Homophone2Vec: Embedding Space Analysis for Empirical Evaluation of Phonological and Semantic Similarity

Sophie Wu and **Anita Zheng** and **Joey Chuang**
McGill University
sophie.wu@mail.mcgill.ca
luo.b.zheng@mail.mgcill.ca
ching-i.chuang@mail.mcgill.ca

Abstract

This paper introduces a novel method for empirically evaluating the relationship between the phonological and semantic similarity of linguistic units using embedding spaces. Chinese character homophones are used as a proof-of-concept. We employ cosine similarity as a proxy for semantic similarity between characters, and compare relationships between phonologically-related characters and baseline characters (chosen as similar-frequency characters). We show there is a strongly statistically significant positive semantic relationship among different Chinese characters at varying levels of sound-sharing. We also perform some basic probing using t-SNE and UMAP visualizations, and indicate directions for future applications of this method.

1 Introduction

Homophones – linguistic units with the same sound but different meanings – evidently produce semantic ambiguity within language. However, certain functional linguistic theories suggest that ambiguity may actually allow for greater linguistic efficiency, by enabling language learners to better use finite phonological space (i.e. limitations on the word length and sounds in a language) (Piantadosi et al., 2012; Wasow et al., 2005). It remains uncertain to what degree linguistically ambiguous input such as homophones require highly differentiable semantic/syntactic contexts for processing, or whether this is generalizable across languages. Studies in French and English have indicated that homophony may have either an insignificant or inhibitory effect on language processing(Ferrand and Grainger, 2003; Rubenstein et al., 1971). Fieldwork in these languages have also shown that homophones with different syntactic contexts and semantic meanings are easier for children to memorize (Dautriche et al., 2018). These studies widely assert that homophones should have different semantic and syntactic functions to survive in a language. However, in Chinese, a language where many characters have high frequency homophone mates, studies have actually indicated that semantically similar homophones can be facilitate lexical decision-making (Chen et al., 2009), and acquisition of new words (Liu and Wiener, 2020).

Do homophones necessitate high semantic dissimilarity? This paper proposes a novel method of using word embedding spaces to empirically investigate this question by using embedding space properties to determine statistically significant relationships between phonological and semantic similarity. Our method involves comparing the cosine similarity between embeddings of homophone pairs to baseline similarities, where we find baselines using similar-frequency characters. We choose a pre-trained embedding space optimized to encode both semantic and syntactic information, and then use this space to look for a statistically significant difference in homophone and baseline similarity.

This methodology can be extended to other languages and linguistic units, but we first turn to Chinese character homophones, which offer an interesting avenue of investigation into homophony due to the previous literature arguing for their unique role in language. The densely packed phonological space of Chinese characters, along with the ease of accessing standard sounds from Chinese characters, also provide a straightforward proof-of-concept for our method.

2 Method

2.1 Embedding spaces

Word embeddings transform linguistic units into numerical vectors within a continuous vector space. In Chinese natural language processing, these spaces have been trained successfully to evaluate semantic hierarchies (Fu et al., 2014) and word similarity (Pei et al., 2016), indicating that both

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 287–292

syntactic and semantic context can be captured successfully through these embeddings. In this paper, we look to use these embeddings to evaluate phonological comparisons, which is a novel application of this architecture. We use a pre-trained model that produces competitive results on the task of Chinese word segmentation by combining radical information within a dual LSTM network to capture deeper semantic meaning between words (He et al., 2018). This downstream task – placing characters closer together if they are more likely to form linguistic constituents – is useful for our project because it captures both deeper semantic meaning and knowledge of syntactic context effectively. We access these embeddings through the `RADICAL_CHAR_EMBEDDING_100` version of the word2vec model from `hanlp`, a multilingual NLP package (He and Choi, 2021). There are 9074 unique Chinese characters in this space, which produce a high number of phonological and baseline comparisons for each of our experiments.

2.2 Evaluating homophone relationships using cosine similarity

To extract groupings of homophones from the characters present in our embeddings, we used the `pypinyin` package[1] which converts characters to pinyin (a romanized representation that allows for the categorization of characters by their oral sound along with tone). Based on this information, we define *true homophones* as different characters that exhibit the same sound and tone. We also investigate the general effect of phonological similarity on the semantic relationships between words on the following levels: *pseudo-homophones*, defined as words which share the same sound but may exhibit different tones, *characters that share an initial sound*, and *characters that share vowels*. These were selected to account for fundamental structural elements of all Chinese characters, as shown below in Table 1.

