Inducing Word Clusters from Classical Chinese Poems

John Lee[1], Mengqi Luo[2]
1. Department of Linguistics and Translation, City University of Hong Kong, Hong Kong SAR, China
2. School of Information Management, Wuhan University, China
jsylee@cityu.edu.hk, lakeygtgz@163.com

Abstract
Parallelism is a literary device that is frequently used in Classical Chinese poetry. Within the two lines of a parallel couplet, the words in one line are expected to mirror those in the other in terms of syntax and meaning. Judicious selection of pairs of related words is thus important for poem composition. This article investigates statistical approaches for word clustering, such that all words in each cluster can serve as candidates to form appropriate word pairs. We compare three corpus-based methods for computing word similarity relatedness, and apply a graph-based clustering algorithm to induce word clusters. We evaluate the quality of the automatically induced clusters with respect to a gold standard proposed by a literary scholar. Experimental results show that similarity scores estimated by the word2vec model lead to more accurate clusters than pointwise mutual information and chi-square, reaching 61.2% precision, 70.8% recall and 61.2% purity. Our work lays a foundation to support further studies on parallelism in Classical Chinese literature, and to provide training data for computer-assisted poem composition.

Keywords
Classical Chinese; parallelism; Chinese poetry; word clustering.

1. Introduction

In many literature traditions, “couplets” are a common feature in poems. A couplet consists of two consecutive lines in a poem, and the two lines are “parallel” in the sense that the words in the first line mirror those in the second, both syntactically and

semantically. Typically, they have comparable grammatical constructions, and bear similar or opposite meaning. A well-known example is ancient Hebrew poetry (Kugel, 1981). Consider the following couplet, taken from Proverbs 27:8 in the King James Version of the Bible:

> As a bird that wandereth from her nest,
> so is a man that wandereth from his place.

As an example of synonymous parallelism, the word 'bird' in this couplet corresponds to 'man', and 'her nest' to 'his place'. Other kinds of parallelism, such as antithetical and synthetical, are also extensively employed. Indeed, parallelism, or "coupling", is said to be a universal principle behind poetic structure (Levin, 1962), and has been studied in many languages, from Ugaritic (Segert, 1983), Russian (Jakobson, 1966), to Chinese (Wang, 2003), which is the focus of this study.

A typical Classical Chinese poem consists of two or four couplets; all lines have the same number of characters, usually five or seven. As an example, consider *Climbing Crane Tower*, a poem by Wang Zhihuan taken from the *Complete Tang Poems* (Peng, 1960), the best known anthology of Classical Chinese poems from the Tang Dynasty, a golden age of poetry. Table 1 shows its first couplet, which contains two lines of five characters each. One can analyze the parallelism therein by examining the "character pairs", i.e., the two characters that occupy the same position in the two lines. The first character on Line 1, *bai* 'white', and the first character on Line 2, *huang* 'yellow', form the first character pair; both characters in this pair refer to colors. The second character on Line 1, *ri* 'sun', and the second character on Line 2, *he* 'river', form the second pair; both are objects in the natural world. So are the characters in the fourth character pair, *shan* 'hill' and *hai* 'sea'. The third character pair, *yi* 'rest' and *ru* 'enter', consists of verbs; and so does the fifth pair, *jin* 'end' and *liu* 'flow'. Hence, in all five pairs, the characters have the same part-of-speech and have related meaning.

Line 1	Chinese	白	日	依	山	尽
	Pinyin	*bai*	*ri*	*yi*	*shan*	*jin*
	Gloss	'white'	'sun'	'rest'	'hill'	'end'
	Translation	"White sun rests on mountains — and is gone."				
Line 2	Chinese	黄	河	入	海	流
	Pinyin	*huang*	*he*	*ru*	*hai*	*liu*

	Gloss	‘yellow’	‘river’	‘enter’	‘sea’	‘flow’
	Translation	“Yellow River enters sea — and flows on.”				

Table 1: An example parallel couplet from a four-line poem by Wang Zhihuan, with English translation from Cai (2008)

There is no strict requirement for every character pair to be parallel. Nonetheless, poets often favor parallelism since it is perceived as a hallmark of an elegant poem. Judicious selection of character pairs, therefore, is an important skill. For poets, it would be useful to have access to groups of characters, from which they can draw any two characters to form appropriate character pairs. Past efforts to compile such groups or clusters have produced classics such as *Weiwendi Shige* and *Shiyunhebi*, as well as a more recently proposed semantic taxonomy (Wang, 2003). These manual efforts, however, are limited in coverage. Wang (2003) provided 863 example characters to illustrate his semantic taxonomy (Table 2); these characters cover only 10.4% of the types in the *Complete Tang Poems*. To the best of our knowledge, there is not yet any reported attempt to exhaustively induce such clusters from large corpora. These clusters would not only provide assistance to poets, but also support computer-assisted poem composition tools as well as literary studies on parallelism.

