1V-1

単語の難易度比較を用いた文章簡略化システム

野崎 徹郎 † 奥村 紀之 †
† 長野工業高等専門学校 電子情報工学科

1 はじめに

これまでに要約や言い替えの手法の研究は，自然言語処理の分野で盛んに行われてきており [1][2]，様々な観点から新しい手法の提案されてきた．本研究では，それらの技術を組み合わせ，文章を簡略化する手法を提案する．照応解析により主語と目的語を明確にしたテキストを要約し，テキスト内の単語の語義の曖昧性を解消した後に単語の言い替えを行うという手順でシステムを構築する．

2 提案手法

本稿で提案する手法は，図 1 に示す手順でテキストを処理することにより簡略化された文章を生成するものである．類義語の抽出は概念辞書の WordNet を用いて文中の名詞が持つ全ての概念を抜き出し，その概念に含まれる単語を抽出することにより行い，コーパスはブラウンコーパスを使用した．

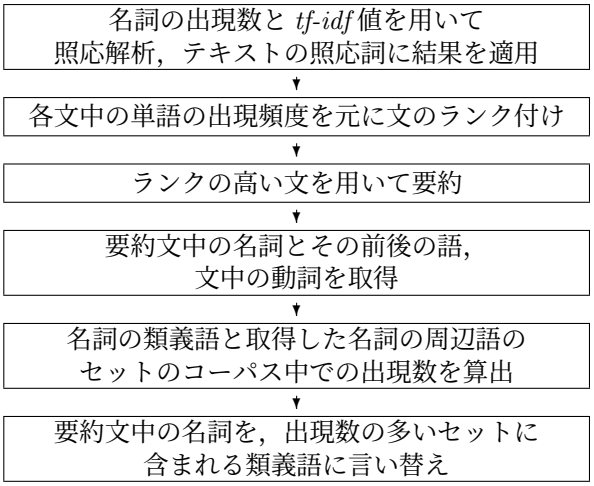

図 1: 提案手法の手順

3 評価実験

提案する手法で実装した照応解析・言い替え・語義の曖昧性解消の精度の評価を行うため，Project Gutenberg と BBC のホームページで公開されているテキスト [3][4] を簡略化する実験の結果を比較した．以下，照応解析 (実験 1) と言い替え（実験 2)，語義の曖昧性 (実験 3) の精度を調査する実験について述べる．

3.1 実験 1

照応解析は，文章中に含まれる名詞の *tf-idf* 値と出現回数を組み合わせた値を元に行った。具体的には各文について，その文より前の文とその文自身に含まれる名詞のそれらの値のうち最も適切なものを先行詞と同定する手法である。この精度を，手作業での判定により算出した．構築したシステムにより照応解析をした結果と，手作業によりテキストに出現する照応詞の先行詞同定を行った結果とを比較し，一致するものを正解とした．このシステムの精度を調査した結果を表 1 に示す．このシステムでは先行詞の候補を単純に候補となる語の *tf-idf* 値とテキスト中での出現回数のみで比較しており，あまり高い精度は得られなかった．

表 1: 照応解析の精度

テキスト	正答数/全照応回数	精度
テキスト 1	72/183	39.3 %
テキスト 2	11/45	24.4 %

3.2 実験 2

言い替えを試行した全てのパターンについて，適切であるか否かの判定を手作業で行った結果のうち，言い替えの試行数が多い 1 つ目のテキストに関するものを表 2 に示す．*tf-idf* 値から *idf* 値へ使用する値を変更したことで精度は落ちたが，語が出現する文書の数のみに着目すべきであり語の分布が特異な文書に引きずられるべきではないので *idf* 値を用いて実験を行った。不適切な言い替えの主な判定理由として他の品詞への誤変換，文中で使用されているものと異なる意味の類義語の取得が挙げられた．照応解析と語義の曖昧性解消を適用することによりこの 2 つのシステムを適用する前と比較して精度が向上した．

3.3 実験 3

要約文中での語義の曖昧性解消の全試行について，適切に行われたか否かの判定を手作業により行った結果を表 3 に示す．このシステムは入力となる名詞と，

A Simplification Method Using Comparison of Word Difficulties.

