言語横断質問応答に適した機械翻訳評価尺度の調査

杉山享志朗[†]・水上　雅博[†]・Graham Neubig[†]・吉野幸一郎[†]・
鈴木　　優[†]・中村　　哲[†]

質問応答システムが高い精度で幅広い質問に解答するためには，大規模な知識ベースが必要である．しかし，整備されている知識ベースの規模は言語により異なり，小規模の知識ベースしか持たない言語で高精度な質問応答を行うためには，機械翻訳を用いて異なる言語の大規模知識ベースを利用して言語横断質問応答を行う必要がある．ところが，このようなシステムでは機械翻訳システムの翻訳精度が質問応答の精度に影響を与える．一般的に，機械翻訳システムは人間が与える評価と相関を持つ評価尺度により精度が評価されている．そのため，この評価尺度による評価値が高くなるように機械翻訳システムは最適化されている．しかし，質問応答に適した翻訳結果は，人間にとって良い翻訳結果と同一とは限らない．つまり，質問応答システムに適した翻訳システムの評価尺度は，人間の直感に相関する評価尺度とは必ずしも合致しないと考えた．そこで本論文では，複数の翻訳手法を用いて言語横断質問応答データセットを作成し，複数の評価尺度を用いてそれぞれの翻訳結果の精度を評価する．そして，作成したデータセットを用いて言語横断質問応答を行い，質問応答精度と翻訳精度との相関を調査する．これにより，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明らかにする．さらに，自動評価尺度を用いて翻訳結果のリランキングを行うことによって，言語横断質問応答の精度を改善できることを示す．

キーワード：言語横断質問応答，機械翻訳，自動評価尺度

An Investigation of Machine Translation Evaluation Metrics in Cross-lingual Question Answering

Kyoshiro Sugiyama[†], Masahiro Mizukami[†], Graham Neubig[†], Koichiro Yoshino[†],
Yu Suzuki[†] and Satoshi Nakamura[†]

Through using knowledge bases, question answering (QA) systems have come to be able to answer questions accurately over a variety of topics. However, knowledge bases are limited to only a few major languages, and thus it is often necessary to build QA systems that answer questions in one language based on an information source in another language (cross-lingual QA: CLQA). Machine translation (MT) is one tool to achieve CLQA, and it is intuitively clear that a better MT system improves QA accuracy. However, it is not clear whether an MT system that is better for human consumption is also better for CLQA. In this paper, we investigate the relationship

[†] 奈良先端科学技術大学院大学情報科学研究科, Nara Institute of Science and Technology, Graduate School of Information Science

between manual and automatic translation evaluation metrics and CLQA accuracy by creating a data set using both manual and machine translation, and performing CLQA using this created data set. As a result, we find that QA accuracy is closely related with a metric that considers frequency of words, and as a result of manual analysis, we identify two factors of translation results that affect CLQA accuracy. One is mistranslation of content words and another is lack of question type words. In addition, we show that using a metric which has high correlation with CLQA accuracy can improve CLQA accuracy by choosing an appropriate translation result from translation candidates.

Key Words: *Cross-lingual Question Answering, Machine Translation, Evaluation Metrics*

1 はじめに

質問応答とは，入力された質問文に対する解答を出力するタスクであり，一般的に文書，Web ページ，知識ベースなどの情報源から解答を検索することによって実現される．質問応答はその応答の種類によって，事実型（ファクトイド型）質問応答と非事実型（ノンファクトイド型）質問応答に分類され，本研究では事実型質問応答を取り扱う．近年の事実型質問応答では，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情報源として用いる手法が盛んに研究されている (Kiyota, Kurohashi, and Kido 2002; Tunstall-Pedoe 2010; Fader, Zettlemoyer, and Etzioni 2014)．知識ベースは言語によって規模が異なり，言語によっては小規模な知識ベースしか持たない．例えば，Web 上に公開されている知識ベースには Freebase[1] や DBpedia[2] などがあるが，2016 年 2 月現在，英語のみに対応している Freebase に収録されているエンティティが約 5,870 万件，多言語に対応した DBpedia の中で英語で記述されたエンティティが約 377 万件であるのに対し，DBpedia に含まれる英語以外の言語で記述されたエンティティは 1 言語あたり最大 125 万件であり，収録数に大きな差がある．知識ベースの規模は解答可能な質問の数に直結するため，特に言語資源の少ない言語での質問応答では，質問文の言語と異なる言語の情報源を使用する必要がある．このように，質問文と情報源の言語が異なる質問応答を，言語横断質問応答と呼ぶ．

こうした言語横断質問応答を実現する手段として，機械翻訳システムを用いて質問文を知識ベースの言語へ翻訳する手法が挙げられる (Shimizu, Fujii, and Itou 2005; Mori and Kawagishi 2005)．一般的な機械翻訳システムは，人間が高く評価する翻訳を出力することを目的としているが，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．Hyodo ら (Hyodo

[1] https://www.freebase.com/
[2] http://wiki.dbpedia.org/

and Akiba 2009) は，内容語のみからなる翻訳モデルが通常の翻訳モデルよりも良い性能を示したとしている．また，Riezler らの提案した Response-based online learning では，翻訳結果評価関数の重みを学習する際に質問応答の結果を利用することで，言語横断質問応答に成功しやすい翻訳結果を出力する翻訳器を得られることが示されている (Riezler, Simianer, and Haas 2014; Haas and Riezler 2015)．Reponse-based learning では学習時に質問応答を実行して正解できたかを確認する必要があるため，質問と正解の大規模な並列コーパスが必要となり，学習にかかる計算コストも大きい．これに対して，質問応答に成功しやすい文の特徴を明らかにすることができれば，質問応答成功率の高い翻訳結果を出力するよう翻訳器を最適化することが可能となり，効率的に言語横断質問応答の精度を向上させることが可能であると考えられる．さらに，質問と正解の並列コーパスではなく，比較的容易に整備できる対訳コーパスを用いて翻訳器を最適化することができるため，より容易に大規模なデータで学習を行うことができると考えられる．

本研究では，どのような翻訳結果が知識ベースを用いた言語横断質問応答に適しているかを明らかにするため，知識ベースを利用する質問応答システムを用いて 2 つの調査を行う．1 つ目の調査では，言語横断質問応答精度に寄与する翻訳結果の特徴を調べ，2 つ目の調査では，自動評価尺度を用いて翻訳結果のリランキングを行うことによる質問応答精度の変化を調べる．調査を行うため，異なる特徴を持つ様々な翻訳システムを用いて，言語横断質問応答データセットを作成する（3 節）．作成したデータセットに対し，4 節に述べる質問応答システムを用いて質問応答を行い，翻訳精度（5.1 節）と質問応答精度（5.2 節）との関係を分析する（5.3 節）．また，個別の質問応答事例について人手による分析を行い，翻訳結果がどのように質問応答結果に影響するかを考察する（5.4 節）．さらに，5.3 節および 5.4 節における分析結果から明らかとなった，質問応答精度と高い相関を持つ自動評価尺度を利用して， 翻訳 N ベストの中から翻訳結果を選択することによって，質問応答精度がどのように変化するかを調べる（5.5 節）．このようにして得られる知見は日英という言語対に限られたものとなるため，さらに一般化するために様々な言語対で言語横断質問応答を行い，言語対による影響を調査する（5.6 節）．

