Generating Chinese Classical Poems with Statistical Machine Translation Models

Jing He*
Tsinghua University
Beijing, China, 100084
hejing2929@gmail.com

Ming Zhou
Microsoft Research Asia
Beijing, China, 100080
mingzhou@microsoft.com

Long Jiang
Microsoft Research Asia
Beijing, China, 100080
longj@microsoft.com

Abstract

This paper describes a statistical approach to generation of Chinese classical poetry and proposes a novel method to automatically evaluate poems. The system accepts a set of keywords representing the writing intents from a writer and generates sentences one by one to form a completed poem. A statistical machine translation (SMT) system is applied to generate new sentences, given the sentences generated previously. For each line of sentence a specific model specially trained for that line is used, as opposed to using a single model for all sentences. To enhance the coherence of sentences on every line, a coherence model using mutual information is applied to select candidates with better consistency with previous sentences. In addition, we demonstrate the effectiveness of the BLEU metric for evaluation with a novel method of generating diverse references.

Introduction

This paper presents a novel approach to generating Chinese classical poetry using the quatrain form as a test bed. Chinese classical poetry is an important cultural heritage with over 2,000 years of history. There are many genres of classical Chinese poetry but quatrain (绝句) and lüshi (律诗) are the most popular. A quatrain is a poem consisting of four lines and a lüshi is a poem of eight lines. Each line of a quatrain or a lüshi contains five or seven characters. Strict tonal pattern, rhyme scheme and structural constraints are required on the quatrain and lüshi as explained below.

a) Tonal pattern

In the traditional Chinese language, every character consists of one syllable and has one tone which can be either "Ping" (level tone), or "Ze" (downward tone). The general principle for the tonal pattern in a Chinese classical poem is that these two kinds of tones should be interleaved in each sentence. There are four common tonal patterns for the 5-char and 7-char quatrain (Wang, 2002). We show one of the tonal patterns widely used for a 5-char quatrain, where "+" indicates the "Ze" tone, "-" indicates the "Ping" tone, and "*" indicates that either tone is acceptable for the position.

* + - - +
- - + + -
* - - + +
* + + - -

b) Rhyme scheme

In Chinese poetry, rhyming characters have the same ending vowel. The rhyming constraint for a Chinese quatrain is that the ending characters of the second and the fourth sentences should rhyme.

c) Structural constraint

The structure of a Chinese quatrain often follows the "beginning, continuation, transition, summary" template which means that the first sentence starts a topic, the second sentence continues the topic, the third sentence expresses something new on the starting topic, and the fourth sentence expresses or implies a general conclusion.

Below is an example of 5-char quatrain.

白 日 依 山 尽，(- + - - +)
white sunlight along hill fade
黄 河 入 海 流。(- - + + -)
Yellow River into sea flow
欲 穷 千 里 目，(+ - - + +)
wish exhaust thousand mile eyesight
更 上 一 层 楼。(+ + + - -)
More up one story tower

This quatrain, entitled "On the Stork Tower" was written by Zhihuan Wang, a famous poet of the Tang Dynasty. The following is the English translation.

*This work was done when the first author was a visiting student at Microsoft Research Asia.

*This work was supported in part by the National Basic Research Program of China Grants 2011CBA00300, 2011CBA00301, and the National Natural Science Foundation of China Grants 61033001, 61061130540, 61073174.

As daylight fades along the hill,
The Yellow River flows into the sea.
Seeking a thousand mile view,
I mount another story of the Tower.

The quatrain describes the author's feeling when he climbed the Stork Tower beside the Yellow River. As we can see, it fully satisfies the above-mentioned constraints of the quatrain form.

In this paper, we propose a statistical approach to generating Chinese classical poetry with the following steps:

1. Given a set of keywords provided by a writer as the description of his/her intent, we use a template-based model to generate candidates for the first sentence of the poem and apply a poetry language model to select the best one.
2. We then use a phrase-based statistical machine translation (SMT) model to generate the remaining sentences one by one. The SMT model takes as input the current sentence as the source language and outputs the next sentence as the target language.

Our contribution includes:

1. As every sentence of a poem plays a certain function in the semantic structure of a poem, sentence position-sensitive SMT models are employed, with each model generating a sentence with a particular semantic function.
2. Furthermore, information from all previously generated sentences is considered by the model when generating the next sentence, in order to ensure high coherence between sentences.
3. The BLEU metric is applied to the evaluation of the generated poems, and the method of generating references is introduced. The BLEU score of our system shows high correlation with the scores given by human experts.

