Sophie Wu (261010381)
Joey Chuang (260893876)
Anita Zheng (260955154)

COMP 550 FINAL PROJECT PROPOSAL

Applying Character Embeddings for Analyzing Homophone Phenomena in Chinese

The effect of homophones is a highly contested topic in linguistics, with various studies debating exactly how homophones impact the development and applied usage of languages. Chinese is an exceptionally homophone dense language — in Standard Mandarin, for example, 1,600 unique sounds cover an estimated 100,000 existing Chinese characters, and our project aims to use **word2vec character embeddings to test various linguistic hypotheses about homophone related phenomena in Chinese.**

We first aim to **test whether Chinese speakers attempt to avoid linguistic ambiguity by placing homophone characters farther apart in their use of the language.** To investigate this, **we will produce vector embeddings of Chinese characters using word2vec and evaluate differences between embedded homophone distances and a predetermined baseline**. If, on average, the distances between homophone mate word embeddings is significantly larger than the baseline distance, this would suggest that individuals do tend to avoid linguistic ambiguity in the context of homophone characters. This is due to the nature of word2vec, which generates its vector embeddings based on the distributional hypothesis, and thus gives words used in similar contexts word embeddings a lower distance from each other. We will compare the distance between our homophone pair, $d(h,\ h')$, to our proposed baseline $mean(d(h,\ c_i'))$, where we choose c_i' from a set of characters with the same ranked frequency in our corpus as h'. This method normalizes our results in terms of character frequency, since embeddings for characters with high frequency are more likely to exhibit smaller distances to all other characters.

To produce our character embeddings, we will use the word2vec algorithm built into the Gensim library. In order to organize characters by their oral sounds, we will preprocess our data using the pypinyin package. We plan on using a variety of Chinese datasets, including the MAGICDATA Mandarin Chinese Read Speech Corpus, which includes a variety of native speaker oral transcriptions, the LCSTS Dataset, which consists of short texts from the popular microblogging website Sina Weibo, and the HKCanCor, which includes transcripts from various spoken Cantonese sources. **By varying the test corpora for which we produce our embeddings, we can also observe whether our homophone-baseline comparisons change depending on the context of the language (i.e. spoken versus written, Cantonese versus Mandarin).**

Since Chinese is a language which may have evolved to reduce homophones through diminutive markers , we also plan on looking at clusters of words around homophones to explore the correlation between monosyllabic homophones and disyllabic words with the same start character, examining whether disyllabic words sharing the same initial homophone character, such as 'huo' (火fire) and 'huo' (活alive), exhibit similar distribution patterns, and whether a clear boundary can be established among disyllabic words starting with these homophone characters. One possible approach is to identify all disyllabic words beginning with the same pinyin and to look for distinct clusters of words within this set. Hopefully, we can verify if all the distinction sets represent one homophone character in Chinese, for example, if one cluster set is huo (fire), and the cluster of another set represents huo (alive).

Finally, we will implement our word2vec model as a basis for a text prediction engine, and **compare the perplexity to a similar model which uses a romanized pinyin version of our corpora (and so does not distinguish between homophones)**. Since larger sample spaces tend to have higher perplexity, we will use perplexity-per-word (PPW) to normalize and compare our results. Using this information theory approach, we can estimate the effect of homophones on Chinese speaker's understanding of the language (i.e. a higher PPW for the homophone-free model would indicate that homophones may actually help speakers learn context more efficiently).

Since our project can take on a variety of different methods right now, we are open to any possible critiques and suggestions for our work, especially on recommendations for what to focus on. Thank you!