

## CHAPTER 4

### AUTOMATICALLY IDENTIFYING ENGLISH LOANWORDS IN KOREAN

#### 4.1 Overview

This chapter deals with the task of automatically classifying unknown words according to their etymological source. It focuses on identifying English loanwords in Korean, and presents an approach for automatically generating training data for use by supervised machine learning techniques. The main innovation of the approach presented here is its use of generative linguistic rules to produce large quantities of training data, circumventing the need for manually labeled resources.

Being able to automatically identify the etymological source of an unknown word is important for a wide range of NLP 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 (i.e., English and Korean), the appropriate course of action 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 necessary 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, 2000b; 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., Llitjós and Black, 2001; Yoon and Brew, 2006). In Korean, for example, a phonological consonant tensification rule applies to semantically transparent compounds of Sino-Korean origin. For example, the Sino-Korean syllable *pyeng* corresponds to two homographic morphemes *illness* and *anger*, both of which have two pronunciations in compounds: untensed initial /p/ (e.g., 화병 *hwapyeng* [hwapyəŋ] *vase*, 호리병 *holipyeng* [horibyəŋ] *genie's bottle* and 지병 *ciipyeng* [ʧibyəŋ] *terminal illness* and tensed initial /p/ (e.g., 콜라병 *khollapyeng* [kʰol:ap\*yəŋ] *cola*, 화병 *hwapyeng* [hwap\*yəŋ] *anger disease* and 허리병 *helipyeng* [həlip\*yəŋ] *backache*) (Yoon and Brew, 2006: 367). In addition, words of English origin often undergo /s/-tensification that is not orthographically indicated (e.g., 세일 *seyil* [s\*eil] ‘sale’, 펄스 *phelsu* [pʰəlsi] ‘pulse’ (Yoon and Brew, 2006: 372).

The sections that follow describe and evaluate statistical approaches to identifying English loanwords in Korean. Section 4.2 describes previous work on identifying English loanwords in Korean. Section 4.3 lays out the current approach and describes the supervised learning algorithm used in the experiments that are presented in Section 4.4.

## 4.2 Previous Research

Identifying foreign words is similar to the task of language identification (e.g., Beesley, 1988), 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* provide reliable 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 *테트* *theytu* and *außerhalb* becomes *아우서할프* *awusehalpu* (Li, 2005: 132). 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 corpus of over 100,000 words in which each syllable was manually tagged as foreign or Korean. Oh and Choi (2001) reports precision and recall values ranging from 96% to 98% for identifying foreign word tokens in their corpus, but is not clear whether these values are obtained from a disjoint train/test split of the data or indicate performance of their system on trained data.

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 a foreign word dictionary, a native word dictionary, and a list of 2000 function words obtained from a manually POS-tagged corpus. Kang and Choi (2002) uses their method to extract a set of 799 potential foreign terms from their corpus, and restrict their analysis to this set of terms. Kang and Choi (2002) reports precision and recall for foreign word extraction over this candidate set of 84% and 92%, respectively. While these results are promising, the burden of manually labeling data has not been eliminated, but deflected to external resources.

The experiments presented in the next section describe an accurate, easily 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). It does not require tagging and uses corpus data that is easily obtainable from the web, for example, rather than hand-crafted lexical resources.

### 4.3 Current Approach

While statistical approaches have been successfully applied to the language identification task, one 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. The classifier used here is a sparse logistic regression model. The sparse logistic regression model has been shown to provide state of the art classification results on a range of natural language classification tasks such as author identification (Madigan, Genkin, Lewis,

Argamon, Fradkin, and Ye, 2005a), verb classification (Li and Brew, 2008), and animacy classification (Baker and Brew, accepted). This model is described in the next section.