最後に，言語横断質問応答に適した機械翻訳システムを実際に構築する際に有用な知見をまとめ，今後の展望を述べて本論文の結言とする（6 節）．

2　本調査の概観

本論文では 2 種類の調査を行う．1 つ目は言語横断質問応答に対する翻訳結果の影響に関する調査である．翻訳結果の訳質評価結果と言語横断質問応答精度の関係を求め，その結果からどのような特徴を持つ翻訳結果が言語横断質問応答に適しているかを明らかにする．2 つ目は 1 つ目の調査結果から，言語横断質問応答に適応した翻訳をできるかについての調査である．具

体的には1つ目の調査で質問応答精度との相関が高かったスコアを用いて翻訳結果のリランキングを行い，質問応答精度がどのように変化するかについて調べる．これにより，質問応答精度との相関が高いスコアを用いた翻訳結果によって質問応答精度を改善できることを確認する．

2.1　言語横断質問応答精度に影響する翻訳結果の調査

1つ目の調査では，翻訳結果がどのように言語横断質問応答精度に影響を与えるかを調べる．実験の概要を図1に示す．本調査は，以下の手順で行う．

翻訳を用いたデータセット作成　質問応答に使用されることを前提として作成された英語質問応答データセットを用意し，その質問文を理想的な翻訳結果と仮定する．まず，理想的な英語質問セットを人手で和訳し（図中の「人手翻訳」），日本語質問セットを作成する．続いて，これらの日本語質問セットを，様々な翻訳手法を用いて英訳し（図中の「翻訳手法 $1\sim n$」），英語質問セットを作成する．

翻訳精度測定　作成した英語質問セットについて，複数の評価尺度を用いて翻訳精度の評価を行う（翻訳精度評価システム）．この時，参照訳は理想的な英語質問セットに含まれる質問文とする．

質問応答精度測定　理想的な英語質問セットと，作成した英語質問セットそれぞれについて，同一の質問応答器による質問応答実験を行い，質問応答精度を測定する．

分析　複数の翻訳精度評価尺度それぞれについて，どのような特徴を持つ評価尺度が質問応答精度と高い相関を持つかを調べる．また，質問セット単位ではなく，文単位でも翻訳精度と質問応答精度との相関を分析する．この際，正確な翻訳であっても正解するのが難しいと思われる質問が存在することを考慮するため，理想的な質問文で正解したかどう

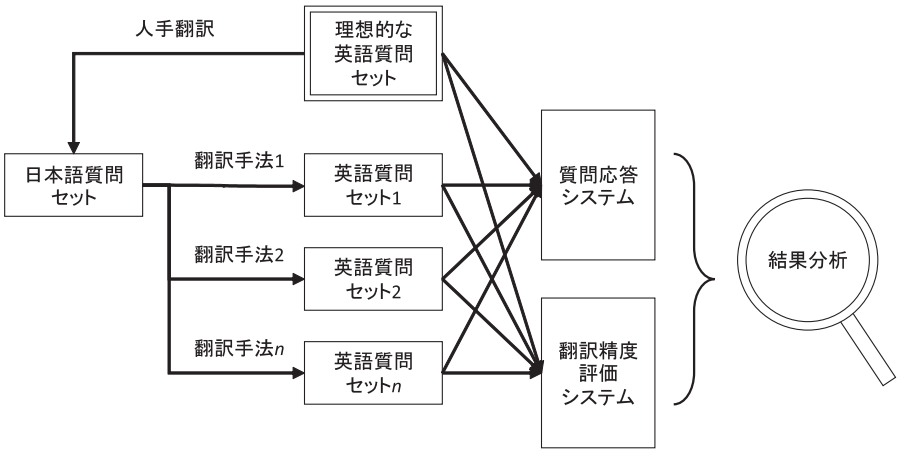

図 1　質問応答精度に影響する翻訳結果の調査実験概要

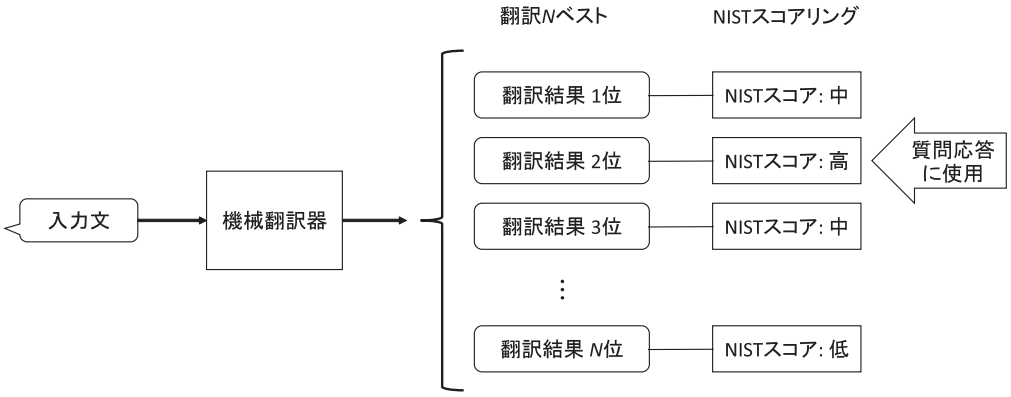

図 2　自動評価尺度を用いた翻訳結果選択

かで 2 グループに分けて分析する．さらに，個別の質問応答事例について人手で確認し，どのような翻訳結果が質問応答の結果を変化させるかを考察する．

2.2　自動評価尺度を用いた翻訳結果選択による質問応答精度改善

前節に述べた実験により得た知見を元に，できる限り既存の資源・システムを用いて言語横断質問応答精度を向上させる可能性を探る．図 2 に調査方法の概要を示す．まず，翻訳結果をもっともらしいものから N 通り出力する N ベスト出力を行う．質問応答精度と高い相関を持つ評価尺度を用いて，N ベストから翻訳結果を選択することによって質問応答精度の向上が見られれば，そのような評価尺度が高くなるように翻訳結果を選択することで質問応答システムの精度が向上することが期待できる．

3　データセット作成

本調査では，日英言語横断質問応答を想定した実験を行うため，基本となる英語質問応答セットとそれを和訳した日本語質問応答セット，日本語質問応答セットから翻訳された英語質問応答セットという 3 種類の質問応答セットを用いた．本節では，これらのデータセットの作成方法について述べる．

3.1　作成手順

基本となる英語質問セットとして, Free917 (Cai and Yates 2013) 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，知識ベースを用いた質問応答の研究に広く利用されている (Cai and Yates 2013; Berant, Chou, Frostig, and Liang

表 1 各質問セットに含まれる質問文と正解クエリの例

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	

2013). このデータセットは 917 対の英語質問文と正解で構成され，各正解は Freebase のクエリの形で与えられている．先行研究 (Cai and Yates 2013) 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．まず，OR セットに含まれる質問文を和訳し，日本語質問セット（JA セット）とした．和訳は，1 名による人手翻訳で行った．なお，今回は日本語の人手翻訳を各セットに対して 1 通りのみ用意するが，この人手翻訳における微妙なニュアンスが以降の機械翻訳に影響を与える可能性がある．次に，JA セットに含まれる質問文を後述する 5 種類の翻訳手法によって翻訳し，各英語質問セット (HT, GT, YT, Mo, Tra) を作成した．質問応答セットの一部を表 1 に示す．

3.2 比較した翻訳手法

本節では，質問セット作成で比較のため用いた 5 種類の翻訳手法について述べる．

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．人手による翻訳結果は人間にとってほぼ最良の翻訳であると考えられ，人間が高く評価する翻訳結果が言語横断質問応答にも適しているかを調べるために HT セットを作成した．

