NLP Final Project : Building an International Phonetic Alphabet Classifier

Viona Lam, Szeyin Lee

December 15, 2014

1 Introduction

This project is motivated by our interest in the domain of speech recognition. We are interested in how manchine processes the sounds receieved from a speaker, and understands the sounds as natural language. The scope of our project focuses on one particular area of speech recognition: classifying words in International Phonetic Alphabet format into different languages. In other words, given a word in its IPA form that is either English, Chinese or Japanese, our model classifies which language it is based on labeled data.

2 Experimental Setup

2.1 Models

Instead of coding the models ourselves from scratch, we leverage the existing models that have been built and optimized, available in the Stanford JavaNLP Classifier library[1]. A classifier is a machine learning tool that takes data points and classify them into one of k classes. For this project, we are using two classifiers: Naive Bayes Model and Multinomial Logistic Regression Model. For both of the models, we use the same set of features: unigram, bigram, trigram, and counts.

2.1.1 Model 1: Naive Bayes (NB)

Naive Bayes is a generative model that models the joint distribution of p(label, features). It is important to note that it has the naive Bayes assumption that the features are independent given the class.

2.1.2 Model 2: Multinomial Logistic Regression (MLR)

Multinomial Logistic Regression is a discriminative model. It models the conditional distribution p(label—features) directly and does not have the strong independent assumptions as Naive Bayes.

[1] http://nlp.stanford.edu/software/classifier.shtml

2.2 Data

2.2.1 English Data

English data of 44,460 of the most common words from the NY Times was obtained from the Bag of Words Dataset provided by UC Irvine[2]. These were then converted into IPA form using Tom Brondstod's English to IPA converter.[3] To ensure accuracy, the results were also partially cross-checked with another converter.[4] In addition, as English speakers, the IPA sounded right to us.

2.2.2 Mandarin Chinese Data

Mandarin Chinese data of 61,698 of the most common words was obtained from the Modern Chinese Frequent Vocab List[5] a text published by the People's Republic of China's State Language Commission in November 2008. These were then converted into pinyin using the NJStar software[6], and then into IPA form by referring to available documentation[7] from cjklib[8], a publicly accessible python library. Tones were stripped from the conversions, as they would make the decision of whether a word is Chinese or not incredibly trivial. The results were partially cross-checked with another converter[9] to ensure accuracy.

2.2.3 Japanese Data

Japanese Data of 123,332 of the most common words was obtained from a selection of Japanese novels online[10], as processed by Michiel Kamermans. These were then converted into IPA form by referring to available documentation[11][12] and our understanding of the nature of how Japanese characters are represented[13]. To maintain the equal distribution of sample data from each language, we randomly selected 60,000 from the processed data to operate on.

2.3 Evaluation

For all datasets, we randomized the data, reserve 10% as the testing set, and train the models with the remaining 90% data points. We compared the label output from both Naive Bayes and Multinomial Logistic Regression models,

[2] **Bag of Words Data Set:** https://archive.ics.uci.edu/ml/datasets/Bag+of+Words

[3] **English to IPA Converter 1:** http://tom.brondsted.dk/text2phoneme/

[4] **English to IPA Converter 2:** http://lingorado.com/ipa/

[5] **Modern Chinese Frequent Vocab List:** http://vdisk.weibo.com/s/ueoM8g6c-sm2o (click the blue button with the downwards arrow to download)

[6] **Chinese Word Processor:** http://www.njstar.com/cms/njstar-chinese-word-processor-download

[7] **Pinyin to IPA Mapping:** https://github.com/cburgmer/cjklib/blob/master/cjklib/data/pinyinipamapping.csv

[8] **CJKLib:** https://code.google.com/p/cjklib/

[9] **Limited Chinese to IPA Converter:** http://easypronunciation.com/en/chinese-pinyin-phonetic-transcription-converter

[10] **Most Common Words in Japanese Novels:** http://pomax.nihongoresources.com/index.php?entry=1222520260

[11] **Katakana to IPA mapping:** http://en.wikipedia.org/wiki/Transcription_into_Japanese

[12] **Additional Japanese IPA information:** http://en.wikipedia.org/wiki/Help:IPA_for_Japanese

[13] **Regarding Katakana to IPA:** The data set contained the words in their original forms, as well as parsed Katakana representations. As Katakana is a Japanese syllabary, it is not complicated to convert from Katakana to IPA. The hardest part was getting Windows to cooperate with UTF-8.