We report our work towards this goal by exploring statistical, corpus-based methods to induce word clusters, such that all words in a cluster can serve as candidates to form appropriate word pairs. For computing semantic relatedness, we investigate statistical measures that directly leverage frequency counts of word pairs, as well as a neural language model. We then apply a graph-based clustering method on an undirected graph weighted by these similarity scores. Evaluation results show that the best model achieves 61.2% precision, 70.8% recall and 61.2% purity with respect to the semantic taxonomy proposed by Wang (2003).

The rest of the article is organized as follows. In the next section, we provide the necessary background on Classical Chinese poetry and relevant research on natural language processing on Chinese literature. In Section 3, we discuss previous work in word clustering. Section 4 gives details on our approach, including the computation of similarity scores and clustering. Section 5 presents our evaluation metrics. Section 6 describes our training data and experimental setup, and then reports experimental results. Section 7 concludes and sketches future work.

2. Chinese poetry and NLP

In this section, we discuss the phenomenon of parallelism in Chinese poetry (Section 2.1),

and then summarize previous applications of natural language processing (NLP) on analysis of Chinese literature (Section 2.2).

2.1. Parallelism

Parallelism, a literary device that is frequently used in Classical Chinese literature, can be generally defined as a pair of sentences that have the same number of characters and similar syntax (Chen, 1957). In poetry, this device manifests itself in parallel couplets. The two lines in a couplet are expected to have the same sentence structure; characters occupying the same positions on the two lines must have matching parts-of-speech (POS), and have related meaning (Feng, 1990).

Although there is no precise definition of "related meaning", a number of guidelines have been proposed. The book *Weiwendi Shige* offered an 8-category semantic taxonomy for nouns. In the 19th century, the book *Shiyunhebi* gave a more fine-grained, 37-category taxonomy. In modern scholarship, among the most well-known is the semantic taxonomy developed by Wang (2003). This taxonomy consists of 22 categories[1], exemplified by 863 characters, covering mostly nouns, but also some adjectives, adverbs, conjunctions, and particles (Table 2).

Semantic category	Size	Description and examples
Celestial	32	Heavenly bodies and other phenomena in the sky e.g., 日 *ri* 'sun', 月 *yue* 'moon', 風 *feng* 'wind'
Seasonal	30	Terms for periods of time e.g., 夜 *ye* 'night', 春 *chun* 'spring', 年 *nian* 'year'
Geographic	81	Geographic entities on earth e.g., 山 *shan* 'mountain', 海 *hai* 'sea', 江 *jiang* 'river'
Architectural	54	A type of building, or a component thereof e.g., 殿 *dian* 'palace', 樓 *lou* 'building'
Instruments	79	A broad topic including tools, utensils, vehicles, household objects, weapons, etc. e.g., 劍 *jian* 'sword', 琴 *qin* 'violin'

[1] We omitted two categories, 'Personal names' and 'Place names', because no example characters were provided for them.

Clothing	32	e.g., 帶 *dai* 'belt', 環 *huan* 'ring'
Food	32	e.g., 茶 *cha* 'tea', 糕 *gao* 'cake'
Products of civilization	27	Objects associated with artistic pursuits, including stationary tools and musical instruments e.g., 筆 *bi* 'pen', 紙 *zhi* 'paper'
Literary	38	Terms related to literature e.g., 書 *shu* 'book', 詩 *shi* 'poem', 信 *xin* 'letter'
Flora	61	Plants e.g., 草 *cao* 'grass', 花 *hua* 'flower', 柳 *liu* 'willow'
Fauna	72	Animals e.g., 鳥 *niao* 'bird', 魚 *yu* 'fish', 龍 *long* 'dragon'
Body parts	53	Parts of the human body and things produced by the body e.g., 眼 *yan* 'eye', 影 *ying* 'shadow', 聲 *sheng* 'voice'
Human emotions	44	Human activities and sentiments e.g., 歌 *ge* 'song', 舞 *wu* 'dance', 意 *yi* 'desire', 心 *xin* 'heart'
Human relations	41	kinship, titles and professions e.g., 兄 *xiong* 'brother', 相 *xiang* 'minister', 兵 *bing* 'soldier', 農 *nong* 'farmer'
Pronouns	18	e.g., 我 *wo* 'I', 爾 *er* 'you'
Locations	14	e.g., 北 *bei* 'north', 中 *zhong* 'middle', 外 *wai* 'outside'
Numbers	25	e.g., 三 *san* 'three', 雙 *shuang* 'pair'
Colors	27	e.g., 紅 *hong* 'red'
Coordinates	22	Words used in a numbering system related to the calendar

		e.g., 甲 *jia* ‘one’
Adverbs	56	e.g., 亦 *yi* ‘also’, 不 *bu* ‘not’
Conjunctions	13	e.g., 和 *he* ‘and’ 共 *gong* ‘together’
Particles	13	Usually placed at end of a phrase for emphasis, with no concrete meaning, e.g., 也 *ye*

Table 2: Description and examples of the semantic categories used in our evaluation (Wang, 2003), with English translations of the category names follow Harvey and Kao (1979)

2.2. Quantitative analyses on Classical Chinese literature

The computational linguistics community has shown increasing interest in performing quantitative analysis on ancient and medieval Chinese literature (e.g., Huang et al., 2002; Wong and Lee, 2016). Past research in poetry has dealt with two main topics. The first is centered on word usage. Huang (2004) constructed and compared ontologies based on the *Three Hundred Tang Poems* and poems by Su Shi. Fang et al. (2009) built a parser for imagistic language on the basis of an ontology of imagery for Classical Chinese poems (Lo, 2008). There have also been various studies on noun collocations in common themes, such as colors, seasons and sentiments (e.g., Lee and Wong, 2012; Hou and Frank, 2015; Liu et al., 2015).