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[3]と Yahoo!翻訳[4]を利用して日英翻訳を行った．これらの枠組みや学習に用いられているデータの詳細は公開されていない．Google 翻訳の翻訳結果を用いて作成した英語質問応答セットを 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．これらの機械翻訳システムは商用目的に作成されており，実用的な品質を持つと考えられるため，機械翻訳の精度についての目安となることを期待して使用した．

[3] https://translate.google.co.jp/, 2015 年 1 月アクセス
[4] http://honyaku.yahoo.co.jp/, 2015 年 2 月アクセス

フレーズベース翻訳　統計的機械翻訳で最も代表的なシステムである Moses (Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, and Herbst 2007) 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．学習には，英辞郎例文[5]，京都フリー翻訳タスクの Wikipedia データ[6]，田中コーパス[7]，日英法令コーパス[8]，青空文庫[9]，TED 講演[10]，BTEC，オープンソース対訳[11] を利用した．また，辞書として英辞郎，WWWJDIC[12]，Wikipedia の言語間リンク[13] を利用した．合計で，対訳コーパス約 255 万文，辞書約 277 万エントリーである．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．

Tree-to-string 翻訳　Tree-to-string 機械翻訳システムである Travatar (Neubig 2013) を用いて質問文を英訳した．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．

Mo セット，Tra セットの作成に用いた翻訳器は，翻訳過程に用いられる手法が明らかであり，翻訳過程という観点からの分析に必要であると考え，これらのセットを作成した．

4　質問応答システム

本研究では，質問応答を行うために SEMPRE[14]という質問応答フレームワークを利用した．SEMPRE は，大規模知識ベースを利用し，高水準な質問応答精度が示されている (Berant et al. 2013)．本節では SEMPRE の動作を述べ，言語横断質問応答に利用する場合に，どのような翻訳が各動作に影響を与えるかを考察する．図 3 に SEMPRE フレームワークの動作例を示し，その動作についてアライメント，ブリッジング，スコアリングの三段階に分けて説明する．

アライメント (Alignment)　アライメントでは，質問文中のフレーズからクエリの一部となるエンティティやプロパティを生成する．このためには，レキシコン (Lexicon) と呼ばれる，自然言語フレーズからエンティティ／プロパティへのマッピングを事前に作成する必要がある．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共起情報などを元に作成される．本研究では先行研究 (Berant et al. 2013) に従い，ClueWeb09[15] (Callan, Hoy,

[5] http://www.eijiro.jp/
[6] http://alaginrc.nict.go.jp/WikiCorpus/index.html
[7] http://www.edrdg.org/wiki/index.php/Tanaka_Corpus
[8] http://www.phontron.com/jaen-law/index-ja.html
[9] http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/align/index.html
[10] https://wit3.fbk.eu/
[11] http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/manual/index-ja.html
[12] http://www.csse.monash.edu.au/ jwb/wwwjdicinf.html#dicfil_tag
[13] https://en.wikipedia.org/wiki/Wikipedia:Database_download
[14] http://nlp.stanford.edu/software/sempre/
[15] http://www.lemurproject.org/clueweb09.php/

Yoo, and Zhao 2009) と呼ばれるデータセットに含まれる新聞記事のコーパスと Freebase を用いて作成されたレキシコンを用いた．図 3 の例では，*“college”* から Type.University のエンティティが生成され，*“Obama”* から BarackObama のエンティティが生成されている．

アライメントに最も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，誤って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)　アライメントによって作成されたエンティティ／プロパティの系列について，隣接するエンティティやプロパティを統合し，知識ベースに入力するクエリを生成する．ブリッジングは隣接する論理式から新たな論理式を生成し，統合する動作である．図 3 の例では，Type.University と BarackObama が隣接しており，両者を繋ぐ論理式として 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．したがって，翻訳結果の語順が誤っていた場合，ブリッジングでの失敗を引き起こすと予想される．

スコアリング (Scoring)　アライメントとブリッジングでは，網羅的に組合せを試し，多数のクエリ候補を出力する．スコアリングでは，評価関数に基づいて候補の導出過程を評価し，

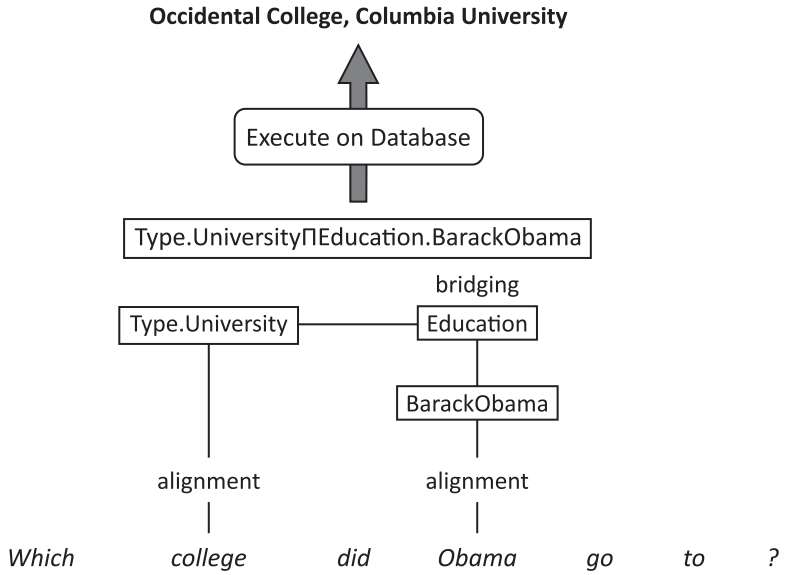

図 3　SEMPRE フレームワークによる質問応答の動作例

最も適切な候補を選択する．図 3 の例では，Type.University⊓Education.BarackObama というクエリ候補のスコアを，「*“college”* から Type.University を生成」し，「*“Obama”* から BarackObama を生成」し，「Education でブリッジする」という導出過程に対して決定する．質問応答システムの学習では，正解を返すクエリを導出することができた導出過程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．そのため，本論文ではこれに関する調査は行っていない．

5 実験

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質問応答精度に影響を及ぼすかを調査した．そのために，3 節で述べたデータセットと 4 節で述べた質問応答システムを用い，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．

5.1 実験 1: 翻訳された質問セットの訳質評価

翻訳精度と質問応答精度の関係を調査するため，まず翻訳結果の訳質を評価した．

5.1.1 実験設定

本実験では，JA セットの質問文から翻訳された 5 つの英語質問応答セットに含まれる質問文の訳質をいくつかの自動評価尺度および人手評価によって評価した．自動評価尺度の参照訳としては，OR セットの質問文を用いた．これは，JA セットの質問文の理想的な英訳が OR セットの質問文であると仮定することに相当する．評価尺度には，4 つの訳質自動評価尺度 (BLEU+1 (Lin and Och 2004)，NIST (Doddington 2002)，RIBES (Isozaki, Hirao, Duh, Sudoh, and Tsukada 2010)，WER (Leusch, Ueffing, and Ney 2003)) と，人手による許容性評価 (Acceptability) (Goto, Chow, Lu, Sumita, and Tsou 2013) を用いた．

BLEU+1　BLEU+1 は，最初に提案された自動評価尺度である BLEU (Papineni, Roukos, Ward, and Zhu 2002) の拡張（平滑化版）である．BLEU は，参照訳と翻訳仮説との間の n-gram 適合率を基準とした評価を行うため，局所的な語順を評価する評価尺度であると言える．短い訳出には参照訳の長さに応じたペナルティを与えることで極端な翻訳に高いスコアを与えないよう設計されている．BLEU はコーパス単位の評価を想定した評価尺度であるが，BLEU+1 は平滑化を導入することで文単位での評価でも BLEU と比べて極端な値

