Introducing a Large Corpus of Tokenized Classical Chinese Poems of Tang and Song Dynasties

Chao-Lin Liu Ti-Yong Zheng Kuan-Chun Chen Meng-Han Chung

National Chenghi University, Taiwan

{chaolin,107753037,105753014}@nccu.edu.tw

Abstract

Classical Chinese poems of Tang and Song dynasties are an important part for the studies of Chinese literature. To thoroughly understand the poems, properly segmenting the verses is an important step for human readers and software agents. Yet, due to the availability of data and the costs of annotation, there are still no known large and useful sources that offer classical Chinese poems with annotated word boundaries. In this project, annotators with Chinese literature background labeled 32399 poems. We analyzed the annotated patterns and conducted inter-rater agreement studies about the annotations. The distributions of the annotated patterns for poem lines are very close to some well-known professional heuristics, i.e., that the 2-2-1, 2-1-2, 2-2-1-2, and 2-2-2-1 patterns are very frequent. The annotators agreed well at the line level, but agreed on the segmentations of a whole poem only 43% of the time. We applied a traditional machine-learning approach to segment the poems, and achieved promising results at the line level as well. Using the annotated data as the ground truth, these methods could segment only about 18% of the poems completely right under favorable conditions. Switching to deep-learning methods helped us achieved better than 30%.

1 Introduction

Word segmentation is an important step for understanding Chinese texts because the Chinese language do not include explicit word delimiters, like the spaces in English, in the texts. Different segmentations of the same statements can lead to different interpretations, so segmenting Chinese strings into correct word sequences is crucial for understanding and processing Chinese in computer systems. Classical Chinese poems typically consist of sequences of short verses, so the quality of word segmentation influences the reading of poems significantly. The segmented poems can facilitate further analysis and applications, e.g., poem styles (Jiang, 2008; Qian and Huang, 2015).

The literature has seen a wide variety of approaches to the problem of word segmentation for vernacular Chinese in the past many years, e.g., (Chen and Liu, 1992; Huang et al., 2007; Chen, Zheng, and Chen, 2015; Deng et al. 2016). Annotated corpora have been created for research and competition as well (Ma and Chen, 2003; Sproat and Emerson, 2003; Emerson, 2005).

In contrast, relatively few researchers of Chinese linguistics and literature discussed word segmentation for classical Chinese poems. Wang (1972) examined the problem from both syntactic and semantic perspectives, while Tsao (2004) argued that the perspective of semantic interpretation should be more natural for native speakers. Jiang (2008) inherited and emphasized more on the semantic viewpoints. Relying on modern databases of Tang poems, Hu and Yu (2001) and Lo (2005) can access and compare more poems conveniently, and they adopt the observations discussed in the previous literature for the word segmentation task.

Beyond conceptual discussion, it is harder to segment words in corpora of classical poems in large scale. Lee and colleagues discussed the topics of annotating part-of-speech tags (2012) and of creating dependency trees (2012) for classical Chinese poems. When they analyzed some interesting syntactic patterns in classical poems, they mentioned around one thousand poems (Lee, Kong, and Luo, 2018).

In this paper, we report a relatively larger scale of work for annotating word boundaries in two collections of classical Chinese poems. At the time of writing, we have annotated 32399 classical Chinese poems of the Tang and Song dynasties.[1] We evaluated our annotations in some different ways. First, we conducted inter-rater agreement (IRA) analysis, and the results are convincing. We applied machine learning methods for segmenting words in classical Chinese poems, and have

[1] Tang is a Chinese dynasty that governed China during 618-907CE. Song is a Chinese dynasty that governed China during 960-1279CE. Both dynasties are very influential for the development of Chinese literature. Compared with the availability of linguistic data of modern days, the amount of available data for classical Chinese poems is extremely scarce.

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 135–144
November 20, 2022. ©2022 Association for Computational Linguistics

achieved and published some preliminary results when we annotated only thousands of poems (Liu and Chang, 2019). We have improved the quality of our word segmenters significantly by using more annotated data and embracing the technology of deep learning. In addition, we compared our annotations with relevant information in a well-known website, and found that our annotations have a reasonable consensus.

We provide information about data sources, define the task of word segmentation, and discuss some domain-dependent heuristics in Section 2. We explain methods for measuring the quality of word segmentation in Section 3. We introduce our annotation team and their annotations, and report a basic statistical analysis of the annotated poems in Section 4. We explored different perspectives for IRA analysis in Section 5. We introduce the probabilistic classifiers for word segmentation in Section 6. We compared the performances of two different designs of the probabilistic classifier in Section 7, and wrap up this paper in Section 8.

2 Data Sources and Problem Definition

We provide a brief introduction to the forms of classical Chinese poems in Section 2.1, and define the task of marking word boundaries in Section 2.2.

2.1 Data Sources: Three Poem Collections

We present two actual poems so that readers can acquire some basic knowledge and relevant terminology about classical Chinese poems.

