WWWからDescriptive知識を抽出・提示するシステムMurasakiの試作

川村 佳史　桝井 文人　河合 敦夫　井須 尚紀
三重大学工学部情報工学科
{*yosi, masui, kawai,isu*}*@ai.info.mie-u.ac.jp*

1　まえがき

蓄積され続ける膨大な電子化テキストから効率よく知識を取り出すための技術として「情報アクセスのためのテキスト処理」技術が重要視されている．この技術の基本は，大規模なコーパスから情報要求に応じた情報のみを効率よく探し出すタスクをこなすことである．大規模なコーパスから情報を抽出する技術の典型例として，WWW 検索がある．WWW 検索は，World Wide Web という膨大な電子化テキストの中からユーザが指定したクエリに適するテキスト集合のみを実時間で提示する．

ただし，WWW 検索は情報検索技術の応用であり，ユーザの要求に対するレスポンスは，あくまでも大雑把に適合する文書の提示にとどまる．これに対して，ユーザの要求に関して答えそのものを返す技術として，質問応答が注目されている．質問応答は，これまで名詞や固有表現を回答する factoid タスクに関する研究が進み，一定の成果をあげている．[1, 2].

ところが，質問応答技術の実用的応用を考えた場合，上記の factoid 型タスクよりも高次な（理由や定義を回答する）タスクの方が潜在的なニーズは高い．例えば, 角ら [3] は WWW マイニングによって構築したオントロジーを利用した協調型合意形成支援システムの提案に際してユーザの誤解の解消や, 未知の知識の補完の重要性を指摘している．黒橋 [4] は, 大規模なテキスト知識ベースを参照する質問応答システムにおけるユーザとのインタラクティブなインターフェイスにおいて，ユーザにとって未知の知識を扱う際の質問応答の重要性を指摘している．

このような背景において，最近，対話を指向するタスクや，why 型や how 型の質問に答えるタスク，定義を答えるタスクが検討されている [5]．しかしながら，これらのタスクに対応するためには，これまでの factoid 型タスクの延長線上の技術では対応できない問題が多い．例えば，対話施行タスクでは照応処理や文脈処理は必須の課題である．定義や理由を問うタスクでは照応処理や文脈処理に加えて，自動要約技術を包含した形で研究を進める必要があり，実用化における難易度は高い．

本論文では，これらの高次のタスクのうち，定義タスクに対して description で回答する試作システム Murasaki について報告する．Murasaki は，クエリと比較を示す定型パターンを組み合わせて新たな検索表現を生成する．次に，生成した検索表現を WWW 検索することによってクエリと descriptive 知識の共起知識を得る．WWW から得た共起頻度に基づいて descriptor の集合の尤度を決定，共起知識を再構成する．最後に，再構成した共起知識を尤度順に視覚化して表示する．

以下，2 章で基本的なコンセプトについて述べ，3 章で Murasaki の構成について説明する．4 章でシステムの評価実験について述べ，5 章で評価結果に対する考察を行う．

2　基本的な考え方

本章では，Murasaki の設計における基本的な考え方について述べる．

WWW からの知識獲得

我々は，以下のような理由から，World Wide Web(WWW) を有効な知識源として利用できると考えている．まず，WWW は現存するうちの最大規模のテキストコーパスである．それ故，WWW から適切に知識を取り出すことができれば，非常に網羅性の高い知識が得られるはずである．また，WWW 上の記述データはダイナミックに変化している．よって，WWW から動的に知識を取り出すことができれば，時事的概念や一般的なイメージの変遷を反映した知識を得ることになる．

概念の descriptive な表現

概念の定義を表現するためには，構文解析や文脈把握その他の複雑高度な処理を施す必要があり，技術的難易度が高い．そのため，これらの複雑な処理は計算量増加の要因となり，動的な処理を考えた場合に不都合となる．

川崎 [6] は, プログラム理解システムを対象として, システムに事前知識がない場合に, 類似した知識を用いて言い換えることで, 理解した場合と同等の効果を得られることを示している．そこで，概念に対する定義を過不足なく回答するのではなく，定義がイメージでき，理解を促すに十分な補足知識を提供することに力点を置き，概念を複数の descriptor の集合で表現することを考える．descriptor とは，概念の特徴を表現する連想語である．例えば,「ライブドア」については,「新興企業」「マネーゲーム会社」「ポータルビジネス」のような descriptor が考えられる．我々は，descriptor 集合の提示によって連想に基づく理解を促し，

