文書間類似度について

浅原 正幸[†]・加藤 祥[†]

文書間類似度は、内容の類似度と表現の類似度の二つの側面を持っている。自動要約や機械翻訳ではシステム出力の内容評価を行うために参照要約（翻訳）との類似度を評価する尺度が提案されている。一方、表現を対照比較するための手段として、形態素（列）を特徴量とする空間上の計量が用いられる。本稿では、さまざまな文書間類似度について、距離・類似度・カーネル・順序尺度・相関係数の観点から、計量間の関係や同値性を論じた。さらに内容の同一性保持を目標として構築したコーパスを用いて、内容の差異と表現の差異それぞれに対する各計量のふるまいを調査し、文書間類似度に基づく自動評価の不安定さを明らかにした。

キーワード：文書間類似度，距離空間，カーネル，順序尺度，文体

On Document Similarity Measures

MASAYUKI ASAHARA[†] and SACHI KATO[†]

Document similarity measuring techniques are used to evaluate both content and writing style. Evaluation measures for comparing the summary or translation of a system-generated source text with that of human-generated text have been proposed in text summarization and machine translation fields. The distance metrics are measures in terms of morphemes or morpheme sequences to evaluate or register different writing styles. In this study, we discuss the relations among the equivalence properties of mathematical metrics, similarities, kernels, ordinal scales, and correlations. In addition, we investigate the behavior of techniques for measuring content and style similarities for several corpora having similar content. The analysis results obtained using different document similarity measurement techniques indicate the instability of the evaluate system.

Key Words: *Document Similarity Measures, Distance Metrics, Kernels, Ordinal Scale, Writing Style*

1 はじめに

文書間類似度がはかるものとして「伝える内容の一致」（内容一致）だけでなく「伝える表現の一致」（表現一致）がある。文書間類似度は自動要約や機械翻訳ではシステム出力の内容評価を行うために参照要約（翻訳）との差異を評価する指標として用いられる。一方、文書間類似度は表現の差異を評価することを目的としてテキストの文体の計量比較にも用いられる。

[†] 人間文化研究機構国立国語研究所, National Institute for Japanese Language and Linguistics

本稿では、文書間類似度の数理的構造の説明し、様々な内容もしくは文体が同じであることが想定されるテキストを用いて、各計量の特性について検討する。(難波, 平尾 2008) は 2008 年時点での自動要約の評価指標についての評価をまとめている。2008 年以降に提案された語順を考慮した内容評価のための指標を含めて、語順に対する順序尺度を含めた距離空間・類似度・カーネル・相関係数などの尺度を用いて、数理的構造について整理する。

具体的には、一致部分文字列による尺度・一致部分列による尺度・ベクトル型順序尺度・編集型順序尺度の四つに分類し議論する。これらの四つの尺度に基づき、内容一致（内容の同一性）と表現一致（文体の類似性）の観点から、言語生産過程の多様性を評価する。複数人が同一課題を実施した場合の各評価尺度の分散や、同一人が同一課題を繰り返し実施した場合の各評価尺度の分散などを検討する。生産過程においては口述・筆術・タイプ入力の種類について評価し、課題においては要約・語釈・再話について評価する。要約は長い文書を同等の内容で短く言い換えることを目的とする言語生産過程であるが、語釈は短い単語が指し示す意味と同等の内容を長く言い換えることを目的とする言語生産過程であることから、要約は語釈の逆写像の一般化ととらえることができる。また、再話は長い文書を再度同等の内容でそのまま提示することを目的とする言語生産過程であることから、要約の一般化であるととらえることができる。

この評価を通して、四つの指標における差異がどのような生産過程の差異に現れるのかを調査する。また同一言語生産課題に対する生成物の多様性についても議論する。

表現一致をつかさどるものとして、情報の提示順序を含む修辞法 (rhetoric)・使用域 (register) や位相 (phase)¹に内在する文体・個人に内在する文体などが考えられる。要約を評価するにあたり、内容一致は重要であると考ええるが、表現一致はどの程度重要であるのだろうか。さらにこれらはどの評価尺度に表出するのだろうか。対照比較を介して、各言語生産過程に共通のふるまいを示す評価尺度と課題に特有のふるまいを示す評価尺度について調査する。

自動要約評価のための参照文書は一般に口述筆記の専門家や記者経験者などにより作成され、統制された少数のものが提供される。自動翻訳評価においても職業翻訳家等により限られた数の参照文書が作成される。統制は距離空間上の凸問題として課題を設定し、その課題設定の枠組内で評価したい工学研究者の都合で行われているものである。さらに、工学研究者は参照文書の差異がユークリッド距離空間上に規定され、文書間類似度で比較可能なレベルで統制できるものだと考えているきらいがある。一方、文書を介したコミュニケーションにおいて、言語生産者ではない者による受容過程は統制されるものではなく、複数の受容者間で共有されるものではない。一人の受容者においても時間的経過などで統制できるものでもない。本稿では、

¹ ここでは「使用域」と「位相」は数学の用途ではなく文体論の用語として用いている。

人間の要約作成時の不安定な言語受容過程²において文書の重要箇所選択がどの程度ゆれるものなのかを評価するとともに、そのゆれは評価指標を構成するどの尺度に表れるのかを調査する。この調査を通して、本来誤りでないものが課題設定の時点で誤りになっている可能性があるという実態を明らかにする。

本稿の貢献は以下のとおりである：

- 既存の文書要約や機械翻訳の自動評価に利用される評価指標と、距離空間・類似度・カーネル空間・順序尺度・相関係数などの尺度との関係を整理した
- 同一課題について複数人の言語生産者間で生成される文書のゆれを定量的に評価した
- 課題ごとに同一人の言語生産者の課題試行間で生成される文書のゆれを定量的に評価した
- 上に述べたゆれの評価に基づき、内容評価と表現評価の尺度上のふるまいの不安定さを明らかにした

尚、本稿では、「評価指標」と「尺度」を区別して用いる。自動要約や機械翻訳ではシステム出力の内容評価を行うための ROUGE や BLEU など広く知られているものを表す際に「評価指標」と呼び、「評価指標」を構成する距離空間・類似度・カーネル・相関係数などを「尺度」と呼ぶ。「評価指標」が単一の「尺度」から構成されることもあり、「評価指標」＝「尺度」である場合もある。

以下、2節では既存の自動評価指標を距離・類似度・カーネル・順序尺度・相関係数により説明することで、文書間類似度を四つに分類し整理する。3節では尺度を適用して比較するさまざまな言語生成過程を記録した言語資源について説明する。4節では評価尺度の定性的な評価について示す。5節にまとめと今後の研究の方向性について示す。

2 評価指標と距離・類似度・カーネル・順序尺度・相関係数

2.1 本節の趣旨

本節では、過去に提案されている自動要約と機械翻訳の評価指標を距離・類似度・カーネル・順序尺度・相関係数などの尺度により説明することを試みる。この説明の過程で、いくつかの評価指標が擬距離の公理の対称性、三角不等式や、距離の公理の非退化性を満たさないことに言及する。

まず、部分文字列 (substring) と部分列 (subsequence) の違いを明確にするため、2.2節で部分文字列と部分列に基づく類似度について解説する。2.3節で先行研究で言及されている評価指標について解説する。2.4節で関連するカーネル・順序尺度について示す。2.5節で指標の一般化について述べる。

² ここで言語受容過程とは、要約作成時に元文書を読む過程のことを指す。

評価指標においては、二つのデータ比較という観点から、単一の参照テキストと単一のシステム出力テキストの対の文書間類似度に限定して議論する。複数の参照テキストを考慮する場合には退化性等を考慮する必要がある。尚、本節で用いる用語や記号の定義は A 節にまとめている。

2.2 LCSustr と LCS

2.2.1 記号列と文字列と部分文字列と部分列

評価指標の議論を始める前に、記号列、文字列、部分文字列、部分列の違いについて確認する。

何らかの全順序が付与されている記号集合のことを**記号列**と呼ぶ。本稿では記号列ベクトル $s = \langle s_1, \dots, s_m \rangle, t = \langle t_1, \dots, t_m \rangle$ などと表現する。参照テキスト、システム出力テキストは、ともに文字 (character) ベースの記号列もしくは形態素解析後の形態素 (morpheme) ベースの記号列とみなすことができる。

評価する記号列上の連続列のことを**文字列 (string)**と呼ぶ。記号列の要素が文字 (character) である場合を「文字ベースの文字列 (character-based string)」, 記号列の要素が形態素 (morpheme) である場合を「形態素ベースの文字列 (morpheme-based)」と呼ぶこととする。

記号列に対して隣接性と順序を保持した部分的記号列のことを**部分文字列 (substring)**と呼ぶ。長さ n の部分文字列を特に n -gram 部分文字列と呼ぶ。記号列 s の i 番目の要素からはじまる n -gram 部分文字列を $s_{i \dots i+n-1}$ で表現する。

記号列に対して順序を保持した部分的記号列のことを**部分列 (subsequence)**と呼ぶ。隣接性は保持しなくてよい。長さ p の部分列を特に p -mer 部分列と呼ぶ。記号列 s の p -mer 部分列を、インデックスベクトル $\vec{i} = \langle i_1, \dots, i_p \rangle (1 \leq i_1 < i_2 < \dots < i_p \leq |s|)$ を用いて、 $s[\vec{i}]$ と表す。

2.2.2 最長共通部分文字列 (Longest Common Substring: LCSustr) 長

最長共通部分文字列 (Longest Common Substring) の略称は LCS だが[§]、一般には 2.2.3 節に示す最長共通部分列 (Longest Common Subsequence) のことを LCS と呼ぶことが多い。本稿では前者を LCSustr、後者を LCS と呼び、区別する。

記号列 s, t を与えた際の最長共通部分文字列を次式で定義する：

$$\text{LCSustr}(s, t) = \arg \max_{s_{i \dots i+n-1} | \exists j, s_{i \dots i+n-1} = t_{j \dots j+n-1}} n$$

記号列 s, t を与えた際の最長共通部分文字列長 (LCSustr 長) を次式で定義する：

$$|\text{LCSustr}(s, t)| = \max_{\forall i, \forall j, s_{i \dots i+n-1} = t_{j \dots j+n-1}} n$$

これを $[0,1]$ 区間に正規化すると以下ようになる：

$$\text{Score}_{\text{LCSubstr}}(s, t) = \frac{2 \cdot |\text{LCSubstr}(s, t)|}{|s| + |t|}$$

2.2.3 最長共通部分列 (Longest Common Subsequence: LCS) 長と Levenshtein 距離

記号列 s, t を与えた際の最長共通部分列 (Longest Common Subsequence: LCS) を次式で定義する：

$$\text{LCS}(s, t) = \arg \max_{s[\vec{i}] \exists \vec{j}, s[\vec{i}] = t[\vec{j}]} |\vec{i}|$$

記号列 s, t を与えた際の最長共通部分列長 (LCS 長) を次式で定義する：

$$|\text{LCS}(s, t)| = \max_{\forall \vec{i}, \forall \vec{j}: s[\vec{i}] = t[\vec{j}]} |\vec{i}|$$

$[0,1]$ 区間に正規化すると、以下ようになる：

$$\text{Score}_{\text{LCS}}(s, t) = \frac{2 \cdot |\text{LCS}(s, t)|}{|s| + |t|}$$

なお、挿入のコストを 1、削除のコストを 1、代入のコストを 2（もしくは代入を禁止）とした場合の Levenshtein 距離（編集型距離）と LCS 長の関係は以下ようになる：

$$d_{\text{Levenshtein}}(s, t) = |s| + |t| - 2 \cdot |\text{LCS}(s, t)|$$

さらに LCS は 2.4.2 節で示すとおり、対称群上の編集型距離のうちの Ulam 距離と深く関連し、一種の順序尺度であるとも考えられる。

2.2.4 ギャップ加重最長共通部分列長による指標

部分列 LCS は部分文字列 LCSubstr と異なり、ギャップを伴う。ギャップが多い LCS に減衰させた値を割り当てるために、「LCS の記号列上の長さ」に対して加重を行うことができる。「LCS の記号列上の長さ」は参照テキスト³側 ($|\text{LCS}(C, R)|_R$ で表す) とシステム出力テキスト⁴側 ($|\text{LCS}(C, R)|_C$ で表す) とで異なるために、それぞれ計算する必要がある。

$$|\text{LCS}(C, R)|_R = \arg \max_{(j_{|\vec{j}|} - j_1) | \forall \vec{i}, \forall \vec{j}, C[\vec{i}] = R[\vec{j}]} |\vec{j}|$$

$$|\text{LCS}(C, R)|_C = \arg \max_{(i_{|\vec{i}|} - i_1) | \forall \vec{i}, \forall \vec{j}, C[\vec{i}] = R[\vec{j}]} |\vec{i}|$$

³ 本稿では参照テキストを R と定義する。

⁴ 本稿ではシステム出力テキストを C と定義する。

参照テキスト側で重みを付けて正規化する再現率的な指標を $R_{\text{WLCS}}(C, R)$ とし, システム出力テキスト側で重みを付けて正規化する精度的な指標を $P_{\text{WLCS}}(C, R)$ とすると以下のようなになる.

$$R_{\text{WLCS}}(C, R) = \frac{\alpha^{|\text{LCS}(C,R)|_R - |\text{LCS}(C,R)|} \cdot |\text{LCS}(C, R)|}{|R|}$$

$$P_{\text{WLCS}}(C, R) = \frac{\alpha^{|\text{LCS}(C,R)|_C - |\text{LCS}(C,R)|} \cdot |\text{LCS}(C, R)|}{|C|}$$

全体を正規化すると以下のようなになる.

$$\text{Score}_{\text{WLCS}}^{(\gamma)}(C, R) = \frac{(1 + \gamma^2) R_{\text{WLCS}}(C, R) P_{\text{WLCS}}(C, R)}{R_{\text{WLCS}}(C, R) + \gamma^2 P_{\text{WLCS}}(C, R)}$$

ここで γ は R_{WLCS} と P_{WLCS} のどちらを重視するかの混ぜ合わせ係数である.