Related Work

As a famous example for automatic Chinese poetry generation, the Daoxiang poem generator (http://www.poeming.com/web/index.htm) is one of several popular web sites providing a poem generation service. It uses the manually made linguistic template-based approach. In the Daoxiang system, a user needs to input a title and select a rhyming template. Then, the system acquires from a manually created table a list of phrases or words related to the keywords in the given title. The system then randomly selects a few words or phrases from the list and inserts these into the user-selected rhyming template to generate lines and finally form the poem. The poems generated do satisfy the tonal and rhyming constraints, but every sentence generated is not natural, and the entire poem has difficulty delivering a centralized meaning that logically corresponds with the user's intent as reflected by the title.

In the area of computer-assisted Chinese poetry generation, which is different with automated poetry generation that this paper focuses, Luo has developed a tool that provides the tonal templates for genres of Chinese poetry and a dictionary of the tone of Chinese characters. (http://cls.hs.yzu.edu.tw/MakePoem/HOME.HTM) In a study of Chinese couplet generation, which could be considered a narrowed case of Chinese poetry generation, (Jiang and Zhou, 2008) achieve promising results with a statistical machine translation model for generating the second sentence given the first one of a couplet. But no simple extension of this approach can accomplish strong coherence among four lines while meeting global structural constraints, such as with "beginning, continuation, transition, summary" templates. This paper intends to solve it. We also study other systems that generate English poems and Spanish poems (Oliveria, 2009), but their methods are quite different from ours.

Overview of the Proposed Approach

As a simplified expression, a poem's topic can be represented by a few keywords. After the author has formed the outline of his or her thoughts, he/she chooses the genre of the poem, for instance, he selects quatrain. Then he needs to select a tonal pattern. Next the chosen tonal template is filled in with qualified characters to complete a line. Then the author continues to write the next line until the entire poem is finished. This is an iterative process in which the author can always change ideas, or try another tonal pattern, or swap the sequence of lines or change characters in a line. Our generation system follows these phases of human poem writing. The system accepts the users' intent concerning the content and format and generates sentences one by one to create a poem. The following information is used to express a user's intent: 1) Several keywords that a user inputs to describe the topic of the poem. For example, "spring", "mountaineering", and "climb" can be used to express the topic of climbing a mountain in spring. 2) The length of each line. The author can select five or seven characters. 3) The tonal pattern.

To simplify the experiments, we constrain the keywords that users may enter to describe the topics to the poetic phrase taxonomy (Liu 1735) "ShiXueHanYing" (诗学含英), a dictionary used for teaching classical poetry to children. In this taxonomy, 41,218 phrases (34,290 unique) of length ranging from 1 to 5 characters have been classified into 1,016 clusters. A phrase can be classified into multiple clusters. Each cluster is assigned a concept such as "spring", "mountaineering", etc. As a result, the cluster of "spring" would contain those phrases widely

used in Chinese poems depicting the spring season. 1,016 concepts cover most of the common topics in ancient Chinese poems.

In our experiments, after the user inputs the above-mentioned information, we use the template-based method to generate the first sentence and then use a statistical machine translation model to generate the other three sentences one-by-one.

Template-based Generation of First Sentence

After the user selects the tonal pattern and several keywords, an intuitive way to generate the target poem is to find some phrases related to the selected keywords and assemble them into four sentences according to the tonal pattern. Since we have constrained the user to select keywords from the cluster names in the taxonomy of "ShiXueHanYing", to get candidate phrases related to the user selected keywords, all phrases included in the corresponding clusters are selected. Then, for generation of each sentence, we build a phrase lattice by putting every candidate phrase into every position of the tonal pattern if the phrase satisfies the tonal constraint for the position. The tone of each character in the phrase is obtained using a tone and rhyming dictionary of Chinese characters, called "Pingshui Rhyme."(Named 《平水韵》 in Chinese, which is a tone and rhyme table for poetry writing compiled by Yuan Liu in the Song Dynasty of China.)

To search for the best sentences in each lattice, we introduce a language model that scores all paths in the lattice. Here, a character-based trigram language model trained with Katz back-off from the training data is used. Training details are covered in Section 6. The Forward-Viterbi-Backward-A* algorithm is used to find the N-Best candidate sentences to facilitate the user's selection working in the interactive model.

