情 報 学 基 礎 63－5
（2001．7．24）

TSC2(Text Summarization Challenge 2) の目指すもの

奥村 学*, 福島孝博**
(*) 東京工業大学 精密工学研究所, (**) 追手門学院大学 文学部

[概要]
本稿では，NTCIR Workshop 3 のサブタスクの一つであるテキスト自動要約タスクについて，課題，評価方法等詳細を，アメリカの動向も踏まえ，述べる．

The Goal of Text Summarization Challenge 2

Manabu Okumura* and Takahiro Fukushima**
(*) Precision and Intelligence Laboratory, Tokyo Institute of Technology
(**) Department of English, Otemon Gakuin University

Abstract
In this paper, we explain the task description and evaluation methods of Text Summarization Challenge 2, the automatic text summarization sub-task in the NTCIR Workshop 3.

1 はじめに

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，アメリカでは現在，DARPA TIDES プロジェクトの一貫で，DUC という要約の評価を行なう会議が毎年開催されるようになっている．日本でも，日本語テキストの要約の評価を目指す動きが昨年本格化し，NTCIR Workshop 2 のタスクとしてテキスト自動要約が行なわれた [2].

本稿では，NTCIR Workshop 3 のサブタスクの一つとして引続き行なわれる予定であるテキスト自動要約タスク，Text Summarization Challenge 2(TSC2) について，課題，評価方法等を説明する．

以下ではまず，昨年度 NTCIR Workshop 2 の元で行なわれた TSC1 について簡単に紹介し，アメリカで開催されている DUC(Document Understanding Conference) について触れた後，今回のテキスト自動要約タスク TSC2 の概要を説明する．

2 TSC1

TSC1 の概要，結果等の詳細については，[7, 2, 8] で述べられているので，ここでは簡単に課題，評価方法を紹介する．なお，TSC1 の課題はどれも，単一の新聞記事を要約対象としている．

1. 課題

   課題 A: 要約対象となるテキストと，作成する要約率 (要約の長さ) が与えられ，参加者は，それを元に要約を作成し提出する．1つのテキストに対する要約率は複数与えられる．

   - 課題 A-1. 重要文抽出型要約
     テキスト中の文に要約率分だけ印をつけたものを提出する．
   - 課題 A-2. 人間の作成した要約と比較可能な要約
     要約を plain text で作成し提出する．

   課題 B: 提示した検索要求と，その検索結果としてのテキストを元に，要約を作成し提出する．要約の長さは自由とする．

2. 各課題における要約の評価方法

   - intrinsic な評価
     課題 A では，別途作成する人間の要約データを用いた評価を行なう．課題 A-1 の提出結果は，重要文抽出に基づいて作成された要約が想定されるため，人間が選択した重要文との間の一致度を元に評価を行なう．評価尺度としては，再現率，精度，F 値の 3 つを用いる．

     課題 A-2 の提出結果は，単に重要文抽出しただけではない要約が想定される．そのため，厳密な評価は行なわないが，人間の自由作成要約および，人間が重要個所を抽出した要約との間の比較を行ない，その結果を参加者にフィードバックする．

     (a) 内容に基づく評価 (content based evaluation)
     人間の作成した要約およびシステムの作成した要約をともに Juman で形態素解析し，内

容語のみを抽出する．そして，人間の作成した正解要約の単語頻度ベクトルとシステムの要約の単語頻度ベクトルの間の距離を計算し，どの程度内容が単語ベースで類似しているかという値を求める [1].

(b) 主観評価
人間の評価者に，原文および，人間の要約 (自由作成要約，重要個所抽出要約)，システムの要約，ベースラインシステムの要約の 4 つを提示し，原文の重要な内容をどの程度要約がカバーしているか，要約の読み易さの2つの評価基準で，要約を順序付けてもらう．

- extrinsic な評価
課題 B では，情報検索タスクに基づく評価を行なう．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．原則的に SUMMAC[4] と同じ方法で評価を行なう．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なえたかを示す指標として，再現率，精度，F 値を用いる．

3 DUC

一方，アメリカで始まった TIDES プロジェクト[1]でも，要約の評価を行なう会議 DUC[2] が NIST の主催で開催されるようになり，DUC 2001 は，現在 formal run の最中である (ワークショップは SIGIR に併設して，9月に開催予定)．

ここでは，簡単に DUC 2001 の課題，評価方法を紹介する．

- 課題

1. 単一記事要約
100 語の generic な要約を作成．
2. 単一話題に関する複数記事要約
長さの異なる (400, 200, 100, 50 語)4 つの generic な要約を作成．
3. Exploratory Summarization
課題 1, 2 で与えられるデータを利用しても良いし，独自のデータを利用しても良い．自由課題．要約における何らかの問題に対する新しい解決策，評価法の提案を行なう．

- 評価方法
課題 1, 2 の評価は，単一，複数記事要約ともに，intrinsic な評価のみで，主に，人間による評価に基づく．人間による評価は，要約作成者が行ない，要約のテキストとしての善し悪し (grammaticality, cohesion, organization) に関する主観評価と，要約の内容の善し悪しに関して，人間の要約とシステムの要約の比較を行なう．

