3D1-2

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について

On the seasons clustering of Chinese poetry based on the stochastic information of “Shigo Shyuu”

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

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

Haiku poetry has “kigo” which must be included in its phrase and shows one of the four seasons that the haiku poem involves. But there is no rule like “kigo” in Chinese poetry. So we must understand the seasons of the Chinese poem by the Chinese words used in its lines. This document discusses on the seasons clustering of Chinese poetry based on the stochastic information within Chinese poetic words dictionary, what is called , “Shigo Shyuu”.

1. はじめに

俳句には季語があり、俳句に含まれる季語を抽出することにより作品の季節を言い当てることができる。一方漢詩には季語の規則がないため,文中に用いられている詩語から作品の季節を総合的に判断しなければならない。幸いなことに、漢詩には漢詩愛好家のために、推奨される詩語を集めた詩語集がある。代表的な太刀掛重男[太刀掛 99]や田川瑞穂[田川 00]の詩語集は、詩語を季節別、詩題別に分類編集している。本研究では詩語集に収録されている、詩語の季節別、詩題別分類情報を用いて、七言絶句作品の季節を推定する方法について論じる。

2. 本研究の目的

芸術作品を分類するためには、分類目的に即した特性と基準を定め、分類する作品からその特性を定量的に抽出し評価する必要がある。芸術作品を分類する研究には、例えば楽曲のクラスタリングの例[大久保 07]がある。本研究の目的は、漢詩作品に使用される詩語のもつ季節特性を詩語集の季節分類情報から抽出し、その基準値をもとに、漢詩を自動的に季節クラスタリングできるかどうかを明らかにすることである。

3. 季節感の定義

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と考えられている。ここ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区分とした。七言絶句は4句からなり、各句は2個の2文字熟語と1個の3文字熟語で構成され、計 7 文字の漢字が使われる。また、七言絶句は起承転結の4句構成である。

今回の実験では、漢字に100ﾎﾟｲﾝﾄ、2文字詩語、3文字詩語には、それぞれ200ﾎﾟｲﾝﾄ、300ﾎﾟｲﾝﾄ、また、漢字 28 文字で構成される七言絶句には、2800 ﾎﾟｲﾝﾄの季節ﾎﾟｲﾝﾄを付与し、付与した季節ﾎﾟｲﾝﾄの季節別分布状態を各々の詩語または作品がもつ季節感と定義した。

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

4. 詩語表からの詩語の季節感抽出

季節推定の基準となる詩語の季節ﾎﾟｲﾝﾄの設定は、国内で最も広く利用されている太刀掛重男の詩語集[太刀掛 99]を用いた。この詩語集では2字詩語、3字詩語、4字詩語を春夏秋冬雑の5区分に大別し、季節にふさわしい詩題ごとに、約 21、000個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回（新春偶成、春日郊行、雨中送春）、夏の部で1回（梅雨書懐）、秋の部で1回（晩秋閑居）登録されている。

詩語のグループ別登録語数、見出し語数、平均重複回数は表 1の通りである。グループ別に文字数が増加するにしたがって、平均重複回数は減少している。従って、文字数の多い詩語ほど基準季節ﾎﾟｲﾝﾄの高い信頼性が求められることがわかる。

表 1　登録詩語数と見出し語数及び平均重複回数

	登録語数	見出し語数	平均重複回数
2字詩語	9,635	5,563	1.73
3字詩語 (押韻句)	8,305	6,796	1.22
3字詩語 (転句)	2,967	2,518	1.18
4字詩語	384	373	1.03
計	21,291	15,250	1.40

また、古典七言絶句約 700 首について、登録詩語の使用状況を調べてみた。ﾋｯﾄ率は表 2に示すように必ずしも高くない。従って、使用される詩語が基準詩語に見つからない場合には、使用漢字の基準季節ﾎﾟｲﾝﾄを、判定基準に加えることとした。

漢字の基準季節ﾎﾟｲﾝﾄは、全登録詩語 21,291 件を漢字に分解し、見出し語漢字の季節別出現頻度を基に算出した。

表 2　古典七言絶句の登録詩語ﾋｯﾄ率

	詩語表詩語登録数	七言絶句熟語見出数	七言絶句熟語差分数	七言絶句熟語重複数	ヒット率(%)
2字詩語	9,635	8,379	7,087	1,292	15.4
3字詩語 (押韻句)	8,305	2,047	1,900	147	7.2

漢字には季節感に無関係なものも多く、品詞区分によっても季節ﾎﾟｲﾝﾄの特性が異なると思われる。従って、漢字の基準季節ﾎﾟｲﾝﾄの設定には、統計的評価が必要である。詩語表から抽出した漢字とその見出し語数、平均重複回数を表 3に示す。

表 3　抽出漢字文字数と見出し語数及び平均重複回数

	抽出文字数	見出し語数	平均重複回数
漢字	54,622	2,554	21.4

詩語または漢字の季節ﾎﾟｲﾝﾄは、各熟語グループの季節別母集団の大きさによる出現頻度への影響を補正し、次式により計算している。

$$P_{ij} = \frac{N_{ij}}{N_i} \times \frac{T_{ij}}{T} \times 100 \quad \text{where}\,(j = 1 \sim 5) \qquad (1)$$

j は春夏秋冬雑の季節区分を、Ni は i グループの詩語または漢字 Pi の総出現回数、Nij は j 季節区分内における Pi の出現回数を示す。また、T は Pi が属する詩語の総数であり、Tij はその i グループ j 季節区分内の詩語ないし漢字の総数である。