We list a poem of a famous Tang poet, Li Bai, in the following.[2]

鳳凰臺上鳳凰遊，鳳去臺空江自流。
吳宮花草埋幽徑，晉代衣冠成古丘。
三山半落青天外，二水中分白鷺洲。
總爲浮雲能蔽日，長安不見使人愁。

This poem has eight lines, each of which has seven Chinese characters. The names of this form of poems are *regulated heptametric octaves* (RHO, henceforth) in English and 七言律詩(qi1 yan2 lu4 shi1) in Chinese. If a poem has only four lines, and each line has seven characters, it is in the form of *heptametric quatrains* (HQ, henceforth) and 七言絕句(qi1 yan2 jue2 ju4). Extended forms of heptametric poems (EFHP, henceforth) may have more than eight lines, e.g., 10, 12, 14 lines. Such poems are called 七言長律(qi1 yan2 chang2 lu4) or 七言排律(qi1 yan2 pai2 lu4) in Chinese.

	items	poets	RPO	HQ	RHO	EFPP	EFHP
CTP1	25990	123	11309	5004	7343	1789	545
CSP1	6409	71			6409		

Table 1: Basic statistics about the annotated poems

We list a poem of another famous Tang poet, Du Fu, in the following.[3]

國破山河在，城春草木深。
感時花濺淚，恨別鳥驚心。
烽火連三月，家書抵萬金。
白頭搔更短，渾欲不勝簪。

This poem also has eight lines, each of which has five Chinese characters. The names of this form of poems are *regulated pentametric octaves* (RPO, henceforth) in English and 五言律詩(wu3 yan2 lu4 shi1) in Chinese. If a poem has only four lines, and each line has seven characters, it is in the form of *pentametric quatrains* (PQ, henceforth) and 五言絕句 (wu3 yan2 jue2 ju4). Extended forms of heptametric poems (EFHP, henceforth) may have more than eight sentences, e.g., 10, 12, 14, etc. lines. Such poems are called 五言長律(wu3 yan2 chang2 lu4) or 五言排律(wu3 yan2 pai2 lu4).

In this research, for the Tang poems, we consider only the poems in volumes 30 through 888 in the *Complete Tang Poems* (CTP, Quan Tang Shi, 全唐詩). CTP has 900 volumes, and is the most representative and important collection of Tang poems for the studies on Chinese literature. Volumes 30 through 888 are the ordinary poems. We also annotated the poems in the *Complete Song Poems* (CSP, Quan Song Shi, 全宋詩).

Due to the limited budget for human annotation, we focus on the word segmentation for poems that have only five-character or seven-character lines. These types of poems represent more than 90% of the poems in the CTP. Similarly, 87% of the poems in the CSP consisted of only five-character or seven-character lines.

As a pioneer work, we did not find known principles to select the poems for annotation. As a consequence, we abide by some basic principles. First of all, we wanted to have reasonably many poems of different types of poems. We annotated the majority of the RPO, HQ, RHO, EFHP, EFPP poems that appeared in volumes 30 through 888 in CTP. Table 1 provides statistics about the annotated data. At this moment, we have annotated only part of the RHO poems in CSP.

Table 1 provides the amounts and types of our annotated poems in CTP and CSP. In total, we have 25,990 annotated CTP poems and 6409 annotated

[2] The poet is 李白. The title of the poem is 登金陵鳳凰臺.

[3] The poet is 杜甫, and the title of the poem is 春望.

CSP poems. Some of the CTP poems were repeatedly annotated by different annotators for IRA analysis. The CTP poems belonged to 123 Tang poets, and the CSP poems belonged to 71 CSP poets. The columns RPO, HQ, RHO, EFPP, and EFHP show the amounts of poems of different types. In Table 1, we use CTP1 to refer to the annotated CTP poems and CSP1 to refer to the annotated CSP poems.

For studying the temporal changes and heritage of the Chinese language, we are working on the annotation of thousands of poems in the *Complete Taiwan Poems* (TWP, Quan Tai Shi, 全臺詩) (Shi, 2011).

2.2 Problem Definition

For human annotators, the goal of word segmentation for classical Chinese poems is to add markers between words. If given a line "吳宮花草埋幽徑", the annotators may produce "吳宮=花草=埋=幽徑", where "=" is the marker for word boundaries.

Technically, we treat the word segmentation problem as a classification problem. Given a line "吳宮花草埋幽徑", an annotator attempts to determine whether or not a character in the string is the last character of a word. If the character is not the last character of a word, we assign it to the category of *non-terminal*. If it is, we assign it to the category of *terminal*. We will use *N* and *T* to denote non-terminal and terminal, respectively, in our discussions. Using this notation, the annotators may produce "NTNTTNT" if "吳宮=花草=埋=幽徑" is the correct segmentation for "吳宮花草埋幽徑".

2.3 Domain-Dependent Heuristics

Over the years, based on the experience in studying classical Chinese poems, researchers have proposed practical heuristics about word segmentation that are useful for reading classical Chinese poems. Although the researchers that we cited in the Introduction may not have a consensus on the implications of the popular patterns, they all discussed the high frequencies of the common patterns.

For poems that have 5-character lines, i.e., PQ and RPO, the most common patterns for segmentation are 2-2-1 or 2-1-2. Here, an individual digit represents the number of characters in a segmented word. Hence, the 2-2-1 pattern indicates that we segment a five-character line into three words in the order of a 2-character word, another 2-character word, and a 1-character word. Hence, one may segment "野鶴隨君子，寒松揖大夫" as "野鶴=隨=君子，寒松=揖=大夫", and these are examples of 2-1-2 lines.

