GDAタグを利用した回答抽出システムの提案

鈴木 潤
慶應義塾大学大学院
理工学研究科 計算機科学専攻
junjun@nak.ics.keio.ac.jp

橋田 浩一
電子技術総合研究所
hasida@etl.go.jp
ソニーコンピュータサイエンス研究所
hasida@csl.sony.co.jp

1　はじめに

大量の電子化文書に容易にアクセスできるようになるにつれ、その中から必要とする情報のみを的確に抽出する技術が必要になってくる。従来の情報検索手法として、キーワードで与えられた検索要求 (query) に対するbool演算による検索が主流に扱われてきた。しかし、検索精度には大幅な改善の余地が残されており、情報検索に自然言語処理技術を適用して検索精度を改善しようとする試みもある。

文書内の意味構造を利用した情報検索に注目する。文書中の単語間の係り受け構造を利用した検索 [1, 2] や、シソーラスによる語の拡張等に関する研究が行なわれている。これらの研究は、検索対象である文書集合内の意味理解が可能となれば、より的確にユーザの検索要求に適合した回答を抽出することが可能になるというアイデアに基づいている。しかし、現在の計算機による自動意味解析技術は発展途上であり、応用分野にそのまま適用するまで解決すべき課題が多い。

本研究では、自然言語処理による意味解析そのものを研究するのではなく、解析結果である意味構造を文書中にタグとして埋め込み、これを情報検索に利用することを考える。このようにタグを用いる利点は多岐にわたる。まず、自動解析の精度を補うために人手によって意味構造を修正することができる。その他に、意味解析に関する基礎研究と意味情報を利用した応用研究開発を明確に分離できる点、タグ情報のみを利用した解析手法によって言語に依存しない汎用的なシステムを構築できる点、タグ付き文書が容易に再利用できる点、また、タグを標準化することによってさまざまな研究成果を統合できる点等が挙げられる。

以下では、GDA (Global Document Annotation) [3, 4] タグ付き文書に対する回答抽出システムについて述べる。特に、意味構造を利用した回答抽出手法を提案する。回答抽出時の解析にはGDAタグに含まれる情報のみを用いる。ここで言う意味情報は、語の間の依存関係、照応関係、およびシソーラス内での語の間の類似関係である。

1.1　回答抽出

回答抽出 (Answer Extraction)[8] とは、情報検索の一種である。通常の情報検索と回答抽出との違いは、情報検索の目的がユーザの情報に適合する文書を抽出するのに対して、回答抽出は、ユーザの検索要求に適合する文を基本単位として回答を提示するという点にある。同時に、情報検索では、文書単位の階層構造や、類似文書の分類などによる検索精度の向上を扱うのに対し、回答抽出システムでは、情報検索より一階層細かい文章中の語や文節単位の構文構造、意味構造等を利用する。

回答抽出では、MUC [9] などで提案されている情報抽出 (Infomation Extraction) と同様に、文書内の文または語単位の解析を行う。しかし、情報抽出では予め決められた種類 (会社の名前や日付など) の情報を抽出するのに対し、回答抽出では抽出すべき情報にそのような限定がない。

回答抽出はTREC [10] で取り上げられている質問応答 (Question Answering) の枠組に近いものであるが、一問一答を目指すのではなく、ユーザの質問に対して、適合すると思われる部分を全て抽出し提示するユーザ支援システムに相当する。よって、「誰が」、「何処に」等の疑問詞に相当する語が何であるかといった言語知識を必要とする解析は行っていない。

計算機が人間と同等の知識処理を行なうことが可能なら、一問一答で回答が得られる質問応答により情報を得ることができる。しかし、現在の技術でそれを完全に実現するのは不可能である。これを補うには、インタラクションを介して人間の知的能力を利用する必要がある。つまり、ユーザとシステムの能力を互いに補完して回答に達するのである。ユーザの発する質問に対して適合すると思われる箇所を部分的に提示し、それが実際に適合しているか否かはユーザの判断で決定する。その判断を支援するのが以下で提案する回答抽出システムであり、これとインタラクションの機能を統合することによってさらに高精度の情報検索が可能になると考えられる。

