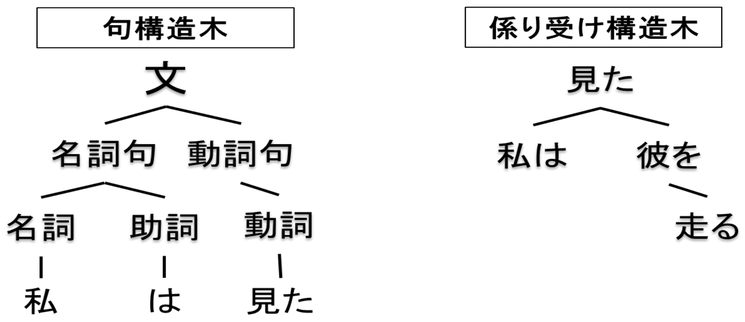
構文に着目した文体類似性の数値化
~情報量木カーネルと木カーネルによる分析~

関西学院大学大学院　理工学研究科　金川絵利子　佐原諒亮　岡留剛

導入

目的：**構文の違い**に着目して
作家の文体の類似度を考える

構文木：



(情報量)木カーネル

定義：

$$K(T_1,T_2)=\sum_i \lambda^{size(i)} h_i(T_1)h_i(T_2)$$

λ　：木の大きさに対する依存度を低くするパラメータ
$size(i)$：i番目の部分木の深さ
$h_i(T)$：i 番目の部分木が木T に出現する回数
p_i　：i 番目の部分木の生成確率

	1番目の部分木 S / NP VP	2番目 S / VP	3番目 NP / N	…	i番目 VP / V	…
生成確率	p_1	p_2	p_3		p_i	
T_1	1	0	0		1	
T_2	1	0	1		1	

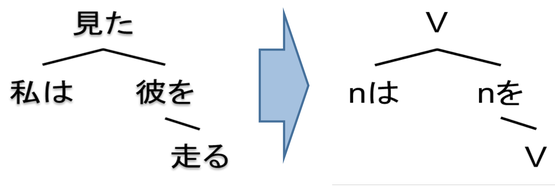
情報量木カーネル値：$K_I(T_1,T_2)=(-\log p_1)+(-\log p_i)$

木カーネル値：$K(T_1,T_2)=2$

還元的縮約・係り受けの確率計算

還元的縮約：

- 句構造：木の葉を削除
- 係り受け構造木：
  単語を品詞と形態素に変換



係り受けの確率計算：

- 係り受け抽出：Cabocha
- 係り受け構造：<係り元,係り先,文節間距離>
- 使用コーパス：9,067,897文

(青空文庫　5,909作品 NHK NEWS WEB　60日分　毎日新聞3年間分)

実験・結果

実験：青空文庫の比較的作品数多い31作家
著名な5作家

実験方法：

ー木カーネル
各作家ランダムに100文選択
総当たり平均を10回行った平均値
λ=0.4

ー情報量木カーネル
各作家ランダムに100文選択
カーネル値上位100の平均を10回行った平均値
λ=1.0

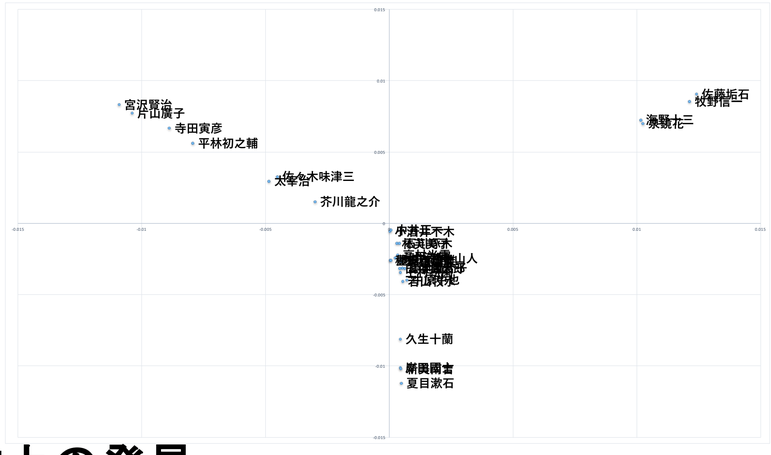
木カーネル：　(×10^{-3})

	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

情報量木カーネル：

	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

ラプラシアン固有マップ法(31作家)：



文学上の発見：

- 芥川龍之介：文節間距離の小さい、直列な文を多く書く
- 太宰治：文節間距離の大きい、並列の長い文を書く
- 夏目漱石：一文の中に同じ係り受けを2回使うことが多い
- 宮沢賢治と新美南吉：
  同じ児童作家でも、宮沢のほうが文節間距離の大きい、複雑な文を書く