NLP Final Project

Viona Lam, Szeyin Lee

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

Due to ease of use, and relevance to our project goal of classifying words into their source languages, we chose to use the linear classifiers available in the ___ library[1], ___ and ___.

2.1.1 Model 1: Add stuff here

2.1.2 Model 2: Add stuff here

2.2 Data

2.2.1 English Data

English data of 44460 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 38285 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

[1] Add Source Here

[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)

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 123332 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].

2.3 Evaluation

Talk about how the accuracy thing is calculated in the algorithms? I don't think it's super complicated? :P

3 Results

We ran ____ models on _____ of each of the data sets. We randomized each data set, took __% of it for training the models, and __% for testing purposes. 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 test file of ___ words, we obtained __% accuracy for the ___ model, and __% for the ___ model. Our results are ____

Insert Graphs here. See spreadsheet file.

[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.

[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 ___ and ___. Given more time and easier access to IPA converters, we could expand this project to other languages.

Add more here. :D

5 Code

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

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