1N2-3in1

構文構造に着目した文体の類似度の数値化

情報量木カーネルと木カーネルによる分析

An Analysis of Similarity between Auther's Writing Style by Information Tree Kernel

金川絵利子 *1　佐原諒亮 *1　岡留 剛 *1
Eriko KANAGAWA　Ryosuke SAWARA　Takeshi OKADOME

*1関西学院大学大学院理工学研究科
Graduate School of Science and Engineering, Kwansei Gakuin University

The information tree kernels proposed here give measures of the syntactic similarity of sentences. For two trees, they are defined as the total amount of information contained in their common subtrees, where the amount of information of a common subtree is calculated using the occurrence probability for the subtree. The information tree kernels defined by the kernels enable us to capture the syntactic similarities and differences in Japanese famous 5 authors' writing styles.

1. はじめに

本研究は，構文構造に着目して，作家間の文体類似性や，作家の文体特徴を捉えることを目的とする．そのため，木カーネルを用い，また新たに情報量木カーネルを定義し導入することで，構文構造の類似性を数値で表現するアプローチをとる．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，1) 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や，2) 読点の直前の格助詞の出現頻度や特定の文節の出現頻度といった構文情報のある側面を表現している量が利用されてきた．(例えば，[金 94]). 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，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 14].

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構造そのものの違いについてはほとんど議論されてこなかった．それは主に構文構造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構造に焦点をあてる．

主に機会学習の分野で，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用いられている [Collins 01]．例えば，[Moschitti 06] 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，構文木の類似性を測る尺度として木カーネルを採用する．

一般に，文体比較としては文や句の長さや名詞の頻度など統計量を用いたアプローチと，作品中の稀出語や稀出表現を取り上げて検討するアプローチがある [前川 95]．前者はまれにしか出現しない語や表現は誤差と考え，より多く使われる言葉の違いや平均的な違いを調べる手法といえる．それに対し後者は，まれな表現は作家の特徴を表わすと考え，それらについて議論する．

木カーネルは，2 文の構文木に共通の部分木の数でその値が定まり，とりたててまれな構造を抽出するわけではない．その意味で，木カーネルを用いた分析は構文の統計的・平均的な解析といえる．一方，出現頻度が低い同じ構造が 2 つの文書間で出てくれば，それが構文間の類似性を強く反映するという考え方もあり，本研究では，部分木の出現頻度を反映する木カーネル，すなわち情報量木カーネルを定義し，作家間の特徴を表現する構造を取り出すことも試みる．

連絡先: 氏名：金川 絵利子
所属：関西学院大学大学院理工学研究科
住所:〒 669-1337 兵庫県三田市学園 2-1
メールアドレス：eriko.k@kwansei.ac.jp

2. 関連研究

[Goncalvel 08] 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．

[金 02] は，助詞の n-gram パターンを用いた書き手の識別法を提案しており，一般人の書いた短い文章に対しても高い判別率で書き手の判別が可能であるという結果を得ている．

3. カーネルを用いた類似性分析手法

3.1 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 [Collins 01]．木カーネルには，subset trees kernel や subtrees kernel ，あるいは [鹿島 06] で提案されたラベル付き順序木カーネルを拡張したものなど，いくつか種類がある．本研究では，subtrees kernel を用いる．ここでは，subtrees kernel の定義とその意味を与える．すなわち，木構造 T_1，T_2 に対して，subtrees kernel は以下の式で定義される．

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

ここで，S は部分木 τ であり，τ はすべての固有木の集合である．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．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わ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．

3.2 情報量木カーネル

情報量木カーネルを導入する．以下では，各エッジにその生成確率が付与された構文木（生成確率付き構文木）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と仮定する．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．

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

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．また，$size(i)$ は i 番目の部分木の深さを表わし，λ は $0 < \lambda \leq 1$ を満たすパラメータであり，木の大きさに対する依存度を低くする効果を持つ．

木 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．

3.3 予備実験

情報量木カーネルが，実際に文の構造を捉えることができるかという予備実験を行なう．現代日本語書き言葉均衡コーパス (以下 BCCWJ と表記する)[山崎 14] の中から，ベストセラーと Yahoo!知恵袋・法律・白書のカテゴリーに対し，各カテゴリーの全作品からランダムに 100 文抽出し，情報量木カーネル値を求める．BCCWJ の中から，C-XML のサンプル長可変のデータを用いる．前処理と出現確率・カーネル値の計算については第 4 章で詳しく述べる．情報量木カーネル値は，抽出した 100 文の組の総当たりで計算し，カーネル値上位 100 の値の平均値とした．これを 10 回行なったものの平均を用いる．パラメータ λ の値は，詳しくはやはり第 4 章で述べるが，情報量木カーネルでは 1.0 とする．その理由も 4 章で詳しく述べる．結果を表 1 にまとめる．

表 1: BCCWJ の 4 カテゴリーの STs(Sub Trees) の情報量木カーネル値.

