社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－FI－84　(11)
2006－NL－175 (11)
2006／9／12

Basic Elementを用いた質問応答の自動評価

福本 淳一†　加藤 恒昭‡　桝井 文人†‡　森 辰則‡‡　神門 典子†‡‡

立命館大学†　東京大学大学院‡　三重大学†‡　横浜国立大学‡‡　国立情報学研究所†‡‡

fukumoto@media.ritsumei.ac.jp†　kato@boz.c.u-tokyo.ac.jp‡

masui@ai.info.mie-u.ac.jp†‡　mori@forest.eis.ynu.ac.jp‡‡

kando@nii.ac.jp†‡‡

あらまし

質問応答技術においては，factoid 型の名称が回答となる質問についての評価が主に行われてきた．しかしながら，factoid 型を超える質問は回答が文レベルなど長い表現となり，あらかじめ与えられた正解とのマッチングのみでは評価が難しいといった問題があった．本稿では，要約の自動評価に用いられてきた Basic Element を質問応答の評価に適用することでこの問題を解決する試みについて述べる．本評価は NTCIR-6 QAC (QAC4) において行われる評価であり，QAC4 のタスク設定および評価に関するこれまでの経過についても報告する．

キーワード

質問応答, 自動評価, Basic Element

An Automatic Evaluation of Question Answering using Basic Element

Jun'ichi Fukumoto†,　Tsuneaki Kato‡,　Fumito Masui†‡,
Tatsunori Mori‡‡ and　Noriko Kando†‡‡

Ritsumeikan University†　University of Tokyo‡　Mie University†‡
Yokohama National University‡‡　National Instisute of Informatics†‡‡

Abstract

In this paper, we will discuss an evaluation method of question answering which exceeds factoid type questions. Answer expressions of non-factoid question will be longer ones than named entities and have difficulty in evaluation. Automatic evaluation method using Basic Elements proposed by Hovy et.al. has been used for evaluation of summaries in DUC. We are planning to apply this method to question answering in NTCIR-6 QAC (QAC4). We will report our current status of evaluation and task description of QAC4.

Keywords

Question Answering, automatic evaluation, Basic Element

1 はじめに

我々は質問応答技術に関する評価プロジェクトとして NTCIR の一環で Question and Answering Challege (QAC) [1] を行ってきた [1][2][3][4]．これまでの QAC では名称を回答とする factoid 型質問について，5 位までの順位で回答を返すタスク，リストタスク，Information Access Dialogue (IAD) タスクを行ってきた．また，factoid については，NTCIR において多言語の質問応答タスクである CLQA タスク[2] も設定され評価が行われてきている．しかしながら，factoid を超える質問として why,how,definition などの質問については対象外であった．そこで，これまで対象にしてこなかったタイプの質問について次回の QAC において扱うことを提案し，現在 NTCIR-6 QAC (QAC4) として評価プロジェクトを進めている．factoid 型を超える質問については回答が名称でなく，文レベルの長い表現になることも想定できることから，これらの回答の評価を行うためには，これまでにようにあらかじめ準備された回答とのマッチングでは評価が難しいといった問題がある．

TREC の QA トラックにおいては，relationship task として，連続した質問の最後に other type question として，それまでに述べられた質問に関連するものを取り出すタスクが設定され評価が行われている [5]．そこでの評価は，システムの回答として集められたものを意味的なパーツである information nuggets に分割し，vital, ok として，より重要な情報が多く含まれているものが良い回答であるという評価が行われてきている．また，DUC においては要約文を Summary Content Units (SCUs) に分割し，それらにより評価を行う Pyramid method と呼ばれる方法が使われてきた [6]．また，Hovy らは SCU の単位の認定が一定でないという問題点もあり，要約の自動評価のため，Basic Elements という最小の意味的な単位を設定し，DUC においても自動評価の 1 つの手法として利用されてきている [7]．

[1] http://www.nlp.is.ritsumei.ac.jp/qac/
[2] http://clqa.jpn.org/

本稿では，質問応答技術において，factoid を超える質問の評価のため，要約において用いられている BE に基づいた回答の評価手法について考察する．factoid を超える質問については，回答が長いものになるため，これを BE に分割し，正解との BE の一致度により評価を行うものである．

