3C4-08

質問応答における回答絞り込み手法の比較

A Comparison of Answer Ranking Methods in Question Answering

日高 直哉* 桝井 文人*
Naoya HIDAKA Fumito MASUI

* 三重大学工学部
Faculty of Engineering, Mie University

This paper describes a comparison between two answer ranking methods in Question Answering (QA). One is an analysis-based method. The other is a statistic-based method. In the analysis-based method, score based on similarity of syntax information and structural distance ranks answer candidates. The similarity of syntax information is between a question sentence and extracted sentences including important words from a question sentence. The structural distance is between answer candidates and the important words. In the statistic-based method, answer candidates are ranked by score based on TF·IDF of answer candidates. Experiments that evaluate these two methods were conducted with 103 questions of QAC1. The results of experiments showed that the analysis-based method is effective to detect only one answer and that the statistic-based method has lower time cost. From the results, a hybrid method having advantages of both methods is examined.

1. はじめに

近年，質問応答技術が注目されてきている．質問応答とは，自然言語で記述された質問文に対して，組織されていないテキスト集合から適した答えを捜し出す技術である．本論文で述べる質問応答システムは，質問文解析，文書検索，回答特定の 3 つの技術から構成される．質問文解析部では，質問文を解析し，重要語と回答タイプを抽出し，文書検索部で，重要語をもとに対象文書群から，適切な文書の検索をする．回答特定部では，検索された文書の中から回答タイプに基づく回答候補から回答を絞り込み，特定する．

質問応答における回答絞り込み手法は，主として解析に基づく手法（解析ベース）と統計に基づく手法（統計ベース）がある．解析ベースにおいて，村田ら[1]は，係り受け関係を用いて，対象文書中から抽出された文と質問文との類似度を定量化する手法を提案している．宮口ら[2]は，質問文と構文構造の類似性と，回答候補からの重要語の係り受け階層距離を用いて，回答を絞り込む手法を提案している．統計ベースにおいて，李ら[3]は，意味的パターンを用いて回答候補を抽出し，回答候補と意味的パターンに一致した語との距離にスコアを与え，回答を絞り込む手法を提案している．

しかしながら，両手法の比較については，詳細な議論がなされていない．そこで本論文では，同条件下で，質問応答の回答絞り込みにおける、解析ベースによる手法と統計ベースによる手法の精度の比較実験を行う．解析ベースは，質問文の構文構造の類似性と質問文中の重要語との係り受け階層距離を考慮した回答の絞り込み手法を用いる．統計ベースは，TF･IDF に基づいた回答の絞り込み手法を用いる．比較により，各手法の特徴，優位性を明らかにし，両手法の長所を生かしたより精度の高い回答絞り込み手法について吟味する．

以下，2 章で解析ベースによる回答絞り込み手法，3 章で統計ベースによる回答絞り込み手法，4 章で両手法の実験と結果を述べ，5 章で考察を行う．

連絡先：日高直哉，三重大学工学部，〒514-8507 三重県津市上浜町 1515，naoya@ai.info.mie-u.ac.jp

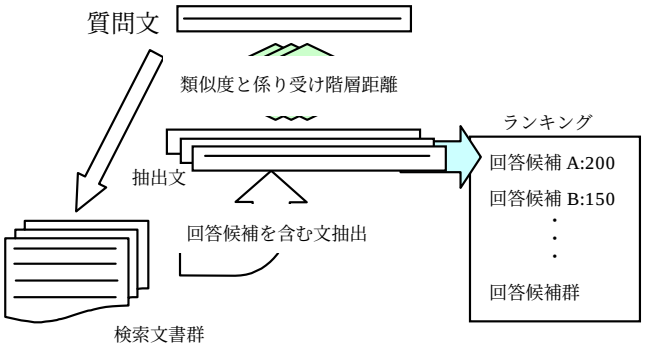

図 1 解析ベースによる回答絞込み手法

2. 解析ベースによる回答絞り込み手法

本章では，質問応答に解析ベースを用いた回答絞り込みについて説明する．本論文では，解析ベースによる回答絞り込み手法として，宮口らが提案した手法を利用する[2]．本章で示す回答候補とは，質問文中の疑問詞部分または文末表現から既定される回答の種別である．例えば「～は誰ですか」という質問文ならば，回答候補は，『人名』と判断される。

