

意味変化率を考慮した文書短縮方法

辻 孝子† 安藤 一秋†
 獅々堀 正幹† 青江 順一†

形態素置換による文書短縮法では、文解釈の意味変化と処理の偏りが生じる。本論文では、意味変化の問題に対して、意味変化率の低い順に短縮規則を分類し、この順位に従った短縮アルゴリズムを提案する。また、偏りの問題に対しては、短縮処理が文書全体に一律となる新しい基準を定義して、短縮アルゴリズムに導入する。提案手法の有効性は、20人の被験者で行った実験結果により、実証される。

A Method of Reducing Texts Using a Meaning Distortion Rate

TAKAKO TSUJI,† KAZUAKI ANDO,† MASAMI SHISHIBORI†
 and JUN-ICHI AOET†

Text reduction assistance using morpheme replacements is useful for text editing, but it has two problems: one is, that the sentence's meaning is sometimes altered and another is, that reductions are restricted to only some particular paragraphs of a text. In order to solve these problems, this paper defines the priority of a meaning distortion rate and the criterion of reducing texts uniformly. A reduction algorithm is presented and the effectiveness is evaluated by the simulation results for a variety of texts.

1. ま え が き

ワードプロセッサやデスクトップパブリッシングにおいて、文書の編集支援は重要な課題である。制限サイズに文書を収めるための文書短縮も1つの重要な支援であり、津田ら¹⁾は、形態素置換に基づいた短縮法を提案した。しかし、この方法(対象法と呼ぶ)では文解釈に意味変化が生じる場合がある。この問題は、短縮の判断を利用者に委ねることで解決できるが、短縮支援は最終的編集作業で有用なので、意味変化率の少ない短縮候補を優先して提示するのが望ましい。また、対象法では処理が一部の段落に偏り、短縮後の漢字比率の増加などから生じる表現上の堅さも偏るので、短縮処理を文書全体で一律に行うことが望ましい。

本論文では、まず短縮規則に対する意味変化率の順位を決定する。そして、この優先順位に従った短縮法を提案し、意味変化の問題を解決する。また、偏りの問題に対しては、処理段落を均一にする新しい基準を

定義し、この基準を短縮法に導入して解決する。

2. 短縮規則と意味変化の順位

2.1 対象法の概要

段落 P の最終行の文字数 REAR(P) より多くの文字数を短縮すれば、段落 P は 1 行削減できるが、段落 P の文字列長 LEN(P) が長いほど短縮候補が多いので、対象法では、1 行削除の容易度 $PRI(P) = LEN(P) / REAR(P)$ を定義し、 $PRI(P)$ が大きい段落 P の処理を優先する。偏りの問題は、この容易度の基準により生じる。

2.2 短縮規則の分類

本論文では、意味変化率が議論しやすいように、文法的に類似した定型表現^{3),8)}に基づき規則を拡充・分類する。replace(x) は定型表現 x の置換候補を表し、名詞を表す形態素を N と記述する。

【大分類 1】関係表現に基づく規則

助詞に短縮できる定型表現を関係表現⁶⁾を 1-K と 1-S に小分類し、1-S の広義な解釈として、さらに小分類 1-T, 1-X, 1-Y, 1-Z を提案する。

【小分類 1-K】活用語を含まない関係表現 K.

† 徳島大学工学部知能情報工学科

Department of Information Science and Intelligent Systems, The University of Tokushima

- N K → N replace(K)
 事故が原因で → 事故で
 [小分類 1-S] 活用語を含む慣用的な関係表現 S (活用語の意味が弱い慣用的表現).
 N S N' → N replace(S) N'
 文化に関する研究 → 文化の研究
 [小分類 1-T] 活用語を含む関係表現 (S 以外) T.
 N T N' → N replace(T) N'
 外国に住んでいる友人 → 外国の友人
 [小分類 1-X] T の関係表現で活用語が次の名詞 N' も含めて名詞句に置換できる定型表現 X.
 N X → N replace(X)
 機器を使うこと → 機器の使用
 画像が美しいことを → 画像の美しさを
 [小分類 1-Y] T の定型表現で N' が名詞以外の品詞で、それも含めて置換できる定型表現 Y.
 N Y → N replace(Y)
 転部を希望したけれど → 転部の希望に反し
 [小分類 1-Z] 疑問詞節を含む定型表現 Z
 N Z → N replace(Z)
 問題をどう対処するか → 問題の対処方法
 【大分類 2】 助述表現と機能動詞に基づく規則
 用言の補助的な意味を付加する定型表現を助述表現と呼び、首藤⁶⁾によるもの 2-J に分類する。機能動詞⁷⁾は、実質の意味を前の名詞にあずけている表現であり、2-D に分類する。用言を M として記述する。
 [小分類 2-J] 首藤⁶⁾による助述表現 J
 M J → M replace(J)
 書いたものである → 書いた
 [小分類 2-D] 機能動詞表現 D による規則⁵⁾.
 M D → M replace(D)
 検索を実行する → 検索する
 【大分類 3】 複合語への短縮規則
 宮崎⁵⁾の複合語の文法規則から短縮規則を定義する。
 [小分類 3-B] N と N' の間の定型表現 B の消去。
 N B N' → N N'
 受験するための資格 → 受験資格
 【大分類 4】 同義語への短縮規則
 [小分類 4-R] 定型表現 R を同義語に短縮する場合で、同一品詞への置換、R 自体の削除などがある。
 R → replace(R)
 しかしながら → しかし (接続詞の同義語)
 これらのソフトは → これらは (名詞の省略)
- ### 2.3 意味変化率の順位
- 短縮処理で生じる意味変化を調査するために、EDR の日本語コーパスなどを用い、以下に示す短縮候補