† Tetsuro NOZAKI
† Noriyuki OKUMURA
Nagano National College of Technology, Department Electronics and Computer Science, noriyuki_okumura@ei.nagano-nct.ac.jp (†)

表 2: 文献 [3] のテキストでの言い替えの精度

利用する値	照応解析	語義の曖昧性解消	適切な言い替え/言い替え試行数	精度
tf-idf	不使用	不使用	25/44	56.8 %
idf	不使用	不使用	21/44	47.7 %
idf	使用	不使用	63/87	72.4 %
idf	使用	使用	75/90	83.3 %

ある属性をもつ語とのセットのコーパス内での出現回数をパラメータとし，そのパラメータによって語義の曖昧性を解消している．その属性をもつ語として文中での名詞の前後の語，名詞の含まれる文中の動詞を利用している．

表 3: 語義の曖昧性解消の精度

テキスト	正答数/試行数	精度
テキスト 1	74/92	80.4 %
テキスト 2	24/26	92.3 %

4 考察

提案した手法による言い替えの一例を以下に示す。

many of these people felt
a genuine allegiance to the Great Leader.

↓

many of these *people* felt
a **true loyalty** to the *Great Leader*.

言い替えられた語は太字，元の語が維持されたものは斜体で示している。システム内でより難易度の低い単語に言い替えられると判断された場合のみ別の単語に言い替えられている。言い替え・語義の曖昧性解消については，今回の実験で利用したパラメータに加え，更に別の有効な属性を持つ語を利用することで精度が向上する可能性がある．また，照応解析の有無に関する言い替えの精度の変化は，表 2 におけるその前後での試行回数と成功数を見る限り，照応詞の先行詞の殆どについてうまく照応解析される状況が偶然できたことによるものであると考えられる．そのため，先行詞のばらつき方によっては照応解析による精度の向上はあまり期待できず，語義の曖昧性解消による向上のみが目立つという場合も考えられる．不適切な言い替えが行われた原因を調査した結果から考えられる言い替え・語義の曖昧性解消に対する今後の課題は，名詞以外の品詞を取得することの回避と極めて広い意味と用法を持つ一般的な語 (have や get 等) に対する対処法の模索の 2 点である．

5 おわりに

本研究では，照応解析により主語と目的語を明確にした文章を要約し，要約文中の名詞の語義の曖昧性を解消した後に名詞を言い替えることにより文章を簡略化する手法を提案した．この手法の中で，照応解析は 25〜40 %，言い替えは 85〜95 %の精度で行われ，照応解析を手作業で行えば自然な文章を生成することができるという結果が得られた．また，照応解析の精度は，その次の手順の要約は各文中で出現する単語により前後の文や全体の中でのその文の重要度を決定する手法であるために要約結果に深く関わることになる．そこで，簡略化後の文章をより適切なものにするためには，照応解析の精度を向上させることも必要となる．今後の課題としては，照応解析の方法を根本的に見直して改良することと語義の曖昧性解消に用いるパラメータを増やし，パラメータごとに重み付けを行うことでそれぞれの精度を向上することが挙げられる．

謝辞

本研究の一部は科研費 (23720222) の助成を受けたものである。

参考文献

[1] 「*Applying Co-Training to Reference Resolution*」Christoph, M., Stefan, R., Michael, S.: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 352–359,2002.

[2] 横山晶一，菅野崇，西原典孝:主題・焦点リンクを用いた重要文抽出システム，情報処理学会，NL, No.156, pp.1–6,2003.

[3] 「*Through The Looking-Glass*」Lewis Carroll：Project Gutenberg (http://www.gutenberg.org/ebooks/12).

[4] 「*How genuine are the tears in North Korea?*」Tom Geoghegan：BBC News (http://www.bbc.co.uk/news/magazine-16262027).