E-023

図書抄録縮約における読みやすさ向上方式の検討

Improving readability in sentence extraction

小峰　恒†　　絹川 博之†　　中川 裕志‡
Hisashi Komine　Hiroshi Kinukawa　Hiroshi Nakagawa

1．はじめに

近年，技術の発展に伴い，Web ブラウザを搭載した携帯電話や PHS などの携帯端末が増えてきた．それに伴い，携帯端末を用いての図書検索のニーズが高まってきている．ところが，既存の図書検索システムのほとんどがパソコンなど大きな画面を用いて検索結果を見ることを前提として作られているため，図書抄録はおよそ 200 字程度のものとなっており，画面の小さい携帯端末では一画面ですべてを表示することが出来ない．そのため，一画面で内容を把握できず，ボタンを何度も押す必要があり，扱いづらいものとなっている．

そこで，我々は無駄な箇所を省き重要な箇所を抽出して，携帯端末に表示できるように図書抄録を短く縮約する抄録縮約方式を提案している[1]．本稿では，提案した方式により抽出した縮約文の読みやすさ向上の方式を検討する．

2．図書抄録縮約方式

我々の提案している図書抄録縮約方式では，重要文抽出によって得られた文を要約としており，本方式は以下のようになっている．

2.1　抽出すべき抄録単位

図書抄録を調べた結果，重要個所は，複数の文に含まれることがわかった．ところが，2 文抽出すると，100 字を超えることが多く，携帯端末に表示するには長すぎる．そこで，抽出する抄録単位は文ではなく，節にした．なお，本方式では，節は句読点で区切られた単位と定義する．ただし，以下の条件を満たす場合，区切らないこととする．

① 直前が接続詞，係助詞である読点
　接続詞は前後をつなぐ物なので，後の節の一部とした．係助詞の場合は当該文節が主格になることが多いからである．

② 連続した名詞，未知語の区切りとして使われている読点
　読点は列挙の区切りとして使われている場合があり，文として分けられないからである．

以上の条件で区切られた単位を節と呼び，抽出する抄録単位とする．

2.2　正解データ

抽出すべき節を決めるため，以下のように正解データを作成した．

(1)　三人の正解データ作成者に抄録を読んでもらい，図書の特徴を示している節を二節選択してもらう．
(2)　(1)の結果から，正解データ作成者の多数決により，上位二節を正解の節とする．

本方式では，これによって選択された正解の節を正解節と呼ぶことにする．

2.3　抄録縮約処理方式

我々は，重要節抽出の手法として，文書頻度と節長を用いた方式を提案している．

提案する抄録縮約方式の処理は次のようになっている．

(1)　形態素解析器の茶筌[2]を用いて抄録を形態素解析し，品詞情報を取り出す．
(2)　(1)の品詞情報を用いて，抄録を節に分割する．
(3)　文書頻度による単語の重み付けを行う．対象となる単語は品詞が名詞もしくは未知語である．ただし，非自立語の名詞は含めない．
(4)　節ごとに単語重み計算の対象となる単語の個数を算出する．本方式ではこれを節長と見なす．
(5)　節ごとに，単語の重みの和と節長を組み合わせて，節の重みを計算する．
(6)　(5)の節の重みから，上位二節を抽出する．

以上の処理によって選択された二節がシステムによって正解節だと判断された節となる．

2.4　抽出節の読みやすさ向上

本方式では，節抽出による要約文の作成を行っているため，読みやすさを向上させるためには以下が必要である．

(1)　語彙補填
　節の区切りが文の途中で終わっている場合に，文として成り立たせるため，語彙を補填する．
(2)　重複出現語の削除
　異なる文中にある節を抽出することによって，重複する単語などが出てくる場合があり，文として冗長となるので重複出現語を削除する．
(3)　指示語の同定

3．語彙補填処理方式

3.1　方式の検討

読みやすさ向上に関し，本稿は語彙補填のうち，動詞補填方式を検討する．この問題は，文の途中で節が区切られてしまい，重要である節が区切られた文の前半部分である場合などに起こる．文の前半部分だけを抽出するため，述語などがない文になってしまい，文として完結しておらず，意味が通じない．解決方法として，抽出されなかった文の後半部分の節から動詞を補う，もしくは，抽出された節の末尾の動詞を終止形に変形させる方法が考えられる．

本稿では，以上の二つの方法を用いた動詞補填処理方式を提案する．

3.2　動詞補填処理方式

動詞補填方式の処理の流れを以下に記す．なお，抽出された節のうち，先に出現する節を節 A，後に出現する節を節 B とする．

(1)　形態素解析を用いて節 A の末尾に動詞があるか調べる．ただし，次の場合は例外として処理をする．

† 東京電機大学大学院 工学研究科
‡ 東京大学 情報基盤センター

① サ変動詞で直前にサ変接続名詞がある場合，サ変接続名詞とサ変動詞を組み合わせて動詞と見なす．
② サ変接続名詞と句点の組み合わせを動詞と見なす．

①の場合，動詞『する』は直前のサ変接続名詞と組み合わせることにより一つの意味になるため，直前の名詞も含める．②は体言止で使用されており，動詞『する』が省略されていると考えられるため，動詞と見なす．

