1K3-2

知識ベースに基づく言語横断質問応答における訳質の影響の調査

The Effect of Translation Quality on Knowledge-based Cross-lingual Question Answering

杉山 享志朗 Kyoshiro Sugiyama　Graham Neubig　Sakriani Sakti　戸田 智基 Tomoki Toda　中村 哲 Satoshi Nakamura

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

To investigate how translation affects knowledge-based cross-lingual question answering (KBCLQA) quality, we translate Japanese queries to English in 3 different ways and compare the results on KBCLQA. Specifically, one data set is translated by humans, and the others are translated by machine translation systems. According to our experimental results, we found that NIST score, which focuses on translation accuracy of content words, has the highest correlation with KBCLQA quality out of four evaluation measures.

1. はじめに

質問応答は、自然言語の質問文に対して検索対象から回答を検索する技術である。言語横断質問応答は、質問と検索対象の言語が異なる場合の質問応答である。言語横断質問応答の実現には、質問文を機械翻訳する手法が広く用いられており、そのために機械翻訳を最適化するという研究もされている [1, 2]。また、機械翻訳された外国語の対話完成問題を人間が解くためにはどのような翻訳が求められるかを調査した例もある [3]。

質問応答の分野で特に近年注目されている研究として、大規模知識ベースを対象としたオープンドメイン質問応答が挙げられる [4, 5, 6]。大規模知識ベースを用いることにより、ユーザの多岐に渡る質問に回答することが可能となる。その一方、大規模知識ベースの構築に膨大な労力がかかるため、知識ベースが存在する言語は限られており、最も大規模なものは英語でしか存在しない。例えば、日本語の知識ベースとして DBpedia[*1] が知られているが、登録エンティティ数は 210 万に留まり、英語の知識ベース Freebase[*2] のエンティティ数が 2300 万であることを考えると非常に小規模と言わざるを得ない。このため、言語横断質問応答はますます重要となってきている。

このような背景の中、大規模知識ベースを用いた言語横断質問応答において、翻訳がどのように質問応答性能に影響しているかを調査した例は見られない。我々は、知識ベースに基づく質問応答において、質が高いと思われる人手による翻訳でも質問応答性能に影響を与えるかどうかを調べ、また翻訳の質の影響を調査するため、2 つの機械翻訳システムを加えて検証を行った。

2. 質問セット

我々は翻訳の影響を検証するため、質問応答用のデータセットから、それを翻訳したデータセットを作成し、両者の質問応答の結果を比較した。本節では、翻訳されたデータセットの作成手順について主に述べる。

連絡先: 杉山 享志朗，奈良先端科学技術大学院大学情報科学研究科，奈良県生駒市高山町 8916-5，080-1947-3064，sugiyama.kyoshiro.sc7@is.naist.jp

∗1 http://ja.dbpedia.org/

∗2 https://www.freebase.com/

実験では、Free917 [4] と呼ばれる質問セットを用いた。このセットは、知識ベース “Freebase” を用いた質問応答の為に作成された質問セットで、917 対の「質問文」と「その質問を正しく表すクエリ (正解クエリ)」のペアから成る。Freebase は無料で公開されている大規模知識ベースで、ユーザによって編集可能で、広い知識をカバーしているという特徴がある。このセットを先行研究に従い、train セット (512 文、56%)、dev セット (129 文、14%)、テストセット (276 文、30%) の 3 つに分けた。以降、翻訳前のテストセットを OR セットと呼ぶ。

翻訳の質が質問応答の精度に及ぼす影響を検証するために、まず OR セットに含まれる質問文を人手で日本語に翻訳し (TR セット)、さらに次に述べる各手法によって再翻訳し、翻訳後テストセットとした。日本語への翻訳は 1 名が人手で行い、再翻訳は、機械翻訳 2 種類 (GT, YT) と翻訳業者による翻訳 (HT) で行った。翻訳後テストセットの正解クエリは OR セットと同一のものとした。作成したテストセットの一部を表 1 に示す。

表 1: テストセットのサンプル

セット	質問文	正解クエリ
OR	what is europe 's area	(location.location.area en.europe)
TR	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	

