7B−9

テキストの特徴を利用した要約作成法の検討*

原 正巳†

NTTデータ通信株式会社‡

1. はじめに

技術論文や新聞記事などテキストの内容を把握する際に、要約はキーワードとともに有効な情報の一つとなる。大量のテキストを効率良く運用するうえで要約の重要性が大きくなってきているが、要約の作成は人手に頼ることが多く困難を伴う。要約を自動的に作成できれば、テキストの管理・運用に有効である。

本検討では、対象を特許明細書に限定し、テキストの書式と特徴を利用して要約を作成する手法の検討を行なったので報告する。

2. 要約自動作成技術

2.1 要約自動作成の必要性

テキストデータベースが巨大化するにつれ、その中から必要な情報を迅速に検索するためのテキスト処理技術が重要となってきている。

その一手段として要約の利用が考えられる。短時間に多くの情報を得るうえで要約は有効である。また、要約に含まれる語はキーワードとして利用できる。

2.2 従来の要約自動作成技術

要約とはテキストの内容を簡潔に表現するものであり、自動作成には内容を把握することが必要となる。このため意味理解技術や文脈理解技術を用いた要約自動作成の研究が行なわれている[1, 2, 3]。しかし、要約を自動的に作成するには問題があり十分ではない。

1. 文の意味や文脈を理解する技術や、それらを利用した文生成技術がいまだ確立していない。
2. 文の重要箇所を決定する手法が確立していない。
3. テキスト全てを解析するためには膨大な時間を要する。

これらの問題に対処する手段として、キーワードや語の出現頻度により要約を得る方法も試みられている[4, 5]。

3. テキストの特徴による要約自動作成

3.1 本検討の特徴

我々は、テキストの特徴を利用して要約を作成する手法を検討してきた。本検討では、処理対象を特許明細書に限定し、書式と表現の特徴を利用して要約候補をテキストから抽出し、次に不要箇所の削除ルールに従い文字数を絞り込むことによって要約を得る手法の検討を行なった。

特許明細書や技術論文など定型フォーマットのテキストでは、重要箇所の出現位置や表現に特徴が見られる。テキストの特徴を利用することで文全体の意味や文脈の解析を回避して重要箇所を決定することが可能となる。さらに、処理対象を容易に絞り込むことが可能となるため、処理速度の向上につながるというメリットもある。

3.2 本検討での要約

明細書の要約は「発明の意義を明確に示す文の集合」と定義できる。しかし先に2.2で述べたように、意味を考慮して要約を自動作成するためには現在の自然言語処理技術では不十分である。

そこで、本検討では以下のようなアプローチで要約を作成する。

- テキスト中から特定箇所を文単位に抜き出したものを要約とする。

特許明細書において、要約は“目的要約”と“構成要約”と分類されるが、今回は目的要約の自動作成について考察する。

3.3 要約作成概要

要約自動作成の流れを以下に述べる。

1. 要約候補の絞り込み1
   絞り込みにはデリミタを利用する。デリミタには

   【発明の名称】、【特許請求の範囲】
   【発明の詳細な説明】、【図面の簡単な説明】

   の4項目がある。このうち要約候補が多く含まれる【発明の詳細な説明】部を明細書より抜きだし、要約含有ブロックとする。

2. 要約候補の絞り込み2
   要約含有ブロックから特有の表現を含む段落を抽出し、要約候補とする。要約候補抽出ルールの一部を次に示す。

   (a) 「欠点」＋「改善」＋「ため」～「以下」＋「説明」＋「示す」の直前
   (b) 「問題点」＋「改良」、「以上」＋「説明」の直前

   「」内の語の類義語(「欠点」ならば、/欠点/問題点/問題/課題/...など)の組合わせパターンによって、要約候補を抽出する。

3. 要約候補から不要箇所を削除
   不要箇所削除ルールにより要約候補から不要箇所を削除し、全体で250字程度とする。削除ルールの一部を次に示す。

---

*An approach for making a summary by using specific words in text.

†Masami HARA

‡NTT DATA COMMUNICATIONS SYSTEMS CORP.