The second category is concerned with sentence structure. Lee et al. (2018) present a quantitative analysis on the distribution of imagistic language and propositional language. They corroborate some of the qualitative observations by Kao and Mei (1971), e.g., that parallelism is more pronounced in the middle couplets compared to those at the beginning and the end. Cao (1998) created a database with 1,000 couplets to study the extent of parallelism in Classical Chinese poems. Yuan (2005:236-239) analyzed 39 poems by Du Fu, and Huang (2006:100-107) computed the percentage of parallel couplets in 28 poems by Du Mu. Various systems for computer-assisted poem composition have also been developed (Jiang and Zhou, 2008; Zhang and Lapata, 2014; Lee et al., 2016). Applying natural language generation techniques for both word usage and sentence structure analysis, these systems automatically or semi-automatically compose couplets that take parallelism into account.

3. Word clustering

Clustering is a common task in natural language processing (NLP). A clustering

algorithm takes as input a set of words $\{w_1, \ldots, w_N\}$, the target number of clusters C, and a similarity matrix $(s_{ij})_{ij=1,\ldots,N}$. Each element s_{ij} in the matrix is a non-negative number that indicates the degree of similarity between w_i and w_j. The algorithm aims to determine an assignment of the words into C clusters, such that the degree of similarity between words within clusters is optimized. Word clustering is a useful step for many NLP tasks, including language modeling (Brown et al., 1992), document classification (Matsuo et al., 2006), and relation extraction (Sun et al., 2011). We now review previous work in the two components of word clustering: first, an algorithm to group words together based on their degree of similarity (Section 3.1); and second, a method to estimate the degree of similarity between two words (Section 3.2).

3.1. Word clustering algorithms

There are two main kinds of word clustering algorithms. A divisive algorithm starts with one cluster containing all words, then recursively splits it. An agglomerative algorithm begins with singleton clusters, and then successively merges them. The optimal clusters to be split or merged can be identified with various measures, such as mutual information between clusters (Brown et al., 1992) or minimum description length (Li and Abe, 1996).

Graph-based clustering, a particular class of algorithm that casts clustering as a graph partition problem, has been found effective not only in word clustering (Matsuo et al., 2006) but also a variety of NLP tasks including coreference resolution (Chen and Ji, 2009) and word sense disambiguation (Agirre et al., 2007). A vertex in the graph represents a word, and an edge is weighted with the similarity score between the words corresponding to its two vertices. The algorithm aims to partition the graph into subgraphs, so that the edges within each subgraph have heavier weights than those that connect different subgraphs. A number of algorithms make use of the modularity score to measure the strength of connections between the vertices within subgraphs compared to those across different clusters. Newman clustering chooses the two subgraphs that would yield the greatest increase in modularity score (Newman, 2004). Matsuo et al. (2006) reported that it outperforms average-link agglomerative clustering on the chi-square similarity measure. Further, Ichioka and Fukumoto (2008) found that it outperforms k-means clustering on Japanese onomatopoetic word clustering. The Louvain method (Blondel et al., 2008) is another modularity-based clustering method that has been successfully used in various clustering tasks, for example in detecting user communities (Lee et al., 2012). It first looks for small communities by optimizing modularity locally, and then aggregates nodes belonging to the same community and builds a new network whose nodes are the communities.

3.2. Word Similarity scores

To construct the similarity matrix, past research has made use of both knowledge bases and unstructured corpora to estimate the degree of similarity between two words. With knowledge bases such as WordNet and HowNet, which organize words into a hierarchical semantic taxonomy, one can calculate the semantic distance between two words (Li and Li, 2007). However, HowNet and other similar knowledge databases for Chinese (Gan and Wong, 2000; Choi et al., 2004; Chen et al., 2002; Xu et al., 2008) all focused on Modern Chinese rather than Classical Chinese.

An alternative is to use corpus-based methods, which assume that words occurring in similar contexts have similar meaning (Harris, 1954). Mutual information and chi-square are both standard metrics for measuring the strength of association between two words, taking into account the probabilities of encountering both words in a text unit, one word but not the other, or neither. Matsuo et al. (2006) and Garcia and Mena (2008) exploited the web as corpus, computing the number of times two words co-occur on the same web page. Bollegala et al. (2007) proposed a supervised classification method that combines both page counts and sentence-level contexts in snippets.