各翻訳後セットの評価には、BLEU [7], RIBES [8], NIST [9], Word Error Rate [10] の 4 つを用いた。BLEU は予め用意された良質な翻訳 (以降、参照訳) との n-gram の一致率に基づいた評価尺度で、機械翻訳の自動評価尺度としては最も広く用いられており、亜種も多い。RIBES は大局的な語順に着目し、訳語の些細な違いに寛大な評価尺度で、日英など翻訳時に語順の入れ替えが起こる言語対に対して有効とされる。これら 2 つの評価値は 0 から 1 の実数で、1 に近づくほど良い評価であることを示す。NIST は BLEU と同様 n-gram 一致率に基づいた評価尺度だが、単語に重み付けを行うことで、機能語よりも内容語を重視した評価を行う。評価値は 0 以上の実数で、大きいほど良い評価となる。WER は、評価対象訳と参照訳との編集距離を語数で正規化したもので、語順に非常に敏感な特性を持つ。評価値は 0 から 1 の実数で、0 に近づくほど良い評価となる。

翻訳前の OR セットの質問文を参照訳とした各セットの質問

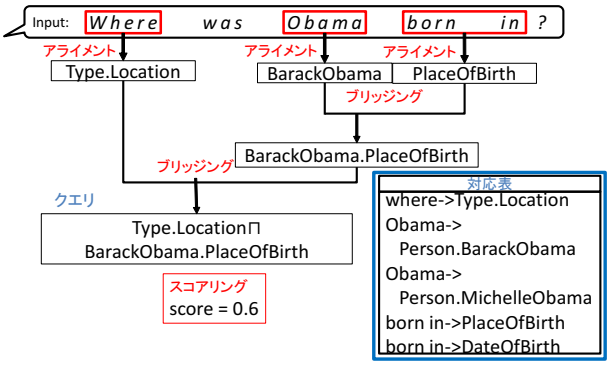

図 1: 意味理解フレームワーク

文のを表 2 に示す。NIST は上限が設定されない尺度のため、参照訳と評価対象訳を同一にした場合の数値を用いて正規化した。

表 2: 各 test セットの自動評価尺度値

	OR	HT	GT	YT
BLEU	1.000	0.384	0.149	0.142
RIBES	1.000	0.838	0.652	0.690
NIST	11.46	6.771	4.503	4.097
NIST(正規化後)	1.000	0.591	0.393	0.358
WER	0.000	0.517	0.905	0.922

HT の評価値は GT,YT に比べると全体的に高く、人手翻訳の質が高いことが読み取れる。GT と YT の 2 つの機械翻訳の間では、特に RIBES と NIST に差が現れている。このことは、GT は内容語の一致率が高く、YT は大局的な語順に関して比較的正しく翻訳できていることを示唆している。また機械翻訳 2 種では WER が 0.9 を超え、参照訳と全く同じ訳が少ないことを示している。

3. 質問応答器

実験では、質問応答器として Sempre [5] を用いた。従来の知識ベース型質問応答器が学習の際に教師データとして正解クエリを必要とするのに対し、Sempre は質問文と正解のみで学習が可能であるという特徴を持っている。これにより、学習用データセットの作成が従来に比べ容易であり、今後の展望が見込まれることから使用した。

本節では、Sempre の意味理解フレームワークをアライメント、ブリッジング、スコアリングと学習の 3 つに分けて述べる。図 1 に意味理解フレームワークの概要を示す。

3.1 アライメント

Sempre に質問文を入力した際、まず質問文に含まれるフレーズを対応表を基に論理形式の系列を生成する。この動作をアライメントと呼ぶ。図 1 におけるアライメントでは、“*Where*”,“*Obama*”,“*born in*” のフレーズからそれぞれ“Type.Location”,“BarackObama”,“PlaceOfBirth”の論理形式が生成されている。

アライメントに用いる対応表は、Berant らによって ClueWeb09*3[11] から抽出された 3 項関係と Freebase から作成されたものを用いた [5]。ClueWeb09 は 10 億以上の web ページをクロールして作成された文書集合で、このデータから対応表を作成することにより、様々な固有表現をカバーすることが可能となる。

