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

金川絵利子 † 佐原諒亮 †† 岡留 剛 †
† 関西学院大学理工学部 †† 関西学院大学大学院理工学研究科
eriko.k@kwansei.ac.jp

1 はじめに

作品に基づく作家の分類や特徴づけは古くから興味を持たれさまざまな研究が行なわれてきた．それらにおいては，作品を特徴づける量として，文書中に含まれる文の長さや読点の数の平均値・単語の出現頻度といった文の表層的な統計量が一般的に利用されている (例えば，前川, 1995; 金, 1994)．一方，よく言われる「作家の文体」いう言い回しにおける「文体」という表現は，字面の表層的情報だけでなく，作品の意味内容や，さらには書かれている媒体さえも含んでいるという主張さえある (山本, 2014)．しかし，作品を構成する文の表層的統計量と作品の意味内容とのちょうど中間に位置づけられる文の構文上の違いについてはほとんど議論されてこなかった．それは主に構文の違いを数値化する困難さに起因していたと思われる．本研究では，作家の文体を特徴づける重要な要因として作品を構成する文の構文情報に焦点をあてる．

非数値的構造データの類似度を測る尺度としてさまざまなカーネルが提案されてきた．その 1 つに木構造を入力とする木カーネルがあり，言語解析でそれが用いられている (Collins and Duffy, 2001)．例えば，Mocshitti (2006) は，ラベルづけられた文に対し，木カーネルを用いて SVM により文書分類を行なっている．本研究でも，木カーネルを基本技術として出発点に据える．ただし，文書中の各文に対し木カーネルを用いて文書分類を行なった場合，単語の違いを無視すると文の句構造が同じであれば，出現頻度が低い部分木と高い部分木で同じ類似度となる．出現頻度が低い同じ構造が 2 つの文書間で出てくればそれらの類似度は高いと考えられ，部分木の出現頻度を反映されるカーネルの構築が課題となる．

本研究は，構文情報による文書類似度の定義づけを行なうため，部分木の出現頻度を反映するカーネルを定義し，それを用いた新たな文書カーネルを構築することを目的とする．そのため，木カーネルに部分木の出現確率を組み込み，木カーネルの概念を拡張するアプローチをとる．なお，本稿では，構文木という用語は，句構造を表現する木や，係り受け関係を表現する木など広い意味での構造木を指す．

2 関連研究

Goncalvel and Quaresma (2008) は，ポルトガル語で書かれた文書を木カーネルを用いて分類し，構文構造は分類に適さないという結果を得ている．文を木構造に展開したとき構文木の葉は単語となり，その木をカーネルの入力として用いた場合，構文木の骨格の違いよりも単語の違いが強調される結果となる．彼らの分析では，単語を含む構文木を用いており，そのため木構造そのものの違いが反映されていない可能性が高い．また，彼らが採用した分類クラスは，文化やスポーツといった意味的カテゴリであり，構文上の違いが意味的カテゴリに反映されなかったことも考えられる．

3 情報量木カーネル

木カーネルは，木構造データに対して定義されたカーネルである．与えられた２つの文の構文木における共通の部分木の個数を数え上げカーネル値する．一般に，カーネルは，引数である２つの対象間の類似度を表す一つの指標である．木カーネルも２つの木のある類似度を表現するが，木に含まれるすべての部分木を対等に扱うため，共通部分木の数という意味での類似度になっている．ここで，一般的にはあまり用いられることがない共通の独特の構文を持つ言い回しをしばしば使う２人の作家を考えよう．この共通の構文をこの２人以外の作家はあまり用いないということは，この構文こそが２人の作家の文体を特徴付ける一つの重要な要因であるといえる．しかし，木カーネルを直接構文の類似度として用いたのでは，このような特徴を浮かび上がらせることはできない．具体例で示そう．以下の Tiny Einglish における文 s_1 と s_2 のカーネル値と，文 s_3 と s_4 のそれは等しい．一方，s_1 と s_2 の構文に含まれる部分木の生成確率は，s_3 と s_4 のそれに比べると大きく，s_1 と s_2 の構文はごく普通に文書中に現れるが，s_3 と s_4 の構文は比較的まれに使われる構文と言える．

