# 2

# Text Summarization Challenge
## ー自動要約の評価型ワークショップー

福島 孝博
追手門学院大学文学部　fukusima@res.otemon.ac.jp

奥村　学
東京工業大学精密工学研究所　oku@pi.titech.ac.jp

難波 英嗣
広島市立大学　nanba@its.hiroshima-cu.ac.jp

## はじめに

情報社会と言われて久しく，情報が日常的に溢れるようになった現代において，多量の情報を「要約」する技術は，ますます重要となっている．

テキスト（文章の集り）をコンピュータを使い自動的に要約する技術である「自動要約」の研究の歴史は1950年代に遡ることができ，コンピュータの誕生当初からこの分野に関する研究が行われていた．

従来の研究は，与えられたテキストにおいて，まず，重要な文を見つけ出し，それらの重要な文を集めることにより元のテキストの要約とする「重要文抽出」による自動要約が主であった．しかし，最近はこれだけではなく，テキストの構造，読み手の視点を考慮しての要約や，より自然な要約文を作成する研究，単一のテキストではなく複数のテキストを対象とした自動要約の研究が進められている[12), 13)]．

自動要約システムの性能を高めるには，自動的に作成された要約をどのように評価するかが重要となる．自動要約システムの要約した結果をどのように評価するかは，最近の自動要約研究の重要な課題の１つとなっている．要約の評価方法には大きく分けて２つある．１つ目は，要約そのものをその内容や読みやすさなどの点から評価する内的な（intrinsic）評価と，要約を利用して他の作業を行いその作業の達成率を計ることにより要約の評価をしようとする外的な（extrinsic）な評価がある．

対象となるテキストを人間が要約を行い，システムの結果と比較して評価を行う内的な評価が従来から行われてきた．しかし，同一テキストを要約するにしても要約を行う人間により，要約結果が違うことがある．また，同じ人が行う場合でも，読む視点によって要約結果が違ってくる場合もあり，人間の要約との単純な比較をしての評価には問題があるといえる．外的評価は，近年になり後述する評価型ワークショップにおいて採用されている．評価にどのような作業を選ぶのかという問題があるが，要約を何に役立てるのかという目的がはっきりとしている評価だといえる．

### ◎評価型ワークショップ

言語処理研究のいくつかの分野では，最近の研究の特徴として，特にアメリカにおいてそうであるが，「評価型のワークショップ」を開催して研究を進めることが行われている．この評価型ワークショップは，以下のような流れで進められる．

(1) 共通の課題を設定する．課題は1つだけに限られず，いくつかあるのが普通である．
(2) 一般からの参加を募る．課題に対する結果を提出する限り海外の研修者であっても参加は自由にすることができる．
(3) 参加者は，どの課題に参加するかを決定して，その課題に対する結果を提出する．参加者には，主催者からシステムの開発に必要なデータが提供されるのが普通である．
(4) 主催者側は，参加者から提出された結果を集計して，評価を行い，評価結果を公表する．
(5) 公表はワークショップ形式で行われ，主催者と参加者，関係者が一堂に集まり，課題の内容や評価方法，今後の進むべき道などについて議論，検討を行う．参加者は，自分たちのシステムについての説明をここで行う．
(6) ワークショップの結果は，論文集やWWWにて公開される．

評価型ワークショップは1回限りではなく，何年かにわたり複数回開催され，対象の分野での要素技術を明確にし，その技術の精度を高めていくような方向で運営される．たとえば，新聞記事などのテキストからあらかじめ決められた重要だとされる情報を取り出してくる技術である「情報抽出（Information Extraction）」の分野では，Message Understanding Conferences（MUC）という評価型ワークショップが米国国防省の支援のもとに，1980年代後半より1990年代を通して約12年間，合計7回のワークショップが開催された．このMUCの開催により英語で書かれた新聞記事を対象とする情報抽出に関する要素技術が確立され，研究課題が明確となり，情報抽出の研究が進められた[17]．

このような手順で行われる評価型ワークショップであるが，自動要約研究でのその役割について考察すると，次のような3つの役割を果たしているといえる．

- 自動要約システム自体の評価の実施
- 評価方法の確立
- 要約データの作成と蓄積

今まで，自動要約システムは開発されてきていたが，その評価については，システム開発者が内部で行うものが普通であり，異なる要約システム間の評価，比較は困難であった．また，評価のための手法も，研究分野内で確立されたものがなく，種々の評価方法が試されてきていた．評価のための要約データも，研究者が共有して使えるものがなく，研究者独自に作成することが通常であった．

しかし，評価型ワークショップを行うことにより，共通の課題について共通の評価尺度を適用でき，異なる要約システム評価が可能となった．評価方法については，いくつかの方法を複数の要約システムに対して試すことができるようになった．同時に，その評価のための要約データを作成して，公開するため，研究者が共有して利用可能な要約データの蓄積がなされて，研究を進める環境が整うことになった．