*3 http://www.lemurproject.org/clueweb09.php/

3.2 ブリッジング

アライメントによって生成された論理形式の系列をクエリに統合する動作をブリッジングと呼ぶ。ブリッジングでは、隣り合った 2 つの論理形式を 1 つの論理形式に統合することを繰り返して、最終的に 1 つの論理形式をクエリとして出力する。

3.3 スコアリングと学習

アライメントとブリッジングによって生成されるクエリは、1 つの質問文に対して一意に定まるとは限らない。例えば、図 1 の例では、“*Obama*” から生成される論理形式は “BarackObama” と “MichelleObama” のどちらでも有り得、またブリッジングの順番も変わり得る。このため、アライメントやブリッジングの良さを評価し、比較する必要がある。

スコアリングでは、質問文、アライメント、ブリッジングを一つの導出としてまとめ、導出の特徴を元にスコアを付ける。学習では、正解を得ることができるクエリに高いスコアが付くよう、導出の特徴に対する重み付けを最適化する。

4. 実験結果

2. 節の手順に従ってテストセットを作成し、train セットと dev セットを用いて 3. 節に示す質問応答器を学習させた。学習の繰り返し回数は 5 回とした。

学習させた質問応答器を用いて各テストセットを解かせた時の結果を表 3 に示す。各項目は以下のように定義されている：

correct 最大スコアの回答が正解と完全に一致した割合

oracle 正解と完全に一致する回答が回答候補に含まれた割合

partCorrect 最大スコアの回答の正解に対する F 値の平均

partOracle 回答候補中最大の正解に対する F 値の平均

表 3: 実験結果

	OR	HT	GT	YT
correct	0.525	0.351	0.312	0.257
oracle	0.761	0.583	0.464	0.402
partCorrect	0.529	0.358	0.314	0.259
partOracle	0.766	0.589	0.471	0.413

表より、OR>HT>GT>YT の順で回答精度が高いことが読み取れる。各質問での最大スコアの回答の F 値 (partCorrect) について片側 t 検定を行ったところ、OR-GT 間と GT-YT 間に統計的な有意差が確認された ($p < 0.05$) が、HT-GT 間には有意差が確認できなかった ($p = 0.129$)。また、全てのセットにおいて誤答もしくは回答できなかった問題は 93 問、全てのセットにおいて正解できた問題は 36 問であった。

5. 考察

5.1 翻訳精度と質問応答精度の相関

考察にあたって、本稿では質問応答の精度を表すのに partCorrect を用いる。これは、正解がリストで返される場合があり、correct では部分正解を扱えないためである。

我々は各翻訳結果の自動評価値を元に、翻訳全体の特徴と質問応答の精度の関係を調べた。まず、自動評価尺度と質問応答精度の関係を図 2 に示す。

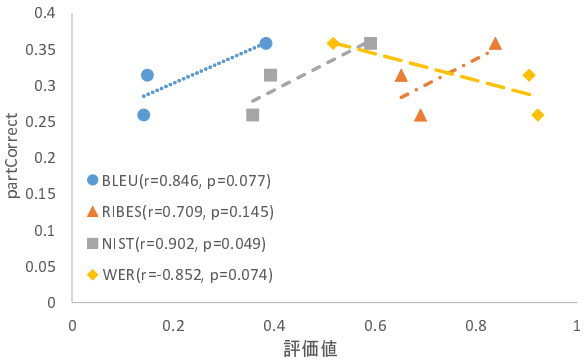

図 2: 自動評価値と正解率

最も高い相関を示し、統計的有意 ($p < 0.05$) に相関が認められた自動評価尺度は NIST であった。NIST は BLEU と同じく n-gram 一致率に基づいた評価尺度であるが、NIST と BLEU の最大の相違点は、個々の n-gram に情報量に基づく重み付けがされ、機能語よりも内容語を重視した評価をする点である。NIST と高い相関を持つという事は、機能語よりも内容語を適切な表現に翻訳することが重要であるということを示唆している。

5.2 実例

次に、実際に翻訳の影響を受けて回答が変化した実例を挙げる。

表 4: 実例 1

○	OR when was interstate 579 formed
-	TR 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