Tiny English
s_1：I love you.　s_2：he likes books.
s_3：beautiful weather.　s_4：good music.
句構造規則：生成確率
S → N VP：0.8　S → A NP：0.2
VP → V N：1.0　NP → N：1.0
A → beautiful：0.5　A → good：0.5
N → I：0.2　N → you：0.2
N → He：0.2　N → books：0.2
N → weather：0.1　N → music：0.1
V → love：0.5　V → likes：0.5

この Tiny English では，s_1 と s_2 の木カーネル値と s_3 と s_4 のそれは両者とも 3 であるのに対し，s_1 と s_2 の共通部分木の生成確率は 0.8 であり，s_3 と s_4 のそれは 0.2 で前者と大きく異なる．そこで構文中の部分木の生成確率を加味したカーネル，すなわち，情報量木カーネルを提案する．

以下では，各エッジにその生成確率が付与された構文木（生成確率つき構文木）であり，1 つの木におけるそれぞれの部分木の生成は独立であることを仮定する．2 つの文 1 と文 2 のそれぞれの構文木を T_1,T_2 とし，N_1 を T_1 のノードの集合, N_2 を T_2 のノードの集合 とする．T_1 と T_2 が与えられたとき，T_1 と T_2 の**情報量木カーネル**を以下のように定義する．

$$K_i(T_1,T_2) = h(T_1)\cdot h(T_2)$$
$$= \sum_i \lambda^{\mathrm{size}(i)} h_i(T_1)h_i(T_2)(-\log p_i)$$
$$= \sum_{n_1\in N_1}\sum_{n_2\in N_2}\sum_i \lambda^{\mathrm{size}(i)} I_i(T_1)I_i(T_2)(-\log p_i)$$
$$= \sum_{n_1\in N_1}\sum_{n_2\in N_2} \Phi(n_1,n_2).$$

ただし，

1. (n_1をルートとする部分木) ≠ (n_2をルートとする部分木) のとき
$$\Phi(n_1,n_2) = 0,$$
2. (n_1をルートとする部分木) ≠ (n_2をルートとする部分木) で, かつ 終端記号の前であれば
$$\Phi(n_1,n_2) = \lambda(-\log p_i),$$
3. それ以外のとき (Subset Trees)
$$\Phi(n_1,n_2) = \lambda(2^{nc(n_1)}(-\log p_i) + \sum_{j=1}^{nc(n_1)} 2^{nc(n_1)-1}\Phi(ch(n_1,j),ch(n_2,j)))$$

ここで，$h_i(T)$ は，すべての木に 1 から番号をつけたとして，i 番目の部分木が木 T に出現する回数である．p_i は i 番目の部分木の生成確率であり，生成確率の低い部分木に対して大きいカーネル値を与えるために驚き度合いである情報量を用いている．$I_i(n)$ はノード n をもつ部分木中に i 番目の部分木が，存在するとき 1，それ以外 0 となる指示関数である．$nc(n)$ はノード n が持つ子ノード数を表し，$ch(n,j)$ はノード n を持つ部分木の j 番目の子ノードを示す．また，$size(i)$ は i 番目の部分木を生成するために適用した生成規則数で，λ は $0<\lambda\leq 1$ を満たし，木の大きさに対する依存度を低くする効果を持つパラメータである．

情報量木カーネルは，ヒルベルト空間 (l_2) での内積になることを示すことで，カーネルであることを証明できる．

T_1 と T_2 の情報量木カーネル値，は共通する部分木の情報量の和となり，生成確率が低い共通の部分木が 2 つの木で出現するほど値は大きくなる．先に挙げた Tiny English における s_1 と s_2 の情報量木カーネル値は 0.322 bit であるのに対し，s_3 と s_4 のそれは 2.322 bit と約 7 倍となる．