#### 4.3.1 Bayesian Multinomial Logistic Regression

At a very basic level of description, learning is about observing relations that hold between two or more variables and using this knowledge to adapt future behavior under similar circumstances. Regression analysis models this type of learning in terms of the way that one variable  $\mathbf{Y}$  varies as a function of a vector of variables  $\mathbf{X}$ . This function is represented in terms of the conditional distribution of  $\mathbf{Y}$  given  $\mathbf{X}$  and a set of weighted parameters  $\beta$ . Bayesian approaches to regression modeling involve setting up a distribution on the parameter vector  $\beta$  that encodes prior beliefs about the elements of  $\beta$ . The prior distribution should be strong enough to allow accurate estimation of the model parameters without overfitting the model to the training data (e.g., Genkin, Lewis, and Madigan, 2004; Gelman, Carlin, Stern, and Rubin, 2004: 354). The statistical inference task involves estimating the parameters  $\beta$  conditioned on  $\mathbf{X}$  and  $\mathbf{Y}$  (Gelman et al., 2004: 354). The simplest and most flexible regression model is the normal linear model (Hays, 1988; Gelman et al., 2004), which states that each value of  $\mathbf{Y}$  is equal to a weighted sum of the corresponding values of the predictors in  $\mathbf{X}$ :

$$(4.1a) \quad \mathbf{Y}_i = \beta_0 + \sum_{p=1}^P \beta_p \mathbf{X}_{ip}$$

In Equation (4.1a),  $i$  indexes over examples in the training set, and  $\beta_0$  is the  $y$ -intercept or bias, which is analogous to the prior probability of class  $k$  in a naive Bayes model. This formulation assumes that the true relationship between  $\mathbf{Y}$  and  $\mathbf{X}$

falls on a straight line, and that the actual observations of these variables are normally distributed around it. Equation (4.1a) is often expressed in equivalent notation as

$$(4.1b) \quad \mathbf{Y}_i = \sum_{p=0}^P \beta_p \mathbf{X}_{ip} \quad \text{where } X_{i0} \equiv 1$$

or in matrix notation as

$$(4.1c) \quad \mathbf{Y}_i = \boldsymbol{\beta} \mathbf{X}_i \quad \text{where } \mathbf{X}_{i0} \equiv 1.$$

The regression function for model (4.1a) expresses the expected value of  $\mathbf{Y}$  as a function of the weighted predictors  $\mathbf{X}$ :

$$(4.2) \quad E\{\mathbf{Y}_i\} = \beta_0 + \sum_{p=1}^P \beta_p \mathbf{X}_{ip}$$

In simple linear regression the expected value of  $\mathbf{Y}_i$  ranges over the set of real numbers. However, in classification problems of the type considered here, the desired output ranges over a finite set of discrete categories. The solution to this problem involves treating  $\mathbf{Y}_i$  as a binary indicator variable where a value of 1 indicates membership in a class and a value of 0 indicates not belonging to that class.

When  $\mathbf{Y}_i$  is a binary random variable, the expected outcome  $E\{\mathbf{Y}_i\}$  has a special meaning. The probability distribution of a binary random variable is defined as follows:

$\mathbf{Y}_i$	Probability
1	$P(\mathbf{Y}_i = 1) = \pi_i$
0	$P(\mathbf{Y}_i = 0) = 1 - \pi_i$

Applying the definition of expected value of a random variable (Kutner, Nachtsheim,

and Neter, 2004: 643, (A.12)) to  $\mathbf{Y}_i$  yields the following:

$$E\{\mathbf{Y}_i\} = \sum_{y \in \mathcal{Y}} yP(y) \quad [Definition\ of\ Expectation]$$

$$(4.3) \quad E\{\mathbf{Y}_i\} = 1(\pi_i) + 0(1 - \pi_i) = \pi_i$$

$$= P(\mathbf{Y}_i = 1)$$

Equating (4.2) and (4.3) gives

$$(4.4) \quad E\{\mathbf{Y}_i\} = \beta_0 + \sum_{p=1}^P \beta_p \mathbf{X}_{ip} = \pi_i = P(\mathbf{Y}_i = 1)$$