Analogously, the researchers believe that 2-2-2-1 and 2-2-1-2 are common patterns for lines in HQ and RHO poems. "雨中=草色=綠=堪染，水上=桃花=紅=欲然" is an example of the 2-2-1-2 pattern.

These heuristic principles are usually right, but there are exceptions. "翻經=謝靈運，畫壁=陸探微" needs the 2-3 pattern to mention person names. One may prefer to read "綠浪東西南北水，紅欄三百九十橋" as "綠浪=東西南北=水，紅欄=三百九十=橋" because of the direction words and the Chinese numbers.

3 Evaluation Measures

3.1 Quality of Word Segmentation

We may measure the quality of word segmentation with four types of measures that are gradually more challenging. Since we are categorizing each character in a poem into two types, it is natural and conventional to measure the classification results with precision, recall, and F_1 measure (Manning and Schütze, 1999; Alpaydin, 2020).

A more practical interest for the task of word segmentation is about word identification. To identify a word, we need to correctly find the beginning and ending of the word, which requires at least two correct classifications. Hence, the percentage of word recovery, PWR, is more challenging than the traditional measures for classification tasks.

We can view the classification of characters as character-level decisions, and view the word recovery as word-level decisions. From here, we can image that there are line-level decisions and poem-level decisions. We may want to measure how well our annotators segment a line completely correct and how well our annotators segment a poem completely correct. Therefore, it should be natural to measure the percentage of perfectly segmented lines, PSL, and the percentage of perfectly segmented poems, PSP. Given a set of *L* lines and *P* poems, if our annotators segment *L'* lines and *P'* poems perfectly, PSL will be *L'*/*L* and PSP will be *P'*/*P*.

We can compare the word segmentations produced by our annotators with the word segmentations annotated by human experts, and compute the precision, recall, F_1, PWR, PSL, and PSP to measure the quality of our classifiers.

3.2 Metrics for IRA Analysis

If we have an expert who will annotate the poems and provide the most reliable annotation of the word boundaries, there would not be a very good reason to ask many annotators to repeat the annotation task. We do not have such an expert yet. More importantly, there might not be just one way to segment a poem because it is possible to segment and interpret poems in different ways. Hence, there might not gold standards for segmenting all classical Chinese poems, at least for some poems.

Therefore, we chose to avoid subjectively decide which annotator is more reliable when comparing the annotators' annotations. We used the Dice coefficient (Dice, 1945) to compare the annotations of a poem that were produced by the annotators.

Let A_1 and A_2 denote the annotations of two annotators. Let C_{12} denote the annotations that both annotators agree. The Dice coefficient for the annotations A_1 and A_2 is defined in (1).

$$\text{Dice}(A_1, A_2) \equiv \frac{2\times|C_{12}|}{|A_1|+|A_2|} \qquad (1)$$

Here, $|A_1|$ and $|A_2|$ are respectively the amounts of annotations (for characters in poems) of A_1 and A_2. Since the annotators are annotating the characters of the same collection of poems, $|A_1|$ and $|A_2|$ must be the same. $|C_{12}|$ is the number of agreed annotations, so $|C_{12}|$ must be smaller or equal to $|A_1|$ (and $|A_2|$). The Dice coefficient doubles $|C_{12}|$ to make the coefficient fall into the range of [0, 1]. When two annotations perfectly agree, the Dice coefficient is 1. When two annotations completely differ, the coefficient will be zero.

Take the annotation for the string "ABCDE" for example. Assume that A_1 is NTNTT and that A_2 is NTTNT, i.e., annotator 1 and 2 segment "ABCDE" into AB=CD=E and AB=C=DE, respectively. The annotators agreed on three character-level decisions, so the Dice coefficient for the character-level decisions is $\frac{2\times3}{5+5}$=0.6. For the word-level decisions, annotator 1 suggests three words, and annotator 2 suggests three words, but they agree on only one word, i.e., AB. Hence, the PWR is $\frac{2\times1}{3+3} = 0.\overline{33}$.

We can reuse the definitions for PSL and PSP in Section 3.1 for inter-rater agreement studies. For PSL and PSP, the annotations for a line or for a poem of two annotators either completely agree or do not agree, so there is no need to arbitrarily choose the ground truth, and we may reuse the original definitions of PSL and PSP.

4 Annotated Poems

4.1 Annotating the Poems

We have seven annotators, and all of them major in Chinese Literature. Four of them are affiliated with the University of Taipei (UT, henceforth), and three are with the National Taipei University (NTPU, henceforth). We intentionally recruited annotators from different universities. Annotators who were trained at different universities and did not know each other may add a bit more independence in their annotation-related decisions.

We could not afford to annotate all of the poems in CTP and CSP because of time limits and budget constraints. In total, CTP and CSP have more than 210,000 items of poems. Sometimes, an item contains multiple poems. We have listed the basic statistics about the current annotated poems in Table 1. The 123 poets for the CTP poems were selected because they were the leading contributors to CTP (Liu, Mazanec, and Tharsen, 2018). In addition to considering the amounts of contributions when selecting the CSP poets, we also considered whether the poets lived in the Northern Song or the Southern Song periods.[4] The poets were selected so that we balanced the poems from these two periods, when huge changes took place in China.