### ◎自動要約の評価型ワークショップ

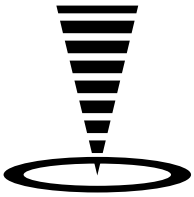
自動要約の評価型ワークショップとしては，アメリカ国防省の支援するTipsterプロジェクトの一部としてSUMMAC（Text Summarization Evaluation Conference）が開催された[10), 16)]．最近は，同国防省が支援するTIDES（Translingual Information Detection, Extraction, and Summarization）プロジェクトの一環として，NIST（National Institute of Standards and Technology）の主催するDocument Understanding Conferences（DUC）が行われている[6), 7)]．DUCは，第1回目が2001年，第2回目が2002年に開催されている．DUCでどのように研究を進めていくのか長期的な計画も発表されており，Roadmapとして公開されている[8)]．これらの一連のワークショップでは，自動要約システムの評価を行い，また，評価方法について種々の取り組みがなされ報告されている．

日本においては，国立情報学研究所主催のNTCIR（NII Test Collection for Information Retrieval and Text Processing）ワークショップの枠組みにおいて，日本語テキストを対象とした自動要約の評価型のワークショップが2回行われている[5), 9)]．この自動要約の評価型ワークショップはText Summarization Challenge（TSC）と呼ばれ，TSC1（第1回目），TSC2（第2回目）として開催された[4)]．

次章にてTSCについて，続いてTSC2についてより詳しく解説する．また，最後に今後の展望を述べる．

## TSC1

TSC1は，1999年から2001年の初めにかけてNTCIR Workshop 2のタスクの1つとして開催された[1), 2), 14)]．以下に課題，評価方法，TSC1にて作成した要約データおよびTSC1から得られた知見について述べる．



### ◎課題

課題は大きく分けて2つあった．1つ目の課題（課題A）の中に2つの課題（課題A-1，A-2）が設定された．

課題**A-1**：与えられたテキストから重要文を取り出してくる．重要文を取り出す割合は，10%, 30%, 50%である．たとえば，対象となるテキストが10文で構成されているのであれば，1, 3, 5文を重要文として選び出す．

課題**A-2**：与えられたテキストから自由に要約を作成する．作成された結果は，人手で作成された要約と比較をして評価される．文字数で計算しての要約の割合（要約率）は，20%, 40%である．1,000文字のテキストであれば，200文字までの要約と400文字までの要約を作成することになる．

課題**B**：検索要求（情報検索における質問）とその検索結果としてのテキストを元にして要約を作成する．要約の長さは自由である．

なお，TSC1では，単一テキストが要約の対象であり，複数テキスト要約は課題としていなかった．

### ◎評価方法

課題Aは，内的な評価を行い，課題Bでは外的な評価を行った．

課題Aでは，まず，人手による要約を作成しておく．課題A-1では，人手により重要だとして選ばれた文とシステムの選んだ文がどの程度一致しているかで評価を行った．課題A-2では，人手による要約を2種類使った．人間が重要だと思う個所を選んでの重要個所要約と自由に要約をした自由要約である．この2種類の要約を使い，以下の2つの評価を行った．

(1) 内容に基づく評価（**content-based evaluation**）

人手の作成の要約とシステムの要約結果を，単語（正確には形態素）に分割し，名詞，動詞，形容詞などの内容があるとされる語のみを対象として，人間の要約とシステムの要約がどの程度近いかを計算し評価を行った．

(2) 主観評価

評価者に，元のテキストに加えて，人手作成の重要個所要約，自由要約，システムの要約結果およびベースラインシステム（基本的な要約手法を用いてのシステム）の要約という4種類の要約を見せて，原文の重要な内容がどの程度保持されているか，要約文の読みやすさの2つの観点から順位を付けてもらった．

課題Bでは，外的な評価，つまり，情報検索の作業に基づく評価を行った．評価者に検索要求とシステムが作成した要約（検索要求結果のテキストの要約）を見せる．評価者は，その要約を読むことによって，そのテキストが検索要求に適合しているかどうかを判断する．米国で開催されたSUMMACにて同様の評価が英語のテキストを対象として行われた．

評価は，DryrunとFormal runの2度行われた．前者は，試験的な評価であり，課題に慣れるために行われたものである．Dryrunを行うことにより，主催者にとっても，参加者にとっても，課題を遂行するにあたって何が必要とされるかが分かるというメリットがある．後者は，本番の評価である．本番の評価に参加したシステム数は，課題A-1に10，課題A-2に9，課題Bに9であった．また，評価には，社説，経済面からの30記事がDryrunにおいて，社説，社会面からの30記事がFormal runで用いられた．