2.3 自動評価指標

次に自動要約と機械翻訳の自動評価指標を確認するが, 基本的には文単位の評価かつ参照テキストが一つであるという仮定をおく.

2.3.1 要約の評価指標

ROUGE-L (Lin 2004) は, システム出力テキストと参照テキストの最長共通部分列 (LCS) 長を指標として正規化したものである.

$$\text{Score}_{\text{ROUGE-L}}^{(\gamma)}(C, R) = \frac{(1 + \gamma^2) \cdot R_{\text{LCS}}(C, R) \cdot P_{\text{LCS}}(C, R)}{R_{\text{LCS}}(C, R) + \gamma^2 P_{\text{LCS}}(C, R)}$$

ここで再現率に相当する $R_{\text{LCS}}(C, R)$ と精度に相当する $P_{\text{LCS}}(C, R)$ は以下のように定義する:

$$R_{\text{LCS}}(C, R) = \frac{|\text{LCS}(C, R)|}{|R|}$$

$$P_{\text{LCS}}(C, R) = \frac{|\text{LCS}(C, R)|}{|C|}$$

上記指標は文単位のものであり, 文書レベルに拡張するために, システム出力テキスト中の文 $c_i \in C$ と参照テキスト中の文 $r_j \in R$ の LCS 記号列中の記号の集合和を用いて評価する. 同様の議論が他の指標においても行われているが, 以下本稿ではこの議論を省略する.

ROUGE-W ROUGE-W (Lin 2004) は, ギャップ加重最長共通部分列長に似た概念である. 違いとしては「LCS の記号列上の長さ」を参照テキスト側とシステム出力テキスト側 $|\text{LCS}(C, R)|_R + |\text{LCS}(C, R)|_C$ でとった上で, 加重関数 $f(x) : f(x+y) > f(x) + f(y), x > 0, y > 0, x \in N, y \in N$ (N は自然数) を別に定義して「LCS の記号列上の長さ」に対して加重を行う. ROUGE-W の実

装では $f(x) = x^\alpha$ という多項式を用いており, ギャップ加重最長共通部分列長 $\text{Score}_{\text{WLCS}}^{(\gamma)}(C, R)$ の変種と考えることができる.

ROUGE-N ROUGE-N (Lin and Hovy 2003; Lin 2004) は n-gram の一致度を指標として用いるものである.

$$\text{Score}_{\text{ROUGE-N}}^{(R)}(C, R) = \frac{\sum_{e \in (\text{n-gram}(C) \cap \text{n-gram}(R))} \min(|e|_C, |e|_R)}{\sum_{e \in \text{n-gram}(R)} |e|_R}$$

但し, $\text{n-gram}(C)$ はシステム出力テキスト C 中の n-gram 集合, $\text{n-gram}(R)$ は参照テキスト R 中の n-gram 集合, $|e|_C$ は C に含まれる e の要素数 (のべ出現数), $|e|_R$ は R に含まれる e の要素数とする.

ROUGE-S(U) ROUGE-S (Lin 2004) は, 2-mer の部分列の一致度を指標として用いるものである.

$$\text{Score}_{\text{ROUGE-S}}^{(\gamma)}(C, R) = \frac{(1 + \gamma^2)P_S(C, R)R_S(C, R)}{R_S(C, R) + \gamma^2 P_S(C, R)}$$

ここで精度に相当する $P_S(C, R)$ と再現率に相当する $R_S(C, R)$ は以下のように定義する:

$$P_S(C, R) = \frac{\sum_{e \in (2\text{-mer}_C \cap 2\text{-mer}_R)} \min(|e|_C, |e|_R)}{\sum_{e \in 2\text{-mer}(C)} |e|_C}$$

$$R_S(C, R) = \frac{\sum_{e \in (2\text{-mer}_C \cap 2\text{-mer}_R)} \min(|e|_C, |e|_R)}{\sum_{e \in 2\text{-mer}(R)} |e|_R}$$

但し, $\text{p-mer}(C)$ は C 中の p-mer 集合, $\text{p-mer}(R)$ は参照テキスト R 中の p-mer 集合とする.

ROUGE-SU は上の ROUGE-S の $p = 2$ を $p \leq 2$ に拡張したものである.

ESK ESK (平尾, 奥村, 磯崎 2006) は畳み込みカーネルの一つである拡張文字列カーネルのうち, ギャップ加重 p-mer 部分列カーネルを評価指標として定義したものである.

$$\text{Score}_{\text{ESK}}^{\text{p-mer}}(C, R) = \frac{\sum_{u \in \text{p-mer}(C)} \sum_{v \in \text{p-mer}(R)} \lambda^{|u|-p} \delta(u, v) |u||v|}{\sqrt{\left(\sum_{u, u' \in \text{p-mer}(C)} \lambda^{(|u|-p)} |u||u'| \right) + \left(\sum_{v, v' \in \text{p-mer}(R)} \lambda^{(|v|-p)} |v||v'| \right)}}$$

(平尾 他 2006) では 2-mer の部分列に制限するほか、文単位に比較し精度重視の指標と再現度重視の指標の二つの重みつき調和平均 ($0 \leq \lambda \leq 1$) を定義している。

ESK は他に各形態素に付与される意味ラベルを考慮した評価指標を提案しているが、本稿で用いる ESK は意味ラベルを考慮しない形態素基本形に基づくものとする。

2.3.2 翻訳の評価指標

BLEU BLEU (Papineni, Roukos, Ward, and Zhu 2001) は機械翻訳評価のための指標で、 n の値を変えた n -gram の精度系指標の重み (ω_n) 付き相乗平均により指標を定義する。

$$P_{\text{BLEU}}^{\text{n-gram}}(C, R) = \frac{\sum_{e \in (\text{n-gram}(C) \cap \text{n-gram}(R))} \min(|e|_C, |e|_R)}{\sum_{e \in \text{n-gram}(C)} |e|}$$

$$\text{Score}_{\text{BLEU}}(C, R) = BP(C, R) \cdot \exp\left(\sum_{n=1}^N \omega_n \log P_{\text{BLEU}}^{\text{n-gram}}(C, R)\right)$$

ここで相乗平均の計算を簡単にするために $\sum_{n=1}^N \omega_n = 1$ という制約がある。

短いシステム出力テキストに対して高い精度が出やすいこの精度系の指標に対し、精度と再現率の重み付き調和平均という方法を取らず、Brevity Penalty (BP) という項を入れて補正している。

$$BP(C, R) = \begin{cases} 1 & \text{if } |C| > |R| \\ \exp(1 - \frac{|R|}{|C|}) & \text{if } |C| \leq |R| \end{cases}$$

IMPACT IMPACT(Echizen-ya and Araki 2007) は LCS に基づく指標ではなく、LCSubstr の再帰的な取得による指標である。

$$R_{IP}(C, R) = \left(\frac{\sum_{r=0}^{\text{RN}} (\alpha^r \sum_{e \in \text{LCSubstr}(C^{(r)}, R^{(r)})} |e|^\beta)}{|R|^\beta} \right)^{\frac{1}{\beta}}$$

$$P_{IP}(C, R) = \left(\frac{\sum_{r=0}^{\text{RN}} (\alpha^r \sum_{e \in \text{LCSubstr}(C^{(r)}, R^{(r)})} |e|^\beta)}{|C|^\beta} \right)^{\frac{1}{\beta}}$$

ここで α はイテレート回数 r ($r \leq \text{RN}$) に対する重み ($\alpha < 1.0$)、 β は LCSubstr 長に対する重み ($\beta > 1.0$)、 $C^{(1)} = C$ 、 $R^{(1)} = R$ 、 $C^{(r)} = C^{(r-1)} \setminus \{\text{LCSubstr}(C^{(r-1)}, R^{(r-1)})\}$ 、 $R^{(r)} = R^{(r-1)} \setminus \{\text{LCSubstr}(C^{(r-1)}, R^{(r-1)})\}$ とする。

$$\text{Score}_{\text{IP}} = \frac{(1 + \gamma^2)R_{\text{IP}}P_{\text{IP}}}{R_{\text{IP}} + \gamma^2P_{\text{IP}}}$$

この指標は 2.4.1 節に示す文字列長加重全部分文字列カーネルに関連がある。文字列長加重全部分文字列カーネルに対して、再帰的に LCSubstr を選択する際に既選択の LCSubstr を排除し、再帰の回数を RN で制限するという制約を入れたものである。

RIBES RIBES (平尾, 磯崎, 須藤, Duh, 塚田, 永田 2014) は, システム出力テキストと参照テキストのアラインメントをとったうえで, 語順の編集型順序尺度を考慮したものである。

$$\text{Score}_{\text{RIBES}} = \left(d_{\text{Kendall}}(\text{1-gram}_{\text{align}}(C, R)) \right) \cdot \left(P_{\text{RIBES}}(C, R) \right)^{\alpha} \cdot \left(\text{BP}(C, R) \right)^{\beta}$$

ここで $d_{\text{Kendall}}(\mu, \nu)$ は 2.4.2 節で定義する順位ベクトル μ, ν に対する Kendall 距離, $\text{1-gram}_{\text{align}}(\mu, \nu)$ は元論文 (平尾 他 2014) の **worder** で出力されるアラインメントされた二つの順序ベクトルの対を表す。右辺 2 項目は 1-gram (単語ベースのもの) 精度とよび $P_{\text{RIBES}}(C, R) = \frac{|\text{1-gram}_{\text{align}}(C, R)|}{|C|}$ とする。 $|\text{1-gram}_{\text{align}}(\mu, \nu)|$ は **worder** で出力されるアラインメントされた順序ベクトルの長さ (二つ出力されるが等しい) である。

α は 1-gram 精度に対する重み, β は BLEU で用いられた BP に対する重みである。

なお, $P_{\text{RIBES}}(C, R)$ は, それぞれの記号列に重複する記号がない場合, 以下が成り立つ:

$$P_{\text{RIBES}}(C, R) = \text{Score}_{\text{ROUGE-1}}^{(P)}(C, R) = \frac{\sum_{e \in (\text{1-gram}(C) \cap \text{1-gram}(R))} \min(|e|_C, |e|_R)}{\sum_{e \in \text{1-gram}(R)} |e|}$$

LRscore LRscore (Birch and Osborne 2010) も同様に, アラインメントをとったうえで, 語順の順序尺度を考慮したものである。順序尺度としてベクトル型である Hamming 距離と編集型である Kendall 距離を用いている。

$$\text{Score}_{\text{LRscore}}^{\text{Hamming}}(C, R) = \alpha \cdot \text{BP}(C, R) \cdot d_{\text{Hamming}}(C, R) + (1 - \alpha) \text{Score}_{\text{BLEU}}(C, R)$$

$$\text{Score}_{\text{LRscore}}^{\text{Kendall}}(C, R) = \alpha \cdot \text{BP}(C, R) \cdot d_{\text{Kendall}}(C, R) + (1 - \alpha) \text{Score}_{\text{BLEU}}(C, R)$$

α は語順をどの程度考慮するか重みづけ係数。

2.4 関連するカーネル・順序尺度

上に述べた指標は、基本的には以下のカーネルおよび順序尺度の組み合わせで構成することができる。以下では、各種指標に関連するカーネルおよび順序尺度について確認する。

2.4.1 カーネル・距離（文字列の共有）

畳み込みカーネルのうち系列データに対するカーネルは、共通する部分文字列・部分列を数え上げる。いずれも効率よく計数する方法が提案されている (Shawe-Taylor and Cristianini, 大北 訳 2010)。また、適切に正規化することにより部分文字列・部分列の共有についての距離や指標を規定することができる。