宮口らの手法は，質問に類似した文は，質問に対する回答をも含む可能性が高いという考えに基づいている．質問文から抽出された重要語（固有表現，サ変名詞･形容動詞語幹，一般名詞）に基づいて絞り込まれた文書群中から，質問文から得られた回答タイプに一致する重要語（回答候補）を含む文を抽出文として取り出す．さらに，(1) 質問文との類似度および，(2) 回答候補と他の重要語との係り受け階層距離に基づいてスコア付けを行い，スコアのランキングによって回答を絞り込む（図 1）．スコアは式 1 によって計算される．

$$Score(c) = \sum \{Sim(q) + Near(q)\} + \alpha \qquad (1)$$

ここで，c はスコアを求める回答候補を示し，q は質問文中の重要語を示し，Sim(q)は抽出文と質問文の q に基づく類似度の

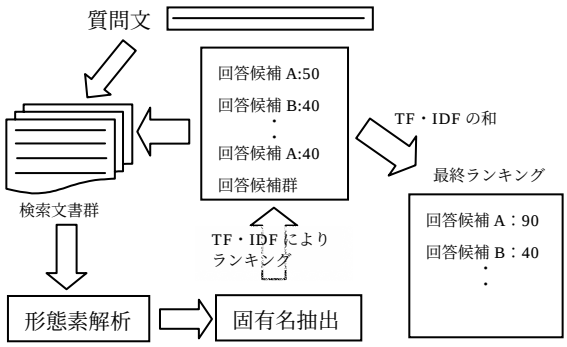

図 2　統計ベースによる回答絞り込み手法

スコアを示し、Near(q)は回答候補と q の係り受け階層距離に基づくスコアを示す．α は，回答候補が固有表現である場合に，与えるスコアであり，正解回答は，固有表現が多いという判断に基づいている．

3. 統計ベースによる回答絞り込み手法

本章では，質問応答に統計ベースを用いた回答絞り込みについて説明する．本章で示す回答候補も，2 章で示したものと同様のものとする．また正解回答は，固有表現の場合が多いと判断し，回答候補が固有表現の時にはスコア α を与える．

ここで提案する手法は，検索文書中で，その文書を特徴付ける語句が正解である可能性が高いという考えに基づいている．文書を特徴付ける語句を判断する要素として，TF･IDFを用いる．TFは文書中に出現する単語の頻度を示す．IDFは，対象文書中のどのくらいの文書に出現するかを表す尺度で，式 2 は次のようになる．

$$IDF = \log \frac{\text{対象文書の総数}}{\text{出現文書数}} + 1 \qquad (2)$$

IDF は出現文書が少ない語ほど値は高くなる．よって TF･IDF の値が高い語句は，その文書には多く出現するが，他の文書にはあまり出現しないような語句であるので，その文書を特徴付ける語であるといえる．

本手法を具体的に示すと，はじめに，質問文に対して，検索された複数の文書から 1 文書を抽出する。次に，抽出された文書に対して，形態素解析を行い，文書中の語句を抽出する．しかし形態素解析だけでは，文書中の固有表現はうまく抽出されず，例えば『1968年』が『1』『9』『6』『8』『年』というように，固有表現が分かれて抽出されてしまう．このままでは，正解回答が固有表現の場合には，正解回答を絞り込むことができない．そこで，固有表現をひとつの語句として抽出するために，固有表現抽出を行う。これにより，分かれてしまった固有表現が，元のひとつの語句として抽出される．よって，形態素解析と固有表現抽出を行うことで，文書中の語句が，適切に品詞情報とともに抽出される．

質問応答において，正解回答となるのは名詞もしくは固有表現であるので，検索文書中から，回答候補となる名詞と固有表現のみを抽出する．そして，抽出された回答候補を TF･IDF の値の高い順にランキングし，上位 10 位以内を抽出する。このとき，それぞれの TF･IDF を各回答候補のスコアとする。同様の処理を検索された全文書に対して行う．

最後に，各検索文書からランキングされた回答候補に対し，同じ回答候補がある場合には統合し，スコアの和の高い順に回答候補をランキングし，上位 5 位以内を抽出し，最終的な回答とする．このとき，品詞は異なるが同じ回答候補は，同一文書中に出現する場合は，意味に相違がないと考えられるので統合した．(図 2)

