Vol.2015-NL-223 No.12
2015/9/28

言語横断質問応答に適した機械翻訳評価尺度の検討

杉山 享志朗[1,a)]　水上 雅博[1,b)]　ニュービッグ グラム[1,c)]　吉野 幸一郎[1,d)]
サクティ サクリアニ[1,e)]　戸田 智基[1,f)]　中村 哲[1,g)]

概要：大規模な知識ベースを用いることで、質問応答システムは幅広い質問に回答することができるようになる。しかし、知識ベースは主要な言語に限定されているため、機械翻訳などを用いた言語横断によって質問応答を行う必要がある場合がある。機械翻訳を利用して言語横断を行う際、翻訳性能が質問応答の精度に影響を与えることは明らかである。一般的な機械翻訳は、人間の評価に相関を持つよう設計された自動評価尺度によって評価・最適化されているが、そのような尺度が質問応答に適しているとは限らない。本稿では、複数の翻訳手法を用いて質問応答データセットを作成して質問応答を行い、複数の評価尺度と質問応答精度との関係を調査する。その結果、質問応答精度に影響を与える翻訳の要因や、質問応答精度と相関が高い評価尺度を発見した。

キーワード：質問応答, 機械翻訳, 自動評価尺度

1. はじめに

質問応答は、質問文に対する解答を情報源から検索するタスクであり、一般に、文書、Web ページ、知識ベースなどが情報源として用いられる。こうした情報源は言語によって偏りが存在し、質問文と情報源の言語が異なる場合の質問応答を特に言語横断質問応答と呼ぶ。

単言語での質問応答では、様々な話題の質問に解答するために構造化された大規模な知識ベースを用いる手法が盛んに研究されている。知識ベースは少数の主要な言語にしか存在しないため知識ベースを質問応答に利用する際には言語横断が特に重要となる。こうした言語横断の手段として、機械翻訳が用いられている。

一般的な機械翻訳は人間の利用者を想定しており、人間の評価と高い相関を持つ評価尺度を定義することが重要である。機械翻訳分野では人間の評価と関係が深い機械翻訳結果の特徴をどのように自動評価尺度に反映させるかが広く研究されてきた。しかしながら、人間にとって良い翻訳が必ずしも質問応答に適しているとは限らない。例として、兵藤ら [1] の研究では、文書や Web ページを用いた言語横断質問応答タスクにおいて、コーパスの文全体を用いて学習した翻訳モデルに比べ、機能語を除いて学習した翻訳モデルがより高い質問応答精度を実現している。また、翻訳器を言語横断質問応答に最適化する研究も行われている [2], [3] が、これらの手法は質問と解答の大規模な並列コーパスを必要とするため、多様な言語に適用することが困難である。これらに関連し、言語横断質問応答の精度を向上させる機械翻訳の設計の端緒として、質問応答精度に影響を与える翻訳結果の要因を調査することは有用であると考えられる。

本研究では、知識ベースを用いた言語横断質問応答における翻訳の影響を調査するため、基となる質問応答評価データセットの質問文を様々な翻訳システムを用いて 5 通りに翻訳し、新たに 5 つの評価データセットを作成した。また、作成したデータセットの訳質評価と、それを用いた言語横断質問応答を行い、訳質評価尺度と質問応答精度の関係を調査した。結果として、質問応答精度は、単語の出現頻度を考慮した尺度である NIST スコアと高い相関を示した。この結果は、言語横断質問応答において出現頻度の低い単語が重要な役割を持っていることを示している。加えて、人手による分析の結果、言語横断質問応答に影響を及ぼすいくつかの翻訳要因を特定した。

2. データセット

本研究では、質問応答における翻訳の影響を調査するた

[1] 奈良先端科学技術大学院大学
NAIST, Takayamacho 8916-5, Ikoma, Nara 630–0101, Japan
[a)] sugiyama.kyoshiro.sc7@is.naist.jp
[b)] masahiro-mi@is.naist.jp
[c)] neubig@is.naist.jp
[d)] koichiro@is.naist.jp
[e)] ssakti@is.naist.jp
[f)] tomoki@is.naist.jp
[g)] s-nakamura@is.naist.jp

