特　集　「NTCIR: 情報アクセスに関わるテキスト処理技術の評価ワークショップ」

テキスト処理研究の動向
—情報抽出・自動要約・質問応答における評価ワークショップの重要性—

Trends in Text Processing Research
— Role of Evaluation Workshop in Information Extranction, Automatic Text Summarization and Question Answering —

福島　孝博　追手門学院大学文学部
Takahiro Fukushima　Faculty of Letters, Otemon Gakuin University.
fukusima@res.otemon.ac.jp

奥村　学　東京工業大学精密工学研究所
Manabu Okumura　Precision and Intelligence Laboratory, Tokyo Institute of Technology.
oku@pi.titech.ac.jp, http://oku-gw.pi.titech.ac.jp/~oku/

加藤　恒昭　東京大学大学院総合文化研究科
Tsuneaki Kato　Graduate School of Arts and Sciences, The University of Tokyo.
kato@boz.c.u-tokyo.ac.jp

Keywords: evaluation workshop, text processing, text summaraization, information extraction, question answering.

1. はじめに

我々が大量の文字情報を扱うとき，重要な情報だけを取り出し，要約し，または，質問に答える形で取り出せれば，大変有用である．これらを研究テーマとする3分野，つまり，新聞記事などのテキストから重要な情報を取り出す「情報抽出」，テキストの本文を自動的に要約する「自動要約」，テキストを利用してさまざまな質問に自動的に答える「質問応答」は，50年代から研究がされているものもあるが，ここ十数年は，特に米国において評価ワークショップを中心にして進められている．以下に情報抽出，自動要約，質問応答の順に，各分野の技術内容の紹介と，評価ワークショップの経過を簡単に説明し，それが担ってきた役割を考察する．

2. 情報抽出

情報抽出（Information Extraction）は，米国におけるMUC（Message Understanding Conference）と呼ばれる一連の評価ワークショップを中心にして，研究が盛んになってきた [Cowie 96, MUC, 若尾 96].

MUCの情報抽出は，新聞記事などの対象となるテキストからあらかじめ定められた事柄や出来事に関する情報を取り出し（抽出し），データベースの形にする．MUCでの抽出すべき情報は，次の4種類であった．

- 固有名詞を中心としたテキスト中の重要要素（Entities）
- Entityがもつ属性（Attributes）
- Entity間の基本的な関係（Facts）
- Entityが関係する出来事（Events）

例えば，新聞での人事異動に関する記事を対象とする情報抽出では，Entityとしては，会社名，人名，地名の固有名詞と日付情報が，属性は，Entityの名前，種類（人名か社名か地名か），人名であれば，その人の役職名などとされた．Factsとしては，人名と会社名の関係，会社名と地名の関係が認定され，Eventとしては，誰が役職を退き誰が新しくその役職に就いたかなどが抽出の対象とされた．次にその例（英語）を示す．

"OGU named Mr. J. Tanaka, 45, currently vice president of OGU, president of its flagship company, OGU Japan in January 2002. Mr. Tanaka succeeds Mr. Kato." という文から，表1のような情報が抽出されることになる．

MUCの情報抽出で使用された基本的な技術はパターンマッチングと言われ，関係情報を含むと思われる文に対して，あらかじめパターンを用意しておき，情報を抽出するものである．Entityであれば，'Mr.' などのタイトルを示す語の直後には，人名が来るというパターン，'president of' の直後の大文字語は組織名とするパター

表1 抽出されるデータ

Entities

Person	J. Tanaka	Kato
Organization	OGU	OGU Japan
Date	January 2002	

Attributes

Name	J. Tanaka
Descriptor	currently vice president of OGU
Category	PERSON
Name	Kato
Category	PERSON
Name	OGU
Category	ORGANIZATION
Name	OGU Japan
Category	ORGANIZATION

Facts

PERSON	Employee of	OROGANIZATION
J. Tanaka	Employee_of	OGU
	Employee_of	OGU Japan
Kato	Employee_of	OGU Japan

Events

SUCCESSION_EVENT

ORGANIZATION	OGU Japan
POST	president
WHO_IS_IN	J. Tanaka
WHO_IS_OUT	Kato

ンを用意し，テキストの文に適用してEntityを抽出する．これらを例の第一文に適用すると，'J. Tanaka'が人名となり，'OGU'が組織名となる．

Eventでは，Entityの情報を生かして，「〈組織名〉named〈人名〉...〈役職名〉」などの動詞を中心としたパターンを用意し，同様にテキストの文に適用して人事異動の内容を抽出する．第一文にこのパターンを当てはめると，J. Tanaka氏がOGU Japanの社長となるSuccession Eventが抽出される．

