E-006

句構造に着目した作家の文体の類似性
A Writing Style Similarity based on Phrase Structure

佐原 諒亮 † 金川 絵利子 † 岡留 剛 †
Ryosuke Sawara Eriko Kanagawa Takeshi Okadome

1. はじめに

作家の文や文章の特徴づけには，文の長さや句読点の間隔・用いる品詞などがよく用いられる．一方,「作家の文体」に焦点をあてた場合，文の句構造や係り受け構造が重要となる．本研究では，句構造に着目して解析を行ない，作家間の文構造の類似度に着目する．本研究では，木カーネル [1] を用いて文の類似性を議論する．文章中の文の順序も重要な特徴となり得るが，文構造に焦点をあてるために，文の順序は考慮せず bag of sentences に基づいて解析を行なう．

2. 木カーネル

木カーネルは，2 つの木構造データ間の共通している構造として，部分木を用いるカーネルであり，共通する部分木の個数を数えることで値が決定される．木カーネルは，木構造 T_1，T_2 に対して，以下の式で定義される．

$$K(T_1, T_2) = \langle \boldsymbol{\phi}(T_1), \boldsymbol{\phi}(T_2) \rangle = \sum_{S \in \tau} \phi_S(T_1)\phi_S(T_2),$$

ここで，S は部分木である．τ はすべての固有木の集合で，また，$\phi_S(T)$ は，木 T が S を部分木として含むときは 1，含まないときは 0 となる．これにより，T_1 と T_2 の共通の部分木の数え上げを実現している．

3. 評価

文の構文を表現する方法として，句構造文法や係り受け構造によるものがある．本研究は句構造を用いて作家の文を解析する．

3.1. 前処理

評価を行なうに当たり，用いる文書のクリーニングを行なう．すなわち，半角・全角スペースなどの空白文字を削除し，会話文は「」を削除した文を使用する．その他の記号に関しては原文通り使用した．

また，用いたテキストに対して形態素解析と句構造解析を行なうため，形態素解析器の Mecab と，長谷川 [2] の文法を拡張した文脈自由文法と，Berkeley Parser[3] を用いた．2 つの文の木カーネル値は葉である単語に大きく依存する．本研究では，構文構造の類似度に焦点を当て，用いられている単語の違いは極力排除したいため，各単語を品詞と形態素情報を表す記号に還元的に縮約したコーパスを用いる．例えば,「僕は学校まで走る」という文の還元的縮約は,「n は n まで v」となる．ただし，この場合の n と v は，それぞれ名詞と動詞を表している．

加えて，カーネル値は文の長さに依存してしまう傾向があり，文が長いほど大きな値を取りやすくなる．従って，本実験では，全作家の文の長さの最頻値（14 文字），中央値（28 文字），平均値（35 文字）を基準とし，それぞれでその値付近の文を収集したコーパスを用いる．カーネル値を算出する際は，最頻値を基準として生成したコーパスであれば，そのコーパス同士を用いて値を算出する．

3.2. 実験

本研究の実験では，まず芥川龍之介，太宰治，宮沢賢治，夏目漱石，新美南吉の 5 作家を対象としたため，各作家の全作品からランダムに 100 文抽出し，木カーネル値の総当たり平均を求める．結果はこれを 10 回行なったものの平均を用いている．また，木カーネルの計算は Moschitti のプログラム [4] を用いて計算した．木の深さに応じて重みづけをするパラメータ λ を設定する必要があるが，今回はデフォルトの 0.4 とした．この際，Subset Trees (SSTs) を用いる方法と，Sub Trees（STs）を用いる方法とがあり，今回は SSTs による解析について述べる．各作家ごとの木カーネル値の総当たり平均値を表にまとめた（表 1，表 2，表 3）．

4. 議論

表からわかったことに加え，木カーネル値を用いて行なった分析をもとに議論を行なう．

表を見てみると，芥川と夏目の値はどの文の長さの基準値の場合でも他の作家と比べて大きな値を取っていることがわかる．これは修飾語と被修飾語を用いた構造をお互い多くもちいるためであるが，それでも値に差が出ている．これは芥川に比べて夏目の方が被修飾語に動詞を用いることが多いためである．そもそも芥川は 1 文の中に用いる動詞が少なく，名詞句を長くする傾向があることもわかる．また，どの文の長さの基準値の場合でも宮沢と新美の値が比較的小さな値を取っていることがわかる．お互い同じ自動作家であるにもかかわらず，あまり似ていないのではないかという結果を得た．

芥川と夏目に加えて，どの作家においても基本となる構造として，修飾語と被修飾語を用いた構造を用いて文を書いていることもわかった．しかしながら，この修飾語と被修飾語の構造に特徴のある作家が太宰である．1 文ではあるが太宰の例を挙げると,「母は観念して、下の子を背負い、上の子の手を引き、古本屋に本を売りに出掛ける。」という具合に，1 文の中に被修飾語となるものが動詞である部分が多いことがわかる．図で表すと，太宰は以下のような図 1 の句構造で文を構成しやすい．ここで，vadt は「して」を表し，動詞である．