め、標準的な質問応答評価データセットを基に、人手翻訳と機械翻訳を用いて新たなデータセットを作成した。本節では、このデータセットの作成方法について述べる。

基となるデータセットとして、Free917[4] を用いた。Free917 は Freebase と呼ばれる大規模知識ベースを用いた質問応答のために作成されており、質問応答の研究に広く利用されている [4], [5]。このデータセットは 917 対の質問文と正解で構成される。各正解は、Freebase に入力したときの出力が質問に対する正答となる論理式で与えられる。先行研究 [4] に従い、このデータセットを train セット (512 対)、dev セット (129 対)、test セット (276 対) に分割した。以降、この翻訳前の test セットを OR セットと呼ぶ。

次に、5 種類の翻訳手法による翻訳結果を用いて質問データセットを作成した。まず、OR セットに含まれる質問文を人手で和訳し、日本語質問文と正解論理式の対から成るデータセットを作成した (JA セットと呼ぶ)。次に、JA セットの質問文を以下に述べる各手法で英訳し、翻訳された英語質問文と正解論理式の対から成るデータセットを作成した。各テストセットの例を表 1 に示す。

人手翻訳　翻訳業者に日英翻訳を依頼し、質問文の日英翻訳を行った。これによって作成したデータセットを HT セットと呼ぶ。

商用翻訳システム　Web ページを通して利用できる商用翻訳システムである Google 翻訳 [*1] と Yahoo!翻訳 [*2] を利用して日英翻訳を行った。これらの翻訳システムの詳細は公開されていないが、Google 翻訳は大規模な統計的機械翻訳、Yahoo!翻訳はルールベース機械翻訳を基にしたものと思われる。Google 翻訳の翻訳結果を用いて作成したデータセットを GT セット、Yahoo 翻訳の翻訳結果を用いて作成したものを YT セットと呼ぶ。

Moses　Moses[6] を用いて作成されたフレーズベース機械翻訳を用いて質問文を翻訳した。Moses の学習には、277 万文のコーパスを用いた。Moses による翻訳結果を用いて作成したデータセットを Mo セットと呼ぶ。

Travatar　Travatar[7] によって作成された、構文木ベースの機械翻訳システムを用いて質問文を英訳した。構文木ベースの機械翻訳は、日英・英日翻訳において高い性能を示すことが知られている。学習に用いたデータは Moses と同様である。Travatar による翻訳結果を用いて作成したデータセットを Tra セットと呼ぶ。

3. 質問応答システム

質問応答を行うため、本研究では SEMPRE[5] というフレームワークを利用した。[*3] SEMPRE は、Freebase のような大規模知識ベースを利用する質問応答フレームワークである。本節では、SEMPRE の動作を簡潔に述べ、どのような翻訳が影響を与えると考えられるかを述べる。**図 1** に SEMPRE フレームワークの動作例を示す。

アライメント (Alignment)　SEMPRE は、自然言語のフレーズと対応する論理式のマッピングを必要とする。これをレキシコン (Lexicon) と呼ぶ。レキシコンは大規模なテキストコーパスと知識ベースを用いて構築される。本研究では ClueWeb09[*4][8] と呼ばれるテキストコーパスと Freebase を用いる。アライメント (alignment) と呼ばれる処理では、レキシコンを用いて質問文中のフレーズを論理式に変換する。

アライメントに影響を及ぼすと考えられる翻訳の要因は、単語の変化である。質問文の中の部分文字列はアライメントにおける論理式の選択に用いられるため、誤翻訳された単語はアライメントでの失敗を引き起こすと考えられる。

ブリッジング (Bridging)　知識ベースに入力するクエリを生成するため、隣接した論理式を統合する。ブリッジングは隣接する論理式から述語となる論理式を生成する動作である。図 1 の例では、BarackObama と Type.University から Education が生成されている。

ブリッジングに影響を及ぼすと考えられる翻訳の要因は、語順の変化である。語順が異なるとアライメント

表 1　各テストセットに含まれる質問文と正解論理式の例

Set	Question	Logical form
OR	what is europe 's area	(location.location.area en.europe)
JA	ヨーロッパの面積は	
HT	what is the area of europe	
GT	the area of europe	
YT	the area of europe	
Mo	the area of europe	
Tra	what is the area of europe	

