

NLP Final Project : Building an International Phonetic Alphabet Classifier

Viona Lam, Szeyin Lee

December 10, 2014

1 Introduction

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 library¹. 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 $p(\text{label}, \text{features})$. 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 $p(\text{label}|\text{features})$ 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 Irvine². These were then converted into IPA form using Tom Brondsted’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 List⁵ 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 software⁶, and then into IPA form by referring to available documentation⁷ from cjklb⁸, 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 converter⁹ 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 online¹⁰, 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 represented¹³. 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/cjklb/blob/master/cjklb/data/pinyinipamapping.csv>

⁸**CJKLib:** <https://code.google.com/p/cjklb/>

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.

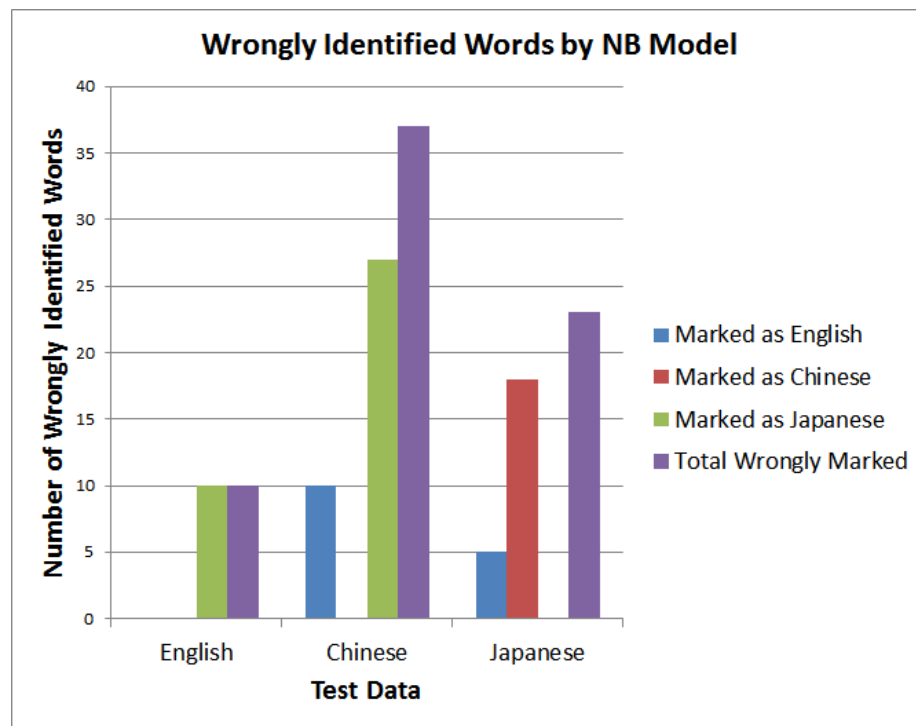
with the known label from testing set. The baseline is 33.33% (correct by guessing randomly one of the three languages).

