

EasyChinese: Simplifying the Chinese Writing System Using Clustering Algorithms and Ambiguity Metrics

Kai Xu, Department of Computer Science, Yale University
 Advisor: Robert Frank, Department of Linguistics, Yale University

1. THE PROBLEM

Chinese is a difficult language to learn to read and write, requiring the memorization of thousands of characters. Our goal is to invent a new way of writing Chinese that is both **easy to read** and **semantically unambiguous**.

We can achieve this by mapping every Chinese character to a combination of **pronunciation** and **semantic category**:

— (one) → yī (pronunciation) + number (category) → #yī
 五 (five) → wǔ (pronunciation) + number (category) → #wǔ
 衣 (shirt) → yī (pronunciation) + clothing (category) → Nyī
 裙 (dress) → qún (pronunciation) + clothing (category) → Nqún

2. HOW TO WRITE PINYIN BETTER

Option 1: Pinyin without spaces: ambiguous 25% of the time.

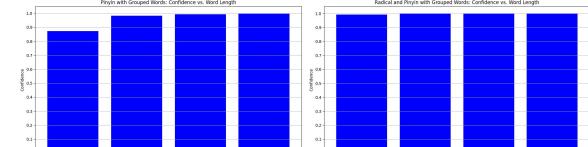
- wōhénxiūhānbīngqínlín

Option 2: Pinyin with spaces: ambiguous 5% of the time.

- wō hěn xiūhān bīngqínlín

Option 3: **Spaced pinyin with radicals**: ambiguous <1% of the time.

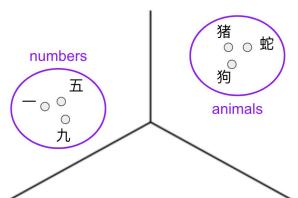
- 戈 wō 𠂇 次 xi 次 huān 𠂇 bīng 水 qí 水 lín



Problem: radicals do not provide the best semantic categories.

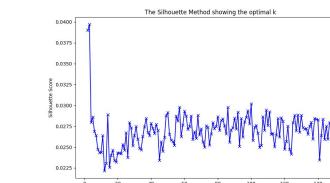
For example, the numbers 1 to 10 are split among seven radicals.

Solution: apply clustering algorithms to Chinese character embeddings to create better semantic categories.



3. CLUSTERING CONSIDERATIONS

- Reasonable cluster size.** Some clustering algorithms like DBSCAN have no check on cluster size. Others like Gaussian mixture models tend to generate fairly even clusters.
- High intra-cluster similarity and low similarity across different clusters.** We use the **silhouette score** to measure a character's similarity to characters in its own cluster versus those in other clusters.
- Reasonable number of clusters.** We set cluster count to 214 (the number of radicals) for meaningful performance comparisons. However, cluster count can be optimized by plotting the silhouette score of different cluster counts.



Clustering Algorithm	Mean Char. Silhouette	Median Cluster Size	Cluster Size Std. Dev.
Radicals	-0.18	10	65.3
DBSCAN	0.332	2	10.0
Ward hierarchical	0.004	40	29.3
K-means	0.036	40.5	28.0
Gaussian mixture	0.030	42	24.9

4. IMPROVING CLUSTERING RESULTS

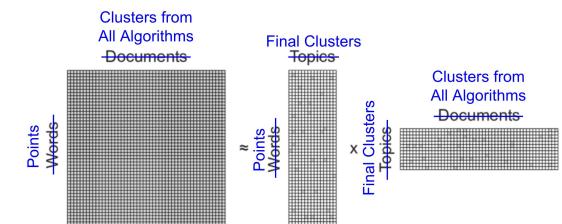
- Removing outliers** (characters that are semantically distant from other characters) before clustering. We developed **geometric ranked outlier score** (measure of outlier status).
- Incorporating human feedback:** we developed an experimental way to revise clustering results based on human feedback. It currently takes very long to converge.
- Consensus clustering:** aggregating the results of multiple clustering algorithms to create better clusters. We discovered a novel method for **conducting consensus clustering with topic modeling algorithms**.

