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

Kirk Baker
Chris Brew
Ohio State University

LREC 2008
May 28-30, 2008

Guessing Etymology

- Identifying the etymological source of an unknown word is important for:
  - Machine transliteration
  - Text-to-speech synthesis
  - Cross-lingual information retrieval
  - ?Testing theories of language contact?

THE OHIO STATE UNIVERSITY

Prior work

- similar to language identification
- more difficult in that
  - words are (almost always) nativized by the target language phonology
  - words are (often) disguised by rendering in the target language orthography

Why Korean?

- English borrowings comprise substantial proportion of Korean lexicon, esp. for technical or pop-culture writing
- Korean orthography, which is alpha-syllabic, affects how borrowed words are rendered
- Korean phonology, which differs substantially from English, affects how borrowed words are pronounced

Previous work

- Oh and Choi (2001)
  - Hangul strings are modeled by a hidden Markov model where
    - states indicate 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

Previous work

- Kang and Choi (2002) employs a similar Markov-based approach that alleviates the burden of manually syllable tagging a huge corpus, but relies instead on dictionaries that distinguish English and Korean words
- Both these approaches are too much work, and unportable to new languages

Our approach

- train a statistical classifier to distinguish English and Korean words
- use a small number of phonological conversion rules to generate potentially unlimited examples of English-like quasi-borrowings.
- use these as data to train the classifier to distinguish actual English and Korean words

Data

- 10,000 English-Korean loanwords
- 10,000 Korean words
- most of the data are from the National Institute of the Korean Language's publicly available word lists; some loanwords we collected ourselves

THE OHIO STATE UNIVERSITY

Features

- Trigrams with counts (most counts being 1 because words are short)
- example: yu-jeo 'user'
- ##y:1, #yu:1, yu-:1, u-j:1, -je:1, eo#:1, o##:1

Data set

- 2276 total features
- English words contained 1431 unique trigrams
  Korean words contained on 1939 unique trigrams
- baseline for all experiments is 50%

Classifier

- sparse logistic regression classifier
- Bayesian Binary Logistic Regression (Lewis and Madigan, 2005)
- `http://www.stat.rutgers.edu/~madigan/BBR/index.html`

Experiment 1: Labeled Data

- 10,000 (real) English loanwords
- 10,000 (real) Korean words
- 10-fold cross-validation, 90/10 train/test split
- 96.2% classification accuracy
- naïve Bayes classifier gives 91.1% accuracy
- take these figures as a reasonable upper bound for what fake training data might do

Experiment 2: Pseudo-loanwords

- pseudo-English loanwords were generated from the list of phonological rewrite rules given by the Korean Ministry of Culture and Tourism (1995)
- describe the changes English phonemes undergo when they are borrowed into Korean
- these rules were used to hallucinate a set of plausible but unattested English loanwords for Korean

THE OHIO STATE UNIVERSITY

Rules

- 'eu' is inserted after word-final and pre-consonantal voiced stops
  - bulb 'beol-beu'
  - land 'laen-deu'