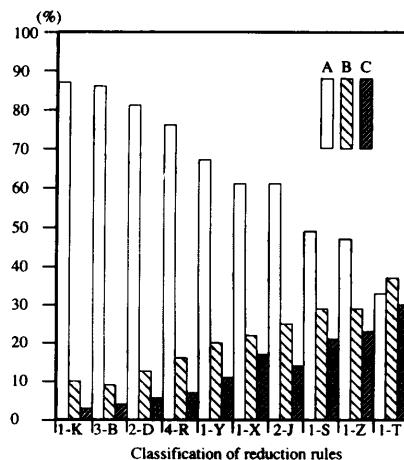


図1 意味変化の実験結果

Fig. 1 Experimental results of meaning transformation.

(約 1,000 個) を被実験者 20 人に対して実験を行った。
 【短縮候補の例】 この 10 年間における [の (S)], パーソナル・コンピュータの普及にはめざましいものがあった [はめざましかった (J)].

下線部分を [] 内の表現に短縮したとき、意味変化なしは A; 意味は少し変化するが、文の解釈には影響が少ない場合は B; 意味変化により文の解釈に問題が生じる場合は C と評価する。この実験より (図 1), 意味変化率の優先順位を A の比率が多い順で決定した。さらに、順位が最も低い分類 1-T に対して、T の前後の名詞組 (N1, N2) の意味素性の組⁸⁾を分析し、曖昧性の少ない順位を次のように決定した。

- 順位 1: (職業, 人間); 順位 2: (人間, 贈り物);
- 順位 3: (組織, 人間); 順位 4: (衣服類, 人間);
- 順位 5: (材料, 制作物); 順位 6: (目的, 手段);
- 順位 7: (場所, 人間); 順位 8: (人間, 具象物)。

順位 1 の例 “医師の兄” では、“医師になった兄”, “医師である兄” など推測が限定されるが、順位 8 の例 “母の絵” では、購買 (母が買った絵), 所有 (母が持っている絵), 作成 (母が描いた絵) など曖昧である²⁾ので、優先順位を低くすべきである。なお、分類 1-S にもこの順位が適用できる。

3. 短縮処理アルゴリズム

2.3 節で提案した細かい順位で応用すると、同種の規則が偏って使用されるので、意味変化が少なくても偏りの問題が生じる。したがって、次の 3 段階のグループに分けて、この偏りを緩和する。

- 1 段階: 1-K, 3-B, 2-D, 4-R
- 2 段階: 1-Y, 1-X, 2-J
- 3 段階: 1-S, 1-Z, 1-T

この段階を全域変数 LEVEL で表し、1 段階から処理を進める。また、各段落 P には解析位置 (先頭からの文字列長) を全域変数 LOC(P) で表し、未処理量の比率が高い段落を優先する次の基準を導入する。

$$\text{NEW_PRI}(P) = \text{LEN}(P) / \text{LOC}(P)$$

【短縮アルゴリズム】

入力：短縮目標行数 NUM

出力：減少した行数 NUM

手順 1：LEVEL=0 とする。

手順 2：LEVEL が最終段階 3 になれば、NUM を返し終了する。LEVEL < 3 ならば LEVEL をインクリメントし、すべての LOC(P) を 1 に初期化する。

手順 3：LOC(P) が段落の長さ LEN(P) を越えていないすべての段落 P に対して、NEW_PRI(P) が最も大きい段落 P を見つけ、次へ進む。ただし、現在 LEVEL で段落候補 P がなくなれば、手順 2 へ進む。

手順 4：手続き MATCH(P) の短縮処理で、NUM が零になれば、NUM を返して終了する。そうでなければ、手順 3 へ進む。(アルゴリズム終了)