Statistical Quatrain Generation with SMT models

(Jiang and Zhou, 2008) proposed a statistical machine translation (SMT) method to generate the second sentence given the first sentence of a Chinese couplet (which can be considered a special kind of two-sentence poetry). Inspired by this, we applied a similar SMT-based approach to generate the next sentences given the previous sentences of a quatrain. When generating the second sentence, we regard the first one as the source language in the MT task and regard the second one as the target language sentence. This approach continues to generate the third and the fourth sentences in the same way. Using an SMT model for quatrain generation has several advantages. The statistical learning based method can easily learn the knowledge of poem generation by training over a large poetry corpus.

The direct use of this SMT approach, however, faces the difficulties of assuring the coherence between all generated sentences and the constraints of semantic structure over a whole poem. To handle these challenges, firstly, because the SMT model relies on an input sentence to generate the next sentence and thus it cannot generate the first sentence with user's keywords, we use the template-based approach to generate the first sentence. Secondly, to maintain a meaningful semantic structure throughout the poem, we use three different SMT models, to generate the 2nd sentence, the 3rd sentence, and the 4th sentence, respectively, rather than using one single model for all these three sentences. These are thus position-sensitive models. Thirdly, to improve the coherence among the sentences, we incorporate a coherence model using the keywords in all previous sentences into the SMT framework to rank the current sentence candidates so that the current sentence preserves a strong association with the previous sentences.

Basic SMT Models

Specifically, given a sentence of a quatrain F denoted as $F = \{f_1, f_2, ..., f_n\}$ (i = 1, 2, 3), our objective is to seek the next sentence S, where fi and si are Chinese characters, such that p(S|F) is maximized. Following (Och and Ney, 2002), we use a log-linear model as follows.

$$S^* = \arg\max_S p(S \mid F) = \arg\max_S \sum_{i=1}^{M} \lambda_i \log h_i(S, F) \quad (1)$$

Where hi(S, F) are feature functions and M is the number of feature functions. In our design, characters are used instead of words as translation units to form phrases because in ancient Chinese, most words consist of only one character and Chinese word segmentation may introduce unexpected errors. To apply the features, S and F are segmented into phrases $\bar{s}_1 ... \bar{s}_I$ and $\bar{f}_1 ... \bar{f}_I$, respectively. We assume a uniform distribution over all possible segmentations.

$h_1(S,F) = \prod_{i=1}^{I} p(\bar{f}_i \mid \bar{s}_i)$	Phrase translation model
$h_2(S,F) = \prod_{i=1}^{I} p(\bar{s}_i \mid \bar{f}_i)$	Inverted phrase translation model
$h_3(S,F) = \prod_{i=1}^{I} p_w(\bar{f}_i \mid \bar{s}_i)$	Lexical weight
$h_4(S,F) = \prod_{i=1}^{I} p_w(\bar{s}_i \mid \bar{f}_i)$	Inverted lexical weight
$h_5(S,F) = p(S)$	Language model

Table 1. Features Used in the SMT Model.

Following (Jiang and Zhou, 2008), the five features listed in Table 1 are selected in our model.

This is the basic model for the generation of the next sentence. As we explained before, however, we have to make several important enhancements to improve semantic structure and coherence. In the following subsections, we explain each of the extensions in detail.

Position-sensitive SMT Models

As introduced above, the functions of the four sentences in a Chinese quatrain can be roughly summarized as "beginning, continuation, transition, and summary" respectively. Therefore the relationship between adjacent sentences at different positions of a quatrain will differ. A naïve translation model may not handle such dependencies well. In our method, we design three specific translation models for generating the second, third, and fourth sentences respectively, where the translation models actually includes the first four feature functions listed in Table 1. Each model is trained with the sentence pairs at that position. For instance, to train the model for the second sentence generation, we only use the sentence pairs of the first sentence and second sentence in the poem corpus. To alleviate data sparseness, each model is interpolated with a general model trained with all sentence pairs regardless of the positions in a poem. When using the translation probabilities in decoding, we empirically assign higher weights α_1 to the probabilities from the specific model and lower weights $(1-\alpha_1)$ to those from the general model.

