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
Improving Translation Model for SMT-based Cross Language Question Answering

兵藤 達浩* 秋葉 友良*
Tatsuhiro Hyodo Tomoyosi Akiba

1 はじめに

質問応答は、自然言語の質問文に対して組織化されていない検索対象文書から回答を求める技術である。言語横断質問応答とは、質問文と検索対象文書の言語が異なる場合の質問応答である。CLEF[1] や NTCIR[2] にて factoid 型質問を対象とした言語横断質問応答システムの評価が行われてきている。

従来、言語横断質問応答システムを実現するためには、翻訳が鍵となる。前処理によって質問文と検索対象文書の言語を統一すれば、単一言語の質問応答の問題に帰着する。従来法では、質問文を検索対象文書に翻訳してから、単言語の質問応答システムで回答抽出する手法が主であった。一方、先行研究 [3] では、統計翻訳モデルをシステムに組み込み、原言語から直接回答抽出を行う言語横断質問応答を提案している。

そこで本論文では、先行研究 [3] の翻訳モデルおよびそれを用いた類似度計算手法を改善することで、質問応答性能を向上させることを目的とする。なお、本論文では、言語横断質問応答の焦点の 1 つである未知語への対応は行っていない。既知語のみを用いたパッセージ検索の性能改善に焦点をあてる。

2 統計翻訳に基づく言語横断質問応答

先行研究 [3] で提案した、統計翻訳を用いた英日の言語横断質問応答システム全体の構成を、従来の前処理として翻訳を用いる手法とともに図 1 に示す。両者の違いは、前処理として翻訳モデルを用い、単一言語に統一してから質問応答を行うか、質問応答プロセスに翻訳モデルを組み込むことで、前処理の翻訳を行わずに、言語の異なるまま質問応答を行うか、である。前者が図 1 の従来手法 (左)、後者が先行研究手法 (右) となる。

英語質問文が入力されると、システムはまず質問文解析により質問タイプを得る。従来手法では、ここで英語質問文を日本語質問文に翻訳してから質問文を検索キーとして文書検索を行うが、提案手法では英語質問文のまま文書検索を行い、検索スコアの高い日本語文書を得る。次に検索された日本語文書から回答候補を抽出する。その後、回答候補のスコアリングを行うために、従来手法は日本語質問文と日本語文書、提案手法は英語質問文と日本語文書における回答周辺の文脈 (パッセージ) の類似度の計算と、回答候補の質問タイプとのマッチングを並列に行う。スコアリングの後、スコアに基づき回答候補の順位付けを行い、結果を出力する。

* 豊橋技術科学大学

入力質問を翻訳せずに原言語をそのまま用いることによる従来手法からの変更点は、図 1 中の太枠箇所の「文書検索」および「パッセージ類似度計算」である。文書検索では、索引付けに翻訳モデルの単語翻訳確率 $t(e_j|j)$ を用いることで英語質問文から直接日本語文書の検索を行う。パッセージの類似度計算では、質問文と回答周辺の文脈 (パッセージ) の類似度を「パッセージから質問文へ翻訳される確率」と見なして計算する。

2.1 文書検索

提案手法では、英語の質問文から直接日本語文書を検索するために、日本語の検索対象文書を英語で索引付けする。その際、単語翻訳確率を用いて、日本語の索引語頻度から英語の索引語頻度を求める。

文書 D にて索引付けた日本語単語を j、翻訳先の英単語を e とすると、D を英語で索引付けした場合の英単語 e の出現頻度 $tf(e,D)$ は式 (1) で推定できる。

$$tf(e,D) = \sum_{j \in D} tf(j,D)t(e|j) \quad (1)$$

$tf(j,D)$ は D における j の単語頻度、$t(e|j)$ は j から e への単語翻訳確率である。単語翻訳確率は統計的機械翻訳の翻訳モデルと同様に対訳コーパスから求める。$tf(e,D)$ を用いれば、日本語で索引付けした場合と単語の出現頻度 (TF) の点で整合性が保持されるので、ベクトル空間モデルに基づく既存の検索エンジンを用いて英語質問から日本語質問への検索が可能になる。

