5 NTCIR Workshop 2の新しいタスクの紹介
−テキスト自動要約タスク−

奥村　学　東京工業大学精密工学研究所／北陸先端科学技術大学院大学情報科学研究科
oku@pi.titech.ac.jp
福島 孝博　追手門学院大学文学部　fukusima@res.otemon.ac.jp

■背　景■

本稿では，NTCIR Workshop 2で新しくタスクとして取り上げられたテキスト自動要約タスクについて述べる．

テキスト自動要約は，1950年代から研究されている研究分野であるが，1990年代後半から急速に研究が活発になり，今日に至っている．しかし，システムの出力である要約をどのように評価するかに関しては明確な基準がなく，従来評価が難しいとされてきた．しかし，研究が活発化するに伴い，評価方法を議論し，基準を明確にしようという動きも活発になり，アメリカでは，1998年5月DARPA Tipsterプロジェクト（Phase III）の一貫で，SUMMACという要約の評価を行う会議が開催されるに至った[3), 4)], ☆1．日本でも，これらの動きに刺激され，日本語テキストの要約の評価を目指す動きが本格化し，今回NTCIR Workshop 2のタスクとしてテキスト自動要約を行うこととなった．

以下では，要約の定義，種類などを概説した後，要約の評価方法を紹介する．そして，今回のテキスト自動要約タスクの概要を説明する．なお，テキスト自動要約全般に関しては，いくつか解説があるので[9), 11)]，詳細はそちらを参照していただきたい．

■要約の種類■

要約は，大意を保持したまま，テキストの長さを短くする処理，あるいはその結果のテキストということができる．近年研究が活発化するとともに，要約を細分類して整理する傾向が強い．本章では，そのいくつかを紹介する．

要約を研究するに当たって考慮するべき要因として，以下の3つが提示されている[7)]が，

1. 入力の性質−テキストの長さ，ジャンル，分野，単一／複数テキストのどちらであるか，など
2. 要約の目的−どういう人が（ユーザはどういう人か），どういう風に（要約の利用目的は何か），など
3. 出力の仕方

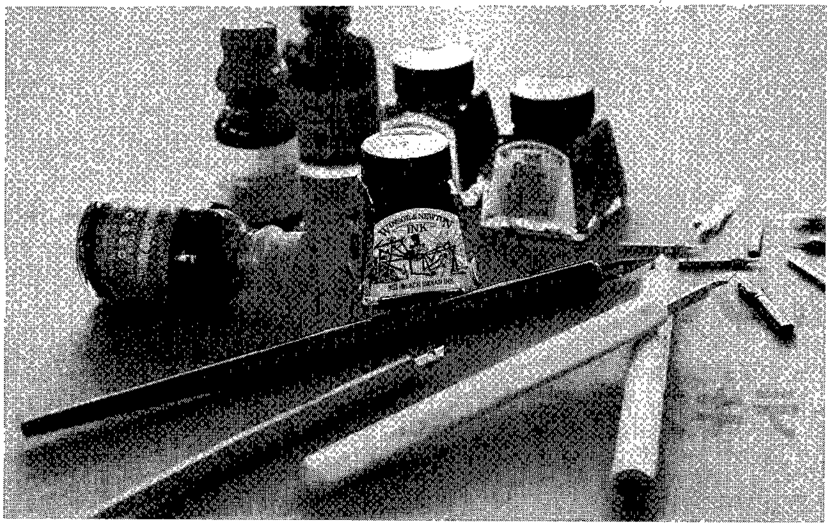
これらの要因に伴ってまず，要約はいくつかの観点で分類することが可能である．要約対象のテキストが1テキストなのか複数テキストなのかにより，単一テキスト要約−複数テキスト要約，また，特定のユーザに特化した要約なのか，特定のユーザを想定しない要約なのかにより，"user-focused"な要約☆2−"generic"な要約という区分がされる．

利用目的に応じて，要約を次の2つのタイプに分けることも多い．

指示的（indicative）：原文の適合性を判断するなど，原文を参照する前の段階で用いる
報知的（informative）：原文の代わりとして用いる

また，これまでのテキスト自動要約手法の多くは，テキスト中の文（あるいは，形式段落）を1つの単位とし，それらに何らかの情報を基に重要度を付与し，その重要度で順序付け，重要な文（形式段落）を選択し，それらを寄せ集めることで，要約を作成する．すなわち，要約は重要文