More recently, Mikolov et al. (2013) developed word2vec, which learns the vector representations of words in high-dimensional vector space and calculates the cosine distance between two words. Word embeddings learned by word2vec have been shown to benefit a variety of tasks involving semantic relatedness, ranging from lexical substitution (Melamud et al., 2015), to similarity in biomedical terms (Muneeb et al., 2015), and the generation of distractors in multiple-choice questions (Jiang and Lee, 2017).

4. Approach

Following the best practices described in Section 3, we approach clustering as a two-step process: First, given a large text corpus, we estimate the degree of similarity between two words (Section 4.1). Then, we apply a clustering algorithm to induce word clusters (Section 4.2). Our gold standard (Section 5.2) annotates individual characters, reflecting the fact that most words in Classical Chinese, unlike those in modern Chinese, consist of only one character. For this reason, we treat characters, rather than words, as the unit of analysis. However, we will continue to use the term “word” in the rest of the article in order to adhere to standard, language-independent terminology.

4.1. Semantic relatedness

We compare two approaches for computing similarity scores. The first directly leverages

frequency statistics on parallelism (Section 4.1.1). The second, in contrast, uses a neural language model that considers all words within a context window (Section 4.1.2).

4.1.1. Parallelism

Since our goal is to induce clusters that reflect the parallelism in the *Complete Tang Poems*, word pairs that are frequently attested in couplets should be assigned a higher similarity score. Most past work defined co-occurrence as the appearance of two words within an *N*-word context window or within the same web page. In our context, we modify the definition of co-occurrence to be occurrence as word pairs, i.e. to appear in the same position on the two lines in a couplet. For example, the pair *bai* 'white' and *huang* 'yellow' in Table 1 would contribute one count of co-occurrence. We compute co-occurrence statistics from the *Complete Tang Poems*, and then derive similarity scores using pointwise mutual information (PMI) and chi-square.

Pointwise mutual information (PMI). The PMI between two words w_1 and w_2 is defined as:

$$PMI(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)\,p(w_2)} \quad (1)$$

Suppose N is the total number of tokens in the corpus, then $p(w)$ is estimated by the number of occurrences of w, divided by N; $p(w_1, w_2)$ is estimated by the number of occurrences of w_1 and w_2 as a word pair, divided by $N/2$, i.e., the total number of word pairs in the corpus. Table 3 shows the word pairs with the highest PMI.

Rank	Word pair	Rank	Word pair
1	南 *nan* 'south', 北 *bei* 'north'	6	雨 *yu* 'rain', 風 *feng* 'wind'
2	山 *san* 'hill', 水 *sui* 'water'	7	風 *feng* 'wind', 月 *yue* 'moon'
3	雲 *yun* 'cloud', 月 *yue* 'moon'	8	水 *sui* 'water', 雲 *yun* 'cloud'
4	玉 *yu* 'jade', 金 *jin* 'gold'	9	西 *xi* 'west', 東 *dong* 'east'
5	青 *qing* 'green', 白 *bai* 'white'	10	東 *dong* 'east', 北 *bei* 'north'

Table 3. Top ten word pairs in terms of pointwise mutual information in the *Complete Tang Poems*

Chi-square. We also calculated chi-square values between two words as their similarity

score. After computing the contingency table (Table 4), we determined the chi-square value for two words w_1 and w_2 by:

$$\chi^2(w_1, w_2) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)} \quad (2)$$

Word pair	Contains w_2	Does not contain w_2
Contains w_1	*a*	*b*
Does not contain w_1	*c*	*d*

Table 4: Contingency table for calculating chi-square for two words w_1 and w_2

Although parallelism is a frequent phenomenon, not every couplet is expected to be parallel. For example, for poems of the "regulated verse" type that have eight lines, only the second and third couplets are expected to be parallel; parallelism is optional elsewhere (Cai, 2008). Other types of poems, unfortunately, do not have fixed patterns. In principle, one can reduce the noise in the training set by filtering out the first and fourth couplets of the regulated-verse poems; however, there is no reliable method to identify regulated-verse poems. Instead, we exploited the statistical strength of all couplets in the corpus to try to overcome the noise from non-parallel couplets.

4.1.2. Word Vector

We used the word2vec toolkit to calculate word vectors with the Continuous Bag-of-Words (CBOW) model. We then computed word cosine distance between word pairs to serve as similarity scores. We tried three different sizes for the context window, namely 3, 5, and 7. Recall that most lines in Classical Chinese poems have either five or seven characters (Section 1). A 5-word window allows the model to consider word pairs in couplets with five-character lines; e.g., in Table 1, the word pair *bai* and *huang* are in the context window of one another. A 7-word window covers word pairs in couplets with 7-character lines as well, at the expense of introducing more noise. In contrast, a 3-word window, too short to cover character pairs, considers adjacent words only.

4.2. Word Clustering

Given a similarity matrix computed with PMI, chi-square, or word2vec (Section 4.1), we constructed an undirected graph whose vertices represent words, and whose edges are weighted with the corresponding value in the similarity matrix between two words. We

then performed clustering on the graph with the Louvain community detection algorithm (Blondel et al., 2008). We used the default parameters in the implementation provided in Gephi (http://gephi.org/), an open-source software for visualizing and analyzing large network graphs.

