情 報 学 基 礎　64−12
自 然 言 語 処 理　145−12
（2001. 9. 11）

SAIQA: 大量文書に基づく質問応答システム

佐々木 裕　磯崎秀樹　平 博順　平尾 努　賀沢秀人　鈴木 潤　国領弘治　前田英作

日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

〒619−0237 京都府相楽郡精華町光台 2−4
TEL: (0774) 93-5360
E-mail: sasaki@cslab.kecl.ntt.co.jp

あらまし　従来の TREC 流の質問応答（Question Answering: QA）では，質問に対する解答を含む 250 バイトまたは 50 バイトのパッセージを答えていた．しかしながら，質問応答技術の最も興味深い点は，質問に対して解答を含む文書やパッセージを返すのではなく，解答そのもの（exact answer）を判定し，答えることができる点にある．そこで，本稿では，大量の文書集合に基づいて質問に答える日本語質問応答システム SAIQA について述べるとともに，2000 問の質問文について評価を行なった結果を示す．評価には，解答の根拠として，文書全体を与えた場合と，質問に適応した要約（Question-Biased Text Summarization: QBTS）を与えた場合の評価も含む．評価の結果，5 位以内の正解率は約 50 %であり，正解の場合には，要約の約 87 %が根拠として正しいことが明らかになった．

SAIQA: A Japanese QA System Based on a Large-Scale Corpus

Y. Sasaki, H. Isozaki, H. Taira, T. Hirao, H. Kazawa, J. Suzuki, K. Kokuryo and E. Maeda

NTT Communication Science Laboratories, NTT Corporation

2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237 Japan
TEL: (0774) 93-5360
E-mail: sasaki@cslab.kecl.ntt.co.jp

Abstract　Conventional TREC-style *Question-Answering (QA)* involves extracting *passages* (250 bytes or 50 bytes) that contain answers to a question. The most attractive feature of QA is that it can provide *exact answers* to the question, rather than a list of ranked passages, paragraphs, or documents. This paper describes a Japanese QA system SAIQA which finds exact answers to a question from a large-scale Japanese text corpus and an experimental evaluation on exact answer extraction. The experiments include evaluations of answers to 2000 questions with justification by article summarization. The results show that around 50% of correct answers can be found from the top of five answers and that over 87% of short summaries can be used for answer justification (or *evidence*) instead of full texts.

1 はじめに

本稿で述べる質問応答 (Question Answering: QA) とは，新聞記事 10 年分といった大量文書集合から質問に対する答えを探し出し，ユーザに提示する技術である．たとえば，「英国のビクトリア女王が即位したのは何年ですか？」という質問に対して，新聞記事を参照することにより「1837 年」と答える．

与えられた特定の単一の文書に関する質問に答えるシステムの研究は，以前から行なわれており，DUALS[3][5] の研究などがある．DUALS は 200 文程度までの特定の国語の文章題の内容を理解し，問いに答えることを目標にしたシステムである．DUALS は，たとえば，「旅客機が飛行中にエンジンから煙りが出た」という内容の小学校 3 年生向けの文章が与えると，「き長はだれですか」「ロールさんがはっとしたのはなぜですか」といった質問に答える．

1999 年から始まった TREC (Text REtrieval Conference) の QA TRACK では，DUALS のように対象を絞って深い理解を行うことにより質問に答えられるかを探るのではなく，情報検索の研究の延長として，大量の文書を使ってどの程度質問に答えられるかを探るアプローチをとっている．

1999 年の TREC-8 の QA Track の概要は，以下のようにまとめられる [8]．

- 約 528,000 記事を対象．
- 正解が短い事実情報となるような質問文 200 問を与える．
- 解答文字列とその文字列が現れる記事 ID のペアを 1 位～5 位までランキングして答える．
- 50 バイトで答える課題と 250 バイトで答える課題の 2 つの課題について評価する．

このような TREC 流の QA 研究については，日本語と対象とした研究として文献 [4] がある．また，我々が行なった日本語 QA システムの比較・評価に関する研究 [7] においては，50 問の質問文セット NTT-QA-200-08-22-FRUN により，いくつかの QA 手法を比較した．この実験において，5 位までの正解率で見ると最高が 54 % であった．また，この研究の中では，比較・評価を行なう上での問題点や困難な点も探っている．さらに，日本語質問応答技術の比較・評価に関しては，国立情報学研究所主催の NTCIR ワークショップ 3 における新たな課題として，Question and Answering Challenge (QAC)[1] が開催され，QA 技術の比較・評価が行なわれる予定である．

