2G-6

WEB 検索による知識文の獲得と 意味グラフ照合推論による質問応答システム Metis

加藤 裕平† 蒲生 健輝† 古川 勇人† 韓 東力‡ 原田 実‡
†,‡青山学院大学 理工学部 情報テクノロジー学科

1. はじめに

インターネットという「知識の宝庫」内に，日常よくある質問に対する解答を含む文章が多く含まれている．しかし，通常のキーワード検索では，キーワード同士の複雑な関係を指定できず，検索されたページを逐一開いて答えを見つけるのにかなりの時間を要している．本研究では質問文を意味解析し得た意味グラフ Q から SVM を用いて抽出した検索キーワードを用いて WEB 検索により知識文を獲得し，これを同様に意味解析して得た知識グラフ K と Q を照合して回答を導き出す高精度の質問応答システム Metis を構築する．

2. 自然言語の意味グラフ表現

意味グラフとは Sowa の概念グラフ[1]を拡張したもので，日本文に表現されている全ての事象を語とその間の深層関係(リレーション)で表現したものである．ノードは一般的に(属性)→[Type label：referent]で表示される．Type label は語の内延で EDR の語意（概念 ID）で表し，referent は語の外延（インスタンス）である．必要ならば，referent はインスタンスの多重度も表現できる．リレーションは語と語の深層関係であり，名前つきアークとして表現される．属性はノードの持つ様相（テンス，ムード，アスペクト）を表現するためのものである．

3. グラフの照合定理

照合の正しさは，「ある質問グラフ Q が真であるためには，真であることが確定している知識グラフ K(または K の部分グラフ)への特殊化の系列が存在することである」という定理に基づく[1]．つまり，Q が真であるためには Q を特殊化し(または特殊化しなくても)K の部分グラフと合同になればよい．しかし，時には知識グラフに余分な修飾が含まれ，この条件が厳しすぎることがある．そこで，グラフ照合の際にノード飛び越しを許すことにし，合同から外れていない程度を数値的に反映するグラフ類似度を導入した．このグラフ類似度はシステムが抽出した解の信頼性の尺度といえる．グラフ類似度はノード類似度とリレーション類似度の和で定義され，下式で計算する．

グラフ類似度 = ノード類似度＋リレーション類似度

$$\text{ノード類似度} = \frac{\sum \text{照合ノードペアの概念 類似度}}{\text{質問グラフのノード数}} \times 50$$

$$\text{リレーション類似度} = \frac{\sum \text{照合ペア間のリレーション ペアの類似度}}{\text{質問グラフのリレーション数}} \times 50$$

3. 質問応答システム Metis のデータフロー

本システムのデータフローを図 2 に示す．本システムでは入力された質問文に対し本研究室で開発した意味解析ソフト SAGE を用いて解析を行って得られる意味グラフを用いて処理を行っていく．

Metisのデータフロー

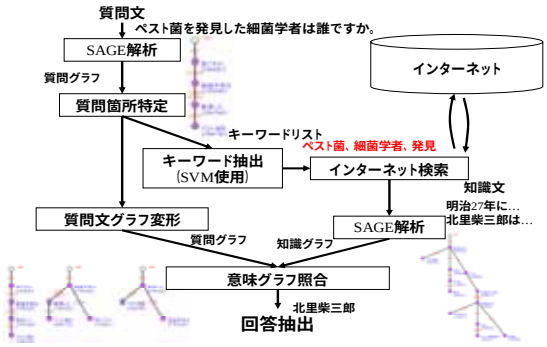

図 1 Metisのデータフロー

4. WEB 検索による知識文の獲得

SVM を用い，図 1 に示すように質問文からキーワードを抽出し，線形判別式の得点順にソートし，キーワードの組み合わせを作成する．組み合わされたキーワードを使い検索エンジン Google で答えを含んでいそうな知識文があるページの URL 複数個を取得する．取得したページから HTML パーサーを使いテキスト情報とタイトル情報を抜き出す．抜き出したテキストを文単位に分割する．これらの文をキーワードを含む個数と文の長さでソートし，検索結果の知識文リストとする．