以下では，BE による評価手法について述べた後，今回の質問応答タスクである NTCIR-6 QAC (QAC4) のタスクについて述べる．そして，最後に現在までに行ってきた評価手法について述べた後，いくつかの課題と今後の進め方について述べる．

2 Basic Element を用いた要約の自動評価

自動評価のアプローチはこれまで要約を対象に行われてきている．Hovy らは Basic Element (BE) と呼ばれる最小の意味的な要素を用いた評価法を提案し [7]，要約文書の内容的な評価を行うことを目的に研究が行われ，現在 DUC において 1 つの評価手法として用いられている．BE とは意味的な最小の構成素として定義され，正解の要約文書を BE に分割し，分割された BE の集合を評価対象の文書から分割された BE の集合と比較することで評価することを目指したものである．

2.1 BE Breaker

Hovy らはこれまでのいくつかの BE の分割実験から BE を次のように設定している．

- 主となる構文的要素のヘッドとなるものとして名詞，動詞，形容詞，副詞句などの要素
- それらヘッドとなる要素と依存関係にあるものをとり，ヘッドと依存要素とそれらの間の関係をもつものを BE とする

以上の設定された BE の集合に文書を分割するため，与えられた文書を複数の構文解析器を

用いて構文解析し，得られた構文解析木を元にBEに分割を行っている．構文解析器の種類といくつかのルールにより次の3つのBE Breakerを開発し実験を行っている．

- BE-L: Chaniak parser +CYL cutting rules
- BE-F: Minipar + JF cutting rules
- Chunker: cutting rulesを含むsyntactic-unit chunker
- Microsoft parser + cutting rules

各BE Breakerは評価対象文書の各分を入力とし，それぞれの構文解析器の出力からcutting ruleを用いてBEに分割を行う．実験からはこれらのBE Breakerの出力であるBEの一致度は約40%であった．これは分割する要素の単位やBE Breakerによって扱う関係名が異なっているためである．

現在，BE BreakerはBE Packageとして配布[3] されており，このPackageではBE-FがBE Breakerとして含まれている．BE-Fにおいては，Miniparによる出力である文の依存構造から要素間の関係として *subj (subject)*, *obj (object)*, *compl (complement)*, *mod (modifier)* などの関係を持つ要素をBEとして分割を行っている．Miniparにおいては，複合名詞や動詞イディオムについては，まとめて1つのノードとしてまとめられている．例えば，“Secretary General” や “turn over” は1つのノードとして扱われているため，BEの要素としても1つのものとして扱われる．また，前置詞句については前置詞句が修飾する要素間に前置詞の表現を関係として持つようなBEとして分割する．例えば，*“sanction against Libya”* は $BE[sanction|Libya|against]$ のように関係名againstと2つの要素から構成されている．さらに，埋め込み文の場合には，埋め込み文が修飾する文の主動詞と埋め込み文の主動詞が修飾関係を持つBEとして分割されている．

2.2 評価方法

要約のreference summaryと評価対象の文書を共にBE BreakerによってBEの集合に分割を行い，それらの一致度によって要約の精度の評価を行っている．現在のバージョンでは，BE Breakerによって分割された各BEはすべて同じ重みが与えられている (重みを変えた実験についてはまだ行われていない)．DUCでは複数のreference summaryが与えられており，それらの中にいくつか存在するようなBEについてはその頻度を重みとして与えることで各BEに重要度が与えられている．

BEのマッチングについては (1) 文字レベルの完全一致，(2) 要素の原型レベルでの一致，(3) シノニムレベルの一致，(4) 句レベルの言い換えの一致，(5) 意味的な一致 が提案されている．さらに要素間の部分一致や要素間の参照関係なども扱う必要があるとされているが，現在のレベルは，文字レベル，原型レベルまでの一致を扱っている．