人間の要約とシステムの要約の比較には，ISI/USC の Chin-Yew Lin が作成した SEE[3] をツールとして利用．評価者が，ユニット単位で，人間の要約に含まれる内容がシステムの要約に含まれるかどうかを判断していく．

これを行なうため，訓練用，評価用それぞれ 30 セットのデータを用意している．

[1] http://www.darpa.mil/ito/research/tides/index.html
[2] http://www-nlpir.nist.gov/projects/duc/
[3] http://www.isi.edu/ cyl/SEE/

各セットは，テキスト集合 (平均 10 テキスト)，人間が作成した各テキストに対する要約 (約 100 語)，人間が作成した複数テキスト要約からなる．複数テキスト要約は 4 種類．まず 400 語の要約．それを元にして作成された 200 語，100 語，50 語の要約．セットの定義 (タイプ) はさまざま (ある出来事とそれに関係する原因・結果, ある人に関するテキスト集合, 台風，フェリーの沈没のような，あるタイプの複数の出来事, あることに関する複数の意見, ...)．

4 TSC2

昨年度行なった TSC1 の経験および，アメリカにおける DUC の動向を踏まえ，TSC2 では，以下の課題，評価方法を現在検討している．

4.1 課題と評価方法

- 課題 A: single
  TSC1 の課題のうち，課題 A-2 を継続．なお，要約は plain text で提出するため，重要文抽出のみを行なうシステムも，この課題に参加できる．

  評価方法としては，TSC1 の課題 A-2 で行なった主観評価を継続するとともに，新たな評価の指標として,「システムの要約に対する修正の度合」を別途導入したい．内容，可読性に関して，システムの要約を評価者 (3 人) に添削してもらう．添削は，挿入，削除，置換の 3 つの操作のみで行なう．その度合を指標とする．ただし，新しい指標も，TSC1 の課題 A-2 の評価と同じく，厳密な評価ではなく，あくまでも参考程度の指標に過ぎない．

- 課題 B: multi(複数記事を対象にした要約)
  DUC 2001 の課題 2 と同様，いくつかの種類のテキストセットを対象とし，それらのテキスト集合の要約を作成する．人間の要約作成者 (3 人) にも同様にテキストセットから要約を作成してもらう．ただし，DUC 2001 の課題 2 とは異なり，この際，セットを用意するのに用いた情報 (クエリ等) も合わせて，システム，要約作成者に与えることとする[4]．セット中のテキスト数は，少数 (2-3)，中くらい (5-7)，比較的多い (10-) 場合を対象としたい．

  評価方法としては，課題 A と同様，内容，可読性に関する主観評価とともに，添削に基づく指標を用いる．

課題 B に関しては，本稿執筆時まだ確定的でない部分も多い．少なくとも，

- 要約の長さはどの程度のものとするか
- どういう種類のセットを用意するか
- テキストの対象領域はどういうものとするか

に関しては今後検討する必要がある．

また，課題 B については，ベースラインシステムとしてどのようなものを用意するかも今後検討が必要である．

4.2 日程

TSC2 の現在予定されている日程は以下の通りである．

[4]DUC 2001 では，この情報を要約作成者に与えなかったため，作成された要約に非常に大きなゆれが生じたとのことであった．

2001.7 CFP(Call for Participation)
9 dryrun 課題公表，結果提出
10,11 評価，評価公表
12,2002.1 分析
2002.2 round table
4 formal run 課題公表，結果提出
5,6 評価，評価公表
7,8 分析
9 round table
10 Workshop

今回は，TSC1 では NTCIR Workshop の日程の制約で実現できなかった 'round table' を実現したいと考えている．TSC2 の評価は，どちらの課題も，システムの比較を実際にできるような絶対的な数値を提示する厳密な評価ではなく，あくまでもシステムの善し悪しを把握する参考となる指標程度の意味合いしか果たしていない．各参加者が，システム，評価の結果等を持ち寄り，各システムの良い点，問題点等を議論し合うことで，システムの「評価」を行なう 'round table evaluation' の形式を取ることが，テキスト自動要約のように，依然評価方法が未確定の分野では良いのではないかと考える．

5 今回のテキスト自動要約タスクの目指すもの

前節の TSC2 の課題を見ると，DUC 2001 の課題とほとんど同じという印象を持たれるかもしれない．確かに，課題の内容はほとんど同じであるが，しかし，我々は評価方法の相違が重要であると考えている．

上で述べたように，テキスト自動要約の分野は，評価方法がまだ確定しておらず，模索を続けている段階と言える．残念ながら，DUC 2001 でも，状況は同じである．このような状況から，我々は，TSC の目的の 1 つとして，要約の評価方法の提案，検討を考えている．TSC1 では，要約の順序づけによる方法 (ranking 法) と内容に基づく評価 (content-based evaluation) を用い，この 2 つの評価方法の間の相関を調査した [8]．TSC2 では，ranking 法と，添削に基づく指標 (editing 法) を用い，同様にこの 2 つの評価方法の相関等を分析する予定である．