表 4 の例では OR の質問文にのみ正解できた。この例で特に異なるのは interstate 579 の表現である。HT の interstate highway 579 は、interstate highway という音楽アルバムとして理解され、そのリリース日を回答していた。GT の文では、has been が音楽アルバムの名前として解釈され、そのリリース日を回答していた。YT の文では interstate が expressway に変化しており、州間高速道路と解釈できず回答に失敗していた。州間高速道路 579 号は Freebase 上では interstate_579 という名前で登録されており、この表現が変化することでアライメントが弱くなり、他の回答候補が優先されたと考えられる。

表 4 のような回答の変化は、内容語の表現の変化によるものと考えられ、このような例は、訳出の内容語を適切な表現に言い換えることで回答精度が向上する可能性を示している。

逆に、翻訳による表現の変化によって正解できなかったものが正解できた例も見られた。以下の例では、GT と YT の質問文に正解できた。

OR の文では、Mike Schmidt の打撃成績を回答していた。これは、“*play*” が強く評価された結果だと考えられる。HT の文では、“*hold*” が上手く解釈できず、回答を返すクエリを得ることができていなかった。GT と YT の文は同一で、“*position*” と “*mike schmidt*” から正解を導くことができていた。この例

表 5: 実例 2

×	OR what position did mike schmidt play
-	TR マイク・シュミットのポジションは何でしたか
×	HT what position does mike schmidt hold
○	GT what was the position of mike schmidt
○	YT what was the position of mike schmidt

では、翻訳によって内容語である “*play*” や “*hold*” といった語が翻訳によって消えたことによって正解できるようになっており、適切でない内容語が及ぼす影響が現れているといえる。

6. まとめ

本稿では、知識ベースに基づいた言語横断質問応答において、翻訳性能が及ぼす影響を調べるため、質問応答用に作成された質問文を一度日本語に翻訳し、複数の手法で再度英語に翻訳することで翻訳の影響を受けた質問セットを作成し、比較を行った。その結果、翻訳精度と質問応答精度の関係の調査において、内容語を重視する NIST スコアが質問応答精度と最も高い相関を示した。また、内容語の表現による回答の変化を確認し、適切な表現への統一による性能向上の可能性を示した。

今後は、日本語以外の言語と英語との間の翻訳の影響の調査や、さらに詳細な実験結果の考察、本調査により得られた知見を利用しての言語横断質問応答の性能向上を行っていく予定である。

謝辞

本研究を進めるにあたり、ご指導を頂いた知能コミュニケーション研究室 Graham Neubig 助教授、議論を通じて多くの知識や示唆を頂いた水上雅博氏他、知能コミュニケーション研究室の皆様に感謝いたします。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] Tomoyosi Akiba, Kei Shimizu, and Atsushi Fujii. Statistical machine translation based passage retrieval for cross-lingual question answering. In *IJCNLP*, pp. 751–756, 2008.

[2] 兵藤達浩, 秋葉友良. E-017 統計翻訳を用いた言語横断質問応答における翻訳モデルの改善 (自然言語・音声・音楽, 一般論文). 情報科学技術フォーラム講演論文集, Vol. 8, No. 2, pp. 289–292, 2009.

[3] 藤田彬, 松崎拓也, 登藤直弥. 中学生は機械翻訳された英語対話文完成問題を解けるか?(言語理解とコミュニケーション)–(第 6 回集合知シンポジウム). 電子情報通信学会技術研究報告= IEICE technical report: 信学技報, Vol. 114, No. 366, pp. 17–21, 2014.

[4] Qingqing Cai and Alexander Yates. Large-scale semantic parsing via schema matching and lexicon extension. In *ACL*, pp. 423–433. Citeseer, 2013.

[5] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, pp. 1533–1544, 2013.

[6] Haas Carolin and Riezler Stefan. Response-based learning for machine translation of open-domain database queries. In *NAACL HLT*, 2015.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318. Association for Computational Linguistics, 2002.

[8] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *EMNLP*, pp. 944–952. Association for Computational Linguistics, 2010.

[9] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT*, pp. 138–145. Morgan Kaufmann Publishers Inc., 2002.

[10] Gregor Leusch, Nicola Ueffing, and Hermann Ney. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit IX*, pp. 240–247, 2003.

[11] Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. Clueweb09 data set, 2009.