様々なカーネルの説明に入る前に、 $[0,1]$ 区間正規化について示す。カーネルの $[0,1]$ 区間正規化はカーネルの研究分野でよく用いられており以下の式により行われる：

$$\text{Score}_{K_-}(s, t) = \frac{K_-(s, t)}{\|K_-(s, s)\| \cdot \|K_-(t, t)\|}$$

各種指標のように、再現率-精度間の重み γ を入れたい場合には以下のようにする：

$$\text{Score}_{K_-}^{(\gamma)}(s, t) = \frac{(1 + \gamma^2)K_-(s, t)}{\sqrt{(K_-(s, s))^2 + \gamma^2(K_-(t, t))^2}}$$

全部分文字列カーネルと文字列長加重全部分文字列カーネル 全部分文字列カーネル (All String Kernel or Exact Matching Kernel) は共通する全ての部分文字列の数を数える。

任意の長さの部分文字列 u の出現数を座標軸とする特徴量空間 $F_{\text{all_str}}$ を考える。

$$\begin{aligned} \Phi_{\text{str}}^* : \sigma^* &\rightarrow F_{\text{all_str}} \sim R^{|\sigma|^*} \\ \Phi_{\text{str}}^*(s) &= (\phi_u^*(s))_{u \in \sigma^*} \\ \phi_u^*(s) &= |\{i | s_{i...*} = u\}| \\ K_{\text{all_str}}(s, t) &= \langle \Phi_{\text{str}}^*(s), \Phi_{\text{str}}^*(t) \rangle_{F_{\text{all_str}}} \\ &= \sum_{u \in \sigma^*} \phi_u^*(s) \phi_u^*(t) \end{aligned}$$

カーネル関数を直接計算すると以下ようになる：

$$K_{\text{all_str}}(s, t) = \sum_{n=1}^{\min(|s|, |t|)} \sum_{i=1}^{|s|-n+1} \sum_{j=1}^{|t|-n+1} \delta(s_{i...i+n-1}, t_{j...j+n-1})$$

ここで δ はクロネッカーのデルタとする。言語処理の場合、得られる n -gram に対して加重をかけることが一般に行われている。文字列長に対して加重をかけたものを文字列長加重全部分文

字列カーネル (Length Weighted All String Kernel or Length Weighted Exact Matching Kernel) と呼ぶ.

$$K_{\text{w_all_str}}(s, t) = \sum_{n=1}^{\min(|s|, |t|)} \sum_{i=1}^{|s|-n+1} \sum_{j=1}^{|t|-n+1} \omega_{|s|} \delta(s_{i\dots i+n-1}, t_{j\dots j+n-1})$$

ここで ω_n は長さ n に対する重みを表す.

2.3.2 節で述べた IMPACT はこのカーネルの特殊形とみなすことができる.

このカーネルと次の n -スペクトラムカーネルは Suffix Tree を用いて効率よく計算する方法が提案されている.

n -スペクトラムカーネル n -スペクトラムカーネル (Spectrum Kernel) は共通する長さ n の部分文字列 (n -gram) の数を数える.

長さ n の部分文字列 u の出現数を座標軸とする特徴量空間 $F_{n\text{-gram}}$ を考える.

$$\begin{aligned} \Phi_{\text{str}}^n : \sigma^n &\rightarrow F_{n\text{-gram}} \sim R^{|\sigma|^n} \\ \Phi_{\text{str}}^n(s) &= (\phi_u^n(s))_{u \in \sigma^n} \\ \phi_u^n(s) &= |\{i | s_{i\dots i+n-1} = u\}| \\ K_{n\text{-gram}}(s, t) &= \langle \Phi_{\text{str}}^n(s), \Phi_{\text{str}}^n(t) \rangle_{F_{n\text{-gram}}} \\ &= \sum_{u \in \sigma^n} \phi_u^n(s) \phi_u^n(t) \end{aligned}$$

直接計算すると以下のようになる :

$$K_{n\text{-gram}}(s, t) = \sum_{i=1}^{|s|-n+1} \sum_{j=1}^{|t|-n+1} \delta(s_{i\dots i+n-1}, t_{j\dots j+n-1})$$

ROUGE-N は, 分子に $K_{n\text{-gram}}(C, R)$ より小さい値を持ち, 分母に参照テキストののべ出力 n -gram 数を持つことから, 再現率として正規化したものに相当する. 通常 of 正規化した $K_{n\text{-gram}}(s, t)$ は再現率と精度の調和平均と解釈できる.

また 1-gram スペクトラムカーネルは 1-mer 部分列カーネルと同値で, これらは近似的に BLEU などと利用されている BP 相当の値を計算すると考える.

全部分列カーネル 全部分列カーネルは共通するすべての部分列の数を数える.

任意の長さの部分列 v の出現数を座標軸とする特徴量空間 $F_{\text{all_seq}}$ を考える.

$$\begin{aligned} \Psi_{\text{seq}}^* : \sigma^* &\rightarrow F_{\text{all_seq}} \sim R^{|\sigma|^\infty} \\ \Psi_{\text{seq}}^*(s) &= (\psi_v^*(s))_{v \in \sigma^*} \end{aligned}$$

$$\begin{aligned}\psi_v^*(s) &= |\{\vec{i} | s[\vec{i}] = v\}| \\ K_{\text{all_seq}}(s, t) &= \langle \Psi_{\text{seq}}^*(s), \Psi_{\text{seq}}^*(t) \rangle_{F_{\text{all_seq}}} \\ &= \sum_{v \in \sigma^*} \psi_v^*(s) \cdot \psi_v^*(t)\end{aligned}$$

$K_{\text{all_seq}}(s, t)$ は以下のように再帰的に計算することにより $O(|s||t|)$ で計算することができる. ϵ を空記号列とすると $K_{\text{all_seq}}(s, \epsilon) = K_{\text{all_seq}}(t, \epsilon) = 1$ とし, $K_{\text{all_seq}}(s, t)$ が求まると $K_{\text{all_seq}}(s \cdot a, t) = K_{\text{all_seq}}(s, t) + \sum_{1 \leq i \leq |t|, j: t_j = a} K_{\text{all_seq}}(s, t_{i \dots j-1})$ と s 再帰的に定義できる. さらに $\tilde{K}_{\text{all_seq}}(s \cdot a, t) = K_{\text{all_seq}}(s, t_{i \dots j-1})$ とすると, $\tilde{K}_{\text{all_seq}}(s \cdot a, t \cdot b) = \tilde{K}_{\text{all_seq}}(s \cdot a, t) + \delta(a, b)K_{\text{all_seq}}(s, t)$ と t 再帰的に定義できる.

固定長部分列カーネル 固定長部分列カーネルは共通する長さ p の部分列 (p-mer) の数を数えあげる.

長さ p の部分文字列 v の出現数を座標軸とする特徴量空間 $F_{\text{p-mer}}$ を考える.

$$\begin{aligned}\Psi_{\text{seq}}^p : \sigma^p &\rightarrow F_{\text{p-mer}} \sim R^{|\sigma|^p} \\ \Psi_{\text{seq}}^p(s) &= (\psi_v^p(s))_{v \in \sigma^p} \\ \psi_v^p(s) &= |\{\vec{i} | s[\vec{i}] = v\}| \\ K_{\text{p-mer}}(s, t) &= \langle \Psi_{\text{seq}}^p(s), \Psi_{\text{seq}}^p(t) \rangle_{F_{\text{p-mer}}} \\ &= \sum_{v \in \sigma^p} \psi_v^p(s) \cdot \psi_v^p(t)\end{aligned}$$

ROUGE-S は, 分子に $K_{2\text{-mer}}(C, R)$ より小さい値を持ち, 分母に参照テキストののべ出力 2-mer 数を持つことから, 再現率として正規化したものに相当する. ROUGE-SU は, 分子に $K_{1\text{-mer}, 2\text{-mer}}(C, R)$ より小さい値を持ち, 分母に参照テキストののべ出力 1-mer, 2-mer 数を持つことから, 再現率として正規化する. 通常正規化した $K_{\text{p-mer}}(s, t)$ は再現率と精度の調和平均と解釈できる.

ギャップ加重部分列カーネル ギャップ加重部分列カーネルは p-mer の部分列の数え上げの際に隣接性を考慮して重み λ を加重する. ESK (平尾 他 2006) は, このカーネルを用いた尺度である.

長さ p の部分列 v を座標とする特徴量空間 $F_{\text{p-mer}}$ を考える.

$$\begin{aligned}K_{\text{gap-p-mer}}(s, t) &= \langle \Psi_{\text{seq}}^{\text{gap-p}}(s), \Psi_{\text{seq}}^{\text{gap-p}}(t) \rangle_{F_{\text{p-mer}}} \\ &= \sum_{v \in \sigma^p} \psi_v^{\text{gap-p}}(s) \cdot \psi_v^{\text{gap-p}}(t)\end{aligned}$$

ここで $\psi_v^{\text{gap-p}}(s) = \sum_{\vec{i}: v=s[\vec{i}]} \lambda^{l(\vec{i})}$ とし, $l(\vec{i}) = |s_{i_1 \dots i_{|v|}}| (i = \langle i_1, \dots, i_{|v|} \rangle)$ とする.

2.4.2 順序尺度

以下では順序尺度について考えるが, 文献 (神嶋 2009) に詳しい解説がある. 基本的には同じ長さ m の二つの順位ベクトル $\mu, \nu \in S_m$ に対する 2 種類の距離を考える.

順位ベクトル型距離 一つ目の距離は「順位ベクトル型」の距離で順位ベクトルを m 次元空間中の点を表すベクトルとみなし, ベクトル空間上の距離を定義する. ベクトル空間上の θ -ノルムを用いると以下のようになる:

$$d_{||\text{Rank}||_\theta}(\mu, \nu) = \left(\sum_{i=1}^m |\mu(i) - \nu(i)|^\theta \right)^{1/\theta}$$

ここで $\theta = 1$ の場合, 特に Spearman footrule と呼ぶ.

$$d_{\text{Footrule}}(\mu, \nu) = \left(\sum_{i=1}^m |\mu(i) - \nu(i)| \right)$$

$\theta = 2$ の場合は通常の Euclid 距離だが⁵, この Euclid 距離を 2 乗したものを特に Spearman 距離と呼ぶ.

$$d_{\text{Spearman}}(\mu, \nu) = \left(\sum_{i=1}^m |\mu(i) - \nu(i)|^2 \right)$$

Spearman 距離は, 距離の公理のうち対称性と正定値性を満たす. しかし, Euclid 距離を 2 乗したもののなので三角不等式を満たさないが, 慣習的に距離として扱われる. さらに $[-1, 1]$ 区間に正規化したものは Spearman の順位相関係数 ρ として知られている.

$$\text{Spearman's } \rho = 1 - \frac{6 \cdot d_{\text{Spearman}}(\mu, \nu)}{m^3 - m}$$

この値は順序尺度に基づく二つの順位ベクトル μ, ν の Pearson 相関係数と等しい⁵.

その他, 順位ベクトルの同一順位のものと同じ要素である要素数を数えた Hamming 距離がある.

$$d_{\text{Hamming}}(\mu, \nu) = \sum_{i=1}^m \delta(\mu(i), \nu(i))$$

Hamming 距離は文字列上で代入 (コスト 1) のみを許した編集距離としても解釈できる.

また, 距離 $d_{||\text{rank}||_\theta}$, d_{Footrule} , d_{Spearman} , d_{Hamming} に対応するスコア $\text{Score}_{||\text{rank}||_\theta}$, $\text{Score}_{\text{Footrule}}$, $\text{Score}_{\text{Spearman}}$, $\text{Score}_{\text{Hamming}}$ を次のように規定することができる.

$$\text{Score}_- = \frac{1}{1 + d_-}$$

⁵ ここで順序尺度とは, 間隔に意味がある間隔尺度を順位のみに変換していることを前提にしている.

対称群上の編集型距離 二つ目の距離は「編集型」の距離である.

順序ベクトルを記号列とみなした場合, 順位ベクトル μ をもうひとつの順位ベクトル ν に変換するために必要な最小操作数を意味する Levenshtein 距離について述べた. 以下では, 順序ベクトルを対称群とみなした場合の編集型距離について述べる. 編集に許される操作によっていくつかの距離のバリエーションがある.