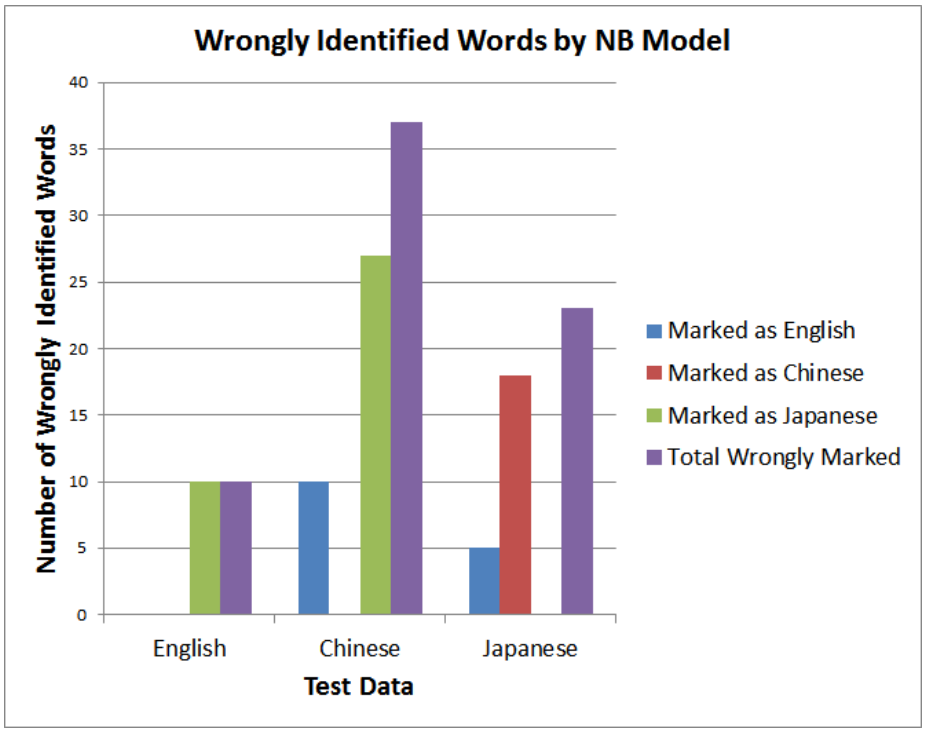
with the known label from testing set. The baseline is 33.33% (correct by guessing randomly one of the three languages).

3 Results

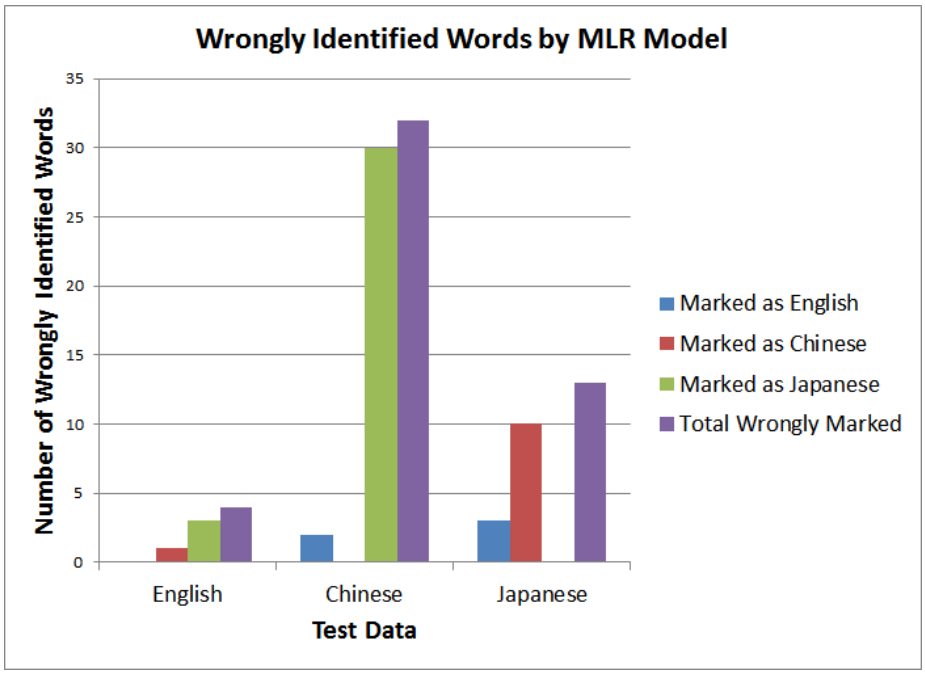
Having earlier done a smaller scale study[14] where we obtained 98% accuracy, we were looking to improve by using more data, and did so by using the above-mentioned data sets.