	ベストセラー	Yahoo!知恵袋	法律	白書	平均
ベストセラー	236.27	6.98	12.83	11.34	10.38
Yahoo!知恵袋	6.98	153.12	8.33	8.12	7.81
法律	12.83	8.33	805.70	18.20	13.12
白書	11.34	8.12	18.20	637.96	12.55

他のカテゴリーとの類似度を比較した場合，最も値が大きいのが法律と白書であり，最も値が小さいのがベストセラーと Yahoo!知恵袋という結果になった．値が大きい方が類似度が高いと考えるため，直感に一致している結果であると考えられる．

4. 文体の類似性に関する実験

構文の表現には，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 などがある．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，係り受け構造に着目して作家の文を分析する．

前処理及び還元的縮約は [金川 15] 4 章参照.

4.1 出現確率

生成確率 (出現確率) として，出現回数に基づく相対頻度を用いる．毎日新聞 3 年間分 (2010 年から 2012 年) と，NHK の NEWS WEB60 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 2015 年 7 月 14 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，BCCWJ から成るコーパス (文数 $9,067,897$，文節数 $71,504,383$，単語数 $198,608,659$) から係り受けの出現確率を計算した．出現確率の求め方は [金川 15] の 4.2 参照.

4.2 情報量木カーネルによる分析

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する．

与えられた二つの文 S_1 と S_2 の構文木から，共通部分木を取り出す．共通部分木を取り出すと，同じ係り受けでも文節間距離によっての確率が異なる．そのため，共通部分木の生成確率は，S_1 の共通部分木部分の生成確率と S_2 の共通部分木部分の生成確率の平均とする．

情報量木カーネルを用いた解析は，作家特有の表現に着目した，まれな現象に注目した解析といえる．そのため，各作家の全作品からランダムに 100 文抽出し，各作家 1 文ずつ順に情報量木カーネル値を総当たりで計算する．総当たりで計算した情報量木カーネル値の平均を結果としてしまうと，情報量木カーネルを用いた解析は，作家特有の表現に着目した，まれな現象に注目した解析であるということに反する．より，まれな現象，情報量木カーネル値の大きなものに着目するために，総当たりの平均ではなく，総当たりで計算した情報量木カーネル値の内，上位 100 の情報量木カーネル値の平均を結果とする．この実験を 10 回行ない，各結果の平均値を情報量木カーネルを用いた解析の結果とした．

また，パラメータ λ は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．λ が小さいほど木の大きさに対する依存度が低くなる．しかし，情報量木カーネルは生成確率を用いり，大きな木が一致した場合，生成確率が非常に小さくなるため，情報量木カーネル値は大きくなる．よって，生成確率に基づく情報量と λ は逆の性質を持っていることいえる．例を挙げて具体的に説明する．図 1 に示すような木 T_1 と T_2 が与えられたとする．以下で述べる実験と同様に，与えられた木に対して部分木集合を生成する．

部分木 ST_1 の情報量が 10，部分木 ST_2 と ST_3 の情報量が 2 であるとする．ここで，部分木集合の中から木 ST_1 と ST_2 に着目する．木 ST_1 と ST_2 の $size(i)$ はそれぞれ 2 と 1 である．$\lambda = 1.0$ の場合，それぞれの部分木に着目した情報量木カーネル値は 10 と 2 となり，$\lambda = 0.1$ の場合，情報量木カーネル値は 0.1 と 0.2 となる．よって，λ の値により情報量木カーネル値の大小関係が逆転する場合がある．情報量木カーネルを用いた解析はまれな現象に注目した解析といえるため，情報量の値が重要であり，木の大きさに情報量木カーネル値が変わること

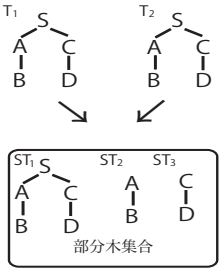

図 1: 木 T_1，T_2 が与えられた時の部分木集合．それに対応する句構造規則と情報量.

