主題・焦点に基づく文統合工程を実装した要約システム

天野 禎章　　横山 晶一　　橋本 力
山形大学大学院理工学研究科

1. はじめに

インターネットの目まぐるしい普及により情報洪水と呼ばれて久しい昨今、夥しい量の文書の海から有用性を即断することは、人手では不可能と言える。そこでユーザの負荷軽減と迅速な判断の補助として、新聞には見出しが、ニュースにはテロップが、検索結果にはリンク先テキストの一部が提示されている。これらは皆、内容を要約した結果である。

算出した重要度を基に文単位で抜き出す重要文抽出では、要約率を高めると、出力結果が首尾一貫性を欠くという問題点が生じる。これは先行詞の消失や、連文の欠落による話題の飛躍などで、文間の繋がりが悪化するからである。

したがって、より人間らしい自然な要約を生成するには、重要文抽出で明らかに不要な文章をカットした後、一文レベルで文短縮する必要がある。つまり、重要文抽出だけでは不充分であり、二つのプロセスを経ることで重要文抽出での要約比率を減らし、話題が跳躍せずに文書を縮められるのである。

この考えに基づき、本システムでは、冗長性の少ない要約生成を目指して、重要文抽出結果からプロトタイプを合成し、自由作成要約時に人が行う編集処理を施した。

2. 自由作成要約に基づく編集操作

過去の研究結果[1]によると、人は自由作成要約時に、次の編集操作を行うと述べられている。

(A)複数文の統合　(B)構文的変形
(C)不要表現の削除　(D)語彙的言い替え
(E)抽象化と具体化　(F)文の並べ替え

先行研究では、自由作成要約と原文の対応コーパスから得られた情報を活用し、分析して生み出した規則で(A)を、大規模な文法知識と文脈情報(WordNetによる類義、反義の関連性など)と統計情報で(C)を、実装している。

本研究では、コーパス情報に依らず、主題と焦点を要素に日本語彙大系[2]を用いて(A)と(B)を同時に行い、定義したパターンマッチで(C)を、形態素タグとパターン辞書の一致による(D)を処理するシステムを実装した。

主題・焦点の定義

重要度算出とプロトタイプ作成、並びに文統合基準に、下記で定義する主題と焦点[3]を利用した。

主題:文中で話題となっている要素であり、前述された既知の情報(名詞)
焦点:その文で新しく導入された情報(名詞)
例)通常の人の**インフルエンザは**、のどや気管の粘膜だけに感染するが、全身の細胞に感染する強毒性のタイプも知られている。
(主題:**太字下線部**　焦点:二重下線部)

前文の主題(既知の情報)と後文の焦点(新情報)、あるいは前文の焦点と後文の主題に強い類似性がある場合は、同じ話題について展開されている(冗長性有)と見なせる。よって、共通話題部の削除が可能であり、一文にまとめられる。

3. システムの流れ

重要文抽出結果を入力にして、システムを実行する。プログラムの流れは、次の通りである。

①:プロトタイプの作成　②:プロトタイプの拡張
③:複数文の統合と構文的変形
④:不要表現の削除と語彙的言い替え
これら四つのプロセスを、逐次的に行う。

3.1 プロトタイプの作成

主題をS、焦点をO、述語成分をVと見立て、最小単位の文の成型を意図し、プロトタイプ(「主題＋焦点＋述語成分」)を作成した。述語成分には、最後に出現した主題・焦点の後半部分を繋げている(「--」が区切り文字)。

例)--インフルエンザは--タイプも知られている。

3.2 プロトタイプの拡張

作成されたプロトタイプをそのまま利用するには、あまりにも欠落が著しいので、次の二つの条件に当てはまる場合に拡張を実行した。

ⅰ) 主題・焦点が形式名詞

ⅱ) プロトタイプが閾値を満たさない

閾値には、要約率を満たしながら、かつ充分な拡張を行うため、次の数式を採用した。

$$\frac{PTの文字数}{INの文字数} \geq (要約率 \times 重要度)$$

重要性の高低でプロトタイプの長さが調節されるように、式には重要度(限界値を越えないために、場合分けして設定した)を取り入れた。これを満たすまで補完することで、指定要約率に準じた要約生成が可能となる。