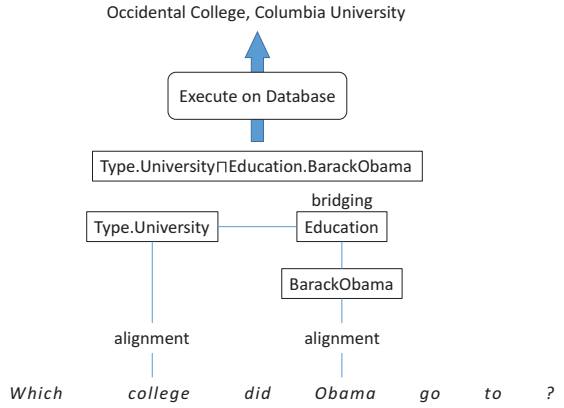

図 1　SEMPRE フレームワークによる質問応答の動作例

[*1] https://translate.google.co.jp/
[*2] http://honyaku.yahoo.co.jp/
[*3] http://nlp.stanford.edu/software/sempre/
[*4] http://www.lemurproject.org/clueweb09.php/

で生成される論理式の順序が変化するため、隣接する論理式の組み合わせが変化する。

スコアリング (Scoring) システムはアライメントとブリッジングから多数の候補を出力し、スコアリングで評価関数に基づいて候補の良さを計算する。質問応答器の学習では、正解を返すことができた候補に高いスコアが付くよう評価関数を最適化する。

言語横断質問応答に最適な評価関数は単言語質問応答と異なる可能性があり、翻訳はこの処理にも影響する可能性がある。しかしながら、言語横断質問応答に最適化するよう学習するためには翻訳された学習データセットが必要であり、その作成には大きなコストがかかるため、本稿ではこれに関する調査は行っていない。

4. 実験設定

本実験では、言語横断質問応答においてどのような翻訳の要因が影響を及ぼすかを調査した。そのために、2 節で述べたデータセットと 3 節で述べた質問応答器を用い、日本語の質問文を翻訳システムで英語の質問文に変換し、英語の単言語質問応答器によって解答を得るという状況を想定した実験を行った。

4.1 実験結果 1：訳質評価

まず、4 つの訳質自動評価尺度 (BLEU+1[9]、WER[10]、NIST[11]、RIBES[12]) と人手による許容性評価 (Acceptability)[13] を用いて作成したデータセットの訳質を評価した。以下に各評価尺度の詳細を示す。

BLEU+1 BLEU[14] は広く利用されている機械翻訳自動評価尺度である。BLEU+1 は BLEU に平滑化を導入し、単文でも極端な結果が出にくいよう拡張した尺度である。これらの尺度は n-gram の一致率に基づいており、評価値は 0 から 1 の実数で、参照訳と完全に一致した場合に 1 となる。

WER Word error rate(WER) は二つの文の間の編集距離を文章長で正規化したものである。評価値は 0 以上の実数で、参照訳よりも長い文長を持つ翻訳仮説が入力された時は 1 を超える場合がある。WER は BLEU と同様に単語単位での一致に着目した評価尺度であるが、BLEU よりも語順に厳格である。WER は誤り率であるため、評価値が低い方が高評価となる。そのため、軸の方向を他の評価尺度と揃えるために 1−WER の値を用いる。

RIBES RIBES は翻訳仮説と参照訳の単語の順位相関係数に基づく尺度である。この尺度は日英・英日のような語順に大きな相違がある言語対に対して他の尺度よりも有効であることが示されている。評価値は 0 から 1 の実数で、大きいほど良い評価となる。

NIST NIST は n-gram の重み付一致率に基づいた尺度である。珍しい n-gram により高い重みが付くように設定される。したがって、頻出する “of” や “in” のような機能語と比べて、内容語の一致に重点をおいた評価を行う。評価値は 0 以上の実数で、大きいほど良い評価となる。