1.2　質問要求

回答抽出システムにおいては、自然言語による検索要求を行う。これにより、自然言語により与えられた語と語の関係を利用して回答を抽出する。ユーザがキーワードとそれらの間の依存関係を何らかの仕方で入力すれば同様の検索ができる。しかし本研究では、ユーザに負担がかからず、より扱い易いシステムにする必要から、自然言語で与えられた検索要求を自動で解析し、回答を抽出するシステムを目指す。

特にユーザが精通していない分野での検索の場合など、欲しい情報を特徴的なキーワードの組合せとして明確に表現できないことが多い。その際の検索要求の具体化が最も困難な作業の一つである [11, 12]。本稿で提案する回答抽出システムは、キーワードによる bool 検索では抽出が困難である質問にも柔軟に対応できる。そのための具体的な手法は、検索要求内の各語のシソーラスによる拡張と、語の間の意味関係が作る構造に関する柔軟な照合である。

しかし、ユーザの検索要求が具体化されていない場合に対処する手法としてインタラクティブな検索手法が提案されており [13, 14]、この手法がより柔軟な情報検索には必須と考えられる。本稿ではインタラクションには詳しく立ち入らないが、今後はより的確なユーザ支援を行なうためインタラクティブな検索をサポートしたい。

2 GDA

GDA は、電子化文書の意味的な構造を明示する XML のタグ集合 (GDA タグ集合) を策定、公開し、これに基づく多用途の知的コンテンツをサポートする応用技術の開発と普及を推進することにより、インターネット等でこのタグ集合を広めることを目指すプロジェクトである。

これにより、文書の意味を計算機に理解可能な形式にし、それを利用した応用分野での研究、開発での利用を高精度で実現することができる。

GDA タグを用いた応用分野での研究報告がなされている [6, 7]。本稿では、回答抽出システムとして、情報検索への応用の一例を示す。

3 特徴と利点

本稿の回答抽出システムの特徴と利点として以下の 5 つを挙げる。

1. XML 形式タグに適用した検索システム
   今後、インターネット上に流れる情報の多くが XML で記述されることを考えると、検索エンジン自体が XML 文書の構造を認識し、処理する利点は大きい。
2. GDA タグ情報の利用自然言語で記述された文書集合に対して GDA タグを付加し、回答抽出時の解析には GDA タグ情報のみを利用したシステムとする。
3. 活性拡散を用いた類似度計算
   本システムで扱う情報は、主に質問要求 (検索要求)、シソーラス、被検索文書集合の 3 つの階層から成る。質問要求及び被検索文書集合内の意味情報を表すネットワークをシソーラスで結んだ統合ネットワーク上での活性拡散を用いて活性値を計算し、質問要求と被検索文書集合内の意味構造の類似性を測る尺度に用いる。
4. 類似度による回答抽出
   活性拡散により計算された活性値を用いて、質問要求と抽出文との類似性を判定する。質問要求と抽出文の類似性をコサイン距離により計算し、抽出された回答文の適合度として使用する。閾値を用いて類似度が高い抽出文のみを回答として提示する。
5. 確率的照応タグ
   厳密な意味でのタグ付けは一対一対応をつけることであるが、厳密な照応解析が困難であることから照応先の尤度をタグに付加する。これにより、不正確な照応解析を用いた抽出処理が可能になる。

4 システムの概略

システムの前提条件として、ユーザが発する質問要求、被検索文書集合は全て自然言語で記述されていることとする。また、ユーザに提示する回答は、文書集合内の文書ではなく、各文書内の文を単位とする。

本稿で述べるシステムは、機能別に分類した場合、ユーザの質問要求および被検索文書集合に GDA タグを付加する GDA タグ付け部、照応関係を与える簡易照応解析部、質問要求を拡張するシソーラス拡張部、被検索文書集合から回答を抽出する回答抽出部で構成される。

4.1 GDA タグ付け

まず前準備として被検索文書集合に対して形態素解析と構文解析を施す。これには既存の形態素解析ツール juman3.61 と構文解析ツール knp2.06b を使用する。その出力を GDA 形式に変換する。ここでの GDA 形式とは、knp の出力の統語情報のみを扱い、意味情報までは付加していない。与えられる情報は、品詞タグ、読み属性、基本形、表層格の属性である。また、語の係り受け構造を表す属性を付加する。その後、簡易照応解析の結果をタグに付加する。