MUCは米国国防省関係機関の支援により1987年から開始された評価ワークショップであり，1999年までの12年間に合計7回のワークショップが開催された．MUCでの抽出の対象となるテキストは，初期の頃が，米国海軍の艦船に関する短い記述であったが，第3回からは新聞記事が中心となり，第3回，第4回では，中南米におけるテロリストの活動に関する記事，第5回以降は，企業間の提携など，ビジネスに関係する記事，また，半導体製造など科学技術に関する記事であった．

情報抽出に関する研究は，米国内でもMUCが始まる以前から多少あったが，評価ワークショップのMUCの開催により，盛んに研究されるようになった．特に，テキスト中の固有名詞を中心とした重要要素（Entity）を抽出する技術は，分野を特定すれば，人間の精度に近い程度（90％以上）で抽出できるところまで進み，実際にソフトウェアとして商品化されるまでになった．しかし，Eventレベルでの情報抽出については，当初よりは，精度が向上したものの頭打ちとなり，ベストのシステムでも50～60％の精度であり，人手に及ばない結果に終始した．その原因として，記事中のEventに関する記述の多様なバリエーション（例えば，名詞句の組合せによってEventが書かれている場合）にパターンマッチングだけでは対応しきれなかったことなどがあげられる．

情報抽出分野における評価ワークショップであるMUCの貢献は，それまでテキスト処理のシステムの評価がもっぱらシステム単独で行われていたのに対して共通の尺度によって複数の違ったシステムを評価したことであり，また同時に，その共通の評価を実現するために，研究者が共有できるテストコレクション（具体的には，新聞記事と情報抽出の正解となるデータ）が作成されたことである．MUCへの参加は自由であり，北米，英国を中心としたヨーロッパ，アジア地域の大学，政府研究機関，企業の研究所などから参加があり，多様なメンバにより研究が進められた．

ただし，情報抽出の精度を示す数字でシステムの順位が公表されるためシステム間の競争が行き過ぎることが見受けられた．このため，問題の本質的な解決に向けての新しいアイディアが試しにくい雰囲気があったことは，マイナス面であったといえる．文の構造の解析をして，その結果を情報抽出に役立てる試みは，MUCの初期の段階では用いられたが，MUC4以降は，パターンマッチングが主流となり，文の意味まで理解をするには，必要であるとされる構文解析を利用するシステムは見られなくなった．

日本においては，Entityの抽出を中心課題の一つとした評価ワークショップ，IREX（Information Retrieval and Extraction Exercise）が，1999年9月に開催された[IREX]．MUCと同様に，日本語の新聞記事を対象とした情報抽出においても，かなりの精度（80％以上），でEntityが抽出可能であることが判明した．IREXのEntity抽出では，日本語の新聞記事を題材とした場合に，英語と同様にパターンマッチングの手法が有効であるのかどうかが試されといえ，効果があることが判明した．

3. 自動要約

テキストからその重要な部分を抜き出す，または，要約する技術に関する研究は，古く1950年代からある[奥村 99]．「要約」は，原文の意味を保ちながら，原文より短く簡潔にするものであるといえるが，技術的に見て，テキスト中の重要な文だけを抜いて集めることにより要約とするもの（重要文抽出）と，テキストの内容を理解し，要約文を新しくつくり出すものに大別される．

重要文抽出は従来から研究されており，以下その技術内容を簡単に紹介する．重要文抽出では，テキスト中の文を単位として，何らかの基準に従って各文の重要度を計算し，その重要度で順位を付け，重要な文を選択して集めることにより要約を作成する．

重要文を特定する情報として利用されるものはいくつ

かあるが [Paice 90]，ここでは，基本となる四つを説明する．

(1) テキスト中のキーワード
(2) テキストのタイトル
(3) テキストあるいは，段落での位置
(4) 手がかり語

(1) では，テキスト中で出現回数が多いものが重要だとしキーワードとするものや，情報検索の技法を使い，対象テキスト固有のキーワードを算出する方法が利用されている．

(2) は，もしテキストにタイトルがついていれば，そのタイトルに使われている語は重要とするものであり，(3) は，重要な情報が書かれている場所が想定できる場合は，その場所の情報（位置情報）を利用するものである．例えは，新聞では，記事の始め，特に第一文に重要な情報が書かれていることが大半であり，新聞記事からの重要文抽出では，このような位置情報が利用される．