4　情報量文書カーネル

文書の類似度を定義するため，情報量木カーネルを用いた文書カーネルを定義する．二つの文書中の文どうしの情報量木カーネルの平均値を用いているため，本研究で提案するカーネルは平均情報量文書カーネルと呼ぶことにする．平均情報量文書カーネルは bags of sentences を仮定し，1 つ目の文書の各文と 2 つ目の文書の各文との情報量木カーネル値の平均値を平均情報量文書カーネル値とした．2 つの文書 D_1 と D_2 は有限個の文の集合で，各文の構文木の各エッジにはその生成確率が付与されていると仮定する．このとき文書 D_1 と D_2 の**平均情報量文書カーネル** $K_D(D_1,D_2)$ を以下のように定義する．

1. $D_1=\phi$ または $D_2=\phi$のとき
$$K_D(D_1,D_2) = 0,$$
2. $$K_D(\{s\},D_2) = \frac{1}{|D_2|}\sum_{\bar{s}\in D_2} K_i(T_s,T_{\bar{s}}),$$
3. $$K_D(D_1\cup\{s\},D_2) = \frac{1}{|D_2|+1}(K_D(D_1,D_2)+K_D(\{s\},D_2)).$$

ここで $K_i(T_s,T_{\bar{s}})$ は文 s と $\bar{s}$ の情報量木カーネル値であり，$|D|$ は文書 D の文数を示す．Haussler (1999) の lemma 1 により，平均情報量文書カーネルがカーネルであることが証明できる．平均情報量文書カーネルは，1 文と 1 文あたりの共通する部分木の情報量の平均であるという意味合いを持つ．

5　評価

文の構文を表現する方法には，句構造文法によるものや，係り受け解析によるもの・意味論的構造も考慮した LFG や HPSG など様々ある．本研究では純粋に構文構造の違いに焦点を当てるため，句構造と係り受け構造とで文の構造を表現する．しかし，現在のところ，さまざまな作家の作品を構文解析できるだけ十分

に強力で一般的な日本語句構造文法は存在しない．そのため，本研究では，係り受け構造にしぼり作家の文を分析する．

5.1 前処理

まず，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は削除し，また,「」内の会話文は「」を削除し会話文の本文のみを使用する．その他の記号に対しては原文通り使用した．

1 文ずつ Cabocha(工藤・松本, 2002) を用いて形態素解析と係り受け解析を行なった．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，骨格としての構文構造の類似性に注目するため，用いられる単語の違いによるカーネル値への影響は極力排除したい．そのため，Cabocha の形態素解析をもとに，長谷川 (1994) に基づいて単語を，品詞と形態素情報を表す記号に還元的に置換した．例えば，(私は 音楽を 聴きながら， 大好きな 本を 読んだ．) の縮約的還元は，(n は n を vccr j n を v) となる．ただし，n，vccr，j，v はそれぞれ名詞,「ながら」が語尾に付いた動詞，形容詞，動詞を表す非終端記号である．

