Statistical Identification of English Loanwords in Korean Using Automatically Generated Training Data

Kirk Baker, Chris Brew

Departments of Linguistics, Computer Science and Engineering
The Ohio State University
Columbus, OH
kbaker@ling.osu.edu, cbrew@acm.org

Abstract

This paper describes an accurate, extensible method for automatically classifying unknown foreign words that requires minimal monolingual resources and no bilingual training data (which is often difficult to obtain for an arbitrary language pair). We use a small set of phonologically-based transliteration rules to generate a potentially unlimited amount of pseudo-data that can be used to train a classifier to distinguish etymological classes of actual words. We ran a series of experiments on identifying English loanwords in Korean, in order to explore the consequences of using pseudo-data in place of the original training data. Results show that a sufficient quantity of automatically generated training data, even produced by fairly low precision transliteration rules, can be used to train a classifier that performs within 0.3% of one trained on actual English loanwords ($\approx 96\%$ accuracy).

1. Introduction

Identifying the etymological source of an unknown word is important for a wide range of language applications. For example, automatically translating proper names and technical terms is a notoriously difficult task because these items can come from anywhere, are often domain-specific and are frequently missing from bilingual dictionaries (e.g., Knight and Graehl, 1998; Al-Onaizan and Knight, 2002). In the case of borrowings across languages with unrelated writing systems and dissimilar phonemic inventories (e.g., English and Korean), an appropriate treatment for an unknown word may be transliteration or back-transliteration (Knight and Graehl, 1998). However, in order to transliterate an unknown word correctly, it is often useful to first identify the originating language of the unknown word. Etymological classification also plays a role in information retrieval and cross-lingual information retrieval systems where finding equivalents between a source word and its various target language realizations improves indexing of search terms and subsequently document recall (e.g., Kang and Choi, 2000; Oh and Choi, 2001; Kang and Choi, 2002). Source language identification is also a necessary component of speech synthesis systems, where the etymological class of a word can trigger different sets of letter-to-sound rules (e.g., Llitjos and Black, 2001; Yoon and Brew, 2006).

Identifying foreign words is similar to the task of language identification (e.g., Beesley, 1998), in which documents or sections of documents are classified according to the language in which they are written. However, foreign word identification is made more difficult by the fact that words are nativized by the target language phonology and the fact that differences in character encodings are removed when words are rendered in the target language orthography. For example, French and German words are often written in English just as they appear in the original languages – e.g., tête or außerhalb. In these cases, characters like ê and ß indicate with a high degree of reliability cues to the etymological source of the foreign word. However, when these same words are transliterated into Korean, such character level differences are no longer maintained: tête becomes <te-teu> and außerhalb becomes <a-u-seo-hal-peu> (Li, 2005:32). Instead, information such as transition frequencies between characters or the relative frequency of certain characters in known Korean words versus known French or German words can be used to distinguish these classes of words.

Oh and Choi (2001) describes an approach along these lines to automatically identifying and extracting English words from Korean text. Oh and Choi (2001) formulates the problem in terms of a syllable tagging problem – each syllable in a hangul orthographic unit is identified as foreign or Korean, and each sequence of foreign-tagged syllables is extracted as an English word. Hangul strings are modeled by a hidden Markov model where states represent a binary indication of whether a syllable is Korean or not. Transitional probabilities and the probability of a syllable being English or Korean are calculated from a hand-tagged corpus of over 100,000 words. Kang and Choi (2002) employs a similar Markov-based approach that alleviates the burden of manually syllable tagging an entire corpus, but relies instead on dictionaries that distinguish English and Korean words. These statistical approaches deliver fairly promising results. However, the burden of tagging words was not eliminated but pushed onto professional lexicographers.

While statistical approaches have been successfully applied to the language identification task, a major drawback to applying a statistical classifier to loanword identification is the requirement for a sufficient amount of labeled training examples. Amassing a large list of transliterated foreign words is expensive and time-consuming. We address this issue by using phonological conversion rules to generate potentially unlimited amounts of pseudo training data at very low cost. Although the rules themselves are not highly accurate, a classifier trained on sufficient amounts of this automatically generated data performs as well as one trained on actual examples. We demonstrate the technique by identifying English words used in Korean.

