機械翻訳と Web による固有名詞の対訳を併用した英日言語横断質問応答システム

川岸 将実 † 森 辰則 ‡
† 横浜国立大学 大学院 環境情報学府 ‡ 横浜国立大学 大学院 環境情報研究院
E-mail: {kawagisi,mori}@forest.eis.ynu.ac.jp

1 はじめに

近年，文書情報に対するアクセス技術として，質問応答が注目されている．質問応答は，利用者が与えた自然言語の質問文に対し，その答を知識源となる大量の文書集合から見つける技術である．

知識源となる文書は種々の言語で書かれているため，単言語だけではなく言語を横断するような検索技術が求められてきている．本稿では，上記の背景の下，質問文が英語，知識源が日本語である英日言語横断質問応答の一手法について提案する．質問文を機械翻訳するとともに，その不備を補うために，Web に存在する対訳情報を用いて固有名詞の翻訳を行なう．これにより機械翻訳の対訳辞書にない語に対応することが可能と考える．また，パタンに基づいた原言語での質問文タイプの決定を行ない，翻訳誤りに対処する手法を提案する．

2 関連研究

言語横断質問応答では，言語依存の質問応答を用いるか，言語非依存の質問応答を用いるかという問題がある．

言語依存の質問応答を流用したシステムとして，関根は英語とヒンディ語における言語横断質問応答について報告している [4]．このシステムでは既存の質問解析部分を流用しており，解決定には知識源側言語のシステムを用いているが，簡単な処理である程度の精度が得られているとしている．

佐々木は言語横断質問応答に対する新しいアプローチとして，機械学習を用いて，質問文の特徴，文書の特徴から質問文タイプを用いずに解答を抽出する手法を提案している [7]．この手法では質問文タイプを用いることをせず，かつ機械学習を用いて解を抽出するため，新たに知識源側の質問応答システムを用意する必要がない．しかし，この方法では問題と解を対応付けた大量のデータを用意しなくてはならない．

我々は日本語に依存した質問応答システムを開発してきており [1]，日本語質問応答システムを基本としたシステムを提案する．英語の質問文を日本語に翻訳したものを入力とみなし，日本語質問応答システムを用いる．

固有名詞の翻訳に関して，辻らはサーチエンジンを利用した人名の翻字手法を提案している [3]．これは，対訳辞書から翻字パタンを学習し，複数の日本語訳候補をヒット件数を元に決定する手法である．一方，我々の提案手法は，英語表記に対応する片仮名表記を見つける翻字に留まらず，日本語のローマ字表記に対応する漢字表記や，英語名称に対応する (逐語訳ではない) 正式な日本語名称を見つけることができる点に特徴がある．

3 提案手法

我々の日本語質問応答は，質問文解析，文書検索，パッセージ抽出，命題照合の４つのモジュールからなっている．質問文解析では，質問文の特徴となる語（以下，キーワードと呼ぶ）と質問が何について尋ねているのか（以下，質問文タイプと呼ぶ）を決定する．文書検索では，キーワードを手がかりに質問に関連する文書を検索する．パッセージ抽出では，検索文書の中から，正解を含む可能性の高い連続する数文（パッセージ）を抽出する．そして命題照合において，質問文タイプ，質問文の解析結果を手がかりに解候補を決定する．この解候補は質問文解析の結果などからスコアリングを行ない，スコアの高い方から優先して出力される．このシステムでは，文書検索モジュール以外は日本語依存となっている．

英日言語横断質問応答とは，質問文が英語で知識源となる文書が日本語の場合の質問応答のことをさす．そのため，パッセージ抽出，命題照合においては既存の日本語質問応答システムのものをそのまま利用することができる．そのため，質問文解析部を変更することで英日言語横断質問応答が可能と考える．

質問文解析部の変更にあたっては以下のような手法が考えられる．

- 質問文（英文）を機械翻訳して，質問文解析モジュールの入力とする．
- 英文のまま質問文解析し，質問文タイプとキーワードを決定する．キーワードは対訳辞書等を用いて翻訳する．

前者は，既存の機械翻訳システムを前提とすれば処理は簡単だが，翻訳できなかった語（以下未翻訳語）の存在や，翻訳誤りのために質問文を正しく解析できなくなり，質問文タイプが正しく決定できない場合がある．一方後者は，質問文の構文情報が使用できなくなり命題照合に支障をきたすため，既存のシステムが活かせなくなってしまう．

