審査の結果の要旨

氏名　　岡崎　直観

本論文は「RESEARCH ON INFORMATION AGGREGATION AND INTEGRATION FOR MULTI-DOCUMENT SUMMARIZATION（複数文書自動要約における情報の集約と統合に関する研究)」と題し，英文で記されており，7 章から成る．

第 1 章「Introduction（序論)」では，WWW（Web）等の発展により大量の電子化情報の流通と共有が進んできていることにより，その効率的な利用を可能にするための複数文書要約技術が重要になるという本研究の背景を述べている．

第2章「Automatic Text Summarization（文書自動要約)」では，関係する文書自動要約技術について纏めている．最初に基本となる考え方と留意点を述べ，指示的（indicative）要約と報知的（informative）要約の考え方があることを示している．複数文書要約に関しては，文書間の関係タイプの考慮や情報集約に際して冗長性の排除などの配慮が必要になることを述べ，幾つかの具体的システムを紹介している．

第3章「Text Summarization Challenge (TSC)（文書自動要約のワークショップ)」では，まず文書要約技術性能を客観的に比較し，その発展を促進するために世界で催されている評価型ワークショップについて記している．本研究の複数文書要約システムも，日本の国立情報学研究所（NII）が主催する評価型ワークショップであるNTCIR（NII-NACSIS Test Collection for IR Systems）の複数文書要約タスクに参加して評価を受けており，開発して参加したシステムの全体構成を示している．要約の元となるテキスト文書も，この評価型コーパスで採用された1998年と1999年の毎日新聞及び読売新聞の記事データを用いており，1 話題についての記事数は5〜19（平均として11.7）である．この評価コンテストで要約性能を評価するのに用いる主な項目は，システムが出力した文の正解率 (precision)，システムが要約として含むべき文をどのくらいカバーしたか（coverage)，要約の質に関するアンケート調査（quality question）であることを記している．

第4章「Sentence Extraction (文抽出)」では，要約の主要コンポーネントである重要文抽出について，関連手法について記した後，本研究で考案しシステムに用いた手法を記している．本研究では，重要文抽出問題を複数文書に含まれる重要な情報断片を，できるだけ多く要約文に含める文の組み合わせを見出す最適化問題と捉え，効率的な手法を開発している．すなわち，係り受け関係にある自立語単語ペアを基礎情報要素とし，その重要度を計算し，その結果に基づき各文の重要度計算を行っている．そして重要度の高い文の順に抽出を行い，抽出された文に含まれる上記の基礎情報要素の重要度を0にして各文の重要度を再計算し，重要度の高い文の抽出を要約文として許容される長さまで反復する．この係り受け関係にある自立語単語ペアを基礎とする重要文抽出法は，よく用いられる自立語の単語 (bag-of-words 法と呼ばれる)を基礎とする方法や，文中に共起する自立語単語ペアを基礎とする方法に比べて，良好な抽出性能が達成できることを示している．本手法によるシステムは，前述のNTCIR 評価型ワークショップにおいて多くの労力を投入している企業からのシステムも含む参加10システムの中で，第3位の好成績を挙げたという結果を報告している．本システムは要約に含まれる冗長情報が少ない点でも優れていることを示している．

第5章「Structuring Extracts (要約文の並び替え)」では，複数文書から抽出した重要文の順序付けに関して新手法を提示している．従来研究では，抽出した文が記された日付の古い順に並べるという chronological ordering (時間順序) がよく用いられていた. 本研究はまず，それぞれの文が元の記事の中でどのような情報を前提として書かれていたのかを解析し，時間順序を改善するアルゴリズムを提案している．さらに，時間順序，前提情報，後置情報，トピックの類似性の４要素を教師あり機械学習アプローチで統合し，２つの文の順序関係の向きと強さに基づいて，階層化クラスタリングと同様の方法で全体の文の並び順を構築する手法を提案している．また，文の並び順の評価方法として，スピアマンの順位相関係数，ケンドールの順位相関係数に加え，連続性の指標を提案し，連続性の指標が文の並び順の評価を行ううえで，優れた指標であることを示している．これらの評価指標を用いて提案手法を評価し，時間順の並びよりも提案手法が優れていることを定量的に示している．

第6章「Abbreviation (略語)」では，略語認識の新たなアプローチを提案している．略語が要約システムにもたらす問題点として，略語が新聞記事でよく用いられる表現であること，短縮形と完全形の関係を認識したうえで，表現の統一を図る必要があることを挙げている．そして，まず英語の略語の短縮形と完全形のペアを獲得する手法として，従来研究である文字列マッチングに依存する手法に対し，括弧表現の外側と内側の表現の共起強度に基づき，略語の短縮形と完全形を抽出する新手法を提案している．医学系文献データベースであるMEDLINEを評価コーパスとして，精度－再現率による評価を行い，先行研究よりも提案手法の方がはるかに良い性能（99%の精度，85-95%の再現率）を示すことを示している．

提案手法を日本語の新聞記事に適用する場合の問題点として，括弧表現の内側と外側の表現が強く共起しても，それらが略語の短縮形と完全形の関係を持たないことが多いことを述べた上で，括弧表現の内側と外側の表現に言い換えの関係が存在するかどうかを調べる手法を新たに提案し，F尺度で60%程度の性能を達成したことを報告している．

第７章「Conclusion（結論）」では，本論文の成果を纏めている．

以上を要するに，本論文は複数文書自動要約に関して，係り受け関係にある自立語単語ペアを基礎情報要素として重要文を抽出する技術，抽出した重要文の要約文中での順序付けを，各文が元の記事中でどのような情報を前提として書かれたか等を考慮することにより決定する新手法を考案し，これらに基づくシステムを作成し，その性能を評価型コンテストなどを通じて客観的に評価し，優位性があることを実証している．また，略語認識に関して，括弧表現の外側と内側の表現の共起度強度に基づき，略語の短縮形と完全形を抽出する新手法を考案し，その性能を実験的に実証している．これらの研究成果により本論文は電子情報学上貢献するところが少なくない.

よって本論文は博士（情報理工学）の学位論文として合格と認められる.