大域知能へ向けての係り受け関係に基づくグラフ構造を用いた質問応答機構

石塚 満　　(協力者：倉田岳人，岡崎直観)
情報理工学系研究科　電子情報学専攻

概 要

情報流通，情報共有の基幹的インフラストラクチャになってきた WWW (World Wide Web) に知的能力を付与し，広域知能基盤に成長させるための一つのアプローチとして，本研究では質問応答機能の研究を行った．現在のサーチエンジンが与えられたキーワードを含む Web ページを検索，提示するのに対し，質問応答機能によれば自然言語文の質問に対する所望の回答をピンポイントで提示することが可能になる．しかし，その精度はまだ必ずしも十分なレベルではない．ここでは新聞記事を対象にして行った研究を示しているが，今後 Web 情報空間への適用を図っていく予定である．

1　まえがき

近年，計算機性能の向上や様々な電子化された文書の整備により，自然言語処理に関する研究が盛んに行われている．質問応答とは，自然言語で与えられた質問文に対して大量文書中から適切な解答を導き出す技術であり，TREC[1] や NTCIR[2] などの評価型ワークショップも開催され，近年注目されている．

質問応答を実現する場合，多くの解答候補に対して，適切な順位付けを行うことが必要となる．しかし，この処理に関して優れたアルゴリズムはいまだ確立されていない．我々は，係り受け解析に基づくグラフ構造を用いることにより，従来手法より高精度のシステムを構築することができたので，ここに報告する [3].

2　日本語質問応答に関する従来手法

2.1　質問応答の流れ

日本語質問応答を実現するための一般的な流れを図 1 に示す．

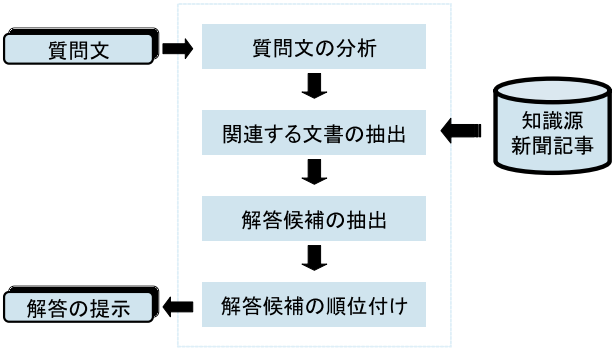

図 1: 質問応答システムの一般的な流れ

2.2　関連研究

従来行われていた質問応答に関する研究では，主に Answer Selection の部分を改良することがよく行われていた．その中のいくつかを簡単にまとめる．

2.2.1　木構造の類似度に基づく順位付け

テキストの構文的類似度を求める尺度がいくつか提案されている [4]．高橋等は質問文と解答候補を含む文との類似度に従って質問応答を実現する手法を提案している [5]．しかし，これらの手法は計算量が膨大になり，また高い精度を得るには至っていない．

2.2.2　単語の属性に従ったルールを大量に記述する順位付け

Lee らは，単語の属性などを Lexico-Semantic Pattern という形で大量に記述する手法を提案している [6]．この手法は非常に精度の高い結果を残しているが，質問文などを非常に多くのパターンに分類し手間のかかる手法であると言える．

2.2.3　解答候補と検索語の距離に基づく順位付け

検索語と解答は近い位置に現れる，ということを前提とし，解答候補と検索語の距離に基づいて順位付けを行う手法がいくつか提案されている．福本らは，解答候補と検索語の位置関係に基づいた手法を提案している [7]．また単語距離に着目した手法もいくつか提案されている．ここで前提として与えられている,「検索語と解答の位置は近い」ということは非常に重要と考えられるが，単語間距離に着目した手法で高精度の結果が得られている手法はない．

2.3　従来手法の問題点

前述したように，単語間距離に基づく順位付けは非常に有効であると考えられる．しかし，単語間距離を用いて，高精度な結果を残しているシステムは未だに構築されていない．これは，単純な単語間距離を用いると，余計な文節などが間に入ったりすることがよく起こり，解答候補と検索語の距離が大きくなる場合があることが原因と考えられる．