が出づらくなっている．評価値は 0〜1 の実数で，参照訳と完全に一致した文の評価は 1 となる．

RIBES RIBES は単語の順位相関係数に基づいた評価尺度であり，大域的な語順を捉えることができる．その特性から，日英・英日のように大きく異なる文構造の言語対の翻訳評価で人間評価と高い相関が認められている．評価値は 0〜1 の実数で，参照訳と完全に一致した文の評価は 1 である．

NIST NIST スコアは，BLEU や BLEU+1 と同じく n-gram 適合率に基づいた評価尺度であるが，各 n-gram に出現頻度に基づいて重み付けをする点でそれらと異なる．低頻度の語ほど大きな重みが与えられ，結果として頻出する機能語よりも低頻度な内容語に重点を置いた評価尺度となる．評価値は正の実数で与えられ，上限が設定されない．本研究では，参照訳と完全に一致した文の評価値で除算することで，0〜1 の範囲に正規化した値を用いる．

WER WER（Word Error Rate: 単語誤り率）は参照訳と翻訳仮説の編集距離を語数で割ることで得られる尺度で，BLEU や RIBES より厳密に参照訳との語順・単語の一致を評価する．WER は誤り率を表し，低いほどよい翻訳仮説となるため，他の評価尺度と軸向きを揃えるために $1 - WER$ の値を用いた．

許容性 (Acceptability) 許容性は人間による 5 段階の評価である．この評価尺度では，意味的に正しくなければ 1 と評価され，意味理解の容易さ，文法的な正しさ，流暢性によって 2 から 5 の評価が行われる．評価値は 1〜5 の整数であるが，他の評価尺度と比率を合わせるため，0〜1 に正規化した値を用いる．

5.1.2 実験結果

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 4 に示す．

図 4 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のものよりも高いことが読み取れる．次に，GT と YT に着目すると，BLEU と NIST では GT が高く，RIBES と許容性では YT が高い．これは先行研究 (Isozaki et al. 2010) と同様の結果となっており，日英翻訳での RIBES と人手評価による許容性との相関が高いという特性が確認された．また，Mo と Tra を比べると，Tra の翻訳精度が劣っている．通常，日英間の翻訳では，文構造を捉える Tree-to-string 翻訳の精度が比較的良くなるとされているが，今回は翻訳対象が質問文であるため，通常と異なる文型に偏っていることと， 各入力文がそれほど長くなく構造が単純である傾向があるため，文構造を捉える長所が生かされなかったことなどが原因と考えられる．次節で，このような特性が人間相手ではなく質問応答システムの入力として用いた場合でも同様に現れるかどうかを検証する．

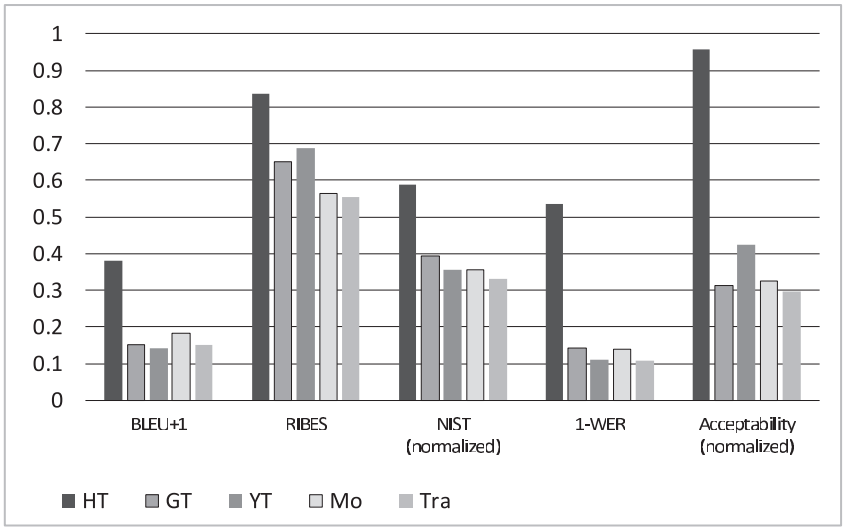

図 4　訳質評価値（平均）

5.2　実験 2: 翻訳された質問セットを用いた質問応答

次に，翻訳精度との関係を調査するため，作成したデータセットを用いて質問応答を行い，質問応答精度を測定した.

5.2.1　実験設定

本実験では 4 節で述べた質問応答フレームワーク SEMPRE を用いて，3 節で述べた手順で作成した 4 つの質問セット及び OR セットの質問応答実験を行い，各セットの質問応答精度を測定した．レキシコンには，ClueWeb09 の新聞記事コーパスと Freebase から構築されたものを使用した．また，評価関数の学習には，Free917 の Train セットと Dev セットを用いた．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なかったため，これらを除いた 264 問の結果を用いて質問応答精度を測定した.

5.2.2　実験結果

各データセットの質問応答の結果を図 5 に示す．図 5 より，元のセット（OR セット）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他のデータセットと比較して高いことが読み取れる．しかしながら図 4 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水準 5%で有意に低いという結果となった（対応有り t 検定）．機械翻訳で作成したセットの中では，GT が最も

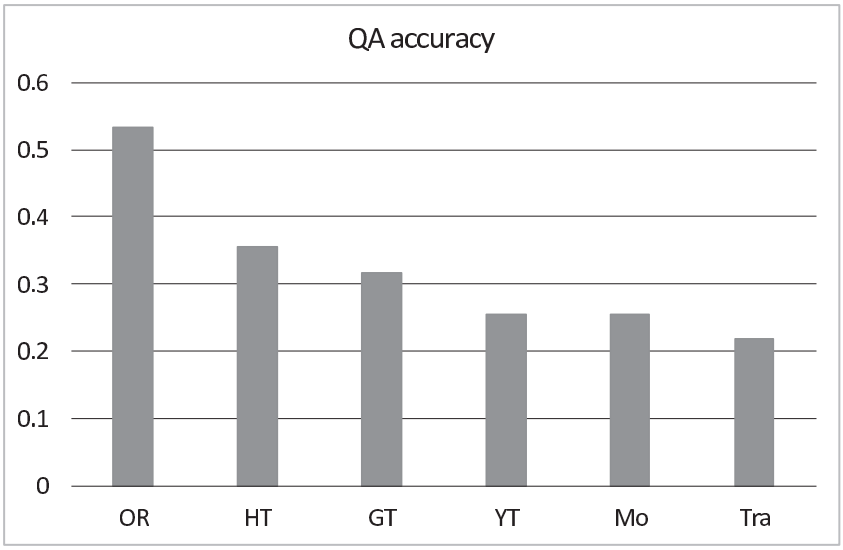

図 5　各データセットにおける質問応答精度

質問応答精度が高く，HT セットの結果との差は有意と言えない結果となった．また，YT は Acceptability において GT を上回るが，質問応答精度は GT より有意水準 5%で有意に低かった．これらの結果は，人間にとって分かりやすい翻訳結果は必ずしも質問応答に適する翻訳結果であるとは限らないことを示唆している．5.3 節，5.4 節で，これらの現象について詳細な分析を行う．

5.3　質問応答精度と機械翻訳自動評価尺度の関係

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，訳質評価値と質問応答精度の相関を文単位で分析した．まず，図 5 に示すように，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いない．参照訳で正解できていない質問は翻訳の結果に関わらず正解することが難しいと考え，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 $141 \times 5 = 705$ 問からなるグループであり，「不正解グループ」は残りの 123 問の翻訳結果 $123 \times 5 = 615$ 問からなるグループである．