Thus, when  $\mathbf{Y}_i$  is binary, the mean response  $E\{\mathbf{Y}_i\}$  is the probability that  $\mathbf{Y}_i = 1$  given the parameterized vector  $\mathbf{X}_i$ . Since  $E\{\mathbf{Y}_i\}$  represents a probability it is necessary that it be constrained as follows:

$$(4.5) \quad 0 \leq E\{\mathbf{Y}_i\} = \pi \leq 1$$

This constraint rules out a linear regression function, because linear functions range over the set of real numbers instead of being restricted to  $[0, 1]$ . Instead, one of a class of sigmoidal functions which are bounded between 0 and 1 and approach the bounds asymptotically are used (Kutner et al., 2004: 559). One such function having the desired characteristics is the logistic function or logit (Agresti, 1990; Christensen, 1997), defined as

$$(4.6) \quad \pi = \frac{e^\eta}{1 + e^\eta}$$

and having the shape shown in Figure 4.1.

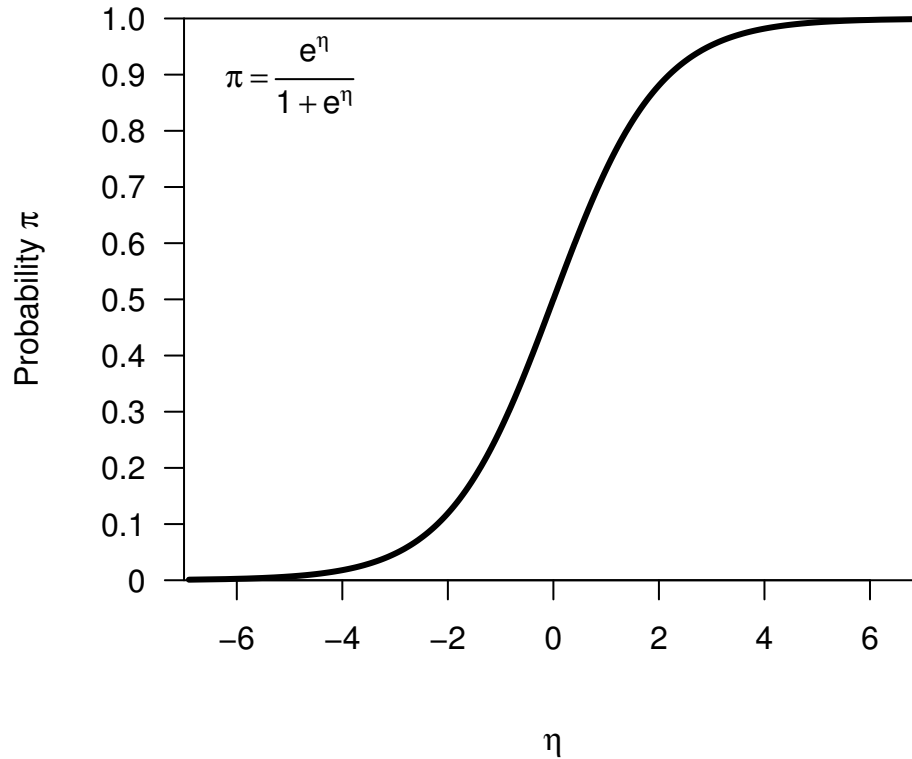


Figure 4.1: Standard logistic sigmoid function

A regression model which assumes a bounded curvilinear relationship between  $\mathbf{X}$  and  $\mathbf{Y}$  is known as a generalized linear model (e.g., Ramsey and Schafer, 2002). A generalized linear model is a probability model that relates the mean of  $\mathbf{Y}$  to  $\mathbf{X}$  via a non-linear function applied to the regression equation. Generalized linear models are linear in the predictors and non-linear in the output. Logistic regression models are a type of generalized linear model.

