漢詩からの季節推定について

On the seasonal prediction from Chinese poetry

石田 勝則[*1]　角 康之[*1]　西田 豊明[*1]
Katsunori Ishida　Yasuyuki Sumi　Toyoaki Nishida

[*1] 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

Summary

To interpret the poetry, the estimation of the season from the works is a fundamental task. “Haiku” which is the Japanese poetry has “kigo” that is the seasonal term included in it. Because “kigo” works on a 1 to 1 match of the season, it is easy to estimate the season from Haiku. On the other hand, seasonal identification from a Chinese poem is not easy since there is no “kigo” rule in Chinese poetry. Fortunately, there is a “shigosyuu” which is the diction of idiomatic phrases called “shigo” of Chinese poetry. But in a “shigosyuu”, the “shigo” has sometimes been registered to some poetic dictions of seasons, or more than once has been registered in the different titles of the same season. Therefore, the direct reference of “shigo” in the “shigosyuu” is not efficient to estimate the season from the Chinese poetry. This paper presents a method for seasonal estimation from the Chinese poetry by using the extended poetic diction which is installed with the seasonal points by measuring the seasonal contribution ratio of “shigo” and added by the unregistered “shigo” with proper classification. This paper discusses the result and evaluation of the seasonal estimation of the Chinese poetry by using the extended poetic diction.

要約

漢詩を解釈する場合、その漢詩から季節を推定することは基本的な作業である。同じ詩文である俳句には季語があり、俳句歳時記に収録された季語と作品中の季語を一対一に対応付けることにより、容易に作品の季節を推定することができる。一方、漢詩には季語という習慣がなく、俳句のように作品の季節を特定することは簡単ではない。幸い漢詩には、慣用的な漢詩熟語を季節別に編纂した詩語集があり、作品の季節推定のための情報源として活用することが期待できる。しかし、詩語集には、ある詩語が複数の季節に同時に登録されており、また、同じ季節の異なる詩題に重複して登録されているので、詩語集を直接参照して季節推定を行うことができない。本論文は、詩語集に収録されている詩語の登録情報を基に詩語の季節寄与属性を数量化し、古典漢詩作品から抽出した詩語で補強した拡張詩語辞書を用いた漢詩の季節推定法について論じる。

連絡先：京都大学大学院情報学研究科知能情報専攻
京都市左京区吉田本町,Tel:075-753-5387
E-mail:k-ishida@ii.its.i.kyoto-u.ac.jp

1. はじめに

詩文を解釈する場合に、作品の季節を推定することは基本的な作業である。俳句には“季語”を作品中に詠い込むことにより、短い表現ながら季節を読み取ることができるように工夫されている。一方、漢詩には季語という習慣がなく、作品の解釈における季節推定は俳句ほど簡単ではない。コンピュータに漢詩作品の季節を推定させるためには、漢詩文に適した詩語を季節別や詩題別に編集した詩語辞書を活用することが考えられる。
しかし、詩語辞書に登録されている詩語と季節は、俳句歳時記のように1：1に対応していないので、辞書に存在する季節情報を何らかの方法で抽出して利用する必要がある。
本研究では、国内で最も広く利用されている[太刀掛 90]の詩語集を元に、2字詩語，3字詩語，4字詩語の詩語辞書を作成し、この詩語辞書の詩語に、詩語辞書の季節情報から抽出した季節ポイントを付与した季節ポイント付き詩語辞書を用いて季節推定を行っている。本論文では、この季節ポイント付き詩語辞書の構成方法、この詩語辞書を用いた漢詩七言絶句作品の季節推定実験と評価、さらに、この辞書に辞書中の登録詩語の偏り補正および新規詩語の補強を施した拡張詩語辞書を用いた漢詩七言絶句作品の季節推定実験と評価、漢詩の季節推定における拡張詩語辞書の有効性について論じる。

2. 季節推定の準備

2．1 季節の定義

中国では一年を二十四節気に区分する習慣があり、春は立春から立夏まで、夏は立夏から立秋まで、秋は立秋から立冬まで、冬は立冬から立春までを意味する。一般の詩語集の季節分類もこの基準に準拠しているので、本研究では、詩語集の季節区分に従って、旧暦の1月～3月を春、4月～6月を夏、7月～9月を秋、10月～12月を冬とし、季節にかかわらないものを雑に区分した。七言絶句は 4 句からなり、各句は 2 個の2文字詩語と 1 個の3文字詩語で構成され、計 7 文字の漢字が使われる。詩語が持つ季節属性を定量化する方法として、本研究では、各漢字に 100 ポイント、2文字詩語、3文字詩語に、それぞれ 200 ポイント、300 ポイントの季節ポイントを付与した。したがって漢字 28 文字で構成される七言絶句の漢詩は、28×100＝2,800 ポイントの季節ポイントを持つこととなる。この 2,800 ポイントの季節別分布状態をその作品の季節属性と定め、季節ポイントの最大値を示す季節区分をその作品の季節と定義した。