2.2 パッセージ類似度計算

英語質問文と日本語文書における回答周辺の文脈 (パッセージ) の類似度を、パッセージが質問文に翻訳される確率で計算する。

質問文 Q と、検索対象文書中の回答候補 A が含まれる文 S の類似度を、次式で求める。

$$sim(Q,S|A) = \max_{D \in H(S)} P(Q|D-A) \quad (2)$$

ここで、$P(Q|D-A)$ は単語列 $D-A$(パッセージ D から回答候補 A を除いた単語列) が Q に翻訳される確率、$H(S)$ は S に関するパッセージ候補の集合である。図 2 に質問とそれに対するパッセージの例を示す。質問文 Q とさまざまな組み合わせのパッセージ候補集合 $H(S) = \{\{S\}, \{S_H S\}, ..., \{S_H S_{-1} S S_{+1}\}\}$ の各要素について類似度を計算し、最も高い類似度を選択する。

$P(Q|D-A)$ を計算するモデルとして、先行研究では次の IBM Model 1 を用いている。

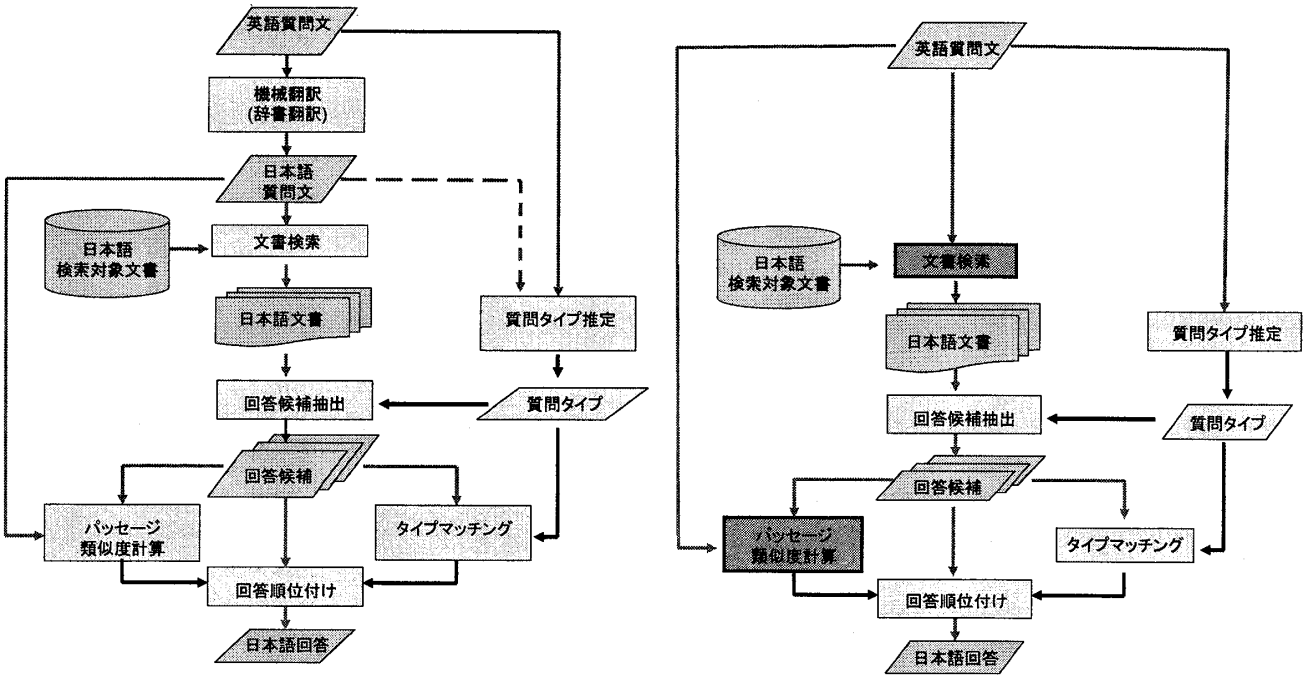
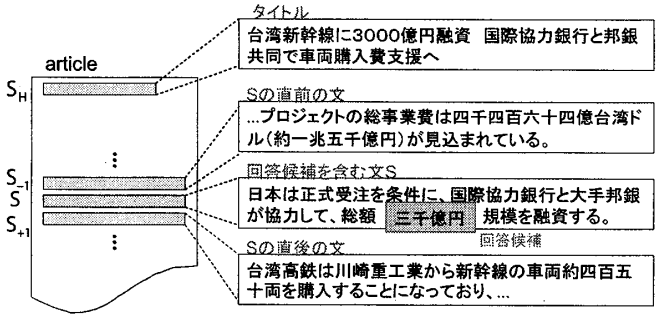