【照合手続き MATCH(P)】

手順 (M-1)：段階 LEVEL の規則照合を LOC(P)^{*1} から開始し、短縮規則 α が見つければ次の手順 (M-2) へ。規則照合なしに最後まで解析が進めば、終了する。

手順 (M-2)：短縮する^{*2}場合は手順 (M-3) へ。短縮しない場合は、手順 (M-1) へ戻り、解析を再開する。

手順 (M-3)：規則 α で短縮し、REAR(P) と LEN(P) を補正^{*3}後、手続き終了する。ただし、REAR(P) ≤ 0 ならば (1 行の短縮)、NUM をデクリメントし、手続きを終了する。(手続き終了)

提案法では、未処理の残量率が多い段落を優先する新基準 NEW_PRI(P) により、処理対象となる段落を短縮ごとに変更することで、処理の偏りを緩和する。

4. 提案手法の評価

4 種類の文書 (詳細は表 1 の下) に対する被験者 20 人の実験結果を表 1 に示す。短縮行数は、20 人の平均値、短縮率は全行数に対する短縮行数の割合 (%) である。本論文では、対象法より規則を拡充したので、短縮率は対象法短縮率 (%) より向上した。規則の適用回数は、20 人の選択総数である。

図 2 は、短縮候補数と 20 人の平均選択率の関係を示し、太実線が提案法、細実線が対象法を表す。図 2

^{*1} 実際の LOC(P) の制御位置は、形態素解析を正しく再開するために、短縮処理の行われた文末とする。

^{*2} 置換するか否かの判断は、利用者が行う。

^{*3} READ(P), LEN(P) とともに、短縮分だけ減少する。

表 1 文書短縮処理の実験結果

Table 1 Experimental results of text reductions.

文書名	文書 1	文書 2	文書 3	文書 4
文書情報				
形態素数	2,995	2,530	3,903	4,682
段落数	38	33	42	36
文数	103	77	135	114
行数	466	299	333	341
1 行文字数	12	16	25	24
短縮結果				
短縮行数	28.3	47.6	35.1	47.2
短縮率 (%)	6.0	15.7	10.5	13.8
対象法短縮率 (%)	4.5	9.7	6.3	7.0
短縮規則の適用回数				
1-K	301	589	283	486
3-B	723	996	479	358
2-D	86	102	132	158
4-R	613	1,204	2,748	2,321
1-Y	19	182	159	101
1-X	134	141	381	545
2-J	669	1,663	9,987	10,357
1-S	239	289	804	997
1-Z	61	23	41	59
1-T	241	631	526	797

文書 1：毎日、朝日、読売各社の 1991/3/25 朝刊“選抜高校野球”記事および“兵庫県農高答案改ざん”記事

文書 2：日経パソコン、1991/2/4, pp.89, 164-168; 1991/2/18 p.159

文書 3：青江：ダブル配列による高速デジタル検索アルゴリズム, bit, Vol.21, No.6, pp.776-784

文書 4：中村ら：ニューラルネットによる英単語品詞予測モデル, 信学論 D-II, No.1, pp1-7 (1991).

より、提案法の選択率は前半に高く、後半には低下する。しかし、推敲作業後の記事では、5%程度の行数が短縮できれば有効¹⁾とのことであるので、文書 1 の新聞記事^{*4}を除けば、選択率の高い短縮候補だけで短縮処理が終了できる。したがって、提案法は、意味変化の問題に対して有効であるといえる。

短縮候補総数と処理した段落数の割合 (%) を図 3 に表す。ただし、比較を容易にするために、提案法の 3 段階指定をなくした結果を示す。図 3 より、提案法の処理段落数は対象法に比べて非常に早く増加するので、偏りの問題に対する有効性が分かる。

対象法の PRI(P) は短縮処理を早く終える基準であるが、提案法の NEW_PRI(P) は、対象段落を分散させる基準であるので、目標行数 NUM を零にする処理時間が提案法では長くなるように思える。しかし、NUM を全体行数の 5%程度とするならば、意味変化率の導入の効果により、短縮候補の選択率が向上する

^{*4} 新聞記事は、記述が簡潔なので、短縮率は高くない。

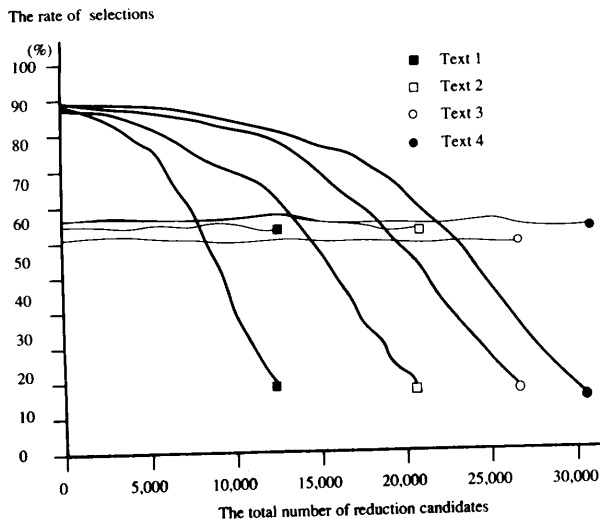


図2 短縮候補総数と選択率の関係
Fig. 2 Relationships between the total number of reduction candidates and the rate of selections.