Multinomial logistic regression is an extension of the binary regression model described above to multiple classes. The basic method for handling more than two outcomes for  $\mathbf{Y}$  is to compare only two things at a time, i.e., to model multiple binary comparisons (Christensen, 1997). In essence, this requires constructing a separate

logit model for each class and choosing the model which assigns the highest probability to  $\mathbf{X}_i$ . The multinomial logistic regression model has the form

$$\begin{aligned}
(4.7) \quad \hat{\pi}_{1K} &= \log \frac{P(\mathbf{Y}_1 = 1 | \mathbf{X} = \mathbf{x})}{P(\mathbf{Y}_1 = K | \mathbf{X} = \mathbf{x})} = \beta_{10} + \sum_{p=1}^P \beta_{1p} \mathbf{x} \\
\hat{\pi}_{2K} &= \log \frac{P(\mathbf{Y}_2 = 1 | \mathbf{X} = \mathbf{x})}{P(\mathbf{Y}_2 = K | \mathbf{X} = \mathbf{x})} = \beta_{20} + \sum_{p=1}^P \beta_{2p} \mathbf{x} \\
&\vdots \\
\hat{\pi}_{(K-1)K} &= \log \frac{P(\mathbf{Y}_{K-1} = 1 | \mathbf{X} = \mathbf{x})}{P(\mathbf{Y}_{K-1} = K | \mathbf{X} = \mathbf{x})} = \beta_{(K-1)0} + \sum_{p=1}^P \beta_{(K-1)p} \mathbf{x}
\end{aligned}$$

The ratio inside the log function represents the odds of obtaining class  $k$  relative to class  $K$ . The choice of denominator is arbitrary in so far as the estimates  $\hat{\pi}_k$  are equivariant once the denominator is fixed (Hastie et al., 2001; Kutner et al., 2004).

The classifier used in this dissertation is an implementation of the pooled response model (Christensen, 1997: 152) specified in Madigan, Genkin, Lewis, and Fradkin (2005b) and compares  $\mathbf{Y}_k$  to the total of all other classes  $\mathbf{Y}_{k' \neq k}$ , e.g., model

$$(4.8) \quad \log \frac{P(\mathbf{Y}_k)}{\sum_{k' \neq k} P(\mathbf{Y}_{k'})}, \quad k = 1, \dots, K$$

which represents the odds of getting class  $k$  relative to not getting class  $k$ .

The multinomial logistic regression model used in this dissertation is a conditional probability model of the form shown in 4.9 (Madigan et al., 2005b: 1, Equation (1)).

$$(4.9) \quad P(y_k = 1 | \mathbf{x}, \mathbf{B}) = \frac{\exp(\boldsymbol{\beta}_k^T \mathbf{x})}{\sum_{k' \neq k} \exp(\boldsymbol{\beta}_{k'}^T \mathbf{x})}, \quad k = 1, \dots, K$$

The model is parameterized by the matrix  $\mathbf{B} = [\beta_1, \dots, \beta_K]$ , where the columns of  $\beta$  are parameter vectors that correspond to one of the classes  $k$ :  $\beta_k = [\beta_{k1}, \dots, \beta_{kP}]^T$ . That is,

$$\mathbf{B} = \begin{bmatrix} \beta_{11} & \dots & \beta_{k1} & \dots & \beta_{K1} \\ \vdots & & \vdots & & \vdots \\ \beta_{1p} & \dots & \beta_{kp} & \dots & \beta_{Kp} \\ \vdots & & \vdots & & \vdots \\ \beta_{1P} & \dots & \beta_{kP} & \dots & \beta_{KP} \end{bmatrix}$$

Classification of a new instance is based on the vector of probability estimates produced by model (4.9) (or equivalently (4.7)). The class with the highest conditional probability estimate is chosen (Madigan et al., 2005b: 2):

$$\hat{y}(\mathbf{x}) = \underset{k}{\operatorname{argmax}} P(y_k = 1 | \mathbf{x})$$

Estimates for the values of  $\mathbf{B}$  are obtained from the training set via the method of maximum likelihood (e.g., Kutner et al., 2004: 27-32). Maximum likelihood estimation involves choosing values of  $\mathbf{B}$  that are most consistent with the sample data, e.g., the likelihood of  $\mathbf{B}$  given a data set is maximized. The basic idea behind maximum likelihood estimates of  $\mathbf{B}$  involves the fact that each observation  $\mathbf{Y}_i$  is expressed in terms of the expected value of the parameter vector  $\beta_i$  applied to the observed values of  $\mathbf{X}_i$ , i.e.,  $E\{\mathbf{Y}_i\} = \beta_i^T \mathbf{X}_i$  (Equation 4.4).

