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

清水 慧[†]　秋葉 友良[†]　藤井 敦[‡]
[†] 豊橋技術科学大学　[‡] 筑波大学

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、事実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。

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できる。翻訳に用いる手法によって、従来の手法は機械翻訳システムを用いる手法 [3]、対訳辞書を用いる手法 [4] の 2 つに分類できる。

本論文では、統計的機械翻訳 [5] を利用した言語横断質問応答手法を提案する。提案手法は、前処理として翻訳を用いる従来手法とは異なり、統計翻訳モデルを質問応答プロセスに組み込むことで言語横断質問応答を実現する。提案手法は、どの言語対へも適用可能であるが、本論文では英日、つまり英語質問文から日本語文書を検索する言語横断質問応答を対象とする。

近年、情報検索の分野において、検索を確率モデルの推定問題として扱う言語モデリングアプローチが提案され、活発な研究分野となっている。それらの一手法として、統計的機械翻訳からの連想による翻訳モデルを用いた検索モデルが提案され、単一言語文書検索 (Mono-lingual IR) [6]、言語横断文書検索 (Cross-lingural IR) [7]、単一言語質問応答 (Mono-lingual QA) [8] に適用されている。提案法は、翻訳モデルに基づく検索モデルを言語横断質問応答 (Cross-lingual QA) に適用した手法と位置づけることができる。

また、言語横断質問応答の研究では、質問に現れる未知語 (辞書にない固有名、略語、外来語、表記の揺れ、など) への対処法に焦点を当てた研究が多い。例えば、Web を用いた固有名の翻訳 [3] や外来語の翻字手法 [4] が提案されている。しかし、本論文では、未知語対策は対象としない。[1]

[1] 評価に用いたシステムでも未知語対策は行っていない。

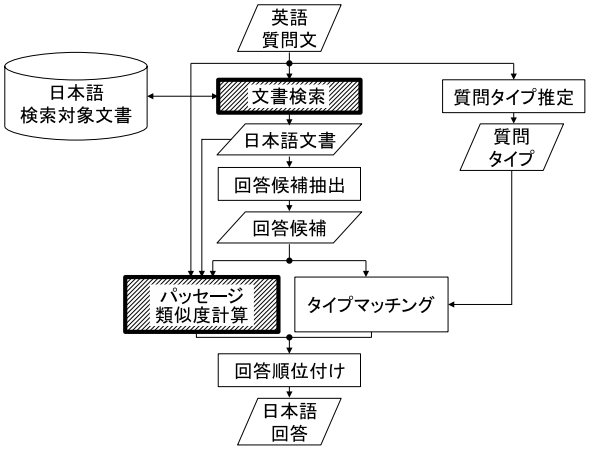

図 1: 言語横断質問応答システムの構成

2　提案手法

提案手法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図 1 に示す。図 1 の構成は、入力が英文であることを除けば、単一言語質問応答システム [9] とまったく同じである。

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。同時に、質問文を検索キーとして文書検索を行い、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、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。

入力が英文になったことによる単一言語質問応答システムからの変更点は、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 $t(e_{\mathrm{j}}|j)$ 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。

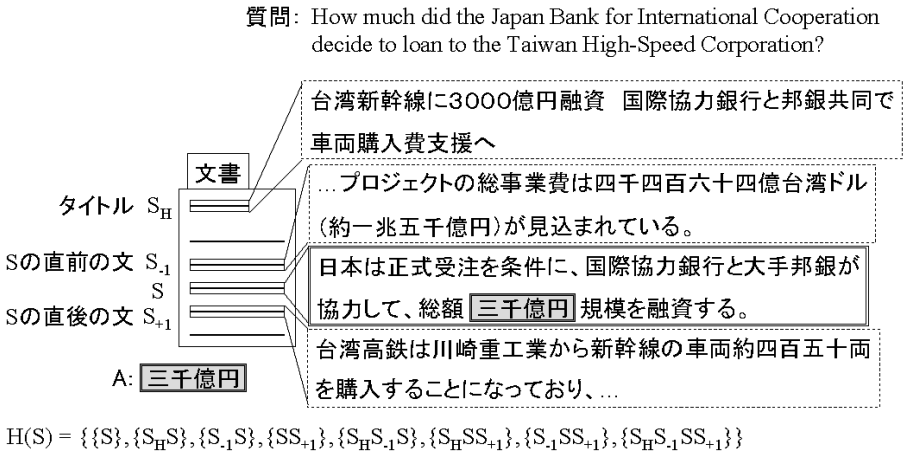

図 2: 質問とパッセージの例

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 $\mathfrak{D}$ 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、$\mathfrak{D}$ を英語で索引付けした場合の英単語 e の出現頻度 $tf(e,\mathfrak{D})$ は式 (1) で推定できる。

