Wikipedia Templateから抽出した 意味的関係インスタンスによる質問応答手法

真嘉比 愛　　Stijn De Saeger　　鳥澤 健太郎　　呉 鍾勲　　山本 和英

長岡技術科学大学 電気系
{makabi, yamamoto}@jnlp.org
情報通信研究機構 ユニバーサルコミュニケーション研究所 情報分析研究室
{stijn, torisawa, rovellia}@nict.go.jp

1　はじめに

Wikipedia は語彙網羅性，即時更新性に優れた大規模な知識源であり，WiQA(Question Answering based on Wikipedia) [1]を初めとした数多くの質問応答研究で利用されている [1, 3]．しかしそれらの研究の大半はWikipedia の構造情報 (カテゴリ情報，リダイレクト情報，曖昧さ回避ページ情報，etc.) に着目した手法であり，使用する情報源は Wikipedia 上に存在するものに限定されている．

本稿では，日本語 Web6 億ページの非構造化データと，構造化データである Wikipedia Template の組み合わせからなる質問応答手法を提案する．Wikipedia Template データは「記事名，Template 名，属性名，属性値」で構成されており，本手法では Template 名と属性名の二つ組を“関係名”(e..g Film，出演者)，記事名と属性値の二つ組を“関係インスタンス”(e.g. アバター，サム・ワーシントン) として定義する．これらのデータを用いて質問応答を行うために，例えば

(1) アバターに出ているのは誰ですか？
(2) アバターで出演しているのは誰ですか？
(3) サム・ワーシントンで有名な映画は何？

といった多種多様な質問文と，質問文が表現する関係名「Film 出演者」を正しく対応させる必要がある．そこで本手法では，それぞれの関係名について関係インスタンスが共起する構文パタン (e.g. A に出ている B) を Web6 億ページ中から自動的に獲得し，関係名を表現する代表的な構文パタンのスコアが高くなるよう構文パタンに対し重み付けを行うことで，各関係名ごとに関係名を表現する構文パタンの順にランキングされた集合を構築する．提案手法は，質問が与えられるとその集合をもとに質問文の構文パタンから該当する関係を特定し，回答となる関係インスタンスを提示する．本手法の特徴は，「タイタニックで有名な人は誰？」といった Wikipedia Template 中に回答が明示されない質問に対しても，適切な回答が得られる点である．本システムの出力例を表 1 に示す．

提案手法を用いることで，Web 上にある大量の非構造化データを利用して，構造化データである Wikipedia Template が持つ意味的関係に対し多種多様な換言表現を学習し，与えられた質問に対し高い精度で回答を提示することができる．

2　提案手法

2.1　関係名を表現する構文パタンの取得

まず，Wikipedia 日本語版の全記事 (2011 年 10 月 23 日版)[2]から，{{}} で囲まれておりかつ 2 つ以上属性を持つデータを，パタン照合により WikipediaTemplate データとして自動的に抽出する．その結果，約 7 割の抽出精度で 33,097 個の関係名，4,562,511 個の関係インスタンスを得た．

次に，これらの関係名及び関係インスタンスを用いて，形態素解析器 JUMAN[3] と係り受け解析器 KNP[4]で解析した 6 億ページの Web 文書から，関係インスタンスが一文内で共起する際の構文パタンを抽出する．その際，データ過疎性への対策として以下の条件で関係インスタンスを換言した単語対を構文パタンの検索に用いる．

1. 関係インスタンスをなす 2 つの単語のうちの 1 つを，2 文字よりも長い末尾の文字列で置き換えたも

[1] http://ilps.science.uva.nl/WiQA/Task/index.html

[2] http://dumps.wikimedia.org/jawiki/20111023
[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

表 1: システムの出力例

質問：タイタニックで有名な人は誰？		質問：小学館から出ている作品は何？		質問：大林組が携わった建物は何？	
関係名	出力結果	関係名	出力結果	関係名	出力結果
Film 出演者	レオナルド・ディカプリオ	radiodrama 発売元	BASARA	体育館 施行	名古屋市総合体育館
Film 出演者	ケイト・ウィンスレット	novel 出版社	ふしぎ遊戯	体育館 施行	大阪市中央体育館
Film 監督	ジェームズ・キャメロン	novel 出版社	超時空要塞マクロス	体育館 施行	府中市立総合体育館
Film 音楽	ジェームズ・ホーナー	novel 出版社	ブラック・ラグーン	hotel 設計	ホテルエンパイア
Film 製作	ジョン・ランドー	novel 出版社	ハヤテのごとく！	ダム 施行業者	上郷ダム
Film 製作総指揮	レイ・サンキーニ	novel 出版社	ケータイ少女	ダム 施工業者	内村ダム
Film 編集	コンラッド・パフ	novel 出版社	人類は衰退しました	ダム 施工業者	美和ダム
Film 編集	リチャード・A・ハリス	novel 出版社	ホットギミック	ダム 施工業者	八汐ダム
Film 製作国	アメリカ合衆国	novel 出版社	神のみぞ知るセカイ	ダム 施工業者	七色ダム
Film 配給	20 世紀フォックス	novel 出版社	レヴィアタンの恋人	ダム 施工業者	天ヶ瀬ダム