3　提案手法の概要

本報告では，上述したような単純な単語間距離に基づく手法の問題点を克服できるように，グラフ構造を用いた距離尺度を導入する．グラフ構造を導入することにより，単純な単語間距離の問題点以外にも，質問応答を実現するための様々な問題点に対処することができる．これらについては 6 においてまとめることとする．本節では，本報告で提案するグラフ構造を用いた日本語質問応答システムについて説明する．図 2 に，提案手法の概要を示した．

以下に，処理の流れを示す．

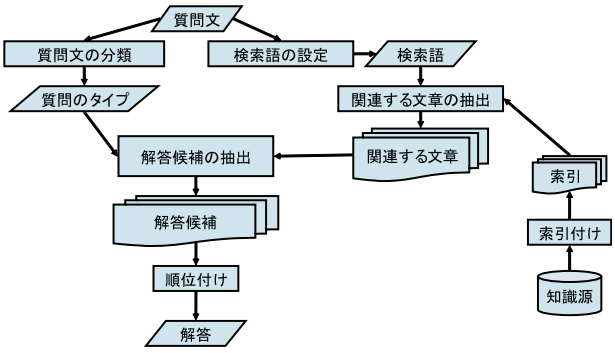

図 2: 構築したシステムの概要

3.1　新聞記事の索引付け

質問応答システムでは，大量のテキストを知識源として扱う．今回実験で用いた NTCIR-3 の QAC1 の場合，新聞記事 2 年分を扱った．このような大量の記事を扱うために，事前にどの単語がどの文書に含まれるかを調べ，索引付けを行った．今回は，各文書の各文ごとに索引付けを行った．索引付けには “Namazu” の *mknmz* コマンドを用いた．また，索引付けを行う際に必要となる日本語の分かち書きには，辞書の整合性を保つために，“茶筌” を用いた．

なお，新聞一年分のデータを文ごとに切り分け，“茶筌” で解析し，索引付けをするために要した時間は，CPU：Pentium4 2.8GHz，メモリ：1GB のマシンを用いて，およそ一日程度であった．

3.2　質問文の分析

3.2.1　検索語の設定

自然言語で記述された質問文から，索引付けされた新聞記事を検索するためのキーワードを選択する．検索語の選択手順の概要を以下に示す．

1. 質問文を “茶筌” で解析する．
2. “茶筌” での解析結果から得られる最初の検索語セット $K = (k_1, k_2, \ldots, k_n)$ で検索を行う．
3. 検索語が多すぎて “Namazu” が検索結果を返さない場合，K から検索語を減らした検索語セット K' で再び検索を行う．
4. 文章が得られるまで,（3）の操作を繰り返す．
5. 検索語セット $K''^{\cdots\prime}$ がキーワードを全く含まなくなった時点で，質問に対する解答を発見できなかったことをユーザに提示する．

ここで,（2）でどのような検索語セットを作成するか,（3）で検索語を減らしていくアルゴリズムなどは，本論文の範疇を逸脱するため，ここではふれないこととする．

3.2.2　推定される解答の形に基づく質問文の分類

質問文を幾つのタイプに分類するかに関しては様々な手法が提案されている．NTT が構築している質問応答システム SAIQA[8] では 30 種類程度に，SAIQA-2 では 100 種類程度に分類を行っている．しかし，質問文に対する分類の数を多くすることは，それだけ手動でのルールの設定が多く必要になり，また，システムの肥大化も招く．そのため，本報告では，4 種類のタ

イプに分類することとした．以下に各々の概要を示す．

Type 1 このタイプには以下の様な質問が属する．

⋆ 木星は何個の衛星を持っていますか．
⋆ 江戸幕府は何年続きましたか．

このタイプの質問は,「何 + 単位」もしくは「何 + 接尾語」という形での質問を行っている．その結果，解答となる表現の単位，もしくは接尾語が限定されることとなり，解答の候補の数が非常に少なくなる．