Due to some historical reasons, a poem may have different versions (Owen, 2007; Liu, Mazanec, and Tharsen, 2018). For this reason, we keep the poems that were recorded relatively more consistently in different sources in our studies, hoping to enhance the authenticity of our data.

We stated that we annotated 25990 CTP poems in Section 2.1. In fact, we have annotated more than 25990 items of Tang poems, and chose only this amount in our study. Originally, we have annotated 28137 Tang poems. We compared our poems with the Tang poems that were also listed in the Chinese Text project[5], the Scripta Sinica database[6], and the Cold-Spring website[7], and kept only those items that differ at most one Chinese character with a corresponding item in these reference sites. By comparing and filtering our poems, we hope that the remaining Tang poems are qualified to be used in our empirical evaluation. In the following presentation, we will refer to "items of poems" as

[4] The Song dynasty had two main periods. The Northern Song existed during 960-1127CE, and the Southern Song existed during 1127-1279CE.

[5] CTEXT: https://ctext.org/

[6] http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm

[7] http://skqs.lib.ntnu.edu.tw/dragon/

CTP 5-char poems	**Patterns**	**2-1-2**	**2-2-1**	**2-3**	**1-2-2**	**1-1-2-1**	
	Percentage	56.61	42.11	0.52	0.42	0.19	
	Patterns	**2-1-1-1**	**3-2**	**1-1-1-2**	**1-1-3**	**3-1-1**	**others**
	Percentage	0.06	0.03	0.02	0.01	0.01	< 0.02
CTP 7-char poems	**Patterns**	**2-2-1-2**	**2-2-2-1**	**2-2-3**	**3-1-2-1**	**2-1-2-2**	
	Percentage	58.77	39.05	0.94	0.23	0.23	
	Patterns	**3-1-1-2**	**2-1-1-2-1**	**2-3-2**	**1-2-1-1-2**	**1-2-1-2-1**	**others**
	Percentage	0.20	0.06	0.06	0.05	0.05	< 0.04
CSP RHO poems	**Patterns**	**2-2-1-2**	**2-2-2-1**	**2-2-3**	**2-1-2-2**	**3-1-1-2**	
	Percentage	63.54	34.01	0.96	0.53	0.13	
	Patterns	**1-2-1-1-2**	**2-2-1-1-1**	**3-1-2-1**	**1-1-2-1-2**	**2-1-1-2-1**	**others**
	Percentage	0.11	0.11	0.09	0.07	0.05	< 0.4

Table 2: Distributions of line patterns of annotated CTP and CSP poems (“%” not shown)

“poems” directly because their distinction is not very important for the current study.

4.2 Patterns of the Annotated Poems

We can inspect the patterns of the lines in the annotated poems, and Table 2 shows the distributions of the patterns for the annotated CTP and CSP poems. In the table, we show the percentages of the most frequent 10 patterns for the CTP and CSP poems. We do not show the “%” symbol for succinctness. Based on the statistics in Table 1, we have annotated 13098 (11309+1789) CTP poems that have 5 characters in their lines, and we have annotated 12892 (5004+7343+545) CTP poems that have 7 characters in their lines. We have annotated CSP poems that have 7 characters in their lines.

The experience reported in the literary studies about the common patterns predicts the distributions extremely well (Hu and Yu, 2001; Yu and Hu, 2003; Lo, 2005). More than 98% of the annotated CTP poems that have 5-character lines were annotated as 2-1-2 or 2-2-1 pattern, and we observed 13 patterns for poems that have 5 characters in their lines. More than 97% of annotated CTP poems and CSP poems that have 7-character lines were annotated as 2-2-1-2 or 2-2-2-1 pattern. Both the CTP and CSP 7-character poems have 33 different patterns.

Mathematically, one may have expected that 5-character and 7-character lines may have as many as 16 and 64 different patterns, respectively. A normal classical Chinese poem should follow quite a few phonological, syntactic, and semantic rules, so not all of the patterns are acceptable. Hence, the patterns of the lines are not uniformly distributed. For instance, although possible, the pattern 1-1-1-1-1 for a 5-character line would be very unusual.

	Items	**RPO**	**HQ**	**RHO**	**EFHP**	**EFPP**
UT	20495	8879	4376	5276	1684	280
NTPU	5495	2430	628	2067	105	265

Table 3. Workloads of the annotators

Our statistics support a phenomenon that was discussed circa 1700CE but was not mentioned in modern literature for computing technologies (Hu, 2003 reprint).[8] Frequent patterns like 2-1-2, 2-2-1, 2-2-1-2, and 2-2-2-1 can be expected, but the large proportions of these patterns may be surprising. The 2-3 pattern is many times more frequent than the 3-2 pattern in Table 2.

5 Inter-Rater Agreement Analysis

We report results of our inter-rater agreement analysis in this section, and argue that the observed agreements are not just results of the annotators’ accepting the heuristics that were explained in Section 2.3.

5.1 Results of the Analysis

To further understand our annotated poems, we conducted an IRA analysis using the annotated Tang poems. Table 3 lists statistics for the annotations that were completed by the UT and NTPU annotators. Hence, the amounts of poems listed in Table 3 must agree with the amounts of poems for the CTP in Table 1. For instance, in Table 1, we have 11309 annotated RPO poems, of which 8879 items were annotated by the UT annotators and 2430 were annotated by the NTPU annotators.