In the normal regression model, each  $\mathbf{Y}_i$  is assumed to be normally distributed with standard deviation  $\sigma$ . The likelihood of obtaining a particular value of  $\beta_i$  can

be assessed with respect to the probability of seeing that value given a normal distribution with mean  $E\{\mathbf{Y}_i\}$ . Maximum likelihood estimation uses the density of the probability distribution at  $\mathbf{Y}_i$  as an estimate for the probability of seeing that observation. For example, Figure 4.2 shows the densities of the normal distribution for two possible parameterizations of  $\beta_i$ . If  $\mathbf{Y}_i$  is in the tail (4.2b), it will be assigned

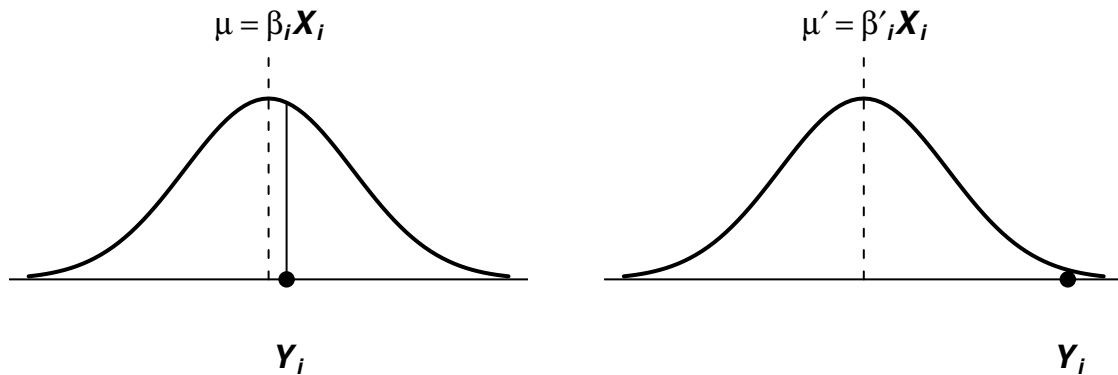


Figure 4.2: Normal probability distribution densities for two possible values of  $\mu$

a low probability of occurring. On the other hand, if it is closer to the center of the distribution (4.2a), it will be assigned a higher probability of occurrence. The method of maximum likelihood estimates for  $\beta_i$  involves choosing values of  $\beta_i$  that favor a value of  $\mathbf{Y}_i$  that is near the center of its probability distribution. The parameters must be optimized over all of the observations in the training sample.

Bayesian approaches to logistic regression involve specifying a distribution on  $\mathbf{B}$  that reflects prior beliefs about about likely values of the parameters. In the typical classification setting involving large data sets in a high dimensional feature space, a reasonable prior distribution for  $\mathbf{B}$  is one that assigns a high probability that most

entries of  $\mathbf{B}$  will have values at 0 (Krishnapuram, Carin, Figueiredo, and Hartemink, 2005; Madigan et al., 2005b). In other words, it is reasonable to expect that many of the features are redundant or noisy, and only a small subset are most important for classification. The goal of such so-called sparse classification algorithms is to learn a model that achieves optimal performance with as few of the original features as possible.