(4) では，'in summary'，'in conclusion' などの慣用的な表現を手がかりとして，重要な情報を見いだすものである．

実際の自動要約のシステムでは，これらの情報を組み合わせて，総合的に利用している．

近年は，重要文抽出は勿論のこと，原文を言い直しての要約やよりユーザの必要性に合った要約の生成の研究がされている．まず，言い直しての要約作成のために，辞書などの定義文を利用して，言い換えや抽象化を行う研究がある [Kondo 97]．例えば，

停滞する：ある所にとまり，先に進まなくなること

という辞書の定義文がある場合に，

「会議は彼の発言でとまり，進まなくなった.」

のような文は，上の定義を利用することにより「会議は，停滞した.」と要約することができる．

利用するユーザの要求を反映した形で要約を作成する研究もされている [Ochitani 97]．ここでは，要約を単に与えられるものとしてではなくて，そのユーザの要求を考慮して，より要求に沿う要約の作成を目指す．具体的には，ユーザの要求を示す文が与えられる場合に，そこで使われている単語がテキスト文中にどの程度現れるかを加味して重要さを算出して要約を作成している．

また，単一の記事から要約をつくる単一記事要約だけでなく，複数の記事をまとめることを試みる複数記事要約システムの研究が始まっている．

複数記事を対象とした要約作成では，単一記事の要約とは異なり，複数の記事要約を扱う場合独特の問題に対処する必要がある．手順としては，まず，重要部分を検出し，複数記事中の共通部分を検出し，冗長性をなくす．これだけでは十分ではないので，その後記事間の相違点をはっきりとさせて，文体の統一を図り，最終的な要約を得ることになる．

複数記事要約の研究は，新聞記事や学術論文を対象として行われてきている．学術論文を対象とした研究では，論文間の参照情報（参照している論文についての記述部分）を利用して論文間の共通点や相違点を明らかにしてサーベイ論文を作成する助けとしている [難波 99]．

応用的な研究としては，テレビニュース番組用の字幕を作成することを前提として，ニュース文の不要箇所を削除して字幕用の要約をつくる技術についても研究が進められている [若尾 97]．ここでは，重要箇所を特定するのではなく，不要である，または，削除してもよいと判断できる箇所を特定して削除を行うルールを用意して，それらのルールを適用することにより字幕用要約を作成する．

例えば，「…定員が大幅に増加することになりました.」というニュース文が与えられた場合に，文末を省略して言い切る形にするルールを適用すると「…定員が大幅に増加.」となり，簡潔な字幕に適する表現となる．

自動要約の研究が多様化する一方で，要約システムの評価は，情報抽出における評価ほど定まったものがなく，自動要約システムをどう評価するか自体がこの分野の研究の重要な課題の一つとなってきている．

従来は，人が作成した要約を正解とし，システムが作成した要約と比較をすることにより評価を行ってきた．しかし，要約そのものが人にとっても必ずしも容易な作業ではないため，人手で作成した要約が高い度合いで一致するとは限らず，この評価方法は問題があると指摘されている．

この分野における評価のワークショップは，米国における Tipster プロジェクト（Phase Ⅲ）での SUMMAC が大規模なものとしては初めてであり，1998 年 5 月に開催された [SUMMAC]．ここでの評価方法は，上記のように評価するのではなく，要約とは別の作業（与えられた記事をその内容に適合したカテゴリーに分類するテキスト分類など 3 種類のタスク）を行う際にどれだけ要約が役立つかをみて評価が行われた．

米国では，SUMMAC の後，2000 年より DUC (Document Understanding Conference) が開始され，第 1 回目の大会が昨年あった．現在は，第 2 回目の評価大会に向けての準備が進められている [DUC]．DUC においては，単一記事要約，複数記事要約の評価，また，SUMMAC で行われたような，要約が別のタスクを行う際にどれだけ役立つかを見て評価を行うことも検討されている．

日本でも自動要約に関する研究が近年盛んとなり，評価ワークショップが，2001 年に開催された第 2 回 NTCIR ワークショップのサブタスクの一つ，要約タスク（別名を TSC: Text Summarization Challenge）として実現した [NTCIR 01, TSC]．TSC1 では，人手により新聞記事を要約したものを用意し，それとの比較を，重要な情報を漏れなくカバーしているのかという点と読みやすさの点から行い評価を行った．また，情報検索タス

クに基づく評価も実施した．現在第3回NTCIRワークショップのサブタスクとしてTSC2を行う予定である（[神門02]のTSC2の説明参照）．

自動要約の分野では，評価ワークショップは，自動要約の研究そのものを進める側面と同時に，評価の方法自体を研究する場となる側面をもち，今後評価ワークショップを続けることにより両側面ともに進むことが期待される．