- Kendall 距離 :

Kendall 距離 $d_{\text{Kendall}}(\mu, \nu)$ は順序ベクトルを対称群とみなした際に隣接互換で置換する最小回数によって定義される. 言い換えると隣接する対象対を交換 (Swap) する操作の最小回数をういたものである. Kendall 距離は, 二つの順位ベクトル中の $\frac{m(m-1)}{2}$ 個の対象対のうち逆順になっている対の数に等しい.

$$d_{\text{Kendall}}(\mu, \nu) = \min(\arg \max_{\bar{q}} \delta((\Pi_{q=1}^{\bar{q}} \pi_2(k_q, k_q + 1)) \cdot \mu, \nu))$$

$$d_{\text{Kendall}}(\mu, \nu) = \sum_{i=1}^m \sum_{j=i+1}^m \chi(i, j)$$

ここで, k_q は順位ベクトル μ のインデックス. また, χ は対象対 $\langle i, j \rangle$ が同順のとき 0, 逆順のとき 1 を返す指示関数 :

$$\chi = \begin{cases} 1 & \text{if } (\mu(i) - \mu(j))(\nu(i) - \nu(j)) < 0, \\ 0 & \text{if } (\mu(i) - \mu(j))(\nu(i) - \nu(j)) \geq 0 \end{cases}$$

$\pi_2 = (i, i+1)$ は隣接する二つの元のみを入れ替えて他の元は変えない操作である隣接互換を意味する.

これを指標として使いやすくするために $[0, 1]$ 区間の範囲に正規化すると以下のようになる :

$$\text{Score}_{\text{Kendall}}(\mu, \nu) = 1 - \frac{2 \cdot d_{\text{Kendall}}(\mu, \nu)}{m^2 - m}$$

これを $[-1, 1]$ 区間の範囲に正規化したものは Kendall の順位相関係数 τ として知られている.

$$\text{Kendall's } \tau = 1 - \frac{4 \cdot d_{\text{Kendall}}(\mu, \nu)}{m^2 - m}$$

- Cayley 距離 :

Cayley 距離 d_{Cayley} は順序ベクトルを対称群とみなした際に互換で置換する最小回数によって定義される. 言い換えると隣接していなくても良い対象対を交換 (Swap) する最小回数をういたものである.

$$d_{\text{Caylay}}(\mu, \nu) = \min(\arg \max_{\bar{q}} \delta((\prod_{q=1}^{\bar{q}} \pi_2(k_q, l_q)) \cdot \mu, \nu))$$

ここで, k_q, l_q は順位ベクトル μ のインデックス. $\pi_2 = (i, j)$ は二つの元のみを入れ替えて他の元は変えない操作である互換を意味する.

- Ulam 距離 :

Ulam 距離 d_{Ulam} は順序ベクトルを対称群とみなした際に連続した順序ベクトル部分列 $\langle i, i+1, \dots, j-1, j \rangle$ の巡回置換の操作のみで置換する最小回数によって定義される. これは「本棚の本の入れ換え」で例えられる. 順位ベクトル μ で並んでいる本棚の本を順位ベクトル ν に並べ替えるために, ある要素を抜いて別の場所に挿入するというを行う.

Ulam 距離は同じ要素が記号列に存在しないという前提のもと, 最長一致部分列長と以下の関係にあることが知られている.

$$d_{\text{Ulam}}(\mu, \nu) = m - |\text{LCS}(\mu, \nu)|$$

これを $[0,1]$ 区間の範囲に正規化すると以下のように正規化最大共通部分列と同じになる :

$$\begin{aligned} \text{Score}_{\text{Ulam}}(\mu, \nu) &= 1 - \frac{d_{\text{Ulam}}(\mu, \nu)}{m} \\ &= \frac{|\text{LCS}(\mu, \nu)|}{m} \\ &= \text{Score}_{\text{LCS}}(\mu, \nu) \end{aligned}$$

図 1 に順序ベクトルによる置換により表現した編集型距離の例を示す. 編集型距離の定義で許される編集の回数を数えると, 順序ベクトル $(1, 4, 3, 2)$ と $(1, 2, 3, 4)$ の Kendall 距離は 3, Caylay 距離は 1, Ulam 距離は 2 となる. また, 順序ベクトル $(2, 3, 1, 4)$ と $(1, 2, 3, 4)$ の Kendall 距離は 2, Caylay 距離は 2, Ulam 距離は 1 となる.

以下は, 我々の意見だが, 言語生産時の編集作業の工数を評価する場合には, (Nivre 2009) の swap に代表されるような Kendall 距離のような編集よりも Ulam 距離のような編集を考慮すべきであると考え. 言語生産時に, Kendall 距離で考慮される列内絶対位置よりも, Ulam 距離で考慮される列内相対位置を考えながら編集を行う方が人にとって自然な処理であると考え. **順序尺度間の関係** ベクトル型の Spearman's ρ と Kendall's τ との間には以下の Daniels の不等式が成立する :

$$-1 \leq \frac{3(m+2)}{m-2}\tau - \frac{2(m+1)}{m-2}\rho \leq 1$$

$m \rightarrow \infty$ の極限をとると $-1 \leq 3\tau - 2\rho \leq 1$ が成り立つ. このことから二つの相関係数の間の

Kendall

$d_{\text{Kendall}}((1, 4, 3, 2), (1, 2, 3, 4)) = 3$

$$\begin{pmatrix} 1 & 4 & \underline{3} & \underline{2} \\ 1 & 4 & \underline{2} & \underline{3} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & \underline{4} & \underline{2} & 3 \\ 1 & \underline{2} & \underline{4} & 3 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 2 & \underline{4} & \underline{3} \\ 1 & 2 & \underline{3} & \underline{4} \end{pmatrix}$$

$d_{\text{Kendall}}((2, 3, 1, 4), (1, 2, 3, 4)) = 2$

$$\begin{pmatrix} 2 & \underline{3} & \underline{1} & 4 \\ 2 & \underline{1} & \underline{3} & 4 \end{pmatrix} \Rightarrow \begin{pmatrix} \underline{2} & \underline{1} & 3 & 4 \\ \underline{1} & \underline{2} & 3 & 4 \end{pmatrix}$$

Caylay

$d_{\text{Caylay}}((1, 4, 3, 2), (1, 2, 3, 4)) = 1$

$$\begin{pmatrix} 1 & \underline{4} & 3 & \underline{2} \\ 1 & \underline{2} & 3 & \underline{4} \end{pmatrix}$$

$d_{\text{Caylay}}((2, 3, 1, 4), (1, 2, 3, 4)) = 2$

$$\begin{pmatrix} \underline{2} & 3 & \underline{1} & 4 \\ \underline{1} & 3 & \underline{2} & 4 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & \underline{3} & \underline{2} & 4 \\ 1 & \underline{2} & \underline{3} & 4 \end{pmatrix}$$

Ulam

$d_{\text{Ulam}}((1, 4, 3, 2), (1, 2, 3, 4)) = 2$

$$\begin{pmatrix} 1 & \boxed{4} & \boxed{3} & \underline{2} \\ 1 & \underline{2} & \boxed{4} & \boxed{3} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 2 & \boxed{4} & \boxed{3} \\ 1 & 2 & \underline{3} & \boxed{4} \end{pmatrix}$$

$d_{\text{Ulam}}((2, 3, 1, 4), (1, 2, 3, 4)) = 1$

$$\begin{pmatrix} \boxed{2} & \boxed{3} & \underline{1} & 4 \\ \underline{1} & \boxed{2} & \boxed{3} & 4 \end{pmatrix}$$

図 1 対称群上の編集型距離

相関が高いことが示される。

距離の観点からは, $d_{\text{Caylay}} \leq d_{\text{Kendall}}$ が成り立つ。さらに Footrule 距離と Kendall 距離と Cayley 距離の間に以下の不等式が成り立つ (Diaconis-Graham inequality):

$$d_{\text{Kendall}} + d_{\text{Caylay}} \leq d_{\text{Footrule}} \leq 2 \cdot d_{\text{Kendall}}$$

また Spearman 距離と Kendall の距離の間には以下の不等式が成り立つ (Durbin-Stuart inequality):

$$\frac{4}{3} d_{\text{Kendall}} \left(1 + \frac{d_{\text{Kendall}}}{m} \right) \leq d_{\text{Spearman}}$$

つまり, 評価指標のデザインにおける順序尺度の選択による差異は, これらの不等式の範囲によって制限される.

2.5 指標の一般化

以上, 評価指標・距離・カーネル・相関係数を議論してきた. まとめて付記 B 表 2 のようになる.

各指標と人手の評価結果をできるかぎり合わせるという観点からすると, (平尾, 奥村, 安田, 磯崎 2007) のように, 表 2 にあげたすべての尺度 $\text{Score}_- \in \{\text{Score}_*\}$ の加重相乗平均 (下式) を考え, 加重 ω_- と各指標に付随するパラメータを, 各指標の従属性や相関に注意しながら人手の評価結果との回帰により求めれば良い.

$$\overline{\text{Score}_*} = \sum \omega_- \sqrt{\Pi \text{Score}_-^{\omega_-}}$$

$$\log \overline{\text{Score}_*} = \frac{1}{\sum \omega_-} \left(\sum \omega_- \cdot \log \text{Score}_- \right)$$

この指標のあり方については注意すべき点はいくつかある.

- substring (部分文字列: n-gram 系) と subsequence (部分列: p-mer 系) との違いを踏まえる.
- 最長一致部分長は対称群上の編集型距離である Ulam 距離と深く関連する.
- 順序に対する順位ベクトル型距離と編集型距離の間には 2.4.2 節に示される関係が成り立つ.

本稿では, 先に述べた四つの尺度がそれぞれどのような特性があるかを明らかにすることを目的としており, 最適な指標の組み合わせについては検討を行わない. 次節以降, 各尺度がさまざまな言語資源上でどのようなふるまいをするのかについてみていきたい.

3 評価に用いる言語資源

ここでは様々な言語生成過程を記録した言語資源におけるテキスト対の尺度の差異を検証することにより, 各尺度がとらえようとしているものが何なのかを分析する.

表 1 に, 利用する言語資源について示す. まず言語生産過程として, 要約 (BCCWJ-SUMM) と語釈 (GLOSS) と再話 (RETELLING) の 3 種類の言語資源を用いる. 要約は長い元文書を短くする情報提示手法である. 語釈は短い単語を長い文書で説明する情報提示手法である. 再話は長い元文書をできるだけその内容を保存したまま示す情報提示手法である. 情報提示手法を比較することで, 各尺度が何を評価しているのかを明らかにすることを試みる.

要約と語釈については, クラウドソーシングにより安価で大量にデータを得る手法 (タイプ

表 1 指標評価に使う言語資源

言語資源名	収集場所	生成過程	繰り返し	取得人数	摘要
BCCWJ-SUMM_C	クラウドソーシング	タイプ入力	なし	100-200	19 文書の要約
BCCWJ-SUMM_L	実験室	筆述	3 回	のべ 47	8 文書の要約
GLOSS_C	クラウドソーシング	タイプ入力	なし	71・111・113	鶏・兎・象の語釈
GLOSS_L	実験室	筆述	4 回	7・6・3	鶏・兎・象の語釈
RETELLING_I	実験室	口述	10 回	5	インタビュー
RETELLING_K	実験室	口述	3 回	3・3・3	怪談 3 種の再話
RETELLING_M	実験室	筆述	4 回	10	物語「桃太郎」の再話

入力)と実験室にて被験者に繰り返し同一課題を依頼してデータを得る手法(筆述)の2種類の方法を用いた。再話のデータについては既存のデータを用い、筆述による形態と口述による形態のデータを準備した。言語生産形態として、タイプ入力・筆術・口述の3種類のデータを対照比較する。これは評価尺度が、要約の内容の類似度だけでなく、個人の文体の類似度を評価してしまう部分を分析するために準備した。それぞれ文体の統制が可能なレベルが異なっており、評価尺度に影響を与えるものだと考え、これを評価することを試みる。さらに、大勢の実験協力者に同じタスクを行わせる場合の協力者間の尺度のふるまいと、同一の実験協力者に同じタスクを複数回行わせる場合の尺度のふるまいを検証し、どの尺度にゆれが生じるかを明らかにする。

以下各言語資源について解説する。

3.1 BCCWJ-SUMM_C

BCCWJ-SUMM_Cは『現代日本語書き言葉均衡コーパス』(Maekawa, Yamazaki, Ogiso, Maruyama, Ogura, Kashino, Koiso, Yamaguchi, Tanaka, and Den 2014)(BCCWJ)の新聞記事(PN サンプル)の要約をYahoo! クラウドソーシング(15歳以上の男女)により被験者実験的に作成したもの(浅原, 杉, 柳野 2015)である。

BCCWJの1サンプルには複数の記事が含まれており、それを記事単位に分割したうえで元文書集合19文書を構築した。元文書集合はBCCWJ コアデータ PN サンプル(優先順位A)から選択した。40文字毎に改行した元文書を画像として提供し、実験協力者に50-100文字に要約せよという指示で収集した。自動要約の本来のあり方としては、文字数の削減ではなく、読み手の読み時間の削減が本質であると考えるが、実験の都合上、文字数による制限を課した。実験協力者の環境はPCに限定した。元文書毎に約100-200人の実験協力者が要約に従事した。実験実施時期は2014年9月である。