許容性 (Acceptability) 許容性は人間による 5 段階の評価である。この尺度は、1 から 3 の評価で意味的正しさを、3 から 5 の評価で文法的正しさを評価することで、流暢さと自然性を両方考慮した尺度となる。

JA セットの質問文を入力とし、OR セットの質問文を参照訳とした時の各翻訳結果の評価値を**図 2** に示す。NIST スコアは上限が設定されない評価値のため、参照訳と同様の文を入力した場合の評価値で正規化した。

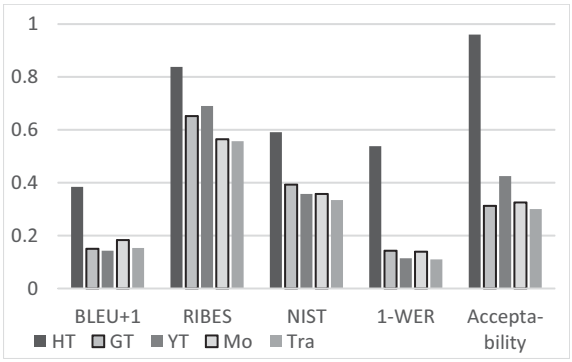

図 2 訳質評価値 (平均)

図 2 より、人手翻訳の訳質は全ての評価尺度において機械翻訳よりも高く、上質であることが読み取れる。次に、GT と YT に着目すると、BLEU と NIST では GT が高く、RIBES と許容性人手評価では YT が高い。これは先行研究 [12] と同様の結果となっており、日英翻訳での RIBES の許容性人手評価との相関が高いという特性が確認された。次節で、このような特性が人間相手ではなく言語理解を行う機械相手でも同様に現れるかどうかを検証する。

4.2 実験結果 2：質問応答結果

次に、作成したデータセットを用いて質問応答を行った。ここで、テストセットとして使用した質問 276 問のうち 12 問で、正解論理式を Freebase に入力した際に出力が得られず、これらを除いた 264 問の結果を用いて行った。

各データセットの質問応答の結果を**図 3** に示す。

図より、元のセット (OR) であっても約 53%の精度に留まっていることがわかる。また、HT セットの精度は機械翻訳で作成したデータセットに比べ高いことが読み取れる。しかしながら、4.1 に示したように高い訳質を持つ HT セットであっても、OR セットと比べると質問応答精度は有意に低いという結果となった ($p < 0.01$)。また、YT は人手許容性評価において GT を上回るが、質問応答精度は GT に劣っている ($p < 0.05$)。これらの結果は、人間にとって

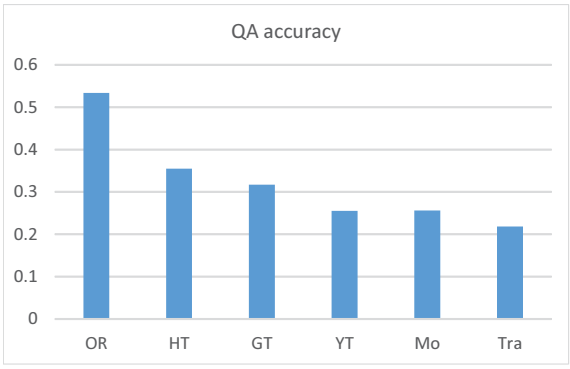

図 3　各データセットにおける質問応答精度

良い翻訳と質問応答に適する翻訳は異なるということを示唆している。次節で、これらの現象について詳細な分析を行う。

5. 分析と考察

5.1 質問応答精度と訳質の関係

質問応答精度に影響を及ぼす翻訳結果の要因をより詳細に分析するため、まず訳質評価値と質問応答精度の相関を文単位で分析した。しかしながら、たとえ質問応答用に作成されたデータセット (OR セット) であっても約半数の質問は正解できておらず、翻訳の結果に関わらず正解することが難しい質問があると考えられる。この影響を考慮に入れるため、質問を 2 つのグループに分けた。「正解グループ」は、OR セットにおいて正解することができた 141 問の翻訳結果 $141 \times 5 = 705$ 問から成るグループであり、「不正解グループ」は残りの 123 問の翻訳結果 $123 \times 5 = 615$ 問から成るグループである。