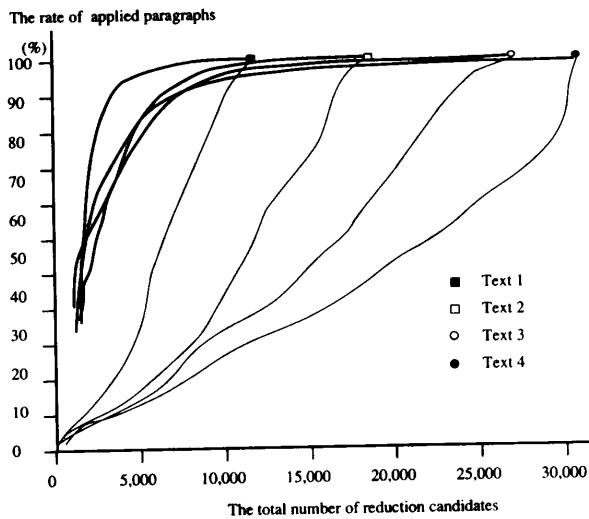


図3 短縮候補総数と適用段落数の比率の関係
Fig. 3 Relationships between the total number of reduction candidates and the rate of applied paragraphs.

ので、逆に処理時間は早くできる。

5. まとめと今後の課題

以上、文書短縮で生じる意味変化と偏りの問題の解決法を提案し、その有効性を実証した。優先順位の低い意味素性の組(“母の絵”)は、曖昧性が高いことを意味するので、校正支援⁴⁾での指定対象としては、逆に高い優先順位として有用である。今後は構文意味解析を利用した高度な短縮規則の考案が課題となる。

参考文献

1) 津田和彦, 中村雅巳, 青江順一: 形態素置換によ

る文書短縮法, 電子情報通信学会論文誌, Vol. J75-D-II, No.3, pp.619-627 (1992).

2) 島津 明, 内藤昭三, 野村浩郷: 助詞「の」が結ぶ名詞の意味関係の subcategorization, 情報処理学会自然言語処理研究会, 53-1 (1986).
3) 新納浩幸, 井佐原均: 疑似 N グラムを用いた助詞的定型表現の自動抽出, 情報処理学会論文誌, Vol.36, No.1, pp.32-40 (1995).
4) 林 良彦, 菊井玄一郎: 日本文推敲支援システムにおける書き換え支援機能の実現方式, 情報処理学会論文誌, Vol.32, No.8, pp.962-970 (1991).
5) 宮崎正弘: 係り受け解析を用いた複合語の自動分割法, 情報処理学会論文誌, Vol.25, No.6, pp.970-979 (1984).
6) 首藤公昭: 文節構造モデルによる日本語の機械処理に関する研究, 福岡大学研報 45 自然科学編, 6, pp.88-119 (1980).
7) 村木新次郎: 日本語の機能動詞表現をめぐって, 国立国語研究所報告 65 研究報告集 (2), pp.17-75 (1980).
8) 辻 孝子, 安藤一秋, 獅々堀正幹, 青江順一: 活用語を含む助詞的定型表現の分析, 第 52 回情報処理学会全国大会論文集, 3 分冊, B-2, pp.101-102 (1996).

(平成 8 年 7 月 22 日受付)

(平成 8 年 9 月 12 日採録)

辻 孝子 (正会員)



平成 2 年上智大学文学部心理学科卒業。野村證券(株)入社。退社後現在徳島大学工学部受託研究員。自然言語処理の研究に従事。

安藤 一秋 (正会員)



平成 4 年徳島大学工学部知能情報工学科卒業。平成 6 年同大学大学院博士前期課程修了。現在同大学大学院博士後期課程在学中。自然言語処理の研究に従事。

獅々堀正幹 (正会員)



平成 3 年徳島大学工学部知能情報工学科卒業。平成 5 年同大学大学院博士前期課程修了。現在同大学工学部知能情報工学科助手。自然言語処理の研究に従事。情報処理学会第 45 回全国大会奨励賞受賞。