5. 質問箇所の特定

質問文の意味グラフから疑問詞を表す語意をもつフレームを探索し，what・who・where・when・other のどれかに分類するとともに照合のための代替語意を付加する．

6.質問文の言い換え

マッチング処理を容易にするため質問文の言いかえを行う．言い換えを行うにあたり質問文を以下の図のようなタイプに分類し，その文のタイプに適した変形を行う．

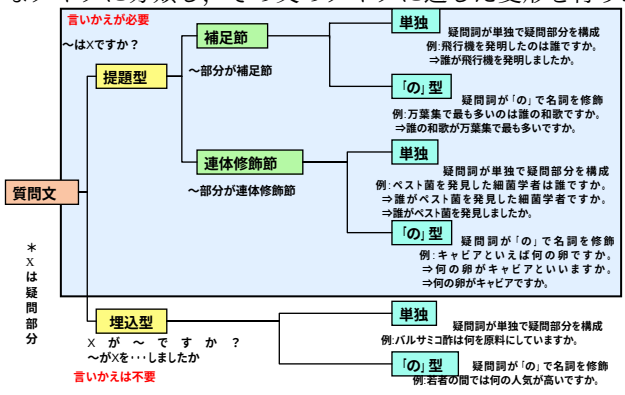

図 2 質問文のタイプ

・**「提題型-補足節」**の言い換えについて

例えば「飛行機を発明したのは誰ですか？」という質問

A Question-Answering System based on Semantic Graph Matching with Knowledge Sentences Discovered from the Web.
Yuhei Kato*, Toshiaki Gamo*, Yuto Furukawa*, Dongli Han ** and Minoru Harada **
* Department of Integrated Information Technology, Aoyama Gakuin University.
**Department of Integrated Information Technology, Faculty of Science and Engineering, Aoyama Gakuin University.

文は「誰が飛行機を発明しましたか？」という文に意味グラフ上で変換する．

・**「提題型-連体修飾節」**の言い換えについて

例えば「ペスト菌を発見した細菌学者は誰ですか？」という質問文は「誰がペスト菌を発見した細菌学者ですか？」や「誰がペスト菌を発見しましたか？」という文に意味グラフ上で変換する．

7. 質問グラフと知識グラフの照合

7.1. 前処理

グラフの照合に先立ち，アルゴリズムの高速性や簡潔性のために前処理を行う．質問グラフの主述語ノード qst があり，qst と照合ペアとなる知識ノード kst を発見し，それぞれのグラフを qst，kst を根とする張る木に変換する．連体修飾における修飾子から被修飾子への辺や並列を表す辺は削除するが，後の照合時にはこれらも用いる．

4.2. ノード飛び越し

グラフ照合において余分な修飾による合同から外れた不一致部分を照合対象から除外し，質問文に照合するノードの探索を継続する為にノードの飛び越し処理を行う[2]．ノードの飛び越しは，アークのつなぎ換えによって実現する．つなぎ変えたアークには飛び越したノード間の深層格情報をリストとして持たせる．ノード飛び越しを行ったことをグラフ類似度の減少として反映する．図４に二種類のノード飛び越しの様子を示す．