We compared the annotations completed by the UT and by the NTPU annotators. A poem that was annotated by a UT annotator and a NTPU annotator is considered as a pair in the IRA studies, and we have 5217 pairs. We compared these 5217 pairs

[8] Both Tsao (2004, p. 59) and Jiang (2008, p. 166) cited Hu (2003, reprint): “五字句以上二下三為脈，七字句以上四下三為脈，其恆也。有變五字句上三下二者，。。。，皆蹇吃不足多學。” Hu was born in the late 16th century.

	Dice for characters	Dice for words	PSL	PSP
observed	95.2	93.0	87.7	42.8
inferred	82.9	70.8	50.0	0.39*

Table 4: Inter-rater agreement analysis ("%" not shown, 0.39 is for regulated octaves)

and calculated the metrics for IRA analysis as we explained in Section 3.2.

The "observed" row in Table 4 lists the statistics for our IRA analysis. The annotators of UT and NTPU showed very high agreement in their decisions as to character and word level decisions. The Dice coefficient for the character classification is 0.952, and the Dice coefficient for common words is 0.930. The percentage that the annotators perfectly agreed on a line is 87.7%, and the percentage that the annotators agreed perfectly on the segmentation of whole poems is only 42.8%.

5.2 A Theoretical Analysis

In this subsection, we derive theoretical estimators, shown in the "inferred" row, for the "observed" row in Table 4 to show that our annotators must not agree with each other only because they might have common belief on the frequent patterns that we explained in Section 2.3. Instead, the expertise and personal judgements of the annotators have also influenced, for otherwise the statistics in the "observed" row could fall as low as those listed in the "inferred" row. We will show the details about this inference procedure in an extended report.

6 Simple Probabilistic Classifiers

Recall that the task of word segmentation can be viewed as classifying characters as a terminal or non-terminal character for a word.

6.1 Directional Pointwise Mutual Information

If we temporarily assume that all the lines of RPO poems used the 2-2-1 or the 2-1-2 pattern and that all the lines of RHO poems used the 2-2-2-1 or the 2-2-1-2 patterns, word segmentation becomes an extremely simplified task. Given these heuristic principles, a simple-minded word segmenter could randomly choose one of the 2-2-1 and 2-1-2 patterns for an RPO poem and choose one of the 2-2-2-1 and 2-2-1-2 patterns for an RHO poem.

A better method is to rely on the directional pointwise mutual information (DPMI) measure to make decisions. Our DPMI is very similar to the traditional pointwise mutual information. The DPMI measures the strength of the closeness of two characters, and we use DPMI(XY) to denote the DPMI of two *consecutive and ordered* characters X and Y.

We can train the DPMI value of two given characters with unannotated poems easily. We define the DPMI value of X and Y, based on their individual appearances and consecutive collocations in poems.

$$DPMI(XY) \equiv log\frac{\Pr(XY)}{\Pr(X)\Pr(Y)} = log\frac{\Pr(Y|X)}{\Pr(Y)} \quad (2)$$

In (2), $\Pr(X)$ and $\Pr(Y)$ are, respectively, the probabilities of reading the unigrams X and Y in the poems, and $\Pr(XY)$ denotes the probability that we see an ordered bigram XY in the poems. Our definition of DPMI is a slight variation of the original definition of pointwise mutual information (PMI) (Manning and Schütze, 1999; Cover and Thomas, 2006), where the computation typically does not consider the orders of X and Y.

Given a line, say "ABCDE" of an RPO poem, we could compare the DPMI measures of CD and DE to determine whether we segment the line into AB-CD-E or AB-C-DE. If DPMI(CD) is larger than DPMI(DE), we choose AB-CD-E; otherwise, we choose AB-C-DE. Given an RHO line, say "ABCDEFG", we segment the line into AB-CD-EF-G if DPMI(EF) is larger than DPMI(FG) and into AB-CD-E-FG otherwise.

6.2 Weighted DPMI

To actually determine the DPMI for a bigram XY, we need to estimate the probability values of $\Pr(X)$ and $\Pr(Y)$ based on a training dataset. We simply employ the maximum likelihood estimator for this task (Alpaydin, 2020; p. 68).

Although we may determine the probability of the bigram XY, $\Pr(XY)$, with the maximum likelihood estimator as well, we chose to add weights to particular bigrams by considering the domain-dependent heuristics that we discussed in Section 2.3.

Given a line of five characters, say "ABCDE", we could consider two different segmentations, and they are AB=CD=E or AB=C=DE. Under this presumption, we assign a base weight, $\boldsymbol{\beta}$, to all of the bigrams in "ABCDE", i.e., "AB", "BC", "CD", and "DE", and we give extra weights to "AB", "CD", and "DE" because of their positions in the line. If the segmentation of "ABCDE" must be either "AB=CD=E" or "AB=C=DE", we essentially have assumed that "AB" is a bigram, so we give a starting weight, $\boldsymbol{\sigma}$, to the starting bigram of each line. We give an additional weight, $\boldsymbol{\alpha}$, to "CD" and "DE" because one of them should be a bigram.

Given a line of “ABCDE”, “AB” will gain β+σ in its total weight, “BC” will gain β, “CD” will gain β+α, and “DE” will gain β+α. If the assumptions about the patterns are reasonable, we hope that the values of the weighted DPMI will be more informative than the raw frequency that is used for maximum likelihood estimators.