Type 2 このタイプには以下の様な質問が属する．

⋆ 日本人で初めて大リーガーになったのは誰ですか．
⋆ スペイン村は三重県のどこにありますか．

このタイプの質問文は，疑問詞が解答の属性を指示している．例えば,「誰」という質問に対する解答は「人名」である，ということがわかる．このタイプの質問に関しては上述した固有表現抽出を用いることができる．今回は，以下の表 1 示した疑問視についてこのタイプに分類することとした．

表 1: Interrogative Words in Type 2

疑問詞		解答
だれ，誰	⇒	人名
どこ，何処	⇒	地名，会社名，機関名
いつ，何時	⇒	時間，日付

Type 3 このタイプには以下の様な質問が属する．

⋆ 東京湾アクアラインの全長はどのくらいですか．
⋆ リニアモーターカーの走行試験で出た最高速はどのくらいでしたか．

このタイプに属する質問は,「どのくらいですか」のような表現で，解答として数字表現を求めている．つまり解答の候補としては，数字表現を抽出すればよい，ということとなる．ここで，数値表現に対する単位がわからないことが，Type 1 との違いとして挙げられる．

Type 4 このタイプには以下の様な質問が属する．

⋆ マカオはポルトガル語でどのように表しますか．
⋆ DVD とは何のことを指しますか．

このタイプに属する質問は，疑問詞や「何 + 接尾語」のような形で解答の属性や表現に関する情報を明示的には示していない．つまり解答の候補の絞り込みを行うことができず，解答の候補の数が非常に大きくなり，正解を提示することが困難となる．

3.3 コーパスからの記事の抽出

提案手法では，質問に対する解答は，コーパス中の検索語を含む文，もしくはその近傍に存在するということを仮定する．そのため，解答を含む可能性のある文として，質問文中の検索語を含む文をまずはじめに抽出した．そして，検索語を含む文に解答の候補が含まれない場合は，その文の前後の文を抽出することとした．ただし，前後の文を抽出する場合，代名詞の問題が生じる．なお，“Namazu” で検索を行った場合，tfidf に従って順位付けを行って出力が行われるため，大量の文が検索結果として返された場合はより順位の高い文を選択した．このコーパスからの文の検索の流れを図 3 に示す．

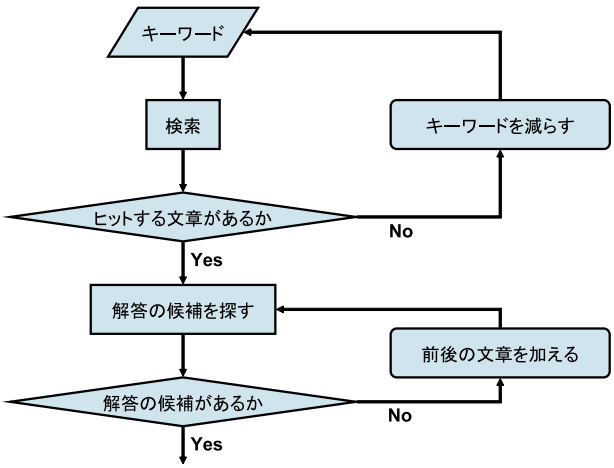

図 3: 文章の検索の流れ

3.4 解答候補の抽出

3.2.2 で推定された解答の形に基づき，3.3 で抽出された文から解答の候補を抽出する．以下に，各タイプ毎の解答候補の抽出方法を示す．

Type 1 質問文から解答候補の形は「XXX 時間」などのように推定されている．よって，抽出された

文から,「XXX 時間」のような形でパターンマッチングを行い，解答の候補を抽出する．

Type 2 質問文から解答がどのような固有表現であるかがわかっている．よって，3.3 で抽出された文に対して固有表現抽出を行い，質問文から推定された属性を持つ単語を解答候補として抽出する．