On the larger data set, with all three langauges and an overall of 155,588 training examples, and 16,570 testing examples, we obtained 99.58% accuracy for the Naive Bayes model, and 99.70% for the Multinomial Logistic Regression model. Our results are significantly above baseline of 33% and almost approaches 100%.

The raw output is found in the Appendix. The following are two graphs that describe the mistakes that the model made on classifying data.



[14]**Status Report:** https://github.com/violxy/nlpfinal/blob/master/StatusReport.txt



As can be seen, most of these errors come from mistakenly classifying Chinese as Japanese, and vice versa. This makes sense, considering how close some Japanese words are to Chinese words, having partially originated from there. Furthermore, the phonemes used in Japanese do have a large overlap with those of Chinese.

4 Conclusion

It is possible to differentiate between languages using simple linear classifiers such as Naive Bayes and Multinomial Logistic Regression to a high level of accuracy (95%). Given more time and easier access to IPA converters, we could expand this project to other languages.

5 Code

The code for this project can be found at the shared Github repository.[15]

6 Appendix

Raw terminal output:

[15] **Github Repository:** https://github.com/violxy/nlpfinal/

```
Incorrect: manman   chinese classified as : japanese
Incorrect: ɕii  chinese classified as : japanese
Incorrect: moni chinese classified as : japanese
Incorrect: fɛɪpɑŋ   chinese classified as : english
Incorrect: tanin    chinese classified as : japanese
Incorrect: paikei   chinese classified as : japanese
Incorrect: aɪjō chinese classified as : english
Incorrect: inman    chinese classified as : japanese
Incorrect: minnan   chinese classified as : japanese
Incorrect: inei chinese classified as : japanese
Incorrect: ii   chinese classified as : japanese
Incorrect: mən  chinese classified as : english
Incorrect: namo chinese classified as : japanese
Incorrect: əi   chinese classified as : english
Incorrect: wɑwɑ chinese classified as : english
Incorrect: maiin    chinese classified as : japanese
Incorrect: matai    chinese classified as : japanese
Incorrect: paimi    chinese classified as : japanese
Incorrect: itai chinese classified as : japanese
Incorrect: iɕinii   chinese classified as : japanese
Incorrect: ɕii  chinese classified as : japanese
Incorrect: i    chinese classified as : japanese
Incorrect: ərmu chinese classified as : english
Incorrect: ama  chinese classified as : japanese
Incorrect: ii   chinese classified as : japanese
Incorrect: fənə chinese classified as : english
Incorrect: iin  chinese classified as : japanese
Incorrect: nanmin   chinese classified as : japanese
Incorrect: i    chinese classified as : japanese
Incorrect: ɕiin chinese classified as : japanese
Incorrect: sanman   chinese classified as : japanese
Incorrect: i    chinese classified as : japanese
Incorrect: fɛɪlaɪ   chinese classified as : english
Incorrect: aita chinese classified as : japanese
Incorrect: tsni chinese classified as : english
Incorrect: keii chinese classified as : japanese
Incorrect: fənfeɪ   chinese classified as : english
Incorrect: net  english classified as : japanese
Incorrect: i:   english classified as : japanese
Incorrect: h    english classified as : japanese
Incorrect: meg  english classified as : japanese
Incorrect: bek  english classified as : japanese
Incorrect: hek  english classified as : japanese
Incorrect: bi:set   english classified as : japanese
Incorrect: ni:  english classified as : japanese
Incorrect: hen  english classified as : japanese
Incorrect: kanpan   japanese classified as : chinese