2 SAIQA の概要

この節では，日本語質問応答システム SAIQA (System for Advanced Interactive Question Answering) の構成について述べる．

SAIQA は次の 4 つのモジュールを順に実行する．

質問解析モジュール (QAM) 質問文を解析し，質問タイプ，質問対象の検索語，補助語，単位語，意味カテゴリを判定する．

テキスト検索モジュール (TRM) パラグラフを検索語によりスコアリングし，スコアの上位 N 件を検索する．[2]

解答抽出モジュール (AEM) 質問に対する解答をパラグラフから取り出す．

テキスト要約モジュール (TSM) 解答の根拠となる要約を，検索語と解答から作成する．

SAIQA は各質問について，解答と要約を 5 位までランキングして答える．各モジュールの詳細は以下の通り．

2.1 質問解析モジュール (QAM)

質問文リスト Qs が与えられると，各質問文 $q \in Qs$ について，QAM は次のような情報を取り出す．

- 質問タイプ: QT
- 検索語: KT
- 補助語: AT
- 単位語: AU
- 意味カテゴリ: SC

意味カテゴリには，日本語語彙大系の意味体系 [2] を使用している．意味体系は，日英翻訳システム ALT-J/E のために開発された，木構造をした一種の概念シソーラスである．ノードが意味カテゴリを表し，リンクが *is_a* 関係や *has_a* 関係を表す．12 段の深さがあり，約 30 万語の単語が約 3000 の意味カテゴリにリンクされている．

質問は以下のような手順で解析される．

[1] http://www.nlp.cs.ritsumei.ac.jp/qac/
[2] 現在 N は 20 に設定されている．

表 1: 主要な質問タイプ一覧

質問タイプ	説明	例
PERSON	人名	小泉，ブッシュ
LOCATION	地名	北アメリカ，米国，名古屋
ORGANIZATION	組織名	アメリカ政府, NTT，関西国際空港会社
ARTIFACT	製品名，作品のタイトル	カローラ,「徹子の部屋」
DATE	日付	1月，クリスマス，5月4～7日
TIME	時間	午後3時，7:30AM
PERIOD	期間	10年間，1分5秒
MONEY	金額	$5，1000円，20ペソ
PERCENT	割合	5%，半分，3分の1
PTITLE	役職，職業	副社長，県議
LENGTH	長さ	50ヤード，5マイル，2cm
...		
ANY	その他	

1. 質問 q を翻訳システム ALT-J/E の形態素解析部を使い解析する．このとき，名詞には日本語語彙大系の意味カテゴリ（複数可）が与えられる．

2. q の質問タイプ QT を質問パターン辞書とマッチさせることにより決定する．たとえば,「～の会場はどこですか?」というパターンに合えば，質問タイプは LOCATION となる．質問タイプは，IREX[3] の固有表現の分類を拡張して使っている．[6] 質問タイプの一覧を表 1に示す．なお，QT は複数の質問タイプを含んでもよい．例えば,「～したのはどこですか?」は「どこ」が組織名を指している場合と，場所を指している場合がある．その場合は $QT = \{ORGANIZATION, LOCATION\}$ となる．

3. q に現れる自立語を検索語 KT とする．但し,「誰」「名前」などのストップワードは KT から除く．

4. かぎ括弧で囲まれた文字列およびドット（・）で結合されたカタカナの列をすべて取りだし，補助語 AT とする．

5. リットルや個など単位を表す語を単位語 AU とする．また，特定の質問タイプの正解に頻繁に含まれる特徴的な語尾表現も単位語に追加する．たとえば，質問タイプが PTITLE の時は，単位語に「～家」「～ニスト」を追加する．

6. パターンマッチを使ったルールにより，q の質問対象の意味カテゴリ SC を判定する．たとえば,「質問文が『X はどこですか』というパターンにマッチし，かつ X の意味カテゴリ s_X が，施設を表す意味カテゴリに含まれれば，$SC = s_X$」という規則により,「ミロのビーナスがある美術館はどこですか？」の SC は「公共機関，博物館」となる．

