

Foreign Word Classification: A Statistical Model

Joseph Blaylock <jrbl@jrbl.org>

Introduction

How is it that individuals with no conscious knowledge of a language can still classify it relatively accurately? Why can I hear Chinese or Korean and correctly differentiate them from one another, even though I don't speak either? These are the questions that I wanted to answer in my final project. Obviously, I can explain for any single language what features set it apart from others – the Chinese tone system has a quite distinctive sound. That, however, is just the point: how is it that I can tell that Chinese tones are so distinctive? How can I know what the “Chineseness” of Chinese is?

Language is hard. Nobody knows how it works, why it works, or even why it is, exactly. Machine translation, natural language processing, and plain text data mining are all rich areas of active research. Yet, the challenges posed by human language to computer scientists, cognitive scientists, and even linguists seem insurmountable. With this project, I wanted to try to understand one small part of a seemingly vast framework, and satisfy myself that I had found at least a possible explanation.

Motivating me is the idea that language acquisition is at least in part a statistical process. Saffran et al., have made a reasonably strong case that children can be using an essentially statistical process to chunk phoneme streams into individual words. Iverson et al. have suggested that the process of language acquisition actually changes the perceptual apparatus, “tuning” it to one native language. It seems to require a short step, then, to suggest that children effectively statistically profile the sounds that they hear and form generalizations of sounds that belong to their language long before they actually begin to acquire words. If this is the case, then I would expect natural languages to display a significant amount of statistical variance in the selection and arrangement of phonemes that they prefer, and that that variance is

exploitable using relatively unsophisticated tools.

Methods

For this project, I've chosen to take a model-building approach. I wanted to write a small piece of software which could be given a selection of words in a variety of languages and correctly differentiate the languages of each sample. To achieve this, I wrote a straightforward implementation of a naïve Bayesian classifier in the Python programming language, and used it to process the Russian and Polish Swadesh lists, rendered in the International Phonetic Alphabet (IPA). A *run* consists of one instance of dividing the Swadesh lists into testing and training sets, building a frequency table from the training set, and then doing lookups into the frequency table for all of the testing cases in an attempt to classify them. A *test* consists of many hundreds of runs performed concurrently. Across runs in a test, information about how many test set items were classified correctly is aggregated, along with how many test set items were attempted. The ratio of these numbers is called the system's *accuracy*.

The Algorithm

Bayes's Rule is a tool used to find the probabilities that some event belongs to a given classification, given a priori information about the event and the classification. In its most general form

$$P(c_i|F) = P(F|c_i)P(c_i)$$

where c_i is some classification, F is some collection of features,
 $P(x)$ is the probability function

We can make some simplifying assumptions, however. The data sets input are the same size for each classification (we have the same number of Russian and Polish words) so we end up with a uniform distribution of examples. This means that the value of $P(c_i)$ is a constant, and so can be dropped.

This leaves the value of $P(F|c_i)$ which can be calculated in different ways, depending on what kind of statistical power we require. For this model, we're going to use the naïve Bayes rule, which says:

$$P(F|c_i) = \prod_k P(f_k|c_i)$$

*where k is an index into the set of all features F ,
 f_k is some particular feature $\in F$*

This means that once we define what the “features” we want to study actually are, we can ignore the relationships between the features and do a straightforward product of their probabilities to determine the overall probability of a collection of features. Then we can apply the rule of *maxima a posteriori* to select the most likely classification.

The Data

This begs the question, however, of what we're going to consider to be a feature, and more importantly, where we're going to get testing and training data from. I've chosen to use the Swadesh lists, which were developed by the linguist Morris Swadesh in the 1940s and 1950s. They're intended to represent the set of words with the least ambiguity across languages – words which always mean the same thing, such as “and” or “with”. Originally developed for lexicostatistical dating, they fell out of favor in the 1970s and are now mostly used for other tasks, e.g., defining core concepts of all languages. The Swadesh lists are usually presented in, or along with, International Phonetic Alphabet transcriptions of the words for the target languages, so that phonology can be compared. This was particularly important to this project, as differences in orthography make classification of written documents a much easier task (i.e., it would be cheating).

The data set used consisted of the Russian and Polish Swadesh lists, rendered in the International Phonetic Alphabet, sorted in English alphabetical order by their English translations. These data files

were selected because they were readily available, and because I speak neither Russian nor Polish, which I felt would be important to avoid influencing the data. Also, Russian and Polish are very similar languages. Speakers of each are commonly mutually comprehensible to one another. This makes the task of teasing out language features which differentiate the two more challenging. I also considered using Spanish and Portuguese lexicons for the same reasons.

The data files were organized with one word per line of the file, and processed with a word as the fundamental unit. Features were extracted from words as single characters and digraphs of characters. This is because in the IPA, each sound is given a unique symbol, so each character from the input represents a unique sound. I also chose to use a sliding window of two characters to make the set of word features more rich, assuming that different languages may use the same set of phonemes, but different *pairs* of phonemes may be normative in each of them. As an example, then, the word **zwi** has a total of five distinct features: **z**, **w**, **i**, **zw**, **wi**.

Testing vs. Training

In every test of the classifier, the testing and training word sets were built by a process of random selection from the input data. In the reference test, there were 206 training words and 208 testing words. In subsequent tests to determine accuracy-versus-training-set-size, the proportions of the total data set devoted to training decreased from 1/2 to 1/3 to 1/4, etc.

Results

In the reference test of 800 runs, the program was given 166400 testing words to classify. Of these, it gave the correct classification to 141267 of them. This is an overall accuracy of 84.90%. In a

follow-up test, the proportion of the dataset devoted to training was iteratively shrunk, from 1/2 to 1/3, etc. down to 1/15. The results are shown in the table below.