### ◎要約データ

TSC1において使用したテキストは，毎日新聞記事データベースからの記事であった．作成された要約データは，新聞記事180記事それぞれに対して，10%, 30%, 50%の割合での重要文を選択したもの，および20%, 40%の要約率の重要個所要約と自由作成要約である．これらの要約データは現在，国立情報学研究所のWebページにて公開されており，研究目的での利用が可能となっている[5)]．

### ◎TSC1の知見

TSC1は，日本語を題材とした自動要約の評価型ワークショップとしては，初めてのものであった．評価結果を簡単にまとめると以下のようである．

- 課題A-1：10%でのシステムの評価値に多少差が見られたが，30%, 50%では，システム間の差は少なかった．大体のシステムがベースラインシステムより良い評価値を示した．
- 課題A-2：内容評価では，ベースラインを含めてシステム間の差があまりない結果となった．主観評価で

は，多少の差があるものの，全体としては，人手の自由要約，重要個所要約，システムまたはベースラインの要約の順で評価が良く，人手の要約とシステムの要約の差がはっきりしていた．

- 課題Bにおいても，システム間の差があまり見られなかった．SUMMACでも確認されたことであるが，日本語のテキストを対象とした場合でもシステムの要約が情報検索の作業を行う上で有効であることが確認された．

ワークショップの運営面では，要約データの作成および評価の遂行に，人手による作業を多く必要とするため，予想以上に時間がかかることが分かった．

また，TSC1 開催後であるが，TSC1 で作成された要約データを使用して重要分抽出型の要約の評価に関する研究が行われ，より良い評価のための検討がなされている[11]．

## TSC2

TSC1 にて引き続いて開催されたTSC2 における課題の内容，評価方法，TSC2 にて作成した要約データについて述べる[3), 15)]．

TSC2 は 2001 年から 2002 年にかけて開催された．TSC1 のときと同様に，NTCIR Workshop 3 のタスクの 1 つとして行われた．

### ◎課題

課題は大きく分けて 2 つである．単一記事の要約と複数記事の要約である．後者は，TSC2 において初めて行われるものである．

**課題A**：単一記事を対象とした自由要約である．TSC1 の課題A-2 と同様の内容である．要約率も同様に，20%, 40%の 2 種類である．

**課題B**：複数記事を対象とした要約を行う．複数記事は，あるトピックに関して集められた複数の新聞記事の集合である．トピックは，川崎公害訴訟，2000 年問題，花粉症，ノーベル賞，全日空のストライキなど多岐に渡っている．システムおよび人間の要約作成者には，記事集合を作成する際に使った情報（検索に用いた語や語句）を与える．要約は，短いものと長いものの 2 種類を設定した．

### ◎評価方法

評価方法は，課題A, 課題Bとも共通で，2 種類の内的評価を行った．

(1) 順位付け評価

これは，TSC1 での主観評価と同様のものである．ただし，TSC1 では，人手による要約作成時に，要約の対象となる 1 つの新聞記事に対して 1 つの要約を作成しただけであったが，TSC2 では，1 つの記事に対して 3 人の要約作成者に要約を作成してもらい，そのうちの 1 つを評価値の上限を示す指標（アッパーバウンド）として評価に取り入れている．

(2) 添削評価

TSC2 にて初めて導入された評価である．この評価では，システムの要約に対してどの程度の修正がなされるかを測ることにより，評価を行うものである．実際には，要約の内容および読みやすさに関して，評価者に添削をしてもらうが，添削は削除，挿入，置換の 3 種類のみに制限されている．削除は不要だと判断される文字を消すことであり，挿入は，反対に，必要な文字を新たに入れることであり，置換は，ある文字や文字の並び（文字列）を言い換えることに相当する．

評価は，TSC1 と同様にDryrunとFormal runの 2 度行われた．本番の評価に参加したシステム数は，課題Aに 8, 課題Bに 9 であった．また，評価には，課題A用に社会面からの 30 記事，課題B用に 16 のトピックがDryrunにおいて使われた．Formal runでは，課題A用に社会面からの 30 記事と課題B用に 30 のトピックが用いられた．

TSC2 では，Dryrunと Formal runの後に，round-table discussionとして，課題の内容，評価方法について議論，検討する会を主催者と参加者が集まってワークショップとは別に行った．これは，研究分野において評価の手法が確立されていない中，主催者と参加者が互いに議論，検討をすることにより，評価結果に見られる数値だけではない緩やかな評価を行い，より良い評価方法を探ろうとするのが目的である．

Formal run後のround-table discussionでは，順位付け評価に関しての問題点と思われる点が指摘され，別の評価方法が提案された．また，今後どのような評価方法，要約データが望まれるかについて検討がなされた．