[3]Information Retrieval and Extraction Exercise (http://cs.nyu.edu/cs/projects/proteus/irex/)

2.2 テキスト検索モジュール (TRM)

TRM は検索語 KT を含むパラグラフを見つける．事前に，与えられた文書集合のインデックスを作成しておく．インデックスは，文書中の各単語からその単語が現れるすべてのパラグラフ番号への連想リストである．

パラグラフ中の語 w のスコア WS_w は次のように計算される．

$$WS_w \stackrel{\text{def}}{=} \log(\min(2, tf_w) + 1.0) \times idf(w)$$

ここで，$idf(w)$ は次のように定義される．

$$idf(w) \stackrel{\text{def}}{=} \log\left(\frac{D}{df(w)}\right) \quad (1)$$

$df(w)$ は 単語 w の document frequency であり，D は文書数である．tf_w は 単語 w のパラグラフ中の出現数である．パラグラフ p のパラグラフスコア PS_p は次の通り．

$$PS_p \stackrel{\text{def}}{=} \sum_{w \in (KT \cup AT) \cap p} WS_w$$

パラグラフのスコア PS_p により，上位２０件のパラグラフを選択する．

2.3 解答抽出モジュール (AEM)

AEM は検索語 KT と隣接して現れる語を解答候補 TT として抽出する．検索されたパラグラフ P に現れる語に対して，次のようなヒューリスティックスを適用することにより TT を収集する．

- 単位語 AU を含む語を TT に追加する．たとえば,「何個のチョコレートを食べましたか?」という質問の $AU = \{$ 個 $\}$ のとき，パラグラフ中の「５個」といった語を TT に追加する．
- 意味カテゴリ SC に合致する語を TT に追加する．たとえば,「～がある美術館はどこですか?」について，意味カテゴリが「公共機関，博物館」に含まれる語が追加される．
- かぎ括弧で括られた文字列は TT に追加する．たとえば「ロミオとジュリエット」を TT に追加される．これは，固有表現抽出が作品名などをあまり高い精度で取れないことと，かぎ括弧が強調を表している場合に質問の対象となりやすいためである．
- 質問タイプ QT に含まれる固有表現を TT に加える．たとえば,「～は誰ですか?」の質問タイプは PERSON と判定されるので，パラグラフ中の PERSON の固有表現を TT に加える．

文中の検索語 k と解答候補 t の距離 $\Delta_{k,t}$ は次のように定義される．

$$\Delta_{k,t} \stackrel{\text{def}}{=} k \text{ と } t \text{ の間の単語数} + 1$$

検索語 k_i と最も近い解答候補との距離 δ_{k_i} の定義は次の通り．

$$\delta_{k_i} \stackrel{\text{def}}{=} \min_{t_j \in TT \cap p} \Delta_{k_i,t_j}$$

対称的に，解答候補 t_j と最も近い検索語との距離 δ_{t_j} は，

$$\delta_{t_j} \stackrel{\text{def}}{=} \min_{k_i \in (KT \cup AT) \cap p} \Delta_{k_i,t_j}$$

パラグラフ p のスコア EPS は以下のように定義される．

$$EPS_p \stackrel{\text{def}}{=} \frac{1}{n} \sum_{k_i \in (KT \cup AT) \cap p} \frac{1}{\exp(0.3\delta_{k_i})} \omega_{k_i}$$

但し，$n = |\,(KT \cup AT) \cap p\,|$，また ω_{k_i} は，$k_i \in KT$ のとき 1，$k_i \in AT$ のとき 2 である．EPS_p は検索語が近くに現れる解答候補を多く含むパラグラフ程，スコアが高くなる．

以上により，次のようにパラグラフ p の解答候補 t_i の抽出スコア ES の定義を与える．

$$ES_p(t_i) \stackrel{\text{def}}{=} EPS_p - 0.01\delta_{t_i} + PS_p$$

つまり，解答候補は基本的には検索スコアとパラグラフスコアによりランキングされるが，検索語との距離が遠いものは下位に来るように調整される．

解答候補 TT のうち，$ES_{p_i}(t_i)$ のスコアの高い上位５件が解答として選択される．

2.4 テキスト要約モジュール (TSM)