Phonological Relationship	Examples
Homophone	鱼 (yú), 愉 (yú), 渔 (yú)
Pseudo-Homophone	腿 (tuǐ),推 (tuī) ,退 (tuì)
Initial sound	会 (huǐ), 哈 (hā), 很 (hěn)
Vowels	乖 (guāi), 段 (duǎn), 挂 (guǎ)

Table 1: Example groups exhibiting level of phonological similarity investigated.

[1] https://pypi.org/project/pypinyin/

We calculate homophone similarity as follows (the process is analogous for all other levels of similarity): let H be the set of all unique homophone pairs in our list of characters, where we use $k_i \in H$ to denote the pair of homophone character embeddings $\{h_{i1}, h_{i2}\}$. We evaluate the cosine similarity by calculating the cosine between the character embeddings for homophone h_{i1} and its homophone mate h_{i2}:

$$sim(k_i) = cos(h_{i1}, h_{i2}) \quad (1)$$

For each homophone comparison produced, we also generate two baseline comparisons. The baseline we chose was cosine similarity between a homophone and the characters of most similar frequency to its homophone mates. For each homophone pair $\{h_{i1}, h_{i2}\}$ in each homophone group, we find character b_{i1} that has most similar frequency to character h_{i1} so that we can compare b_{i1} to h_{i2}, and similarly we find character b_{i2} that has similar frequency to character h_{i2}. Let B be the set of all appropriate baseline comparisons that we can make to our original homophone characters. We then evaluate the cosine similarity between the homophones and their corresponding similar-frequency comparison, where we denote each possible pair as $l_i \in B$, as follows:

$$sim(l_{i1}) = cos(h_{i1}, b_{i2}) \quad (2)$$

$$sim(l_{i2}) = cos(b_{i1}, h_{i2}) \quad (3)$$

Using words of similar frequency as our baseline normalizes our results, since within a high-dimensional embedding space, higher frequency tokens generally exhibit smaller distances to all other tokens on average, and lower frequency tokens generally exhibit higher distances to all other tokens on average. Given this, a statistically significant difference in overall baseline and homophone comparisons would indicate a relationship between homophony and embedded similarity, since similar-frequency words – all else equal – should be most likely to exhibit the same average distances from homophone mates if there is no real underlying effect of homophony on context-sharing.

We extract the frequency of characters using `wordfreq`, a library containing word frequencies in various languages (Speer, 2022). We assume this to be a good proxy for the original frequency since the original corpus was trained on a scraped version

of Chinese Wikipedia from 2017, and `wordfreq` obtains its corpora for evaluating word frequency from Wikipedia alongside a variety of other sources such as newspapers, books, and websites. *Similar frequency character* in our experiments is thus extracted as the character in our embedding space that precedes or follows the target character in the sequence of all characters arranged in increasing order of frequency.

The baseline cosine similarities computed are then compared to the cosine similarities between homophones. We calculate the difference between the average baseline and homophone similarities as follows:

$$\textbf{Diff} = \frac{1}{|B|}\sum_{l_i \in B} sim(l_i) - \frac{1}{|H|}\sum_{k_i \in H} sim(k_i) \quad (4)$$

Finally, we record each individual similarity result alongside a binary value for whether the similarity is a baseline comparison or not. These results were employed in a probit model to test if similarity is a statistically significant predictor of a comparison being a baseline or homophone comparison.

2.3 Testing statistical significance

To test whether there is a statistically significant relationship between the similarities of characters and their status as homophone pairs, we fit a probit model to the relationship between our computed similarities and the homophone/baseline status of all of our comparisons. The model uses the similarity as the independent variable and fits a cumulative distribution function of the standard normal distribution to predict the effect of similarity on the probability that a word is either a baseline comparison or a homophone comparison.

3 Results

Table 2 shows the average cosine similarities calculated from our experiments, with the target column displaying the results of comparing characters that exhibit the indicated relationship.

Based on our similarity comparisons, we find that there is a highly statistically significant effect of homophony and pseudo-homophony on increasing cosine similarity, and a slightly less significant effect of characters with the same initial sound (in the opposite direction), but no significant effect on characters which share the same vowels. Results are shown in Table 3.

