E-011

構文解析を用いた日本語論文の読みやすさ判定法

A Method for Readability Assessment of Japanese Paper Using Syntactic Analysis

山中 翔太 †　山崎 高弘 ††　常盤 欣一朗 ††　長谷川 哲子 †††
Shota Yamanaka　Takahiro Yamasaki　Kin-ichiroh Tokiwa　Noriko Hasegawa

1 はじめに

近年，自然言語を対象とし，膨大な文書から有益な情報を抽出及び分類する研究が盛んに行われている．分類対象はWikipediaやブログなどのWEB上の情報やアンケート情報など多岐にわたっている．また，分類をするための基準は，評価情報や信頼性など様々である．しかしながら，文書の難しさや面白さなど人の主観が強い評価基準に対する分類手法については，報告はあまりなされていない．

本研究では，文書の難しさを評価基準とした分類方法の検討を行う．文書の難しさを文書の読みやすさとして考え，それによる判定方法の提案を行う．分類対象を日本語論文とすることで，文書に一定の統一性を与えることにする．日本語論文から，語句の出現頻度などの統計情報および文節の接続構造を特徴として抽出し，SVMを用いて分類を行う．

2 文書の読みやすさ

文書の難しさには，大別すると二つの解釈があると考えられる．一つ目は，文書の内容が理解できないことによる難しさである．この難しさは，文書に書かれている語句によって決まり，読み手の知識量や理解力でも左右される．二つ目は，文書の構造が複雑であることによる難しさである．一文の長さや係り受けの複雑さ等によって決まり，構造が複雑な文書は何度も読み返すことが必要なため理解することが難しくなる．

前者の難しさについては，文書中の語句の難易度レベルを用いて判定する手法が提案されている [1]．本研究では，後者の難しさを読みやすさと定義し，文書の複雑さをもとに読みやすさの判定を行う．

日本語文書における読みやすさとは，以下に示す三つの要因からなるとする．

- 読み手が文書を見たときの印象
- 文書の係り受け構造と論理構成
- 記述されている内容

まず，文書の印象は文の長さ，使われている文字の種類等で決まる．一つの文が長すぎると読み直しが必要になり，読みにくく感じることが多い．また，ひらがなばかりの文書，逆に漢字ばかりの文書は読みやすいとはいえない．次に，文書の構造では，一つの文における文節の係り受け構造と，複数の文にわたっての文同士の論理構成を考える．文節の係り受けが複雑である文や，文と文とのつながりが明確にわからない文書は，一読しただけでは読みにくいと感じられる．最後に，文書の内容は，事前の知識量によって決まり，読み手によっては読みにくいことがある．文書の内容は意味を分析する必要があるので，提案手法では考慮に入れず，文書の印象と文書の構造から読みやすさを判定する．

† 大阪産業大学大学院 工学研究科 電子情報通信工学専攻
†† 大阪産業大学 工学部 電子情報通信工学科
††† 大阪産業大学 教養部

3 統計情報と文構造を用いた読みやすさの評価 (提案手法)

SVMを用いて読みやすさの判定を行うために，日本語論文の統計情報と文構造を用いて特徴量のパラメータ化を行う．文書の印象による読みやすさに関しては，文の長さや文字種の割合等の統計情報を用いて評価する．一つの文が適当な長さで区切られている文書のほうが一般的に読みやすいと感じられる．また，日本語文書では文字の種類がひらがな，漢字，カタカナ，アルファベットの4種類であるが，特にひらがなと漢字の割合によって読みやすさは変化する．これらを文書中から統計情報として抽出し，文書の特徴量として利用する．

文書の構造による読みやすさに関しては，文の構造と文同士の構造をそれぞれ分析することで評価する．簡単な構造か複雑な構造かで読みやすさは決定される．複雑な構造の文の判定は，文節と文節のつながりである係り受けを解析することで可能となる．係り受けが離れている場合は，修飾語と被修飾語が離れており，読みにくいといえる．

文書の論理構成による読みやすさは，文同士がどのようにつながっているかで決まる．そのため，接続詞に注目する．接続詞とは，文と文との橋渡しのような働きを持つ品詞で，例えば,「また」や「よって」等である．接続詞の後の文の展開に適合していない場合や接続詞の有無によって読みやすさは変化する．

本研究では，日本語論文の印象を表す特徴量として，文の長さ (P1) と文字種の文字数比 (P2) の統計情報を用いる．さらに文書の構造を表す特徴量として，構文解析による係り受け情報 (P3, P4) を用いる．それぞれの特徴量を以下のようにパラメータ化する．

まず，文の長さ P_l は句点から句点までのひらがな，漢字，カタカナの文字数とし，文書における一つの文の平均値を用いる．文字種の割合は，文に含まれる割合の多いひらがなと漢字の文字数比 P_r を用いる．次に，文の構造はある一つの文節に対する係り受けの回数 P_f と文節同士の係り受けの遠さ P_d で表すことができる．