A common choice of prior is the Laplacian (Figure 4.3), which favors values of  $\mathbf{B}$  of 0 (Krishnapuram et al., 2005; Madigan et al., 2005b). The basic idea behind specifying a Laplacian prior on  $\mathbf{B}$  is illustrated in Figure 4.3, which compares the Laplacian distribution to the normal distribution with the same mean and variance. Compared to the normal distribution, the Laplacian is more peaked at the mean,

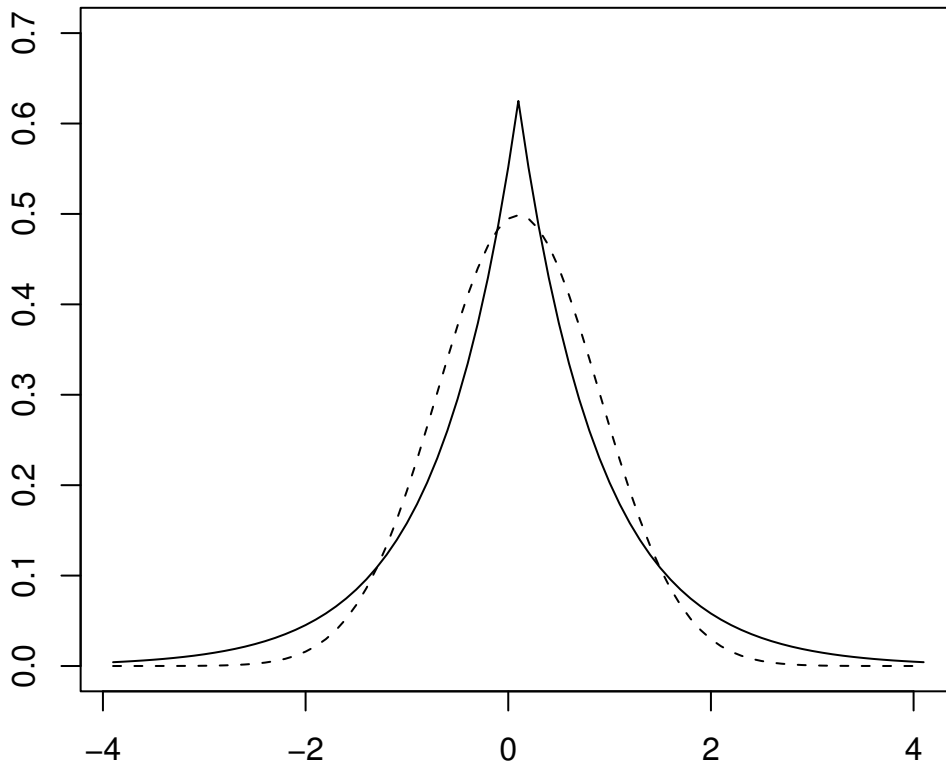


Figure 4.3: Density of the normal (dashed line) and Laplacian distributions with the same mean and variance

while the normal distribution is relatively flat and wide around the mean. When the distribution is relatively flat in the region around the maximum likelihood estimate, the maximum likelihood estimate is not as precise because a large number of values of  $\beta_i$  are nearly as consistent with the training data as the maximum likelihood estimate itself (Kutner et al., 2004: 29-30). Because the Laplacian is sharply peaked about the mean, estimates of  $\beta_i$  that are slightly away from the mean will receive drastically lower probabilities than an estimate of 0. Only those features which receive strong support in the training data will receive a non-zero estimate. Thus, when applied to the original features, automatic feature selection is obtained as a side-effect of training the model (Krishnapuram et al., 2005: 958). The Laplacian prior embodies a bias which allows for efficient model fitting in situations where the number of predictor variables is large and exceeds the number of observations. Because of this property it is expected to be suitable for large scale natural language classification tasks.

There are a number of algorithms in use for fitting regression models. Unlike for ordinary least squares regression, a closed-form analytic solution for training a multinomial regression model does not exist (Kutner et al., 2004; Mitchell, 2006). Instead, iterative methods for finding approximations to the roots of a real-valued function are used (e.g., iteratively reweighted least squares (Krishnapuram et al., 2005)). These methods produce a converging sequence of approximations to the actual root that can be used as approximations of the actual values of  $\mathbf{B}$ . Detailed discussion of the algorithmic details and computational techniques involved in training a logistic regression classifier are provided in Hastie et al. (2001), Gelman et al. (2004), Krishnapuram et al. (2005), and Madigan et al. (2005b).