を避ける必要があると考える．よって，情報量木カーネル値を求める場合の λ は 1.0 とする

4.3 木カーネルによる分析

各作家の全作品からランダムに 100 文抽出し，各作家 1 文ずつ順に木カーネル値を総当たりで計算する．木カーネルを用いた解析は統計的・平均的な解析といえるため，すべての文を平等に扱い，すべての木カーネル値の平均値を結果とした．これを 10 回行なったものを平均を木カーネルを用いた解析の結果とした．

共通部分木を求める場合の文節間距離の扱いは情報量木カーネルと同様である．

パラメータ λ は木の大きさに対する依存度を低くする効果を持つため，パラメータ λ が小さいほど木の大きさに対する依存度が低くなる．木カーネルの場合，大きな木が一致した場合ほど共通部分木の数は増える．係り受けの木構造では，木の大きさは主に単語数に依存する．よって，情報量木カーネル同様 $\lambda = 1.0$ とすると，木カーネル値は単語数に依存した値となるため，パラメータ λ の値は，木カーネル値を求める [Moschitti 06] のデフォルトの 0.4 とした．

4.4 結果

青空文庫の作家の中から比較的作品数の多い 34 作家で実験を行なった．各作家の全作品からランダムに 100 文抽出し，部分木集合に対して，情報量木カーネル値と木カーネル値を求めた．今回はスペースの関係上，著名な 5 作家の芥川龍之介と太宰治・夏目漱石・新美南吉・宮沢賢治について議論を行なう．結果を表にまとめた (表 2 ・表 3) .

平均とは，自分以外の他の 4 作家とのカーネル値の平均である．表 2 の情報量木カーネル値の結果より，他の作家との情報量木カーネル値が一番大きいのは芥川である．最も情報量木カーネル値の小さいのは宮沢と新美のペアであり，自分自身との情報量木カーネル値が一番大きいのは夏目である．表 3 の木カーネル値の結果より，他の作家との木カーネル値が一番大きいのは夏目である．最も木カーネル値の小さいのは宮沢と新美のペアであり，自分自身との木カーネル値が一番大きいのは夏目である．他の作家とのカーネル値の平均に注目すると，情報量木カーネル，木カーネル共に，芥川と夏目が比較的カーネル値が大きいグループであり，太宰と宮沢と新美が比較的カーネル値が小さいグループである傾向を持つ.

表 2: 代表 5 作家の STs(Sub Trees) の情報量木カーネル値の上位 100 の平均値.

	芥川	太宰	宮沢	夏目	新美	平均
芥川	271.48	13.47	11.65	13.90	11.45	12.62
太宰	13.47	278.58	10.27	12.72	10.56	11.75
宮沢	11.65	10.27	214.18	11.37	8.57	10.46
夏目	13.90	12.72	11.37	341.37	11.19	12.30
新美	11.45	10.56	8.57	11.19	202.78	10.44

表 3: 代表 5 作家の STs(Sub Trees) の木カーネル値の総当たり平均値. $(*10^{-3})$

	芥川	太宰	宮沢	夏目	新美	平均
芥川	12.46	2.42	3.36	6.41	3.24	3.86
太宰	2.42	11.07	2.27	3.68	3.07	2.86
宮沢	3.36	2.27	13.95	4.14	2.44	3.05
夏目	6.41	3.68	4.14	22.37	4.74	4.74
新美	3.24	3.07	2.44	4.74	12.92	3.37

5. 議論

5.1 結果に関する議論

他の作家との情報量木カーネル値の平均では比較的カーネル値の小さいグループであるが，自分自身との情報量木カーネル値は大きな太宰に注目する．他の作家との情報量木カーネル値は比較的小さく，自分自身との情報量木カーネル値が大きいということは，太宰が他の作家は用いない出現確率の低い部分木を用いた文を多く書くと考える．この推測を裏づける具体例として，太宰の “二人ならんでお母さまの枕もとに坐ると、お母さまは、急にお蒲団の下から手をお出しになって、そうして、黙って直治のほうを指差し、それから私を指差し、それから叔父さまのほうへお顔をお向けになって、両方の掌をひたとお合せになった。” という文がある．還元的縮約した構文木では図 2 のようになる．単語数が多いのに対し，木の深さが浅く単語を並列に並べてていることが分かる．単語を並列に並べると，部分木の出現確率はそれぞれの係り受けの出現確率の積となるため，出現確率が小さくなり，情報量木カーネル値の値が大きくなる．よって，太宰は単語を並列に並べる文を多く書くという特徴があり，他の作家はこのような構文をあまり書かないと考えられる．また，太宰のこのような特徴は自分自身との木カーネル値が小さいが，情報量木カーネル値では比較的大きな値をとることからもいえる．単語を並列に並べることで，同じ単語数で直列に並べたもとと比較すると，部分木数は非常に少なくなるが，情報量は大きくなる.

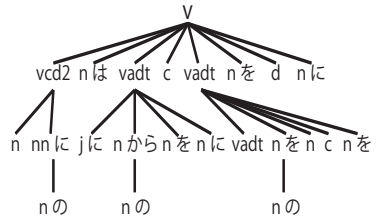

図 2: 太宰の構文上の特徴を表す文の構文木.