5. Evaluation metrics

To evaluate the quality of the automatically induced clusters, we utilized the semantic taxonomy proposed by Wang (2003). As shown in Table 2, it consists of 22 categories with a variety of sizes, ranging from 13 members for 'Conjunctions' and 'Particles', to 81 members for 'Geographic'. We compared the automatically induced clusters to the gold standard using three metrics: purity, precision, and recall.

Purity expresses the proportion of members in a cluster that should belong to the same category. To measure the purity of an automatically inferred cluster, we associated each cluster in the system output with the gold standard category in Wang's (2003) taxonomy that contains the largest proportion of its members. Purity is then the number of correctly assigned words divided by the total number of words in the cluster.

Our calculation of precision and recall is based on the Rand Index, which measures the similarity between two partitions of the same dataset. Specifically, the clustering process is viewed as a set of decisions, one for each of $N(N\text{-}1)/2$ pairs of types in the corpus, where N is the total number of words. Two words are "similar" if they belong to the same category in the gold standard, and "dissimilar" if they belong to different categories.

- A true positive (TP) decision assigns two similar words to the same cluster.
- A false negative (FN) decision assigns two similar words to different clusters.
- A false positive (FP) decision assigns two dissimilar words to the same cluster.

Precision is defined as TP/TP + FP; recall is defined as TP/TP + FN.

6. Experiment

We implemented the approach described above, and evaluated it by inducing word clusters from the *Complete Tang Poems*. We provide details on the experimental setup (Section 6.1), and then present the results (Section 6.2).

6.1. Experimental setup

We used as training data the set of poems in the *Complete Tang Poems*, which contains 2.4M tokens and 8K types. We represented the 863 characters in the taxonomy of Wang (2013), and their similarity scores computed from the three approaches outlined above

(Section 4.1), as a co-occurrence graph. For word2vec, the size of the word context window was set at 3, 5 and 7. Our gold standard has 22 clusters; since *a priori* knowledge about the optimal number of clusters cannot be assumed, however, we evaluated the performance of our approach when the output included from up to 25 clusters. As the Louvain algorithm is non-deterministic, we repeatedly executed the algorithm in Gephi 10 times, and selected the output with the minimum number of "small" clusters. A cluster was considered "small" if it contained fewer than 5 words.

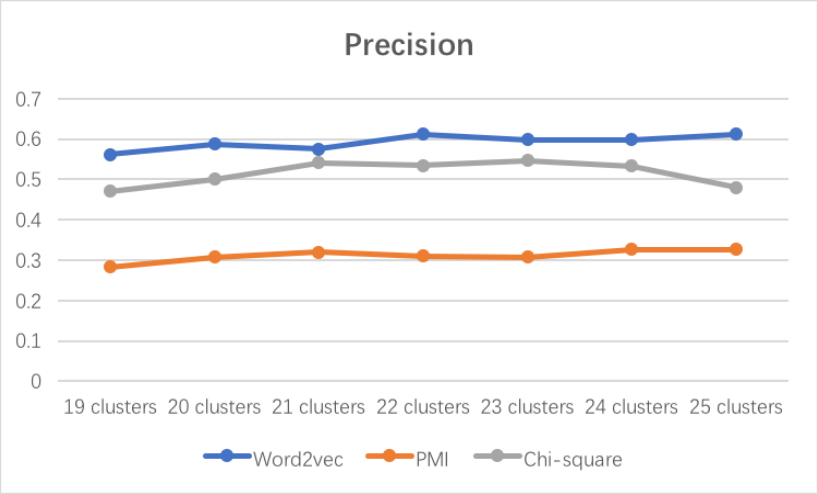

Figure 1. Precision of the automatically induced clusters, with the target number of clusters ranging from 19 to 25.

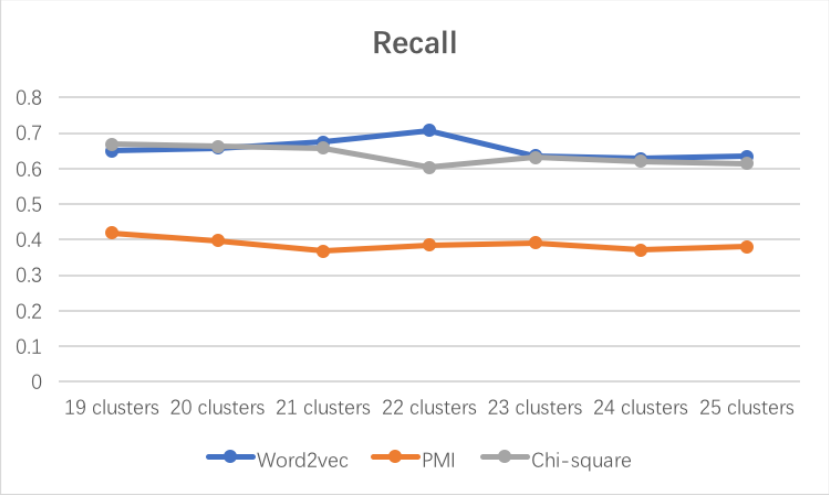

Figure 2. Recall of the automatically induced clusters, with the target number of clusters ranging from 19 to 25.