Examples

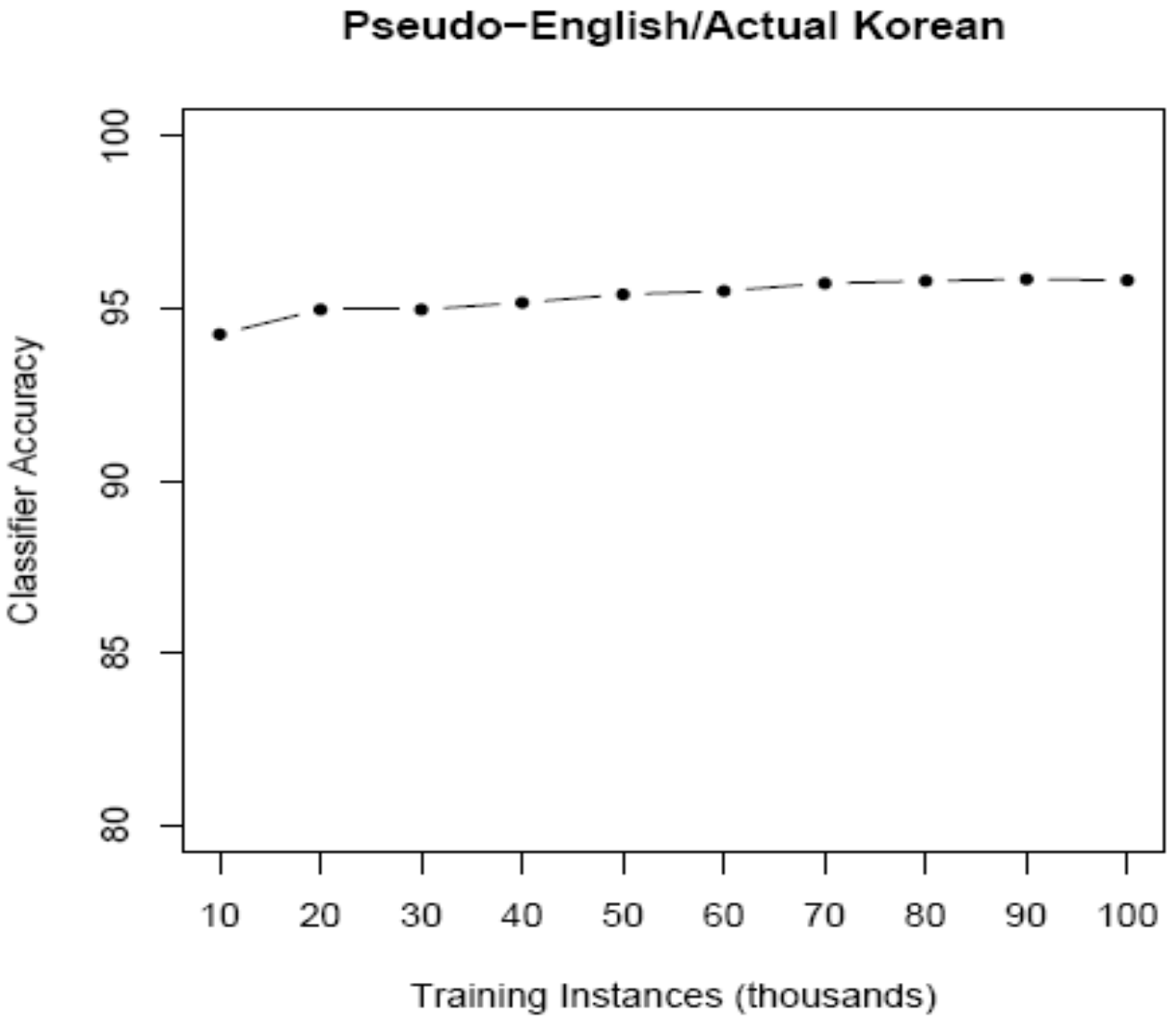
- ExamplesWord-final [S] is written as 'si', preconsonantal [S] is written as 'syu'
  - flash 'peul-lae-si'
  - shrub 'syu-leo-beu'

Implementation

- Total of 30 rules were implemented as regular expressions in a Python script
- applied to the pronunciations in the CMU Pronouncing Dictionary
- Generated up to 100,000 training examples

Method

- Tested on all 20,000 items from first experiment
- Removed training items that occurred in the test set i.e. ones where the rules produced an attested (loan) word
- Classification accuracy asymptoted at around 90,000 instances of each class, within 0.3% (95.8% correct) of the classifier trained on actual English loanwords.


Pseudo−English/Actual Korean
Classifier Accuracy
100
95
90
85
80
10
20
30
40
50
60
70
80
90
100
Training Instances (thousands)

It works

- 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 investigated the feasibility of approximating a label for the Korean words as well.