図 1　従来手法 (左) と先行研究手法 (右) のシステム構成

Q *How much did the Japan Bank for International Cooperation decide to loan to the Taiwan High-Speed Corporation?*



$H(S) = \{\{S\}\{S_H S\}\{S_{-1} S\}\{SS_{+1}\}\{S_H S_{-1} S\}\{S_H SS_{+1}\}\{S_{-1} SS_{+1}\}\{S_H S_{-1} SS_{+1}\}\}$

図 2　質問とパッセージの例

$$P(Q|D-A) = \prod_{j=1}^{m} \frac{1}{n-(q-p+1)+1} \sum_{i=0,1,2,\ldots p-1,q+1,\ldots,n} t(q_j|d_i) \quad (3)$$

ここで、$Q = q_1 \ldots q_m$ は質問文の単語列、$D = d_1 \ldots d_n$ はパッセージ ($H(S)$ の一要素)、$A = d_p \ldots d_q$ は S 中の回答候補単語列である。$D - A = d_1 \ldots d_p - 1 d_q + 1 \ldots d_n$ はパッセージから回答候補を除去した単語列である。質問応答の性質から、質問文中には回答に相当する語が現れないので、D ではなく $D - A$ を用いた。

3　翻訳モデルの改善手法

2.2 節の類似度を計算する式 (3) は翻訳モデルのみに依存する式なので、ユニグラム、逆方向の翻訳モデル、IDF を用いて改善する手法を示す。

3.1　日英翻訳とユニグラム線形補間

前述の式 (3) に、英単語のユニグラム (事前確率) 線形補間を行う。

$$P(Q|D-A) = \prod_{j=1}^{m} \frac{1}{n-(q-p+1)+1} \sum_{i=0,1,2,\ldots p-1,q+1,\ldots,n} \{\lambda t(q_j|d_i) + (1-\lambda) u(q_j)\} \quad (4)$$

$u(q)$ がユニグラムで、2.1 節の文書検索の際に付けられた $tf(e, D)$ を元にした英単語の出現数比率 (出現数/総単語数) である。

$$u(e_j) = \frac{\sum_i tf(e_j, D_i)}{\sum_i \sum_j tf(e_j, D_i)}$$

この手法は、類似度の比較にあたり、翻訳確率の総計が 0 となるのを回避するために導入し、スムージングを行うことで、再現率の向上を図っている。

3.2　英日翻訳とユニグラム線形補間

翻訳モデルの方向を逆転した英日方向の翻訳確率と、検索対象文書中の日本語単語の出現頻度を用いたユニグラムを線形補間して、類似度を測る。

$$P(D-A|Q) = \prod_{i=1,2,\ldots p-1,q+1,\ldots,n} \frac{1}{m+1} \sum_{j=0}^{m} \{\lambda t(d_i|q_j) + (1-\lambda) u(d_i)\} \quad (5)$$

式 (4) と比較して、日英の翻訳確率が英日の翻訳確率に、英単語ユニグラムが日本語単語ユニグラムになっている。ユニグラム $u(d)$ は、検索対象文書 (日本語) に出現する日本語単語の出現数比率 (出現数/総単語数) であり、日英方向と同様の、スムージングの役割をする。ここでの注意として、式 (5) で得られる $P(D-A|Q)$ を比較するには、パッセージ (検索対象の、回答候補が含まれる日本語文) の長さが異なるので、正規化を行う必要がある。よって、$P(D-A|Q)^{\frac{1}{n-(q-p+1)}}$ とした値で類似度比較を行う。

3.3 英日ユニグラムと IDF

式 (4),(5) は質問文中の単語を同等に扱った。これを単語の重要度を考慮して重み付けすることを考える。

IDF(逆出現頻度) は単語の重要度を示す数値であり、対象とする単語の出現する文書が、多いほど小さな値、少ないほど高い値となる。そこで、質問文中の単語の IDF 比率を式 (5) に掛け合わせることで重みとし、重要な単語の翻訳確率を重視することで性能の向上を狙っている。

$$P(D-A|Q) = \prod_{i=1,2,..p-1,q+1,...,n} \sum_{j=0}^{m} \frac{idf(q_j)}{\sum_k idf(q_k)} \{\lambda t(d_i|q_j) + (1-\lambda)u(d_i)\} \quad (6)$$

$$idf(q) = log\{\frac{N}{df(q)}\} \quad (7)$$

式 (5) では各単語を同じ重み ($\frac{1}{m+1}$) で平均していたところを、$\frac{idf(q_j)}{\sum_k idf(q_k)}$ で単語ごとに重みを付けている。式 (7) で N は総文書数、$df(q)$ は単語 q の出現する文書数である。また、ここでも 3.2 節と同様のパッセージの長さに基づく正規化を行ってから、比較をする。