3 質問応答評価プロジェクトQAC

3.1 QAC1, 2, 3の評価タスク

我々はNTCIRプロジェクト[4] のNTCIR-3, 4, 5において質問応答タスク QACの評価を行ってきた．これまでのQAC，NTCIR-3 QAC (QAC1)，NTCIR-4 QAC (QAC2) においては日本語の新聞記事データ数年分を用いて与えられた質問から新聞記事中に存在する名称をexact asnwerとして返す3種類のSubtask1,2,3を設定した．

Subtask1では，回答候補を5つまで順位付きで返し，より上位の順位のものが正解になるほど得点が与えられるものである (順位つきタスク)．この評価方法は，TRECのQAトラックで採用されている手法と同様であり，これにより質問応答の基本的技術の評価を目指したものである．Subtask2は，与えられた質問に対して正解のみをすべて返すものである (リス

[3] http://www.isi.edu/~ cyl/BE/

[4] http://research.nii.ac.jp/ntcir/

トタスク) [1] [2]．正解がない場合も含めて複数の正解となる質問を設定し，複数の回答がある場合にはそれをすべて返すものであり，質問応答の回答抽出の際に，正解とそうでないものをどのように区別するのかについての評価を目指したものである．Subtask3 は，お互いに関連する一連の質問の並びに対して答えを返すものである (コンテキストタスク)[3]．ただし，連続する質問において最初の質問の答えに従って後続質問が変るものは，タスク設定の制約から設定できないため，あらかじめ決められたシナリオに従った連続する質問を設定することで擬似的な対話を実現したタスクとした．回答は Subtask2 と同様にリスト形式とした．

NTCIR-5 QAC (QAC3) では，これまでの subtask3 を拡張し [4]，連続する質問によりレポート作成を目指した情報獲得をシミュレートした Information Access Dialogue (IAD) タスクとし，gathering type と browzing type の 2 つの連続する質問をタスクとして設定し，評価を行ってきた．

3.2 QAC4 の評価タスク

これまでの QAC において対象としてきた質問は名称を質問対象とする factoid 型のみであり，それを超えるタイプの質問は対象外であった．QAC-4 において設定する質問は，名称を対象とした質問文を超え任意の回答を前提としたものを対象にすることとした．[5] 例えば,「なぜ」といった理由を尋ねるものや「何ですか」のようにものの説明を尋ねるものが回答の対象となる質問がこれに含まれる．さらに「どのようにして」のように手順を尋ねたり，意見を尋ねるような質問も含まれる．これらの質問に対してなるべく簡潔な回答を得るためにはどのような技術が必要であるのかの評価を行うことを目的とした．

評価方法に関しては，これまでの QAC では対象が factoid 型の質問であることから，あらかじめ準備された回答セット，もしくは，システムの結果をプーリングして評価したものを蓄えておき，それらとのシステムの結果の回答とのマッチングを取ることで正解かどうかを判定することが可能であった．しかしながら，名称を超える範囲のものを対象にした場合，準備された正解とのマッチングのみでは正解の判定が困難であり，我々はこのような名称を超える質問の回答の評価に BE を用いた評価手法をとることとした．これにより，あらかじめ準備された正解例とシステムの回答をそれぞれ BE に分割し，分割された BE の集合の一致度により評価を行うことを目指す．

QAC4 では，以上の任意のタイプの質問応答の評価だけでなく，システムの回答の評価手法についても評価手法の研究を行うことを目指す方には評価に参加していただき，さまざまな観点からの評価手法の研究も QAC4 の目的である．

以下に QAC4 の評価について述べる．評価は以下の 2 つのトラックを設定している．

1. QA トラック

- 任意の質問文に対する質問応答として，1 つの質問に対する 1 セットの回答を返す．
- 想定する質問としては，従来までの factoid 型質問だけでなく，why, how, definition などのすべての質問文を対象とする．
- 質問文としては 100 問程度を準備し，システムによる回答と人手による回答を返す．
- 回答の評価については，人手による評価をオーガナイザによって行い，ひとつの評価結果として返す．

2. 評価トラック

- 人手による評価や自動評価手法の提案などに関する評価手法を広く求める．
- 評価対象としては，QA トラックで提出されたデータを用いて質問応答の精度の評価を行う．

