質問応答における探索制御付き命題照合の精度向上

太田 知宏[†] 藤畑 勝之[†] 公文 隆太郎[†] 森 辰則[††]
[†] 横浜国立大学 大学院 環境情報学府 [††] 横浜国立大学 大学院 環境情報研究院
E-mail: {tomo_o,fujihata,kumon,mori}@forest.eis.ynu.ac.jp

1 はじめに

近年、世界規模のネットワーク拡大に伴う情報過負荷への対策として、質問応答システムが注目されている。質問応答システムは、ユーザーから自然言語で与えられた質問の回答を返すシステムである。現在提案されている多くのシステムでは、あらかじめ文書集合全体に対して形態素解析や固有表現抽出等の計算コストの高い処理を行ない、結果を保存している [PRA00]。ところがこの方法は、新聞記事のように固定化された文書群を対象とした場合には有効であるが、WWWのように日々刻々と変化する膨大な文書群にその場で対応するには、システムに非常に大きな計算能力が要求される。

我々は、このような前処理を行なわずに、かつ計算コストを大幅に削減する探索制御付き命題照合を提案している [志賀 02]。そこでは，構文解析、固有表現抽出などコストの大きいいくつかの処理に注目し、これらの処理に対して探索制御を適用することで、解の精度を落さずに計算コストを削減することを実現している。ただし、このシステムで採用した照合の各項目は基本的なもので、絶対的な解の精度はそれほど秀でたものではなかった。

そこで本稿では、上記枠組に対し、より高い解の精度を実現するために、キーワード重要度、ベクトル表現による係り受け照合、複数文を考慮した命題照合などの手法を導入する。そして、このように精度向上を目的とする複雑な照合を導入した命題照合においても、探索制御が有効に機能することを実証する。

2 提案システムの概略

提案システムの全体構成および動作例を図1に示す。本システムは主に4つのモジュールから構成されており、質問文解析モジュール、文書検索モジュール、パッセージ検索モジュール、そして命題照合モジュールに分類される。

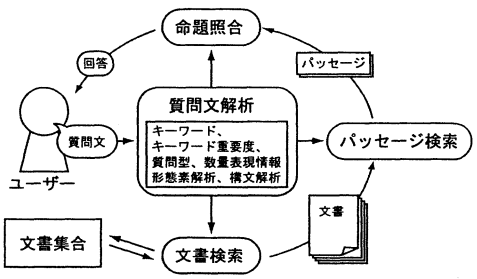


図 1: 質問応答システムの概要

2.1 質問文解析

2.1.1 解析の内容

ユーザーが入力した質問文から質問応答に有用な情報を抽出するのが質問文解析の役割である。解析により得られる情報を以下に示す。

- 形態素解析および構文解析の結果
  始めに質問文が形態素解析、構文解析され、その結果を利用して後続の処理がなされる。
- キーワードおよびその品詞
  文書検索やパッセージ検索、命題照合に用いるキーワードを質問文から抽出する。
- キーワード重要度 (節 2.1.2 参照)
- 質問型
  質問の求める内容がどのような固有表現あるいは数量表現に該当するかを表1のように判定する。
- 数量表現抽出
  質問型が数量に該当するとき、質問文の疑問詞に対応する「もの」と「属性」の関係を数量表現抽出器 [藤畑 01] によって抽出する。

表 1: 質問型の例

種別	質問型	
固有表現	人名	PERSON
	地名	LOCATION
	組織名	ORGANIZATION
数値 (数量表現抽出非使用)	日付	date
	時間	time
	割合	rate
数量 (数量表現抽出使用)	長さ	length
	金額	money
	数量一般	vol

2.1.2 キーワードの重要度計算

質問文から得られたキーワードには、質問の解に直結する重要なものが含まれる一方、それほど重要でないものもある。そこで、以下の重み W_{Kn} をキーワード Kn の重要度として導入する。

