季語を用いた作品季節推定システムの評価

Performance Evaluation of Season Estimation System
Using Seasonal Haiku Indicators

伝住颯夏　　ジェプカ ラファウ　　荒木健治

Satsuna Denzumi　　Rafal Rzepka　　Kenji Araki

北海道大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

１．まえがき

通信分野をはじめとした技術の発達により，情報の蓄積・流通に必要なコストが過去と比較して大きく低下し，古今東西様々な作品を得ることが容易となった．そのため，現代では利用者の希望・嗜好に沿った情報をスムーズに提供する推薦システムの重要度が増している[1]．その一つである書籍推薦システムの例として，Amazon.com から書籍情報の特徴を取得し，それに基づいた内容ベースの推薦を行う Libra[2]がある．Libra が取得する書籍情報の中には作品のあらすじも含まれるが，あらすじの記述が非常に簡潔である，作品の内容と直接関係のない情報が含まれる，あらすじ自体が存在しないなど，作品の内容を反映した推薦を行うには十分な情報量ではないという問題点がある．

その問題を解決し，作品の中身をより反映した推薦を実現するため，作品を構成する 5W1H の中から【Who】【Where】【When】の 3 要素に着目した内容依存推薦システム[3]を提案した(詳細は 2 章で解説する)．このシステムは作品全体の単語を解析し，3 要素それぞれの頻出上位単語をその作品のキーワードとして登録し，それを基準として推薦を行う．しかし取得するキーワードの中で，作品の時間を表す【When】の要素は作品内の季節をはじめとする時間情報を反映することができていないケースが多い．このことから，高頻出の時間に関する単語を単純に取得するだけでは作品の時間情報を十分に表せず，推薦に反映させることは難しいとわかる．

ここでは，作品中で最も描写が丁寧である，あるいは最も描写量が多い季節がその作品を代表する季節であるという仮定のもと，俳句を詠む際に用いられる季語を本文全体から抽出し，その出現頻度に着目した作品季節推定システムの構築と有効性の確認を行う．提案手法は，数百年の歴史を通して多くの人に育まれ，洗練されてきた俳句を構成する季語に着目しそれらを利用することで，直接的な時間に関する単語の描写がなくとも作品の季節推定を効率よくかつ正確に実現すると考える．

２．関連研究

作品中の時間情報を推定する研究の例として，奥村・瀧本による物語文章の時系列推定手法が挙げられる[5][6]．ただし，この研究は物語文章の自動要約生成を目的としたものであり，要約と推薦では必要とする時間情報の種類が若干異なる．要約を行う際は作品中の時系列が非常に重要であるが，推薦を行う場合は作品の舞台となる季節が分かる程度で十分であり，詳細な時系列情報は必ずしも必要ではないと考えられる．また，概念ベースや関連度計算方式を用いることで未知語を処理する機能を備えているとはいえ，予め人手によって作成された時間に関する単語のデータベースには時間帯単位，年月日単位，季節単位，人生単位の四種類計 511 語しか登録されておらず，著者や時代により様々な文体・形式で記述される作品に対応するにはデータ不足ではないかという不安も挙げられる．

1 章で取り上げた内容依存推薦システムは，ある入力作品に対して最も似た要素を持つ作品を推薦するシステムである．システムは大きく分けてキーワード取得処理とキーワード比較処理の二つに分かれ，前半の取得処理では日本語形態素解析システムである JUMAN[4]を用いて本文全体を形態素解析し，JUMAN 辞書から該当するカテゴリ情報を持つ単語を【Who】【Where】【When】の 3 要素に分類，多くの作品に共通して出現する語をストップワードとして設定した上で各要素の頻出上位 5 単語をその作品のキーワードとして取得する．そして後半の比較処理ではキーワードとその頻出頻度を他作品と比較して推薦を行う．実際に「蟹工船」のキーワード取得例を表 1 に，「蟹工船」を入力作品とした時の比較例を表 2 に示す．下線がついた語が入力作品と一致するキーワードであり，この場合は「無人島に生きる十六人」が推薦作品として選択される．

表 1 キーワード取得例

蟹工船		
【Who】	【Where】	【When】
漁夫	海	度
船	日本	瞬間
監督	工場	初め
夫	函館	か月
船長	会社	毎日

表 2 キーワード比較例

無人島に生きる十六人			党生活者		
【Who】	【Where】	【When】	【Who】	【Where】	【When】
船	島	時分	同志	工場	毎日
魚	海	毎日	我々	会社	最近
あざらし	日本	午後	女工	警察	その後
伝馬船	国	冬	母親	仲間	昼
長	矢倉	晩	母	組織	時期

ただし表 1，表 2 から分かるように時間を表す【When】のキーワードは「毎日」や「その後」，「瞬間」など作品中の時間情報の推定に役立つ語が少なく，更に短編作品ではキーワード自体がほぼ存在しないケースも確認されている．また，春夏秋冬などの季節を表すキーワードが取得されても，作品中の季節の一部しか取得できていない場合，あるいは作品中の季節と完全に異なる場合があり，キーワードだけで時間情報を推定することは難しい．表 2 でも「無人島に生きる十六人」のキーワードに「冬」という単語があるが，実際の作品中の季節は 1 年を通した四季であり，このキーワードだけでは本当に作品の代表となる季節が冬なのか，判断基準が乏しく推定結果に不安が残る．