4.2 簡易照応解析

現段階での一般的な照応解析の精度は、7 割から 8 割程度と言ったところである。意味解析の困難性から、現段階では厳密な照応解析モジュールの作成にはコストと実現性の問題点があるため、今回は明確な照応解析を目指すのではなく照応先の候補を列挙し、その尤度をタグに埋め込むことで解決する。

本稿では、指示詞、代名詞、ゼロ代名詞に関しての照応解析を扱う。照応関係にあると推測される語を、文献 [15, 16, 17] の手法を参考にし列挙する。列挙する際に与えられるスコアにより照応の指示対象を決定するが、本研究では解を一意に限定するのではなく列挙された候補全てがある尤度を持って指示対象であると仮定し、その値をネットワーク上の照応リンクの重みとして使用する。

実際の計算は、以下のように照応先と照応元の距離に比例して減衰する式として与える。

$$P = \frac{N}{l} \tag{1}$$

ここで、l は照応先と照応元の距離、N は列挙時に与えられたスコアである。

このように、厳密に照応解析を行っていない。しかし、回答抽出時の計算手法により、照応解析の曖昧さに柔軟に対処する。

4.3 シソーラス拡張

シソーラスを使用して質問要求を拡張する。これは、被検索文書集合内に表れる表現と、実際に質問要求と

して使用される表現とは必ずしも同じではないことから、語の類似性を考慮し、類似性の高い語も検索対象に含めて検索を行う事で再現率が向上する効果を期待するものである。

シソーラスによる類似語を拡張する手法は様々な方法が提案されているが、本稿ではシソーラス内の各二単語間の意味距離を用いて拡張を行う。意味距離の計算には、シソーラス内の各概念間の繋がり及び意味的な繋がり (反義、多義など) でネットワーク構造を作成し、シソーラス中に表れる各二単語間の意味距離を活性拡散で計算する。ここで求められた値を質問要求と被検索文書集合内の対応する語のリンク重みとして使用する。

回答抽出時にシソーラス拡張を含めた活性拡散による計算を行うと検索時間に多大な影響を与えるので、シソーラス拡張は検索前に事前に計算しておく。これは、シソーラス内の意味距離は質問要求に依らず一定であることによる。

4.4 意味構造を利用した回答抽出手法

回答抽出には、GDA タグ内に記述されている意味的依存関係のリンクからなるネットワーク構造を用いた活性拡散を行い、そこで得られた活性値による意味構造の類似度を比較することで、質問要求と抽出された回答文の尤もらしさの判断基準とする。意味的依存関係には以下の関係を使用する。

- 自立語の係り受け関係
- 照応関係

回答抽出の簡単な手順は以下の通りである。

1. 質問要求から自立語を抽出
2. 質問要求の意味構造を解析
3. 抽出した自立語をシソーラスを用いて拡張
4. 拡張した語が 1 つでも出現する文を全て抽出 (以下抽出文とする)
5. 抽出した文毎に活性拡散
6. 得られた活性値を基に質問要求と類似度計算
7. 類似度の高い順に回答を列挙
8. 閾値を用いて不要な回答を足切り

よって、活性拡散によって得られた活性値を用いて類似度を計算する。つまり、活性拡散は質問要求と抽出文の類似性を測る尺度として用いる。

4.4.1 活性拡散

GDA タグ付き文書集合に対する要約に活性拡散を用いた手法が提案されている [5, 6, 7]。回答抽出では文単位での抽出を目的とするため、通常の活性拡散で計算できる活性値を用いて質問要求と抽出文の類似性の判断基準とするのは難しい。そこで、活性拡散時の活性値を単一値からベクトル値に改良する。ベクトル内の各要素が表すものは、検索要求内の自立語の出現順に対応おり、検索要求に現れる自立語の数によりベクトルの次元は変化する。これにより各活性値ベクトルが表すものは、あるベクトル内の要素に対応する自立語の意味距離と解釈することができる。活性拡散後、あるノードでの活性値ベクトルを見ることで、そのノードとベクトル内の該当する要素が表す検索要求の自立語との意味的結合度を表した数値として解釈できる。ここで、ノードは GDA タグのエレメントに相当する。