TSM は各解答について，根拠となる要約を作成する．要約手法は「質問に適応した要約」(Question-Biased Text Summarization: QBTS)[1] を用いている．ここでは，紙面の制限により，要約の手法を [1] に沿って簡単に述べる．

ハニング窓関数

ウインドウサイズを W，l をウインドウの中央位置とする．ウインドウ中の位置 i について，ハニング窓 $f_H(i,l)$ は次のように定義される．

$$f_H(i,l) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{2}\left(1 + \cos 2\pi \frac{i-l}{W}\right) & (|i-l| \le W/2) \\ 0 & (|i-l| > W/2) \end{cases} \tag{2}$$

中央位置 l のスコア $S(l)$ は次の通り．

$$S(l) \stackrel{\text{def}}{=} \sum_{i=l-W/2}^{l+W/2} f_H(i,l) \cdot a(i) \tag{3}$$

但し，$a(i)$ は以下のような定義である．T を AEM により抽出された解答とする．

$$a(i) \stackrel{\text{def}}{=} \begin{cases} idf(k) & if \text{ 位置 } i \text{ の単語が } k(\in KW) \text{ である} \\ \alpha & if \text{ 位置 } i \text{ の単語が } T \text{ である} \\ 0 & otherwise \end{cases} \tag{4}$$

ここで，α は解答に対する重みである．$S(l)$ の l を各パラグラフの先頭から末尾まで動かしながら，

表 2: 2000 問の分類

質問タイプ	
PERSON	323
ORGANIZATION	321
LOCATION	340
DATE	377
ARTIFACT	175
TIME	25
PERIOD	52
MONEY	50
PERCENT	38
PTITLE	54
上記以外	245
合計	2000

$S(l)$ を計算し，このスコアによりウインドウをランキングする．上位のウインドウに含まれる文を順に取りだし，150 文字に最も近くなる文数を選択し，要約として出力する．詳しくは文献 [1] を参照されたい．

3 実験と評価

SAIQA は毎日新聞 9 年分約 100 万記事をもとに質問に答えられるが，ここでは 9 9 年 1 年分の記事を対象に実験を行なった．

3.1 質問セット

TREC-8 流の質問仕様に加えて，文献 [7] と同様に，固有表現を答える 2000 問の質問文を，毎日新聞 99 年を参照しながら作成した．なお，SAIQA システムと質問セットは独立に作成された．各正解の固有表現種別による，質問数の分布は表 2のようになっている．

3.2 評価

1 つの質問について，SAIQA は正解と要約およびその正解が含まれる記事番号を，1 位から 5 位までランキングして答える．評価には次のような尺度を用いた．

Top5 スコア $Top5 \stackrel{\text{def}}{=} R/|Qs| \times 100$ 但し，$|Qs|$ は質問数，R は 5 位以内に正解が含まれた質問数である．

NIST スコアは質問について，ランクの 1 位から 5 位まで順に正解かどうかをチェックしていき，最初に正解と判定されたランク n のポイント $1/n$ を与え，質問数で平均したもの．これらの 2 つの評価尺度において，正解の判定には，根拠を含めて正しいとする評価と含めない評価の 2 通りを行なった．

要約有用性 (Summarization Utility: SU) は記事全体の代わりに要約が正解の根拠となった割合を示す．

3.3 実験

質問セットの 2000 問のうち，ランダムに抽出した 300 問について，人手により評価を行なった．1 つの答えは 1 人の採点者により評価した．また，参考のため，2000 問に対する評価も行なった．2000 問全体に対する評価は作業が膨大になるため，今回は，あらかじめ用意した正解セットにより自動的に評価を行なった．但し，現在，正解セットは 1 つの質問につき正解と記事を 1 ペアしか用意できなかったため，別解を考慮しない荒い評価にとどまっている．

3.4 結果

表 3 に 2000 問についての評価結果を示す．Top5 スコアは記事番号のチェックをしない場合，54.4%であり，記事番号のチェックを含めると 42.2%であった．NIST スコアは，記事番号のチェックなしで 38.3，チェックありで 28.8 であった．