$$TM = \alpha_1 TM_s + (1-\alpha_1) TM_b \qquad (2)$$

Coherence Models

The models used in Formula (1) only consider the information between adjacent sentences of a quatrain. Consequently, the generated four sentences do not have strong meaningful association to each other. In a good quatrain, however, all four sentences should be coherent in order to deliver a centralized meaning. Coherence is also considered useful in text summarization for handling macro-level relations between clauses or sentences (Mani et al., 1998). In order to keep the next sentence coherent with all the previously generated sentences, the Mutual Information (MI) score is added as the sixth feature $h_6(S,F)$ to the SMT model, which measures the association between the next sentence and the sentences already generated before. Let *Set*-A be the characters appearing in the next sentence to be generated and *Set*-B be the characters in all previously generated sentences. All mutual information values between any two characters in *Set*-A and *Set*-B are summed in the coherence model:

$$h_6(S,F) = \sum_{i,j} MI(s_i, s_j) = \sum_{i,j} \log \frac{p(s_i, s_j)}{p(s_i)\,p(s_j)} \qquad (3)$$

Where S_i and S_j indicate characters in Set-A and Set-B respectively, and the parameters $p(S_i, S_j)$, $p(S_i)$, and $p(S_j)$ indicate the probabilities of S_i and S_j co-occurring, S_i occurring, and S_j occurring in a poem, respectively.

Decoding

After the first sentence of the target quatrain is obtained with the template-based approach, the remaining three sentences are generated one by one with the proposed SMT-based approach. For each sentence, we use a phrase-based decoder similar to that used by (Koehn et al., 2003) to generate an N-best list of candidates.

Because there is no word reordering operation in our next sentence generation, our decoder is a monotonic decoder. In addition, poetic sentences are often shorter than typical source sentences in the machine translation task, so our decoder is more efficient than traditional translation decoders. Moreover, translation options are pruned if they violate the author-specified tonal pattern.

According to the rhyming constraints of the Chinese quatrain, the ending character of the fourth sentence should follow the same rhyme of the ending character of the second sentence. In order to satisfy this constraint, when decoding for the fourth sentence, we remove those translation options at the last position of the sentence whose ending character does not rhyme with that of the second sentence. Note that the process of our poetry generation can be run in an interactive mode, which means that for each of the four sentences, the N-best candidates can be presented for user-selection. The system will use the user-selected sentence as input to continue to generate the N-best candidates of the next sentence. With this mode, the user can control the generation such that the final quatrain best conveys his or her leading idea.

Model Training

To estimate the weights in formulas (1), (2), and (4), we use the Minimum Error Rate Training (MERT) algorithm, which is widely used for phrase-based SMT model training (Och, 2003). The training data and criteria (BLEU) for MERT will be explained in Section 7. The training data for the translation and language models are discussed below.

Data for Translation Model Training

We downloaded from the Internet the <Tang Poems>, <Song Poems>, <Ming Poems>, <Qing Poems>, and <Tai Poems>, which amount to more than 3,500,000 sentences. From one quatrain, we extracted three sentence pairs (first-second, second-third, and third-fourth) and from an eight-sentence lüshi, we get seven pairs. Altogether, we constructed more than 1,200,000 sentence pairs as training data for our translation models. The smoothing details of the general translation models are shown in Section 5.2.

Data for Language Model Training

The data for training the language model includes two parts: One is the corpus of poems mentioned above, which contains more than 3,500,000 sentences, and the other is a corpus of ancient articles obtained from the Internet, including 《文选》, 《唐宋八大家散文选》, etc, which contains about 12,000,000 sentences, and serves as the complementary data. Using the BLEU evaluation, we trained two character-based trigram models on the two parts of data, denoted by p1(s) and p2(s), which stand for the poem language model and article language model, respectively. Then we linearly combined the two models to get the final one as follows, where the weights are automatically trained by the Minimum Error Rate Training (MERT) algorithm, See Section 7.4.