Type 3 質問文から解答として数字表現を求めていることがわかっている．よって，3.3 で抽出された文から，数字表現を抽出し，それらを解答候補とする．

Type 4 質問文から解答に関する情報は得られない．そのため，3.3 で抽出された文の中の名詞，未知語のうちで，tfidf 値上位 20 個を解答候補とした．

3.5 解答候補の順位付け

3.4 で抽出された解答候補に対して，どの解答候補が最も解答らしいかという点に関して順位付けを行う．提案手法では以下の様にして順位付けを行った．

1. 3.4 で抽出された文に対して係り受け解析を行う．今回は係り受け解析に CaboCha[10] を用いた．
2. 複数の文から得られた文節間の係り受け関係に従い，各文節をノードとするグラフ構造を作成する．
3. グラフ内で，質問文から抜き出された検索語を含むノードに関しては，検索語とその他に分割する．具体的には，キーワードに「発明」があり，グラフ中に「発明品」というノードがあれば,「発明 → 品」という形にする．
4. 係り受け関係から作成されたグラフは有向グラフであるが，これらをすべて無向グラフにする．このようにして作成されたグラフ構造の例を次頁の図 4 に示した．
5. ノード間のリンク数に従って，隣接するノード間のコストを定める．ここで隣接するノード A，B 間のコスト $Cost(A,B)$ は式 1 に従って定めた．

$$Cost(A,B) = 1/(N_{link(A,B)})^2 \qquad (1)$$

ただし，$N_{link(A,B)}$ はノード A，B 間のリンク数とする．

6. Dijkstra のアルゴリズム[1]に従い，解答候補と検索語の最短 距離を算出する．そして，ある解答候補とすべての検索語との距離の和を，その解答候補のスコアとし，そのスコアに従って順位付けを行った．

$$Score(Candidate) = \sum_{All\ keywords} Distance(Candidate, Keyword) \qquad (2)$$

ここで $Candidate$ は特定の解答候補，$Keyword$ は検索語を表し，ノード X，Y の最短距離 $Distance(X,Y)$ はダイクストラのアルゴリズムにより，式 3 の様に定められる．

$$Distance(X,Y) = \min \sum Cost \qquad (3)$$

[1]Dijkstra のアルゴリズムはグラフ理論の基本的なアルゴリズムであり，閉じたグラフ内のある始点から他の点までの最短距離を求めることができる．

4 評価実験とその評価

我々の研究室では，NTCIR-4 QAC2 に参加している．しかし，まだ NTCIR-4 QAC2 のデータは利用不可能であるため[2]，昨年度行われた NTCIR-3 QAC1 のデータセットを用いて，提案したシステムに対する評価を行った．以下の表 2 に実験の条件を示した．

4.1 評価方法

表 2 に示した様に，今回は Task 1 の条件に従って評価を行った．ここで，Task 1 の評価方法について簡単に述べる．

Task 1 では，システムは一つの質問に対して，順位を付けて 5 個の解答を返す．ここで，正解を返した最も上位の順位の逆数 RR をその設問の得点とする．そしてその平均値 MRR をシステムの評価とする．

$$MRR = \frac{\sum_{i=1}^{n} RR_i}{}$$
$$RR_i = \frac{1}{Rank}$$

4.2 実験結果

次頁の図 3 に，4.2.2 で分類したタイプ別の MRR，およびすべての設問での MRR を示す．また，比較のために，単純な頻度に基づく順位付けを行った結果も示す．