1. n 個目のキーワード Kn を除き、残りのキーワードのみを使って文書検索を行なう (50文書程度)。
2. 得られた文書集合全体における Kn の出現頻度を基に次式により重要度 W_{Kn} を計算する。

$$W_{Kn} = \frac{1}{\left(\frac{Kn\text{の出現頻度}}{\text{検索文書数}} + 1\right)} \quad (1)$$

3. すべての Kn について W_{Kn} を計算する。

この重要度は、そのキーワードを除いて文書検索したときに出現頻度の低いものほど大きな値となる。この計算手法は検索結果文書のみを用いるので、第三者の検索エンジンを利用する場合でも利用可能であり、ドメインにも依存しない。

2.2 文書検索

質問文中のキーワードにより文書検索を行う。評価実験においては、TFIDF による語の重みづけとベクトル空間法による類似度尺度を用いた基本的なエンジンを利用した。

2.3 パッセージ検索

文書検索で得られた関連文書の中でも、質問の解となる情報が書かれているのはその一部だけである。パッセージ検索は文書集合から正解を含む可能性の高い文書部分 (パッセージ) を取り出すために行う。キーワードの出現種類数とそのキーワード重要度などからパッセージ毎にスコアが計算され、スコア大きいパッセージが次の命題照合に渡される。以下では抽出されたパッセージ中の各文を検索文と呼ぶ。

3 命題照合

命題照合は、パッセージ検索までの処理で得られたパッセージから、質問の解を抽出する処理である。本システムの命題照合は 2-gram 照合、キーワード照合、係り受け照合、質問型照合の 4 種の照合からなる。各照合において照合の一致の度合に応じて検索文の文字または形態素 (解候補) に部分スコアが与えられ、全ての部分スコアの和がその解候補のスコアとなる。そしてスコアの最も高い解候補から解が生成される。ただし、実際は探索制御が行なわれるため、全ての解候補について順番にスコアを求めなくとも、最もスコアの高い解候補を決定することができる。

3.1 複数文の取り扱い

我々の命題照合は基本的に 1 つの質問文と 1 つの検索文の間で照合を行ない、スコアを計算する。ところが、質問文の内容が検索文の 2 文以上にわかれて出現する場合には、1 文対 1 文の照合では正解を導くことが難しい。そこで、複数文を結合し、仮想的に 1 文であるとみなして質問文との照合を行なう手法を提案する。パッセージの $n(n>0)$ 文目の検索文 L_n と質問文との照合を行なうときに、L_n が次の 2 つの条件を共に満たすならば、L_0 から L_n までを結合する。

1. L_0 から L_{n-1} までの間に、L_n で現れないキーワードが出現している。
2. L_n に、L_0 から L_{n-1} までの間で現れないキーワードが出現している。

3.2 2-gram 照合

質問文から得られる 2-gram が検索文にどれだけ現れるかによりスコアを与える。キーワードの連接や用言の活用形など、キーワードに現れない情報もこれによりに抽出する。

3.3 キーワード照合

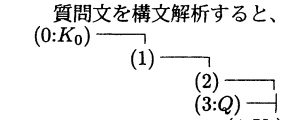
「質問文キーワード」および「質問文キーワード+助詞」が検索文にどれだけ現れるかによりスコアを与える。出現しているキーワードの重要度が高いほど大きなスコアを与える。また、作品の題名や名称の略称を求める質問では、解が検索文中で「」や () に囲まれて出現することが多いので、括弧中にある解候補に対しては高いスコアを与える。キーワード照合のスコアは検索文を形態素解析したのちに得られる。

3.4 係り受け照合

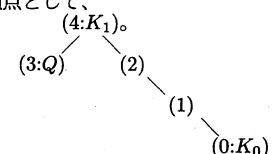
質問文における疑問詞とキーワードの間の係り受け関係と、検索文における解候補とキーワードの間の係り受け関係を照合する。ここで我々は、文の係り受け構造をベクトルで表現し質問文と検索文を照合する手法と、質問文の内容が複数の検索文に分かれて現れたときに複数文を結合する手法を提案する。このスコアは検索文を構文解析したのち得られる。