漢字と 2 字詩語の季節ﾎﾟｲﾝﾄの計算例を表 4に示す。“雨”は夏を中心に、春＞秋＞冬の順にﾎﾟｲﾝﾄが分散している。“細雨”は春をピークに夏＞秋＞冬と分散し、“白雨”は夏にピークが集中している。“細雨”は春“、白雨”は夏というように、漢字“雨”が2字熟語を形成することにより、“細”、“白”の季節ﾎﾟｲﾝﾄと異なる季節ﾎﾟｲﾝﾄに変化していることが分かる。

表 4　季節ﾎﾟｲﾝﾄ計算例　（ ）:出現回数

	春	夏	秋	冬	雑	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
細(44)	28	25	16	24	7	100
白(183)	16	23	26	13	22	100

5. 基準季節ﾎﾟｲﾝﾄの検証

季節感は1年の季節の移り変わりを表わす連続的な値であり春夏秋冬も一定の幅をもつ。しかし、詩語の中には季節を明確に表すものがあり、基準季節ﾎﾟｲﾝﾄの妥当性を検証する一助となる。季節区分が明確な漢字と熟語の季節ﾎﾟｲﾝﾄの例を表 5に示す。尚、各項目の() の数字はそれぞれの出現頻度を示している。

表 5　基準季節ﾎﾟｲﾝﾄ例　（ ）:出現回数

	春	夏	秋	冬	雑	計
春(346)	73	2	0	13	12	100
夏(37)	0	93	0	0	7	100
秋(355)	0	10	74	1	15	100
三月(4)	100	0	0	0	0	200
十月(4)	0	300	0	91	9	200

6. 七言絶句の季節推定実験

以上により設定した基準季節ﾎﾟｲﾝﾄﾃﾞｰﾀを用いて、七言絶句の季節推定を行った。推定精度を評価するために、古典七言絶句約700首の中から詩題に四季を表す春夏秋冬の文字が含まれる作品を選別し、その季節判定結果の正解率を調べた。

漢字の品詞別季節ﾎﾟｲﾝﾄ特性についても比較評価した。七言絶句の4句毎の季節ﾎﾟｲﾝﾄを求め、その合計値をその作品の総合季節ﾎﾟｲﾝﾄとした。詩題に春夏秋冬を含む古典七言絶句106首の、季節クラスタリング結果は表 6の通りである。
ケース1は全ての品詞の基準漢字、ケース2は名詞及び形容詞の基準漢字、ケース3は名詞のみの基準漢字を評価基準に用いた実験結果である。図 1は正解率 93.1%の作品（春）/ケース1の季節推定について、53 件の作品の総合季節ﾎﾟｲﾝﾄを2次元の散布図に表したものである。

表 6　詩題季節作品季節推定結果　（）は誤り件数

	評価件数（作品数）	ケース1 正解率	ケース2 正解率	ケース3 正解率
作品（春）	53 首	93.1%(4)	93.1%(4)	91.4%(5)
作品（夏）	23 首	91.3%(2)	73.9%(6)	73.9%(6)
作品（秋）	24 首	84.0%(4)	100.0%(0)	96.0%(1)
作品（冬）	6 首	83.3%(1)	83.3%(1)	83.3%(1)
平均正解率%	(106 首)	89.6%(11)	89.6%(11)	87.7%(13)

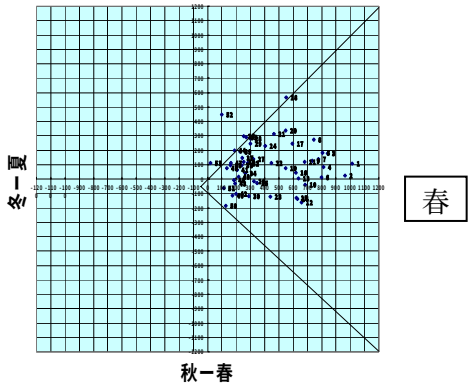

図 1　ケース 1 作品(春)の総合季節ﾎﾟｲﾝﾄ散布図

7. 実験の評価

実験結果では、ケース1, 2, 3 いずれの場合も平均 87%以上の精度で季節を特定できた。漢字は名詞、形容詞、動詞について、全ての品詞を基準に取り入れるほうが平均正解率を高めることが分かった。正解から外れたケースについて精査すると、外れて正しいもの(“秋思”は夏の詩)、まったく基準詩語にﾋｯﾄしないもの（全て基準漢字で評価）等が含まれている。

漢詩作品を適切な季節区分に分類することにより、未収録の詩語を適切な詩語分類区分に収録することにより、漢詩作詞支援システムの詩語検索サービスを充実させるとともに、さらに精度の高い漢詩の季節クラスタリングが可能になると考えている。

尚、下記の公開 WEB ｻｲﾄにアクセスし、漢詩添削システムを用いて、誰でも、七言絶句作品の季節感評価ができる。
http://www.kannshi.net/kannshi/index.html

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)
[川田　00]　川田 瑞穂: 詩語集成（改訂版）松雲堂書店,(2000)
[大久保　07]形式概念に基づく Top-N 楽曲ｸﾗｽﾀﾘﾝｸﾞに関する一考察　人工知能学会第 21 回全国大会, (2007)
[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)