We set β, σ, and α to 0.3, 1, and 0.5, respectively, in our current study. Obviously, we may try other combinations in our experiments. We set β to a relatively small value because it provides a basic weight to all bigrams. Since “AB” is relatively more certain than “CD” and “DE” to form a bigram, the starting weight is not smaller than the additional weight. We set σ to one because, if accepting the heuristics explained in Section 2.3, the staring bigrams of each line are two-character words. We set α to 0.5 because, in an “ABCDE” line, one of “CD” and “DE” will be a word, so they share the starting weight equally.

We use the total weights of bigrams observed in the training set to calculate the probability of bigrams. Every observed bigram in the training set will accumulate their own total weights, and the probability of a bigram, $\Pr(XY)$, is defined as its total weight, $\mathrm{TW}(XY)$, divided by the overall weights of all bigrams in the training set.

$$\Pr(XY) \equiv \frac{\mathrm{TW}\,(XY)}{\sum_{z\,\in\{the\ bigrams\ in\ the\ training\ set\}}\mathrm{TW}(z)} \tag{3}$$

We will refer to this score function as WDPMI. Note that we establish WDPMI from a probabilistic perspective, but we did not verify whether the resulting weights conform to the axioms of probability properly.

When we apply the weighted DPMI for segmenting the test data, we must be prepared for encountering unseen unigrams and unseen bigrams in the test data. This is because we must strictly separate the test data from the training data (Alpaydin, 2020). As a consequence, we need to handle unseen unigrams and bigrams in the test data. For these cases, we assign them the minimum DPMI for the unigrams or bigrams that we have seen in the training data. This choice is inspired by the Good-Turing smoothing method (Good, 1953).

6.3 Training DPMI and WDPMI

Since we do not need labeled data to train DPMI or WDPMI, we can employ more poems for training the classifiers.

Again, although we do not have theoretical rules to follow and select the poems for training, we do abide by some basic principles. First of all, we wanted to have reasonably many poems for training. We can use all of the PQ, RPO, HQ, RHO, EFHP, EFPP poems that appeared in volumes 30 through 888 in CTP for training. We chose to consider only the RPO and RHO poems in CSP for training because the total of these two types was already more than the CTP poems that we could use for training. Here we also have some TWP poems.

	items	poets	PQ	RPO	HQ	RHO	EFHP	EFPP
CTP2	36562	2257	2183	11859	6960	6970	7222	1368
CSP2	74505	3608		32929		41576		
TWP	58267	99	2220	5451	31614	18982		

Table 5. Statistics about more poems

Table 5 reuses the format of Table 1, but lists the number of labeled and unlabeled poems that we have in the CTP, CSP, and CWP. The Tang and Song poems that we listed in Table 1 are subsets of the poems that we listed in Table 5. We use CTP2 and CSP2 in Table 5 to differentiate the different sets in Tables 1 and 5. Notice that, although we have 6970 RHO items in CTP2, we have 7343 annotated RHO items in CTP1 (Table 1). This is because a CTP poem may be annotated multiple times by different annotators, even when we may not annotate all of the poems in CTP2 and CSP2. A repeatedly annotated poem is counted multiple times in CTP1 and is counted only once in CTP2.

7 Empirical Evaluations

Since we discussed the differences between DPMI and traditional PMI, and we claimed the superiority of weighted DPMI (WDPMI) against DPMI. We conducted a wide variety of experiments to verify this projection.

Since we will use the CTP1 and CSP1 as the test data, we will remove the poems in CTP1 and CSP1 from CTP2 and CSP2, respectively, at training time. We do not indicate this exclusion in Table 6. We can use different combinations of unannotated data (Table 5) as the training data and use different annotated data (Table 1) as the test data to check whether WDPMI indeed prevails.

We list 14 such experiments and their results in Table 6. In Table 1, we have two sets of annotated data. CTP1 and CSP1 are for the Tang (618-907CE) and Song dynasty (960-1279CE), respectively. In Table 5, we have three basic sets of unannotated data. In addition to CTP2 and CSP2, we added TWP. Therefore, we can create seven combinations of these three sets for training in different experiments.

Recall the definition for WDPMI and our discussion in Section 6.2. We set β, σ, and α to 0.3, 1, and 0.5, respectively, for the experiments in