そこで，本稿では前者と後者を併用した質問文解析を提案する．具体的には，未翻訳語に関しては，Web 上の情報を用いて対訳語を抽出し，質問文タイプは原文から決定する．これにより，質問文タイプを正しく選択でき，かつ，質問文の構文情報を使用でき，既存のシステムを有効に活かせる．システムの概略を図 1に示す．

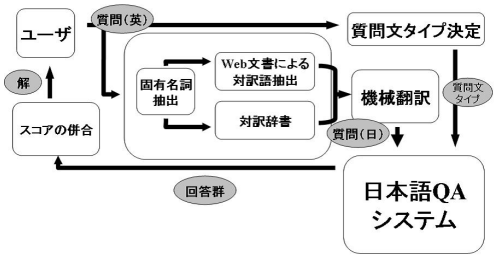

図 1: 提案する英日言語横断質問応答システム

ユーザが英語の質問文を入力すると，機械翻訳システムを用いて日本語に翻訳すると同時に，英語の質問文に

基づいて質問文タイプを決定する．そして，翻訳された質問文と質問文タイプを質問文解析へと渡す．以後の処理は日本語質問応答と同じ流れになる．

Web を用いた固有名詞の翻訳は節 3.1で，原言語による質問文タイプ決定は節 3.2で詳しく述べる．

3.1 Web を用いた固有名詞の翻訳

先に述べたように，ユーザが英語の質問文を入力すると，機械翻訳システムを用いて日本語へと翻訳される．しかし，翻訳されない語があると，未翻訳語のままキーワードとして判別されてしまい，正しい関連文書を検索することができなくなる．結果として見当外れな解を抽出してしまうことになる．特に，固有名詞が翻訳されずに原文のまま残ってしまう傾向にある．この問題を解決するために， Web から固有名詞の対訳情報を抽出する手法を提案する．特に本手法では，従来扱われてきた翻字に留まらず，日本語のローマ字表記に対応する漢字表記や，英語名称に対応する (逐語訳ではない) 正式な日本語名称を見つけることができる点に特徴がある．

まず，質問文から固有名詞と思われる単語を以下の判別法に従い抽出する．この時，日本語（ローマ字列）と思われる単語と，それ以外の（英語の）固有名詞が区別される．

- ローマ字かな変換できるかどうかの判定を行ない，かな文字で 2 文字以上に変換できる場合，その単語を日本語とする．ただし，単語が小文字で始まっていて，かつ，翻訳辞書にその語がある場合は除外する．
- 大文字で始まっている単語が連続した場合，そのまとまりを英語もしくは日本語の固有名詞とする．
- 上記の条件を両方満たす場合は，日本語単語のローマ字表記であると判断する．

日本語を優先させるのは，それ以外の固有名詞の翻訳法よりも信頼性が高いためである．

また，翻訳を行なう際の情報源として， Google Web APIs[1]を用い， Snippet （Web 文書の数文程度の要約）から対訳語（日本語表記）を抽出する．検索語には英質問文から抽出した語を使用する．

3.1.1 日本語由来の固有名詞の翻訳

日本語（ローマ字列）と判別された語に対しては，訳語間に存在する音を介した関係があるので，日本語表記の読みに着目して，以下の手順で処理を行なう．

1. 抽出された語（ローマ字列）を検索語（以下クエリとする）とし， Web 検索を行なう．
2. 取得した Snippet から，特殊文字・英記号を除去する．
3. 整形された Snippet に対し日本語形態素解析[2]を行ない，形態素ごとに，読みの情報を元にローマ字変換を行なう．
4. クエリ全体とローマ字表記が一致した形態素列を対訳語として抽出する．表層表現が異なる語がある場合は，それらをすべて抽出する．

検索対象となる Web 文書は日本語で記述されたものを指定する．これは，ローマ字列をクエリとして検索した結果得られた日本語の Web 文書には，それに対応する元の日本語表記が含まれている可能性が高いという仮定に基づくものである．形態素の読みの情報を利用することで，従来手法では出来なかった漢字交じりの対訳が可能となる．

[1] `http://www.google.com/apis/`

[2] 形態素解析器は JUMAN3.61 を用いた．

3.1.2 英語の固有名詞の翻訳

英語の固有名詞に対する翻訳の場合，訳語間に音を介した関係がないので読みに基づく手法は使えない．そのため英語の翻訳はパタンマッチのみで行なう．この時，同じテキスト内に英語の対訳となる日本語がある場合，その語の近くに対訳語があるという仮定に基づいた処理を行なう．予備実験を行なった結果，近くに存在する場合は何らかの区切り文字 (() ;,/ など) で分けられていることが多いことがわかった．また，日本語列の場合区切り文字がない場合，どこで切れるか判別しにくいため，区切り文字があるもののみに限定した．実際の手順は以下のようになる．