5.2 生成確率

生成確率として本研究では，出現回数に基づく相対頻度と Cabocha のスコアによる「相対頻度」を用いた．毎日新聞 3 年間分 (2010 年から 2012 年) と，NHK の NEWS WEB 6 0 日分 (2014 年 7 月 20 日から 2014 年 7 月 27 日と，2014 年 9 月 12 日から 2014 年 11 月 3 日), さらに青空文庫の中から比較的作品数の多い 34 作家の 5,909 作品から成るコーパス (文数 5,511,696，文節数 42,024,675，句読点を除く単語数 109,929,329，単語の種類 327,487) から係り受けの生成確率を計算した．すなわちまず，コーパス中のすべての文に対して，Cabocha で形態素解析を行ないさらに還元的縮約を行なう．Cabocha の係り受け解析の結果から，係り元の品詞と係り先の品詞などによるすべての種類の係り受けを列挙し，そのおのおのの出現回数と Cabocha の総スコアを求める．ある文節の係り先の文節の種類も重要であるが，係り受けの文節間距離も重要な構文情報である．文節間の距離は隣り合う文節同士で 1 とし，間に k 個の文節が存在する場合を k とした．しかし，すべての係り受けの種類と文節間距離を考慮し区別すると，係り受けの種類が多く，ほとんどのもの相対頻度が 0 に近くなる．そのため係り受けの種類ごとに，文節間の距離のグルーピングが必要である．そのために，まず，距離に適当なしきい値を決め，そのしきい値より大きい距離を同一とする方略が考えられるが，しきい値より大きい距離の係り受けを持つ相対頻度の総和が，しきい値での相対頻度を上回ってしまう．また，しきい値の決め方も困難である．そこで「Fibonacci 数列」を用いてグループを構成した．すなわち，Fibonacci 数列の各数が 1 つのグループの構成員数となるように，距離 1 のものから順にグループ化する．全係り受けの出現総数と，ある文節間距離を考慮した係り受けの出現回数の比として，相対頻度を計算し，その係り受けに対する生成確率とする．Cabocha のスコアに基づく「相対頻度」からの生成確率も同様に計算する．

5.3 情報量木カーネル値

各文に対して，還元的縮約を行なった確率付き構文木から情報量木カーネルを計算し，平均情報量文書カーネルを求める．情報量木カーネルの実装は，Mocshitti (2006) の木カーネルプログラムを拡張する形で行なった．作成したプログラムの計算量は，2つの木のノード数の積に比例する．

5.4 実験

青空文庫の作家の中から，著名な 5 作家の芥川龍之介と太宰治・夏目漱石・新美南吉・宮沢賢治を選択し実験を行なった．各作家の全作品からランダムに 100 文抽出し，2 作家の平均情報量文書カーネル値を求める．これを 10 回行なったものの平均を情報量文書カーネル値とした．パラメータ λ の値は木カーネル値を求める Mocshitti (2006) のデフォルトの 0.4 とした．出現相対頻度に基づく情報量を用いる方法と，Cabocha のスコアから求めた情報量を用いる方法のそれぞれに対して，Subset Trees(SSTs) Kernel と SubTrees(STs) Kernel の二種類の実験を行なった．計 4 種類の実験を行なったが，今回はスペースの関係上，出現相対頻度に基づく情報量を用いた SSTs のみ議論を行なう．各作家ごとの平均情報量文書カーネルを表にまとめた (表 1) .

5.5 議論

表 1 から分かるように芥川と夏目の情報量文書カーネル値が大きい．それゆえ，この二人は一般的に珍しい係り受けを用いた文を多く書くことで，構文的に似ていると推測できる．珍しい係り受けには，その作家らしさ，作家の文体の特徴が含まれていると考えられる．宮沢と新美の情報量文書カーネル値が小さいことから，この二人は構文的に似ていない文を書くといえよう．

次に，芥川と太宰と夏目の 3 人に着目する．芥川と太宰と夏目は，今回使用した青空文庫中の作品に対して，1 文あたりの平均文節数や 1 文あたりの平均単語

表 1: 出現相対頻度に基づく情報量を用いたSSTs(Subset Trees) の代表５作家の情報量文書カーネル値.