得られたデータには、文字数制限を守っていないもの・実験の趣旨を理解していないもの・既に実験を行った実験協力者から同一回答を提供されたと考えられるものなどが含まれており、

これらを排除したものを有効要約とする。統計分析においてこの有効要約のみを用いる。

3.2 BCCWJ-SUMML

BCCWJ-SUMMLはBCCWJの新聞記事の要約を実験室環境で筆述により作成したもの(浅原他 2015)である。BCCWJ-SUMML_Cで用いた元文書を印刷紙面で提供し、実験協力者に50–100文字に要約せよという指示で収集した。一つの元文書に対して、3回まで繰り返して要約文作成を行った。実験協力者は1回の要約文作成に10分間の時間制限を設定した。各回の間には休憩時間をおかず、早く要約課題が完成した場合には、ただちに次の回の要約作成を行った。尚、各実験協力者は要約対象文書を含む文書群の読文時間を評価する実験を、本実験の前に行っており、要約文作成前に元新聞記事を1回読んでいた。今後、読文時間と要約抽出箇所との評価を進める予定である。繰り返しの際には、特別に「前と同じ要約文を作成してください」といった指示は行わず、質問された場合にも「自由に要約文を作成してください」と教示した。実験協力者は原稿用紙上で筆述(鉛筆と消しゴム利用)で要約を行い、そのデータを電子化した。

本実験の実験参加者は要約作業前に要約元文書の読み時間のデータも取得した。さらに被験者の特性(最終学歴・語彙数・言語形成地・記憶力)などのデータについても取得した。実験実施時期は2014年8月–10月であるが、今後このデータは引き続き拡充していく予定である。

統計分析においては、同一課題について、異なる被験者間のスコア(1回目のみを評価:BCCWJ-SUMML(P))と、同一被験者の回数間のスコア(BCCWJ-SUMML(T))の両方を評価する。

3.3 GLOSS_C

GLOSS_Cは語釈文をYahoo!クラウドソーシング(15歳以上の男女)により被験者実験的に作成したものである。実験実施時期は2014年2月である。

「その動物を全く知らない人がどのようなものかわかるように説明してください」と教示し、同意した実験協力者は兎(単語親密度6.6)・鶏(同6.4)・象(同6.0)の3種類から対象物を選択回答した。単語親密度は(天野, 近藤 1999)による。150文字以上250文字以内で3文字以上の同文字連続は認めない設定とした。実験協力者300名を募集したところ得られた解答数は、鶏:71・兎:111・象:113(295/300)であった。

尚、このデータの質的分析は(加藤 2015b)を参照されたい。

3.4 GLOSS_L

GLOSS_Lは語釈文を実験室環境で筆述により収集したものである。実験実施時期は2013年5月–6月である。

実験協力者8名(20代–50代の男女)に、GLOSS_Cと同様に「その動物を全く知らない人

がどのようなものかわかるように説明してください」と教示した。実験協力者は、10分間で兎（単語親密度 6.6）・鶏（同 6.4）・象（同 6.0）の3種類から2種類の対象物を選択回答した。目安として5分経過時にブザー音を鳴らした。選択した対象物について同様に記述を繰り返すことを4回行った。得られた解答数は、兎7人分×4回、鶏6人分×4回、象3人分×4回である。平均145文字（max 227文字, min 85文字）を得た。

統計分析においては、同一課題について、異なる被験者間のスコア（1回目のみを評価：GLOSS_L(P)）と、同一被験者の回数間のスコア（GLOSS_L(T)）の両方を評価する。

尚、このデータを用いた認知的な分析は（加藤 2015a; 加藤, 浅原 2015）を参照されたい。

3.5 RETELLING_I

最初の再話のデータは「独話 Retelling コーパス」（保田, 田中, 荒牧 2013a, 2013b）である。このコーパスは（宮部, 四方, 久保, 荒牧 2014）でも用いられている。

実験協力者は5名で、同一人が同内容をそれぞれ10回独話を繰り返した。就職活動を前提とした模擬面接の設定で、実験協力者は自ら予め用意した「学生生活で力を入れてきたこと（3分間程度）」についての独話を行った。同内容を繰り返すことや何回依頼するかは知らせていない。5人分×10回（50話分）の独話を取得した。面接官（聴衆）は有無を交互とした。奇数回（1・3・5・7・9回）は聴衆なしの独話、偶数回（2・4・6・8・10回）は聴衆に対する独話である。聴衆には、聴いていることを表すために頷くことのみを許可しており、話者への質問や意見など、発話は一切行わなかった。収録は録音と録画を行い、音声データを書き起こした。

被験者によってインタビュー内容が異なるために、統計分析においては同一被験者の回数間のスコア（RETELLING_I(T)）のみを評価する。

3.6 RETELLING_K

次の再話のデータは怪談を繰り返し口述したものであり、先行研究（保田, 荒牧 2012）によるものである。

実験協力者は3名⁶で、実験は1名ずつ個別に行った。実験協力者は怪談を聞いたのち、その怪談について3回の再話を行った。怪談は3種類を用意したため、各人9回の語りを行った。語りに関しては、「怪談として他の人に伝えるよう話す」との指示をした。既存の物語では、個人の記憶による先入観の影響が予測されたため、4分間程度の新規な怪談を3本作成した。

実験環境は図2のように、ビデオカメラと録音機により、録音と録画を行った。聴衆の影響を除去するために、聴衆は設置しなかった。実験協力者は以下の配置で録音機に向かって話した。

本稿では音声データを書き起こしたものをを用いる。

⁶ 実験協力者1 20代・女性・東京都, 実験協力者2 30代・女性・茨城県, 実験協力者3 20代・女性・神奈川県

図 2 RETELLING_K データの収録環境

統計分析においては, 同一課題について, 異なる被験者間のスコア (1 回目のみを評価: RETELLING_K(P)) と, 同一被験者の回数間のスコア (RETELLING_K(T)) の両方を評価する.

3.7 RETELLING_M

最後の再話のデータは桃太郎の物語を筆述で繰り返し記述したものであり, 先行研究 (保田 2014) によるものである.

実験協力者 10 名 (20 代–50 代の男女) に, 「桃太郎の物語を全く知らない人に向けて記述してください」と教示し, 実験協力者は 10 分間で記述 (筆述) した. 同様に記述を繰り返すことを 4 回行った. 平均のべ 284 語 (min: 150 語 · max: 451 語), 異なり語 107 語 (min: 74 語 · max: 152 語) の「桃太郎」10 人分 × 4 回 (40 話分) を取得した.

統計分析においては, 同一課題について, 異なる被験者間のスコア (1 回目のみを評価: RETELLING_M(P)) と, 同一被験者の回数間のスコア (RETELLING_M(T)) の両方を評価する.

4 尺度の定性的な分析

4.1 尺度の分析方法

本節では前節で述べたコーパスを用いて各尺度がどのように振る舞うかを観察する. 利用する尺度は以下の 30 種類である.

- n-gram スペクトラム (1,2,3,4) (char/mrph)
- n-gram 以下スペクトラム ($\leq 2, \leq 3, \leq 4$) (char/mrph)
- p-mer 部分列 (2,3,4) (char/mrph)
- p-mer 以下部分列 ($\leq 2, \leq 3, \leq 4$) (char/mrph)
- 1-gram スペクトラム+Footrule (char/mrph) (=Spearman)
- 1-gram スペクトラム+Kendall (char/mrph)

付記 C 表 4, 5 に各コーパス中の 2 サンプル間の尺度の平均値 (Mean) と標準偏差 (SD)⁷を示す. スコアについて (char)“_c” は文字単位の記号列として評価したもの, (mrph)“_m” は形態

⁷ サンプル対に規定する同値類内全組み合わせに対する算術平均. 実験室において複数回実施したコーパスについては, 回数を固定した場合 (T) と, 被験者を固定した場合 (P) と部分集合群を規定し, 各部分集合中で全組み合わせに対する算術平均を得た. 標準偏差も同様.

素単位の記号列 (MeCab-0.98+IPADIC-2.7.0 による) として評価したものである。括弧内の数字は部分文字列長 (n-gram における n) もしくは部分列長 (p-mer における p) を示す。シャピロ・ウィルク検定の結果, ほとんどの場合 p 値が 0.05 未満であり, 正規分布とはいえない傾向が見られた。

4.2 尺度のグラフ

図 3 に形態素単位に評価した, n-gram(1), n-gram(2), p-mer(2), Kendall の尺度のタスク毎の平均値グラフを示す。エラーバーは標準誤差を表す。

unigram(n-gram(1)) を用いた場合, 要約と語釈は中程度, 再話はかなり高い値である。GLOSS-L(T) がほぼ再話と同程度の値である一方, BCCWJ-SUMML(T) が低いことから, 要約を繰り返す際の言語生産の特殊性が見られる。要約を繰り返す際には, 回数毎に文章中の重要箇所を変更するサンプル・被験者が存在し, 標準偏差も高くなっている。

bigram(n-gram(2)), skip-bigram(p-mer(2)) を用いた場合, 異なる被験者間と繰り返し間との間に差が見られるようになる。これは何らかの個人の文体差が形態素の接続に影響を与えているのではないかと考える。

bigram(n-gram(2)) と skip-bigram(p-mer(2)) の間の差として, 語釈の場合のみ bigram の値が下がった。語釈という課題の性質上, 物語や要約と異なり, 情報の提示順が変わることも考え

図 3 課題と評価尺度 (n-gram(1), n-gram(2), p-mer(2), Kendall: 形態素単位 平均値と標準誤差)

られる。しかし、順序尺度である Kendall の値では bigram の値ほど顕著な差が見られなかった。単語の隣接性が語釈のみ下がるという値のふるまいについては今後検討していきたい。

クラウドソーシングと研究室内被験者実験との差 (BCCWJ-SUMM.C $\Leftrightarrow$ BCCWJ-SUMM.L(P), GLOSS.C $\Leftrightarrow$ GLOSS.L(P)) については、各尺度・各課題 (要約・語釈) で差が見られなかった。

4.3 課題間の評価

以下、課題間を比較するために、6 種類の評価軸を分析する。ほとんどの場合、正規分布であることも等分散であること (F 検定による) も仮定できない。ここではウィルコクソンの順位和検定 (0.05 未満で 2 群の代表値が左右にずれている) を行う⁸。多重比較に対応するために Bonferroni 法を用いた。付記 C 表 3 に結果のまとめを示す。

- 実験室における複数人の課題間の違いの評価

BCCWJ-SUMM.L(P) $\Leftrightarrow$ GLOSS.L(P) $\Leftrightarrow$ RETELLING.K(P) $\Leftrightarrow$ RETELLING.M(P)

- BCCWJ-SUMM.L(P) $\Leftrightarrow$ GLOSS.L(P)

文字単位の評価の場合 n-gram(2,3,4).char に有意差が見られた。

形態素単位の評価の場合 n-gram(2,3,4).morph に有意差が見られた。

- BCCWJ-SUMM.L(P) $\Leftrightarrow$ RETELLING.K(P)

n-gram(4).char, n-gram(3,4).morph 以外で有意差が見られた。

- BCCWJ-SUMM.L(P) $\Leftrightarrow$ RETELLING.K(M)

n-gram(4).morph 以外で有意差が見られた。

- GLOSS.L(P) $\Leftrightarrow$ RETELLING-{K,M}(P)

全ての尺度について、有意差が見られた。

- RETELLING.K(P) $\Leftrightarrow$ RETELLING.M(P)

全ての尺度について、有意差が見られなかった。

要約 $\Leftrightarrow$ 語釈間は n-gram(1) で有意差が見られなかった。同じ文字・同じ形態素を使うという観点では一致度のレベルが等しいが、語の接続が入ると有意差が見られることがわかった。グラフからは語釈の方が語の接続や順序尺度の一致度が低い。これは語釈の目的としては情報の提示順に重要性のないことが伺える。

要約 $\Leftrightarrow$ 再話、語釈 $\Leftrightarrow$ 再話の間においては有意差が見られた。再話は同じ話をするという特性から一致度が高くなる一方、要約・語釈は目的を達成するがために同じ表現を用いなければならないという制約がなく、一致度が低くなる傾向にある。また同一課題の語釈間では有意差はなかった。

⁸ コルモゴロフ=スミルノフ検定 (0.05 未満で 2 群は異なる分布から取り出されたことを示す) も行ったが、ほぼ同等の結果が得られたため省略する。

- 実験室における単一人の回数間距離の課題間の違いの評価

$BCCWJ-SUMM_L(T) \Leftrightarrow GLOSS_L(T) \Leftrightarrow RETELLING_I(T) \Leftrightarrow RETELLING_K(T) \Leftrightarrow RETELLING_M(T)$

- $BCCWJ-SUMM_L(T) \Leftrightarrow GLOSS_L(T)$

文字単位の評価の場合 $n\text{-gram}(1, \leq 2, \leq 3, \leq 4)\text{-char}$, $p\text{-mer}(2, 3, 4, \leq 2, \leq 3, \leq 4)\text{-char}$ に有意差があった.