☆1 TIDESプロジェクトと名前は変わり，趣旨も若干変更はされるものの，今後もアメリカでは要約の評価を行っていく動きに変化はないものと思われる．
☆2 情報検索システムと併用する場合のように，ユーザの情報が検索要求（"query"）として明示される場合は，"query-biased"な要約などとも呼ばれる．



抽出により行われてきた．そのため，元テキストから抜き出すことで作成した要約をextract（抜粋），そうでない場合をabstractと区別することが多くなっている．また，summary，summarizationはextractとabstractを包含する概念として使われることが多い．

■要約の評価方法■

本章では，単一テキスト要約の場合に，従来どのような評価方法が模索されてきたかを簡単に述べる．

単一テキスト要約の伝統的な評価方法の1つとして，人間の作成した要約を正解とし，それと比較するものがある．従来の要約手法はextract（抜粋）を作成するものがほとんどであるため，人間にも同様に，テキストを読んで，重要と思われる文（個所）を選び出してもらい，それを集めて正解の要約とする．そして，システムの作成した要約がどのくらい一致しているかを計ることで評価を行う．一致度を計る尺度には再現率（recall）と精度（precision）がよく用いられる．

$$\text{再現率} = \frac{\text{システムが選んだ文のうちで正解の文の数}}{\text{人間が選んだ正解の文の総数}}$$

$$\text{精　度} = \frac{\text{システムが選んだ文のうちで正解の文の数}}{\text{システムが選んだ文の総数}}$$

しかし，人間にとっても要約を作成するという作業は必ずしも容易ではなく，人間が作成した要約が必ずしも高い割合で一致するとはいえない．また，この評価方法の前提とする「テキストには，ただ1つ正しい要約が存在する」という仮定が不自然であるという批判が以前からあった．

これに対して，要約を利用して人間がタスクを行う場合の，タスクの達成率が間接的に要約の評価となるという考え方に基づき，評価を行う評価方法がタスクに基づく要約の評価と呼ばれる評価方法である．

DARPA Tipsterプロジェクト（Phase III）の評価においても，上の仮定の不自然さから，タスクに基づく評価方法が採用されている．Tipsterプロジェクトでは，テキストの分類，情報検索における検索テキストの適合性の判断それぞれに要約を利用し，被験者のタスクに要する時間，タスクをどの程度うまく行えたか（再現率，精度）により要約を評価する．

このような，要約の内容に関する評価とは別に，要約の「文章としての読みやすさ」を評価する評価方法も考えられる[6]．人間の受容可能性判断に基づいて要約を評価する方法も提案されている[1]．受容可能性は，人間が，原文と照らし合わせて，内容と読みやすさに関して，受容可能／不可能の判定を要約に対して行い求められる指標である．

要約は，本来このように，内容に関する評価と，読みやすさに関する評価の，両方の次元で評価されるべきであるとはいえ，今後もより良い要約の評価方法の模索は続けられるものと考えられる．

この章でこれまで紹介してきた要約の評価方法は，大きくintrinsic（内的）な評価，extrinsic（外的）な評価の2つに分けられるとされる[8]．intrinsicな評価は，要約を直接分析することで要約の質を判断するもので，人間の要約との比較や，読みやすさの評価，受容可能性を用いた評価がこれに該当する．extrinsicな評価は，他のタスクを実行するのに要約がどのように影響を与えるかに基づいて，要約の有用性を判断するもので，タスクに基づく評価方法がこれに該当する．

■今回のテキスト自動要約タスクの目指すもの■

NTCIR Workshopは，テストコレクションを構築し，それを用いた評価を行うワークショップである．評価を行うワークショップであるだけでなく，今後の研究に役立つテストコレクションを整備することも目的としている．

そのサブタスクの1つであるテキスト自動要約タスクでも，したがって，目的は2つある．1つ目は，日本語テキストに対する要約データを蓄積することである．残念なこ

とに，これまで人手でテキストを要約した言語データは，日本語に対してはごくわずかしか作成されておらず，また，研究に利用可能なものが十分存在するという状況とはいえない☆3．今回のタスクでは，新聞記事を対象に，人手で作成した要約データを大規模に蓄積し，研究目的で利用に供したいと考えている．また，これまで作成されてきた要約データは，新聞記事，特に報道記事に限定されてきた傾向が強い．今回の要約データ作成においては，そのような現状を鑑み，報道記事だけでなく，社説などの論説記事も対象に要約データ作成を試みる．