$$p(s) = \beta_1 p_1(s) + (1 - \beta_1) p_2(s) \quad (4)$$

Evaluation

We use BLEU (Papineni et al., 2002), which is widely used for automatic evaluation of machine translation systems. First we will explain the method of building references for BLEU. Second, we verified the effectiveness of BLEU by computing the correlation with human evaluation. Finally, our systems with the proposed position-sensitive SMT models and coherence model are evaluated with BLEU metric. To verify our method, we conducted experiments over 5 different systems. Following (Papineni et al., 2002), we use 5 systems of different levels for parallel experiments, to exhibit the proper usage of BLEU metric. So first, we built two simple systems with only 1/50 and 1/10 of a fraction of the training data, called "Sys-1/50" and "Sys-1/10" respectively. These two systems use the base SMT approach as "Sys-Basic". The third one, called "Sys-Basic," is trained with all sentence pairs, rather than using position-sensitive models trained with specific sentence pairs at certain positions, and a coherence model. Then, we evaluated the effectiveness of two advanced systems obtained by respectively adding the position-sensitive models and the coherence model to the baseline SMT System with the whole data set. The two advanced systems are called "Sys-Position" that means "Position-sensitive models" and "Sys-Coherence" that denotes "Coherence model." estimate the weights in formulas (1), (2), and (4), we use the Minimum Error Rate Training (MERT) algorithm,

BLEU Metric

The BLEU metric is widely used for automatic evaluation of machine translation systems. It ranges from 0 to 1 and a higher BLEU score stands for better translation quality. In the work of couplet generation with an SMT approach (Jiang and Zhou, 2008), they use BLEU for automatically evaluating generated couplets. We think it's also useful for evaluating poem-generating systems as couplet is a special type of poem. Here we use 1-gram for BLEU because in ancient Chinese most words consist of only one character.

Build Reference for BLEU

The biggest difficulty in using BLEU to evaluate the poems is to get the reference sentences. In order to compute BLEU, we need references made by humans. In (Jiang and Zhou, 2008), given the first sentence of a couplet, they collect a set of second sentences written by humans and then compare the n-gram precision between the next sentences generated by the machine and humans. They use an online forum to collect references for a certain first line. In the online forum, someone challenges the other users by giving a first line of the couplet and users give their own answers for the next line, which are collected as references. Different from couplets, in the poetry generation, given the same key words and the first sentence, the poems generated by humans can be quite diverse, thus a numerous number of next sentences with different writing styles and characters are required to ensure the evaluation quality. Due to there not being a large enough number of poets, an automatic reference selecting method is used based on the following idea: If two sentences share similar keywords, the next sentences can be references for each other. For example, "月明花满地" and "明月临沧海" share keywords "明" and "月", the next sentence of "月明花满地" is "君自忆山阴" while the next one of "明月临沧海" is "闲云恋故山". We see that "君自忆山阴" and "闲云恋故山" also share the similar contents and emotions. Thus "君自忆山阴" can be the reference of "闲云恋故山". We refer to "ShiXueHanYing", the taxonomy mentioned in the previous sections. Given a first sentence S1, we extract several keywords, say keywords A, B and C in the taxonomy, and generate a keyword set consisting of both the extracted keywords from the first sentence and the related keywords in the same directory in the taxonomy, say {A, A_1, A_2, B, B_1, … , C, C_1, C_2, C_3, … .}, which represents the meaning of the first sentence. Then if another sentence S2 has keywords A1, B1 and C3, S1 and S2 are similar ones. The next sentences of S1 and S2 are references for each other. In this way, a numerous number of next sentences with similar keyword sets of the first sentences can be selected from about 30,000 famous classical Chinese poems, such as <Tang Poems>.

With this method, we obtained a testing set of 2050 poems with 6150 sentences pairs. Each poem has 3 sentence pairs to be evaluated, the first-second, second-third, and third-fourth pairs. Results are shown in the following sections.

BLEU vs. Human Evaluation

To the best of our knowledge, there is no previous work about poetry evaluation with the BLEU metric. In order to prove its soundness, we need to conduct human evaluations on poems and compare them with the BLEU results.

Score/ Criteria	Fluency	Rhyme	Coherence	Meaning
3	Totally fluent	Meet all	Good	Good
2	Fluent	Meet most	More or less	Has some meaning
1	Not fluent	Not meet	Not good	not meaningful

Table 2 The criteria of sentence evaluation

A clear criterion is necessary for human evaluation. In the books on poem generation, such as (Wang, 2002), the criteria for evaluating poems are discussed. Based on these books, we formed four simple criteria: "Fluency", "Rhyme", "Coherence", and "Meaning", as shown in Table 2, which the human annotators can easily follow. The annotator only needs to assign different scores according to the four criteria. After that, the score of each sentence is calculated by summing up the four separate scores. According to the obtained sentence-score, a sentence's quality is classified into three grades: A, B and C, with A standing for the best quality, B the middle and C the worst. Grades A, B and C are given if the sentence-score is at least 10, in range [7,9], and in range [4,6], respectively. Then the system-scores of different systems can be calculated by summing up all grades by letting A=3, B=2 and C=1. For example, the "Basic SMT" has 11 sentences with Grade A, 26 sentences with Grade B and 23 sentences with Grade C, thus its system-score is 3*11+2*26+1*23 = 108.