これは、シソーラス展開時に使用した二単語間の意味距離を計算するのと同等の意味を含んでおり、ベクトル値はそれらを一括して扱うための適用である。

ここで、活性拡散時の条件を以下に提示する。

- 各ノードに入ってきたリンクには拡散しない (拡散は往復しない)
- 拡散するリンクが存在しない場合はそこで拡散が終了
- 拡散の始点は質問要求の各自立語のみ
- 質問要求から被検索文書集合に向かうリンクは有向リンク
  そのときのリンク重みはシソーラス内の二単語間意味距離
- 同一層内のリンクは無向リンク

4.4.2 類似度計算手法

次に、質問要求内の意味構造を用いて活性拡散により計算された活性値ベクトルと、抽出文との活性値ベクトルとのコサイン距離を計算する。コサイン距離の計算は質問要求と抽出文の対応する自立語間で行われる。

$$\sigma(d_x, d_y) = \frac{\sum_{i=1}^{T} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{T} x_i^2 \times \sum_{i=1}^{T} y_i^2}} \qquad (2)$$

ここで、d は各ノードに相当し、i は検索要求内の自立語の番号であり、x_i、y_i は各ノードの i 番目の要素を表す。

その後、文内の全てのノードの類似度を合計し、その値を抽出文の類似度として使用する。値が大きい順に回答として適合していると判断しユーザに提示する。

4.5 本手法の利点

この手法は、抽出文に質問要求内の語が全て存在しない場合でも、不足語に対応する活性値ベクトル内の要素は 0 になるが、類似度が 0 になるとは限らない。よって、質問要求に必要な語が不足している場合、または、逆に不要な語が存在する場合でも、柔軟に回答を抽出することができるという利点がある。

また、語を全て数値として扱っているので、bool 検索のように語が出現するかしないかといった検索よりも柔軟に対応することができる。これにより、ユーザからの質問要求の意味的類似文を抽出できる可能性が増えると考える。

5 実験及び結果

本稿で提案する回答抽出システムの性能を評価するための実験を行う。

実験は、人手による修正済み GDA タグを付加された文書集合及び自動 GDA タグ付けによって作成された文書集合 14631 記事を使用した。これに、作成した質問 20 問を用いて実験を行った。質問には、自立語をキーワードとした and 検索では検索できないものを使用する。また、質問に対する回答文は複数あるものを用意した。適合率の計算には抽出された回答を人手で判断し、回答として妥当と判断された場合に正解文として計算している。

表 1: 質問と回答文の例

質問例：	子供のスポーツ選手が活躍した (記事)
回答例：	... 最年少のモンゴルの太極拳少女、 ... ちゃん (9 つ) が十四日、... 国際舞台にデビューする。
質問例：	戦時中に作成された文学
回答例：	... 太平洋戦争にかけての暗い時代 に生まれた俳句作品...
:	「一輪のきらりと花が光る突撃」の赤黄男... 二人の残した作品や日記から、当時の時代...

表 2: 実験結果

	recall	precision
key(and)	0.0	0.0
key(or)+thes	82.6	0.3
key+thes+sem	78.3	36.7

6 考察

活性拡散と類似度計算により、質問要求と意味構造が近似している文が回答抽出時に上位に位置付けされる。これにより、ユーザは類似度が高いものから調査することで検索意図に適合する部分を容易に発見できる。

現在は、抽出文の類似度に閾値を設けることで、不必要な文章をユーザに提示しない手法を採用している。現在の設定値は実験により良好であると判断した値を使用しているが、この閾値を用いて回答の足切りを行っても、再現率はそれほど低下しない。これは、活性拡散と類似度計算によって、必要と思われる回答が上位に位置しているため、閾値を下回ることがあまりないことを表している。

以上の点から、本論文で提案する回答抽出手法のような、数値を用いた抽出手法でも回答抽出システムとして動作する事が分かった。

また、本稿で提案するシステムの特徴として、再現率に対して適合率が低い結果となった。これは、システムの性質が回答と考えられる部分を全てユーザに提示し、その中からユーザの判断で真の正解を選んでもらうユーザ支援を目的としている点からこのような結果になったと考えられる。