今回の要約データ作成ではさらに，2種類の要約を作成しようと考えている．1つは，いわゆる重要文抽出に基づく要約であり，要約作成者に，重要な文を選択してもらい，また，それらの重要文中の重要個所を選択してもらい，それらの重要個所をつないだものを要約とする．2つ目は，自由作成要約である．要約作成者に，原文にとらわれずに，自由に要約を作成してもらう．要約作成者には，編集者，国語教師，記者など，ある程度要約という作業に熟練している方々を依頼する．どちらの種類の要約も数百テキストの規模で作成したい．

タスクの2つ目の目的は，いうまでもなく，自動要約システムの評価である．前章で述べたように，これまで要約の評価方法にはさまざまな議論があり，また我々も実行委員会やメイリングリストでの議論を重ねてきたが，今回は，intrinsicな評価としては，作成した要約データを用いたシステムの評価を予定している．システムの評価は，この原稿を執筆中も依然検討中の部分が多いが，作成した要約の種類に対応して次の2つを候補としている．重要文抽出に基づいて作成された要約を評価に利用する場合，伝統的な評価方法同様，人間の要約との一致度により評価を行う．人間の自由作成要約を評価に利用する場合，人間の要約との間の文字列上の一致度で評価を行うのは困難であることなどから，人間に主観的評価を行ってもらい，システムの要約がどの程度人間の要約に近いかで評価する．評価基準としては，原文の重要な内容をどの程度カバーしているか，読みやすさ，およびその組合せである受容可能性が候補となる．

一方，extrinsicな評価方法としては，システムの出力である要約を情報検索における適合性判定タスクに利用することで評価する方法を採用する予定である．これは，1998年に開催されたTipsterのSUMMACでも評価方法の1つとして採用されたものである[4), 2)]．

情報検索タスクに基づく要約の評価は基本的に次のように行われる．まず，人間の被験者に，検索要求とその検索結果としてテキストの要約を提示する．被験者は各要約を読むことによって，そのテキストが検索要求に合っているかどうか（適合性）の判断を行う．この適合性の判断をどの程度うまく行えたか，判断にかかった時間などを基に，提示された要約がよいかどうかを間接的に評価する．このことから，評価している要約の種類は，"query-biased"で，指示的（indicative）な要約ということができる．

情報検索用テストコレクション（検索要求，テキスト集合，適合性判定の正解）としてどのようなものを用いるのかなど未定の部分もあるが，原則的にSUMMACと同じ方法で評価を行いたいと考えている（http://www.itl.nist.gov/div894/894.02/related_projects/tipster_summac/index.html参照）．評価基準としては，タスクに要した時間および，タスクをどの程度うまく行えたかを示す指標として，再現率，精度，F値を用いる．

$$\text{再現率}=\frac{\text{被験者が正しく適合と判断したテキスト数}}{\text{実際に適合するテキストの総数}}$$

$$\text{精　度}=\frac{\text{被験者が正しく適合と判断したテキスト数}}{\text{被験者が適合と判断したテキストの総数}}$$

$$\text{F　値}=\frac{2\times\text{再現率}\times\text{精度}}{\text{再現率}+\text{精度}}$$

■今後の展開■

CFP（Call for Participation）は，5月末に出る予定であり，その直後にdryrun（予備試験）を行いたいと考えている．本試験は，秋に開催予定であり，その結果や，結果に対する分析，考察などは，2001年2月開催予定のNTCIR Workshop 2で報告される予定である．

この原稿を執筆している5月の時点では，そろそろCFPが出されようかという段階であり，最終的にどの程度の参加者がいるかも定かではない．しかし，テキスト自動要約システムの評価は，SUMMACが一度行われただけであり，日本語テキストに対しては，今回が初めての試みである．今回作成予定の要約データが今後のテキスト自動要約研究において有用であり，また，今回の評価を機に，さらにテキスト自動要約研究が活発になり，今後もサブタスクとしてテキスト自動要約が継続的に行われることを期待したい．