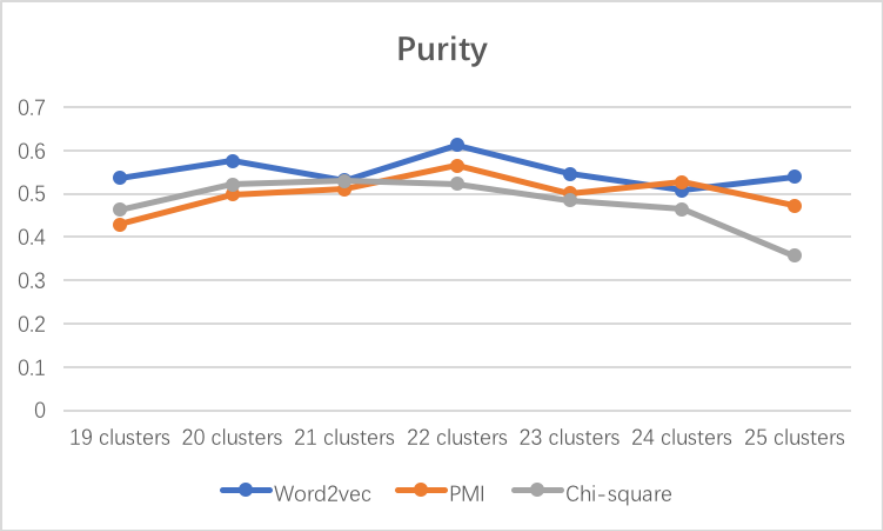

Figure 3. Purity of the automatically induced clusters, with the target number of clusters ranging from 19 to 25.

Similarity score	Window size	Precision	Recall	Purity
PMI	n/a	0.310	0.385	0.564
Chi-square	n/a	0.534	0.604	0.522
Word2vec	3	0.595	0.675	0.588
Word2vec	**5**	**0.612**	**0.708**	**0.612**
Word2vec	7	0.568	0.629	0.574

Table 5. Precision, recall and purity of clusters induced with similarity scores based on PMI, chi-square and word2vec, with the number of clusters set to the gold standard (22)

6.2. Experimental Results

Table 5 shows the quality of the automatically induced clusters in terms of precision, recall and purity when the target number of cluster was set to the gold standard.

Word2vec. In terms of the context window size, the 5-word window produced optimal results (Table 5). As explained above, this window size allowed the model to directly capture the word pairs in parallel couplets in five-character lines. Lengthening this window to 7 enabled direct capture of word pairs in parallel couplets in seven-character lines, but the greater amount of noise led to worse performance, below that of the 3-word

window. With respect to all three metrics, word2vec outperformed both PMI and chi-square.

Figures 1 to 3 illustrate changes in the quality as the target number varies from 19 to 25. The Louvain algorithm appears relatively robust in terms of the target number of clusters. As the target number varied from 19 to 25, the precision, recall and purity figures fluctuated within 1% only.

With the context window size set to 5 and the number of clusters set to 22, the 'Fauna', 'Body parts' and 'Flora' clusters induced by word2vec achieved the highest purity (at 90.1 %, 82.9% and 81.5%, respectively). The clusters with the lowest purity included 'Instrument', 'Food' and 'Human relations' (at 39.2%, 32.0% and 3l.3%, respectively). The 'Instrument' cluster had a substantial number of words from 'Clothing', probably due to similar syntactic usage. The 'Food' cluster tended to mix with 'Products of civilization', reflecting perhaps the close connection between banquets and cultural pursuits. The 'Human relations' cluster had a large proportion of words from 'Literary'.

Clusters for function words are among the smallest in the gold standard, and they turned out to be difficult for all methods. The members of 'Particles' were scattered throughout, and most of 'Conjunctions' were merged with 'Adverbs'. The gold-standard 'Pronoun' cluster is also small, and did not gain the majority in any of the induced clusters; it was instead divided almost evenly between 'Adverbs' and 'Human relations'. Some word pairs with obvious semantic affinity were clustered together due to high scores from word2vec; but, they belong to different clusters in the gold standard. These include, for example, the pair *si* 寺 'temple' ('Architectural') and *seng* 僧 'monk' ('Human relations'); and the pair *dian* 店 'shop' ('Architectural') and *cun* 村 'village' ('Geographic').

PMI and chi-square. Both PMI and chi-square suffered from the tendency to yield overly large or small clusters, leading to less accurate results than word2vec. Chi-square achieved better precision and recall than PMI in most settings, but was slightly outperformed by PMI on recall when the target number of clusters is known. Similar to word2vec, the 'Body parts' and 'Fauna' clusters induced by chi-square were among the purest clusters, while 'Flora' fared worse.