7 ユーザインタラクション

考察でも述べた通り、本稿のシステムでは不必要な回答を提示する量が多いといった問題が残されている。しかし、これはユーザとのインタラクションを行うことにより、改善することができる。前述の通り、類似度が高位の回答を見て新たな質問要求を拡張していくことにより、最終的には、一問一答の形で回答を抽出することができると考えられる。

現在は、インタラクションに関する性能を表す数値としての実験の検証が十分に行われていないため、今回は扱わなかったが、今後これに適用し計算機の知識が不十分でもインタラクションによる解消を用いることで質問応答が可能になるようなシステムを作成してみたいと考えている。

8 まとめ

情報検索の分野で GDA タグを利用したシステムの一例を示した。意味構造にわたる活性拡散を用いた回答抽出手法の有効性を確認できた。

今後、意味構造に関するより高精度のタグ付けを行うことが可能となれば、より高精度の情報検索が可能となる。

また、今後は GDA タグ付けを施した辞書等の知識ベースを利用することによってさらに柔軟かつ高精度の検索を行う方法など、意味構造に基づく抽出に関しては多くの興味深い研究課題がある。

本稿のように GDA タグを用いることにより、言語の意味解析技術と応用分野での意味解析結果の利用とを分離して研究を進めることが可能である。自然言語処理の基礎研究と応用開発の両面における GDA タグの普及が一層進展することを期待する。

謝辞

本研究では、自然言語処理ツールとして juman および knp、実験用データとして毎日新聞 CDROM 版を使用した。これらの作成と公開に尽力された方々に敬意を表し、感謝する。

参考文献

[1] 新美 和彦, 兵藤 安昭, 池田尚志, "係り受け情報を用いた全文検索とその評価", 第 11 回デジタル図書館ワークショップ, pp.27–34, 1998.

[2] 立石 健二, 峯 恒憲, 雨宮 真人, "係り受け構造や語の意味情報を利用した日本語テキスト検索システム", 言語処理学会第 5 回年次大会発表論文集, pp.317–320, 1999.

[3] Koiti Hashida, et al., "Global Document Annotation", http://www.etl.go.jp/etl/nl/gda.

[4] 橋田 浩一, "意味的修飾に基づく多用途の知的コンテンツ", 人工知能学会誌, Vol. 13, No. 4. 1998.

[5] K. Hasida, S. Ishizaki, H. Isahara, "A connectionist approach to the generation of abstracts", Natural Language Generation, pp.140–156, 1987.

[6] 内山 将夫, 橋田 浩一, "GDA タグを利用した複数文書の要約", 言語処理学会, pp. 376–379, 2000.

[7] 長尾 確, 白井 良成, 橋田 浩一, "言語的アノテーションに基づくマルチメディア要約", 言語処理学会, pp. 380–383, 2000.

[8] D.M. Aliod, M. Hess, "On the Scalability of the Answer Extraction System "ExtrAns"", Applications of Natural Language to Information Systems, pp.219–224, 1999.

[9] Message Understanding Conference, http://www.muc.said.com/

[10] Text Retreaval Conference, http://trec.nist.gov/

[11] N. Oddy, "Information retrieval through man-machine dialogue", Journal of Documentation, 33(1), pp.1–14, 1977.

[12] S. Taylor, "Question-negotiation and information seeking in libraries", Collage And Resarch Libraries, 178-194, 1968.

[13] 橋田 浩一, 豊浦 潤, 津高 新一郎, "構造化文書に基づくインタラクティブな意味的情報検索", 情報処理学会研究報告 99-ICS-116, pp.13–16, 1999.

[14] 高橋 敦子, 平井 誠, 北橋 忠宏, "自然言語を用いた対話形式による文書検索における辞典情報の利用", 情報処理学会 自然言語処理研究会, pp.151–157, 2000.

[15] R. Mitkov, "Robust Pronoun resolution with limited knowledge", Pro. of COLING-ACL'98, pp.869-875, 1998.

[16] 村田 真樹, 長尾 真, "用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定", 言語処理学会誌, Vol.4, No.1, 1997.

[17] M. Walker, M. iida, S. Cote, "Japanese Discourse and the Process of Centering", Computational Linguistics, col.20, No.2, pp.193–232, 1994.