今回作成予定の要約データは，複数のジャンルのテキストを対象にしており，また，重要文抽出に基づく要約だけでなく，自由作成による要約も蓄積する予定である．多様なジャンルの要約データを蓄積することで，多様なジャンルに対するテキスト自動要約研究が進むことが期待できる．また，自由作成要約の言語データは，近年研究者の注目を集めている，extract（抜粋）ではなくabstractを作成

☆3 現在利用可能な要約データのリストは，http://galaga.jaist.ac.jp:8000/pub/research/summarization/を参照していただきたい．

する研究（たとえば，人間の要約過程をモデル化し，より読みやすい要約を作成する研究）[10]を進めるうえで重要なデータとなることはいうまでもない．

今回NTCIR Workshop 2でテキスト自動要約をサブタスクとして行う企画を進めるにあたり，我々は1999年テキスト自動要約タスクの実行委員会を組織し，32人の方々に実行委員をお願いし，何回かのミーティングを行い，また，メイリングリスト上で議論を重ねてきた．議論に加わっていただいた実行委員の方々にここで感謝したい．また，北陸先端大の望月源氏にはメイリングリストやWebページの管理をはじめ，このタスクの運営にかかわる多くのことでお手伝いいただいている．ここに感謝の意を表したい．

テキスト自動要約タスクのWebページはhttp://galaga.jaist.ac.jp:8000/tsc/にあり，また，メイリングリストのアドレスは，tsc@recall.jaist.ac.jpである．メイリングリストへの参加希望者は，tsc-request@recall.jaist.ac.jpまでご連絡いただきたい．

近年要約研究者の関心が，複数テキスト要約，言語横断要約（translingual summarization）など，タスクとしてより困難ではあるが，興味深いものに移りつつあるのは事実であるが[5), 10)]，我々は，今回，単一言語の単一テキストの要約をあえて評価の対象とした．これは，複数テキスト要約などではまだまだ評価自体に多くの議論を要することや，タスクとして行うことが，時間的，予算的，また人的資源の制約から困難と考えたからである．

次回以後サブタスクとしてテキスト自動要約を行う際，新たな試みを加えていきたいと考えている．

参考文献

1）Brandow, R., Mitze, K. and Rau, L. F.: Automatic Condensation of Electronic Publications by Sentence Selection, Information Processing and Management, Vol.31, No.5, pp.675-685（1995）.

2）Firmin, T. and Chrzanowski, M. J.: An Evaluation of Automatic Text Summarization Systems, In Mani, I. and Maybury, M., editors, Advances in Automatic Text Summarization, pp.325-336, MIT Press（1999）.

3）Hand, T. F.: A Proposal for Task-based Evaluation of Text Summarization Systems, In Proc. of the ACL Workshop on Intelligent Scalable Text Summarization, pp.31-38（1997）.

4）Mani, I. et al.: The Tipster Summac Text Summarization Evaluation, Technical Report MTR 98W0000138, MITRE（1998）, http://www.itl.nist.gov/div894/894.02/related_projects/tipster_summac/index.html

5）Mani, I. and Maybury, M., editors: Advances in Automatic Text Summarization, MIT Press（1999）.

6）Minel, J., Nugier, S. and Piat, G.: How to Appreciate the Quality of Automatic Text Summarization? Examples of Fan and Mluce Protocols and Their Results on Seraphin, In Proc. of the ACL Workshop on Intelligent Scalable Text Summarization, pp.25-30（1997）.

7）Jones, K. S.: Automatic Summarizing: Factors and Directions, In Mani, I. and Maybury, M., editors, Advances in Automatic Text Summarization, pp.1-12, MIT Press（1999）, http://xxx.lanl.gov/ps/cmp-lg/9805011

8）Jones, K. S. and Galliers, J. R.: Evaluating Natural Language Processing Systems: An Analysis and Review, Springer（1996）.

9）奥村　学, 難波英嗣: テキスト自動要約に関する研究動向, 自然言語処理, Vol.6, No.6, pp.1-26（1999）.

10）奥村　学, 難波英嗣: テキスト自動要約に関する最近の話題, Technical Memorandum, 北陸先端科学技術大学院大学 情報科学研究科, 2000（準備中）. http://www.jaist.ac.jp/~oku/okumura-j.html

11）佐藤理史, 奥村　学: 電脳文章要約術－計算機はいかにしてテキストを要約するか－, 情報処理, Vol.40, No.2, pp.157-161（Feb. 1999）.

（平成12年7月3日受付）