3.4 PLSI

以上の手法は、どれも単語対単語の翻訳確率による類似度を計算し、比較することで検索としている。それゆえ、単語間の関係、文脈情報は使われておらず、各単語の持つ意味は全て独立している。そこで、単語間の関係を検索に使うため、PLSI を導入する。

LSI(latent semantic indexing, 潜在意味インデキシング) は、特異値分解を用いて高次元ベクトルの次元削減を行い、相互に関係のありそうな単語の次元を共通の次元に縮退することにより、表層表現が異なる同じ意味の語、共起する語をまとめ、検索精度の改善を図る技術である。それを確率・統計的な枠組ので再定式化したものが PLSI(確率的潜在意味インデキシング) である。

PLSI の利点は、次元を削減することでひとつの単語から複数の可能性を同時に考慮できること、そして単語間の依存関係を扱えることである。ここでは、以下の式で類似度計算を行う。

$$P(D-A|Q) = \prod_{i=1,2,..p-1,q+1,...,n} \sum_k P(d_i|z_k)P(z_k|Q) \quad (8)$$

$$P(z_k|Q) = \frac{\prod_{j=1}^{m}\{P(q_j|z_k)P(z_k)\}}{\sum_{k=0}^{n}\prod_{j=1}^{m}\{P(q_j|z_k)P(z_k)\}} \quad (9)$$

Q は英語質問文を表し、$Q = q_1, ..., q_m$ である。式 (9) で質問文の持つ潜在的なトピック分布 (確率変数 z の分布) を求め、式 (8) で各トピックと関連する語の出現確率を求め、パッセージと比較している。学習では、対訳コーパス中の互いに対訳となる日本語文と英語文を連結して文書を構成し、その文書集合から PLSI のパラメータ学習を行った。

3.5 内容語の抽出

これまでは統計的機械翻訳にならって、翻訳モデルの学習に全単語を用いてきた。しかし機能語などのストップワードは検索には不要である。そこで、ストップワードを除くために、内容語だけを抽出して対訳コーパスを再構築して、翻訳モデルおよび PLSI を学習した。

そのための内容語の判定には、日本語側は形態素解析器 ChaSen、英語側は品詞タガーである TreeTagger を用いて付与される品詞を使い、日本語側、英語側、両方から「名詞・動詞・形容詞・未知語」に該当する単語を抽出し、対訳コーパスとして再構築した。

4 評価実験

4.1 テストコレクション

言語横断質問応答システムの評価に NTCIR CLQA1 テストコレクション [2] の英日サブタスクを用いた。検索対象文書は読売新聞 2 年分 (2000-2001) である。テストコレクションは factoid 型質問 200 問で構成されている。

4.2 翻訳モデルの学習

翻訳モデルを構築するために、以下の文対応の対訳コーパスを用いて学習を行った。

辞書例文	170740 ペア
日英新聞記事対応付けデータ [5]	114404 ペア
ロイター日英記事の対応付け [5]	56782 ペア

上記コーパスのうち、日英新聞記事対応付けデータは読売新聞とその英語版である Daily Yomiuri の文対応で構成される対訳コーパス [5] であるので、CLQA1 で検索対象文書となる 2000,2001 年の対訳は取り除かれている。対訳コーパスは日本語、英語それぞれに対し、前処理を行い正規化した。日本語文に対しては日本語形態素解析器 ChaSen を用いて形態素ごとに区切り、活用語の標準形化を行った。また英語に対しては、品詞タガーを用いて英単語を原型に直し、全ての語を小文字化した。