太宰は他作家と比較してこのような構造で 1 文を生成することが多い．一方で芥川や夏目などの文は，同様に芥川のみであるが 1 文例にとってみると,「おれの

†関西学院大学大学院理工学研究科

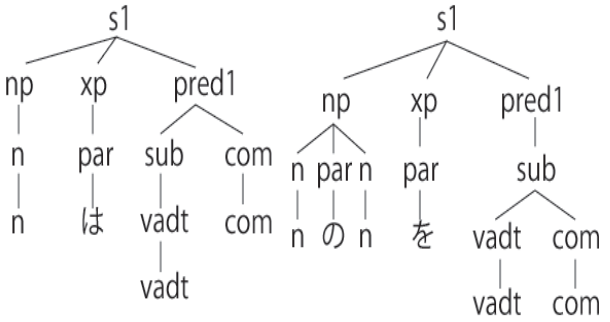

図 1: 太宰の用いやすい句構造

家の二階の窓は、丁度向うの家の二階の窓と向ひ合ふやうになつてゐる。」というように，被修飾語となるものが名詞である場合が多いことがわかる．さらに，芥川は修飾するときに用いる助詞が「の」である場合が比較的多いことがわかる．具体的に芥川は図 2 のような構造で文を書きやすい．

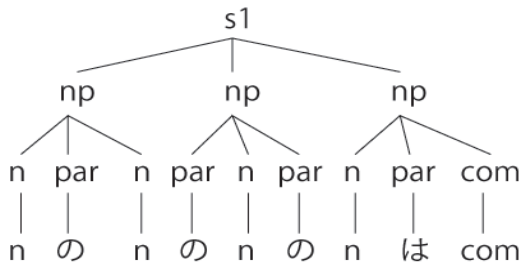

図 2: 芥川の用いやすい句構造

表 1: 木カーネルを用いた SSTs（Subset Trees）の代表 5 作家間のカーネル値（基準：最頻値）

	芥川	太宰	宮沢	夏目	新美
芥川	2.69	2.23	2.24	2.49	2.20
太宰	2.23	2.18	1.96	2.18	2.00
宮沢	2.24	1.96	2.06	2.18	1.96
夏目	2.49	2.18	2.18	2.54	2.16
新美	2.20	2.00	1.96	2.16	2.08

5. 関連研究

係り受けについて分析を行なったものが金川らの研究 [5] である．係り受けの解析によって，各作家がどのような係り受け構造を特徴として文を書いているのかを明らかにした．

6. まとめ

本研究では文書間の類似度を測るため，木カーネルを用いて評価実験を行ない，作家の特徴を分析した．実験では句構造解析を行ない，1 コーパス 100 文と制約を設けたコーパスを用いて実験を試みた．その結果，句

表 2: 木カーネルを用いた SSTs（Subset Trees）の代表 5 作家間のカーネル値（基準：中央値）

	芥川	太宰	宮沢	夏目	新美
芥川	11.07	9.11	9.10	9.86	8.74
太宰	9.11	8.53	7.85	8.52	7.80
宮沢	9.10	7.85	8.32	8.56	7.60
夏目	9.86	8.52	8.56	9.56	8.20
新美	8.74	7.80	7.60	8.20	8.74

表 3: 木カーネルを用いた SSTs（Subset Trees）の代表 5 作家間のカーネル値（基準：平均値）

	芥川	太宰	宮沢	夏目	新美
芥川	19.21	15.52	16.91	17.15	15.75
太宰	15.52	14.37	14.16	14.63	13.83
宮沢	16.90	14.16	16.55	15.83	14.43
夏目	17.15	14.63	15.83	16.75	14.81
新美	15.75	13.83	14.43	14.81	14.36

構造に着目した場合の作家の特徴がどのようなものなのかを分析することができた．句構造解析に関しては，小田 [6] による日本語句構造解析器「Ckylark」があるので，今後これを用いた実験も行なう予定である．

参考文献

[1] Collins, M. and N. Duffy (2001). Convolution kernels for natural language. *Advances in Neural Information Processing Systems 14* [*Neural Information Processing Systems: Natural and Synthetic, NIPS 2001*] , 625-632, MIT Press.

[2] 長谷川守寿 (1994). 日本語の句構造規則, **筑波応用言語学研究**, 59-71.

[3] Petrov,S., L. Barrett, R. Thibaux, and D. Klein (2006). Learning accurate, compact, and interpretable tree annotation, *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, 433-440.

[4] Moschitti, A. TREE KERNELS IN SVMLIGHT. http://dit.unitn.it/moschitt/.

[5] 金川絵利子, 佐原諒亮, 岡留 剛（2015）. 情報量木カーネルとそれに基づく作家の構文類似性解析, **言語処理学会　第 21 回年次大会　発表論文集**, 864-867.

[6] 小田悠介 (2015). 解析失敗の発生しにくい PCFG-LA 句構造構文解析, **言語処理学会　第 21 回年次大会　発表論文集**, 111-114.