JAIST Repository
https://dspace.jaist.ac.jp/

Title	論文間の参照情報を考慮した学術論文要約システムの開発
Author(s)	難波，英嗣
Citation	
Issue Date	1998-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1151
Rights	
Description	Supervisor:奥村 学，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

論文間の参照情報を考慮した学術論文要約システムの開発

難波 英嗣

北陸先端科学技術大学院大学 情報科学研究科

1998 年 2 月 13 日

キーワード： 複数テキストの要約, 参照関係, cue word.

近年、インターネットの整備とともに、オンラインで数多くの電子化されたテキストを入手できるようになった。電子化された多くの情報の中から求める情報を探し出す際、様々な検索手段が利用できる。例えば WWW 上の場合、Altavista、goo といった検索エンジンを用いてキーワード検索をすることが可能である。また Yahoo!や NTT DIRECTORY、CSJ インデックスのようにあらかじめ分類された Web page を、ディレクトリをたどることで目的の Web page を検索する方法もある。このように、目的の情報にアプローチするための手段が増え、益々便利になる一方でいくつかの問題点も浮き彫りになってきている。問題点は大きく以下の 2 つが考えられる。

(1) 検索結果が何万件にも及ぶような場合、その中から目的の情報を捜し出すのが困難である。

(2) WWW のように情報が複数の情報源から発信される場合、重複した内容のテキストが検索される可能性がある [船坂 96]。

(1) のひとつの対処方法として、検索するためのキーワードを拡張することで、目的の情報に絞り込んでいく方法がある。別の対処方法として要約技術を用いる方法がある。[佐藤 96] はネットニュースのダイジェスティングにより情報を効率的に提示可能にしている。同じネットニュースを扱った研究に [McKeown96] がある。ネットニュースの場合は、ひとつのニュースグループ内では、記事が重複することはほとんどないと考えられるが、情報源が複数存在する場合、(2) に示した重複した内容のテキストが存在することを考慮して要約を生成しなければならない。このような場合、複数テキストからひとつの要約を生成する技術が必要とされる。

Copyright © 1998 by Nanba Hidetsugu

本研究では、学術論文を要約対象としている。ある特定の研究分野に関する複数の学術論文の差異に注目し、論文間の参照情報を考慮して複数の関連する論文との違いを明確にする要約を自動的に作成することを試みる。論文間の関係を解析する際、論文の参照情報に着目する。ある論文が他の論文を参照する場合、参照論文について記述してある箇所 (参照箇所) が存在する。その箇所を読むことで、著者がどのような目的で参照しているのか明らかになる。このようにして参照箇所から得られる情報を参照情報と呼ぶ。参照情報を収集し整理することで、ある分野の複数の論文間の関係が明らかになり、またそれらの参照情報が要約生成に利用できると考えられる。

本研究で開発する要約生成システムの枠組を説明する。ある特定の分野の複数の論文を要約対象とする。論文データは Web 上の論文データベース e-Print archive (http://xxx.lanl.gov/cmp-lg) の論文データ (TEX ソース) 約 450 本を使用する。まず論文データベース中から入力論文 (target papers) を参照している論文を検索する。次に検索された論文から入力論文について記述している箇所 (参照箇所) を自動抽出する。参照箇所から、著者がどのような目的で他の論文を参照しているのか (参照タイプ) を自動的に判別する。本研究では、参照タイプを 3 種類 (B、C、O) に定義し、この 3 種類に分類することを試みる。次に複数論文の融合という処理を行う。target papers が直接参照関係にない場合、target papers を共引用している論文に着目し、この論文の参照関係を用いることで target papers を融合、要約生成を行う。ここで、先程の参照箇所抽出、参照タイプの決定という処理結果が利用できる。target papers を共引用する論文がどの参照タイプで target papers を参照しているか、という情報を用いて target papers の融合が可能になる。

参照箇所の抽出実験において、論文の引用箇所の前後の文間の結束性を考慮して 11 種類のルールを作成した。このルールを用いることで、人手で作成した評価用データにおいて約 80%の精度で参照箇所を抽出できるようになった。また、cue word (一種の手がかり語) の並びを考慮して、12 種類のルールを作成した。このルールにより評価用のすべての参照箇所データ (200 参照箇所) のうち約 6 割はタイプ決定ができた。タイプが決定できたもののうち、約 80%は人が割り振った正解データと一致した。

今後の課題として、参照タイプを用いて、複数の target papers とそれらを共引用する論文の要約生成を試みる。さらに、大規模な参照関係の論文、ある分野の論文をまとめて一種のサーベイを自動生成するための手法について検討していく予定である。