の (e.g. 関係インスタンス：NICT，京都府精華町 → NICT，精華町)

2. Wikipedia のリダイレクト情報を用いて関係インスタンスの換言を行った単語対 (e.g. 関係インスタンス：ヤマト運輸，トラック輸送→ クロネコヤマト，トラック輸送).
3. 高度言語情報融合フォーラム (ALAGIN)[5]の日本語異表記対データベース (Version 1.1，リソース ID A-7) と基本的意味関係の事例ベース (Version 1.3，リソース ID A-9) 中にある同義語対を用いて関係インスタンスの換言を行った単語対.

これらの処理により，10,468 個の関係名に対して 2,946,385 個の構文パタンが得られた.

残りの 22,629 個の関係名は，

1. 関係インスタンスが非常に疎な関係名であり，Web6 億ページ中に関係名を表現する構文がない (e.g. 関係名：天体軌道 近日点距離).
2. 関係インスタンスの情報が記号で表現されていたため，Web6 億ページ中のデータと一致しない (e.g. 関係名：駅情報 社色，関係インスタンス：大阪駅，#0072bc).
3. Wikipedia Template の抽出誤りによるもので，関係名として有効でない.

といった理由から，構文パタンが抽出できなかった．これらのデータは処理対象から除外して扱っている.

また例外処理として，全ての関係名に対し「A の (属性名) は B 」(e.g. A の出演者は B) というパタンを加えた．活用の違い等を無視するために，得られた構文パタンから内容語 (未定義語，名詞，形容詞，動詞の基本形のいずれか) のみを抽出した単語集合を構築し，それらをパタンと見なす (e.g. A に出演した B → A B 出演)．更に抽出したパタン群を関係名を表現する代表的なパタンのスコアが高くなるよう重み付けする．ここでは情報検索における一般的な特徴単語の重み付け手法である TF-IDF 法に則り，関係名 r におけるパタン p の重み $\mathrm{wgt}(r,p)$ を式 (1) で定義する.

$$\mathrm{wgt}(r,p) = \frac{n_{p,r}}{\sum_x n_{x,r}} \times \left\{ \log_2 \frac{|R|}{|R_p|} + 1 \right\} \qquad (1)$$

ここで $n_{p,r}$ は関係名 r 中の全インスタンスに対するパタン p の共起頻度，$\sum_x n_{x,r}$ は関係名 r 中のパタン総数，$|R|$ は関係名の総数，$|R_p|$ はパタン p がその関係インスタンスと共起する関係名の総数をそれぞれ表している．例外的に導入した「A の (属性名) は B」というパタンに対しては，関係名中で最も高かったスコアと同等の値がふられている．最後に，こうして得られた重み付きパタン群を高度言語情報融合フォーラムの動詞含意関係データベース (Version 1.3.1，リソース ID A-2)[2] を用いて含意関係にある語で換言し，パタンの拡張を行う (e.g. A B 出演 → A B 演じる)．拡張で得られたパタンには，拡張元のパタンと同等のスコアがふられている.

2.2 質問応答部の処理

提案手法では，まず JUMAN と KNP を用いて入力された質問文から質問のトピック候補となる名詞群，およびそれらをつなぐ構文パタンを取得する．構文パタン抽出のために質問文に含まれる疑問代名詞も考慮する．例えば，「黒澤明が監督した映画は何？」という質問から下記のパタンと名詞対が得られる.