3 Results

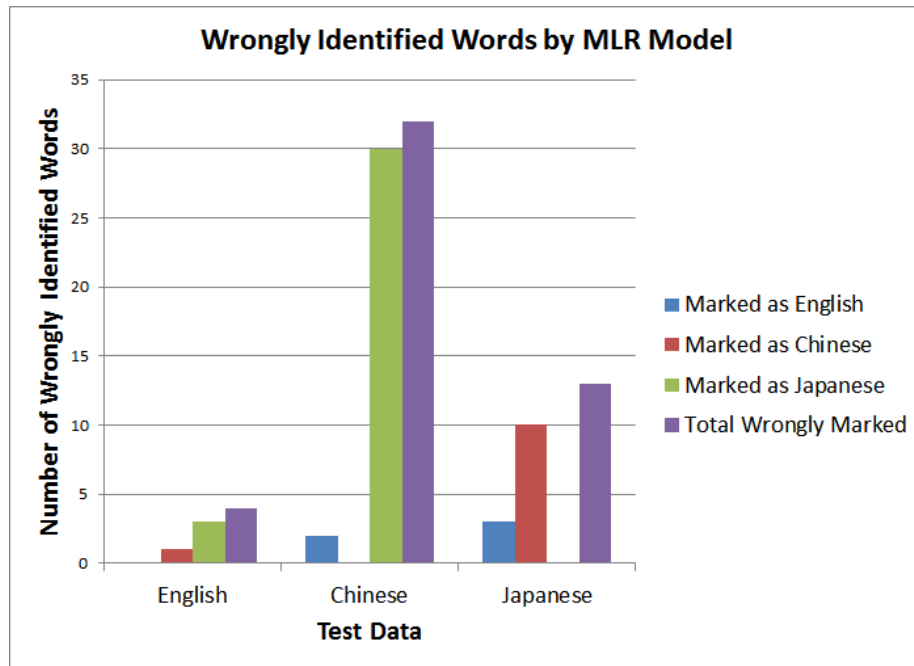
Having earlier done a smaller scale study¹⁴ where we obtained 98% accuracy, we were looking to improve by using more data, and did so by using the above-mentioned data sets.

On the larger data set, with all three languages 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.



¹⁴Status 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/>

1 Incorrect: manman chinese classified as : japanese
2 Incorrect: ġii chinese classified as : japanese
3 Incorrect: moni chinese classified as : japanese
4 Incorrect: feipəŋ chinese classified as : english
5 Incorrect: tanin chinese classified as : japanese
6 Incorrect: paikēi chinese classified as : japanese
7 Incorrect: arjō chinese classified as : english
8 Incorrect: inman chinese classified as : japanese
9 Incorrect: minnan chinese classified as : japanese
10 Incorrect: inei chinese classified as : japanese
11 Incorrect: ii chinese classified as : japanese
12 Incorrect: mən chinese classified as : english
13 Incorrect: namo chinese classified as : japanese
14 Incorrect: əi chinese classified as : english
15 Incorrect: wawa chinese classified as : english
16 Incorrect: maiin chinese classified as : japanese
17 Incorrect: matai chinese classified as : japanese
18 Incorrect: paimi chinese classified as : japanese
19 Incorrect: itai chinese classified as : japanese
20 Incorrect: iġinii chinese classified as : japanese
21 Incorrect: ġii chinese classified as : japanese
22 Incorrect: i chinese classified as : japanese
23 Incorrect: ərmu chinese classified as : english
24 Incorrect: ama chinese classified as : japanese
25 Incorrect: ii chinese classified as : japanese
26 Incorrect: fənə chinese classified as : english
27 Incorrect: iin chinese classified as : japanese
28 Incorrect: nanmin chinese classified as : japanese
29 Incorrect: i chinese classified as : japanese
30 Incorrect: ġiin chinese classified as : japanese
31 Incorrect: sanman chinese classified as : japanese
32 Incorrect: i chinese classified as : japanese
33 Incorrect: ferlar chinese classified as : english
34 Incorrect: aita chinese classified as : japanese
35 Incorrect: tsni chinese classified as : english
36 Incorrect: keii chinese classified as : japanese
37 Incorrect: fənfer chinese classified as : english
38 Incorrect: net english classified as : japanese
39 Incorrect: i: english classified as : japanese
40 Incorrect: h english classified as : japanese
41 Incorrect: meg english classified as : japanese
42 Incorrect: bek english classified as : japanese
43 Incorrect: hek english classified as : japanese
44 Incorrect: bi:set english classified as : japanese
45 Incorrect: ni: english classified as : japanese
46 Incorrect: hen english classified as : japanese
47 Incorrect: kanpan japanese classified as : chinese
48 Incorrect: de japanese classified as : english
49 Incorrect: enden japanese classified as : english
50 Incorrect: in japanese classified as : chinese
51 Incorrect: kanpan japanese classified as : chinese
52 Incorrect: ġi japanese classified as : chinese
53 Incorrect: an japanese classified as : chinese
54 Incorrect: tenten japanese classified as : english
55 Incorrect: de japanese classified as : english
56 Incorrect: ġian japanese classified as : chinese
57 Incorrect: iġi japanese classified as : chinese
58 Incorrect: ġini japanese classified as : chinese
59 Incorrect: ġiġin japanese classified as : chinese
60 Incorrect: ġin japanese classified as : chinese
61 Incorrect: nanġin japanese classified as : chinese
62 Incorrect: ġinġi japanese classified as : chinese
63 Incorrect: ġian japanese classified as : chinese
64 Incorrect: pen japanese classified as : english
65 Incorrect: intai japanese classified as : chinese
66 Incorrect: n japanese classified as : chinese
67 Incorrect: n japanese classified as : chinese
68 Incorrect: ġintai japanese classified as : chinese
69 Incorrect: sanġi japanese classified as : chinese
70 NB Accuracy: 0.995835847917924
71
72 MLR Log prob: -107.20564227801076
73 NB Log prob: -337.0368341477132