The human evaluation are done over 20 5-char quatrains and 20 7-char quatrains with 120 sentence pairs because its time cost is much larger than automatic evaluation. We asked two human judges who are experts in traditional Chinese poems to give scores to the 20 5-char quatrains and 20 7-char quatrains. About 80% of the scores of the testing set are the same as the two judges. We got an average score for the non-consistent ones.

Systems	7 char quatrain				5 char quatrain			
	A	B	C	Score	A	B	C	Score
Sys 1/50	5	21	34	91	2	16	42	80
Sys 1/10	7	25	28	99	9	25	26	103
Sys Basic (whole data)	11	26	23	108	11	24	25	106
+Sys Position	14	26	20	114	9	31	20	109
+Sys Coherence	17	26	17	120	12	30	18	114

Table 3 Human evaluation of different systems

Systems	7 char quatrain	5 char quatrain
	Average BLEU	Average BLEU
Sys 1/50	0.322	0.109
Sys 1/10	0.352	0.130
Sys Basic(whole data)	0.368	0.134
+Sys Position	0.371	0.139
+Sys Coherence	0.380	0.141

Table 4 BLEU evaluation of different systems

We evaluated the five systems, three systems have the same model but different training data and two advanced systems are added new features, to see whether BLEU is well-correlated with human judgments. BLEU evaluation is conducted on the same data set of 40 quatrains. The high correlation coefficients of 0.95 in the 7-char quatrain and 0.97 for the 5-char quatrain indicate that BLEU can track human judgment in poetry generation as well as in machine translation. To illustrate our result, we show a linear regression of the human evaluation scores as a function of the BLEU score over 7-char and 5-char quatrains; see Figures 1 and 2.

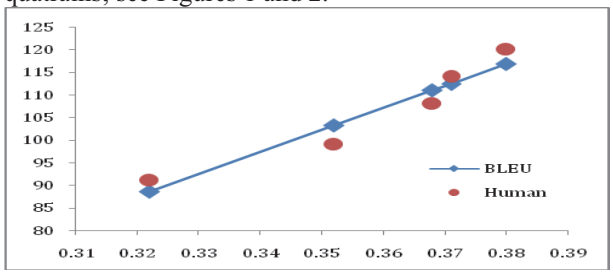

Figure 1 BLEU predicts human judgment in 7 char quatrains

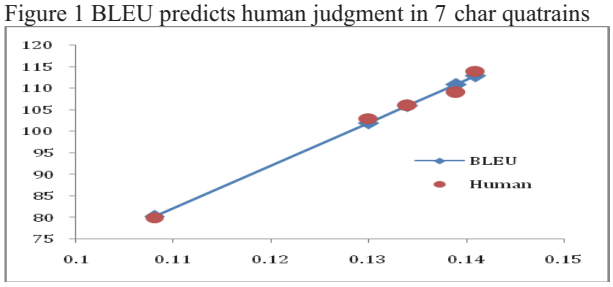

Figure 2 BLEU predicts human judgment in 5 char quatrain

BLEU Results of Our Systems

BLEU is proved to be reasonable for poem evaluation. In addition for the evaluation, another usage of BLEU is to consider it as the criteria for MERT training for weights of different features in the SMT approach and weights of different translation models and language models in our advanced approach in Section 5 and Section 6. We use MERT training to train the weights in formulas (1) (2) and (4). Our training data set for the BLEU metric contains 2000 classical Chinese poems different from the testing set chosen from <Tang Poems>, half of which are 7-char quatrains and other half are 5-char.

Our testing data set for the BLEU metric contains 2050 classical Chinese quatrain poems (with 6150 sentence pairs) that are randomly chosen from the most famous Chinese

poetry anthology <Tang Poems> except the ones in the training set. Half of them are 5-char quatrains and the other half are 7-char quatrains. Each poem has 3 sentence pairs to be tested and evaluated, namely the first-second, second-third, and third-fourth pairs. From Tables 5 and 6, we can see that adding the position-sensitive translation models increases the BLEU score from 0.137 to 0.139 for the 5-char quatrain and from 0.368 to 0.371 for the 7-char quatrain, respectively. With further investigation, we find that the poems generated by adding the position-sensitive translation models have better semantic structure than the basic SMT approach.