3．季語

季語とは「俳句で，季節と結びついて，その季節を表すと定められている語」(デジタル大辞泉)である．基本的に季語は春夏秋冬と新年の五季に分類されており，季語辞典によって掲載される季語の種類，数は異なる場合がある．

まず，現代俳句協会のデータベース[7]から，春夏秋冬それぞれの季節に対応する季語リストを作成する．新年の季語については，時期として考えると冬の季語であるものの，縁起を重視する観点から春の季語としても考えられる語であるため，リストには入れないものとして考える．

データベースは季語 1 語に対してよみがな，旧仮名，傍題・別名の 3 種の情報が付随している．各季節の季語の種類と付随した情報を含めた語数，そして実例を表 3 に示す．四季合計で 2,532 種 10,509 語が登録されている．ただし，この中から季節内で重複して存在している語句，「たか」や「のみ」，「うちは」など本来の意味とは誤った形で抽出されることが多いと判断した語句についてはストップワードとしてリスト化する際に削除する．

推定システムはリストに記された語句を本文中から直接検索して抽出し，季語の出現頻度が最も高かった季節を作品の主な舞台であるとして出力する．日本語形態素解析システムである MeCab[8]や JUMAN などを用いずに直接検索するのは，表 4 に示すように決して少なくない数の季語が，MaCab や JUMAN で形態素解析を行うと必要以上に分解され，正確に抽出することが難しいと判断したためである．

表 3 季語の語数と例

	種類	語数	例
春	616	2,775	鶯・遠足・わかさぎ
夏	815	3,344	ビール・蜜豆・武者人形
秋	546	2,304	凶作・ねぶた・八朔
冬	555	2,086	火事・雪祭・今川焼

表 4 季語の形態素解析例

季語	解析結果
豆桜	豆/桜
端午の節句	端午/の/節句
虫時雨	虫/時雨
鬼は外	鬼/は/外

4．実験

実験では複数の作品について，人手により事前に作品を通して読んだ上で代表する季節を推定し，その結果と提案手法が推定した季節が一致するかということを用いて提案手法の性能を評価する．本実験では推定時に多くの文章を読む必要があり時間と労力を要することから，評価者は第一著者一人である．人手による季節判定は作品中で季節が明確に表記されていればそれに従う．季節が明確に表記されていない場合や，複数の季節から作品が構成されている場合は，四季の中から最も作品を代表する季節であると評価者が感じたものを選択する．中には四季の枠に収まらない無季の作品も存在するが，本実験ではそのようなものはないものとして考える．

対象となる作品は，第一著者が過去に全編読了して内容を把握しており，かつ青空文庫[9]に掲載されている中から選択した 50 作品とする．過去に読書経験がある作品を再読した上で推定を行うことで，読解ミスや理解不足による作品中の季節の誤推定を極力回避する．

実験結果を以下に示す．提案手法と人手による推定の一致率とその内訳については，表 5 に示す．提案手法による季節の推定は 50 作品中 39 作において人と推定結果が一致し，78.0%の一致率を得た．また，季節ごとの一致率を確認すると春の作品が 50.0%，夏の作品が 85.0%，秋の作品が 72.7%，冬の作品が 90.9%という結果となっており，最も一致率が高い冬の推定結果と，最も一致率が低い春の推定結果では 40 ポイント以上の差が出る結果となった．

表 5 提案手法の一致率

季節	一致率
春	50.0% (4/8)
夏	85.0% (17/20)
秋	72.7% (8/11)
冬	90.9% (10/11)
四季	78.0% (39/50)

5．考察

表 5 に示したように，提案手法は人による推定結果と高い一致率を示し，システムによる作品季節推定が有効であることが確認された．

また，内訳を確認すると夏と冬の季節で一致率が高い一方で，春と秋，特に春の季節で一致率が低いことがわかる．気温や天候などの点から，夏や冬は春や秋よりも強い特徴を持った季節であり，その強い特徴が作品中の描写にも反映されやすい可能性がある．また，春の誤推定結果は夏か冬のどちらかに限定されている．春はまだ雪も残る頃から始まり，徐々に気温が上がることで緑が芽吹き始める季節である．そのため「雪」に関する季語を持つ冬や，「緑」に関する季語を持つ夏の影響を強く受けることで，このような誤推定が増加したとも考えられる．しかし，これらの考えは母数となるデータが少ないため，あくまでも仮説に過ぎない．そのため，まずは更に多くの作品を対象に同様の実験を行い，それぞれの季節において十分な量のデータを得ることが必要である．

誤推定に関しては大きく分けて主に三つの原因が考えられる．第一の原因は人とシステムの季節感のずれである．旧暦に基づいて区分された季語を基準として推定を行うシステムに対し，現代を生きる我々は一般的に新暦に基づいて季節を認識するため，表 6 に示すように人間とシステムの間には約一カ月程度の季節認識のずれが発生する．そのため対象となる作品がそのずれた時期を主な舞台とする場合，両者が季節の推定を正しく行えたとしても結果的に異なる季節を推定し，誤りと判断される可能性がある．