1. 抽出された語をクエリとし， Web 検索を行なう．
2. 取得した Snippet から特殊記号を除去する．
3. 整形文中で，以下の条件にあてはまる語を対訳語としてすべて抽出する．
   - (D)?SWDTWD
   - DTWDSWD

ここで， D は区切り文字， SW は翻訳すべき語， TW は抽出すべき対訳語である．また， ()? は括弧内が省略可能であることを表す．パタンマッチにより切り出しているため，漢字交じりの対訳語が抽出できる．

3.1.3 対訳語が複数ある場合の処理

複数の対訳語候補が得られた場合，それぞれの候補について質問文を作成する．例えば，`"When did the Battle of Sekigahara begin?"`という質問に対しては，`Sekigahara`が日本語として判別され，対訳語として { 関ケ原, 関ヶ原, せきがはら, 関が原 } が抽出される．これらの対訳語に対してそれぞれ質問文を作成して質問応答を行ない，解候補のリストをスコアに従って併合する．このとき，違う質問文から同一の解候補が出てきた場合は，それぞれのスコアを比較し，高いスコアの方をその解候補のスコアとする．

3.2 質問文タイプの同定

質問文タイプは，疑問詞に注目したパタンに基づく推定規則により決定する．これは，英語は疑問詞がはっきりしているため，表層から容易にタイプを決定できるからである．例えば，文頭に`When`がくれば日付を問う`date`となり，`Where`がくれば場所を問う`LOCATION`となる．このような正規表現による推定規則を 20 パタン用意し，質問文タイプ同定を行なった．

4 評価実験

4.1 使用データ

NTCIR5 CLQA1 EJ-Task[2] のサンプルテストセット 300 問を開発用として使用した．評価用には NTCIR5 CLQA1 EJ-Task で使用された 200 問を使用した．また，節 4.2で指標として利用している日本語質問応答の質問文として，対になっている CLQA1 JE-Task （日本語の質問に対し，英語の知識源から解答を抽出，提示するもの）のテストセット 200 問を使用した．

また，知識源は読売新聞の 2000 年， 2001 年の記事 2 年分，機械翻訳は市販の翻訳ソフトウェア [5] を，対訳辞書としては EDR 対訳辞書 [6] を使用した．

4.2 システム性能評価

英日言語横断質問応答の評価基準として，Accuracy[3]とMRR 値[4]を用いた．結果を表 1に示す．ここで，Acc は

表 1: システムの性能評価

	Acc	MRR	Acc+U	MRR+U
MT	0.065	0.081	0.090	0.116
EJQA	0.125	0.141	0.155	0.190
JQA	0.170	0.239	0.265	0.373

Acc:Accuracy による評価
MRR:MRR 値による評価
MT: 機械翻訳と対訳辞書のみ使用
EJQA: 提案手法をすべて使用
JQA: 日本語 QA

Accuracy による評価，MRR は MRR 値による評価，+U は Unsupported(解は一致，文書 ID は異なる) を加えたものである．また，EJQA は提案手法をすべて使用したもの，JQA は日本語 QA である．「日本語 QA」は完全に翻訳が成功した場合の評価であり精度の上限となる．

機械翻訳と外部辞書のみを用いたものに比べ，提案手法のすべての機能を使用したものの性能がよくなっていることがわかる．全ての提案手法を適用したシステムの出力と日本語 QA システムの出力を比較すると，日本語 QA の約 5/7 の精度となった．しかし，日本語 QA で正解できていない質問を正解できているものなどがあった．

4.3 固有名詞の翻訳性能評価

節 3.1の手法により取得した日本語列と英語列それぞれについて正しく翻訳できたかを評価した．日本語の異なり語 112 語，英語の異なり語 97 語についてそれぞれ対訳情報抽出処理を行なった．

また，評価尺度を以下のように定義し，その結果を表 2に示す．再現率で対訳辞書にない単語としているのは，翻字処理に対訳辞書を使用しているためである．

翻字対象の同定に関する尺度

$$\text{再現率}(R) = \frac{\text{辞書にない語で正しく同定された翻字対象単語列数}}{\text{対訳辞書にない翻字すべき単語列数}}$$

$$\text{適合率}(P) = \frac{\text{辞書にない語で正しく同定された翻字対象単語列数}}{\text{翻字対象として同定した単語列数}}$$