2. Experiments

2.1. Data Set

Our experiments are based on a list of 10,000 English words attested as loanwords in Korean. The majority of the words (9686) come from the National Institute of the Korean Language's (NIKL) list of foreign words (NIKL, 1991) after removing duplicate entries, proper names and non-English words. Entries considered duplicates in the NIKL list are spelling variants like *traveller/traveler, analog/analogue, hippy/hippie*, etc. The remainder (314) were manually extracted by the first author from a variety of online Korean text sources.

Pronunciations for English words were added to this list of words and were derived from two main sources: the Hoosier Mental Lexicon (HML) (Nusbaum et al., 1984), which contains phonological representations of 20,000 English words based on standard American English, and the Carnegie Mellon Pronouncing Dictionary (CMUDICT) (Weide, 1998), which contains pronunciations of 127,000 words. Differences between the transcription conventions used in the HML and CMUDICT were standardized to produce onsistent phonological representations. Loanwords contained in neither of these two sources were transcribed with reference to an online dictionary[1] using the HML transcription conventions.

10,000 Korean words were randomly selected from the National Institute of the Korean Language's (NIKL, 2002) list of Korean words, which contains frequency and familiarity information for approximately 55,000 Korean words. We did not maintain any distinction between Sino-Korean and native Korean words.

Standard Korean character encodings represent syllables rather than individual letters, so we converted the original hangul orthography to a character-based representation, retaining orthographic syllable breaks. Words are represented as sparse vectors, with each non-zero entry in the vector corresponding to the count of a particular character trigram that was found in the word. For example, the English loanword *user* is produced in Korean as <yu-jeo> and is represented as {∅∅y:1, ∅yu:1, yu-:1, u-j:1, -jeo:1, jeo∅:1, eo∅∅:1}, where ∅ is a special string termination symbol and '-' indicates an orthographic syllable boundary.

The decision to use trigrams instead of syllables was based on the intuition that character level transitions provide important cues to etymological class. 1grams or 2grams are not as informative, while going to 4grams or higher results in severe problems with data sparsity. This trigram feature representation resulted in 2276 total features; English words contained 1431 distinct trigrams and Korean words contained on 1939 distinct trigrams.

2.2. Classifier

We want to learn a classifier $y = f(\boldsymbol{x})$ from a set of labeled training data $D = \{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_i}, y_i), \ldots, (\boldsymbol{x_n}, y_n)\}$. In our case, the vectors $\boldsymbol{x_i} = [x_{i1}, \ldots, x_{ij}, \ldots, x_{ip}]^T$ contain counts of the number of times each trigram in the data set appeared in word $\boldsymbol{x_i}$. Most of the time, any given trigram only occurs one time in a particular word, so $\boldsymbol{x_i}$ is most often a binary vector. $y_i \in \{+1, -1\}$ represents class labels that encode membership $(+1)$ or non-membership (-1) in one of the two etymological classes English or Korean.

A wide range of statistical learning algorithms could be used for the task of assigning words to one of two etymological classes. We used a logistic regression model (Genkin et al., 2004) to automatically classify words as English or Korean in origin. Although logistic regression has not been widely used in the machine learning community (Krishnapuram et al., 2005), it has a long history of use in classical statistics, and we have found that this model consistently outperforms other machine learning algorithms (e.g., support vector machines, naive Bayes, decision trees, SNoW) in head-to-head comparisons on a range of language classification tasks – for example, verb classification (Li et al., 2008) and animacy classification (Baker and Brew, in progress). The logistic regression classifier is a conditional probability model of the form

$$P(y_k = +1|\boldsymbol{x}, \boldsymbol{\beta}) = \psi(\boldsymbol{\beta}^T \boldsymbol{x_i}) = \psi(\sum_j \beta_j x_{ij})$$

which is parameterized by the vector of regression coefficients $\boldsymbol{\beta}$. $p(y = +1|\boldsymbol{x_i})$ represents an estimate of the probability that $\boldsymbol{x_i}$ belongs to the class.