人手による 300 問の評価結果を表 4に示す．質問タイプ別に正解率は，人名の Top5 が 76.5%，NIST が 61.6%であり，日付の Top5 が 85.7%，NIST が 71.4%であった．役職やその他の数値表現については低い値となった．要約有用性 SU はどの種別でも高く平均して 86.6%であった．

4 考察

TREC-8 流の QA システムの評価では，正解が 50 バイト，250 バイトといったパッセージに含まれるかどうかを評価していた．パッセージに含まれるどの部分が解答なのかは人間が判断する必要があった．また，パッセージの中に偶然正解が含まる場合も含んでいる．これは，QA システム結果をさらに言語処理・知識処理システムで利用する場合には障害となる．そこで，SAIQA は解答そのものを出力するような仕様とした．また，解答の評価は，正解と解答が同一かどうかの単純評価でシステムの自動評価が可能となる．実験の結果，SAIQA は 2000 問の Top5 スコアで 42.2%～54.4%となった．

300 問についての人手による評価では，質問タイプにより正解率に差が見られた．LOCATION と ORGANIZATION の正解率があまり良くないのは，

表 3: ２０００問全体の評価結果

	$\|Qs\|$	Top5 (%)	NIST
完全一致のみ	2000	54.4	38.3
完全一致＋根拠記事正解	2000	42.2	28.8

表 4: サンプリングした３００問の評価結果

質問タイプ	$\|Qs\|$	Top5 (%)	NIST	SU (%)
PERSON	102	76.5	61.6	85.9
LOCATION	48	58.3	38.3	96.4
ORGANIZATION	45	71.1	48.0	81.3
ARTIFACT	33	60.6	48.6	85.0
DATE	7	85.7	71.4	83.3
PTITLE	5	20.0	6.7	100
上記以外	60	35.0	24.3	85.7
合計・平均	300	62.0	46.2	86.6

「～はどこですか?」のように質問タイプが LOCATION か ORGANIZATION かを絞り込めない場合があるからである．また，ARTIFACT やその他で正解率が低いのは，表現の幅が広く，対象の特定が難しいためであると考える．

要約有用性が 86.6%であったことで，一般に，150 文字程度の要約により解答の根拠を与えることが可能であることが分かった．

最後に，今回の実験では記事を参照して質問文を作成したため，想定されるユーザによる普通の質問文よりも，解答が容易であった可能性がある．質問文の作成法やその分類，難易度の判定については，今後の研究が必要である．

5 おわりに

本稿では，日本語 QA システムにおける，解答の抽出と要約による根拠の提示について述べた．2000 問の質問文について，大枠の自動評価を行ない，300 問について人手による詳細な評価を行なった．評価結果により，大規模な質問文セットによる評価においても，約 50 %の正解率が得られることと，要約により約 87 %の場合に根拠を与えることが可能であることが分かった．

参考文献

[1] Tsutomu Hirao, Yutaka Sasaki, and Hideki Isozaki, An Extrinsic Evaluation for Question-Biased Text Summarization on QA Tasks, NAACL-2001 Workshop on Automatic Summarization, pp. 61–68, 2001.

[2] 池原 悟, 宮崎正弘, 白井 諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林 良彦: 日本語語彙大系（１ 意味体系）, 岩波書店 (1997).

[3] 向井国昭：談話理解への応用，古川康一，溝口文雄（共編），「自然言語の基礎理論」第 6 章，共立出版，1986.

[4] 村田真樹，内山将夫，井佐原 均: 類似度に基づく推論を用いた質問応答システム，自然言語処理研究会, 2000-NL-135, 2000.

[5] 杉村領一，田中裕一，橋田浩一，向井国昭: 談話理解実験システム DUALS 第 3 版における自然言語処理，人工知能学会誌，Vol. 4，NO. 3，pp. 49–60, 1989.

[6] 佐々木裕：トランスデューサによる日本語固有表現抽出，第 5 回言語処理学会年次大会，pp. 108–111, 1999.

[7] 佐々木 裕，磯崎秀樹，平 博順，廣田啓一，賀沢秀人，平尾 努，中島浩之，加藤恒昭：質問応答システムの比較と評価，信学技報，NLC-2000-10，pp. 17–24，2000.

[8] Voorhees, E. M. and Tice, D., Building a Question-Answering Test Collection, Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval, pp. 192–199, 2000.