ID	TrainD	TestD	WDPMI						DPMI			
			prec	recl	F_1	PWR	PSL	PSP	F_1	PWR	PSL	PSP
7	CTP2	CTP1	90.25	90.43	90.34	86.27	76.32	16.79	89.19	84.63	73.54	13.42
8	CTP2	CSP1	91.17	91.36	91.27	86.86	73.56	11.11	90.41	85.58	71.05	8.07
9	CSP2	CTP1	90.34	90.53	90.44	86.40	76.55	17.01	89.39	84.92	74.03	13.99
10	CSP2	CSP1	91.97	92.17	92.07	88.07	75.97	13.82	91.24	86.82	73.51	10.38
11	TWP	CTP1	89.30	89.48	89.39	84.93	74.04	14.03	88.42	83.56	71.72	11.79
12	TWP	CSP1	91.06	91.25	91.16	86.70	73.27	10.15	90.41	85.57	71.04	8.05
13	CTP2+CSP2	CTP1	90.73	90.92	90.82	86.95	77.48	18.32	89.84	85.56	75.10	15.25
14	CTP2+CSP2	CSP1	92.09	92.29	92.19	88.24	76.33	14.42	91.44	87.12	74.10	11.06
15	CTP2+TWP	CTP1	90.48	90.67	90.58	86.60	76.89	17.51	89.60	85.23	74.55	14.51
16	CTP2+TWP	CSP1	91.76	91.96	91.86	87.75	75.32	13.16	91.10	86.61	73.08	10.06
17	CSP2+TWP	CTP1	90.45	90.64	90.54	86.56	76.81	17.51	89.57	85.18	74.47	14.60
18	CSP2+TWP	CSP1	92.03	92.23	92.13	88.15	76.12	14.01	91.37	87.02	73.89	10.79
19	CTP2+CSP2+TWP	CTP1	90.76	90.95	90.85	86.99	77.55	18.46	89.89	85.63	75.23	15.41
20	CTP2+CSP2+TWP	CSP1	92.22	92.42	92.32	88.43	76.69	14.91	91.51	87.22	74.29	11.47

Table 6. WDPMI consistently offers better performances than DPMI. ("%" not shown)

	prec	recl	F_1	PWR	PSL	PSP
max	1.15	1.16	1.16	1.64	2.78	3.44
median	0.91	0.91	0.91	1.32	2.34	3.06
mean	0.90	0.90	0.90	1.30	2.38	3.03
min	0.75	0.75	0.75	1.12	2.23	2.11

Table 7. Differences in performance when comparing WDPMI with DPMI ("%" not shown)

Table 6. An unweighted version of DPMI can be considered as a special case of WDPMI without giving special weights. Namely, we could set β, σ, and α to 0.3, 0, and 0, respectively. Due to the limitation of page width, we do not show the values of precision and recall for DPMI in Table 6.

We could verify that using WDPMI indeed led to better performances than using DPMI, if we compare the corresponding statistics in Table 6. Each of the statistics in the shaded area in the WDPMI column is larger than the corresponding statistic in the DPMI column.

We could calculate the differences between the metrics of WDPMI and DPMI by subtracting an item for DPMI from the corresponding item for WDPMI. For Exp. 7, the difference in PSP is 3.37. We can calculate the differences in PSP for 14 experiments, and obtain their maximum (3.44), median (3.06), mean (3.03), and minimum (2.11). The rightmost column in Table 7 shows these results. We repeated such a calculation procedure for precision (prec), recall (recl), F_1, PWR, and PSL for Table 6, and show the results in Table 7. The statistics of 14 experiments in Table 7 consistently suggest that using WDPMI led to better performance than using DPMI.

We can compare the performances of WDPMI and DPMI from other perspectives, and we can include more domain knowledge about the classical poems to improve the performances of our probabilistic classifiers in an extended report of our work. Of course, with the annotated poems, we could apply deep learning (Goodfellow et al., 2016) and other machine learning methods to train and test classifiers that may further enhance the quality of word segmentation.

8 Concluding Remarks

The main purpose of this paper is to report the annotation of word boundaries for 32399 classical Chinese poems. Seven annotators of Chinese literature background carried out the task. To investigate the quality of these human annotation, we conducted inter-rater agreement studies. In fact, we have also compared the annotations with some relevant information extracted from the Sou-Yun website[9], which is a highly recommended website for learning classical Chinese poems, but we cannot provide the details here. Based on these further analyses, we gained confidence on the quality of our annotations.

We have used the annotated data to train classifiers for algorithmically segmenting classical Chinese poems. It was relatively easy to segment the lines in poems correctly, but remained challenging to segment poems completely correct. We understand that there may not be "the" correct answer to segment a poem. "The" correct answer

[9] https://sou-yun.cn/

depends on how a reader interpret the poem. Nevertheless, for the studies of computer science, we used the annotated data as the ground truth in our analysis. The annotators achieved perfect agreement for a given poem 43% of the time. Under favorable conditions when domain-heuristics are applicable, using a traditional machine-learning method, we segment a poem completely correctly 18.46% in Table 6. Switching to deep-learning methods, we could improve the results to slightly above 30%. Details about these new experiments can be provided in an extended paper.

Responses to the Reviewers

Although we briefly discussed the challenges to segment the poems for the “ground truth” that typical experts of computer science background would expect at the beginning of Section 3.2, a reviewer still commented for more discussions on this issue. Almost no one who has reasonable experience in reading Chinese poems would deny that poets might intentionally leave a certain degree of ambiguity in poems for beauty, imageries, hidden intentions, etc. We recognize this level of difficulty as well, but we also hope that it is possible that, for a majority of poems, readers may have an acceptable consensus about the interpretation of a poem. Whether our hope will hold from the perspectives of experts in Chinese literature is subject to more further studies.

A reviewer encouraged us to show the usability of our corpus via higher level of tasks for natural language processing, including named entity recognition and slot tagging (Xu and Sarikaya, 2013). We would like to extend our work in those directions after we first establish the position of the current corpus in the academic world via the discussions in this presentation.