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 6 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．例えば，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．図中の R^2 は決定係数である．決定係数は，線形回帰における全変動に対する回帰変動の割合を示し，値が 1 に近いほどよく当てはまる回帰直線である

ことを示す．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 スコアと最も高い決定係数を示した．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ことが確認でき，これを考慮した翻訳を行うことで質問応答精度が改善できると考えられる．これは，内容語が 4 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，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決定係数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考えられる．

Acceptability の図に着目すると，1 → 2 と 3 → 4 で精度の上昇の幅が大きく，2 → 3 や 4 → 5 ではほとんど変化していない．Acceptability における評価値 1 は，「重要な情報が欠落しているか，内容が理解できない文」であることを示し，評価値 2 は「重要な情報が含まれており内容も理解できるが，文法的に誤っており理解が困難な文」であることを表す．このことからも，重要な情報や内容が欠落することは質問応答の精度に大きな影響を与えることがわかる．評価値 2 と 3 の差異は「容易に理解できるかどうか」である．この 2 つの評価値間で質問応答精度が大きく変わらないことは，人間にとっての理解の容易さは，質問応答精度の向上にはそれほど寄与しない可能性を示唆している．評価値 3 と 4 の差異は「文法的に正しいかどうか」である．この 2 つの間でも精度が大きく上昇しており文法が重要な可能性があるが，評価値 4 と評価された文が少ないため誤差が含まれている可能性もある．この点については後の分析で述べる．評価値 4 と 5 の差異は「ネイティブレベルの英語かどうか」である．この間では質問応答精度がほとんど変わらず，評価値 5 の方が少し下がる傾向が見られた．前述したように評価値 4 の文が少ないことによる誤差の可能性もあるが，ネイティブに用いられる言い回しが質問応答器にとっては逆効果となっている可能性も考えられる．

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 7 に示す．不正解グループにおいては，全ての自動評価尺度において正解グループと比較して決定係数が低いという結果となった．この結果より，参照訳で質問応答器が解答できない問題では，翻訳を改善することで正解率を向上させるのが難しいということが言える．これは，言語横断質問応答のための翻訳器を評価する際の参照訳は質問応答器で正解可能であることが望ましいと言うこともできる．また，質問応答成功率を予測できれば，質問応答成功率が高い文を参照訳として機械翻訳を最適化することでこの問題を軽減できると考えられる．しかし，正解グループ・不正解グルー

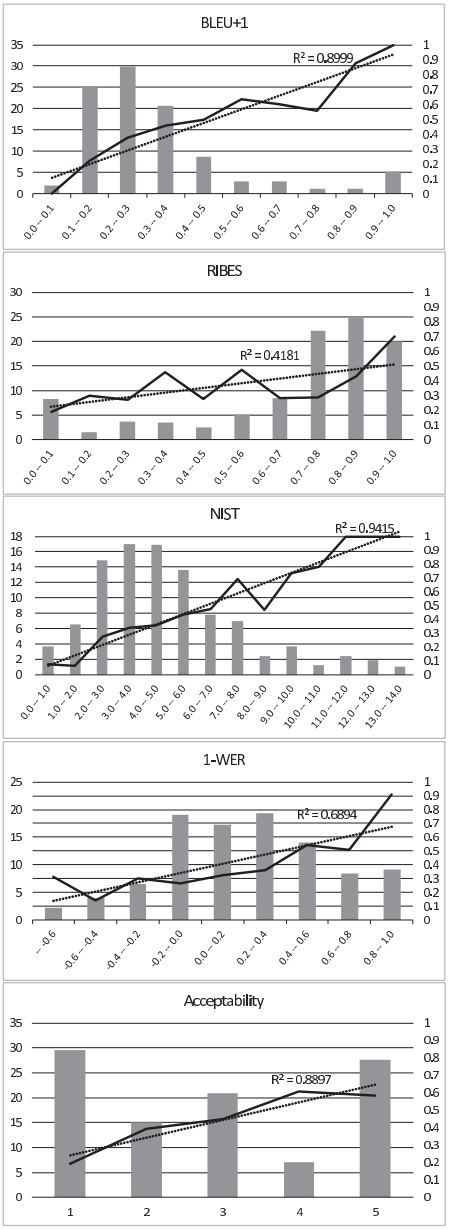

図 6　評価尺度値と質問応答精度の相関（正解グループ）
横軸：評価値の範囲
棒グラフ（左縦軸）：質問数の割合 (%)
折れ線（右縦軸）：質問応答精度（範囲内平均）

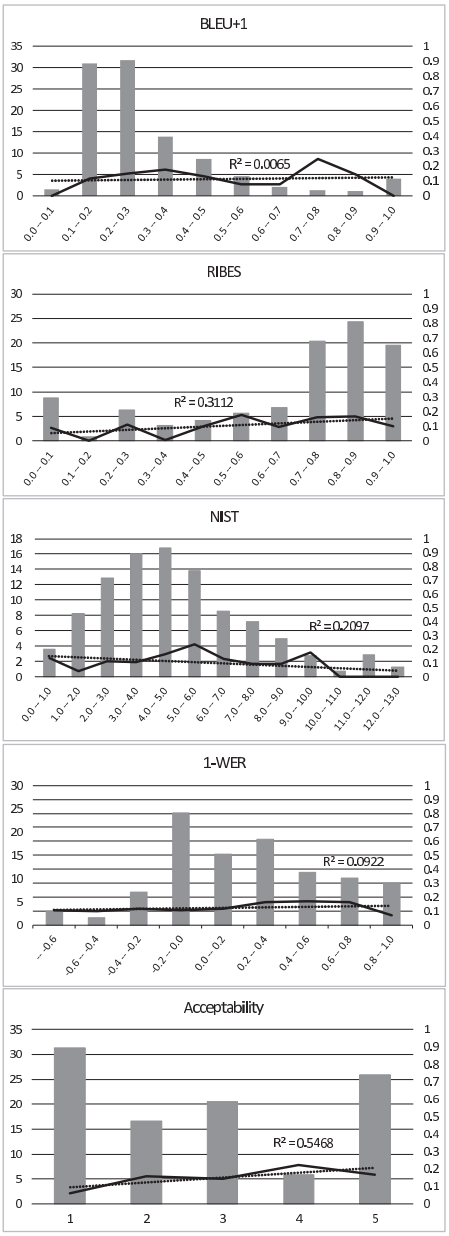

図 7　評価尺度値と質問応答精度の相関（不正解グループ）
横軸：評価値の範囲
棒グラフ（左縦軸）：質問数の割合 (%)
折れ線（右縦軸）：質問応答精度（範囲内平均）

プのどちらにおいても，訳質評価尺度の値に対する質問数の分布は似通っており，訳質評価尺度でこの問題を解決することは困難であると考えられる．

5.4 質問応答事例分析

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．

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．第 1 列は，各質問文での質問応答が成功したかどうかを表す記号であり，○ が成功，× が失敗を表す．表 2 の 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，これらを入力とした場合のみ正しく答えることができている．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変換されていた．例えば，HT に含まれる *“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，*“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が，*“who made magic flute”* では不正解であったことから，*“librettist”* が重要な語であることがわかる．この例でも 1 つ目の例と同様に，librettist という Freebase 内のプロパティと一致する表

表 2　内容語の変化による質問応答結果の変化の例

正誤	セット	質問文
○	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
○	—	librettist magic flute
×	—	who made magic flute