Geometric Ranked Outlier Score						
$r = 0$	$r = 0.5$	$r = 1$	$r = 2$	$r = \infty$		
H-scores Lo-scores	H-scores Lo-scores	H-scores Lo-scores	H-scores Lo-scores	H-scores Lo-scores		
0.544 0.821	0.549 0.858	0.588 0.887	0.584 0.897	0.583 0.899		
0.539 0.823	0.549 0.862	0.571 0.893	0.583 0.897	0.583 0.899	0.811	
0.533 0.824	0.549 0.863	0.571 0.894	0.583 0.898	0.583 0.900	0.814	0.815
0.523 0.824	0.549 0.865	0.571 0.895	0.583 0.898	0.583 0.901	0.814	0.816
0.511 0.849	0.553 0.868	0.571 0.896	0.583 0.899	0.583 0.902	0.814	0.817
0.506 0.849	0.553 0.871	0.571 0.897	0.583 0.900	0.583 0.903	0.814	0.818
0.506 0.849	0.553 0.874	0.571 0.898	0.583 0.901	0.583 0.904	0.814	0.819
0.506 0.849	0.553 0.877	0.571 0.899	0.583 0.902	0.583 0.905	0.814	0.820
0.506 0.849	0.553 0.880	0.571 0.900	0.583 0.903	0.583 0.906	0.814	0.821
0.506 0.849	0.553 0.883	0.571 0.901	0.583 0.904	0.583 0.907	0.814	0.822
0.498 0.854	0.551 0.887	0.572 0.902	0.585 0.905	0.585 0.908	0.815	0.823
0.498 0.854	0.551 0.890	0.572 0.903	0.585 0.906	0.585 0.909	0.815	0.824
0.498 0.854	0.551 0.893	0.572 0.904	0.585 0.907	0.585 0.912	0.815	0.825
0.498 0.854	0.551 0.896	0.572 0.905	0.585 0.908	0.585 0.913	0.815	0.826
0.498 0.854	0.551 0.899	0.572 0.906	0.585 0.909	0.585 0.916	0.815	0.827
0.498 0.854	0.551 0.902	0.572 0.907	0.585 0.910	0.585 0.917	0.815	0.828
0.498 0.854	0.551 0.905	0.572 0.908	0.585 0.911	0.585 0.918	0.815	0.829
0.498 0.854	0.551 0.908	0.572 0.909	0.585 0.912	0.585 0.923	0.815	0.830
0.498 0.854	0.551 0.911	0.572 0.910	0.585 0.913	0.585 0.926	0.811	0.833

Geometric ranked outlier score takes one parameter, r , that controls the extent to which closer points, as opposed to farther points, influence a point's outlier status.

Original	Iteration 1
Cluster 1: 猫 狗	Cluster 1: 猫 狗 狼 狐 鹰
Cluster 2: 狮 虎 狼 狐 鹰	Cluster 2: 狮 虎
Cluster 3: 鸡 鸭	Cluster 3: 鸡 鸭
Iteration 2	Iteration 3
Cluster 1: 猫 鹰	Cluster 1: 猫 狮 虎
Cluster 2: 狮 虎	Cluster 2: 狗 狼 狐
Cluster 3: 狗 狼 狐 鸡 鸭	Cluster 3: 鸡 鸭 鹰

Our experimental method allows humans to specify characters that do not belong in a cluster, and re-clusters accordingly. However, convergence takes a long time.



Consensus clustering can be converted into a topic modeling problem. Instead of assigning words in documents to topics, we assign points in clusters from all clustering algorithms to final clusters. Image adapted from Andrea L. Bertozzi.

5. FUTURE RESEARCH

Our discovery of a novel method for conducting consensus clustering using topic modeling algorithms could be useful for broader clustering tasks. Further investigation is pending.