A number of character pairs received high similarity scores, even though they do not belong to the same cluster in the gold standard. The categories 'flora' and 'celestial', for example, were especially confusable, with characters such as 樹 shu 'tree' and 雲 yun 'cloud' frequently paired in the training data. Other clusters included characters that were related in meronymy, such as 樹 shu 'tree' and 山 shan 'mountain', or 鳥 niao 'bird' and 雲 yun 'cloud'. In both of these cases, a natural object ('hill' and 'cloud') is paired with a typical object in its realm ('tree' and 'bird'). Other pairs, seemingly unrelated in

literal meaning, are nonetheless evocative of certain events or situations. A prime example is 酒 jiu 'wine' and 詩 shi 'poem'. Wine, a catalyst for inspirations to write and interpret poems, served as a central theme for many prominent poets. Despite this tradition, 'wine' and 'poem' are classed in different semantic categories — 'Food' and 'Literary', respectively — in the gold standard. Likewise, at banquets in Ancient China, poets would drink while playing the lute and composing poems or articles. That explains why jiu and 琴 qin 'lute', as well as jiu and 書 shu 'book', which do not belong to the same cluster in the gold standard, nonetheless received high similarity scores.

7. Conclusion and Future Work

We have applied word clustering algorithms on the *Complete Tang Poems* to induce clusters of semantically related words. Specifically, we used the Louvain community detection algorithm (Blondel et al., 2008), a graph-based clustering method, to form clusters so that words within the same cluster can serve as candidate word pairs in a parallel couplet. We compared three methods for computing similarity scores to serve as edge weights. In an evaluation against a gold standard proposed by a literary scholar, we found that word2vec, a word vector representation, outperformed pointwise mutual information (PMI) and chi-square. While word2vec does not directly exploit character pairs in couplets, it is effective in identifying word usage patterns in the context window, yielding the best result of 61.2% precision, 70.8% recall and 61.2% purity for the induced clusters. The noise from non-parallel couplets might have adversely affected the results for PMI and chi-square.

In future work, we plan to further improve the quality of the clusters by incorporating grammatical features in estimating semantic relatedness, and by exploring other clustering methods. In our evaluation, we intend to take into consideration other gold-standard word clusters for Classical Chinese poems, to reflect a broader range of understanding of parallelism in the field of Classical Chinese literary studies. It is hoped that the induced word clusters can support literary studies in parallelism, provide assistance to poets, and also improve the performance of computer-assisted poem composition systems.

8. References

Agirre, E., Martinez, D., de Lacalle, O. L., and Soroa, A., 2007, Two graph-based algorithms for state-of-the-art wsd, in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bollegala, D., Matsuo, Y., and Ishizuka, M. 2007, An integrated approach to measuring semantic similarity between words using information available on the web, in *Proceedings of the Annual Conference of the North American Chapter of the Association*

for Computational Linguistics, pp. 340–7.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E., 2008, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, vol. 10.

Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C., and Mercer, R. L., 1992, Class-based n-gram models of natural language, *Computational Linguistics*, 18(4), p.467-479.

Cai, Z.-Q., 2008, *How to read Chinese poetry*, New York: Columbia University Press.

Cao, F. 曹逢甫, 1998, A linguistic study of the parallel couplets in Tang poetry, *Technical Report*, Linguistics Graduate Institute, National Tsing Hua University.

Chen, W. 陳望道 (1957). 修辭學發凡 [An investigation to Rhetoric], Taipei: Kaiming shuju.

Chen, H.-H., Lin, C.-C., and Lin, W.-C. Lin, 2002, Building a Chinese-English wordnet for trans lingual applications, *ACM Transactions on Asian Language Information Processing*, vol. 1, no. 2, pp. 103—122.

Chen, Z., and Ji, H., 2009, Graph-based event coreference resolution, in *Proc. Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*.

Choi, K.-S., Bae, H.-S., Kang, W., Lee, J., Kim, E., Song, H., and H. Shin, H., 2004, Korean-Chinese-Japanese multi-lingual wordnet with shared semantic hierarchy, in *Proc. Conference on Language Resources and Evaluation (LREC)*.

Fang, A. C., Lo, F., Chinn, C. K., 2009, Adapting NLP and corpus analysis techniques to structured imagery analysis in classical Chinese poetry, in *Proc. Workshop Adaptation of Language Resources and Technology to New Domains*.

Feng, X. 馮興煒, 1990, Duiou zhishi 對偶知識 [Knowledge of parallelism], Beijing: Luyou jiaoyu chubanshe.

Gan, K. W., and Wong, P. W., 2000, Annotating information structures in chinese texts using hownet, in *Proc. 2nd Workshop on Chinese Language Processing*.

Gracia, J., and Mena, E., 2008, Web-based measure of semantic relatedness, in *Proceedings of the 9th International Conference on Web Information Systems Engineering (WISE'08)*, pp. 136–150. Berlin, Germany: Springer-Verlag.

Harris, Z., 1954, Distributional structure, *Word*, vol 10, 146-162.

Harvey, B. L. and Kao, K. S. Y., 1979, Text generation modelling of Chinese regulated verse, *Poetics*, vol 8, pp.459-479.

Hou, Y. and Frank, A., 2015, Analyzing sentiment in Classical Chinese poetry, in *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 15–24.