まず、正解グループにおける質問応答精度と訳質評価値の関係を**図 4** に示す。このグラフにおいて、棒グラフは評価値に対する質問数の分布を表し、折れ線グラフは評価値に対する正答率の変化を表す。例として、BLEU+1 の値が 0.2–0.3 の質問は正解グループの内 30%ほどを占め、それらの質問の正答率は 35%程度である。この図より、本実験に使用した全ての評価尺度は質問応答精度と相関を持ち、言語横断質問応答において訳質は重要であることを示している。また、質問応答精度は NIST スコアと最も高い相関を示した。前述したように、NIST スコアは単語の出現頻度を考慮した尺度であり、機能語よりも内容語を重要視する特徴を持つ。この結果から、内容語が言語横断質問応答において重要な役割を持つと考えられる。これは、内容語が 3 節に述べたアライメントにおける論理式選択において重要であることを考えると自然な結果と言える。また、NIST スコアによってこの影響を自動的に適切に評価できる可能性もこの結果から読み取れる。

一方で、4 節に示したように人手評価との相関が高かった RIBES は、質問応答精度においては相関が低いという結果となった。つまり、大域的な語順が言語横断質問応答のための翻訳にはそれほど重要ではない可能性があると言える。これらの結果を合わせると、語順に影響を受けやすいブリッジングよりも、単語の変化に影響を受けやすいアライメントの方が誤りに敏感であるという可能性が示されている。

次に、不正解グループにおける訳質評価値と質問応答精度の関係を**図 5** に示す。不正解グループにおいては、正解グループと比較すると全ての自動評価尺度において無相関に近く、人手評価のみがわずかに相関を持つという結果となった。この結果は、参照訳での質問応答が失敗している場合、その質問文は負例としてすら不適切であるということを示している。即ち、言語横断質問応答のための翻訳器を評価する際、質問応答器が正解できる質問文を参照訳として評価を行うのが好ましい。[*5]

5.2 事例分析

本節では、翻訳によって質問応答の結果が変化した例を挙げながら、どのような翻訳結果の要因が影響しているかを考察する。

表 2　内容語変化の例

○ OR when was interstate 579 formed
- JA 州間高速道路 579 号が作られたのはいつですか
× HT when was interstate highway 579 made
× GT when is the interstate highway no. 579 has been made
× YT when is it that expressway 579 between states was made
× Mo interstate highway 579) was made when
○ Tra when interstate 579) was built

○ OR who was the librettist for the magic flute
- JA 魔笛の台本を作成したのは誰ですか
× HT who wrote the libretto to the magic flute
× GT who was it that created the script of the magic flute
× YT who is it to have made a script of the the magic flute
× Mo the magic flute scripts who prepared
× Tra who made of magic script
○ - who librettist magic flute

内容語の変化による質問応答結果の変化の例を**表 2** に示す。1 つ目の例では、“interstate 579” という内容語が翻訳によって様々に変化している (“interstate highway 579”、“expressway 579” など)。OR と Tra の文のみが “interstate 579” というフレーズを含んでおり、どちらも正解されている。出力された論理式を見比べると、不正解であった質問文では “interstate 579” のエンティティが含まれておらず、別のエンティティに誤変換されていた。例えば、HT に含まれ