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

$tf(j,\mathfrak{D})$ は $\mathfrak{D}$ 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。$t(e,\mathfrak{D})$ 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文書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。

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

ここで、$P(Q|D)$ は単語列 D 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、様々な組み合わせのパッセージ候補集合 $H(S) = \{\{S\},\{S_{-1}S\},\ldots\{S_{\mathrm{H}}S_{-1}SS_{+1}\}\}$ 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、次の IBM Model 1 を用いた。

$$P(Q|D-A) = \frac{1}{(n+1)^m}\prod_{\mathrm{j}=1}^{m}\sum_{\mathrm{i}=1,2,\ldots\mathrm{p}-1,\mathrm{q}+1,\ldots,\mathrm{n}} t(q_{\mathrm{j}}|d_{\mathrm{i}}) \qquad (3)$$

ここで、$Q = q_1\ldots q_{\mathrm{m}}$ は質問文の単語列、$D = d_1\ldots d_{\mathrm{n}}$ はパッセージ ($H(S)$ の一要素)、$A = d_{\mathrm{p}}\ldots d_{\mathrm{q}}$ は S 中の回答候補単語列である。$D-A = d_1\ldots d_{\mathrm{p}-1}d_{\mathrm{q}+1}\ldots d_{\mathrm{n}}$ 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 評価実験

3.1 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] および CLQA2 テストコレクションの英日サブタスクを用いた。検索対象文書はそれぞれ読売新聞 2 年分 (2000-2001)、毎日新聞 2 年分 (1998-1999) である。テストコレクションはそれぞれ事実型質問文 200 問で構成されている。

3.2 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。

- 英辞郎 例文 44841 ペア
- 新編 英和活用大辞典 例文 90613 ペア
- ランダムハウス英語辞典 例文 35285 ペア
- 日英新聞記事対応付けデータ [10]
  114404 ペア (CLQA1), 150000 ペア (CLQA2)
- ロイター日英記事の対応付け [10] 56782 ペア

上記のコーパスのうち、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であるので、CLQA1 での評価では、検索対象文書と重なる 2000〜2001 年の対訳を取り除いた。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、英文に対しては、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。翻訳モデルの学習には GIZA++ [11] を用い、学習によって IBM Model 4 翻訳モデルを得た。IBM Model 4 のうち、日英方向の単語翻訳確率 $t(e_j|j)$ を、提案する文書検索とパッセージ類似度計算に用いた。

3.3 比較手法

従来法として、機械翻訳による手法と、対訳辞書による手法との比較を行った。機械翻訳を用いた手法では、まず英文を機械翻訳システムで日本語文に翻訳し、次にそれを日本語質問応答システムに入力し、回答を得た。ただし、質問タイプ推定は、提案法と同様に、英文から行った。機械翻訳システムには、市販の機械翻訳ソフト (以下、この手法を***MT*** と記す) [2] と統計的機械翻訳 (***SMT*** と記す) を使用した。統計的機械翻訳では 3.2 節で構築した IBM Model 4 翻訳モデル[3] と、CLQA1 英日サブタスクの検索対象文書である読売新聞 2 年分 (2000-2001) を用いて学習した単語 3-gram 言語モデルを用いた。

対訳辞書を用いる手法として、藤井らによる言語横断文書検索 [12] システムを利用した (***Dict*** と記す)。藤井らの手法には、対象文書の共起情報を利用した複合語の翻訳や、英語と日本語の文字単位の類似度に基づいた翻字手法が用いられている。検索された文書に対して、提案手法と同じプロセスで回答を抽出した。すなわち、図 1 における「文書検索」を藤井らの言語横断文書検索に置き換えたものに相当する。

さらに、単一言語質問応答との性能比較のため、テストコレクションに含まれる質問の日本語訳を入力として日本語質問応答を行った (***JJ*** と記す)。これは理想的な翻訳を行った場合に相当する。

また、提案手法のバリエーションとして、

Proposed : 2 節で説明した手法

Proposed +r : 回答候補のスコアリングに文書検索スコアも反映させた場合

Proposed -p : 回答周辺文脈として常に S だけを用いる場合 (すなわち $H(S) = \{S\}$ の場合)

Proposed -p+r : 上記二つの組み合わせ

について性能を比較した。

3.4 評価結果

評価尺度として、回答候補上位 1 位における精度 (Top1 精度)、回答候補上位 5 位までに正解が含まれる精度 (Top5 精度)、回答候補上位 5 位までの MRR を用いた。表中の**R** は文字列の一致と回答抽出文書の一致の両方で正解判定した場合、**R+U** は文字列の一致のみで正解判定した場合である。CLQA1 はテストコレクションとして配布された正解セットによる評価、CLQA2 は参加チームに対して行われた人手による評価の結果である。評価結果を表 1 に示す。[4]