	芥川龍之介	太宰治	宮沢賢治	夏目漱石	新美南吉
芥川龍之介		321	224	527	200
太宰治	321		147	197	120
宮沢賢治	224	147		123	101
夏目漱石	527	197	123		162
新美南吉	200	120	101	162	

数の値がよく似ている．しかし，芥川と太宰の情報量文書カーネル値と，芥川と夏目のそれは大きく異なる．１文の長さが長ければ，文に含まれる部分木の個数も一般的には増加するため，文の長さに情報量木カーネルが依存しているのであれば，情報量文書カーネル値は大きくなる．しかし，文の長さがほぼ等しいの芥川と太宰の情報量文書カーネル値と，芥川と夏目の情報量文書カーネル値の値に差があることから，情報量木カーネルは，ノード数 (文の長さ) への依存が少なく，構文類似度ををとらえられている考えられる．また，宮沢と新美も，１文あたりの平均文節数や１文あたりの平均単語数の値がよく似ている．しかし，宮沢と新美の情報量文書カーネル値はこの５作家の総当たりで一番小さい．二人とも児童文学を中心としているが，構文的には異なる文を書くと考えられる．

なお，５作家の類似度をバネモデル (Kamada and Kawai, 1989) を用いて可視化した (図１)．その際,「距離」は情報量文書カーネルの逆数とした．

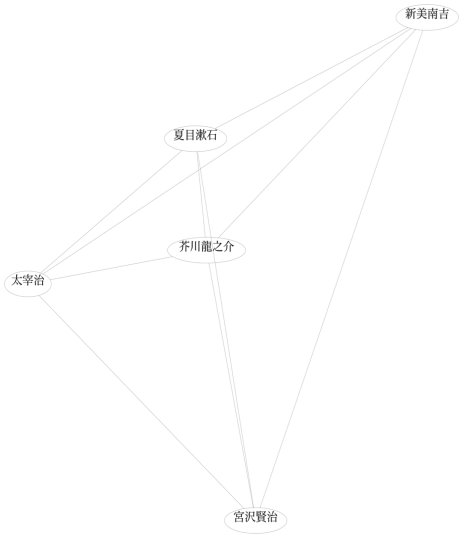

図 1: 出現相対頻度に基づく情報量を用いたSST(subset tree) の情報量文書カーネル値による代表５作家のバネモデルでの表現.

6 おわりに

本研究は，構文情報により文書の類似性を測るため，部分木の出現頻度を反映する情報量木カーネルを定義し，それを用いた情報量文書カーネルを構築した．また，日本を代表する作家の文を用いた評価実験を行なった．前処理では文書のクリーニングと還元的縮約を行ない，係り受けの生成確率を計算し，情報量木カーネル値を求め，情報量木カーネル値に基づき情報量文書カーネルを求めた．実験では，情報量木カーネルが文の長さへの依存度は低いことや，構文的違いをとらえることができることを確認した．

参考文献

[1] 前川守 (1995). **文章を科学する**, 岩波書店.

[2] 金明哲 (1994). 読点の打ち方と著者の文体特徴, **計量国語学, 19**, 7, 317-330.

[3] 山本貴光 (2014). **文体の科学**. 新潮社.

[4] Collins, M. and N. Duffy (2001). Convolution kernels for natural language. *In Advances in Neural Information Processing Systems.* 625-632.

[5] Moschitti, M. (2006). Efficient convolution kernels for dependency and constituent syntactic trees. *Proceedings of the 17th European Conference on Machine Learning* (*ECML*2006), 318-329.

[6] Goncalvel, T. and P. Quaresma (2008). Text classification using tree kernels and linguistic information. *Proceedings of the Seventh International Conference on Machine Learning and Applications* (*ICMLA*'08), 763-768.

[7] Haussler, D. (1999). Convolution Kernels on Discrete Structures. *Technical Report* UCSC-CRL 99-10, University of California.

[8] 工藤拓, 松本裕治 (2002). チャンキングの段階適用による日本語係り受け解析, **情報処理学会論文誌**, **43**, 6, 1834-1842.

[9] 長谷川守寿 (1994). 日本語の句構造規則, **筑波応用言語学研究**, **1**, 55-71.

[10] Kamada, T. and S. Kawai (1989). An algorithm for drawing general undirected graohs. *Information Processing Latters*, **31**, 1, 7-15.