3.4.1 係り受け関係のベクトル表現と照合スコアの計算

質問文を構文解析すると、



のような係り受け構造が得られる。この構文木は、文末を頂点として、



と見ることができる。この木において、疑問詞 Q からキーワード Kn への最短経路をたどるときに、構文木の共通祖先へ向かう距離を Qd_n、共通祖先からキーワードへ向かう距離を Qu_n とすると、Q と Kn の係り受け距離は二次元ベクトル (Qu_n, Qd_n) で表現できる。上の例では、$(Qu_0, Qd_0) = (3,1), (Qu_1, Qd_1) = (0,1)$ となる。同様に検索文においても、解候補 x とキーワード Kn の係り受け距離の関係を $(Au_n(x), Ad_n(x))$ で表せる。

ところで、質問文における疑問詞 Q と検索文における解候補 x を、キーワード Kn との係り受けの距離の関係によって照合するときに、経験的に次のことを言うことができる。

- ベクトル (Qu_n, Qd_n) の長さと、ベクトル $(Au_n(x), Ad_n(x))$ の長さがともに小さいほど Q と x は照合できる可能性が高い。
- ベクトル (Qu_n, Qd_n) が指す座標と、ベクトル $(Au_n(x), Ad_n(x))$ が指す座標が近いほど Q と x は照合できる可能性が高い。

前者は双方のベクトルの絶対値の和を求めることによって、後者は双方のベクトルの差の絶対値を求めることによって、計算することができる。係り受け照合のスコアは、この 2 つに基づく値と、構文解析で得られる格の情報の一致に与えられる値の和で表される。このスコアも、検索文内に出現するキーワードの重要度が高いほど大きな値となる。

3.4.2 複数文を考慮した係り受け照合

検索文が節 3.1 で説明した条件を満たすときには、複数文を考慮した係り受けの照合によって係り受けのスコアが計算される。質問文と複数の文とを照合するには複数の文の構文木を何らかの方法で 1 つに結合しなければならない。前文は次文の主題に係ると考えるのが妥当であり、提題の文節はその文の主題となる語を含んでいると考えられることから、前文の最後の文節を次文の提題の文節に係り受けさせ、構文木を結合する。ただし、文が提題の文節を含まないこともあるので、その場合は前文の最後の文節を次文の最初の文節に係り受けさせる。

3.5 質問型照合

解候補が質問文から得られた質問型に一致するかどうかを照合し、スコアを与える。表 1 に示されている質問型の種類により、固有表現抽出または数量表現抽出の処理が行なわれる。

3.5.1 固有表現抽出

質問型が PERSON,LOCATION または ORGANIZATION のとき、固有表現抽出器 [山田 01] によって検索文の各解候補の固有表現種が判定され、質問型と一致すればスコア 1.0 が与えられる。固有表現種が一致しなかったときも、その解候補の性質 (表層や品詞) によりスコアが与えられる。例えば、形態素解析の品詞細分類と質問型が一致すれば 0.6、カタカナ (固有名になりやすい) には 0.5 が与えられる。質問型の照合スコアはこのポイントに比例して与えられる。

3.5.2 数量表現抽出

質問型が length,money などの数量表現のときには、解候補の表層 (数値＋単位) が質問型に適合するかをパタンマッチで判断する。適合した解候補は、数量表現抽出器 [藤畑 01] によって解候補に対応する「もの」と「属性」の組が抽出され、質問文の「もの」「属性」と照合される。質問型の照合スコアは「もの」「属性」の一致文字数の割合に基づき与えられる。

質問型が date,rate など数量表現抽出器が対応しない数値情報のときには、表層のパタンマッチのみで照合スコアが計算される。

3.6 解生成

解候補の最終スコアは、2gram, キーワード、係り受け、質問型の各照合によって与えられたスコアの和で表され、最終スコアの高い解候補とその周辺の形態素から解が作成される。