まず、CLQA1 と CLQA2 を比較する。"***JJ***"の TOP1 精度を比較すると、**R** について CLQA1 の結果は CLQA2 に比べかなり低い (56%) のに対し、**R+U** では値がほぼ一致している。質問セットを比べても CLQA1 と CLQA2 では質問の傾向や難易度に大きな差はないと考えられるため、CLQA1 の**R** に関する正解セットは信頼できないと考える。そこで、CLQA1 については以降**R+U** での比較を行う。

次に、提案手法***Proposed*** と従来手法を比較する。CLQA1 の**R+U** を見ると、機械翻訳を用いた 2 手法 (***MT***,***SMT***) はほぼ同じ性能を示している。これらに対し、***Proposed*** は、評価尺度により、1.3〜1.5 倍の性能を示した。特に、***SMT*** は提案手法と同じ対訳コーパスで学習した翻訳モデルを用いていることから、前

[2] 日本アイビーエム株式会社「インターネット翻訳の王様バイリンガル Version 5」を使用

[3] ただし、日本語文に対して標準形化を行っていない

[4] CLQA2 については、NTCIR で評価中のため、現時点では参加システムの TOP1 精度のみ結果が得られている

表 1: 評価結果

手法	CLQA1						CLQA2	
	R			**R+U**			**R**	**R+U**
	Top1 精度	Top5 精度	MRR	Top1 精度	Top5 精度	MRR	Top1 精度	Top1 精度
JJ	0.140	0.300	0.196	0.260	0.535	0.354	0.250	0.275
MT	0.020	0.075	0.039	0.065	0.175	0.099		
SMT	0.015	0.070	0.034	0.060	0.175	0.098		
Dict	0.055	0.100	0.070	0.095	0.195	0.134	0.070	0.100
Proposed	0.045	0.125	0.074	0.090	0.225	0.146	0.135	0.170
Proposed +r	0.050	0.140	0.083	0.105	0.285	0.173	0.125	0.160
Proposed -p	0.040	0.120	0.069	0.105	0.245	0.155		
Proposed -p+r	0.055	0.155	0.091	0.120	0.280	0.178		

処理の代わりに翻訳モデルを組み込む提案手法の効果があったと考えられる。対訳辞書を用いた手法 (***Dict***) は、CLQA1 では***Proposed*** に近い性能を示している。しかし、CLQA2 の評価では***Proposed*** が 1.7〜1.9 倍の性能を示した。

続いて、提案手法間で比較する。CLQA1 の**R+U** では、***+r*** と***-p*** の両方とも性能を改善している。しかし、CLQA2 では***+r*** の効果は見られない。さらなる調査が必要と考える。

また、***JJ*** との比較から、提案手法は単一言語質問応答の約半分の性能であることが分かる。

4 まとめ

本稿では、統計翻訳に基づくパッセージ検索を用いた言語横断質問応答システムを提案した。評価実験の結果、従来の機械翻訳を用いる手法や、対訳辞書を用いる手法に比べて、高い性能を得ることができた。

提案手法のパッセージ類似度計算では IBM Model 1 を用いた。今後は、IBM model 2〜5 などのより精密なモデルや、質問応答に適した新たなモデルを検討していきたい。

参考文献

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

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

[3] T.Mori et al. A Method of Cross Language Question-Answering Based on Machine Translation and Transliteration. - Yokohama National University at NTCIR-5 CLQA1 - In Proceedings of the Fifth NTCIR Workshop, pp.215–222, 2005.

[4] H.Isozaki, et al. NTT's Japanese-English Cross-Language Question Answering System. In Proseedings of the Fifth NTCIR Workshop, pp.186–193, 2005.

[5] P.F.Brown et al. The mathematics of statistical machine translation: Parameter estimation. In Computational Linguistics 19(2), pp.263–311, 1993.

[6] A.Berger et al. Information Retrieval as Statistical Translation. In Proceedings of the 22nd Annual Conference on Research and Development in Information Retrieval (ACM SIGIR), pp.222–229, 1999.

[7] J.Xu et al. Evaluating a Probabilistic Model for Cross-lingual Information Retrieval. In Proceedings of the 24th Annual International ACM SIGIR Conference, pp.105–110, 2001.

[8] V.Murdock et al. Simple translation models for sentence retrieval in factoid question answering. In itshape Proceedings of the Information Retrieval for Question Answering Workshop SIGIR, pp.31–35, 2004.

[9] 秋葉 友良 他. 質問応答における常識的な解の選択と期待効用に基づく回答群の決定. 情報処理学会研究報告, 2004-NL-163, pp.131–138, Sep. 2004.

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

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

[12] A.Fujii and T.Ishikawa. Cross-Language Information Retrieval at ULIS. In Proceedings of The First NTCIR Workshop, pp.163–169, 1999.