### ◎要約データ

TSC1 では，180 の記事を対象にして，要約作成者 1 名の要約データを作成したが，TSC2 では，それに追加をして，同一記事に対して第 2，第 3 の要約の作成を進めた．要約の種類は，TSC1 と同様であり，評価では用いなかった重要文選択の要約も含まれている．また，課題B用として，50 の異なるトピックについての記事集合を対象とする長短 2 種類の要約を，要約者 3 名を使って作成している．

## 今後の展望

TSC3 の開催が予定されているが，その内容については，今後TSC3 へ向けての検討会を開催して決定していく計画となっている．

また，TSC3 で実施されるかどうかは別にして，TSC として以下のような課題を検討していきたい．

**(1) 質問応答の研究分野との連携**

ユーザの質問に対して新聞記事などのデータベースを元にしてシステムが答えを出す質問応答（Question Answering）システムの研究分野との関係が注目される．たとえば，「AO 入試とは何？」という質問に対しての回答を得るには，AO 入試に関する新聞記事などのテキストから関係の深い文や段落を抜き出してまとめる必要がある．また，英語でいう How や Why の質問においても同様に，回答が存在するであろう部分を探し出して集め，まとめる必要がある．このように見ると，この種の質問応答では，質問を元としてそれに対する要約を作成する作業と見なすことができる．質問応答自身は，NTCIR Workshop 3 においてもタスクの 1 つとして評価型ワークショップが実行されており，連携の可能性を探っていきたい．

**(2) 外的な（extrinsic）評価**

これは，TSC1 で情報検索作業を用いて行ったが，TSC2 では，内的な評価のみとなった．一方，関連する動向として，自動要約研究の実社会の問題へ応用しようとする試みが始まっている．たとえば，携帯端末での表示のための要約，テレビ番組に付与される字幕を制作するための要約などが挙げられる．このような自動要約研究の実問題への応用も考慮して，外的な（extrinsic）評価としてどのような課題が良いのか検討していきたい．

**(3) 要約の対象となるテキストの多様化**

今までに開催した 2 回の TSC では要約の対象となるテキストはすべて新聞記事であったが，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とした要約も考えられる．また，よりチャレンジングであるが興味深い課題となる，英語で書かれたテキストを日本語で要約するような複数の言語にまたがる要約（translingual summarization）も検討していきたい．

最後に，これまでの TSC の経験および蓄積されたデータが活かされ，さらに自動要約研究の活動が活発になること，そして評価型ワークショップが引き続き開催されることを願いたい．

参考文献
1) Fukusima, T. and Okumura, M.: Text Summarization Challenge - Text Summarization Evaluation at NTCIR Workshop2, In Proceedings of NTCIR Workshop2, pp.45-50 (2001).
2) Fukusima, T. and Okumura, M.: Text Summarization Challenge - Text Summarization Evaluation in Japan, North American Association for Computational Linguistics (NAACL2001), Workshop on Automatic Summarization, pp.51-59 (2001).
3) Fukusima, T., Okumura, M. and Nanba, H.: Text Summarization Challenge 2 - Text Summarization Evaluation at NTCIR Workshop3, In Proceedings of NTCIR Workshop3 (2002).
4) http://oku-gw.pi.titech.ac.jp/tsc/index-en.html
5) http://research.nii.ac.jp/~ntcadm/index-ja.html
6) http://www.darpa.nil/ito/research/tides/index.html
7) http://www-nlpir.nist.gov/projects/duc/
8) http://www-nlpir.nist.gov/projects/duc/roadmapping.html
9) 神門典子: NTCIR とその背景－情報アクセス技術の評価ワークショップとテストコレクション，人工知能学会誌，Vol.17, No.3 (May 2002).
10) Mani, I., et al.: The TIPSTER SUMMAC Text Summarization Evaluation, Technical Report, MTR 98W0000138 The MITRE Corp. (1998).
11) 難波英嗣，奥村　学: 要約の内的（intrinsic）な評価方法に関するいくつかの考察－第 2 回NTCIR ワークショップ自動要約タスク（TSC）を基に－，自然言語処理，Vol.9, No.3, pp.129-146 (July 2002).
12) 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).
13) 奥村　学，難波英嗣: テキスト自動要約に関する最近の話題，自然言語処理，Vol.9, No.4, pp.97-116 (2002).
14) 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 (Aug. 2000).
15) 奥村　学，福島孝博: TSC2（Text Summarization Challenge 2 の目指すもの，情報処理学会情報学基礎研究会，63-1 (July 2001).
16) Proceedings of The Tipster Text Program Phase III, Morgan Kaufmann (1999).
17) 若尾孝博: 英語テキストからの情報抽出，情報処理学会自然言語処理研究会，96-NL-114-12 (1996).

（平成 14 年 10 月 23 日受付）