2．2 季節ポイント付き詩語辞書の構成

利用した[太刀掛 90]の詩語集は、2字詩語、3字詩語、4字詩語を春夏秋冬雑の 5 つの区分に大別し、季節にふさわしい詩題ごとに、約 21、000 語の詩語が収録されている。例えば、“細雨”という2字詩語は、春の部で 3 つの詩題（新春偶成、春日郊行、雨中送春）、夏の部で 1 つの詩題（梅雨書懐）、秋の部で 1 つの詩題（晩秋閑居）に登録されている。この詩語集に収録されている2字、3字、4字詩語の形態別総数、見出し語数、平均重複回数は表 1 の通りである。古典七言絶句約 770 首について、詩語集の登録詩語との一致度を調べてみると、必ずしも高くない 。季節推定の精度を上げるためには、さらに詩語を追加する必要がある。詩語を詩語辞書に追加するには、作品の季節を推定し、詩語の季節属性を決定する必要がある。また、作品の季節推定において作品中の詩語が詩語辞書にみつからない場合には、漢字の季節ポイントを使用することとし、詩語と同様に、詩語表の季節別出現頻度をもとに設定した。 表 2 に[太刀掛 90]の詩語辞書に用いられている漢字とその見出し語数、平均重複回数を示す。

表 1 登録詩語数･見出詩語数･重複回数

	登録語数	見出し語数	平均重複回数
2字詩語	9635	5563	1.73
3字詩語（押韻句）	8305	6796	1.22
3字詩語（転句）	2967	2518	1.18
4字詩語	384	373	1.03
計	21291	15250	1.40

表 2 詩語辞書の出現漢字数と見出し語数及び平均重複回数

	出現漢字数	見出し漢字数	平均重複回数
詩語辞書漢字	54622	2554	21.4

2．3 季節ポイントの付与

i 番目の詩語 S_i の季節ポイント P_i は $P_i = (P_{i1}, P_{i2}, P_{i3}, P_{i4}, P_{i5})$ で表したベクトル値であり、P_{ij} は当該詩語の季節区分別出現頻度をもとに次式で与えた。

$$P_{ij} = \frac{n_{ij}}{Ni} \times \frac{T_{ij}}{T} \times 100 \quad where \ \ (j = 1 \sim 5)$$

尚、j は春夏秋冬雑の季節区分を、N_i は S_i の総出現回数を、n_{ij} は季節区分毎の S_i の出現回数を示す。また、T は S_i が属する詩語区分に登録されている詩語の総数、T_{ij} は季節区分 j の詩語の総数である。表 3 に漢字･2字詩語･3字詩語の季節ポイントの計算例を示す。“雨”は夏をピークに、春＞秋＞冬の順にポイントが分散しているが、“細雨”は春をピークに夏＞秋＞冬の順にポイントが分散し、“白雨”は夏にポイントのピークが集中している。季節毎に降る“雨”に応じて、季節ポイントがそれぞれの季節によく反映していることがわかる。

表 3 詩語辞書からの季節ポイント計算例 （ ）：出現回数

	春pt	夏pt	秋pt	冬pt	雑pt	計
雨(333)	30	42	16	5	7	100
細雨(5)	108	42	36	0	4	200
白雨(4)	0	200	0	0	0	200
雨如糸(4)	90	111	0	0	99	300
烟雨村(2)	0	300	0	0	0	300

3. 季節推定実験

3．1 初回実験結果と評価

漢詩の初回季節判定実験には、収録した七言絶句 772 首の中から、詩題に春夏秋冬の文字が含まれる作品を選択し、季節推定を行った。詩題に春夏秋冬の 4 文字が含まれる作品は古典七言絶句 772 首中に 106 首あり、この七言絶句 106 首に対し季節推定を行った。結果を表 4 に示す。

表 4 詩題季節作品初回評価結果

	作品(春)	作品(夏)	作品(秋)	作品(冬)	平均正解率
作品数	53	23	24	6	
正解率	81.1	82.6	54.2	66.7	74.5

初回実験の結果、たとえば、柳絮は柳の棉花のことで、春の風物であるが、詩語辞書には降雪のことを表現する詩語として、冬の部に収録されており、作詩をする場合の手引書として編纂された詩語集には、詩語登録にかたよりがあることがわかった。詩語集の登録情報を季節推定のデータとして使用するには、詩語登録のかたよりを補正する必要がある。そこで、初回設定季節ポイントを正しい季節ポイントの一次近似と位置づけ、古典作品 772 首の季節推定を行って作品の推定季節と乖離した季節ポイントをもつ詩語を抽出した。