形態素単位の評価の場合 $n\text{-gram}(1, \leq 2, \leq 3)\text{-mrph}$, $Kendall\text{-mrph}$ に有意差があった.

- $BCCWJ-SUMM_L(T) \Leftrightarrow RETELLING_I(T)$

$n\text{-gram}(4)\text{-char}$ ($BCCWJ-SUMM_L(T) \Leftrightarrow RETELLING_I(T)$), $n\text{-gram}(4)\text{-mrph}$ ($BCCWJ-SUMM_L(T) \Leftrightarrow RETELLING_I(T)$) 以外の全ての尺度について, 有意差が見られた.

- $GLOSS_L(T) \Leftrightarrow RETELLING_I(T)$

全ての尺度について, 有意差があった.

- $RETELLING_I(T) \Leftrightarrow RETELLING_K(T)$

$n\text{-gram}(1)\text{-mrph}$ についてのみ有意差があった.

- $RETELLING_I(T) \Leftrightarrow RETELLING_M(T)$

文字単位の評価の場合 $n\text{-gram}(4)\text{-char}$, $footrule\text{-char}$, $kendall\text{-char}$ 以外に有意差があった. 形態素単位の評価の場合, $kendall\text{-mrph}$ 以外に有意差があった.

- $RETELLING_I(T) \Leftrightarrow RETELLING_M(T)$

文字単位の評価の場合全ての尺度に有意差がなかった. 形態素単位の評価の場合, $n\text{-gram}(\leq 2, \leq 3, \leq 4)\text{-mrph}$, $footrule\text{-mrph}$, $kendall\text{-mrph}$ 以外に有意差があった.

複数人間の評価ではなく, 複数回間の評価でも同じ傾向が見られる.

再話課題間については, 形態素単位の評価において, 三課題のうちどの二つ組においても有意差が出る傾向にある. 口述による再話 ($RETELLING_I(T)$) の方が筆述による再話 ($RETELLING_M(T)$) より一致度が高くなる. また口述による再話においては, 自身の体験に基づく再話 ($RETELLING_I(T)$) の方が, 他者から聞いた話の再話 ($RETELLING_K(T)$) よりも一致度の高くなることが認められた.

- クラウドソーシングにおける課題間の違いの評価

$BCCWJ-SUMM_C \Leftrightarrow GLOSS_C$ について, 全ての尺度について, 有意差があった.

クラウドソーシングにおける課題間の違いについても, 前項と同じ傾向が見られる.

- 要約課題においてクラウドソーシングと実験室との違いの評価 (複数人間)

$BCCWJ-SUMM_C \Leftrightarrow BCCWJ-SUMM_L(P)$ について, $n\text{-gram}(2)\text{-char}$, $n\text{-gram}(3)\text{-char}$, $n\text{-gram}(4)\text{-char}$ にのみ有意差があった.

これは、タイプ入力 (BCCWJ-SUMM_C) と筆述 (BCCWJ-SUMM_L(P)) とで、表記ゆれ統制の差の影響が考えられる。

- 語釈課題においてクラウドソーシングと実験室との違いの評価（複数人間）
GLOSS_C $\Leftrightarrow$ GLOSS_L(P) について, n-gram(2,3,4)_char, n-gram(2,3,4)_mrph, Footrule_morph, Kendall_morph 以外について有意差があった。
語釈においては、クラウドソーシングの場合 wikipedia や辞書サイトからのコピーが行われる傾向にある一方、実験室の場合は特にリファレンスもなく筆述で行うために差が出たのではないかと考える。
- 複数人間距離と単一人の回数間距離の違いの評価
BCCWJ-SUMM_L(P) $\Leftrightarrow$ BCCWJ-SUMM_L(T), GLOSS_L(P) $\Leftrightarrow$ GLOSS_L(T), RETELLING_K(P) $\Leftrightarrow$ RETELLING_K(T), RETELLING_M(P) $\Leftrightarrow$ RETELLING_M(T) について、全ての尺度について有意差があった。
基本的に単一人が実施したほうが一致度が高いと考えられるが、統計分析の結果からもそれが確認できる。

4.4 各評価尺度の特性

課題間の議論から考えられる各尺度の特性について論じる。

まず、文字 n-gram はタイプ入力と筆述入力の差として認められることから、表記ゆれレベルで一一致度の下がる特性があると考えられる。形態素 n-gram は再話と繰り返して顕著に高くなったことから、個人の文体などを反映していると考えられる。

p-mer, Footrule, Kendall などは語順の一致を反映していると考えられるが、ストーリー性がある要約・再話で一一致度が高い一方、語釈などにおいては低い傾向にあることがわかった。語順に対して、ストーリーの一致を評価するのか、説明の順序を評価するのかについて深く検討する必要があると考えられる。ストーリーの一致については(加藤, 富田, 浅原 2016)において、被験者実験的に人が何を同一の物語とみなすかについて検討されている。また、語釈などにおいても情報提示順序により伝わりやすさが変わること(加藤, 浅原 2015)が報告されている。自動要約の評価尺度で導入された語順の尺度については、ストーリーの一致(内容一致)を目的とするのか、伝わりやすさの一致(表現一致)を目的とするのかについては言及されていない。今回の調査では、タスクの設定によりこれらを切り分けることを試みたが、タスク間の差異は確認できなかった。

n-gram, p-mer とともに n, p の値が高くなるにつれて尺度の値が低くなる。このために有意差が出にくくなる傾向にある。n-gram, p-mer とともに n (or p) 以下の尺度として設定した場合に、より低い n (or p) の方が一致が多くなる傾向にあるために、より高い n (or p) の差異が見られなくなる傾向がある。これは尺度の自然な解釈であると考えられるが、何らかの用途で長い

n-gram, p-mer を重要視する場合には部分（文字）列長に対して加重を行う必要があるだろう。n-gram(1)* と Kendall* と比較した場合、n-gram(1)*では有意差が出るが、順序尺度を入れた Kendall* では有意差が出ない尺度の組み合わせがいくつかあった。これは文字順・語順の一致度が低い場合に、順序尺度を掛けあわせたがために全体の一致度の差がなくなったことが考えられる。

5 おわりに

本稿では、まず自動要約・機械翻訳で用いられている評価指標の数理的構造を説明した。評価指標がどのカーネル・距離・相関係数などの尺度と対応しているのかを説明し、n-gram 系、p-mer 系、ベクトル型順序尺度、編集型順序尺度の四つに抽象化した。

次に様々な言語資源を用いて各指標を構成する尺度の特性を明らかにした。要約・語釈・再話からなる 7 種類の言語資源を用いて、課題・多人数産出・複数回産出・産出手段（口述・筆述・タイプ）の軸を用いて、どのような分散が観察されるかを確認した。結果、各評価尺度において、表現一致と内容一致の識別は困難であり、評価の識別限界としての分散があることを示した。

今後の展開として以下の五つを考えている。

一つ目は要約評価に求められる尺度とは何かを明らかにすることである。尺度が捉える言語の特性については明らかにしたが、自動要約に必要な内容評価と読みやすさの観点については何も言っていないに等しい。現在、収集した要約に対して、以下の五つの軸で人手による評価を付与している (浅原 他 2015)。

- 文法性 (Grammaticality): 誤字・文法的でない文が含まれていないか
- 非冗長性 (Non-redundancy): 全く同じ情報が繰り返されていないか
- 指示詞の明解さ (Referential clarity): 先行詞のない指示詞（代名詞）が含まれていないか
- 焦点 (Focus): 要約全体と無関係な情報が含まれていないか
- 構造と一貫性 (Structure and Coherence): 接続詞を補ったり削除したりする必要のある箇所はないか

人手による評価を悉皆的に付与したうえで、各評価軸がどの尺度に表れるのかを引き続き分析していきたい。

二つ目は情報構造アノテーションとの重ね合わせである。尺度において、語順の評価を入れるかどうかの一つの論点であった。日本語において語順を決める一つの要素として情報構造がある。情報構造は言語生産者側の観点である情報状態 {speaker-new, speaker-old} と言語受容者側の観点である共有性 {hearer-new, hearer-old} の区別を行い、後者については被験者実験的にアノテーションを行う。これらのアノテーション結果を用いて、なぜ要約文はその順序で情報

を提示する必要があるのかについて検討する.

三つ目は要約文の言語受容者側の観点からの認知的な評価である. 今回は元文書の言語受容者であり要約文の言語生産者側の観点からの認知的な評価を主に扱った. 生産された要約文が他の言語受容者にとって同じ話として認定されるか (加藤 他 2016; 加藤, 浅原 2015) を検討していきたい. 一方, 日本語複数文書要約についての拡張も考えられるが, 元文書側の言語生産者が複数人になるという問題がある. 複数の言語生産者側が考慮している情報構造が, 要約作成者と要約受容者にどのように受容されるか追跡可能な認知実験手法を検討する.

四つ目は人文系の研究者が評価する文体についての尺度を明らかにすることである. 文体の研究は使用域 (register) や位相 (phase) などに着目して行われるが, 現在のところ役割語など単一の語についての研究がほとんどである. 語の連接や語の順序などの使用域や位相を, 先に述べた尺度で捉えることを目標とする.

五つ目は同じ話とは何かということを定量的に評価する手法の提案である. 内容を捉える尺度と表現を捉える尺度を分離することで, 人が内容が一致していると認知できる表現のゆれを捉えることを目標とする. 既に同じ話を構成する要素について, 様々な分析 (保田, 荒牧 2012; 保田 他 2013b; 保田 2014; 加藤 2015c; 加藤, 浅原 2016) を進めているが, これらの結果が尺度にどのように表れるのかについて分析を行う.

謝 辞

要約文の作成および評価については NTT CS 研の平尾努氏の助言を受けました. 本研究の一部は, 国立国語研究所基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」(2011–2015) および国立国語研究所「超大規模コーパス構築プロジェクト」(2011–2015) によるものです. 本研究は JSPS 科研費 基盤研究 (B) 25284083, 若手研究 (B) 26770156 の助成を受けました.

参考文献

- 天野成昭, 近藤公久 (1999). 日本語の語彙特性第 1 期 CD-ROM 版. 三省堂.
- 浅原正幸, 杉真緒, 柳野祥子 (2015). BCCWJ-SUMM: 『現代日本語書き言葉均衡コーパス』を元文書とした要約文書コーパス. 第 7 回コーパス日本語学ワークショップ予稿集, pp. 285–292.
- Birch, A. and Osborne, M. (2010). “LRscore for Evaluation Lexical and Reordering Quality in MT.” In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pp. 327–332.

- Echizen-ya, H. and Araki, K. (2007). “Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum.” In *Proceedings of the MT Summit XI Workshop on Patent Translation*, pp. 151–158.
- 平尾努, 磯崎秀樹, 須藤克仁, K. Duh, 塚田元, 永田昌明 (2014). 語順の相関に基づく機械翻訳の自動評価法. 自然言語処理, **21** (3), pp. 411–444.
- 平尾努, 奥村学, 磯崎秀樹 (2006). 拡張ストリングカーネルを用いた要約システムの自動評価法. 情報処理学会論文誌, **47** (6), pp. 1753–1765.
- 平尾努, 奥村学, 安田宣仁, 磯崎秀樹 (2007). 投票型回帰モデルによる要約自動評価法. 人工知能学会論文誌, **22** (2), pp. 115–126.
- 神畹敏弘 (2009). 順序の距離と確率モデル. 人工知能学会研究会資料 SIG-DMSM-A902-07.
- 加藤祥 (2015a). テキストから対象物認識に有用な記述内容—動物を例に一. 国立国語研究所論集, **9**, pp. 23–50.
- 加藤祥 (2015b). 象は鼻が長い—テキストから取得される対象物情報. 第7回コーパス日本語学ワークショップ, pp. 35–44.
- 加藤祥 (2015c). 同じ話における共通語彙. 社会言語科学会第36回大会.
- 加藤祥, 浅原正幸 (2015). テキストからの対象物認識における情報提示順序の影響. 2015年度日本認知科学会第32回大会, pp. 362–369.
- 加藤祥, 浅原正幸 (2016). 恋愛小説において物語を特徴づける表現—タイトルと帯に見られる表現分析の試み—. 社会言語科学会第38回大会, pp. 128–131.
- 加藤祥, 富田あかね, 浅原正幸 (2016). 物語がその物語であるための要素—何が同じであれば同じで何が違えば違うのか—. 言語処理学会第22回発表論文集, pp. 266–269.
- Lin, C.-Y. (2004). “ROUGE: A Package for Automatic Evaluation of Summaries.” In *Proceedings of Workshop on Summarization Branches Out, Post Conference Workshop of ACL 2004*, pp. 74–81.
- Lin, C.-Y. and Hovy, E. (2003). “Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics.” In *Proceedings of the 4th Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technology*, pp. 150–157.
- Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). “Balanced corpus of contemporary written Japanese.” *Language Resources and Evaluation*, **48**, pp. 345–371.
- 宮部真衣, 四方朱子, 久保圭, 荒牧英治 (2014). 音声認識による認知症・発達障害スクリーニングは可能か?—一言語能力測定システム“言秤”の提案—. グループウェアとネットワークサービスワークショップ2014.