3.3 複数文の統合・構文的変形

拡張されたプロトタイプを条件に応じて一文に集約する。「前行の主題と作業行の焦点」もしくは、「前行の焦点と作業行の主題」が類語関係にある場合を統合条件とした。表1の例では、文番号20の焦点である「タイプも」と文番号21の主題の「タイプは」にあたる(二重下線部)。このとき、重要文抽出で抜け落ちた文も検索したことで、文番号20がたとえ欠落しても、先行詞の補完が可能となった。

類語関係の判定には、日本語語彙大系を利用した。分類番号が一致するときを類語関係と定めた。例では、同じ単語であるため、全て一致する。

表1 入力例と複数文の統合結果

入力例	20 通常の人のインフルエンザは、のどや管の粘膜だけに感染するが、全身の細胞に感染する強毒性のタイプも知られている。	
	21 このタイプは症状が激しく、肺炎や脳症など合併症を引き起こす危険性が高い。	
プロトタイプ	20 通常の人のインフルエンザは、--全身の細胞に感染する強毒性のタイプも知られている。	
	21 このタイプは--肺炎や脳症など合併症を引き起こす危険性が高い。	
	前行処理	通常の人のインフルエンザは、--全身の細胞に感染する強毒性のタイプも知られ、
	作業行処理	--肺炎や脳症など合併症を引き起こす危険性が高い。
統合後	通常の人のインフルエンザは、--全身の細胞に感染する強毒性のタイプも知られ、--肺炎や脳症など合併症を引き起こす危険性が高い。	

前行の統合対象では、行末を次の場合に準じて変化させた。

(a) 動詞文では、「連用形＋読点」に

(b) 形容詞文では、「く＋読点」に

(c) 名詞文では、「で＋読点」に

文の分類には、MeCab[4]による形態素解析結果を用いた。対象となる三品詞が最後に出現する位置を比較し、最大値を取る品詞を分類とした。

作業行の統合対象では、類語関係を満たした主題と焦点のタイプに応じて変化させた。

(1) 主題の場合には、対象となる主題を含むプロトタイプを削除する

(2) 焦点の場合には、対象となる焦点までのプロトタイプを削除する

これらの統合処理を経た結果を互いに連結して出力した(表1の統合後)。

3.4 不要表現の削除

プロトタイプから不要と定義した表現、比喩指標要素[5]が存在する直喩と丸括弧内の文章を不要表現として削除した。

3.4.1 比喩表現の削除

各品詞の比喩指標要素を目印に喩詞(「AのようなB」や「BはAのようだ」のAにあたる)を削除した。このとき、片方しか指標要素が存在しない場合は、プロトタイプの区切り文字までを喩詞の範囲と定めている。

3.4.2 丸括弧内の削除

丸括弧に関する研究は過去に行われており、システム化されている[6]。今回は通し番号としての使用を除き、全てを削除した。

3.5 語彙的言い替え

プロトタイプを入力セグメントとして、形態素解析タグと言い替えパターン辞書とのマッチングを図った。パターン辞書は、類語大辞典[7]の分類と、大辞林[8]の語義文から作成している(用例や記号などは除去してある)。表2が言い替えパターン辞書の一例である。このパターンと入力がマッチしたとき(許容範囲に±1を設定)を言い替え候補とした。

表2 言い替えパターン辞書

コード番号と見出し語	言い替えパターン(名詞、動詞、形容詞、副詞)
0901a27 号泣する	<noun-大声><verb-あげる> <verb-泣き叫ぶ>
0901a28 嗚咽する	<noun-声><verb-詰まる> <verb-泣く><verb-むせび泣く>
0901a29 哭する	<noun-大声><verb-あげる> <verb-泣き叫ぶ>
	<noun-古代><noun-死者> <verb-とむらう><noun-礼> <noun-大声><verb-泣き叫ぶ>