表 6 季語と人の季節区分

	季語	人
春	2～4月	3～5月
夏	5～7月	6～8月
秋	8～10月	9～11月
冬	11～1月	12～2月

第二の原因として，季語リストの内容不足が挙げられる．仮名・片仮名・漢字をどのように用いて表記するか，またどのような同意語が存在するか，季語 1 つに対しても様々なパターンが考えられる．本実験で実際に確認された季語の取りこぼしの例を表 7 に示す．「いちょうの実」と「ぎんなん」は同じ意味である，「ハンケチ」と「ハンカチ」は時代による発音の差であり同じものであると理解することは人にとって決して難しくはないが，システムがそれを理解することは難しい．今回用いた季語は四季合計で 2,000 種 10,000 語を超えるが，「蛙」に対して「かえる」は登録されていても「カエル」は登録されていないなど，俳句の特徴からか特に片仮名表記はリストに記されていない傾向にある．著者ごと，時代ごと，作品ごとに異なる文体で表現される作品に対応するには，単純にこの季語リストを用いるだけではまだ不足であると思われる．

表 7 取りこぼし語とリスト掲載語

取りこぼし語	リスト掲載語
いちょうの実	ぎんなん
うろこ雲	鱗雲
ビイル	ビール
ハンケチ	ハンカチ

第三の原因としては，人名や地名などをはじめとする固有名詞の一部が，季語として誤った形で抽出される場合である．これは不一致例だけではなく一致例の中にも確認されている．結果だけを見ると正しい季節を推定しているものの，内容を確認すると季節とは一切関係ない固有名詞の一部を季語として誤って抽出することで，推定結果を正解へと導いているという問題が発生している．一例を挙げると，ある作品において登場人物である「熊城」から冬の季語である「熊」を 300 回近く抽出し，その結果作品の季節を冬と推定した．実際にその作品は冬の季節を舞台としているが，抽出結果から「熊」を除くと推定される季節は夏となり，抽出が正しく行われていたならば推定は失敗であったことがわかる．仮に季節推定に成功していたとしても，誤った形での季語抽出は季語を取りこぼすのと同等，あるいはそれ以上にシステムとして大きな問題になり得ると考えられる．この問題を解決するためには単純にストップワードを増やすのではなく，季語そのものの抽出をできる限り阻害することなく，その上で季節と関係のないフレーズからの抽出を回避する方法を取り入れる必要がある．誤った形での季語抽出回避は今後システムの推定精度を高めていく上で避けて通れない，最も重要な問題点であると考えられる．

6．まとめ

本稿では俳句において季節と結びつき，季節を表すと定められている季語を春夏秋冬の四季に分類した上で，作品の本文全体から季語を抽出し，それぞれの季節の出現頻度を比較することで作品の代表的な舞台となる季節を推定する手法について述べた．実験の結果，提案手法による推定は 7 割以上の作品において人手による推定結果と一致し，システムによる作品の季節推定は有効であることが確認された．しかし，季語が表す季節と現代の季節感とのずれ，季語抽出の取りこぼし，誤った形での季語抽出，それらを原因とした季節の誤推定，また実験を行う際の評価者数など解決すべき課題も挙げられる．

今後は第一に季語の抽出精度向上を目標とする．第二に「北海道は夏でも涼しい」「沖縄は冬でも暖かい」といった，地域や環境の差による季節の特徴を推定時に反映する手法の導入を試みる．第三に利用者を対象に調査を行い，複数の季節から構成される作品や，季節が存在しない無季の作品が対象となった時について，どのように推定結果へと反映させるのが適当であるか検討する．その上でこの推定システムを利用し，より利用者の希望に沿うことが可能である内容依存推薦システムの実現を目指したい．

参考文献

[1]神嶌敏弘．推薦システムのアルゴリズム(1)，人工知能学会誌，Vol. 22，No. 6，pp.826-837，2007.

[2]R. J. Mooney and L. Roy, Content-based book recommending using learning for text categorization, Proceeding of ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation, pp.195-204, 1999.

[3]伝住颯夏，ジェプカ ラファウ，荒木健治，人物・場所・時間情報を用いた内容依存推薦システムの評価，電気・情報関係学会北海道支部連合大会講演論文集，112，2014-10-26

[4]日本語形態素解析システム JUMAN, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[5]瀧本洋喜，奥村紀之，物語文章における時系列情報の抽出，信学技報，Vol.111，No.119，NLC2011-1，pp.1-6，2011.

[6]奥村紀之，瀧本洋喜，物語文章における時系列推定の拡張，情報科学技術フォーラム講演論文集，12(2)，301-306，2013-08-20

[7]現代俳句協会「現代俳句データベース」，http://www.haiku-data.jp/

[8]MeCab - Japanese morphological analyzer -, https://code.google.com/p/mecab/

[9]青空文庫，http://www.aozora.gr.jp/