4. 実験と比較結果

4.1 実験環境

上記で説明した手法を用いて 2 種類の回答絞り込み手法を実装し，実験を行い結果の比較を行った．今回の実験で用いた質問文は，NTCIR3 QACタスク[4]の formalrun 用の質問文を用いた．回答絞り込みの性能のみを比較するため，検索文書部において検索に失敗した質問文は除外した．その結果，200 文のうち 103 文となった．対象とした文書群は毎日新聞の 1998 年 1999 年の 2 年分である．各質問文に対し最大 5 件の文書を検索し，その結果から回答絞り込み処理を行った．また，本実験では形態素解析器として Chasen ver 2.0 を用い，固有名抽出器として NExT ver 0.63，構文解析器として KNP ver 2.0 を用いた．

評価では，質問応答システムに出力される回答候補群の上位 5 位以内における正解の有無で正解･不正解を判断した．

4.2 実験結果

実験の結果を表 1，表 2，表 3，表 4 に示す．

表 1 と表 2 は，解析ベースと統計ベースの両手法の実験結果に対して，評価を行った結果である．採点に用いられる要素を示すと，Question は実験に用いた質問数，Answer は Question における全正解数，Output はシステムが出力した回答数，Correct はシステムが正解回答を絞り込んだ数(1 つの質問文で 2 つ異なる正解回答を絞り込んだ場合は 2 とする)，Recall，Precision，F-value，MRR はそれぞれ以下の式 3〜6 を用いて計算できる．

$$\mathrm{Re}\,call = Correct / Answer \qquad (3)$$

$$\mathrm{Pr}\,ecision = Correct / Output \qquad (4)$$

$$F\text{—}value = \frac{2 * \mathrm{Re}\,call * \mathrm{Pr}\,ecision}{\mathrm{Re}\,call + \mathrm{Pr}\,ecision} \qquad (5)$$

表 1　両手法の評価 (1)

	Question	Answer	Output	Correct
解析	200	281	758	56
統計	200	281	758	52

表 2　両手法の評価 (2)

	Recall	Precision	F-value	MRR
解析	0.199	0.074	0.108	0.182
統計	0.185	0.069	0.100	0.144

表 3　各手法による正解数

	両手法とも	解析のみ	統計のみ	合計
正解数	33	18	15	66

表 4　各手法の順位別正解数

	1 位	2 位	3 位	4 位	5 位
解析	20	6	3	3	1
統計	13	3	5	3	2

$$MRR = \sum 順位の逆数 / Question \quad (6)$$

Recall は, 正解のカバー率を示し, Precision はシステムの精度を示す. F-value は Recall と Precision をひとまとめにしたものである. MRR は順位の逆数をスコアとしたもので, 正解回答が上位で絞られてくるほど, MRR は高くなる.

表 2 は、評価の対象となった質問文 103 文について, 各手法における正解数(1 質問文に対して 1 つ)を示し,「両手法とも」とは解析ベースと統計ベースの両手法ともに正解回答が絞り込めたものを示し, その正解数は 33 個(50%)あり,「解析のみ」は解析ベースのみで正解回答が絞り込めたものを示し, 正解数は 18 個(27%),「統計のみ」は統計ベースでのみ正解回答が絞り込めたものを示し,その正解数は 15 個(23%)であった. 括弧内の%は正解回答が絞り込めた総数に対する割合である。

表 3 からは, 両手法ともに正解回答を絞り込めた質問文 33 文に対して、解析ベースでは, 1 位で正解を絞り込んだものが 20 個, 2 位で絞り込んだものが 6 個, 3 位で絞り込んだものが 3 個, 4 位で絞り込んだものが 3 個, 5 位で絞り込んだものが 1 個あることが分かり, 統計ベースでは, 1 位で正解を絞り込んだものが 13 個, 2 位で絞り込んだものが 3 個, 3 位で絞り込んだものが 5 個, 4 位で絞り込んだものが 3 個, 5 位で絞り込んだものが 2 個あることが分かる.

5. 考察

本章では, 実験から得られた比較結果について考察する.

表 1 において, 解析ベース, 統計ベースの両手法の Recall, Precision, F-value に大きな差は見られないが, MRR は解析ベースのほうが統計ベースの手法よりも高くなっている. このことから, 解析ベースと統計ベースにおいて, 上位 5 位以内に正解回答を絞り込む数に大きな差はないが, 解析ベースのほうがより上位に正解回答を絞り込むことができたと言える. しかし, 表 1 は, 文書検索に失敗した質問文も含めた質問文 200 文に対する結果であるため, 解析ベース, 統計ベースの性能以外の要因による影響も含まれる. よって, 表 3, 表 4 の結果のほうが, 解析ベースと統計ベースの手法による性能の違いが明らかにできる.