現を含むことが質問応答の精度に寄与することが示唆される例である．

このような例から，内容語が変化することでアライメントが失敗し，正しいエンティティが生成されないことや誤ったエンティティが生成されることが重要な問題であることが確認できる．この問題は，正しいエンティティと結びつきやすい内容語の表現を翻訳の過程で考慮することで改善できる可能性がある．また，これらの結果は本実験で使用した質問応答器の問題であるとも考えられ，言い換えを考慮できる質問応答器を用いることでも改善できる可能性がある．

次に，質問タイプを表す語の誤訳が質問応答結果の変化の原因となる例を表 3 に示す．1 つ目の例では，内容語と考えられる *“tv (television) programs”*，*“danny devito”*（YT は綴りミスあり），*“produce(d)”* の 3 つは全ての翻訳結果に含まれているが，HT 以外は正解できていなかった．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と考えられる．GT や Mo の質問文に対する解答を確認したところ，番組名をリストアップして答えており，正解とされる数と同じ数だけ答えていた．この例より，解答の形式を変化させるような質問タイプを示す語を，正確に翻訳する必要があることがわかる．一方で，2 つ目の例では，*“what”* や *“which”* といった語が含まれていない Mo の質問文でも正解することができている．この例より，質問タイプを表す語であっても重要度が低いものがあると考えられる．したがって，言語横断質問応答のための翻訳器は，解答の形式を変えるような質問タイプ語の一致を重視することが求められる．質問タイプを表す語は内容語と異なり頻出するため，NIST スコアのように頻度に基づいて重要度を決めることは難しく，質問応答固有の指標が必要であると考えられる．

表 3　質問タイプ語の誤訳による質問応答結果の変化の例

正誤	セット	質問文
○	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
○	OR	what weight class was the fight of the century
—	JA	「世紀の一戦」はどの階級でしたか
×	HT	what rank was the fight of the century
×	GT	did any class century of battle is
×	YT	which rank was “the fight of the century”
○	Mo	class is the fight of the century
○	Tra	the fight of the century, ’ which was the class

文法や語順に関連する例を表 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．これらの例は，少なくとも Free917 に含まれるような単純な事実型質問においては，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．ただし，より複雑な事実型質問や，非事実型質問に対して解答する際には，誤った語順の影響が強くなる可能性は否定できない．

これらの例は，使用した質問応答システムが語順の影響を受けづらいものであったことによる可能性も考えられる．これを明らかにするためには様々な質問応答システムを用いて実験を行うことが必要であるが，それは今後の課題とする．

5.2.2 節で述べたように，人間にとってわかりやすい翻訳が質問応答にも成功しやすい翻訳とは限らない可能性がある．実際に質問応答の結果を見ると，質問応答の正誤と Acceptability の評価が反する例が確認された．その一例を表 5 に示す．1 つ目の例では，*“do you”* というフレーズを含むことによって文章の意味が変わっているため Acceptability は 1 と評価されているが，質問応答では正解できている．この例では内容語は正しく翻訳できており，*“do you”* というフレーズを無視することができたため正解することができたと考えられる．2 つ目の例では，主に前置詞の意味の違いによって，GT は 2 という低い評価が付けられている．一方で YT は GT と比較して意味的に正しく翻訳されており 3 と評価されているが，質問応答の結果は不正解で

表 4　文法誤りを含む訳による質問応答結果の例

正誤	セット	質問文
○	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

表 5 許容性と質問応答結果が反する例

正誤	許容性	セット	質問文
○	—	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
○	1	GT	sunset branch library do you belong to any library system
○	—	OR	what are the theme areas at disneyland
—	—	JA	ディズニーランドにはどのようなエリアがありますか
○	2	GT	what are the areas to disneyland
×	3	YT	what kind of area is there in disneyland
○	—	OR	what decision did manny pacquiao vs. timothy bradley end with
—	—	JA	マニー・パッキャオ対ティモシー・ブラッドリーはどの判定で終わりましたか
○	2	GT	did the end in any decision manny pacquiao vs. timothy bradley
×	3	YT	which judgment did mannie pacquiao vs. timothy bradley terminate in

あった．質問応答の過程を見ると，OR と GT の文からは areas というテーマパークのエリアを示すプロパティが得られたのに対し，YT の文からは area という面積を示すプロパティが得られていた．このことから，意味的に正しい文であることよりも内容語の表層的な一致がより重要であることがわかる．3 つ目の例では，YT は固有名詞である *“manny pacquiao”* を *“mannie pacquiao”* としており，質問応答結果が不正解となっている．人間が固有名詞を判断するときには少々の誤字が含まれていたとしても読み取れることから，YT の文に 3 という評価値が付けられたと考えられるが，機械による質問応答においては，特に固有名詞中の誤字は重大な問題であることがこの例により示唆される．

5.5 実験 3: 自動評価尺度を用いてリスコアリングされた翻訳結果を用いた質問応答

5.3 節，5.4 節の分析の結果，質問応答精度と最も高い相関を持つ自動評価尺度は NIST スコアであった．したがって，NIST スコアが高評価となるよう翻訳システムを学習させることで，質問応答に適した翻訳システムとなる可能性がある．そこでまず，多数の翻訳結果から NIST スコアが最も高い翻訳結果を選択することで，質問応答精度が向上するかどうかを調べる．

5.5.1 実験設定

翻訳 N ベストの内，最も NIST の高い翻訳を使用した時の質問応答精度を調査する．本実験では，翻訳器に Moses と Travatar を用い，$N = 100$ とした．また，比較のため BLEU+1 についても同様の実験を行った．

表 6　翻訳 100 ベスト選択実験結果

Moses			
選択基準	質問応答精度	vs. 翻訳器 1 ベスト	vs. BLEU+1
翻訳 1 ベスト	0.253	—	—
BLEU+1	0.285	+0.032 ($p = 0.241$)	—
NIST	0.301	+0.048 ($p = 0.078$)	+0.016 ($p = 0.181$)
Travatar			
選択基準	質問応答精度	vs. 翻訳器 1 ベスト	vs. BLEU+1
翻訳 1 ベスト	0.218	—	—
BLEU+1	0.271	+0.053 ($p = 0.023$)	—
NIST	0.281	+0.063 ($p = 0.009$)	+0.010 ($p = 0.253$)

p 値は質問応答精度についての対応有り両側 t 検定の結果

5.5.2　実験結果

表 6 に実験結果を示す．また，比較のため翻訳システム第一位の結果を用いた場合の精度も表中に示す．表より，翻訳 N ベストの中から適切な選択を行うことで，質問応答の精度が向上することがわかる．特に Travatar を用いた言語横断質問応答において，BLEU+1 および NIST スコアを用いて翻訳結果を選択することで，有意水準 5%で統計的有意に質問応答精度が向上している．また，選択基準に NIST スコアを選んだ場合の正答率は，選択基準に BLEU+1 を選んだ時の正答率よりも向上する傾向にある．これらの結果は，機械翻訳器の最適化によって言語横断質問応答の精度を改善できる可能性を示している．

本実験で使用した選択手法は，質問応答精度の高い参照訳が必要であり，未知の入力の翻訳結果選択に直接用いることはできない．しかし，質問応答精度と高い相関を持つ評価尺度に基づいて翻訳器を最適化することで，質問応答精度の高い翻訳結果を得ることが可能であると考えられる．