Relationship	Target Sim.	Baseline Sim.	Diff
Homophones	0.233	0.220	0.0125
Pseudo-H	0.227	0.219	0.00792
Initial sound	0.215	0.218	-0.00330
Vowels	0.222	0.219	0.00300

Table 2: Average cosine similarities and Diff (difference between average cosine similarities)

Relationship	n	coefficient	z-value	P > |z|
Homophones	9070	0.525	-26.634	0.000
Pseudo-H	755,304	0.345	-28.745	0.000
Initial sound	30,173	-0.154	2.474	0.013
Vowels	31,016	0.118	-1.632	0.103

Table 3: Relationships and levels of statistical significance obtained from probit model

In Table **??**, *n* displays the sample size for each group. This varies because different numbers of comparisons can be made for each relationship, but since we generate baselines proportionately to each target group, this should not impact the results. The *coefficient* can be interpreted as the amount which a one-unit addition in similarity impacts the likelihood of the comparison belonging to the target relationship (e.g. homophone) rather than the baseline comparison. The positive coefficient for homophones and pseudo-homophones, alongside the high level of significance ($p\text{-}value < 0.05$ in all cases except same-vowel comparisons), indicate that increased proximity in the embedding space (which we use as a proxy for increased semantic similarity) increase the probability of a comparison sharing similar phonological features according to our model. Other results are analyzed further in the Discussion section.

3.1 Visualizing embeddings in 2D space

We use t-SNE[2] and UMAP[3] packages to visually assess the local and global structures of homophone embeddings against the baseline. We have selected a specific pair of UMAP and t-SNE plots featuring the character 为 (wèi) to effectively illustrate our intended purpose, where baseline characters are chosen to have similar frequency to our "original homophone" (为).

Both UMAP and t-SNE are dimensionality re-

[2] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

[3] https://umap-learn.readthedocs.io/en/latest/basic_usage.html

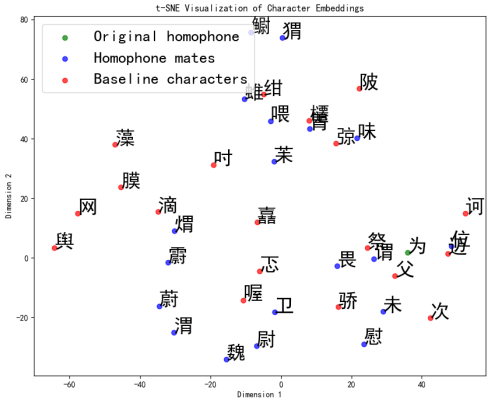

Figure 1: 为 (wèi) t-SNE plot

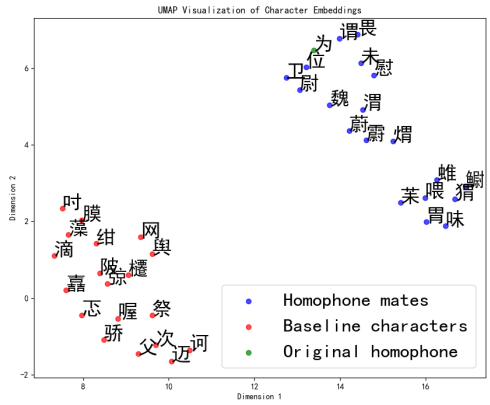

Figure 2: 为 (wèi) UMAP plot

duction techniques designed for handling non-linear data. While t-SNE focuses on finding a lower-dimensional approximation distribution by minimizing divergence between two higher-dimension-agnostic probability distributions[4], UMAP attempts to represent the underlying manifold structure of the data.[5] In simpler terms, t-SNE preserves local structure by retaining relationships between nearby data points in the high-dimensional space, while UMAP preserves global structures by considering the overall patterns of the data.

Our t-SNE plot (Figure 1) shows no distinct pairwise relationship between either homophones or the baseline group. However, our UMAP (Figure 2) analysis reveals that the chosen homophone group forms a distinctly different cluster from the baseline group. This indicates that on a local level, homophone character embeddings may not exhibit high levels of similarity, corroborating our low Diff score, but at a more global level they may exhibit distinct semantic meanings compared to baseline characters.

4 Discussion

Based on our results, we find evidence that phonological similarity and semantic similarity are correlated in Chinese. However, since words with the same initial sound exhibit semantic dissimilarity (albeit with a coefficient of smaller magnitude than for coefficient for homophones or pseudo-homophones, and with slightly less statistical significance), there may be a particular semantic role that is played by characters that share all sounds that cannot just be explained by the general level of phonological similarity. This is especially supported by the lack of statistically significant relationship for characters that share the same vowel sound.