3.7 A*アルゴリズムに基づく探索制御

本システムでは命題照合に A*アルゴリズムに基づく探索制御が導入されている。この探索制御では、上記の各照合の前後において各解候補が最終的に得られるスコアの値を推定することにより、最適解 (最終スコアが最も大きい解) である可能性の高い解候補から先の処理を進める。スコアの推定方法は [志賀 02] で述べた手法とほぼ同等だが、係り受け照合の近似スコアについてのみ、文節の区切りと文節の格だけを推定する、より確実な方法に変更されている。

4 評価実験

提案システムの有効性を検証するために、評価実験を行なった。

4.1 実験方法とシステム

NTCIR Workshop 3 QAC1[NTC01] で提供された AdditionalRun の質問文から 200 問を無作為に抽出し、評価用のテストセットとした。システムの調整には FormalRun の質問文 200 問を利用している。タスク定義は、回答の出力件数を除き QAC1 のタスク 1 に準ずる。なお、文書検索の出力数は 250 文書、パッセージ検索の出力数は 10 パッセージ (30 文) とした。

4.1.1 実験 1 照合方法による精度の比較

次の各システムの精度を比較し、提案手法による命題照合の有効性を検証した。最大・近似両スコアによる A*探索制御を採用し、出力する回答数は 5 件とした。

1. 3 章で述べた照合方法すべてを採用したシステム
2. 複数文照合の対応を除いたシステム
3. キーワード重要度を利用しないシステム
4. 係り受け照合で、和の値を利用しないシステム
5. 係り受け照合で、差の値を利用しないシステム
6. 提案手法による係り受け照合を用いず、代わりに解候補とキーワードとの語の距離に基づいてスコアを与えるシステム

システム (6) は、構文解析を用いずに解候補のスコアを決定する素朴な手法である。この手法では、全てのキーワードからの距離が近い解候補ほど大きなスコアが与えられる。なお、このシステムに限り最大スコアのみによる探索制御を採用した。

4.1.2 実験 2 探索制御の有無による処理時間と正解性能の比較

次の各システムの正解性能および命題照合の処理時間比較し、提案手法による命題照合を導入したときの探索制御の有効性を検証した。各システムでテストセットを 3 回ずつ回答させ、1 問あたりの平均処理時間を計測した。出力する回答数は 2 件とした。

1. 探索制御無しのシステム
2. 最大スコアのみによる A*探索制御付きシステム
3. 最大・近似両スコアによる A*探索制御付きシステム

(2) は A*探索制御における推定スコアとして各照合で得られうるスコアの最大値のみを採用したシステム、(3) は最大スコアに加えて、その時点で得られる情報から計算される近似スコアを採用したシステムである。

4.1.3 実験環境

実験システムは，Xeon 2.2GHz 2CPU の Linux システム上に Perl5 を用いて実装されている。また、形態素解析器として JUMAN 3.61、構文解析器として KNP 2.0b6、固有表現抽出器として SVM を用いたシステム [山田 01]、数量表現抽出器として係り受けの制約と優先規則に基づくシステム [藤畑 01]、知識源として 1998 年、1999 年の毎日新聞記事がそれぞれ採用されている。