4. 質問応答

自然言語で表現された質問に適切に回答する質問応答システムは，自然言語処理研究の黎明期から，自然言語理解技術のテストベッドとして研究されていた．一方，近年，注目を集めているものは，オープンドメインでの質問応答技術で，膨大なテキスト集合を知識源として，分野を限定しない質問を受け付ける．これは，むしろ，情報検索などのテキスト処理技術の延長に位置付けられるものである．このような研究は，米国NIST（National Institute of Standards and Technology）の主催で1990年代初頭から毎年行われている情報検索技術に関する評価ワークショップTREC（Text Retrieval Conference）において，99年のTREC-8にQA Trackが設けられたことで注目されることになった．

TRECにおける質問応答は，「文書ではなく解答を」あるいは「文書の検索から情報の検索へ」と位置付けられ，事実に基づく簡潔な質問に対して，新聞記事集合から，その解答を含む50 byteを抜き出し，五つの候補を順位付きで回答するというものである．例えば，「ビッグ・マディとして知られているのはアメリカの何という川ですか」という質問に「ミシシッピ」を含んだ文字列で回答することが目的となる．昨年のTREC-10（QA-Track第3回）では，最も優れたシステムは約7割の質問に正解できている[TREC]．

一般的に，質問応答システムは質問解析，文書（パッセージ）検索，回答抽出という三つの要素から構成される．質問解析では，入力された質問を解析して，質問タイプの分類などを行い，回答が含まれている文書を検索するための情報と，検索された文書から回答を抽出するための情報を得る．文書検索は，質問解析によって得られた情報を用いて，解答が含まれていると期待される文書もしくはパッセージの検索を行う．回答抽出は，文書検索によって検索された文書と質問解析部で得られた情報を用いて，文書から，解答を含むと判断される部分を抽出する．

例えば，「PostScriptを開発したのはどこですか」という質問に正しく回答するためには，この質問が「どこ」という疑問詞をもつにもかかわらず，場所ではなく組織をたずねていることを認識する必要があるし，文書中のどの部分が組織名を表現しているかを判断できなければならない．さらに，文書中では「開発」でなく「実用化」という語が使われているかもしれないから，それら用語の類似性を判定できなければならない．このために，情報抽出でも使用されるパターンに基づく理解技術やEntityの抽出技術が用いられるし，情報検索分野でのパッセージ検索や質問拡張の技術が，質問応答という新しい文脈で研究されている．現時点での回答抽出は，テキストを語の並びとみなして，意味を考慮した語の近接性を用いるものが主流であるが，統語情報の利用や依存構造から得られる意味情報を用いた推論による方法も提案されている．

このような質問応答研究それ自体が，TRECという評価ワークショップによって提案されたといっても過言ではないし，その中での課題の具体化もTRECという場を通じて行われてきた．事実に基づく簡潔な質問という設定もそうであるが，最初は250 byteもしくは50 byteという回答サイズであったものが50 byteのみと厳しくなったり，確実に解答が存在するという設定だったものから解答が存在しない可能性が考慮されるようになるなど，回を重ねるにつれ，さまざまな点で課題が変化している．質問セット自体も検索エンジンなどへの現実の質問を収集して作成するようになってきている．ここから，現実世界の複雑な要求と研究として達成すべき中間目標が，評価ワークショップでの議論を通じて，研究者の間で擦り合わされ，共有されていく流れが感じられる．ちなみに，今年のTREC QA Trackは，解答そのものを回答させるとのことである．

もちろん，情報抽出技術がMUCでの議論を通じて整理され，その要素技術に関する共通認識がつくられてきたように，質問応答においても，前述のような要素技術の整理が進みつつあるし，質問と解答のセットであるテストコレクションが研究者に共有されていくという点でも評価ワークショップの役割は大きい．

NISTは，今後5年の質問応答研究の展開について，要約技術との融合などを含めたロードマップを提出している[NIST]が，そのような研究の段階的展開を先導し方向付けるものとして評価ワークショップが機能していくと考えられる．また，日本では，第3回NTCIRワークショップのサブタスクとしてQAC（Question Answering Challenge）が，現在，実施されている[QAC]．そこでは，TRECで課題設定された質問応答研究をにらみつつ，独自の強調点をもった研究を加速することが試みられている[神門02]．

5. おわりに

これまで見てきたように，情報抽出，自動要約，質問応答の分野における評価ワークショップは，要素技術の明確化と発展，共有可能なテストコレクションの作成と蓄積に貢献をしている．