また、この形態素解析によって付けられる品詞を元に、3.5 節で述べた品詞制限モデルのための対訳コーパスを組み立てる。

翻訳モデルの学習には GIZA++[6] を用い、学習によって IBM Model4 翻訳モデルを得た。IBM Model4 のうち、日英方向の単語翻訳確率 $t(e|j)$ と英日方向の翻訳確率 $t(j|e)$ を、提案する文書検索とパッセージ類似度計算に用いた。

PLSI の学習にも同じ文対応のコーパスを用いるが、学習時の前処理として、英日の対応する文を一文にまとめ、各単語の出現頻度を求めて学習する。トピックの次元数は 100 とした。

4.3 評価結果

評価には回答候補上位 5 位までの MRR(平均逆順位) を用いた。

$$MRR = \frac{1}{N}\sum_{Q=1}^{N}\max_{r\in\{1,\ldots,5\}}(\frac{A_{Q,r}}{r}) \quad (10)$$

$A_{Q,r}$ は Q 番目の質問の第 r 位の回答を意味し、正解なら 1、不正解なら 0 となる。N は質問数 (今回は N =200 問) である。順位の逆数で集計するため、多くの質問で高い順位が出るほど高得点となり、最終的に、平均第何位で回答が得られるかを表す指標となる。

CLQA1 テストコレクションとして配布された正解セットにより、文字列の一致のみで正解判定をした。結果を表 1 にまとめる。

表 1　評価結果 (MRR)

	通常モデル	内容語モデル
日英翻訳	0.1605	0.1602
日英ユニグラム	**0.1621**	**0.1740**
英日ユニグラム	0.1158	0.1035
IDF+ 英日ユニ	0.1294	0.1200
PLSI	0.0744	0.0653

表 1 より、英日翻訳よりは日英翻訳を用いる方が効果があることが分かった。IDF 重み付けは性能改善を示したが、日英翻訳の手法には及ばなかった。PLSI は良い結果を得られなかった。日英翻訳とユニグラムを組み合わせたときの内容語モデルの結果のみ、通常モデルの結果を上回ることができた。以上のことから、検索に対する IDF 重み、ユニグラム線形補間と内容語の組合せの有効性が示された。

日英翻訳と英日翻訳では、翻訳方向が入れ替わるだけで大きく違いが生じた。単語翻訳確率、計算途中の値など、細部を検証する必要がある。ユニグラム線形補間と IDF 重み付けは、当初の予想通りに改善できた。しかし、ユニグラム比率など、改善の余地はあるので続けて検証を行いたい。PLSI は、各トピックの単語分布を調べたところ、トピック間での分布の差が見られず、学習がうまくいっていないことが分かった。内容語モデルについては、今回は品詞で検索の役に立つと思われる内容語を選択したが、品詞以外の解析情報 (例えば単語頻度の多すぎるものに注目するなど) も使って、もう少し細かく限定することで改善の余地があると考える。

5　まとめ

本稿では、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を行った。評価実験の結果、日英翻訳のユニグラム線形補間で内容語のみから成る翻訳モデルを用いた場合が、最も良い結果となった。

今後は、日英翻訳と英日翻訳の両モデル間の性能差の原因の調査、利用する内容語の選択による内容語モデルの改善、PLSI 学習方法についての再検討を行っていく予定である。

また、本稿では対象としていなかったが、対訳コーパスに現れない未知語への対応の検討、パッセージ検索と回答タイプ一致判定の統合法の検討も行う予定である。

参考文献

[1] B.Mgnini et al. The Multiple Language Question Answering Track at CLEF 2003. In Working Notes for CLEF 2003 Workshop, 2003.

[2] Y.Sakaki et al. Overview of the NTCIR-5 Cross- Langual Question Answering Task(CLQA1). In Proceedings of the Fifth NTCIR Workshop, pp.175-185, 2005.

[3] Tomoyosi Akiba, Kei Shimizu, Atsushi Fujii, and Katunobu Itou. Statistical Machine Translation based Passage Retrieval for Cross-Lingual Question Answering — Experiments at NTCIR-6, NTCIR Workshop 6 Meeting pp.216-221

[4] 清水 慧 他. 統計翻訳に基づくパッセージ検索の言語横断質問応答への適用. NLP-2007, pp.1176-1179, 2007.

[5] M.Utiyama et al. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In ACL-2003, pp.72-79, 2003.

[6] F.J.Och. GIZA++: Training of statistical translation model. http://www-i6.informatik.rwth-aachen.de/Colleague/och/software/GIZA++.html