Moreover, we added both position-sensitive models and the coherence model to the basic system. We observe that the BLEU score of a 7-char quatrain increases from 0.371 to 0.380 while that of a 5-char quatrain increases from 0.139 to 0.141, which shows that the coherence model works well.

Human evaluations are done over a small set of 40 samples from testing data set and the scores are increasing with the advanced models added. See Table 7.

Systems	1 2	2 3	3 4	Average
Sys Basic	0.360	0.316	0.429	0.368
+Sys Position	0.346	0.309	0.459	0.371
+Sys Coherence	0.346	0.323	0.470	0.380

Table 5 The BLEU scores of 7 char quatrain

Cases	1 2	2 3	3 4	Average
Sys Basic	0.137	0.107	0.158	0.134
+Sys Position	0.146	0.106	0.165	0.139
+Sys Coherence	0.146	0.109	0.167	0.141

Table 6 The BLEU scores of 5 char quatrain

Systems	7 char quatrain				5 char quatrain			
	A	B	C	Score	A	B	C	Score
Sys Basic	11	26	23	108	11	24	25	106
+Sys Position	14	26	20	114	9	31	20	109
+Sys Coherence	17	26	17	120	12	30	18	114

Table 7 Human evaluation over 40 sample poems

As an illustration of the quality of the generated quatrains, a 5-char quatrain generated by the advanced SMT approach accepting three keywords “春” (spring), “琵琶” (lute), “醉” (drunk) is shown below.

双眸剪秋水，(- + - - +)	With two eyes, watching the autumn river
一手弹春风。(- + - - -)	With one hand, playing the springtime lute
歌尽琵琶怨，(- + - - +)	With song done and lute exhausted
醉来入梦中。(- - + + -)	I drunkenly fall into a dream

Conclusion and Future Work

Poetry generation is a difficult problem in the field of text generation. We propose a novel approach combining a statistical machine translation model with an ancient poetic phrase taxonomy to generate the Chinese quatrain. According to the evaluation of the generated poems, our method achieves impressive results. As the SMT model we applied is data driven, our method is language independent which means it can be extended to poetry-generation in other languages if the poetry data is readily available at enough scale. The application system can also be improved to be interactive with human in the poem-generating process. We have many issues worth further investigation. Both our template based generation of the first sentence in Section 4 and the generalization of references in Section 7.2 are based on the taxonomy <ShiXueHanYing>. As a poem dictionary collected by human experts in Qing Dynasty, its coverage is very limited and many words are not used now. As a future exploration, we want to reinforce this taxonomy by automatic mining synonyms and associated words, and by learning new concepts from large corpus to improve the current keyword set of <ShiXueHanYing>. We are also interested in extending our methods to English poem generalization.

References

Engelmore, R., and Morgan, A. eds. 1986. *Blackboard Systems.* Reading, Mass.: Addison Wesley.

Long Jiang, Ming Zhou. Generating Chinese Couplets using a Statistical MT Approach. In: The 22nd International Conference on Computational Linguistics, Manchester, England, August 2008.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase Based Translation. HLT/NAACL 2003.

Wenwei Liu. ShiXueHanYing. (《诗学含英》), 1735.

Mani, I., E. Bloedorn, and B. Gates. Using Cohesion and Coherence Models for Text Summarization. In Working Notes of the AAAI’98 Spring Symposium on Intelligent Text Summarization, 69 76. Stanford, CA, 1998.

Franz Josef Och, Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In "Proc. of the 40th Annual Meeting of the Association for Computational Linguistics" pp. 295 302, Philadelphia, PA, July 2002.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In "ACL 2003: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics", Japan, Sapporo, July 2003.

H. G. Oliveira, Automatic Generation of Poetry: an Overview. In: “1st Seminar of Art, Music, Creativity and Artificial Intelligence”, 2009.

K. Papineni, S. Roukos, T. Ward and W. J. Zhu. BLEU: a Method for automatic evaluation of machine translation. In Proc. of the 40th Meeting of the Association for Computational Linguistics, 2002.

Li Wang. A Summary of Rhyming Constraints of Chinese Poems (《诗词格律概要》). Beijing Press. August, 2002.