The description of the model above is based on (Genkin et al., 2004) and the accompanying freely available software[2]. In this implementation, the logistic link function

$$\psi(r) = \frac{\exp(r)}{1 + \exp(r)}$$

is used, giving a logistic regression model.

2.3. Experiment with Labeled Data

The first experiment, with labeled data (10,000 English loanwords; 10,000 Korean words), used a 10-fold, 90/10 train/test split. Baseline accuracy for all experiments was 50%. Mean classification accuracy for the regression classifier was 96.2%.

For the sake of comparing the regression model to a more familiar model, we ran a Bayesian classifier over the same data set. The Bayesian model is widely used for its simplicity and the fact that it is often competitive with more sophisticated models on a wide range of classification tasks. It is typically used to estimate class-conditional probabilities from maximum likelihood estimates approximated with relative frequencies from a set of training data, and has the form

$$c' = \underset{c}{\text{argmax}}\, P(c|\boldsymbol{x}) = \prod_j P(x_j|c)$$

Mean classification accuracy using labeled data was 91.1% for the Bayes classifier. This difference is not unexpected, in accordance with the observation that discriminative models typically perform better than generative ones (Ng and Jordan, 2002).

[1] `http://dictionary.reference.com`

[2] Available for download from `http://www.bayesianregression.org`.

Taking the results of the regression classifier as a reasonable baseline for what can be expected using hand-labeled data, the next experiment looks at using phonological rules to automatically generate English training data.

2.4. Experiment with Pseudo-English Loanwords

The pseudo-English loanwords were generated from the list of context-sensitive phonological rewrite rules presented in (Korean Ministry of Culture and Tourism, 1995) that describe the changes English phonemes undergo when they are borrowed into Korean. These rules map English phonemes onto hangul characters. Example rules are shown below (Korean Ministry of Culture and Tourism, 1995, p. 129: 1(1), 2).

1. after a short vowel, word-final voiceless stops ([p], [t], [k]) are written as codas (b, s, g)
   book [bʊk] → <bug>

2. ɨ is inserted after word-final and pre-consonantal voiced stops ([b], [d], [g])
   signal [sɪgnəl] → <si-gɨ-neol>

We implemented a total of 30 rules as regular expressions in a Python[3] script and applied them to the pronunciations in the CMU Pronouncing Dictionary (Weide, 1998) to create a set of possible but unattested English loanwords in Korean. These items served as training data for the distinction between actual English loanwords and Korean words.

In order to get some estimate of quality of the transliteration rules, we tested them on the set of 10,000 actual English loanwords in Korean. The output of the program was compared to the attested Korean forms, and the proportion of times the rule applied as predicted was calculated for each English word and phoneme. Overall transliteration accuracy, measured as exact word matching, was 50.3%. Fortunately, as we will see, we do not need this figure to be high in order for the rules to be useful to us.

We also evaluated the rule based output in terms of the number of correctly transliterated consonant sequences per word. For example, given the English word *pocket* and actual transliteration of <po-kes>, a predicted transliteration of <pa-kes> would count as containing all correctly transliterated consonants. Phonological rules generate correctly transliterated consonant sequences 90% of the time. This disparity underscores the variability associated with vowel transliteration, which is often highly influenced by orthography (Oh and Choi, 2002).

In general the rules do a good job of predicting the borrowed form of individual English consonants in Korean. The weighted mean proportion of times the consonants appeared as predicted is 0.97. These results are broken down by consonant in Table 1.

The number of training instances ranged from 10,000 to 100,000. The test items were all 20,000 items from Experiment 1. The training data did not include any of the test items. This means that if the phonological conversion rules produced a form that was homographic with any of the actual English loanwords, this item was removed from the training set. Note that this is conservative: in practical situations we would expect that the conversion rules would sometimes manage to duplicate actual loanwords, with the possibility of improved performance.

[3]Distributed under an open source license from `http://www.python.org`.