(P1) 文の長さ

$$P_l = \frac{\text{文字数 (ひらがな，漢字，カタカナ)}}{\text{文の数}}$$

(P2) ひらがなと漢字の文字数比

$$P_r = \frac{\text{ひらがなの文字数}}{\text{漢字の文字数}}$$

(P3) 係り受けの回数

$P_f = n$ 個以上の文節から係り受けがある文節数

(P4) 係り受けの遠さ

$P_d = k$ 文節以上係り受けが離れている文節の組数

(P1)~(P4) の各パラメータが文書の読みやすさにどのような影響を及ぼすことになるかを以下に示す．文の長さ P_l は長すぎると読みにくくなり，さらに係り受けの回数 P_f，係り受けの遠さ P_d にも影響を与えている．また，係り受けの回数 P_f や係り受けの遠さ P_d の値が増えると文の構造が複雑であり，読みにくいといえる．

ここで，「この論文は内容が難しい.」という例文の構文解析の結果を示す．

```
構文解析の結果
  文節の順番
    1        この-D
    2        論文は---D
    3          内容が-D
    4            難しい.
```

二番目の文節「論文は」は四番目の文節「難しい.」に係っている．この場合係り受けが 2 文節離れている．また，四番目の文節「難しい.」に注目すると二番目と三番目の文節「論文は」「内容が」の 2 つの文節から係り受けがある．P_f は係り受けの数が一定以上の文節数を示す．P_d は一定以上離れた文節の組数である．被験者達のアンケート結果より，係り受けの回数 P_f におけるしきい値は $n = 5$ とし，係り受けの遠さ P_d におけるしきい値は $k = 15$ とした．実際に用いた読みやすい論文と読みにくい論文のパラメータの一例を以下に示す．

	読みやすい	読みにくい
P1	66.7	89.6
P2	1.15	0.99
P3	0	2
P4	2	7

この文書例では，読みやすさによって文書中の一つの文あたりの長さが約 22.3 文字，ひらがなと漢字の文字数比が 0.16 の差となっている．また，5 回以上の係り受けがある回数の差は 2 回であるが，15 文節離れている回数の差は 5 回あることがわかる．

4 分類結果

分類を行うために用いたツールを以下に示す．TinySVM は初期設定のまま分類を行った．

- 分類器：TinySVM 0.09
- 構文解析器：CaboCha 0.53

テキスト分類関連の論文を被験者 6 人で 100 件ずつ読み，読みやすさを決定した．その結果，読みやすい論文は 73 件，読みにくい論文は 27 件であった．表 1 に全論文データにおける一つの文あたりの平均文字数と係り受けの平均値を示す．表 1 より，読みにくい方が一つの文あたりの文字数が多いことがわかる．

表 1: 一つの文あたりの平均文字数と係り受けの平均値

	読みやすい	読みにくい
漢字 [文字]	21.0	23.4
ひらがな [文字]	24.5	27.2
カタカナ [文字]	5.5	7.7
一文の長さ [文字]	51.1	58.4
係り受けの回数	0.06	0.07
係り受けの遠さ	0.13	0.15

表 2: 正解率と精度

正解＼判定結果	易	難
読みやすい (易)	A	B
読みにくい (難)	C	D

評価値として，正解率，精度，再現率を用いた．それぞれの定義を表 2 および式 (1)~ 式 (3) に示す．

- 正解率：正解と判定結果が一致している割合

$$\text{正解率} = \frac{A+D}{A+B+C+D} \times 100 \ [\%] \tag{1}$$

- 精度：判定結果で読みやすいと判定されたうち，正解が読みやすい割合

$$\text{精度} = \frac{A}{A+C} \times 100 \ [\%] \tag{2}$$

- 再現率：正解が読みやすいうち，判定結果で読みやすいと判定された割合

$$\text{再現率} = \frac{A}{A+B} \times 100 \ [\%] \tag{3}$$

本研究では，読みにくい論文が読みやすいと判定される疑陽性の数を減らすために，今回は正解率と精度を重視した．精度が低いときには，読みやすいと判定しているにもかかわらず，実際は読みにくい論文である可能性が高い．また，精度が高いときには読みやすいと判定された論文の信頼性が高い．

読みやすい論文から教師データを 20 件と分類したいデータを 7 件，読みにくい論文から教師データを 20 件と分類したいデータを 7 件，それぞれランダムで選び，分類を 1,000 回行い，正解率と精度の平均値を求めた．その結果，平均正解率は 60%，平均精度は 64%となった．

5 おわりに

本研究では，文の長さ，ひらがなと漢字の割合，係り受けの回数，係り受けの遠さの 4 つのパラメータを用いて読みやすさを評価する手法を提案した．提案手法を用いることにより，教師データ 40 件において，平均正解率 60%，平均精度 64%の結果が出たが，平均正解率と平均精度ともに向上させる必要がある．

今後の検討課題として，文同士のつながりを見るために，接続詞を用いた分類手法で正解率，精度ともに向上させることである．

参考文献

[1] 近藤 洋介，松吉 俊，佐藤 理史，“ 教科書コーパスを用いた日本語テキストの難易度推定，” 言語処理学会第 14 回年次大会論文集，pp.1113–1116 (2008).