[*5] どちらのグループにおいても質問の分布は似通っており、訳質評価尺度ではこの問題を解決することは難しいと言える。

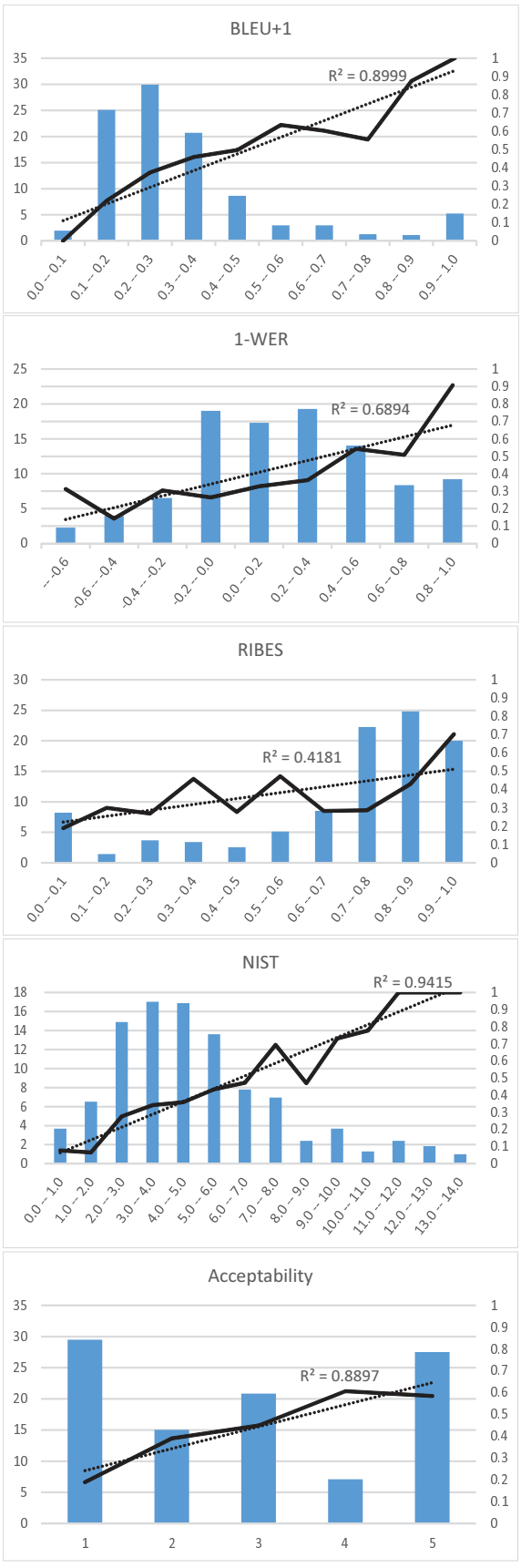

図 4　評価尺度値と質問応答精度の相関 (正解グループ)
横軸：評価値の範囲
棒グラフ (左縦軸)：質問数の割合 (%)
折れ線 (右縦軸)：質問応答精度 (範囲内平均)

