EasyChinese: Making Chinese Writing Phonetic

Kai Xu | Advisor: Professor Robert Frank

Chinese is one of the few languages that is exceedingly difficult to learn to read and write. For most languages, much of reading proficiency can be achieved by learning a small set of phonemes and their visual representations. However, learning to read and write Chinese involves memorizing thousands of characters, posing a significant learning barrier.

To assist with the pronunciation of Chinese, a phonetic tool known as pinyin was created. For instance, the phrase 你好 (hello) can be written in Pinyin as nǐhǎo. So why not write Chinese in all pinyin? The main reason is that one pronunciation can correspond to many characters. For instance, the pinyin nǐ corresponds to the character 你 (you), but also the characters (nihonium) and 拟 (to plan). So using purely pinyin will result in semantic confusion. Another reason pinyin is suboptimal is its lack of efficiency as a writing system. For instance, the phrase chángjiāng (Yangtze River) is 10 characters long, with a combined 21 strokes, despite being only two syllables.

But what if we can reinvent the Chinese writing system in a way that retains the readability of Pinyin, but also allows for semantic distinguishing? I aim to achieve this by assigning each Chinese character to a semantic “category,” allowing every Chinese character to be replaced by a combination of pronunciation and category. So for instance, the phrase 我爱你 (I love you) could be notated as something similar to this:

Character	我	爱	你
English	I	love	you
Pronunciation	wǒ	ài	nǐ
Category	person	emotion	person

Of course, when actually writing under this new system, we wouldn’t be writing “person” next to the pronunciation. Rather, we would use a concise symbol to represent the “person” category, and another symbol to represent the “emotion” category. Under this new system, a character might be broken into top and bottom components. The top could represent the pronunciation, possibly annotated with pinyin or perhaps an even more streamlined and graceful notation. Meanwhile, the bottom would bear the symbol indicative of category. Crafting such a design is an artist's task; my role is to define the categories.

Likewise, the phrase 路边有鹿 (there’s a deer on the side of the road) could potentially be notated:

Character	路	边	有	鹿
English	road	side	has	deer

Pronunciation	lù	biān	yǒu	lù
Category	infrastructure	position	auxiliary verb	animal

The combination of pronunciation and category allows us to distinguish between lù (road) and lù (deer). Now, perhaps one can say that the word “road” falls under the category “geography” rather than “infrastructure,” which is a valid point. Creating the categories is in some ways a science and in other ways an art. There is no “right” way to create the categories, although some partitions would provide a “better fit” than others. Too few categories, and there would be multiple characters with the same category for a given pronunciation. Too many, and memorization will be difficult, defeating the point. The goal is to come up with the right categories, and the right number of categories.

A potential challenge is the prevalence of polyphones in Chinese. A polyphone is a character with pronunciations, often with very different meanings. For instance, the character 乐 could be pronounced lè and mean “happy,” or be pronounced yuè and mean “music.” I am currently unsure how to resolve such instances. One option is to create two separate pronunciation-category combos that map to the same character. Another option is to use only one pronunciation-category combo corresponding to the most common usage, adding an annotation denoting a change in pronunciation if necessary.

To complete my task, I plan to try the following techniques, or a combination of them, spaced out relatively evenly during the 2.5 month period I have to complete this project:

- Statistical analysis of modern Chinese writing in Python: detecting patterns such as Zipf’s law, stroke count and usage correlation, and potential stroke count inefficiencies with current Chinese writing systems.

- WordNet: a tree structure that organizes Chinese characters into a hierarchy of is-a relationships. A subtree of a WordNet could theoretically correspond to a category. In the best-case scenario, this is all we need, but such a scenario is incredibly unlikely. Most likely, the WordNet would provide insights into possible ways to cluster the characters, but more work would be necessary.

- Unsupervised learning: run clustering algorithms such as K-means, hierarchical clustering, and DBSCAN on Chinese character embedding vectors. The anticipated result would not be fully usable, but could potentially provide some valuable insights into which clusters to use.

- Semi-supervised learning: develop an algorithm that learns your clustering patterns and aids with manual clustering. For instance, I know that the characters for “dog,” “cat,” and “cow” would probably fit under the same category for “animal,” so I would assign them to the same preliminary category. The algorithm would recommend other characters that go to the same category, such as “snake” or “lion.”

- Large language models: another option is to feed characters into a large language model, asking it to generate categories for the characters. We could potentially ask several LLMs (or ask one LLM using different prompts) to generate categories to classify a huge list of characters. Then, we can look at which categories are generated by many different models, which would most likely be the most reasonable categories. After first finding each category, we could call the LLM for every character, asking the LLM to place it in the appropriate category. Or, we can do manual clustering.

- (Possible) Crowdsourcing categories: perhaps it may be interesting to post the preliminary categories and character assignments to an online platform. This platform would allow users to vote whether a character is in the right category or not. This could potentially be used to correct errors, especially with very obscure characters.