(a) 括弧で囲まれた部分は削除
(b) / すなわち / つまり /... など、言い換えを示す語以降を削除

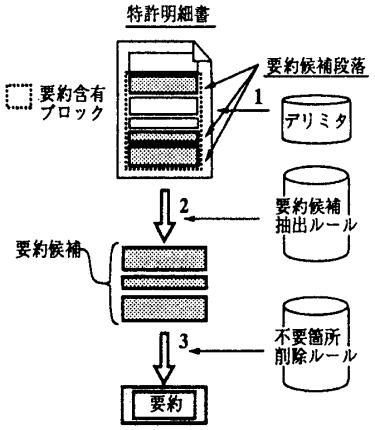

図 1: 要約自動作成フロー

3.4 自動要約作成ルール数

要約候補抽出 / 不要箇所削除ルール数を表 1 に示す。

表 1: ルール数

処理	ルール数
要約候補抽出	42
不要箇所削除	14

4. 精度検証

4.1 検証方法

検証は、JAPIO[1]で作成された抄録 (JAPA 抄) を正解の要約とみなし、JAPA 抄との類似性の評価を以下の 2 つの観点から行なった。

1. JAPA 抄中のキーワードと同一あるいは同義のキーワードが自動作成要約中に何 % 存在するか。
2. JAPA 抄と等しい内容の文が自動作成要約中何 % 程度存在するか。

今回は計算; 計数分野に限定、特許明細書 15 件に関して評価した。

4.2 要約例

JAPA 抄

カナ漢字変換テーブル中に文章を作成する分野における使用頻度を同時に記憶させ、文章内に同音異字語が存在した時には使用頻度の一番高い同音語を検出して変換・出力し、カナ漢字変換効率を高める。

本手法による要約

同音異字語を有する漢字又は熟語登録した変換テーブル中に文章を作成する分野における使用頻度を記録させておき、文章を入力する前又は特定の入力後の前後に対象とする分野の情報を入力することにより、文章内に同音異字語が存在した時は都度、指定された分野での使用頻度の高い順に同音異字語を 1 語づつ選出して変換、表示するものである。

[1](財) 日本特許情報機構

4.3 検証

表 2: キーワード含有率と類似度

キーワード含有率	類似度			
	A	B	C	計
80% 以上	5	2	1	8
60% ～ 80%	2			2
40% ～ 60%		1	3	4
40% 未満			1	1
計	7	3	5	15

A:75% 以上,B:50% ～ 75%,C:%50 未満

本手法で類似度が A または B でかつ、キーワードを 60% 以上含む要約は 15 件中 9 件 (60%) 得られており (図太線部)、本手法の有効性が確認できた。

人間の場合、文章の広範囲から重要な内容やキーワードを抽出し、まとめ上げることで要約を作成している。一方テキストからの抜き出しでは、重要箇所の全てを抽出することは不可能であり、漏れは避けられない。内容的類似性の低い要約の多くはこの問題が原因であると考えられる。この手法で失敗した明細書に対しては、並行して意味解析を導入した要約文の生成も同時に行なう必要がある。

本方式の文字数の絞り込みでは、作成された要約が冗長である場合が多い。表現の特徴からの絞り込みでは重要箇所の確定は限界がある。これ以上文字数を絞り込むには、

1. キーワードによる絞り込み
2. 係受け関係や意味関係での絞り込み
3. シソーラスによる同一概念の統一

などが必要であり、今後検討すべき課題である。

今回は対象テキストの分野を限定して検証したが、書式や表現は分野に強く依存していると考えられ、分野別に表現や書式の特徴を抽出し、分野ごとにルールベース化して、さらに品質の向上を目指す。

5. まとめ

特許明細書に対象を絞り、テキストの書式と表現を利用して要約を自動作成する手法の有効性について述べた。今後は特許明細書への分野別対応に加えて、他の定型フォーマットへの対応や、本方式では救済できない文書への意味情報の利用について検討していく。

参考文献

[1] 稲垣:“事象解析による要約情報の抽出” 電子通信学会 NLC91-9

[2] 邑本, 阿部: “物語文章の要約化処理について” 情処学会自然言語研究会 (1990:78-2)

[3] 田村, 田村:“文章の表現形式に基づいた要約文章の生成について” 情処学会自然言語研究会 (1992:92-1)

[4] 鈴木, 栃内:“キーワード密度方式自動抄録法の改良” 情処学会論文誌 (1988:3)

[5] 間瀬, 大西、杉江:“説明文の抄録作成について” 電子通信学会 NLC89-40