(2) (1) で末尾に動詞があった場合，動詞を終止形にする．ただし，同一文内に節 B がある場合，節を繋げるため，そのままにする．

(3) (1)で末尾に動詞がなかった場合，節 A と同一文内である節のうち，節 A の後にある節で，尚かつ，節 B より前の節を対象に，形態素解析を用いて動詞を探し出す．

(4) (3)で動詞が見つかった場合，その動詞を含んだ文節を抽出し，節 A の後に繋げる．ただし，節 B が同一文内にない場合，動詞 を終止形に直す．(3)で動詞が見つからなかった場合は，そのままにする．

(5) 節 B に対して，末尾に動詞があるが調べる．

(6) (5)で末尾に動詞がある場合，動詞を終止形にする．

(7) (5)で末尾に動詞がなかった場合，節 B と同一文内である節のうち，節 B の後にある節を対象に，形態素解析を用いて動詞を探し出す．

(8) (7)で動詞が見つかった場合，探し出された動詞を含んだ文節を抽出し，動詞を終止形に直し，節 B に繋げる．(7)で動詞が見つからなかった場合は，そのままにする．

(9) 補填済みの節 A と節 B を繋げる．要約文が読点で終わっている場合，読点を句点に変える．

4 ．動詞補填処理方式の実験

3 章の動詞補填処理方式を用いることにより，文として適切な形になるのか実験を行った．図書抄録として，東京大学付属図書館のブックコンテンツ[3]を用いた．

4.1 実験対象

今回の実験対象となる図書データの図書数及び，今回のスムージングの対象となる図書数を表 1 に示す．ここで述べた対象図書とは，3.2 節の処理を用いたことによって要約文が変更されたものを指す．各分野，およそ半分が評価の対象となる．

表 1　実験対象図書データ

分野	図書数	対象図書数
エレクトロニクス	314	142
物理	310	156
法律	339	166

4.2 評価方法

今回の実験では，3.2 節の節 A，節 B は人手で作成した正解節（2.2 節参照）を用いて評価する．

評価は人の手により判断する．動詞補填を行うことにより，文として体裁が整っているか，また，元の文と比較して意味が変わっていないかを調べる．なお，本実験は 2.4 節(1)に対する対処の検討であるため，評価の際に，(2)(3)の問題による読みやすさの低下については考慮しない．

以上の条件で実験を行い，評価を行う．

4.3 実験結果

各分野において，本方式を用いて，要約文に読みやすさ向上が見られた対象図書の数，及び割合を表 2 に示す．

A：動詞の補填が正しく，他の語の補填が不要なもの
B：動詞の補填は正しいが，他の語の補填が必要なもの
C：補填が誤っているもの

括弧内は対象図書を全体としたときの割合を示している．

表 2　動詞補填の精度

分野	A	B	C
ｴﾚｸﾄﾛﾆｸｽ	57（40%）	69（49%）	16（11%）
物理	66（42%）	79（51%）	11（7%）
法律	50（30%）	98（59%）	18（11%）

4.4 実験結果の考察

(1) 実験の結果，動詞の補填は 90%正しい．しかし，そのうちの 60%前後は動詞だけの補填では不十分であることがわかった．

(2) 補填した動詞のヲ格，ニ格，ガ格のうち，当初の抽出節に含まれない文節が，動詞に加えて補填の候補になりうると考えられる．

(3) 正しく動詞が抽出できなかった物として，サ変接続名詞に続く動詞として，「なる」「ある」「できる」などがあり，その場合を考慮に入れなかったため，文として意味が通じなくなる場合がある．また，サ変接続名詞と「する」の間に助詞があり，サ変接続名詞を取り出せないこともあった．

(4) C に分類される原因としては，サ変接続名詞ではない体言止である述語や，形態素解析の誤解析による動詞抽出のミスなどがあった．

5 ．おわりに

5.1 本稿の成果

節単位の抽出による図書抄録縮約方式において，未抽出節の動詞の補填による読みやすさ向上方式の検討を行った．

提案した方式を用いることにより，動詞の補填は約 90%が正しく行うことが出来た．

5.2 今後の課題

抄録縮約方式の読みやすさ向上に関する今後の課題は以下の通りである．

(1) 語彙補填における動詞以外の補填
動詞補填に加えて，目的語や主語などの補填が必要となる場合がある．

(2) 語彙補填以外の問題
重複出現語の削除，指示語の同定の各処理方式は今後検討していく．

読みやすさの向上に際して，携帯端末の表示長の制約を考慮に入れる必要がある．

参考文献

[1]小峰恒,絹川博之,中川裕志：“単語の文書頻度と文の長さを利用した抄録縮約方式” 情報処理学会自然言語処理研究会研究報告,NL-149-11,pp.73-80,May2002.

[2]形態素解析器「茶筌」：
http://chasen.aist-nara.ac.jp/

[3]東京大学付属図書館ブックコンテンツ：
http://contents.lib.u-tokyo.ac.jp/contents/top.html