1 Incorrect: ġii chinese classified as : japanese
2 Incorrect: moni chinese classified as : japanese
3 Incorrect: ġi chinese classified as : japanese
4 Incorrect: tanin chinese classified as : japanese
5 Incorrect: paikēi chinese classified as : japanese
6 Incorrect: arjō chinese classified as : english
7 Incorrect: inei chinese classified as : japanese
8 Incorrect: ii chinese classified as : japanese
9 Incorrect: namo chinese classified as : japanese
10 Incorrect: main chinese classified as : japanese
11 Incorrect: matai chinese classified as : japanese
12 Incorrect: itai chinese classified as : japanese
13 Incorrect: iġinii chinese classified as : japanese
14 Incorrect: ġii chinese classified as : japanese
15 Incorrect: i chinese classified as : japanese
16 Incorrect: ama chinese classified as : japanese
17 Incorrect: ġi chinese classified as : japanese
18 Incorrect: ii chinese classified as : japanese
19 Incorrect: tanta chinese classified as : japanese
20 Incorrect: tanpai chinese classified as : japanese
21 Incorrect: iin chinese classified as : japanese
22 Incorrect: iġin chinese classified as : japanese
23 Incorrect: i chinese classified as : japanese
24 Incorrect: ġiin chinese classified as : japanese
25 Incorrect: sanman chinese classified as : japanese
26 Incorrect: i chinese classified as : japanese
27 Incorrect: aiia chinese classified as : japanese
28 Incorrect: ferlar chinese classified as : english
29 Incorrect: aita chinese classified as : japanese
30 Incorrect: inpei chinese classified as : japanese
31 Incorrect: keii chinese classified as : japanese
32 Incorrect: iġin chinese classified as : japanese
33 Incorrect: i: english classified as : japanese
34 Incorrect: merarən english classified as : chinese
35 Incorrect: ōi:i: english classified as : japanese
36 Incorrect: hen english classified as : japanese
37 Incorrect: manman japanese classified as : chinese
38 Incorrect: in japanese classified as : chinese
39 Incorrect: innai japanese classified as : chinese
40 Incorrect: ġian japanese classified as : chinese
41 Incorrect: ġini japanese classified as : chinese
42 Incorrect: nanġin japanese classified as : chinese
43 Incorrect: ġian japanese classified as : chinese
44 Incorrect: pen japanese classified as : english
45 Incorrect: imapo japanese classified as : chinese
46 Incorrect: n japanese classified as : english
47 Incorrect: taimei japanese classified as : chinese
48 Incorrect: n japanese classified as : english
49 Incorrect: nannan japanese classified as : chinese
50 MLR Accuracy: 0.9970428485214242