If we interpret the embedding similarity purely as a measure of the syntactic environments which these characters are likely to occur in, the fact that homophones are more likely to share syntactic environments challenges the theory that language users rely explicitly on different syntactic context to avoid linguistic ambiguity. Further, if we use the distributional hypothesis to assume that this embedding similarity is representative of the semantic similarity between homophones, our findings dispute the idea that homophones must be semantically distant to survive or be effectively used in a language. This also corroborates the work of linguistic studies that have shown that encoding similar semantic information into words with phonological similarity may be more efficient for learning Chinese.

Our t-SNE and UMAP plots confirm these results, and also indicate that homophones exhibit different levels of similarity at the local and global level of the embedding space. Future probing could potentially determine what this discrepancy implies for the semantic relationship between homophones.

4.1 Conclusion

Our results show that previously existing architectures can be applied to produce fruitful empirical insights into Chinese homophony. Namely, our results indicate that a possible positive relation-

[4] https://tivadardanka.com/blog/how-tsne-works

[5] https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

ship exists between the phonological similarity of character-level homophones and their semantic similarity. Future robustness tests in other languages could further contribute to our understanding of homophony's role in language at large.

4.2 Limitations and Future Work

Since we did not have access to the original training corpus of our embeddings, we estimate character frequency using an external package. Possible discrepancies may exist between our recorded frequency and the true frequency of the character in the original training corpus.

Our study also exclusively relied on a single set of pretrained character embeddings for conducting the experiments. Consequently, the results may vary slightly when employing alternative models, given their capacity to generate distinct embeddings compared to our chosen model.

For ease of comparison, we evaluated a single embedding space that was shown to effectively capture both syntactic and semantic information for a downstream task (word segmentation). Future work could evaluate the robustness of these results across different embedding spaces, especially using embeddings that were optimized for different tasks. Another potential option for extension would be to perform experiments by training new models to produce vector embeddings. This would allow for variation in training corpora, thus possibly investigating if homophony displays different semantic behavior in different contexts. Investigating homophony at the word-level rather than the character-level in Chinese could also provide new insights into the relationship between phonological and semantic similarity within the Chinese language.

Future work extending this form of analysis to other languages could produce interesting and novel linguistic results, as well as improve the robustness of this technique. Agglutinative languages, where sounds can be densely packed together to construct new meanings, may be a particularly interesting avenue for investigation since embedding spaces could be produced at the morpheme and word level.

4.3 Acknowledgements

The authors of this paper thank Jackie CK Cheung for his feedback in the beginning stages of this paper, which first began as a final project in his NLP class at McGill University. We also thank the time of Brendan Gillon, who provided insightful ideas to this work at later stages. Finally, we thank the constructive feedback given to this paper by the anonymous reviewers.

References

Hsin-Chin Chen, Jyotsna Vaid, and Jei-Tun Wu. 2009. Homophone density and phonological frequency in chinese word recognition. *Language and Cognitive Processes*, 24(7-8):967–982.

Isabelle Dautriche, Laia Fibla, Anne-Caroline Fievet, and Anne Christophe. 2018. Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology*, 104:83–105.

Ludovic Ferrand and Jonathan Grainger. 2003. Homophone interference effects in visual word recognition. *The Quarterly Journal of Experimental Psychology Section A*, 56(3):403–419.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Han He, Lei Wu, Xiaokun Yang, Hua Yan, Zhimin Gao, Yi Feng, and George Townsend. 2018. Dual long short-term memory networks for sub-character representation learning. In *Information Technology-New Generations: 15th International Conference on Information Technology*, pages 421–426. Springer.

Jiang Liu and Seth Wiener. 2020. Homophones facilitate lexical development in a second language. *System*, 91:102249.

Jiahuan Pei, Cong Zhang, Degen Huang, and Jianjun Ma. 2016. Combining word embedding and semantic lexicon for chinese word similarity computation. In *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24*, pages 766–777. Springer.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Herbert Rubenstein, Spafford S Lewis, and Mollie A Rubenstein. 1971. Evidence for phonemic recoding in visual word recognition. *Journal of verbal learning and verbal behavior*, 10(6):645–657.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Thomas Wasow, Amy Perfors, and David Beaver. 2005. The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, pages 265–282.