Huang, C.-R., 2004, Text-based construction and comparison of domain ontology: A study based on classical poetry, in Proc. 18th Pacific Asia Conference on Language, Information and Computation (PACLIC).

Huang, L. 黃麗敏, 2006, The study of Classical Poems of Tu-mu. 杜牧古體詩研究 Master's Thesis, National Sun Yat-sen University.

Huang, L., Peng, Y., Wang, H., and Wu, Z., 2002, PCFG parsing for restricted Classical Chinese texts, in *Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing*.

Ichioka, K. and Fukumoto, F., 2008, Graph-based clustering for semantic classification of onomatopoetic words, in *Proc. Workshop on Graph-based Algorithms for Natural Language Processing (TextGraphs-3)*.

Jakobson, R., 1966, Grammatical parallelism and its Russian facet, *Language*, vol. 42.

Jiang, S. and Lee, J., 2017, Distractor generation for Chinese fill-in-the-blank items, in *Proc. 12th Workshop on Innovative Use of NLP for Building Educational Applications*.

Jiang, L. and Zhou, M., 2008, Generating Chinese couplets using a statistical MT approach, in *Proceedings of the 22nd Int'l Conf. on Computational Linguistics*, 377–384.

Kao, Y. and Mei T., 1971, Syntax, diction, and imagery in T'ang Poetry, *Harvard Journal of Asiatic Studies*, 31, p.49–136.

Kugel, J. L., 1981, *The idea of biblical poetry: Parallelism and its history*, New Haven and London: Yale University Press.

Lee, S., Ko, M., Han, K., and Lee, J.-G., 2012, On finding fine-granularity user communities by profile decomposition, in *Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

Lee, J., Hui, Y. C., and Kong, Y. H., 2016, Knowledge-rich, computer-assisted composition of Chinese couplets, *Digital Scholarship in the Humanities*, 31(1), p.152-163.

Lee, J., Kong, Y. H., and Luo, M., 2018, Syntactic patterns in classical Chinese poems: a quantitative study, *Digital Scholarship in the Humanities*, 33(1), p.82-95.

Lee, J. and Wong, T. S., 2012, Glimpses of ancient China from Classical Chinese poems, in *Proc. 24th International Conference on Computational Linguistics (COLING)*.

Levin, S. R., 1962, *Linguistic structures in poetry*, The Hague: Mouton.

Li, H. and Abe, N., 1996, Clustering words with the mdl principle, in *Proc. International Conference on Computational Linguistics (COLING).*

Li, F., and Li, F., 2007, A new approach measuring semantic similarity in HowNet 2000, *Journal of Chinese Information Processing*, 21(3), pp.99-105.

Liu, C.-L., Wang, H., Cheng, W.-H., Hsu, C.-T., and Chiu, W.-Y., 2015, Color aesthetics and social networks in Complete Tang Poems: Explorations and Discoveries, in *Proc.*

29th Pacific Asia Conference on Language, Information and Computation, p.132-141.

Lo, F., 2008, The research of building a semantic category system based on the language characteristic of Chinese poetry (in Chinese), in *Proc. of the 9th Cross-Strait Symposium on Library Information Science*.

Matsuo, Y., Sakaki, T., Uchiyama, K., and Ishizuka, M., 2006, Graph-based word clustering using a web search engine, in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 542—550.

Melamud, O., Levy, O., and Dagan, I., 2015, A simple word embedding model for lexical substitution, in *Proc. Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Mikolov, K., Chen, G. Corrado, and J. Dean, 2013, Efficient estimation of word representations in vector space, in *Proc. ICLR*.

Muneeb, T. H., Sunil, K. S., and Anand, A., 2015, Evaluating distributed word representations for capturing semantics of biomedical concepts, in *Proc. Workshop on Biomedical Natural Language Processing (BioNLP)*.

Newman, M. E. J., 2004, Fast algorithm for detecting community structure in networks, *Phys. Rev. E*, vol. 69, 066133.

Peng, D., 1960, *Quantangshi* [The complete Shi poetry of the Tang], Beijing: Zhonghua shuju.

Segert, S., 1983, Parallelism in Ugaritic poetry, *Journal of the American Oriental Society*, Vol. 103, No. 1.

Sun, A., Grishman, R., and Sekine, S., 2011, Semi-supervised relaion extraction with large-scale word clustering, in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 521—529.

Wang, L., 2003, The metric of Chinese poems 汉语诗律学, Hong Kong: Zhonghua shuju.

Wong, T. S. and Lee, J., 2016, A dependency treebank of the Chinese Buddhist Canon, in *Proc. Conference on Language Resources and Evaluation (LREC)*.

Xu, R., Gao, Z., Pan, Y., Qu, Y., and Huang, Z., 2008, An integrated approach for automatic construction of bilingual Chinese-English wordnet, *Proc. 3rd Asian Semantic Web Conference*.

Yuan, X. 袁行霈 (ed.), 2005, The history of Chinese literature (Vol.2) 《中国文学史（第二卷）》 (in Chinese), Beijing: Higher Education Press.

Zhang, X, and Lapata, M., 2014, Chinese poetry generation with recurrent neural networks, In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.