[5] これまでの factoid 型質問を対象とする質問応答については，CLQA の日-日タスクとして継続されております．

対象テキスト

毎日新聞 1998-2001 年の 4 年分を対象テキストとする．回答を得るために新聞記事以外の知識源を用いることも可能であるが，どのような知識を用いてどのように回答を得たのかについてはNTCIR Workshop にて報告してもらう予定である．

3.3 スケジュールと現在までの活動

現在以下のスケジュールで QAC4 を進めている．

2006.4.15	Call For Participations
2006.5.31	参加申し込み締め切り
2006.6	Sample question set 配布
2006.9	Question set delivery
2006.10	System results due
2006.11.1	Start of Evaluation
2007.2.1	Evaluation results release
2007.3.1	System paper 締切り
2007.5.15-18	NTCIR Workshop 6 meeting

以上のスケジュールにしたがって，2006 年 6 月に 5 問のサンプルの質問と回答のセットを配布し，さらに 2006 年 7 月に 30 問の質問文を参加者に配布を行った．そして，参加者によりサンプル質問の回答の作成を行ってもらい，その回答についての検討会を 2006 年 8 月 8 日に行い，それぞれの質問文についてどのような回答がふさわしいのかの意識あわせと問題点の洗い出しを行った．factoid を超えるタイプの質問応答について以下の意見があった．

情報の種類の網羅性

答えとして与えるべき情報にはいくつかの種類があり，それをどの程度まで与えるのか良いのかといった問題があった．例えば，「世界遺産条約とはどのような条約ですか。」の質問文について，「世界遺産条約」のフルの表現「世界の文化遺産及び自然遺産の保護に関する条約」で答えた場合正解としてどうなるのか．また，条約の目的やいつ採択されたのかの情報をどの程度まで与えるのか回答としていいのかといった意見があった．

情報源について

質問に対する回答を得る場合，意見などの情報を与える必要がある場合，その情報源，つまり，誰がその意見を言っていたのかをその内容である意見と一緒に与える必要があるのかといった問題点があった．

情報の順序について

「どのようにしてですか。」といった手順を問う質問の場合，回答として手順の 1 つ 1 つを答えるだけでなくその順番についても正しくなければ回答として正しくないものもあった．

4 BE に基づく質問応答の自動評価の検討

BE に基づく評価手法を日本語の質問応答に対して適用するため，QAC4 のサンプル質問文について 3 名の人間により回答の作成を行った[6]．そして，それらの回答について内容的な違いにより，回答の種類を分類した．回答によっては 2 つの内容を共に含んでいるものもあった．表 1, 表 2 にその例を示す．表中の各回答例について，その回答例の存在した記事 ID と内容番号を付けている．同じ番号の内容はほぼ同一のものであると判断されたものであり，番号を 2 つ以上あるものについては 2 つの内容を含んでいることを示している．

BE に基づく評価のため，以上の回答例を形態素解析，構文解析を行い，構文解析結果の構文木から依存関係にある 2 つの要素をそれらの間の関係とともに抽出することで，構文解析木を BE の集合に分割を行う．分割された BE を以下に示す．

[6] 本回答作成は QAC4 参加チームの立命館大のデータを用いたものであり，上記の QAC4 の参加者による回答作成作業の一部である．

表 1: QAC4-00001-00: “世界遺産条約とはどのような条約ですか。” の回答例

記事 ID	回答例	内容番号
JA-000101057	世界的な価値を持つ文化遺産や自然遺産は人類共有の財産との認識に立つ。そして、遺産が存在する国の責任を明らかにしながら、国際社会全体が協力して遺産を守ろうとする。	2
JA-001130224	世界遺産条約は、世界的な文化、自然遺産の保護・保存を目的にして、１９７２年の国連教育科学文化機関（ユネスコ）の総会で採択された。	1,2
JA-001207101	１９７２年の第１７回ユネスコ総会で採択された国際条約。	1
JA-001207101	価値ある遺産を国際協力によって保護し、次の世代に伝えていくことをうたっている。	2
JA-001207101	１９７２年の第１７回ユネスコ総会で採択された国際条約。価値ある遺産を国際協力によって保護し、次の世代に伝えていくことをうたっている。	1,2
JA-010219121	１９７２年の第１７回ユネスコ総会で採択された。	1
JA-980227197	「世界の文化遺産及び自然遺産の保護に関する条約」	2
JA-981202054	多様な世界の民族と風土に根差す「驚嘆すべきもの」の保護の盾	?