Experiment 3

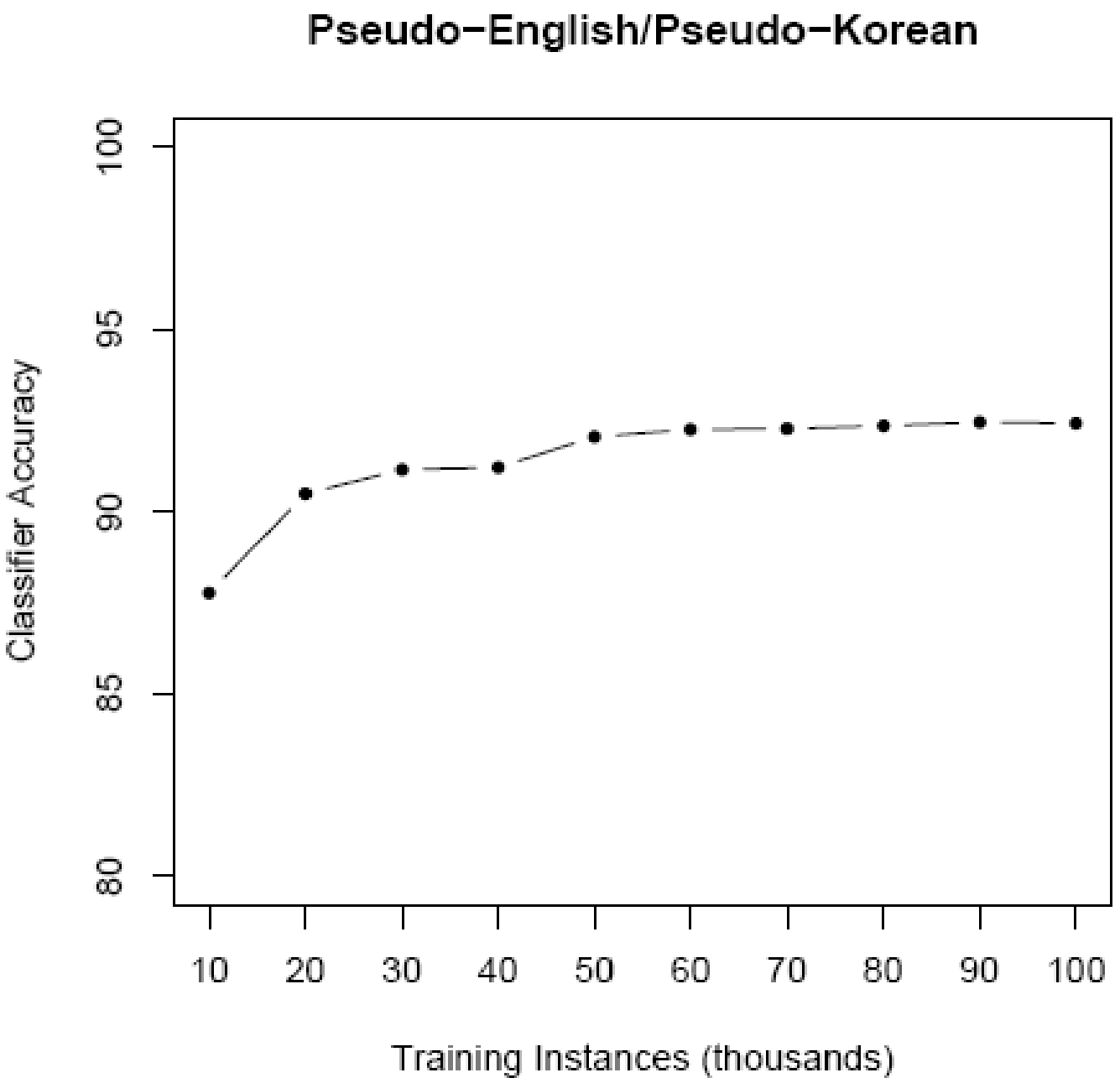
- Based on observations of English loanwords in Japanese and Chinese newswires, we believe that the majority of these items will occur relatively infrequently in comparable Korean text
- we are assuming that there is a relationship between word frequency and the likelihood of a word being Korean, i.e., the majority of English loanwords will occur very infrequently
- we sorted the items in the Korean Newswire corpus 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

Experiment 3

- extracted 23,406,254 Korean orthographic units in the Korean Newswire corpus Cole and Walker (2000)
- Because we believe that high frequency items are more likely to be Korean words, we sampled without replacement from 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

Experiment 3

- 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
- Classifier accuracy asymptoted around 90,000 items per training class at 3.7% below (92.4%) the classifier trained on actual English loanwords.


Pseudo-English/Pseudo-Korean
Classifier Accuracy
100
95
90
85
80
10
20
30
40
50
60
70
80
90
100
Training Instances (thousands)

Data is noisy

- The assumption that frequent items in the Korean Newswire corpus are all Korean is false
- For example, 5 of the 100 most frequent items are English borrowings
- Yeonhab News 30th
- percent 32nd
- New York 89th
- Russia 91st
- Clinton 94

Conclusions

- 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

Conclusions

- These experiments 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

Conclusions

- 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 applied to additional languages for which obtaining labeled training data is difficult

Future

- One way the current research on foreign word identification can be expanded is to consider the identification of borrowings from additional languages
- In practical terms it makes sense to focus on English borrowings because these make up the majority of borrowings in Korean
- However, it would be interesting to look at the performance of an automatic classifier identifying loanwords from multiple languages with respect to the performance of a classifier working with the original source languages

Future

- It would also be interesting to compare the performance of a classifier given a common set of languages but varying the target language
- For example, one could compare the accuracy of a classifier identifying English and Japanese loanwords in Korean versus one identifying English and Korean loanwords in Japanese, etc.
- Work along these lines could be tied to theories of loanword adaptation and the role of phonological systems in perceiving non-native contrasts.