表 3 から, 各手法により正解回答を絞り込めた質問文に違いが生じた理由として、検索文書中における、正解回答の出現のしかたの違いが考えられる.

解析ベースのみで正解回答を絞り込めた質問文について, 検索された文書を調べたところ, ①1 文書中に正解回答の話題以外にも他の話題が出現している文書や, ②複数の検索文書中で正解回答の話題が出現している文書の他に別の話題が出現している文書が検索されていた. このような文書が検索された場合, 回答候補が多くなり, 正解回答のTFも小さくなってしまうために, 統計ベースでは回答として絞り込むことが難しい. しかし, 解析ベースの場合, TF は関係なく, 抽出文と質問文の類似度と質問文中の重要語との距離に基づいて回答を絞り込むために, 検索文書中に正解回答とは関係のない話題が出現したとしても, 抽出文の類似度が低く, 重要語との距離が離れていると考えられるので, そこに出現する回答候補がランキングに上位にくる可能性は低い. よって解析ベースでは、正確な抽出が可能である. 検索文書において, 各回答候補に十分なTFが得られないような文書に対しては, 解析ベースが効果的である.

統計ベースのみで正解回答を絞り込めた質問文について, 検索された文書を調べたところ, 解析ベースにおいて抽出文中に正解回答と質問文中の重要語が出現しない文書や, 重要語と正解回答との多くの単語が出現する文書が検索されていた. 解析ベースでは, 抽出文中に正解回答と質問文中の重要語が出現しない場合, 類似度のスコアが低くなり, 正解回答を絞り込むことは難しい. また重要語と正解回答との多くの単語が出現する場合, 係り受け階層距離が離れスコアが低くなるので, 正解回答を絞り込むことは難しい. しかし, 統計ベースの場合, TF･IDF のみで回答を絞り込むために, 正解回答が検索文書中の, 話題の中心となるような語句であれば, 正解回答を絞り込むことができる.

表 4 より, 解析ベースの方が, 正解回答を 1 位で絞り込む性能が高いことが分かる. 解析ベースでは, 検索文書中に, 質問文との類似度が高い文が存在する場合, 正解回答を 1 位で絞り込むためであると考えられる. 統計ベースでは, TF･IDF でのみ回答を絞り込むため、関係のない語句が回答候補となり, ノイズとして紛れ込むことがある. そのため, 質問文とよく類似した文が文中に見つかった場合には, 解析ベースの手法を用いたほうが効果的であるといえる. 逆に, 質問文と類似した文が見つからなかった場合や, 解析ベースを用いて, 回答候補のスコアに大きな差が見られない場合には, 統計ベースによる絞り込みが有効であると考えられる. また, 質問文 1 問当たりに回答を絞り込むのに要する時間は, 解析ベースが約 3 分かかるのに対し, 統計ベースでは約 20 秒である. よって統計ベースにおいて回答候補のTF･IDFが十分に高い値を示すときには, 統計ベースを用いた方が時間的コストは少ないのでより実用的な質問応答システムになると考えられる.

6. おわりに

本論文では, 同条件下で, 質問応答における解析ベースと統計ベースの精度の比較実験を行った. 実験の結果, 両手法の特徴および優位性を示すことができた. 解析ベースによる手法では, 回答を 1 つに絞り込む精度が高いことが判明した. また, 複数の話題について書かれている文章から, 回答を見つけ出す場合にも優位性が見られた. 一方, 統計ベースによる手法では, 質問文に類似した文が検索文章中に見つからない場合でも, 回答を絞り込めることがわかった. 更に, 解析ベースによる手法に比べ, 時間的コストが少ない.

今後は, さらに他の手法を提案し, それぞれの手法を比較することで, 各手法の特徴･優位性を見つけ, より精度の高い回答絞り込み手法について検討する.

参考文献

[1] 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.”, 情処研報, 自然言語処理研究会報告書, NL-135-24, 2000.

[2] 宮口正行, 桝井文人：“構文構造を考慮した質問応答のための重要文抽出”, 信学技報, 言語理解とコミュニケーション研究報告書, NLC-2002-38, 2002.

[3] S. Lee, G. G. Lee：“SiteQ/J:A Question Answering for Japanese”, In Working Notes of the Third NTCIR Workshop Meeting:QAC1, 2002.

[4] J. Fukumoto, T. Kato and F. Masui：“Question Answering Challenge(QAC1): Question answering evaluation at NTCIR Workshop 3”, In Working Notes of the Third NTCIR Workshop Meeting:QAC1, 2002.