TSC のもう一つの目的は，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．実際，TSC1 では，単一テキストを対象にした要約データとして，報道記事，社説などの論説記事の両方を対象に，重要文抽出，重要個所抽出，自由作成の，3 つの異なる要約を人手で作成し，蓄積することができた．

TSC2 では，この要約データ作成を継続的に行なった方が良いと考え，継続して行なう．ただし，TSC1 では 1 つのテキストに対して 1 人の人間による要約しか作成できなかったのに対し，TSC2 では要約を 3 人に作成してもらうことにする．

また，これまで人手で複数テキストを対象に要約した言語データは，日本語に対してはほとんど作成されておらず，研究に利用可能なものは全く存在しないという状況であり，今回初めて作成を試みる．近年複数テキスト要約に関する研究が活発になっているが [5, 6]，複数テキストを対象とした要約として，どのようなものが望ましい (理想的) かということは，人手で作成した要約データがないこともあり，研究者の間でも不明確であり，研究の進展を阻害していると考えられる．今回複数記事を対象にした要約データを作成することで，望ましい要約というものが明確化されることを期待している．

6 おわりに

NTCIR Workshop 3 のサブタスクの一つである TSC2 について，課題，評価方法等を

述べた.

この原稿を執筆している6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が,多くの研究者の方々の参加を期待したい.

また,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なわれることを期待したい.

テキスト自動要約タスクのwebページは`http://lr-www.pi.titech.ac.jp/tsc/`にあり,また,メイリングリストのアドレスは,`tsc-ml@lr.pi.titech.ac.jp`である.メイリングリストへの参加希望者は,webページに記載されている情報を御参照頂きたい.

最後に,TSCの今後について言及したい.

1. Q&Aとの接点
   近年研究が活発化しているQ&A(Question Answering)(たとえば,[3])は,その質問の種類に応じ,さまざまな研究分野との重なりを持っており,テキスト自動要約とも接点を持っていると考えられる.たとえば,「××について教えて」,「××とは何?」といった,説明を求める質問の場合,その解答は,テキスト中のパッセージあるいは,複数のテキストから集めたパッセージ群(さらに,それを整形したもの)のような形式と考えられる.また,'How','Why'を尋ねる質問の場合も,解答は同様の形式を取ると思われる.このようなQ&Aは,question-biased summaryの作成と近いと言える.このことから,質問に対する解答が要約中に含まれているかどうかを調べることで,要約の評価を行なうという評価方法も今後検討していきたい.

2. 実問題への適用
   近年テキスト自動要約の典型的な応用として,携帯端末へのテキストの表示が注目を集めている.このように,現実の応用において要約手法が用いられるようになるなら,その応用において,extrinsicな評価ができると考えられる.今後このような試みをTSCに加えていきたい.

3. extrinsicな評価方法
   TSC2では,残念ながら,要約をintrinsicに評価することしか試みていない.上で上げたように,現実の応用を想定しないで,要約をextrinsicに評価する方法は,intrinsicな評価方法同様,決め手と言えるものがない.今後模索していきたい.

4. 要約対象テキストの検討
   webページなど,新聞記事以外のジャンルのテキストを対象にすることも今後検討していきたい.また,言語横断要約(translingual summarization),ジャンル横断要約など,まだ全く試みられていないタスクも,次回以後新たな試みとして加えていきたい.

謝辞

本稿で述べたTSCの課題,評価方法に関しては,TSCの実行委員の皆様および,これまでのタスク検討会に参加して下さった皆様との議論が大変参考になっています.議論に加わって下さった皆様に感謝致します.

参考文献

[1] R.L. Donaway, K.W. Drummey, and L.A. Mather. A comparision of rankings produced by summarization eval-

uation measures. In *Proc. of the ANLP/NAACL2000 Workshop on Automatic Summarization*, pp. 69–78, 2000.

[2] T. Fukushima and M. Okumura. Text summarization challenge: text summarization evaluation in japan. In *Proc. of the NAACL2001 Workshop on Automatic Summarization*, pp. 51–59, 2001.

[3] S. Harabagiu and D. Moldovan. Open-domain textual question answering. Tutorial Presented at NAACL-2001, 2001.

[4] I. Mani, et al. The tipster summac text summarization evaluation. Technical Report MTR 98W0000138, MITRE, 1998.

[5] I. Mani and M. Maybury, editors. *Advances in automatic text summarization.* MIT Press, 1999.

[6] 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 is-tm-2000-001, 北陸先端科学技術大学院大学情報科学研究科, 2000.
http://galaga.jaist.ac.jp:8000/pub/papers/oku/summarization2000.ps.gz.

[7] 奥村 学, 福島孝博. Ntcir workshop 2 の新しいタスクの紹介 – テキ スト自動要約タスク –. 情報処理, Vol. 41, No. 8, pp. 917–920, 2000.

[8] 難波英嗣, 奥村 学. 第 2 回 ntcir ワークショップ 自動要約タスク (tsc) の結果および評価法の分析. 情報処理学会自然言語処理研究会報告, 2001. 144-20.