Incorrect: de   japanese classified as : english
Incorrect: enden    japanese classified as : english
Incorrect: in   japanese classified as : chinese
Incorrect: kanpan   japanese classified as : chinese
Incorrect: ɕi   japanese classified as : chinese
Incorrect: an   japanese classified as : chinese
Incorrect: tenten   japanese classified as : english
Incorrect: de   japanese classified as : english
Incorrect: ɕian japanese classified as : chinese
Incorrect: iɕi  japanese classified as : chinese
Incorrect: ɕini japanese classified as : chinese
Incorrect: ɕiɕin    japanese classified as : chinese
Incorrect: ɕin  japanese classified as : chinese
Incorrect: nanɕin   japanese classified as : chinese
Incorrect: ɕinɕi    japanese classified as : chinese
Incorrect: ɕian japanese classified as : chinese
Incorrect: pen  japanese classified as : english
Incorrect: intai    japanese classified as : chinese
Incorrect: n    japanese classified as : chinese
Incorrect: n    japanese classified as : chinese
Incorrect: ɕintai   japanese classified as : chinese
Incorrect: sanɕi    japanese classified as : chinese
NB Accuracy: 0.995835847917924

MLR Log prob: -107.20564227801076
NB Log prob: -337.0368341477132
```

```
Incorrect: ɕii  chinese classified as : japanese
Incorrect: moni chinese classified as : japanese
Incorrect: ɕi   chinese classified as : japanese
Incorrect: tanin    chinese classified as : japanese
Incorrect: paikei   chinese classified as : japanese
Incorrect: arjō chinese classified as : english
Incorrect: inei chinese classified as : japanese
Incorrect: ii   chinese classified as : japanese
Incorrect: namo chinese classified as : japanese
Incorrect: maiin    chinese classified as : japanese
Incorrect: matai    chinese classified as : japanese
Incorrect: itai chinese classified as : japanese
Incorrect: iɕinii   chinese classified as : japanese
Incorrect: ɕii  chinese classified as : japanese
Incorrect: i    chinese classified as : japanese
Incorrect: ama  chinese classified as : japanese
Incorrect: ɕi   chinese classified as : japanese
Incorrect: ii   chinese classified as : japanese
Incorrect: tanta    chinese classified as : japanese
Incorrect: tanpai   chinese classified as : japanese
Incorrect: iin  chinese classified as : japanese
Incorrect: iɕin chinese classified as : japanese
Incorrect: i    chinese classified as : japanese
Incorrect: ɕiin chinese classified as : japanese
Incorrect: sanman   chinese classified as : japanese
Incorrect: i    chinese classified as : japanese
Incorrect: aiia chinese classified as : japanese
Incorrect: fɛɪlaɪ   chinese classified as : english
Incorrect: aita chinese classified as : japanese
Incorrect: inpei    chinese classified as : japanese
Incorrect: keii chinese classified as : japanese
Incorrect: iɕin chinese classified as : japanese
Incorrect: i:   english classified as : japanese
Incorrect: meɪaɪən  english classified as : chinese
Incorrect: ðiːiː    english classified as : japanese
Incorrect: hen  english classified as : japanese
Incorrect: manman   japanese classified as : chinese
Incorrect: in   japanese classified as : chinese
Incorrect: innai    japanese classified as : chinese
Incorrect: ɕian japanese classified as : chinese
Incorrect: ɕini japanese classified as : chinese
Incorrect: nanɕin   japanese classified as : chinese
Incorrect: ɕian japanese classified as : chinese
Incorrect: pen  japanese classified as : english
Incorrect: imapo    japanese classified as : chinese
Incorrect: n    japanese classified as : english
Incorrect: taimei   japanese classified as : chinese
Incorrect: n    japanese classified as : english
Incorrect: nannan   japanese classified as : chinese
MLR Accuracy: 0.9970428485214242
```