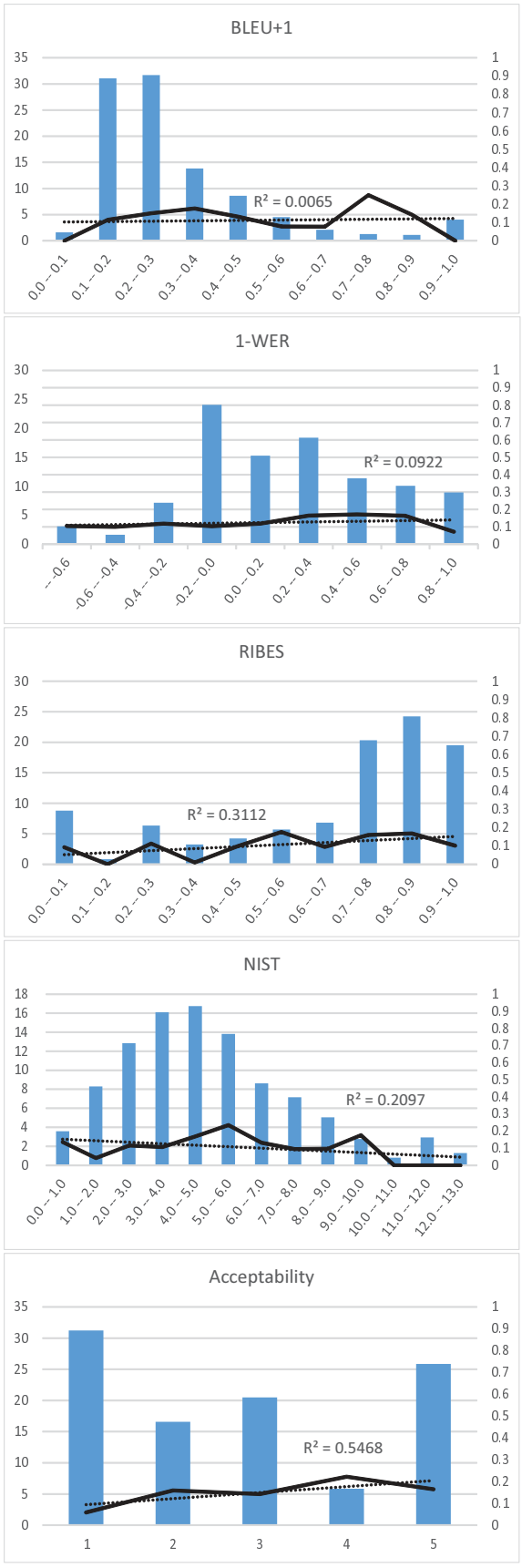

図 5　評価尺度値と質問応答精度の相関 (不正解グループ)
横軸：評価値の範囲
棒グラフ (左縦軸)：質問数の割合 (%)
折れ線 (右縦軸)：質問応答精度 (範囲内平均)

る “interstate highway 579” というフレーズは “interstate highway” という音楽アルバムのエンティティに変換されていた。2 つ目の例も同様に、“librettist” という内容語が翻訳によって様々に変化し、不正解となっている。ここで、新たに “who librettist magic flute” という質問文を作成し質問応答を行ったところ、正解することができた。

このような例は、内容語が変化することでアライメントが失敗し、誤ったエンティティが生成されることが非常に重要な問題であることを示している。この問題は、翻訳の過程で固有表現語彙を素性として組み込み、考慮することで改善できる可能性がある。

表 3 質問タイプを表す語の誤訳の例

○ OR how many religions use the bible
- JA 聖書を使う宗教はいくつありますか
× HT how many religions use sacred scriptures
○ GT how many religions that use the bible
○ YT how many religion to use the bible are there
○ Mo how many pieces of religion, but used the bible
× Tra use the bible religions do you have

○ OR how many tv programs did danny devito produce
- JA ダニー・デヴィートは何件のテレビ番組をプロデュースしましたか
○ HT how many television programs has danny devito produced
× GT danny devito or has produced what review television program
× YT did danni devito produce several tv programs
× Mo what kind of tv programs are produced by danny devito
× Tra danny devito has produced many tv programs

次に、質問タイプを表す語の誤訳が原因となる例を**表 3** に示す。1 つ目の例では、Tra の質問文は OR の質問文の内容語を全て含んでいるにもかかわらず不正解となっている。同じく、2 つ目の例では、“tv (television) programs”、“danny devito”、“produce(d)” の 3 つは全ての翻訳結果に含まれているが、HT 以外は正解できなかった。正解できた質問文とそれ以外の質問文を比較すると、“how many” という質問タイプを表す語を含んでいることが必要であることがわかった。これらの例は、質問タイプを正確に捉え、翻訳することの重要性を示唆している。ここで、質問タイプを表す語は内容語と異なり頻出するため、NIST スコアによって改善することは難しく、質問応答固有の指標が必要であると考えられる。

文法に関連する例を**表 4** に示す。1 つ目の例では、YT 以外の機械翻訳の結果は文法が整っていないにも関わらず全て正解している。一方 2 つ目の例では、OR と HT では文法が正しいにも関わらず不正解となっている。OR と HT の質問応答の結果を調べると、ベーブルースの打撃成績を出力していた。これは、“babe ruth” と “play” が隣接しており、ブリッジングの際に結びついたためと考えられる。

表 4 文法誤りの例

○ OR what library system is the sunset branch library in
- JA サンセット・ブランチ図書館はどの図書館システムに所属しますか
○ HT to what library system does sunset branch library belong
○ GT sunset branch library do you belong to any library system
○ YT which library system does the sunset branch library belong to
○ Mo sunset branch library, which belongs to the library system
○ Tra sunset branch library, belongs to the library system?