3. 2 拡張詩語辞書による実験結果と評価

漢詩鑑賞のために登録した 772 首の中国古典七言絶句について、その作品中に使われている2字詩語および3字詩語と詩語集の詩語とを照会し、照会できた詩語および照会できない新規登録候補詩語に分類した。結果は表 5 の通りである。新規詩語登録は、登録候補から全作品に2回以上出現した詩語について季節を推定し、詩語辞書に追加登録した。また、照合できた詩語の季節ポイントと作品の推定季節ポイントとの一致度を調べ、乖離が大きい詩語に対し、作品の季節ポイントを基準に詩語の追加登録を行い、詩語の季節ポイントの補正を行った。

表 5 古典漢詩作品の詩語抽出と新規登録詩語候補
(七言絶句 772 首)

	二字詩語	三字詩語	転句三字詩語	合計
全抽出数	6176	2316	772	9264
詩語見出語数	4526	2245	752	7523
照会済み詩語	1115 (26%)	193 (7.3%)	67 (2.6%)	1375 (18.3%)
新規登録候補詩語	3411	2052	685	6098
内2回以上出現詩語数	387 (11.3%)	14 (0.6%)	135 (19.7%)	1375 (18.3%)

この拡張詩語辞書の有効性を評価するために、専門家が季節分類した中国古典七言絶句作品を用いて、季節推定を行った。実験結果は表 6 に示すように、良好な結果を得た。

表 6 [石川 01][黒川 88]の季分類作品の季節推定結果

	評価対象	不適合数	適合数	正解率%	(不適合作品番号)
春	74	1	73	98.6	284
夏	50	2	48	96.0	44,705
秋	49	1	48	98.0	574
冬	37	5	32	86.5	56,363,526,583,584
計	210	9	201	95.7	

中国古典の漢詩作品であり、これにより季節ポイントの偏りが完全に除去されたとはいえないが、772 首の作品に使用されている詩語を追加し、詩語登録の偏りを補正した拡張詩語辞書が、漢詩作品の季節推定に有効であることが確認できた。

4. 今後の課題

古典作品に使用される詩語がその作品の季節に相応しいものであれば、全ての季節作品に使われる詩語には季節感がなく、特定の季節作品にしか使われない詩語はその季節を代表していると考えることができる。しかし、この一般的な規則について、次に示すいくつかの例外が発見された。これらの例外については、詩語漢字の季節ポイントを単純に積算するだけでは、季節推定を誤ることとなる。たとえば

①人名、地名、建築物などの固有名詞に季節感の強い漢字が含まれるもの (たとえば楊万里、寒山寺、秋風亭など)

②比喩や比較の対象として異なる季節感をもつ詩語が使われる場合 (たとえば如雪で色の白さを表現し、紅於秋で紅を秋と比較するなど)

③否定や疑問表現で異なる季節の詩語が用いられる場合 (たとえば未見雪で未だ雪を見ずとし、疑是霜で霜のような白さを表現するなど)

などである。①については、まだ固有名詞を自動判別する機能がないので、今回は抽出された新規詩語中の固有名詞を人力で抽出し、雑区分に登録することにより、誤判断を回避している。②、③の場合は、漢字の季節属性がわかったとしても、さらに文脈を理解したうえで、季節ポイントを適用する必要がある。今回は推定した作品の季節属性と、使用された詩語の季節属性の乖離による、詩語集の詩語登録の偏りを自動修正することにより、詩語に設定した季節ポイントの補正を行ったが、②、③の誤りを回避する方法として、同一作品に使用されている詩語の季節ポイントの共起解析により、共起しない詩語について季節判定詩語から除外する方法や、例外詩語を偏りのある詩語ととらえ、作品の当該詩語の使用事例を参考に季節ポイントを付与し、拡張詩語辞書に登録する方法等、いくつかの例外処理拡張詩語辞書に反映し、例外による誤推定率を低減することが今後の課題である。

5. おわりに

作品の季節推定精度を高めるために、例外処理を取り込み、より多くの作品の季節推定実験による詩語季節ポイントの学習を重ねて、季節推定に適した拡張詩語辞書の完成を目指したい。

6. 参考文献

[太刀掛 90] 太刀掛 重雄: だれにでもできる漢詩の作り方, 呂山詩書刊行会 ,(1990)

[石川 01] 石川忠久 漢詩を読む 春の詩100選、夏の詩 100選、秋の詩100選、冬の詩100選 NHK 出版(2000)

[黒川 88] 黒川洋一 中国文学歳時記 春・夏・秋・冬 同朋社 (1988)

[石田 06] 石田 勝則: 漢詩添削サービスにおける詩的表現の評価方法について、人口知能学会全国大会予稿集, (2006)

[石田 08] 石田 勝則: 詩語分類表の統計情報に基づく漢詩の季節クラスタリング 人工知能学会全国大会予稿集, (2008)