5.6　実験 4: 様々な言語対での翻訳精度と質問応答精度の関係調査

実験 1 から 3 では，日英言語横断を行い，訳質と翻訳精度の関係について調査した．次に，日英以外の言語対における言語横断質問応答においても，同様の結果が得られるかどうかを調査する．

5.6.1　データセット作成

Haas らによって作成されたドイツ語版の free917 セット (Haas and Riezler 2015) を入手し，そのテストセットに含まれる質問文を Google 翻訳[16]および Bing 翻訳[17]を用いて英訳し，DE-GT

[16] https://translate.google.co.jp/, 2016 年 6 月アクセス
[17] https://www.bing.com/translator, 2016 年 6 月アクセス

セットおよび DE-Bing セット（独英）を作成した．また，3 章に示す手順に従い，OR セットに含まれる質問文を中国語，インドネシア（尼）語，ベトナム（越）語の母語話者に依頼して人手翻訳してもらい，それぞれの言語の質問セットを新たに作成した．次に，これらの 3 つの質問セットをそれぞれ Google 翻訳および Bing 翻訳を用いて英訳し，ZH-GT セット（中英），ID-GT セット（尼英），VI-GT セット（越英），ZH-Bing セット，VI-Bing セット，ID-Bing セットの 6 つの英語質問セットを作成した．また比較のため，JA セットを Bing 翻訳を用いて英訳し，JA-Bing セット（日英）を作成した．

5.6.2　訳質評価と質問応答精度の関係

作成した 9 つの質問セットを用いて，4 章に示す質問応答システムによる質問応答を行い，質問応答精度を評価した．その結果を図 8 に示す．比較のため，同翻訳手法を用いた日英の質問セット (JA-GT) での結果を合わせて示す．図より，どの言語対においても，翻訳による質問応答精度の低下は起こっており，その影響を緩和するような翻訳結果を得ることは重要であると言える．また，中英セットと越英セットの質問応答精度が他と比較して低いことから，同じ翻訳手法を用いても言語対によって影響に差があることがわかる．

次に，5.1.1 節に示す評価尺度の内，許容性評価を除く 4 つの評価尺度を用いて，前節で作成した 9 つの質問セットの訳質を評価した．また各質問セットについて，5.3 節と同様に参照訳での質問応答が正解できているかどうかで 2 つのグループに分け，各グループ内での各評価尺度と質問応答精度との相関を測定した．ただし本実験では，評価値の範囲で平均するのではなく，各文の評価値と質問応答結果（完全正解で 1，完全不正解で 0）を直接使用した．表 7，表 8 に示す結果より，どの言語対においても不正解グループの決定係数は正解グループに比べて小さ

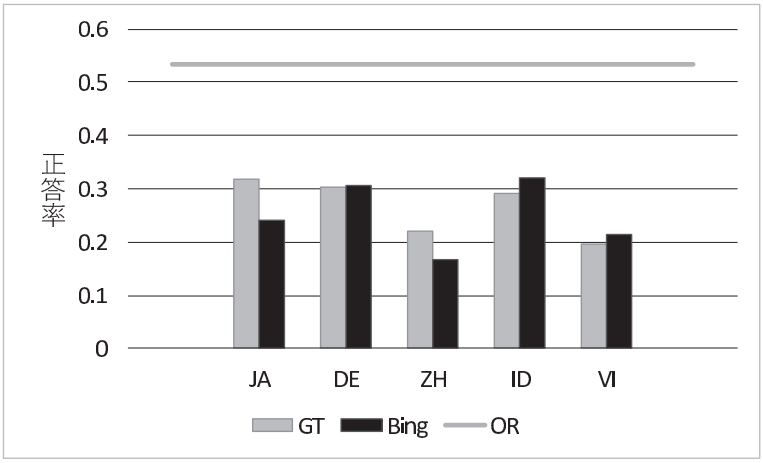

図 8　様々な言語対における質問応答精度

表 7　評価尺度と質問応答精度との決定係数 (GT)（太字は正解グループ内の最大値）

セット	JA-GT（日英）		DE-GT（独英）		ZH-GT（中英）		ID-GT（尼英）		VI-GT（越英）		全言語対	
グループ	正解	不正解	正解	不正解	正解	不正解	正解	不正解	正解	不正解	正解	不正解
BLEU+1	0.120	0.000	0.048	0.000	0.129	0.000	0.069	0.000	**0.115**	0.004	0.086	0.000
RIBES	0.057	0.004	0.020	0.004	0.099	0.004	0.061	0.000	0.016	0.000	0.058	0.000
NIST	**0.147**	0.002	**0.109**	0.003	**0.167**	0.000	0.077	0.000	0.100	0.008	**0.117**	0.000
1-WER	0.095	0.003	0.083	0.005	0.115	0.004	**0.122**	0.010	0.071	0.000	0.094	0.001

表 8　評価尺度と質問応答精度との決定係数 (Bing)（太字は正解グループ内の最大値）

セット	JA-Bing（日英）		DE-Bing（独英）		ZH-Bing（中英）		ID-Bing（尼英）		VI-Bing（越英）		全言語対	
グループ	正解	不正解	正解	不正解	正解	不正解	正解	不正解	正解	不正解	正解	不正解
BLEU+1	0.095	0.006	0.174	0.007	0.067	0.000	0.075	0.006	0.062	0.000	0.122	0.000
RIBES	0.019	0.014	0.090	0.002	0.011	0.008	0.125	0.000	0.011	0.004	0.060	0.000
NIST	**0.104**	0.013	**0.200**	0.001	**0.070**	0.008	0.082	0.005	**0.084**	0.002	**0.140**	0.001
1-WER	0.042	0.005	0.157	0.003	0.031	0.000	**0.128**	0.002	0.050	0.004	0.104	0.000

く，無相関に近いことがわかる．正解グループの決定係数も最大 0.200 となっており図 6 の値と比べると小さいが，これはほぼ 2 値で表現される質問応答結果と連続値で表される評価尺度の間で相関を計算したことが原因であると考えられる．まず，全言語対の結果をまとめて計算した時（表中の右端の列），最も相関が高い評価尺度は NIST スコアであり，本実験で使用したどの言語対においても内容語の表層の一致が重要であることがうかがえる．各言語対の正解グループの決定係数に着目すると，日英と中英では似た傾向がある一方で，尼英では 1-WER が最大の決定係数を持っており，言語対によっては異なった特徴が現れている．また独英では，他言語対と比べて NIST スコアと BLEU+1 の差が大きく，両評価尺度の差である内容語の一致が特に重要であることが予想できる．このことから，全体として NIST スコアが質問応答精度と強く相関するが，言語対の特徴を考慮することでより強い相関を持った尺度を得ることができると考えられる．しかしながら，言語対によって異なる特徴については，現段階では詳細に至るまで分析できておらず，今後さらなる分析が必要とされる．

6　まとめ

本研究では，言語横断質問応答システムの精度を向上させるため，翻訳結果が質問応答の結果に与える影響を調査した．

具体的には，翻訳精度評価（5.1 節）と言語横断質問応答精度の評価（5.2 節）を行い，両者の関係を分析した（5.3 節）．その結果，内容語の一致を重視する NIST スコアが質問応答精度と高い相関を持つことがわかった．これは質問応答において内容語が重要であるという直感にも合致する結果である．一方で，人手評価が NIST スコアや BLEU+1 といった自動評価よりも