構文パタン	名詞対
「A が監督した B」	A:黒澤明，B:映画
「A が監督した映画は B」	A:黒澤明，B:何
「A は B」	A:映画，B:何

これらの名詞に現れる語 (黒澤明，映画，何) を以下では質問のトピック候補と呼ぶ．次にこのトピック候補を片方の名詞として持つような関係インスタンスを全て抽出し，もう片方の名詞を回答の候補と見なす．この回

[5] http://www.alagin.jp/

答候補全ての集合を W_{cand} で表す (ちなみに以上の例にある非固有名詞「映画」「何」のような語は Wikipedia から抽出された関係インスタンスに含まれる可能性が低いので，大きな問題とはならないことに注意されたい)．この段階で回答候補となる単語集合が得られない場合は，Wikipedia Template 中に該当する回答がないと判断して処理を終了する．

次に質問文から取得した構文パタンから内容語以外を除外したパタンを生成する．また，上記 2.1 節と同様に動詞含意関係データベースを用いて換言を行い，パタンの拡張を行う．

こうして得られた質問のパタン群を用いて，質問文が表現する関係名に属す回答候補が上位になるよう回答候補集合 W_{cand} をランキングする．具体的には前述したパタンの重みスコア $\mathrm{wgt}(r,p)$ を用いて，W_{cand} と対応するトピック候補からなる関係インスタンスを持つ各関係名 r に対し，質問文から得られた構文パタン集合 P のスコアの合計値 $\mathrm{score}(r,P)$ を求める．

$$\mathrm{score}(r,P) = \sum_{p \in P_r} \mathrm{wgt}(r,p) \tag{2}$$

回答候補 $w \in W_{cand}$ とトピック候補 n からなる関係インスタンスを $ins(w,n)$ としたとき，$P_r \subseteq P$ は関係 r を持つ関係インスタンス $ins(w,n)$ と質問文中で共起するパタンの集合である．例えば「黒澤明が監督したのは何？」という質問に対し，回答候補「羅生門」とトピック候補「黒澤明」からなる関係インスタンス ins(羅生門，黒澤明) が関係名「Film 監督」を持つとき，パタン「A B 監督」とその拡張パタンが $P_{Film\ 監督}$ に加えられる．

最終的に，全回答候補 $w \in W_{cand}$ に関して，w が属する関係名 r のスコア $\mathrm{score}(r,P)$ によりランキングを行って提示する．トピック候補の名詞 n と同一の関係 r を持つ回答候補 w_1 と w_2 はスコアが同値となり，出力での上下が無作為に決められる．パタンに対応する関係名を 1 つも得られなかった場合は，回答候補集合 W_{cand} を無作為にランキングして提示する．

3 評価実験

本提案手法を用いて，質問文に対し適切な回答を提示出来るか評価実験を行った．以下でその詳細について述べる．

3.1 評価用データの用意

評価実験では，Wikipedia の網羅性による影響をなくすため，回答が Wikipedia Template 内に確実に存在する質問セットを用意した．そのため，本手法による評価は実用上の性能評価とは異なる．まず，Wikipedia Template 中に関係インスタンスが 100 個以上存在する 3,952 個の関係名中から無作為に 100 個を抽出し，それぞれの関係名について無作為に 10 個ずつ関係インスタンスを用意する．この際，以下の条件を満たす関係名・関係インスタンスは除外した．

1. 関係インスタンスの単語の大半が外国語表記になるもの (e.g. 関係名：大統領 各国語表記)
2. 関係インスタンスの単語対が同じ単語になってしまうもの (e.g. 関係名：モデル モデル名，関係インスタンス：相沢紗世，相沢紗世)．
3. 関係インスタンスの単語が一文節を超える固有名詞であるもの (e.g. 関係インスタンス：ハリーポッターと賢者の石，ダニエル･ラドクリフ)

3. に関しては，提案手法で扱う関係インスタンスを一文節に限定しているため処理することができない．今後これらにも対応できるよう提案手法を拡張する予定である．

次に，著者以外の 3 名でそれぞれの関係名について任意に 3 つずつ関係インスタンスを選択し，選択した関係インスタンスの単語対を用いてどちらかが質問，どちらかが回答となるように質問回答の対を作成する．質問を作成する上で以下に示すような条件を設けた．

1. 質問文中に用いる関係名は 1 つに限定する (例えば「アバターを監督し，ゴールデングローブ賞を受賞した 人は誰ですか」といった質問は作成しない)．
2. 1 つの関係名について作成される質問 (3 個) は全て言い回しを変え，単語を並び替えただけ／語尾を少々変えただけの質問にしない．

最終的に得られた 900 個の質問中から重複した質問を削除し，計 893 個の評価セットを用意した．評価実験では，質問から想起できる回答が複数ある場合 (e.g. アバターに出演していたのは誰ですか→サム・ワーシントン，シガニー・ウィーバー，...)，それらの中からどの関係インスタンスが選択されても正解とした．

3.2 実験方法

用意した 893 個の質問に対し提案手法がどれだけの精度で正しい回答を返せるか実験を行う．より具体的には，システムが提示する回答の上位 N (N=1,3,5,10) 件以内の精度及び MAP (Mean Average Precision) を計算する．ここで MAP とは，質問ごとの平均精度を平均