翻字に関する尺度

$$\text{ヒット率}(H) = \frac{\text{翻字候補が得られた語数}}{\text{翻字対象として同定した単語列数}}$$

$$\text{精度}(A) = \frac{\text{正しく同定された翻字対象単語列数}}{\text{翻字候補が得られた語数}}$$

表 2より，日本語と判別された文字列については，再現率が 61.4% と，翻字すべき単語に関して正しく対訳語を抽出できると言えるが，適合率が 31.3% と低く，翻字をしなくてもいい語まで翻字対象として同定していると言える．また，ヒット率，精度が高いことから提案手法により機械翻訳の辞書にない語を適切に補えているといえる．英語と判別された文字列については，再現率，適合率ともに低い結果となり，翻字対象文字列の取得がうまくいっていないことがわかった．また，ヒット率，精度ともに低い結果となり，英語と判別された文字列については，提案手法によって機械翻訳の辞書にない語を適切に補完できなかった．

表 2: 対訳語抽出の評価

	R	P	H	A
日本語	61.4%	31.3%	67.0%	76.0%
英語	7.14%	3.09%	20.6%	30.0%

R: 再現率　　P: 適合率
H: ヒット率　　A: 精度

4.4 原言語での質問文タイプ同定の効果

節 3.2の手法により，質問文から抽出した質問文タイプの精度を評価した．結果を見ると，機械翻訳での質問文タイプ同定精度は 55.6%，原言語の精度は 79.5% となり，原言語でタイプを決定した方が精度がよいことがわかった．特に，数量表現に関する質問においては機械翻訳の結果ではほとんど正しく同定できないのに対し，原言語でのタイプ同定を用いると比較的高い精度で取れることがわかった．

4.5 NTCIR5 CLQA-1 による評価

我々は提案システムにより，NTCIR5 CLQA-1 に参加した．その結果を図 2に示す．

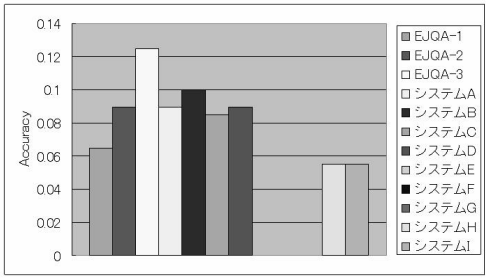

図 2: NTCIR-5 CLQA1 EJ-Task の結果

ここで，EJQA-1 は機械翻訳のみ，EJQA-2 は機械翻訳＋原言語での質問文タイプ同定，EJQA-3 は EJQA-2 に Web による固有名詞の対訳を加えたものである．また，システム A-G は他の参加システムである．A-C は英語で質問文解析を行ない，キーワードを対訳して CLIR を行ない，知識源側の質問応答システムを用いるものである．D は質問文を機械翻訳するとともに原言語で質問文タイプを決定し，日本語質問応答システムの入力とするもので，我々の手法と類似している．E-G は言語依存の質問応答システムを用いず，機械学習による言語非依存の質問応答システムを用いたものである．H,I は質問文を機械翻訳し原言語で質問文タイプを決定するとともに，質問の焦点を原言語で決定し，翻訳したものを日本語質問応答の入力としている．

EJQA-3 は参加システム中最もよい精度となり，提案手法が有効であることを示した．類似した手法の D,H,I に比べて精度が良かったのは，日本語質問応答の精度及び Web による固有名詞の翻訳の効果だと考えられる．

[3] システムは質問に対して解答を一つだけ提示し，全質問に対する正解数の割合を求めるもの

[4] Mean Reciprocal Rank．各問について最上位正解の順位の逆数を評価値とし，それを平均したもの．

5 考察

5.1 日本語の対訳語抽出

翻訳に失敗したものを詳細に調べてみると，以下の3種に大別できる．

1. 英語を日本語として判別している（Morraなど）
2. 形態素解析が誤っている （Meiseiなど）
3. Snippet 中に語が出現していない

1に関しては，辞書に存在していない語を表層のみで日本語か英語かを判別するのは困難なので，処理時間は大きくなるが，日本語として判定した場合の対訳語抽出処理と並行して英語として判定した場合の対訳語抽出処理を行なうことで対応できると考える．2に関しては，形態素解析器の精度の向上が求められる．3に関しては，現在は検索結果の上位 10 件の Snippet のみで処理を行なっているので，実際の HTML ファイルを取得するなど，情報源の拡張を行なうことで対応できると考える．