結果的には定義文を示す場合と同等の効果が得られると考えている．

また，descriptor を重み付きで表現すれば，概念を descriptor のベクトル空間で規定することになる．これは，概念を確率的な概念記述で表現するモデルとほぼ等しい．したがって，WWW を利用して概念記述データベースを構築する手法を応用することによって，descriptor 集合データベースを構築することができるはずである．これによって，処理の複雑さを増大せずに大量の descriptor を自動取得することが可能である．

descriptive 知識の可視化

複雑な情報や大量の情報を可視化する研究が注目されている [7][1]．概念の descriptive な知識も，尤度に基づいてグラフ化したり，リスト表示するなどの可視化を施すことによって，直感的な把握を支援すると考える．

3 Murasaki の基本構成

本章では，descriptive 知識抽出提示システム Murasaki の構成について説明する．

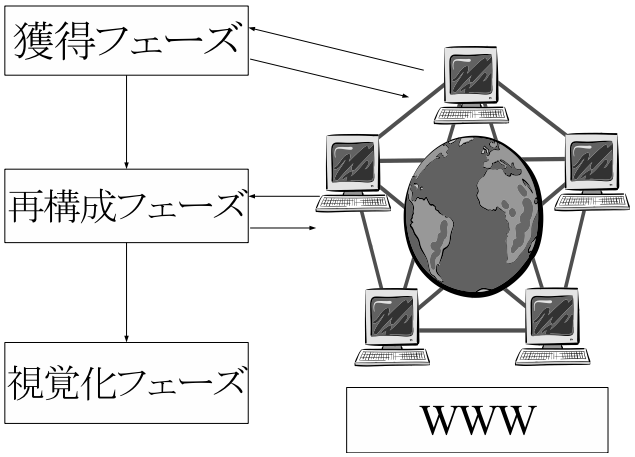

図 1: Murasaki の構成

Murasaki の処理は，WWW から知識を獲得する「獲得フェーズ」と，獲得した知識を再構成する「再構成フェーズ」，再構成された知識を視覚化して提示する「視覚化フェーズ」からなる（図 1）．

以下，各フェーズについて詳述する．

3.1 獲得フェーズ

獲得フェーズは，ユーザが指定したクエリと descriptor の共起知識を WWW から獲得する．ここでは，意味レベルの比較を解析するためのフィードバック手法（桝井ら 2005）を応用する．具体的な処理の流れを示す．

(1) ユーザから入力されたクエリ α を認識する．

(2) α を，あらかじめ準備したフレーム「x_ground」に適用し，検索表現を生成する．このとき，$ground$ は「のような」「のように」「という」などの定型表現である．例えば，α=「メタリカ」が与えられた場合，検索表現として「メタリカのような」が生成される．

(3) 生成した検索表現を WWW 検索し，検索表現を含む文書集合を得る．

(4) 得られた文書集合から検索表現を含む文を抜き出し，さらに，比較表現「x_ground_γ」を認識する．上記の例では，「メタリカのようなバンド」「メタリカのようなスラッシュ系」などが得られる．

3.2 再構成フェーズ

再構成フェーズでは，再構成フェーズで得た比較表現を WWW 上で検証することによって重み付き descriptor 集合を構成する．

(1) 獲得フェーズで抽出した比較表現「$x\_groound\_\gamma$」について WWW 検索を行い，頻度情報を得る．

(2) 頻度情報に基づいて descriptor の重みを計算する．

(3) γ に相当する語句を抽出し，各々に重みを付与して descriptor 集合 $x : \{\gamma_1, \gamma_2, \ldots, \gamma_i\}$ を構築する．上記の例ではメタリカ : { バンド, スラッシュ系, 音, ヘビー, メタル, ...} のような結果が得られる．

3.3 視覚化フェーズ

視覚化フェーズは cgi として実装されており，重み付き descriptor 集合をリスト表示やグラフに変換して提示する．

(1) 獲得フェーズにおいて得られた比較表現をリスト表示する．

(2) 再構成フェーズにおいて得られた重み付き descriptor 集合を重みに基づいたグラフとして表示する (付録:図 5)．

4 実験と評価

Murasaki の基本性能を検証するために，簡単な評価実験を行った．被験者として大学生 20 名を使った．被験者には，自由にクエリを考えてもらい，Murasaki を用いて descriptive 知識を得る．得られた descriptive 知識それぞれに対して評価してもらった．評価は，得た知識が (1) イメージ通りであるか (2) イメージ通りでないかの二値判断をしてもらった．

上記の結果，173 組のクエリ評価結果が得られた．得られた評価結果から無作為に 70 組を選び，それらのランキング性能を示すスコアを計算した．スコアとして，descriptor を尤度順に整列させ，上位 5 位までと 10 位までの MRR(Mean Reciprocal Rank)[2] を計算した．この作業