表 2: 評価実験の条件

知識源	毎日新聞 98 年，99 年
質問セット	NTCIR3 QAC1 Task1 200 問

[2]QAC のタスクでは，結果報告の一週間程度前に質問文のデータが公開される．

Q. テニスの全仏オープン女子シングルスで３年ぶりの優勝を果たしたのは誰ですか。

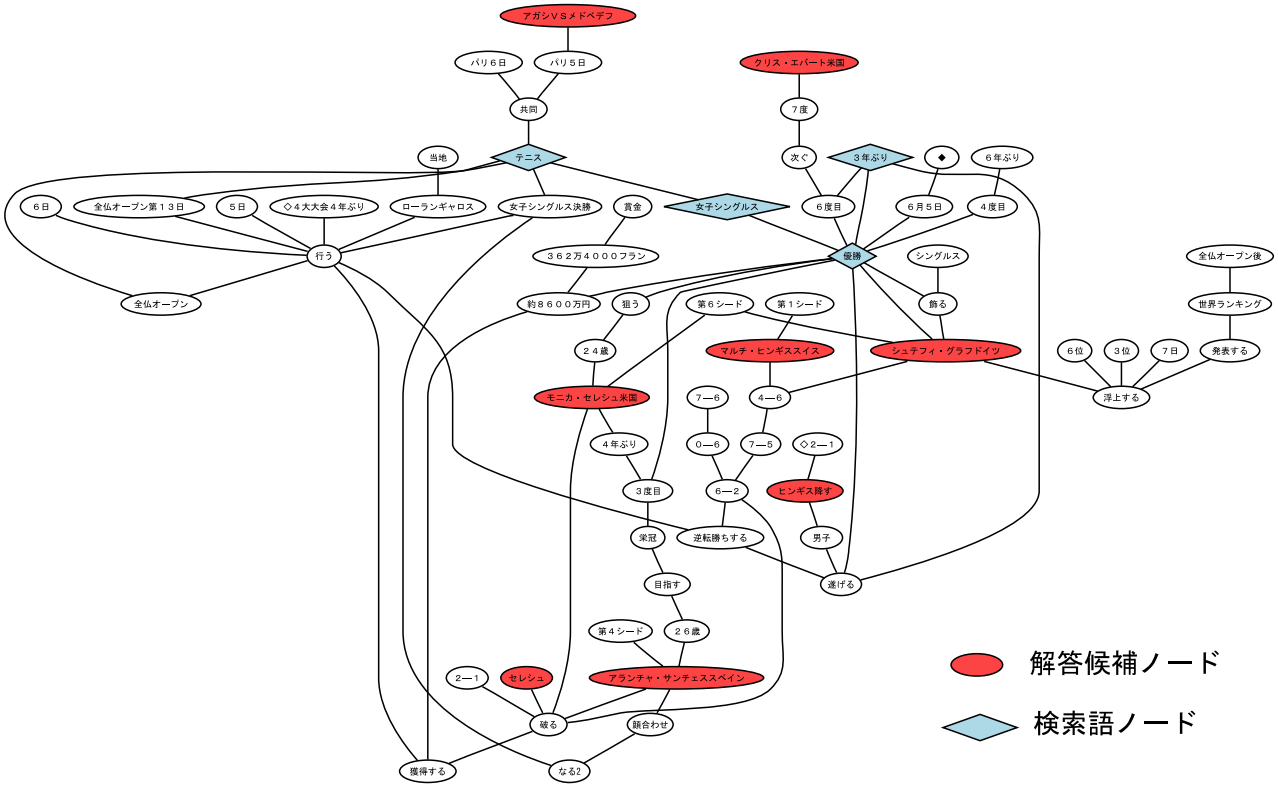

図 4: 作成されたグラフ構造の例

4.3 実験結果に対する考察

表 3 から，Type 1 に属する設問に対する MRR が非常に高いことがわかる．これは 5.2.2 で説明したように，解答の候補を抽出する時点で，設問が求める形の解答候補が少なかったことが原因と考えられる．

それに対して，Type 4 に関しては，“Namazu” による検索で抽出された文中のすべての名詞，複合名詞を解答の候補としたため，非常に低い MRR しか得ることが出来なかった．また，非常に多くの一般語が解答として提示される，という結果となった．

システム全体の評価としては，表 3 に示したように，実際に提案手法を用いることにより，単純な頻度や tfidf を用いて順位をつけた場合よりも高い MRR を得ることが出来たので，提案手法の有効性を示すことができたと考えている．さらに，NTCIR-3 QAC1 において MRR が 0.4 を越えるようなシステムはあまりなく，我々の提案するシステムの優位性を示すことができたと考えている．

また，頻度のみの場合でも NTCIR-3 QAC1 において提案されている多くの手法と比肩する MRR が得られている．これは，質問文の分類手法が適切であることを示唆していると言うことができる．

5 まとめと今後の課題

本報告では構築してた質問応答システムの機構について説明し，実際に評価実験を行い，我々のシステムの優位性を示した．以下に，グラフ構造を用いるための前提と，グラフ構造を導入することにより得られた利点を示し，最後に今後の課題を示す．