4.2 結果と考察

実験 1、実験 2 の結果をそれぞれ図 2、図 3 に示す。

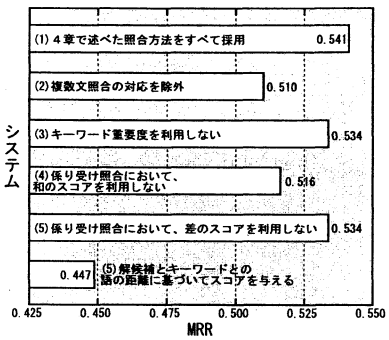


図 2: 照合方法による正解性能の比較

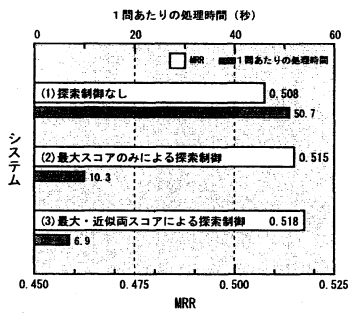


図 3: 探索制御の有無による処理時間と正解性能の比較

実験 1 では、MRR の値において、提案手法をすべて採用したシステムが他のシステムを上回った。特に複数文の対応を除いたシステムとは明らかな差が生じた。一方、キーワード重要度の有無ではそれほど大きな差にならなかった。システムの出力を確認したところ、確かに重要なキーワードに高い重要度が与えられてはいるのだが、検索文では重要度の高いキーワードの代わりにその類義語が出現していたり、キーワードが省略されていたりして、必ずしも重要度の高いキーワードが正解と同じ文に現れるとは限らなかった。係り受け照合の方法では、和の値を利用しないシステム、差の値を利用しないシステム、解候補とキーワードの距離に基づくシステムのいずれに対しても、両方を使用するシステムの結果が上回った。これにより、提案手法の係り受け照合が優れていることが示された。

実験 2 では、最大・近似両スコアによる探索制御付きシステムが、探索制御をしないシステムの約 7.3 倍、最大スコアのみによる探索制御付きシステムの約 1.5 倍の処理速度が得られた。最大・近似両スコアによる探索制御付きシステムの MRR が他のシステムより若干良い結果となったが、これは最終スコアが全く同じになった解同士で順位が入れ替わってしまったために起こった。本質的には、提案システムに探索制御を導入しても精度が全く落ちないことが示された。

4.3 参考記録

提案システムで QAC1 FormalRun の質問 200 問を回答させたときの MRR は 0.525 だった。この値は、FormalRun 参加 15 システム中 2 位相当の成績である。ただし、提案システムは FormalRun の質問の処理結果を元に調整がなされているので、この値は厳密な比較とはならない参考記録である。しかし、未知の質問 (AdditionalRun 200 問) についても同程度の精度を示すことから、同程度の難易度の質問については、この精度で安定して解が得られると思われる。

5 まとめ

本稿では、精度の良い命題照合を実現するために、キーワード重要度の導入や構文構造をベクトルで表現する係り受け照合、命題が複数の検索文に分かれる際に複数文を結合して照合する方法などを提案した。評価実験の結果、提案システムが精度、計算コストともに優れていることが示された。

さらに精度の良い質問応答を実現するためには、詳細な質問型判定の導入や、キーワード重要度を有効に機能させるための照応解析や同義語、類義語への対応、表や箇条書に特化した照合の導入などが課題となる。これらの改善方法の導入により命題照合の精度向上が見込まれるが、それに伴って計算コストも増加するだろう。本提案システムの A*探索制御による命題照合は、これら複雑な照合を導入した際の計算コストの増加も最小限に押えることができる。

参考文献

[志賀 02] 志賀 正裕, 太田 知宏, 藤畑 勝之, 公文 隆太郎, 森 辰則. 実時間質問応答のための探索制御付き命題照合. 言語処理学会第 8 回年次大会発表論文集, 3 月 2002.

[NTC01] NTCIR3 QAC-1 http://research.nii.ac.jp/ntcir/workshop/qac/cfp-ja.html, 2001.

[PRA00] John Pager, Eric Brown, Anni Coden, Dragomir Radev. Question-answering by predictive annotation. In Proceedings of SIGIR2000: 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[藤畑 01] 藤畑 勝之, 志賀 正裕, 森 辰則. 係り受けの制約と優先規則に基づく数量表現抽出. 情報処理学会自然言語処理研究会研究報告 2001-NL-145-18, 情報処理学会,9 月 2001.

[山田 01] 山田寛康, 工藤拓, 松本裕治. Support vector machines を用いた日本語固有表現抽出. 情報処理学会研究報告 01-NL-142-17, 情報処理学会, 3 月 2001.