× OR what teams did babe ruth play for
- JA ベイブ・ルースはどのチームの選手でしたか
× HT what team did babe ruth play for
○ GT did the players of any team babe ruth
○ YT was babe ruth a player of which team
○ Mo how did babe ruth team
○ Tra babe ruth was a team player

これらの例は、少なくとも Free917 に含まれるような単純な質問においては、語順を正しく捉えることは質問応答精度の向上の観点からは必ずしも重要でないことを示している。

6. 結論

知識ベースを用いた質問応答における翻訳の影響を調査するため、複数の翻訳手法を用いて質問データセットを作成し、質問応答精度の比較を行った。結果として、参照訳が正解できた場合に限るものの、内容語の変化に敏感な NIST スコアが質問応答精度と高い相関を持つことが明らかとなった。さらに事例分析によって、質問応答の結果に影響しうる翻訳の要因である 3 つの要因、内容語、質問タイプ、文法を発見した。これらの結果に基づき、言語横断質問応答タスクにおける機械翻訳の評価において、1) NIST スコアか、もしくは別の、内容語を重視できる訳質評価尺度を用いるべきである。2) 参照訳は質問応答で正解可能な文章で構成するべきである。という 2 つのことがわかった。

この結果は SEMPRE という質問応答フレームワークに基づいた結果であり、SEMPRE は通常の質問応答タスクで高い性能を示しているが、今後別の質問応答システムでも調査を行う予定である。例えば、言い換えを考慮した質問応答フレームワーク [15] を用いた場合、表層的な翻訳の差異に頑健な結果が得られると考えられる。また、この分析結果をレスポンスベース学習の枠組みと組み合わせて機械翻訳の最適化を行うことも検討している。

謝辞

本研究の一部は、NAIST ビッグデータプロジェクトおよびマイクロソフトリサーチ CORE 連携研究プログラムの活動として行ったものです。

参考文献

[1] Hyodo, T. and Akiba, T.: Improving Translation Model for SMT-based Cross Language Question Answering, *Proc. of FIT*, Vol. 8, No. 2, pp. 289–292 (2009).
[2] Riezler, S., Simianer, P. and Haas, C.: Response-based learning for grounded machine translation, *Proc. of ACL* (2014).
[3] Haas, C. and Riezler, S.: Response-based Learning for Machine Translation of Open-domain Database Queries, *Proc. of NAACL HLT*, pp. 1339–1344 (2015).
[4] Cai, Q. and Yates, A.: Large-scale Semantic Parsing via Schema Matching and Lexicon Extension., *Proc. of ACL*, pp. 423–433 (2013).
[5] Berant, J., Chou, A., Frostig, R. and Liang, P.: Semantic Parsing on Freebase from Question-Answer Pairs., *Proc. of EMNLP*, pp. 1533–1544 (2013).
[6] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. et al.: Moses: Open source toolkit for statistical machine translation, *Proc. of ACL*, pp. 177–180 (2007).
[7] Neubig, G.: Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers., *Proc. of ACL*, pp. 91–96 (2013).
[8] Callan, J., Hoy, M., Yoo, C. and Zhao, L.: Clueweb09 data set (2009).
[9] Lin, C.-Y. and Och, F. J.: ORANGE: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation, *Proc. of COLING*, pp. 501–507 (2004).
[10] Leusch, G., Ueffing, N. and Ney, H.: A novel string-to-string distance measure with applications to machine translation evaluation, *Proc. of MT Summit IX*, pp. 240–247 (2003).
[11] Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, *Proc. of HLT*, pp. 138–145 (2002).
[12] Isozaki, H., Hirao, T., Duh, K., Sudoh, K. and Tsukada, H.: Automatic evaluation of translation quality for distant language pairs, *Proc. of EMNLP*, pp. 944–952 (2010).
[13] Goto, I., Chow, K. P., Lu, B., Sumita, E. and Tsou, B. K.: Overview of the patent machine translation task at the NTCIR-10 workshop, *Proc. of NTCIR-10*, pp. 260–286 (2013).
[14] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, *Proc. of ACL*, pp. 311–318 (2002).
[15] Berant, J. and Liang, P.: Semantic parsing via paraphrasing, *Proc. of ACL*, Vol. 7, No. 1, pp. 1415–1425 (2014).