- 難波英嗣, 平尾努 (2008). テキスト要約の自動評価. 人工知能学会誌, **23** (1), pp. 10–16.
- Nivre, J. (2009). “Non-Projective Dependency Parsing in Expected Linear Time.” In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 351–359.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). “Bleu: a Method for Automatic Evaluation of Machine Translation.” Tech. rep., IBM Research Report RC22176 (W0109-022).
- Shawe-Taylor, J., Cristianini N., 大北剛 (訳) (2010). カーネル法によるパターン解析 (Kernel Methods for Pattern Analysis), 第 11 章 構造化データに対するカーネル: 文字列, 木など. 共立出版.
- 保田祥 (2014). 同じ話を成立させる語—「桃太郎」を「桃太郎」として成立させる語彙—. 社会言語科学会第 33 回大会発表論文集, pp. 138–141.
- 保田祥, 荒牧英治 (2012). 人が同じ話を何度もするとどうなるか?: 繰り返しによって生じる物語独話の変化. 日本認知科学会第 29 回大会, pp. 217–223.
- 保田祥, 田中弥生, 荒牧英治 (2013a). 繰り返しにおける独話の変化. 社会言語科学会第 31 回大会発表論文集, pp. 190–193.
- 保田祥, 田中弥生, 荒牧英治 (2013b). 同じ話であるとはどういうことか. 社会言語科学会第 32 回大会発表論文集, pp. 30–33.

付録

A 2 節で用いる用語定義

以下 2 節で用いる用語を定義する:

- 記号集合: 本稿では記号の集合を σ で表す.
- 記号列: 何らかの全順序が付与されている記号集合.
本稿では記号列ベクトル $s = \langle s_1, \dots, s_m \rangle, t = \langle t_1, \dots, t_m \rangle$ など で表現する.
- 文字 (character), 文字ベース (character-based): 記号集合 σ の要素の記号 $s_i \in \sigma$ としての文字. 記号集合 σ の要素が文字であること.
- 形態素 (morpheme), 形態素ベース (morpheme-based): 記号集合 σ の要素の記号 $s_i \in \sigma$ としての形態素. 記号集合 σ の要素が形態素であること.
- 文字列 (string): 評価する記号列上の連続列. 記号列の要素が文字 (character) である場合を「文字ベースの文字列 (character-based string)」, 記号列の要素が形態素 (morpheme)

である場合を「形態素ベースの文字列 (morpheme-based string)」と呼ぶこととする.

- 部分文字列 (substring): 記号列に対して隣接性と順序を保持した部分的記号列. 長さ n の部分文字列を特に n -gram 部分文字列と呼ぶ.

記号列 s の i 番目の要素からはじまる n -gram 部分文字列を $s_{i...i+n-1}$ で表現する.

- 部分列 (subsequence): 記号列に対して順序を保持した部分的記号列. 隣接性は保持しなくてよい. 長さ p の部分列を特に p -mer 部分列と呼ぶ. 記号列 s の p -mer 部分列を, インデックスベクトル $\vec{i} = \langle i_1, \dots, i_p \rangle (1 \leq i_1 < i_2 < \dots < i_p \leq |s|)$ を用いて, $s[\vec{i}]$ と表す.
- 参照テキスト (reference): 人間が作成した正解要約／翻訳. 本稿では記号列 R で表す.
- システム出力テキスト (candidate): 要約作成器／機械翻訳器が出力した要約／翻訳. 本稿では記号列 C で表す.
- 距離 (distance): 集合 X 上で定義された 2 変数の実数値関数で, 本稿では $d: X \times X \rightarrow R$ などの記号を使う. 正定値性 ($d(x, y) \geq 0$), 非退化性 ($x = y \Rightarrow d(x, y) = 0$), 対称性 ($d(x, y) = d(y, x)$), 三角不等式 ($d(x, y) + d(y, z) \geq d(x, z)$) を満たす.
- 絶対値 (absolute value): 大きさの一般化概念. 実数については 0 からの距離, 集合については要素数を表すのに用い $|x|$ で表す. また, テキスト C 中の単語/ n -gram/ p -mer の要素数 (のべ出現数) を $|e|_C$ で表す.
- θ -ノルム (norm): ベクトル空間上に距離を規定する長さの一般化概念. ベクトル $x = \langle x_1, \dots, x_n \rangle$ の θ -ノルムを $\|x\|_\theta = (\sum_{i=1}^n |x_i|^\theta)^{1/\theta}$ により定義する. 特に θ を定義しない場合 ($\|x\|$) は 2-ノルムを用いる.
- 内積記号 $\cdot$: 文字列に対しては連結, 整数・実数については積, ベクトルなどについては内積, 対称群については写像の合成 (積) を扱うために用いる.
- 類似度 (similarity): 二つの元の距離は遠さを表すのに対し, 類似度は近さを表す. 距離の逆数は類似度として扱える.
- 相関係数 (correlation): 二つの確率変数の間の相関を表す指標で, 類似度として扱える. $[-1, 1]$ 区間の値をとり, 1 に近い場合は正の相関があると呼び, -1 に近い場合には負の相関があると呼ぶ. 0 に近い場合には相関が弱いという意味がある.
- カーネル関数 (kernel function): 特徴空間中の座標の明示的な計算を経由せずに特徴量空間における内積 (正定値性と非退化性を持ち, 実数ベクトル空間では対称性ももつ) を定義するもの. 本稿では $K(s, t)$ と表記する. 内積を正規化することにより cosine 類似度 ($\frac{K(s, t)}{\|K(s, s)\| \cdot \|K(t, t)\|}$) を定義することができる.
- スコア (score): 類似度などを $[0, 1]$ 区間に正規化したもの. 本稿では score などの記号で示す.
- 接頭辞 (prefix): 記号列の先頭要素を含む連続文字列.
- 接尾辞 (suffix): 記号列の末尾要素を含む連続文字列.

- 部分集合 (subset): 記号列を集合とみなした場合の部分集合. 隣接性と順序は保持しなくてよい. 要素数 k の部分集合を特に k -element 部分集合と呼ぶ.
- 順位ベクトル (rank vector): インデックス i 要素が対象 i の順位を表すベクトル. 本稿では m 次元の順位ベクトル空間を S_m で表し, 順位ベクトル空間の要素である順位ベクトルを $\mu = \langle \mu(1), \dots, \mu(m) \rangle$ で表す. $\mu(i)$ には対象 i の順位を表す自然数が入る.
- 順序ベクトル (order vector): 順位が i 番目である要素がインデックス i の位置に格納されているベクトル. 本稿では m 次元の順序ベクトル空間を T_m で表し, 順位ベクトル $\mu(i)$ に対応する順序ベクトルを $\mu^{-1} = \langle \mu^{-1}(1), \dots, \mu^{-1}(m) \rangle$ で表す. $\mu^{-1}(i)$ には順位が i である要素 (の順位ベクトル上でのインデックス) が入る.
- 同順 (concordant): 二つの順位ベクトル中で対象対 i と j が以下を満たすとき, その対象対が同順であるという.

$$(\mu(i) - \mu(j))(\nu(i) - \nu(j)) \geq 0$$

- 逆順 (discordant): 二つの順位ベクトル中で対象対が同順でないことを逆順という.
- 文字列上の編集: 挿入 (insertion), 削除 (deletion), 代入 (substitution) の三つを規定する.
- 順序ベクトル上の編集: 順序ベクトルを対称群 (symmetric group) と考えて編集する際の操作を規定する.

- 対称群: 並び替えの編集操作 (置換: permutation) を元とする群. 順序ベクトル $\mu^{-1} = \langle \mu^{-1}(1), \dots, \mu^{-1}(m) \rangle$ のうち, $\mu^{-1}(k_1), \mu^{-1}(k_2), \dots, \mu^{-1}(k_r)$ 以外は動かさず, $\mu^{-1}(k_1) \rightarrow \mu^{-1}(k_2), \mu^{-1}(k_2) \rightarrow \mu^{-1}(k_3), \dots$ のように順にずらす置換

$$\begin{pmatrix} \mu^{-1}(k_1) & \mu^{-1}(k_2) & \dots & \mu^{-1}(k_r) \\ \mu^{-1}(k_2) & \mu^{-1}(k_3) & \dots & \mu^{-1}(k_1) \end{pmatrix}$$

のことを巡回置換と呼び, $\pi_r = (k_1, k_2, \dots, k_r)$ で表す. 二つの元のみを入れ替えて他の元は変えないもの (2 元の巡回置換) を互換 (transposition) と呼び, $\pi_2 = (i, j)$ で表す. 隣接する二つの元のみを入れ替えて他の元は変えないものを隣接互換 (adjacent transposition) と呼び, $\pi_2 = (i, i+1)$ で表す.

- クロネッカーのデルタ δ : $\delta(i, j) = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$
- s 再帰 (s -recursive), t 再帰 (t -recursive): それぞれ変数 s, t に対して再帰的に定義すること. 本稿では s, t は文字列を想定し, 1 文字増やした際の文字列を定義する差分方程式の説明に用いる.

B 指標・スコア・距離・カーネル・相関係数の関係まとめ

表 2 指標・スコア・距離・カーネル・相関係数の関係まとめ

	指標	スコア $[0, 1] \uparrow$	距離 $[0, \infty] \downarrow$	カーネル $[0, \infty] \uparrow$ (加重) 全部分文字列 n-スベクトラム	相関係数 $[-1, 1] \uparrow$
部分文字列系 (n-gram)	IMPACT	$\text{Score}_{K_{\text{all-str}}}^{(\gamma)} \uparrow$			
	ROUGE-N	$\text{Score}_{K_{\text{n-gram}}}^{(\gamma)} \uparrow$			
	BLEU LRscore				
部分列系 (p-mer)		$\text{Score}_{K_{\text{all-seq}}}^{(\gamma)} \uparrow$		(加重) 全部分列	
	ROUGE-S(U)	$\text{Score}_{K_{\text{p-mer}}}^{(\gamma)} \uparrow$		p-mer 部分列	
	ESK	$\text{Score}_{K_{\text{gap-p-mer}}}^{(\gamma)} \uparrow$		加重 p-mer 部分列	
順序系 (ベクトル型)		$\text{Score}_{ \text{rank} _{\theta}} \uparrow$	$d_{ \text{rank} _{\theta}}$		
		$\text{Score}_{\text{Footrule}} \uparrow$	$d_{\text{Footrule}(\theta=1)}$		
	RIBES	$\text{Score}_{\text{Spearman}} \uparrow$	$(d_{\text{Spearman}(\theta=2)^2})$		Spearman's ρ
	LRscore	$\text{Score}_{\text{Hamming}} \uparrow$	d_{Hamming}	$\subset$ Pearson's	
順序系 (編集型)	RIBES LRscore	$\text{Score}_{\text{Kendall}} \uparrow$	d_{Kendall}		Kendall's τ
			d_{Caylay}		
	ROUGE-L	$\text{Score}_{\text{LCS}} \uparrow$	d_{Ulam}		
	ROUGE-W	$\text{Score}_{\text{WLCS}}^{(\gamma)} \uparrow$	$d_{\text{Levenshtein}}$		
(加重最長一致部分列長)					
(最長一致部分文字列長)		$\text{Score}_{\text{LCS}_{\text{Substr}}} \uparrow$			

最初の行の横にある $[,]$ は値域を表す. $\uparrow$ は似ている文書対が高い値をとることを表す. $\downarrow$ は似ている文書対が低い値をとることを表す.

C 言語生成過程と尺度

表3に4.3節で行った検定の結果のまとめを示す. 表4に要約課題・語釈課題と各尺度の比較を表5に再話課題と各尺度の比較を示す. 標準偏差はBCCWJ-SUMML(T)が最も大きい. これは繰り返し要約する際に全く同じ要約文を再生産する被験者と全く異なる要約文を再生産する被験者とが存在するからだと考えられる. しかし, 他の再話 (RETELLING_K(T), RETELLING_M(T))でも被験者間の標準偏差と比して高いことから要約文特有の現象ではないと考える.