Stops		Fricatives		Nasals		Glides	
p	0.990	f	0.999	m	1.000	r	0.988
t	0.989	v	0.985	n	0.997	l	0.987
k	0.990	θ	0.978	ŋ	0.983	w	0.967
b	0.996	ð	1.000			j	0.859
d	0.996	s	0.975				
g	0.984	z	0.733				
		ʃ	0.985				
		ʒ	1.000				
		ʧ	0.951				
		ʤ	0.969				
		h	0.983				

Table 1: Accuracy by consonant of transliteration rules. Mean= 0.970

We had a total of 62688 labeled actual Korean words (Sino-Korean plus native Korean). In order to keep the same number of items in the English and Korean classes, i.e., in order to avoid introducing a bias in the training data that was not reflected in the test data, we used a random sampling with replacement sampling model for the Korean words.

Figure 1 shows the classification accuracy of the regression classifier as a function of the amount of training data. Classifier accuracy appears to asymptote at around 90,000 instances of each class within 0.3% (95.8% correct) of the classifier trained on actual English loanwords.

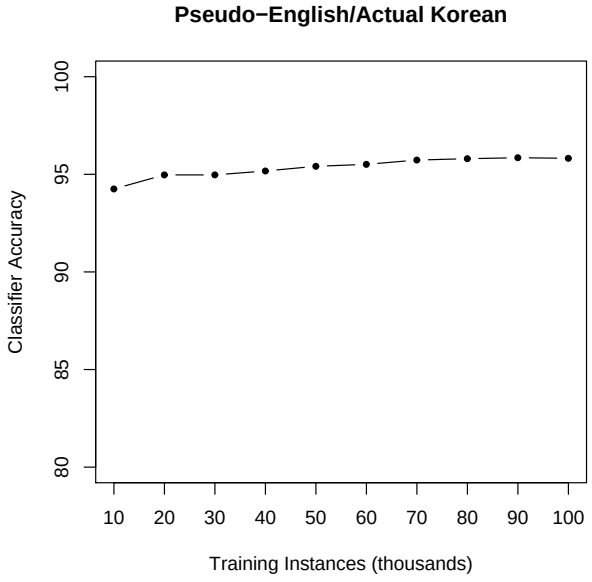

Figure 1: Classifier accuracy trained on pseudo-English loanwords and actual Korean words; classifying actual English and Korean words.

Experiment 2 demonstrates the feasibility of approximating a set of English loanwords with phonological conversion rules. However, it relies on a dictionary of native words, which is a time-consuming and expensive resource to produce. Therefore, we also investigated the feasibility of approximating a label for the Korean wordsas well.

2.4.1. Experiment with Pseudo-English Loanwords and Unlabeled Korean Words

Based on observations of English loanwords in Japanese (Graff and Wu, 1995) and Chinese (Graff, 2003) newswires, we believe that the majority of these items will occur relatively infrequently in comparable Korean text. This means that we are assuming that there is a direct relationship between word frequency and the likelihood of a word being Korean, i.e., the majority of English loanwords will occur very infrequently. Accordingly, we sorted the items in the Korean Newswire corpus (Cole and Walker, 2000) by frequency on the assumption that Korean words will tend to dominate the higher frequency items, and examined the effects of using these as a proxy for known Korean words.

We identified 23406254 Korean orthographic units (i.e., *eojeol*) in the Korean Newswire corpus. Because we believe that high frequency items are more likely to be Korean words, we applied a sampling without replacement sampling scheme to the instances extracted from the corpus. This means that the frequencies of items in our extracted subset approximately match those in the actual corpus, i.e., we have repeated items in the training data. Thus, the classifier for this experiment was trained on automatically generated pseudo-English loanwords as the English data and unlabeled lexical units from the Korean Newswire as the Korean data. Again, the test items were all 20,000 items from Experiment 1. The training data did not include any of the test items.

Figure 2 shows the classification accuracy of the regression classifier as a function of the amount of training data. Classifier accuracy again asymptoted around 90,000 items per training class at 3.7% below (92.4%) the classifier trained on actual English loanwords.