Test No.	Proportion	Test Size	Train Size	Tested	Correct	Correct %
1	1/2	206	208	166400	141336	84.94
2	1/3	138	276	220800	182906	82.84
3	1/4	102	312	249600	203004	81.33
4	1/5	82	332	265600	212620	80.05
5	1/6	68	346	276800	219373	79.25
6	1/7	58	356	284800	224424	78.80
7	1/8	50	364	291200	227263	78.04
8	1/9	46	368	294400	229048	77.80
9	1/10	40	374	299200	231134	77.25
10	1/11	36	378	302400	231541	76.57
11	1/12	34	380	304000	232192	76.38
12	1/13	30	384	307200	232702	75.75
13	1/14	28	386	308800	232916	75.43
14	1/15	26	388	310400	233079	75.09

table 1

Discussion

This numbers show that while the training set size can get very small – in the final case just 13 randomly-selected examples from each language corpus – the overall accuracy of statistical classification remains relatively high. This maps well to the subjective experience of humans, who seem to be able to classify spoken language fairly accurately after only a small number of exposures.

But the correlation to human performance is unclear without reference to studies with human subjects. These results are intriguing, but how good of a model of actual human cognitive processes are

they? Several potential studies come to mind:

- Just how good are humans at learning to classify unfamiliar languages? How many phonemes worth of a language stream are needed to achieve some arbitrary level of accuracy in classification?
- Does the language family of the speaker impact the speed with which they learn to classify arbitrary new languages? That is, do speakers of Slavic languages learn to identify Slavic languages faster than speakers of Semitic languages?
- Statistical learners such as Bayesian classifiers can be manipulated in a variety of ways, by manipulating the training and testing data. For example, one could elect to train on the most statistically characteristic data subset, and achieve artificially inflated accuracy. Or one could elect to train on the most statistically uncharacteristic data subset, and achieve artificially deflated accuracy. Can humans be manipulated in the same ways? That is, when measuring human performance in the classification of an unfamiliar language, can they be made to do better or worse than average by manipulating the samples given to them to characterize the language?

One of the things that made this data set interesting to work with is that Russian and Polish are similar, mutually comprehensible languages. This also provided one of the main frustrations of the data set. 30 of the 207 words (14.49%) were cognates. This means that a significant subset of the incorrect classifications were due simply to words which were trained for one language, but tested for the other. This hardly seems like a “fair” test. In general, it's hard to imagine counting these as errors in a human study. If a person were to say that *i* were Russian, they would be correct. If they were to say that it were Polish, they would also be correct. So what is the uninformed computer model to do? The least sophisticated answer is to simply drop cognates from the testing and training sets, and indeed informal tests imply that a useful increase in accuracy can be achieved by doing so. A more realistic requirement would ask the model to correctly determine when a word was a cognate between the languages under study. This is obviously more complicated.

In further work with this model, it would be useful to generate a larger and more diverse corpus. The Swadesh lists represent a very convenient data set, with valuable orthography neutrality, but it's not

clear what their relationships are to the languages from which they're drawn. Is the Polish Swadesh list highly characteristic of Polish as a whole? Or is it chock full of loan words from other languages? Using larger corpora would help to alleviate these concerns.

Also, the use of IPA is problematic. It represents a “cheat”, in that every phoneme is given a unique character representation. For a human listener attempting to identify unfamiliar language, this is nothing like the data they're given. Instead, they get a relatively unbroken sequence of sounds with no obvious chunking or demarcation between. They have to learn to hear the phoneme clusters of a given language. With IPA representations, that learning is a given. It would be an improvement if phonemes could somehow be represented as a vector in a space of possible sounds, and then presented to the computer as such, so that the classifier would have to deal with noise at the level of phonemes.

Overall, the success of this model implies that at the very least, statistical classifiers could have practical use as a first-pass in a natural language processing engine, to determine which lexicons should be worked with without human intervention. It also provides a useful mental model for thinking about how human language may develop and perhaps shed light on how we begin to form social distinctions (in-group versus out-group determination.) These results only show plausibility, however; they don't necessarily establish a link between human brains and statistical classifiers. The opportunities for future work in this area are rich.

References

This is loosely a works-cited page, but not exactly. R&N would be good, Everything you come up with in the lit. search.

1. Statistical Learning by 8-Month-Old Infants, Jenny R. Saffran; Richard N. Aslin; Elissa L. Newport, *Science*, New Series, Vol. 274, No. 5294. (Dec. 13, 1996), pp. 1926-1928.
2. A Perceptual interference account of acquisition difficulties for non-native phonemes, P. Iverson et al., *Cognition* 87 (2003) B47–B57
3. B351 Class Lecture Notes, Joe Blaylock, 8 November 2006
4. Conversation, Mike Gasser and Joe Blaylock, 21 November 2006
5. “Swadesh list”, Wikipedia, 14 December 2006, http://en.wikipedia.org/wiki/Swadesh_list
6. “Naive Bayes classifier”, Wikipedia, 14 December 2006, http://en.wikipedia.org/wiki/Naive_bayes
7. “Swadesh list of slavic languages”, Wikipedia, 21 November 2006, http://en.wikipedia.org/wiki/Swadesh_list_of_slavic_languages
8. “Polish language/Swadesh list”, Answers.com, 21 November 2006, <http://www.answers.com/topic/polish-language-swadesh-list>
9. Conversation, Rich Knepper and Joe Blaylock, 01 December 2006

Project source code, slides, and this paper are available at:

<http://www.jrbl.org/~jrbl/q270-final/>