複数存在した場合は、より字数が少なくなる方を選択している。

例)彼女は大声をあげて泣き叫んだ。

→彼女は--大声をあげて泣き叫んだ。

→(×)彼女は号泣した。

→(○)彼女は哭した。

4. システム評価

内容に関する評価として、*ROUGE-1*を用いてシステムを評価した。*ROUGE-1* では、正解とする要約中の unigram の数を、システムの出力と正解との両方に含まれるunigram数で割った数値をC_1と定義し、算出している。

$$ROUGE\text{-}1 = \exp(\log C_1)$$

評価用正解データは、NTCIR[9]にて配布されているテストコレクションを用いた。表3に示した数値がその評価結果である。

構築したシステムでは、要約率を50から30へ変化させても評価値の減少が少なく、一部の入力では要約率30の出力の方が非常に高くなった。これは最初に作成したプロトタイプが、自由作成時の要約との一致度が高かったためと思われる。

また、評価値がともに高いデータでは、プロトタイプの作成時の評価も「0.7474」(要約率50)と高く、

表3 *ROUGE-1*による評価結果

入力データ	要約率 30	要約率 50
940101002	0.35	0.5769
940101013	0.5676	0.6145
940219151	0.7762	0.5929
940220030	0.45	0.4497
940220031	0.5	0.6933
940221046	0.5641	0.6515
940221047	0.4725	0.6503
940301017	0.2979	0.5765
940401035	0.75	0.7172
平均値	0.5253	0.6136

表4 Quality Questionsによるチェック数

入力データ	要約率 30	要約率 50
940101002	5	3
940101013	8.5	9
940219151	7.5	8
940220030	4	8
940220031	7	8.5
940221046	6.5	9
940221047	7.5	10
940301017	4	8
940401035	6	7.5
平均値	6.222	7.889

最小となったデータでは、重要文抽出で正解文を抽出したとして文短縮すると、評価が「0.5744」(要約率30)と三割近く上昇した。重要文抽出システムの精度向上と、プロトタイプの作成をより人間に近づけることが必要である。

次に可読性に関する評価として、NTCIR-4のTSC3におけるQuality Questionsの質問項目を用いた。16個あるチェック項目に当てはまる数(複数個あり)を主観的に数えた結果が表4である。

全体的に要約率50の方がおかしな文が多いという回答結果になった。これはプロトタイプの作成に、読みやすさを考慮すべき部分がまだ多いと言える。今後のシステム構築では、可読性の向上を念頭に入れることが望ましい。

5. 問題点と今後の展望

人手作成により近い要約を目指して、文短縮システムを実装した。現段階での評価は高くないものの、今後の土台となるシステムを作成することができた。

評価が高くない一因として、正解との一致度でスコアを算出するため、単語を言い替えると評価が下がる傾向がある点が挙げられる。また語彙的言い替えのシステムには、一部適さない語に言い替える問題点があった。パターン辞書の整備や範囲の特定、候補選別のスコア値設定などの検討が望ましい。

このほか、不要表現の判別精度向上や表現の追加などの改良を押し進め、より人手作成に近い要約を生成するシステムの構築を目指す。

謝辞

本研究のため、大辞林の見出し語と語義文の利用を許諾してくださった株式会社三省堂と、テストコレクションを提供してくださったNTCIRに多大な感謝を申し上げます。

参考文献

[1]H.Jing and K.McKeown, “Cut and paste based text summarization”, In Proc. of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pp.178-185, 2000

[2]吉田悦子, 横山晶一, “主題・焦点を用いた文脈解析の一手法”, 電子情報通信学会技術研究報告, NLC97-29(1997)

[3]池原悟, 宮崎正弘, 白井論, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦, “日本語語彙大系”, 岩波書店(1997)

[4]奈良先端科学技術大学, 形態素解析器「MeCab (和布蕪)」

[5]中村明, “比喩表現の理論と分類”, 秀英出版(1978)

[6]菅野紘平, 横山晶一, 西原典孝, “丸括弧解析システムの構築”, 言語処理学会第11回年次大会, pp.1217-1220, Mar. 2005

[7]柴田武, 山田進, “類語大辞典”, 講談社(2002)

[8]松村明, “大辞林”三省堂(1999)

[9]NTCIR, “http://research.nii.ac.jp/ntcir/index-ja.html”