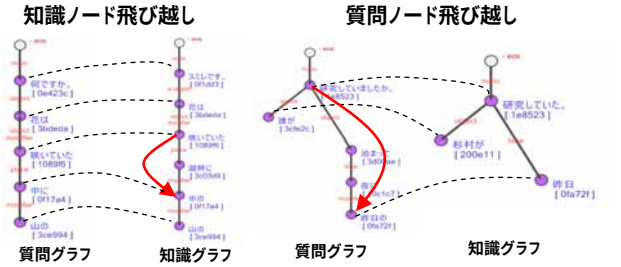

図 3ノード飛び越し

4.3. グラフの照合アルゴリズム

質問グラフを張る木の根 qst からスタートして，質問グラフのノード qn を縦型に訪問しながら知識グラフのノード kn と照合していく．また照合を行いながらグラフ類似度を計算する．ノードの照合において，質問箇所ノードとそれ以外のノードでは条件を替えて行う．一般のノードでの照合条件は表１のものを使用し，質問箇所ノードの場合は表２の条件を使用する．リレーションの照合に関しては対象リレーション同士が同じ深層格を持っているならば類似度1.0とし，リレーションが違う場合においては，対象リレーション同士が表３に示す類似リレーショングループの中で同じグループに属していた場合，そのグループの類似度をリレーションの類似度とする．また，飛び越しを行ってリレーションがリストになっている場合においてはリレーションリスト同士内に同じ深層格があれば類似度を1.0とし，対象リスト内の深層格の１つが同じグループに属していればそのグループに所属していることとする．照合条件がリレーション類似度＋概念類似度としているのはリレーション類似度，概念類似度がともに低かった場合に照合ペアとなるのを防ぐためである．質問箇所のノード照合に関しては概念類似度を求める場合，質問文の特定の際に付与した代替語意を用いて計算を行う．グラフの照合はWEB検索により得られた知識文に対しそれぞれ質問文の言い換えで得られたすべての変換後の意味グラフと照合し，解を抽出する．

$$\text{概念類似度} = \max\left(\frac{2\times d_c}{d_q + d_k}\right)$$

d_q,d_k：それぞれの概念の概念深さ

d_c　：dq,dkの共通概念の概念深さ

表 1 一般ノード対の照合条件

	質問ノード(Qn)	知識ノード(Kn)
概念類似度	0.27以上(経験的に定めた値)	
リファレント	一般概念(*)	何でも良い
	固有概念	質問と同じリファレントを持つ
属性	否定属性の有無が合致	
リレーション	リレーション類似度＋概念類似度 >= 1.27	

表 2 質問箇所ノード対の照合条件

	質問ノード(Qn)	知識ノード(Kn)
概念類似度	0.50以上(経験的に定めた値)	
属性	否定属性の有無が合致	
品詞	-	動詞以外
リレーション	リレーション類似度＋概念類似度 >= 1.50	

表３ リレーショングループ

合致度	グループ名	対象深層格
80%	動作の主体	agent object
	対象説明	object modifier a-object
	・・・	・・・
60%	動作対象･目標	goal beneficiary purpose
	場所	place goal from-to
	・・・	・・・
40%	理由･原因	NULL nil cause logical reason
	・・・	・・・

6. 解の抽出と表示

照合結果は回答とそのスコアという形で保存される．抽出された解をグラフ類似度の順にソートし順位付で解を表示する．

7. 評価実験

クイズ・ミリオネアの問題集から問題文を集め，質問文100セットの評価実験を行った．その結果を表に示す．

表４ Web 検索実験結果

適切な知識文を取得	誤った知識文を取得	
48% (48/100)	52% (52/100)	
	無回答	誤回答
	87% (45/52)	13% (7/52)

表５ グラフ照合実験結果

正答	誤答	無回答
75% (36/48)	6% (3/48)	19% (9/48)

適切な知識文が得られた割合が 48%と低い，これは指定されたキーワード全てを含む文のみを抽出したためだと考えられる．しかし，得られた知識文を用いた照合の正解率は 75%と高く，不適切な知識文からは回答を出力しないことから誤回答率は低い．従って、有用性の観点からの正解を得る確率は 36/100 = 36%だが，インターネットに解がない時の無回答を正解と考えれば，総合的な正答率は(45 + 36) / 100 = 81%といえる．これは，本手法では文章内容を精密に照合して行っているためといえる．

謝辞

本研究の一部は文部科学省科学研究費基盤研究 C『意味ベースの精密な照合を行う高精度質問応答システムの開発研究』の補助金を用いて行われました．ここに感謝いたします．

参考文献

[1] John F. Sowa: “Conceptual Structures: Information Processing in Mind and Machine”, 1984.

[2] 竹原一彰,安部建助,安田智成,韓東力,原田実:“質問応答のための質問文と知識文の間の意味ベースでの精密な照合方式”,情報処理学会第66回全国大会論文集,6U-03,第２分冊pp.173-174(2004.3).