5.2 Cabocha のエラーの影響

情報量木カーネルを用いる場合，部分木の生成確率が必要となる．部分木の生成確率として本研究では，Cabocha の形態素解析と係り受け解析の結果に基づくの係り受けの出現回数の相対頻度を用いた．情報量木カーネルは，出現回数が低い同じ構造が 2 つの文間で出てくれば, それが構文間の類似性を強

く反映する見方をするため，係り受けの出現回数が重要な役割を演じる．Cabocha の形態素解析及び係り受け解析がすべての文に対し正しいならば，係り受けの出現回数は信用できるものであるといえるが，Cabocha の形態素解析及び係り受け解析が間違っているため，出現回数が低くなり確率が小さくなっているとも考えられる．そこで，Cabocha のエラーに関して分析した．

実験に用いた，芥川と太宰・夏目・新美・宮沢の 5 作家に対して，情報量木カーネル値が 500 以上の文に対して，各作家 20 文，合計 100 文を人手で評価した．1 文に対して，5 段階評価を行なった．5 を正しいとし，3 を分からない，1 を誤りとした．結果を表 4 にまとめた．100 文中 43 文で誤りがある

表 4: 100 文にに対する Cabocha のエラー評価.

5	38 文
4	13 文
3	6 文
2	20 文
1	23 文

という結果になった．誤りと判断した 1 と 2 についてより詳しく解析した．

2 として判断した文の内訳は，本文中に旧字が含まれているため誤りが生じているものが 11 文，品詞は正しいが文節の区切りが誤っているものが 5 文，品詞はや文節は正しいが係り受けが誤っているものが 2 文，明らかに 2 文節以上に区切られていなければ不自然なものが 1 文ある．2 として判断した実際の文の例として以下のようなものがある．旧字が原因の例として，“そこへ丁度この清夫のすきとほるばらの実のはなしを聞いたもんですからたまりません。” がある．文中の “すきとほる” の部分が，Cabocha の解析では “すきと” と “ほる” の 2 文節に分かれる．

また，1 として判断した文の内訳は，ひらがなが原因で誤りが生じているものが 19 文，名詞が原因のものが 1 文，動詞が原因のものが 1 文，古文に似た言い回しであるため誤りが生じているものが 1 文，送りがなが現代仮名遣いと異なるため誤りが生じているものが 1 文ある．1 として判断した実際の文の例として以下のようなものがある．ひらがなが原因の例として，“鐘かねにはよしひこさんがひとりついて、まちのこくみんがっこうのこうていまでゆくことになっていた。” がある．文中の “よしひこさんが” の部分が，Cabocha の解析では，“よし” と “ひこさんが” の文節に分かれ，“よし” は副詞，“ひこ” は動詞となる．

しかし，1 文 1 文を確認しているため，43 %という結果になったが，1 文が誤りと判定されている場合でも，ほとんどの文で 1 文節の中の 1 単語のみが間違っており，他の文節は正しい．よって，共通部分木にその誤りの文節が出現する確率は少なく，実験結果の信頼性への影響も少ないと考えられる．

6. おわりに

本研究は，構文情報により作家の文体の類似性を測るため，木カーネルを用いり，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た．情報量木カーネルが実際に文の構造を捉えることができることを予備実験で確認した．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，係り受けの出現確率を計算し，情報量木カーネル値・木カーネル値を求めた．情報量木カーネル値を求める時に必要な係り受けの確率に Cabocha のエラーが重要な影響を与える．そのために Cabocha の係り受け解析のエラー率関しても分析し，影響が少ないことを確認した．

参考文献

[前川 95] 前川守 (1995). **文章を科学する**, 岩波書店.

[金 94] 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.

[金 02] 金明哲 (2002). 助詞の n-gram モデルに基づいた書き手の識別, **計量国語学, 23**, 5, 225-239.

[山本 14] 山本貴光 (2014). **文体の科学**. 新潮社.

[Bille 03] Bille, P(2003). *Tree Edit Distance Alignment Distance and Inclusin*, Technical report TR-2003, 23, IT University of Copenhagen.

[Collins 01] Collins, M. and N. Duffy (2001). Convolution kernels for natural language. *In Advances in Neural Information Processing Systems*. 625-632.

[鹿島 06] 鹿島久嗣, 坂本比呂志, 小柳光生 (2006). 木構造データに対するカーネル関数の設計と解析, **人工知能学会論文誌, 21**, 1, 113-121.

[Moschitti 06] 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.

[Goncalvel 08] 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.

[金川 15] 金川絵利子, 佐原諒亮, 岡留剛 (2015). 作家の文体の類似性, **第 29 回人工知能学会全国大会予稿集**,NO.2K1-1in.

[山崎 14] 山崎誠 [編] (2014). **『 書き言葉コーパス —設計と構築—』講座日本語コーパス 2**, 朝倉書店.

[工藤 02] 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.

[金 96] 金明哲 (1996). 助詞の分布に基づいた文章の著者の認識, **行動計量学会第 24 回大会論文抄録集**, 144-147