表3 4.3節で行った検定の結果のまとめ (有意差があったもの一覧)

実験室における複数人の課題間の違いの評価	
BCCWJ-SUMML(P) $\Leftrightarrow$ GLOSS_L(P)	{n-gram(2,3,4)_char, n-gram(2,3,4)_mrph} 以外の尺度
BCCWJ-SUMML(P) $\Leftrightarrow$ RETELLING_K(P)	{n-gram(4)_char, n-gram(3,4)_mrph} 以外の尺度
BCCWJ-SUMML(P) $\Leftrightarrow$ RETELLING_M(P)	{n-gram(4)_mrph} 以外の尺度
GLOSS_L(P) $\Leftrightarrow$ RETELLING_K(P)	全ての尺度
GLOSS_L(P) $\Leftrightarrow$ RETELLING_M(P)	全ての尺度
RETELLING_K(P) $\Leftrightarrow$ RETELLING_M(P)	—
実験室における単一人の回数間距離の課題間の違いの評価	
BCCWJ-SUMML(T) $\Leftrightarrow$ GLOSS_L(T)	n-gram(1, $\leq 2, \leq 3, \leq 4$)_char, p-mer(2,3,4, $\leq 2, \leq 3, \leq 4$)_char, n-gram(1, $\leq 2, \leq 3$)_mrph, Kendall_mmrph
BCCWJ-SUMML(T) $\Leftrightarrow$ RETELLING_I(T)	全ての尺度
BCCWJ-SUMML(T) $\Leftrightarrow$ RETELLING_K(T)	n-gram(4)_char, n-gram(4)_mrph 以外の尺度
BCCWJ-SUMML(T) $\Leftrightarrow$ RETELLING_M(T)	n-gram(4)_mrph 以外の尺度
GLOSS_L(T) $\Leftrightarrow$ RETELLING_I(T)	全ての尺度
GLOSS_L(T) $\Leftrightarrow$ RETELLING_K(T)	全ての尺度
GLOSS_L(T) $\Leftrightarrow$ RETELLING_M(T)	全ての尺度
RETELLING_I(T) $\Leftrightarrow$ RETELLING_K(T)	n-gram(1)_mrph
RETELLING_I(T) $\Leftrightarrow$ RETELLING_M(T)	{n-gram(4)_char, footrule_char, kendall_char, kendall_mmrph} 以外の尺度の
RETELLING_K(T) $\Leftrightarrow$ RETELLING_M(T)	n-gram(1, $2 \leq, 3 \leq, 4 \leq$)_mrph, p-mer(2,3,4, $2 \leq, 3 \leq, 4 \leq$)_mrph
クラウドソーシングにおける課題間の違いの評価	
BCCWJ-SUMML_C $\Leftrightarrow$ GLOSS_C	全ての尺度
要約課題におけるクラウドソーシングと実験室との違いの評価 (複数人間)	
BCCWJ-SUMML_C $\Leftrightarrow$ BCCWJ-SUMML(P)	n-gram(2)_char, n-gram(3)_char, n-gram(4)_char
語釈課題におけるクラウドソーシングと実験室との違いの評価 (複数人間)	
GLOSS_C $\Leftrightarrow$ GLOSS_L(P)	n-gram(2,3,4)_char, n-gram(2,3,4)_mrph, Footrule_mmrph, Kendall_mmrph 以外の尺度
複数人間距離と単一人の回数間距離の違いの評価	
BCCWJ-SUMML(P) $\Leftrightarrow$ BCCWJ-SUMML(T)	全ての尺度
GLOSS_L(P) $\Leftrightarrow$ GLOSS_L(T)	全ての尺度
RETELLING_K(P) $\Leftrightarrow$ RETELLING_K(T)	全ての尺度
RETELLING_M(P) $\Leftrightarrow$ RETELLING_M(T)	全ての尺度

表 4 要約課題・語釈課題と各尺度

score	BCCWJ-SUMM_C		BCCWJ-SUMML(P)		BCCWJ-SUMML(T)		GLOSS_C		GLOSS_L(P)		GLOSS_L(T)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
n-gram(1)_c	0.63	0.12	0.64	0.11	0.68	0.27	0.68	0.09	0.65	0.06	0.83	0.07
n-gram(2)_c	0.36	0.17	0.33	0.15	0.48	0.32	0.24	0.10	0.22	0.09	0.56	0.16
n-gram(3)_c	0.26	0.17	0.22	0.15	0.40	0.34	0.11	0.07	0.09	0.06	0.41	0.18
n-gram(4)_c	0.20	0.16	0.16	0.14	0.35	0.34	0.05	0.06	0.04	0.04	0.31	0.18
n-gram(≤ 2)_c	0.52	0.13	0.53	0.12	0.61	0.28	0.55	0.09	0.51	0.07	0.75	0.10
n-gram(≤ 3)_c	0.45	0.14	0.45	0.12	0.55	0.29	0.47	0.08	0.42	0.06	0.68	0.11
n-gram(≤ 4)_c	0.40	0.14	0.39	0.12	0.51	0.30	0.41	0.08	0.36	0.06	0.62	0.12
p-mer(2)_c	0.34	0.14	0.35	0.13	0.49	0.33	0.39	0.10	0.34	0.07	0.63	0.14
p-mer(3)_c	0.18	0.13	0.18	0.12	0.37	0.35	0.20	0.08	0.15	0.05	0.45	0.17
p-mer(4)_c	0.09	0.10	0.09	0.10	0.30	0.36	0.09	0.06	0.06	0.03	0.31	0.17
p-mer(≤ 2)_c	0.34	0.14	0.35	0.13	0.49	0.33	0.39	0.10	0.34	0.07	0.63	0.14
p-mer(≤ 3)_c	0.18	0.13	0.18	0.12	0.37	0.35	0.20	0.08	0.16	0.05	0.45	0.17
p-mer(≤ 4)_c	0.10	0.11	0.09	0.10	0.30	0.36	0.10	0.06	0.06	0.03	0.31	0.17
Footrule_c	0.50	0.15	0.50	0.14	0.59	0.29	0.48	0.10	0.45	0.08	0.69	0.14
Kendall_c	0.48	0.14	0.47	0.13	0.57	0.29	0.44	0.08	0.41	0.06	0.64	0.13
n-gram(1)_m	0.60	0.12	0.62	0.11	0.67	0.27	0.62	0.10	0.58	0.07	0.79	0.09
n-gram(2)_m	0.25	0.16	0.24	0.15	0.41	0.34	0.13	0.08	0.11	0.07	0.42	0.18
n-gram(3)_m	0.15	0.15	0.14	0.14	0.34	0.35	0.04	0.06	0.03	0.04	0.27	0.18
n-gram(4)_m	0.10	0.13	0.09	0.12	0.29	0.35	0.02	0.05	0.01	0.02	0.19	0.16
n-gram(≤ 2)_m	0.46	0.13	0.48	0.12	0.57	0.29	0.48	0.09	0.43	0.07	0.67	0.11
n-gram(≤ 3)_m	0.38	0.13	0.39	0.13	0.51	0.30	0.39	0.08	0.34	0.06	0.58	0.12
n-gram(≤ 4)_m	0.32	0.13	0.33	0.12	0.47	0.31	0.33	0.07	0.28	0.05	0.52	0.13
p-mer(2)_m	0.30	0.14	0.32	0.14	0.47	0.33	0.32	0.10	0.27	0.07	0.55	0.15
p-mer(3)_m	0.15	0.12	0.15	0.12	0.35	0.35	0.14	0.07	0.11	0.04	0.35	0.17
p-mer(4)_m	0.07	0.09	0.07	0.10	0.28	0.36	0.06	0.05	0.03	0.02	0.22	0.16
p-mer(≤ 2)_m	0.31	0.14	0.33	0.14	0.47	0.33	0.32	0.10	0.28	0.07	0.55	0.15
p-mer(≤ 3)_m	0.16	0.12	0.16	0.13	0.36	0.35	0.14	0.07	0.11	0.04	0.36	0.17
p-mer(≤ 4)_m	0.08	0.10	0.08	0.10	0.29	0.36	0.06	0.05	0.04	0.02	0.23	0.17
Footrule_m	0.48	0.15	0.48	0.14	0.59	0.30	0.44	0.10	0.42	0.10	0.66	0.14
Kendall_m	0.46	0.15	0.46	0.14	0.56	0.29	0.41	0.09	0.39	0.08	0.61	0.13

表 5 再話課題と各尺度

score	RETELLING_J(T)		RETELLING_K(P)		RETELLING_K(T)		RETELLING_M(P)		RETELLING_M(T)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
n-gram(1)-c	0.96	0.09	0.86	0.05	0.91	0.19	0.87	0.08	0.92	0.13
n-gram(2)-c	0.85	0.12	0.59	0.10	0.78	0.21	0.58	0.20	0.75	0.16
n-gram(3)-c	0.73	0.17	0.39	0.13	0.65	0.22	0.44	0.21	0.64	0.18
n-gram(4)-c	0.61	0.19	0.23	0.13	0.50	0.21	0.33	0.18	0.52	0.18
n-gram(≤ 2)-c	0.94	0.09	0.80	0.06	0.89	0.19	0.80	0.11	0.88	0.13
n-gram(≤ 3)-c	0.91	0.09	0.76	0.07	0.86	0.19	0.74	0.12	0.84	0.13
n-gram(≤ 4)-c	0.89	0.10	0.73	0.07	0.84	0.18	0.70	0.13	0.81	0.14
p-mer(2)-c	0.91	0.13	0.72	0.09	0.85	0.19	0.72	0.13	0.83	0.14
p-mer(3)-c	0.86	0.16	0.60	0.11	0.79	0.21	0.58	0.16	0.74	0.16
p-mer(4)-c	0.81	0.18	0.49	0.13	0.73	0.22	0.45	0.16	0.64	0.19
p-mer(≤ 2)-c	0.91	0.13	0.72	0.09	0.85	0.19	0.72	0.13	0.83	0.14
p-mer(≤ 3)-c	0.86	0.16	0.60	0.11	0.79	0.21	0.58	0.16	0.74	0.16
p-mer(≤ 4)-c	0.81	0.18	0.49	0.13	0.73	0.22	0.45	0.16	0.64	0.19
Footrule-c	0.88	0.11	0.73	0.07	0.83	0.18	0.75	0.13	0.85	0.14
Kendall-c	0.81	0.11	0.66	0.06	0.77	0.17	0.68	0.12	0.78	0.13
n-gram(1)-m	0.94	0.15	0.85	0.06	0.92	0.18	0.83	0.07	0.90	0.13
n-gram(2)-m	0.75	0.18	0.48	0.10	0.72	0.21	0.45	0.18	0.64	0.17
n-gram(3)-m	0.58	0.21	0.21	0.13	0.48	0.19	0.28	0.15	0.47	0.16
n-gram(4)-m	0.39	0.21	0.09	0.07	0.30	0.15	0.10	0.08	0.27	0.15
n-gram(≤ 2)-m	0.92	0.15	0.78	0.08	0.88	0.18	0.73	0.10	0.84	0.13
n-gram(≤ 3)-m	0.89	0.15	0.73	0.08	0.85	0.18	0.66	0.11	0.78	0.13
n-gram(≤ 4)-m	0.86	0.15	0.69	0.09	0.82	0.18	0.60	0.11	0.73	0.13
p-mer(2)-m	0.90	0.18	0.70	0.11	0.86	0.19	0.64	0.11	0.78	0.14
p-mer(3)-m	0.85	0.19	0.56	0.13	0.79	0.19	0.46	0.13	0.66	0.17
p-mer(4)-m	0.79	0.20	0.44	0.14	0.73	0.20	0.32	0.12	0.54	0.18
p-mer(≤ 2)-m	0.90	0.18	0.70	0.11	0.86	0.19	0.64	0.11	0.78	0.14
p-mer(≤ 3)-m	0.85	0.19	0.56	0.13	0.79	0.19	0.46	0.13	0.66	0.17
p-mer(≤ 4)-m	0.79	0.20	0.44	0.14	0.73	0.20	0.33	0.12	0.54	0.18
Footrule-m	0.86	0.16	0.71	0.07	0.83	0.18	0.72	0.10	0.83	0.14
Kendall-m	0.80	0.16	0.63	0.06	0.76	0.17	0.66	0.10	0.77	0.13

略歴

浅原 正幸：2003 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。2004 年より同大学助教。2012 年より国立国語研究所コーパス開発センター特任准教授。国立国語研究所言語資源研究系准教授を経て、現在国立国語研究所コーパス開発センター准教授。博士（工学）。

加藤 祥：2011 年神戸大学人文学研究科博士後期課程修了。2012 年より国立国語研究所コーパス開発センタープロジェクト PD フェロー。現在国立国語研究所コーパス開発センタープロジェクト非常勤研究員。博士（文学）。

(2016 年 5 月 12 日 受付)

(2016 年 8 月 17 日 再受付)

(2016 年 10 月 4 日 採録)