したシステムの検索性能に対する評価尺度である．以下に式を示す．

$$\mathrm{MAP} = \frac{1}{|Q|}\sum_{q\in Q}^{|Q|}\frac{\sum_{k=1}^{n}(\mathrm{Prec}(k)\times \mathrm{rel}(k))}{|A_q|} \quad (3)$$

ここで Q は質問の集合，$|A_q|$ は質問 q の適合回答数，n は質問 q の全回答数，$\mathrm{Prec}(k)$ はトップから k 番目までにおける精度，$\mathrm{rel}(k)$ は k 番目の回答が正解だった場合 1 を，不正解だった場合 0 を返す 2 値符号である．

この結果を以下に示す 2 つのベースライン手法と比較する．

1. 構文パタンを利用した関係名の特定を行わず，回答候補集合 W_{cand} 中から無作為に回答を提示する手法 (比較手法 1)
2. データ過疎性を考慮した構文パタンの抽象化及び動詞含意関係データベースによるパタンの拡張を行わず，構文パタンをそのまま用いる手法 (比較手法 2)

比較手法 1 の結果はタスクの難易度を示すものであり，提案手法を比較手法 1 と比較することで，質問応答における構文パタンを利用した関係名特定の有効性を確認する．また提案手法と比較手法 2 を比較することで，データ過疎性を考慮したパタンの抽象化，および動詞含意関係データベースを用いたパタン換言の有効性を確認する．

3.3 実験結果

実験結果を表 3.3 に示す．

表 2: 評価実験の結果

	比較手法 1	比較手法 2	提案手法
精度@1 (%)	31.3	47.9	53.1
精度@3 (%)	45.3	58.3	65.2
精度@5 (%)	52.5	64.4	71.2
精度@10 (%)	64.3	71.8	76.8
MAP (%)	33.9	52.1	65.7

提案手法は出力上位 10 件以内における精度が 76.8%，MAP 値が 65.7% という結果を示した．質問の構文パタンから回答を含む関係名を特定する手法 (比較手法 2 と提案手法) は，質問の名詞情報のみを利用した比較手法 1 に対し，精度@10 でそれぞれ+7.5 ポイント，+12.5 ポイント，MAP 値でそれぞれ+18.2 ポイント，+31.8 ポイントの向上を見せ，非構造化データから得られた構文パタンを利用することで，構造化データ中から必要な情報を抽出できることが示された．よって，非構造化データと構造化データを組み合わせ，大規模な Web データを用いて Wikipedia 中で定義される関係名を表現する多種多様な構文パタンを自動的に学習する質問応答手法は有効といえる．またデータ過疎性を考慮して構文パタンを内容語に抽象化し，更に動詞含意関係データベースを用いて換言を行った提案手法は，構文パタンをそのまま利用した比較手法 2 に対して精度@10 で+5.0 ポイント，MAP 値で+13.6 ポイントの向上を見せ，パタン抽象化および動詞含意関係データベースを用いたパタン換言の有効性が確認できた．

本手法による関係名抽出がうまくいかなかった原因としては，処理の途中で質問文に含まれる回答の上位語や疑問代名詞を失ってしまったことによる抽出誤り (e.g. 男女共学で学ぶ幼稚園はどこですか→ A B 学ぶ) が挙げられる．これらの問題点を改善する案として，回答候補集合の上位語を特定する手法や，質問文に含まれる疑問代名詞情報を利用して回答となる語の属性を限定する (e.g. 誰→属性：人) 手法が考えられ，今後提案手法を拡張する予定である．

4 おわりに

本稿では，Wikipedia Template のデータに対し関係名と関係インスタンスを定義し，関係名ごとに関係インスタンスが共起する構文パタンを Web6 億ページ中から自動的に獲得した．そして獲得した構文パタンと非構造化データを用いて，構造化データ中から必要な情報を取得する手法を提案した．実験から，提案手法は検索結果上位 10 件における精度 76.8%，MAP 値 65.73% となり，提案手法の有効性を確認した．

参考文献

[1] D. Buscaldi and P. Rosso. Mining knowledge from Wikipedia for the question answering task. In *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 727–730, 2006.

[2] C. Hashimoto, K. Torisawa, and S. De Saeger. Extracting paraphrases from definition sentences on the Web. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011)*, pp. 1087–1097, 2011.

[3] A. Kalyanpur, J. Murdock, J. Fan, and C. Welty. Leveraging community-built knowledge for type coercion in question answering. *The Semantic Web–ISWC 2011*, pp. 144–156, 2011.