5.2 英語の対訳語抽出

翻訳に失敗したものを詳細に調べてみると，以下の4種に大別できる．

1. パタンマッチにより抽出された語が不正確である
2. キーワードと対訳語は隣接しているが，パタンにマッチしていない
3. キーワードと対訳語が離れて存在している
4. Snippet 中にキーワードと対訳語のうち片方しか出現していない

1では，パタンマッチにより抽出された文字列が，正解を含むより長いものであったり正解の一部であったりするということが見られた．これは，日本語は表層ではどこまでが一語なのか判別しにくいことに起因していると考えられる．この対処には，形態素解析などを用いて隣接する対訳語を過不足なく抽出する方法が考えられる．4は日本語の場合と同様に，情報源を拡張することで対処できると考える．2，3は，現在の提案手法では抽出することができないので他の方法を考える必要がある．

5.3 複数個対訳語が得られた場合の処理

複数得られた候補の中に正解と不正解が混在する場合が特に英語と判定された文字列において見られた．得られた対訳語が明らかな翻訳間違いの語で作成された質問文で，質問応答するのは適切ではない．さらに，Web 上の情報源は常に同一であると限らないため，時間により，正しい対訳が得られたり得られなかったりということもある．実際に1月 25 日に取得した Snippet で実験を行なった際は，翻訳，Accuracy ともに精度が下がった．

それらに対処する方法として，確信度付きの翻訳辞書を整備する方法が考えられる．抽出した語を辞書の項目から探し，確信度が閾値以上であれば対訳語を提示し，新たな Web 検索は行なわない．閾値以下，あるいは辞書にない場合にのみ Web 検索を行ない，対訳語が抽出された場合に，確信度づけを行なう．これにより，確信度の高い語のみ翻訳候補として使用されると考えられる．

確信度づけの手法としては，サーチエンジンのヒット件数を使用する方法や，質問応答のスコアによる方法が考えられる．

5.4 原言語での質問文タイプ同定

数量表現に関する質問においては機械翻訳の結果ではほとんど正しく同定できないのに対し，原言語でのタイプ同定を用いると比較的高い精度で同定できることがわかった．これは，機械翻訳が`what percent`を「どのパーセント」と誤訳することで，質問文タイプの推定に失敗するためだと考えられる．

6 おわりに

本稿では，機械翻訳と日本語質問応答を利用した英日言語横断質問応答システムを提案し，問題となる箇所に対して二つの解決手法を提案した．

第1に Web 上の情報を用いて，機械翻訳の未翻訳語を翻訳する手法を提案した．質問文から固有名詞と思われる語を抽出し Snippet から対訳語を探すことで，ローマ字列の翻訳に対して有効に働くことがわかった．その一方で英語の固有名詞の翻訳に対してはさらなる検討が必要であることもわかった．

第2に原言語から質問文タイプを決定する手法を検討し，質問応答の精度向上に寄与することを示した．パタンマッチにより，機械翻訳と既存の質問文解析を使用する手法よりも，比較的高い精度で質問文タイプを決定できることがわかった．その一方で同定しづらい質問文タイプや，対応していない質問文タイプがあることがわかった．

この二つの提案手法を同時に用いた場合に，システム全体の精度が最も向上することがわかった．しかしながら，MRR で評価した時に日本語質問応答システムの約 5/7 程度の精度しかないので今後はその差を埋めていく検討をしたいと考えている．具体的には，Web を用いた翻訳手法のさらなる検討，確信度付きの翻訳辞書の整備，質問文タイプ同定の検討などを行なっていきたいと考えている．

参考文献

[1] Tatsunori Mori. Japanese Q/A system using A* search and its improvement: Yokohama national university at QAC2. In Working Notes of the Fourth NTCIR Workshop Meeting, pp. 345–352, 6 2004.

[2] NTCIR5 CLQA-1. NTCIR Workshop5 言語横断質問応答タスク. http://www.slt.atr.jp/CLQA/, 2005.

[3] 辻慶太, 佐藤理史, 影浦峡. 言対訳人名における翻字・サーチエンジンの有効性評価語. 言語処理学会 第11回年次大会 発表論文集, pp. 352–355, 3月 2005.

[4] 関根聡. 言語横断質問応答システム. 言語処理学会 第10回年次大会 発表論文集, pp. 321–324, 3月 2004.

[5] 日本アイ・ビー・エム株式会社. 翻訳の王様バイリンガル Version5, 2002.

[6] 日本電子化辞書研究所. EDR 電子化辞書, 2001.

[7] 佐々木裕. 統合学習による質問応答システムの新しい構成法 ～ＣＬＱＡに向けて. 自然言語処理研究会報告 2004-NL-163, 情報処理学会, 2004.