The assumption that frequent items in the Korean Newswire corpus are all Korean is false. For example, of the 100 most frequent items we extracted, 5 were English loanwords. These words and their rank are shown in Table 2. However, we believe that the performance of the classifier in this situation is encouraging, and that using a different genre for the source of the unlabeled Korean words might provide slightly better results. This is because of the nature of a news corpus: it reports on international events, so foreign words are relatively frequent compared to a period novel or something like that.

Word	Rank	Frequency
Yeonhab News <yeon-hab-nyu-seu>	30	51792
percent <peo-sen-teu>	32	49367
New York <nyu-yog>	89	19652
Russia <leo-si-a>	91	19162
Clinton <keul-lin-ton>	94	18860

Table 2: Frequent English loanwords in the Korean Newswire corpus.

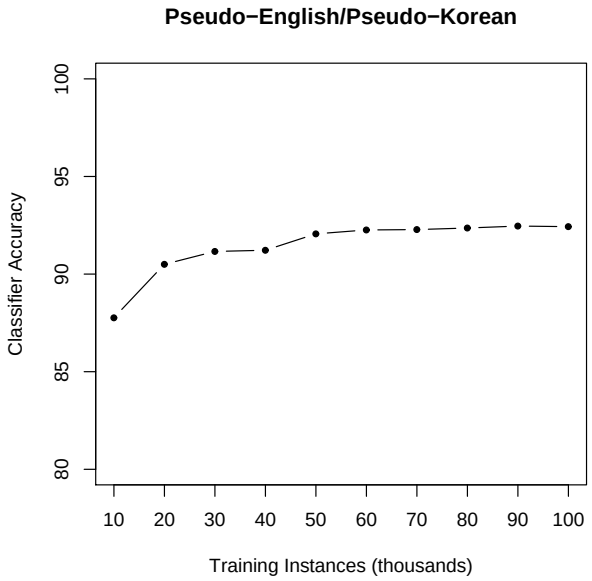

Figure 2: Classifier accuracy trained on pseudo-English loanwords and pseudo-Korean items; classifying actual English and Korean words.

3. Conclusion

This paper addressed the issue of obtaining sufficient labeled data for the task of automatically classifying words by their etymological source. We demonstrated an effective way of using linguistic rules to generate unrestricted amounts of virtually no-cost training data that can be used to train a statistical classifier to reliably discriminate instances of actual items. The basic insight our method makes use of is the fact that there is a substantial research literature dealing with the phonology of loanword adaptations between a large number of contact languages. We take advantage of the work of previous researchers who have codified the various ways in which words from one language change when they are borrowed into another. Although the initial linguistic analysis is non-trivial, a wealth of such information is now available.

The rules describing how words change when they are borrowed from one language to another tend to be relatively few and easy to implement. For example Li (2005) provides a similar number of adaptation rules (around 20-30) for several European languages that have loanwords in Korean. Therefore, the methodology outlined here is not restricted to the English-Korean language pair, but can be widely applied to additional languages for which obtaining labeled training data is difficult. It is also applicable to languages other than Korean that have large numbers of English borrowings (e.g., Japanese).

Other researchers have used manually specified transliteration rules for converting English words to their borrowed form in some other language, especially in the context of

cross language information retrieval. For example, Mettler (1993) and Fujii and Ishikawa (2001) describe methods for converting English words to a set of potential katakana equivalents in Japanese for bilingual English/Japanese document retrieval, and Kang and Choi (2001) and Oh and Choi (2002) use the same set of phonological adaptation rules used in this paper in the context of English-to-Korean transliteration. In this paper we demonstrated that linguistic rules of this nature can be used not only to generate exact cross language equivalents, but to generate large sets of training instances over which further class-based generalizations can be reliably obtained.

Acknowledgments

We acknowledge support from NSF Grant 0347799, and from the Graduate School of the Ohio State University. The conclusions are our own, and do not necessarily represent the opinions of these funding bodies.

4. References

Yaser Al-Onaizan and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 400–408.