[1] 動向情報の要約と可視化に関するワークショップが開催されている．http://must.c.u-tokyo.ac.jp/

[2] 順位の逆数の平均値

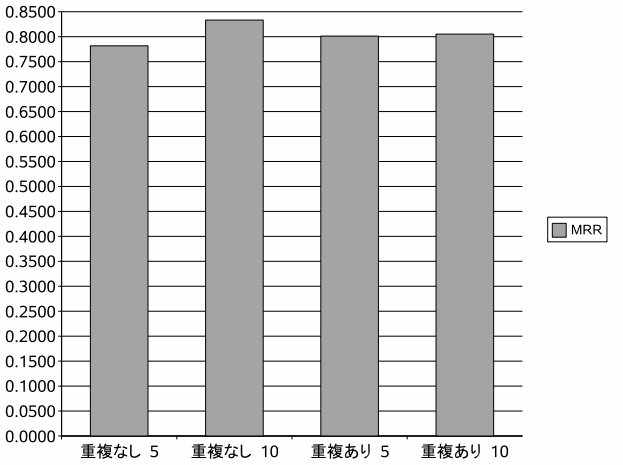

図 2: descriptor 集合のスコア結果

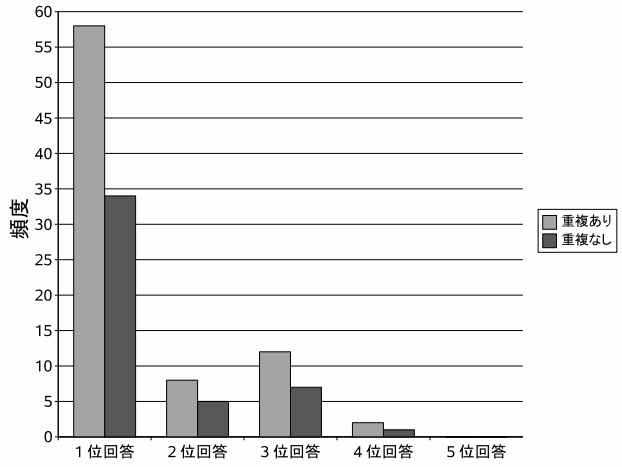

図 3: 正解順位の度数分布

を 57 回繰り返し，各スコアの平均を求めた．スコアを求める際に，同順位の要素が複数現れた場合，評価対象とした計算結果（重複あり）と，評価対象から外した計算結果（重複なし）を求めた．評価対象として計算した結果は，同順位の要素が複数個妥当と判断されても加算はせず，要素一個が妥当であったとみなしてスコアを与えた．

MRR に基づく評価結果を図 2 に示す．全体として，0.8 程度のスコアが得られた．重複に関しては明確な差はみられなかった．上位 5 位までの結果と上位 10 位までの結果についても大きな差はみられなかったが，上位 10 位までの結果の方がわずかにスコアが上回っている．

5 考察

本章では，評価結果に対する考察を行う．

全体的なスコアが 0.8 程度得られたことから，Murasaki によって提示された descriptor 集合の主要なものが概ね妥当であったと考えられる．表 1 に成功例を示す．

表 1: 「ジュビロ」に関する descriptive 知識

descriptor	重み	妥当性
チーム	0.231	○
サッカー	0.154	○
強豪	0.115	○
パス回し	0.115	○
パスサッカー	0.077	○

図 3 は，正解できた順位の度数を示す．この結果をみると，順位 1 位で妥当な descriptor を提示できたケースが多いことがわかる．さらに，ほとんどの場合，順位 3 位までに妥当な descriptor を提示できており，Murasaki のランキング性能が有効に働いていることが示唆される．さらに詳しく分析すると，上位にあげられたものは，「ジュビロ」に対して「チーム」，「ロサンゼルス」に対して「都市」というようにクエリに対して上位概念の関係にあるものとなる傾向が強い．今回実験で行った重み付けは，共起頻度に基づいているため，上位にランクされるものは必然的に一般性が高く，上位概念となることは容易に推測できる．また，上位概念も立派な descriptor といえるので，このこと自体に問題はない．しかしながら，同様の理由で，クエリの特徴をより明確に示す descriptor は，高頻度で出現しない傾向をもつ．これらに対する重みを重視するためには，他のクエリ概念との共起性の薄いものを重視する重み計算が必要である．我々は，現在，$pf * icf$[8] や相互情報量を利用した，より柔軟な重み付け処理を実装し，評価実験を予定している．

以下，適切な特徴情報がうまく得られなかった例について考察する．

表 2: 「ハンバーグ」に関する descriptive 知識

descriptor	重み	妥当性
もの	0.521	×
形	0.129	×
感じ	0.044	×
料理	0.042	○
肉	0.040	○

(1) 多くは，「ハンバーグのようなもの」「ハンバーグのような形」といった，代名詞や一般的な名詞など，抽象度の高い語句が重みの高い descriptor となったために，概念の説明として不十分となる場合であった（表 2）．このような場合，妥当な知識が抽出できていたにもかかわらず，相対的に低い順位となり，性能に寄与しなかったケースが多い．対策として，不要語リストの利用や，一般性の高い語の順位を下げるようなパラメータを設けることが考えられる．