### 4.3.2 Naive Bayes

For comparison purposes, we also use a naive Bayes classifier in the first experiment below. The motivation for including the naive Bayes classifier is its simplicity and the fact that it is often competitive with more sophisticated models on a wide range of classification tasks (Mitchell, 2006). The naive Bayes classifier is a conditional probability model of the form

$$P(C|F_1, \dots, F_n)$$

where  $C$  stands for the class we are trying to predict and  $F_1, \dots, F_n$  represent the features used for prediction. The class-conditional probabilities can be estimated using maximum likelihood estimates that are approximated with relative frequencies from the training data. Therefore, the conditional distribution over the class variable  $C$  can be written

$$P(C|F_1, \dots, F_n) \approx P(C) \prod_{i=1}^n P(F_i|C)$$

This rewrite is possible only under the assumption that the features are independent. When used for classification, we are interested in obtaining the most likely class given a particular set of values of the input features, i.e.,

$$\text{classify}(f_1, \dots, f_n) = \underset{c}{\operatorname{argmax}} P(C = c) \prod_{i=1}^n P(F_i = f_i|C = c)$$

In the experiments reported here we use a balanced data set (i.e., the same number of English and Korean words) and therefore do not include the prior probability of a word being English or Korean in the model.

## 4.4 Experiments on Identifying English Loanwords in Korean

### 4.4.1 Experiment One

#### 4.4.1.1 Purpose

The purpose of this experiment is to establish classification accuracy for identifying English loanwords in Korean using hand labeled data in a supervised learning scenario. The accuracy obtained with hand labeled data will serve as a target for subsequent experiments which utilize automatically generated training data.

#### 4.4.1.2 Experimental Setup

The data in this experiment consisted of the list of 10,000 English loanwords described in Chapter 2 Section 2.1 and 10,000 Korean words selected at random from the National Institute of the Korean Language’s frequency list of Korean words (NIKL, 2002). No distinction between native Korean and Sino-Korean words was maintained. 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. The count of a given trigram in a single word was rarely more than one. For example, the English loanword *user* is produced in Korean as 유저 *yuce* and is represented as

$$(\emptyset\emptyset y : 1, \emptyset y u : 1, y u - : 1, u - c : 1, -c e : 1, c e \emptyset : 1, e \emptyset \emptyset : 1)$$

where  $\emptyset$  is a special string termination symbol and ‘-’ indicates an orthographic syllable boundary.

The decision to use trigrams instead of syllables as in Oh and Choi (2001) and Kang and Choi (2002) was based on the intuition that segment level transitions provide important cues to etymological class that are lost by only considering syllable transitions. Unigrams or bigrams are not as likely to be sufficiently informative, while going to 4-grams or higher results in severe problems with data sparsity. This feature representation resulted in 2276 total features; English words contained 1431 unique trigrams and Korean words contained on 1939 unique trigrams.

This experiment used a 10-fold, 90/10 train/test split. We report identification accuracy, which is computed as the number of correctly classified words in the test set divided by the total number of words in the test set, averaged over ten trials. Baseline accuracy for all experiments is 50%.

#### 4.4.1.3 Results

Mean classification accuracy using labeled data was 91.1% for the Bayes classifier and 96.2% for the regression classifier. This is expected, in accordance with the observation that discriminative models typically perform better than generative ones (Ng and Jordan, 2002). Taking these results 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.

## 4.4.2 Experiment Two

### 4.4.2.1 Purpose

The purpose of this experiment is to use phonological transliteration rules to generate a set of possible but unattested English loanwords in Korean and train a classifier to automatically distinguish actual English loanwords from actual Korean words.

### 4.4.2.2 Experimental Setup

This experiment applied the phonological rule based transliteration model presented in Chapter 3 Section 3.3.1 to the pronunciations of English words 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. The number of pseudo-English training instances ranged from 10,000 to 100,000. The test items were all 20,000 items from the experiment above. 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. 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.

### 4.4.2.3 Results

Figure 4.4 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.