Kirk Baker and Chris Brew. in progress. Animacy classification using Bayesian logistic regression.

Kenneth R. Beesley. 1988. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, pages 47–54.

Andy Cole and Kevin Walker. 2000. Korean Newswire. Linguistic Data Consortium, Philadelphia. LDC2000T45.

Atsushi Fujii and Tetsuya Ishikawa. 2001. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, 35(4):389–420.

Alexander Genkin, David D. Lewis, and David Madigan. 2004. Large-scale Bayesian logistic regression for text categorization. *DIMACS Technical Report*.

David Graff and Zhibiao Wu. 1995. Japanese Business News Text. Linguistic Data Consortium, Philadelphia. LDC95T8.

David Graff. 2003. Chinese Gigaword Third Edition. Linguistic Data Consortium, Philadelphia. LDC2007T38.

Byung-Ju Kang and Key-Sun Choi. 2000. Two approaches for the resolution of word mismatch problem caused by English words and foreign words in Korean information retrieval. In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, pages 133–140.

Byung-Ju Kang and Key-Sun Choi. 2002. Effective foreign word extraction for korean information retrieval. *Information Processing and Management*, 38:91–109.

Byung Ju Kang. 2001. *A resolution of word mismatch problem caused by foreign word transliterations and English words in Korean information retrieval*. Ph.D. thesis, Computer Science Department, KAIST.

Kevin Knight and Jonathan Graehl. 1998. Machine Transliteration. *Computational Lingustics*, 24:599–612.

Korean Ministry of Culture and Tourism. 1995. English to Korean standard conversion rules. Electronic Document. `http://www.hangeul.or.kr/nmf/23f.pdf`. Accessed February 13, 2008.

Balaji Krishnapuram, Lawrence Carin, Mário A. T. Figueiredo, and Alexandar J. Hartemink. 2005. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27.

Jianguo Li, Kirk Baker, and Chris Brew. 2008. A corpus analysis of Levin's verb classification. Paper presented at the American Association for Corpus Linguistics (AACL 2008). Provo, Utah.

Eui Do Li. 2005. Principles for transliterating roman characters and foreign words in Korean. In *Proceedings of the 9th Conference for Foreign Teachers of Korean*, pages 95–147.

Ariadna Font Llitjós and Alan Black. 2001. Knowledge of language origin improves pronunciation of proper names. In *Proceedings of EuroSpeech-01*, pages 1919–1922.

Matt Mettler. 1993. TRW Japanese fast data finder. Tipster Text Program: Phase I Workshop Proceedings. `http://acl.ldc.upenn.edu/X/X93/X93-1011.pdf`.

Andrew Y. Ng and Michael I. Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in NeuralInformation Processing Systems 14*. MIT Press.

NIKL. 1991. Survey of the state of loanword usage: 1990. Electronic Document. The National Institute of the Korean Language, Seoul, Korea. `http://www.korean.go.kr`.

NIKL. 2002. Hyeondae gugeo sayong bindo josa bogoseo. Electronic Document. The National Institute of the Korean Language, Seoul, Korea. `http://www.korean.go.kr`.

H. C. Nusbaum, David Pisoni, and C. K. Davis. 1984. Sizing up the Hoosier Mental Lexicon: Measuring the Familiarity of 20,000 Words. *Research on Speech Perception Progress Report No. 10*, pages 357–376.

Jong-Hoon Oh and Key-Sun Choi. 2001. Automatic extraction of transliterated foreign words using hidden Markov model. In *Proceedings of the 19th International Conference of Computer Processing on Oriental Language*, pages 433–438.

Jong-Hoon Oh and Key-Sun Choi. 2002. An english-korean transliteration model using pronunciation and contextual rules. In *COLING 2002*, pages 1–7.

J. W. Weide. 1998. The Carnegie Mellon Pronouncing Dictionary v. 0.6. Electronic Document, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Kyuchul Yoon and Chris Brew. 2006. A linguistically motivated approach to grapheme-to-phoneme conversion for Korean. *Computer Speech and Language*, 2(4):357–381.