評価ワークショップは，情報抽出において，米国でのMUCが研究の推進役として大きな役割を果たしたといえる．自動要約と質問応答では，米国，日本において現在継続中であり，評価方法の設定やテストコレクションの作成などで成果が出ている．今後どこまで貢献をするかは，主催者および参加者の熱意によるところが大きいであろう．

3分野で使われている技術は，情報抽出でのEntityの抽出技術のように3分野共通で利用可能なものもあるが，それぞれの課題の設定が異なり要求される技術も違うため，独自の評価ワークショップが行われ研究が進められてきている．一方，テストコレクションの共有化や課題を関連付けて設定することなどの分野を超えての協力が重要であることが意識されてきている．

今後日本においては，より長期的な展望に基づいた，異なった分野間の協力や海外での評価ワークショップとの連携を考慮した評価ワークショップのさらなる発展が望まれる．

◇ 参 考 文 献 ◇

[Cowie 96] Cowie, J. and Lehnert, W.: Information Extraction, *Communication of the ACM,* Vol. 39, No. 1, pp. 80-91 (1996)

[Paice 90] Paice, C.: Constructing Literature Abstracts by Computer, *Techniques and Prospects, Information Processing and Managements,* Vol. 26, No. 1, pp. 171-186 (1990)

[DUC] http://www-nlpir.nist.gov/projects/duc

[IREX] http://cs.nyu.edu/cs/projects/proteus/

[神門 02] 神門典子, 安達 淳: 評価ワークショップによるテキスト処理研究—第3回NTCIRワークショップを例として—, 人工知能学会誌, Vol. 17, No. 3, pp. 312-319 (2002)

[Kondo 97] Kondo, K. and Okumura, M.: Summarization with Dictionary-based Paraphrasing, in *Proc. of the Natural Language Processing Pacific Rim Symposium'97* (1997)

[MUC] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html

[難波 99] 難波英嗣, 奥村 学: 論文間の参照情報を考慮したサーベイ論文作成支援システムの開発, 自然言語処理, Vol. 6, No. 5, pp. 43-62 (1999)

[NTCIR 01] *Proc. Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization* (2001)

[NIST] http://www-nlpir.nist.gov/projects/duc/roadmapping.html

[Ochitani 97] Ochitani, R., Nakao, Y. and Nishino, F.: Goal-Directed Approach for Text Summarization, in *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization,* pp. 47-50 (1997)

[奥村 99] 奥村 学, 難波英嗣: テキスト自動要約に関する研究動向(巻頭言に代えて), 自然言語処理, Vol. 6, No. 6, pp. 1-26 (1999)

[QAC] http://www.nlp.cs.ritsumei.ac.jp/qac/

[SUMMAC] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/

[TREC] http://trec.nist.gov/

[TSC] http://oku-gw.pi.titech.ac.jp/tsc/

[若尾 96] 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)

[若尾 97] 若尾孝博, 江原暉将, 白井克彦: テレビニュース番組の字幕に見られる要約の手法, 情報処理学会自然言語処理研究会, 97-NL-122-13 (1997)

2002年3月20日 受理

著 者 紹 介

福島 孝博 (正会員)

1978年大阪外国語大学英語科卒業．1990年ニューヨーク州立大学(バファロー校)コンピュータ・サイエンス研究科修士課程修了．90～93年ニューメキシコ州立大学付属Computing Research Lab研究員，94年英国シェフィールド大学コンピュータ・サイエンス研究科博士課程修了．96年日本電気(株)入社．同年通信放送機構にて研究員．2000年より追手門学院大学文学部英語文化学科．自然言語処理，情報抽出，要約筆記に関する研究に従事．電子情報通信学会，情報処理学会，言語処理学会，ACL各会員．

奥村 学 (正会員)

1962年生まれ．1984年東京工業大学工学部情報工学科卒業．1989年同大学院博士課程修了．同年，東京工業大学工学部情報工学科助手．1992年北陸先端科学技術大学院大学情報科学研究科助教授，2000年東京工業大学精密工学研究所助教授，現在に至る．工学博士．知的情報提示技術，語学学習支援，テキストマイニングに関する研究に従事．情報処理学会，AAAI，言語処理学会，ACL，日本認知科学会，計量国語学会各会員．

加藤 恒昭 (正会員)

1959年生まれ．1981年東京工業大学工学部電気電子工学科卒業．1983年東京工業大学大学院総合理工学研究科電子システム専攻修士課程修了．同年，日本電信電話公社(現NTT)に入社．2000年より東京大学大学院総合文化研究科言語情報科学専攻助教授．自然言語理解，マルチモーダルコミュニケーション，質問応答技術に関する研究に従事．工学博士．電子情報通信学会，情報処理学会，言語処理学会，ACL各会員．