5.1 グラフ構造を導入するための前提

質問応答を実現するために，語と意味のマッピングについて考えた場合，以下の二つの場合が問題となる．

1. 一つの意味に多数の単語がマッピングされている場合
2. 多くの意味が一つの単語にマッピングされている

表 3: 評価実験の結果

	Type 1	Type 2	Type 3	Type 4	Total
分類された質問の数	22	94	16	68	200
提案手法で得られた MRR	*0.635*	*0.526*	*0.625*	*0.205*	*0.427*
頻度のみを用いて得られた MRR	0.543	0.350	0.250	0.223	0.305

場合

（1）は類義語，同義語の問題.（2）はストップワード，多義語の問題，ということができる．

（1）に関してはシソーラスを用いることが解決策として挙げられる．しかし，シソーラスを検索で用いることは容易であるが，グラフ構造の中で用いることは難しい．

（2）に関して，ストップワードは一つの単語に多くの意味がマッピングされている典型例ということができる[3]．そのため，グラフのなかで各々のストップワードが一つのノードにまとまってしまわないようにしなければならない．その他の多義語に関しては，検索を行った時点で一つの閉じたドメインが構成されている，と考えると，この閉じたドメインでは，多義語の多義の中の一つの意味に，その多義語がマッピングされている，と考えることができる．この前提によりグラフ構造を導入することができる，ということができる．

5.2 グラフ構造を用いることの利点

質問応答を行うにあたり，問題となる点がいくつかあるが，提案手法においてグラフ構造を用いることにより，これらの問題をどのように吸収しているかを以下に示す．

5.2.1 言い換え表現の吸収

質問文と知識源となる文書で同じことを異なる表現で行っている場合がある．提案手法では，グラフ構造の縮退により，以下の様な言い換え表現を吸収できる．

- エジソンはフィラメントランプを発明した．
- フィラメントランプはエジソンの発明品だ．

また有向グラフを無向グラフにすることで，以下のような能動態，受動態の言い換えを吸収できる．

- エジソンはフィラメントランプを発明した．
- フィラメントランプはエジソンによって発明された．

[3]厳密に言うとストップワードに与えられているのは意味ではなく機能である

5.2.2 指示代名詞

知識源となる文書がいくつかに別れている場合，指示代名詞が問題となる．グラフ構造でリンクを張ることにより，これらの問題を解消できる場合がある．ただし，厳密な対応を行う場合には他の処理を加える必要がある．

参考文献

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

[2] *NTCIR*. http://research.nii.ac.jp/ntcir/.

[3] 倉田岳人，岡崎直観，石塚満：係り受け関係に基づくグラフ構造を用いた質問応答システム，電子情報通信学会，信学技報「自然言語処理」No.158-011, 2003-11.

[4] M.Collins N.Duffy: “Convolution kerneks for natural language”, *Neural Information Processing Systems*, 2001.

[5] INUI Kentaro TAKAHASHI Tetsuro, NAWATA Kozo: “Applying Structural Mathcing and Paraphrasing”, *Proceedings of the Third NTCIR Workshop*, 2003.

[6] Gary Guenbae Lee Seungwoo Lee: “SiteQ/J: A Question Answering System for Japanese”, *Proceedings of the Third NTCIR Workshop*, 2003.

[7] NIWA Tatsuhiro FUKUMOTO Jun’ich, ENDO Tetsuya: “RitsQA:Ritsumeikan question answering system used for QAC-1”, *Proceedings of the Third NTCIR Workshop*, 2003.

[8] 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: “SAIQA:大量文書に基づく質問応答システム”, 情報学基礎研究会, No.064-12, 2001.

[9] ISAHARA Hitoshi MURATA Masaki, UTIYAMA Masao: “A Question-Answering System Using Unit Estimation and Probablistic Near-Terms IR”, *Proceedings of the Third NTCIR Workshop*, 2003.

[10] *“CaboCha”*. http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/.

[11] Robert S. Taylor: “question-negotiation and information seeking in libraries”, *College & Research Libraries 1968*, pp. 178–194, 1968.