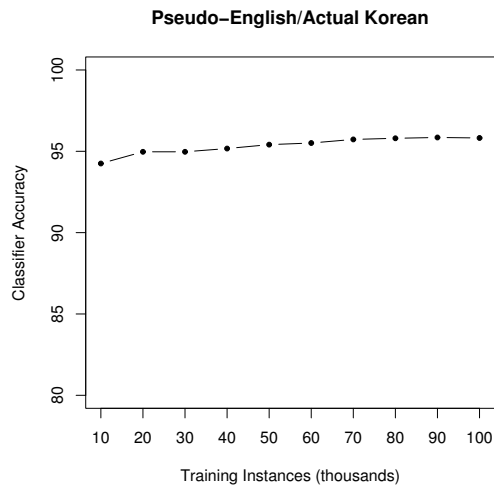


Figure 4.4: Classifier accuracy trained on pseudo-English loanwords and classifying actual English loanwords

While this experiment demonstrates the feasibility of approximating a set of English loanwords with phonological conversion rules, it still relies on a manually constructed dictionary of Korean words. The next experiment investigates the feasibility of approximating a label for the Korean words as well.

### 4.4.3 Experiment Three

#### 4.4.3.1 Purpose

The purpose of this experiment is to examine the performance of the loanword identifier on distinguishing actual English loanwords from actual Korean words when it is trained on pseudo-English loanwords and unlabeled items that serve as examples of Korean words.

#### 4.4.3.2 Experimental Setup

Based on observations of English loanwords in Japanese (Graff and Wu, 1995) and Chinese (Graff, 2007) 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 (Cole and Walker, 2000). 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.

#### 4.4.3.3 Results

Figure 4.5 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.

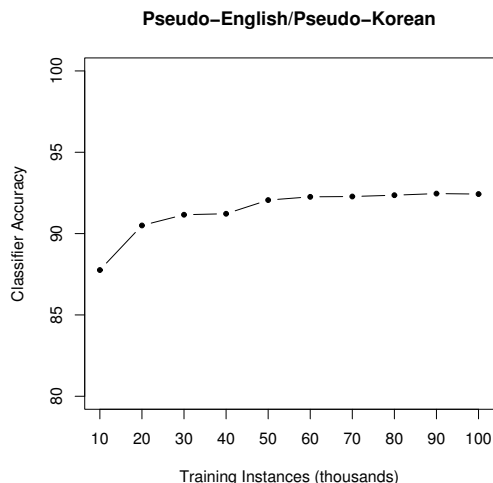


Figure 4.5: Classifier accuracy trained on pseudo-English loanwords and pseudo-Korean items

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 4.1. 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

Word		Rank	Frequency
연합뉴스	Yeonhab News	30	51792
퍼센트	percent	32	49367
뉴욕	New York	89	19652
러시아	Russia	91	19162
클린턴	Clinton	94	18860

Table 4.1: Frequent English loanwords in the Korean Newswire corpus

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.

#### 4.5 Conclusion

The experiments presented here 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. Because the rules describing how words change when they are borrowed from one language to another are relatively few and easy to implement, the methodology outlined here can be widely applied to additional languages for which obtaining labeled training data is difficult.

For example, Khaltar, Fujii, and Ishikawa (2006) describes an approach to identifying Japanese loanwords in Mongolian that is also based on a small number of phonological conversion rules, and Mettler (1993) uses a set of katakana rewrite rules to find English loanwords in Japanese. The current approach is novel in that the identification of loanwords is not limited to those items explicitly generated by

the conversion rules, but generalizes beyond a specific set of input items to identify loanwords that are not contained in the training material. As a point of comparison on the current data set, we can take the performance of the rule-based transliteration models described in Chapter 3 as indicative of a direct rule-based approach to identifying English loanwords on this data set. The phonological rule-based model correctly transliterates (i.e., identifies) about 49% of the loanwords in the data set, and the ortho-phonemic rule-based model finds 78%. The identification model trained on the output of the phonological rule-based model and approximated Korean labels performs about 15% higher than the ortho-phonemic model would, and the model trained on pseudo-English and actual Korean words performs about 18% higher.