表 2: QAC4-00006-00: “住基ネットのメリットとは何ですか。” の回答例

記事 ID	回答例	内容番号
JA-000817207	住民の負担軽減が約２７０億円	1
JA-000817207	国及び地方自治体の事務軽減などが約２４０億円	2
JA-010911147	住民サービスの向上。市区町村が発行する「住民基本台帳カード」というＩＣカードを利用して、居住地以外の自治体で住民票の写しが取得できたり（広域交付）、引っ越し手続きが転居先の自治体だけで済む（特例転出入）	3
JA-010911147	国の行政事務の効率化。恩給の支給や雇用保険、不動産鑑定士や宅地建物取引主任者の登録など本人確認を必要とする１０省庁９３件の国の事務に利用するという。	4
JA-011128129	住民にとって便利	3
JA-011128129	住民は居住地以外でも住民票の交付を受けられる	3
JA-011128129	国は宅地建物取引主任者など９３件の本人確認事務に利用できる	4

質問文QAC4-00001-00の記事IDがJA-001207101とJA-010219121の回答例をBEに分割した例を以下に示す．BEはBE番号 [要素1, 要素2, 関係名] の形式をとっている．

- JA-001207101
  BE1:[第１７回ユネスコ総会, １９７２年, の]
  BE2:[採択された, 第１７回ユネスコ総会, で]
  BE3:[国際条約, 採択された, 採択された]

- JA-010219121
  BE4:[第１７回ユネスコ総会, １９７２年, の]
  BE5:[採択された, 第１７回ユネスコ総会, で]

BE1については，「１９７２年の第１７回ユネスコ総会」から2つの要素が関係「の」で結ばれている．また，BE3では「採択された」が「国際条約」を修飾していることから2つの要素間の関係を修飾している動詞「採択された」が関係名として設定されている．そして，これらの間の一致するBEを多く持つ場合，内容的に近いものを示していると考えられる．以上の例ではBE1とBE4が一致，BE2とBE5が一致しており内容的な類似性は高いと考えられる．

また，同じ質問文について記事IDがJA-000101057とJA-001207101の回答例をBEに分割した例を以下に示す．

- JA-000101057
  BE6:[持つ, 価値, を]
  BE7:[自然遺産, 持つ, 持つ]
  BE8:[守ろうとする, 遺産, を]
  BE9:[価値, 世界的, な]
  BE10:[自然遺産, 文化遺産, や]
  BE11:[立つ, 自然遺産, は]
  BE12:[財産, 人類共有, の]
  BE13:[認識, 財産, との]
  BE14:[立つ, 認識, に]
  BE15:[守ろうとする, そして, そして]
  BE16:[存在する, 遺産, が]
  BE17:[国, 存在する, 存在する]
  BE18:[責任, 国, の]
  BE19:[明らかにしながら, 責任, を]
  BE20:[守ろうとする, 明らかにしながら, 明らかにしながら]
  BE21:[協力して, 国際社会全体, が]
  BE22:[守ろうとする, 協力して, 協力して]

- JA-001207101
  BE23:[ある, 価値, 価値]
  BE24:[遺産, ある, ある]
  BE25:[保護し, 遺産, を]
  BE26:[保護し, 国際協力, によって]
  BE27:[伝えていく, 保護し, 保護し]
  BE28:[世代, 次, の]
  BE29:[伝えていく, 世代, に]
  BE30:[こと, 伝えていく , 伝えていく]
  BE31:[うたっている, こと, を]

この例では，言いかえとして，「価値ある」を「価値を持つ」に，「遺産」を自然遺産」に，「保護する」を「守ろうとする」に言い換えることができた場合，BE6,7,8がBE23,24,25と同一のものであると考えることができる．これによって，内容の中の「遺産を守る」というが同一のものであると捉えることができる．

