Sproat R., Black A., Chen S., Kumar S., Ostendorf M., & Richards C. (1999). "Normalization of Non-Standard Words: WS'99 Final Report", *CLSP Summer Workshop*, Johns Hopkins University. www.clsp.jhu.edu/ws99/projects/normal.

[6] Yarowsky, D., (1996), "Homograph Disambigutation in Text-to-Speech Synthesis." In J. van Santen, R. Sproat, J. Olive, and J. Hirschenberg, eds., *Progress in Speech Synthesis*, pp. 158-172, Springer-Verlag.