(2) クエリが多義語であるため，特徴が一定しない場合があった．例として，「シリウス (＝星の名前，人名，馬名)」や「ＡＩ (＝人工知能という技術，映画タイトル，発音記号，頭辞語など)」があげられる．この問題に大しては，descriptor をクラスタリングすることで，多義性を整理することが可能と考えている．

6 あとがき

本論文では，WWW から descriptive な知識を抽出し，提示する手法の試作システム Murasaki について報告した．確率的概念記述に基づく知識ベース構築手法を応用して，クエリを WWW を利用して構築した重み付き descriptor 集合として表現した．実験の結果，MRR 値で 0.8 と有効性の高い評価結果が得られた．このことから，質問応答における定義タスクに対して，従来と比べて効率良く対応できる見通しが得られたと考えている．

今後は，一般性の高い要素への対応や，多義性を持つクエリの descriptive 知識のクラスタリング手法の考案に取り組み，実用的なシステム構築を目指す．

7 おわりに

本論文では，曖昧な質問に対する質問応答処理の一手法として，比喩指標のパターンを手がかりとして，コーパス中から関連した知識を取り出し，それらを統合して回答する処理を提案した．現状の質問応答技術の課題と，比喩認識における「例え」の仕組みについて触れ，提案手法について説明した．本手法に基づき，人手によるシミュレーション調査を行い，評価した結果，本手法の有効性が示唆された．

しかしながら, いくつかの課題も明かになった．今後は，分析中である類似例抽出に関する評価考察を行い，本論文で明らかになった，被喩辞部における一般的な名詞への対応，修飾語句や接頭辞によるキーワードの意味限定への対応，多義語への対応などについて検討していきたい．さらに，検討結果をふまえ，本手法の実装を進め，比喩による質問への応答にも着手する方針である．

参考文献

[1] Ellen M. Voorhees. The Evaluation of Question Answering Systems: Lessons Learned from the TREC QA Track. In *Proc. of the LREC 2002 Workshop on Question Answering — Strategy and Resources*, pp. 1–4, 2002.

[2] Jun'ichi Fukumoto, Tsuneaki Kato, and Fumito Masui. Question Answering Challenge (QAC1) An Evaluation of QA Tasks at the NTCIR Workshop 3. In *Papers from the 2003 AAAI Spring Symposium "New Directions in Question Answering"*, pp. 1–4, 2003.

[3] 角薫, 溝口理一郎. オントロジー工学と HCI を融合した協調型合意形成支援システム．Sig-fai-a201-04, 人工知能基礎論研究会, 2002.

[4] 黒橋禎夫. 大規模テキスト知識ベースに基づく自動質問応答. NLC-2001-73, 信学技報, 2001.

[5] 加藤恒昭, 福本淳一, 桝井文人, 神門典子. 質問応答から対話理解へ –NTCIR QAC Task3 の提案–. 言語処理学会第 10 回年次大会発表論文集, 2004.

[6] 川崎治夫. 言い換えのプログラム理解への応用. 言語処理学会第 7 回年次大会ワークショップ論文集, pp. 89–92, 2001.

[7] Tsuneaki KATO, Mitsunori MATSUSHITA, and Noriko KANDO. A Workshop on Multimodal Summarization for Trend Information. In *Proceedings of NTCIR-5 Workshop Meeting*, pp. 556–563, 2005.

[8] 桝井文人, 福本淳一, 荒木健治. 比喩解釈を目的とする World Wide Web を利用した属性値の適合性判定. 言語処理学会第 11 回年次大会発表論文集, pp. C2–2, 2005.

付録

図 4: Murasaki のクエリ入力インターフェイス

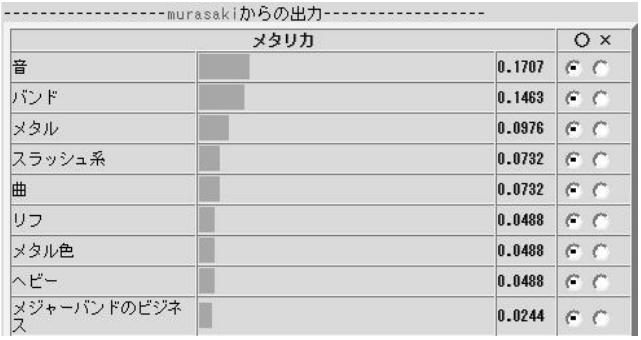

図 5: Murasaki の descriptor 知識可視化画面