以上のBEレベルでの内容の比較を行った結果は一致する部分が少ないという結果となった．表層的に類似しているものについては一致度が高くなるが，意味的に同じ内容で表層表現が異なるものを比較するためにはまだ多くの課題がある．しかしながら，システムの回答として基本的にextractを前提とした場合の正解例との比較であれば，いくつかの言い換え例をあらかじめ準備しておくことで，BEの一致についてはある程度は可能なのではないかと考えられる．

BEによる評価としてBEの一致度と共に重要なものがBEの重要度である．質問応答の回答については前節でも述べたように，回答として内容的にいくつかのものが存在しており，各内容についても詳細度のレベルが異なっているものがある．これらの中でより重要なものについて判断するためには，人手により作成された回答例から多く出現するBEに重みを与えるなど，要約の評価でも用いられてきた手法を用いることが可能である．

質問応答は要約と異なり，複数の回答が存在しており，それぞれ異なる内容の回答とのシステムの回答を比較する必要がある．このため，システムの複数の回答と正解例との比較が必要であると考えられる．

5 おわりに

本稿では，質問応答技術において，factoid を超える質問の評価のため，要約において用いられている BE に基づき，回答の評価手法についてこれまでに行ってきた評価実験について述べた．

QAC4 ではこれから質問応答の評価というスケジュールであり，さらに多くの課題が出てくるものと予想される．factoid を超えるタイプの質問についての回答抽出方法についてだけでなく，評価手法についてもタスクを進めていきながら，参加者間の議論を通じ，QAC4 が質問応答技術に貢献できることを願っている．

謝辞

本稿をまとめるにあたって，多くの貴重な助言を頂きました QAC4 タスクへの参加者，QAC タスク検討会に参加された皆様，QAC-ML で議論に参加された皆様に感謝いたします．

参考文献

[1] J. Fukumoto, T. Kato, and F. Masui. Question and answering challenge (QAC-1) : Question answering evaluation at NTCIR workshop 3. In *Working Notes of the Third NTCIR Workshop Meeting Part IV: Question Answering Challenge (QAC-1)*, pp.1–10, 2002.

[2] J. Fukumoto, T. Kato, and F. Masui. Question answering challenge for five ranked answers and list answers – overview of NTCIR4 QAC2 subtask 1 and 2 –. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.283–290, 2004.

[3] T. Kato, J. Fukumoto, and F. Masui. Question answering challenge for information access dialogue – overview of NTCIR4 QAC2 subtask 3–. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.291–296, 2004.

[4] T. Kato, J. Fukumoto, and F. Masui. An Overview of NTCIR-5 QAC3. In *Proc. of the Fifth NTCIR Workshop Meeting* , pp.361–372, 2005.

[5] E.M.Voorhees. Overview of TREC 2005. In *http://trec.nist.gov/pubs/trec14/t14_proceedings.html*, 2005.

[6] A. Nenkova and R. Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proc. of the HLT-NAACL conference*, 2004.

[7] E. Hovy, C. Y. Lin, L. Zhou and J. Fukumoto. Automated Summarization Evaluation with Basic Elements. In *Proc. of the Fifth Conference on Language Resources and Evaluation (LREC 2006)*, 2006.

[8] 加藤, 福本, 桝井, 神門. 質問応答から対話理解へ – NTCIR QAC task3 の提案 –. 第 10 回年次大会発表論文集, 言語処理学会, pp.317-320, 2004.

[9] 加藤, 福本, 桝井, 神門. 質問応答技術は情報アクセス対話を実現できるか. 情処研究会報告 NL-162-21, pp.145-150, 2004.

[10] 加藤, 桝井, 福本, 神門. リスト型質問応答の特徴付けと評価指標. 情処研究会報告 NL-163-16, pp.115-122, 2004.

[11] 桝井, 福本, 加藤, 神門. 質問応答システム評価用テストコレクションの構築 – NTCIR QAC の取り組み –. 第 10 回年次大会発表論文集, 言語処理学会, pp.313-316, 2004.

[12] 加藤, 桝井, 福本, 神門. 情報アクセス対話に向けた質問応答技術の評価 ふたたび– NTCIR5 QAC3 での試み–. 情処研究会報告 NL-172-8, 情報処理学会, pp.55-62, 2006.