相関が低いこともわかった．この結果より，人間が正しいと評価する翻訳が必ずしも質問応答に適しているとは限らないという知見が得られた．

この結果に対して，質問応答結果の事例分析（5.4 節）を行ったところ，以下の 2 つのことがわかった．1 つ目は，人間が正しいと評価した内容語でも質問応答システムが正しく解答できない場合もあり，翻訳結果に含まれる内容語の正しさの評価基準は人間と質問応答システムで必ずしも一致しないということがわかった．2 つ目は，質問タイプを表す語の中には，正しい解答を出すために重要な語と重要でない語があることがわかった．具体的には，*“how many”* など解答の形式を変化させる語は正しい翻訳が必須であり，*“what”* や *“which”* などの語は翻訳結果に含まれていなくても正しく解答することができている例が確認できた．

また，NIST スコアに基づいて選択された翻訳結果の質問応答実験（5.5 節）により，内容語に重点を置いた翻訳結果を使用することで言語横断質問応答精度が改善されることがわかった．この結果から，機械翻訳器の最適化を行うことで，言語横断質問応答の精度を改善できる可能性を示した．

最後に，日英以外の言語対における言語横断質問応答実験（5.6 節）では，日英以外の 3 言語対においても日英と同様に内容語を重視する訳質評価尺度が質問応答精度と相関が高い傾向が見られた．このことから，内容語を重視した訳質評価尺度と質問応答精度が高い相関を持つという知見は多くの言語対で見られ，一般性のある知見であることが示された．

今後の課題としては，様々な言語対および質問応答システムを用いた言語横断質問応答を行うことでより一般性のある知見を得ることや，質問応答精度と高い相関を持つ評価尺度の作成，そのような尺度を用いて機械翻訳器を最適化することによる質問応答精度の変化を確認することなどが挙げられる．

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ある．また，本研究開発の一部は総務省 SCOPE（受付番号 152307004）の委託を受けたものである．

参考文献

Berant, J., Chou, A., Frostig, R., and Liang, P. (2013). “Semantic Parsing on Freebase from Question-Answer Pairs.” In *Proceedings of EMNLP*, pp. 1533–1544.

Cai, Q. and Yates, A. (2013). “Large-scale Semantic Parsing via Schema Matching and Lexicon Extension.” In *Proceedings of ACL*, pp. 423–433.

Callan, J., Hoy, M., Yoo, C., and Zhao, L. (2009). “Clueweb09 Dataset.”.

Doddington, G. (2002). “Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics.” In *Proceedings of HLT*, pp. 138–145.

Fader, A., Zettlemoyer, L., and Etzioni, O. (2014). “Open Question Answering over Curated and Extracted Knowledge Bases.” In *Proceedings of ACM SIGKDD*, pp. 1156–1165.

Goto, I., Chow, K. P., Lu, B., Sumita, E., and Tsou, B. K. (2013). “Overview of the Patent Machine Translation Task at The NTCIR-10 Workshop.” In *Proceedings of NTCIR-10*, pp. 260–286.

Haas, C. and Riezler, S. (2015). “Response-based Learning for Machine Translation of Open-domain Database Queries.” In *Proceedings of NAACL HLT*, pp. 1339–1344.

Hyodo, T. and Akiba, T. (2009). “Improving Translation Model for SMT-based Cross Language Question Answering.” In *Proceedings of FIT*, Vol. 8, pp. 289–292.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (2010). “Automatic Evaluation of Translation Quality for Distant Language Pairs.” In *Proceedings of EMNLP*, pp. 944–952.

Kiyota, Y., Kurohashi, S., and Kido, F. (2002). “Dialog Navigator: A Question Answering System based on Large Text Knowledge Base.” In *Proceedings of COLING*, pp. 1–7.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). “Moses: Open Source Toolkit for Statistical Machine Translation.” In *Proceedings of ACL*, pp. 177–180.

Leusch, G., Ueffing, N., and Ney, H. (2003). “A Novel String-to-string Distance Measure with Applications to Machine Translation Evaluation.” In *Proceedings of MT Summit IX*, pp. 240–247.

Lin, C.-Y. and Och, F. J. (2004). “ORANGE: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation.” In *Proceedings of COLING*, pp. 501–507.

Mori, T. and Kawagishi, M. (2005). “A Method of Cross Language Question-answering based on Machine Translation and Transliteration.” In *Proceedings of NTCIR-5*, pp. 182–189.

Neubig, G. (2013). “Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers.” In *Proceedings of ACL*, pp. 91–96.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). “BLEU: A Method for Automatic Evaluation of Machine Translation.” In *Proceedings of ACL*, pp. 311–318.

Riezler, S., Simianer, P., and Haas, C. (2014). “Response-based Learning for Grounded Machine Translation.” In *Proceedings of ACL*, pp. 881–891.

Shimizu, K., Fujii, A., and Itou, K. (2005). “Bi-directional Cross Language Question Answering using a Single Monolingual QA System.” In *Proceedings of NTCIR-5*, pp. 455–462.

Tunstall-Pedoe, W. (2010). “True Knowledge: Open-domain Question Answering Using Structured Knowledge and Inference.” *AI Magazine*, **31** (3), pp. 80–92.

略歴

杉山享志朗：2014 年呉工業高等専門学校機械電気専攻卒業．2016 年奈良先端科学技術大学院大学情報科学研究科修士課程修了．同年より，同大学院博士後期課程在学．自然言語処理に関する研究に従事．

水上　雅博：2012 年同志社大学理工学部卒業．2014 年奈良先端科学技術大学院大学情報科学研究科修士課程修了．同年より同大学院博士後期課程在学．自然言語処理および音声対話システムに関する研究に従事．人工知能学会，音響学会，言語処理学会各会員．

Graham Neubig: 2005 年米国イリノイ大学アーバナ・シャンペーン校工学部コンピュータ・サイエンス専攻卒業．2010 年京都大学大学院情報学研究科修士課程修了．2012 年同大学院博士後期課程修了．同年奈良先端科学技術大学院大学助教．2016 年より米国カーネギーメロン大学助教．機械翻訳，自然言語処理に関する研究に従事．

吉野幸一郎：2009 年慶應義塾大学環境情報学部卒業．2011 年京都大学大学院情報学研究科修士課程修了．2014 年同博士後期課程修了．同年日本学術振興会特別研究員 (PD)．2015 年より奈良先端科学技術大学院大学情報科学研究科特任助教．2016 年より同助教．京都大学博士（情報学）．音声言語処理および自然言語処理，特に音声対話システムに関する研究に従事．2013 年度人工知能学会研究会優秀賞受賞．IEEE，ACL，情報処理学会，言語処理学会各会員．

鈴木　　優：2004 年奈良先端科学技術大学博士後期課程修了．博士（工学）．現在，奈良先端科学技術大学院大学情報科学研究科特任准教授．情報検索やクラウドソーシングに関する研究開発に従事．情報処理学会，電子情報通信学会，ACM，IEEE Computer 各会員．

中村　　哲：1981 年京都工芸繊維大学工芸学部電子工学科卒業．京都大学博士（工学）．シャープ株式会社．奈良先端科学技術大学院大学助教授，2000 年 ATR 音声言語コミュニケーション研究所室長，所長，2006 年（独）情報通信研究機構研究センター長，けいはんな研究所長などを経て，現在，奈良先端科学技術大学院大学教授．ATR フェロー．カールスルーエ大学客員教授．音声翻訳，音声対話，自然言語処理の研究に従事．情報処理学会喜安記念業績賞，総務大臣表彰，文部科学大臣表彰，AntonioZampoli 賞受賞．IEEESLTC 委員，ISCA 理事，IEEE フェロー．

（2016 年 4 月 4 日　受付）
（2016 年 7 月 11 日　再受付）
（2016 年 8 月 31 日　採録）