In further experiments, we can elaborate on how using deep learning techniques can outperform the performance of using the heuristics WDPMI. Machines can learn the frequent patterns of classical Chinese poems directly via labeled data, without the need of relying on human’s heuristics.

Acknowledgments

We thank the reviewers for their helpful comments on the original submissions. This research was support in part by projects 107-2221-E-004-009 - MY3 of the Ministry of Science and Technology, 110-2221-E-004-008-MY3 of the National Science and Technology Council of Taiwan, and 111H124D-13 of the National Chengchi University in Taiwan.

References

Alpaydin, E. 2020. *Introduction to Machine Learning*, fourth edition, Cambridge: The MIT Press.

Chen, K.-J. and Liu, S.-H. 1992. Word identification for mandarin Chinese sentences, *Proceedings of the Fourteenth Conference on Computational Linguistics*, 101–107.

Chen, Y., Zheng, Q., and Chen, P. 2015. A boundary assembling method for Chinese entity-mention recognition, *IEEE Intelligent Systems*, 30(6):50–58.

Cover, T. M. and Thomas, J. A. 2006. *Elements of Information Theory*, second edition, New Jersey: Wiley-Interscience.

Deng, K., Bol, P. K., Li, K. J., and Liu, J. S. 2016. On the unsupervised analysis of domain-specific Chinese texts, *Proceedings of the National Academy of Sciences*, 113(22):6154–6159.

Dice, L. R. 1945. Measures of the amount of ecologic association between species, *Ecology*, 26(3):297–302.

Emerson, T. 2005. The second international Chinese word segmentation bakeoff, *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 123–133.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters, *Biometrika*, 40(3/4):237–264.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*, Cambridge: The MIT Press.

Hu, J. and Yu, S. 2001. The computer aided research work of Chinese ancient poems (唐宋诗之计算机辅助深层研究), *Acta Scientiarum Naturalium Universitatis Pekinensis* (北京大学学报(自然科学版)), 37(5):727–733. (in Chinese)

Hu, Z.-H. (胡震亨，1569-1645?) 2003. *Tangyin Tongqian* (唐音統籤) (reprint), Shanghai: Shanghai Guji Chu Ban She (上海古籍出版社). (in Chinese)

Huang, C.-R., Šimon, P., Hsieh, S.-K., and Prévot, L. 2007. Rethinking Chinese word segmentation: tokenization, character classification, or wordbreak identification, *Proceedings of the Forty-Fifth Annual Meeting of the Association for Computational Linguistics*, 69–72.

Jiang, S. 2008. *A Linguistic Research for Tang Poems* (唐詩語言研究), Beijing: Language&Culture Press (語文出版社). (in Chinese)

Lee, J. 2012. A classical Chinese corpus with nested part-of-speech tags, *Proceedings of the Sixth EACL*

Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 75–84.

Lee, J. and Kong, Y. H. 2012. A dependency treebank of classical Chinese poems, *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 191–199.

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):82–95.

Liu, C.-L. and Chang, W.-T. 2019. Onto word segmentation of the *Complete Tang Poems*, *Proceedings of the 2019 International Conference on Digital Humanities*.

Liu, C.-L., Mazanec, T. J., and Tharsen, J. R. 2018. Exploring Chinese poetry with digital assistance: Examples from linguistic, literary, and historical viewpoints, *Journal of Chinese Literature and Culture*, 5(2):276–321.

Lo, F. 2005. Design and applications of systems for word segmentation and sense classification for Chinese poems (詩詞語言詞彙切分與語意分類標記之系統設計與應用), *Proceedings of the Fourth Conference of Digital Archive Task Force* (第四屆數位典藏技術研討會論文集) (in Chinese)

Ma, W.-Y. and Chen, K.-J. 2003. Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff, *Proceedings of the Second SIGHAN workshop on Chinese Language Processing*, 168–171.

Manning, C. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*, Cambridge: The MIT Press.

Owen, S. 2007. A Tang Version of Du Fu: The Tangshi Leixuan, *Tang Studies*, 25:57–90, 2007. (DOI: 10.1179/073750307790779469)

Qian, P. and Huang, X. 2015. The statistical modeling and macro-analysis of Chinese classical poetry (中国古诗统计建模与宏观分析), *Journal of Jiangxi Normal University* (Natural Science), 39(2):117–123. (in Chinese)

Shi, Y.-L. (ed.) 2011. *The Complete Taiwan Poems*, National Museum of Taiwan Literature. (https://www.nmtl.gov.tw/publicationmore_149_306.html) (in Chinese)

Sproat, R. and Emerson, T. 2003. The first international Chinese word segmentation bakeoff, *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 133–143.

Tsao, F.-f. 2004. *Some Linguistic Analyses of Chinese Literature: Three Studies of Tang and Song Poems* (從語言學看文學：唐宋近體詩三論), Taipei: Academia Sinica. (in Chinese)

Wang, L. 1972. The Rhymes in Chinese Poems (漢語詩律學), in the *Collection of Wang Li* (王力全集) (reprint), Shangdong: Shangdong Education Press (山東教育出版社) (in Chinese)

Xu, P. and Sarikaya R. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling, Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, 78–83.

Yu, S. and Hu, J. 2003. Word-based statistical analysis of Chinese ancient poetry (唐宋詩之詞匯自動分析及應用), *Language and Linguistics*, 4(3):631–647. (in Chinese)