Web 関連度と確率的翻訳モデルを併用した質問応答システム

阿部 裕司（東京農工大学大学院 情報工学専攻）
森田 一（東京工業大学 知能システム科学専攻）
古宮 嘉那子（東京農工大学大学院 工学研究院）
小谷 善行（東京農工大学大学院 工学研究院）

Question Answering System Using Web-Relevance and Probabilistic Translation Model

Yuji Abe (Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology)
Hajime Morita (Department of Computational Intelligence and System Science, Tokyo Institute of Technology)
Kanako Komiya (Institute of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institute of Engineering, Tokyo University of Agriculture and Technology)

１．はじめに

質問応答タスクは、与えられた自然文の質問に対し回答自体を検索して出力するタスクである。質問応答における回答候補評価モジュールは、文書検索で取得した回答候補に対して詳細な分析・評価を行なうもので、質問応答システムの回答性能に直結する重要な要素である。回答候補は「内容の関連度（質問文に対しどれだけ近い内容が記述されているか）」と「記述の回答らしさ（質問文の記述形式に対応した記述が含まれているか）」によってスコア付けされる。

本稿では、既存の回答候補評価指標二つを統合する候補評価式を用い、より柔軟に回答候補を評価する手法を提案する。

２．既存研究とその理論

石下(2009)では「内容の関連度」を、Web を適用した擬似適合フィードバックによって評価している。これは、質問文の内容語をクエリにして文書検索を行なった際に検索文書内に頻出する語を関連語とみなし、その関連語を多く含む回答候補は「内容の関連度」が高いとみなすものである。

「内容の関連度」と「記述の回答らしさ」を同時に判定する方法として、Soricut(2006)は確率的翻訳モデルを利用した手法を提案している。この手法では、質問文を翻訳前の文、回答文を翻訳後の文とみなし、個々の単語ごとの翻訳確率を QA コーパスから学習する。そして新しい質問が入力された際に、その質問が回答候補文に翻訳される確率を利用して回答候補評価をおこなう。翻訳モデルとしては、単純だが多くのタスクで有効性が確認されている IBM-Model1(Brown(1993))を改変したものを利用している。質問応答タスクにおいて、本来の IBM-Model1 を用いた定式化は式(4.1),(4.2)のようになる。

$$A^* = \arg\max_A p(A \mid Q) = \arg\max_A p(Q \mid A)p(A) \tag{4.1}$$

$$p(Q \mid A) \approx \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} p(q_j \mid a_i) \tag{4.2}$$

式(4.1),(4.2)において、A^*は最も適切と思われる回答候補文、A $(=a_1, a_2, ..., a_l)$は回答候

補文、$Q(=q_1,q_2,...,q_m)$は質問文、lは回答候補文の単語数、mは質問文の単語数、$p(q|a)$は質問側の単語aが回答側の単語qに翻訳される確率、$p(A)$は回答候補文Aの生成確率、εは回答文から単語数mの質問文が生成される確率である。式(4.2)において、相乗記号の外側の係数はlが小さいほど大きな値となり、したがって回答候補文の単語数が少ないほど文全体の翻訳確率が大きくなってしまうという問題がある。そのため、Soricut(2006)では、式(4.2)の相乗記号の外側の係数を無視した定式化を行なっている。

$$p(Q|A) \approx \prod_{j=1}^{m} \sum_{i=0}^{l} p(q_j|a_i) \tag{4-3}$$

３．既存手法を統合した回答候補評価

「内容の関連度」を評価する際、Soricut(2006)のように回答候補評価に確率的翻訳モデルを用いた場合は同義語や類義語などを柔軟に考慮した評価が可能となるが、複数の語彙に対する共起関係などの情報は利用することができない。一方、石下(2009)のように Web 等を利用した擬似適合フィードバックを用いた場合、複数の語彙に対する共起関係をうまく利用することができるが、類義語などを柔軟に考慮することは困難である。これらを統合的に利用することで、回答候補評価の性能向上が期待できる。

３．１．回答候補評価式

式(4.1)において最大化されるべき部分を$\wp(Q,A)$とおき、石下(2009)が提案する Web 上の内容関連度に基づく回答候補スコアを$Web_relevance(Q,A)$とおく。両スコアを組み合わせた最終的な評価式を次式で定義する。

$$EvalScore(Q,A) = \wp(Q,A)^{1-\gamma} \cdot Web\_relevance(Q,A)^{\gamma} \tag{4-4}$$
$$\wp(Q,A) = p(Q|A)p(A) \tag{4-5}$$

上式において、γは手法の混合比を決める混ぜ合わせパラメータである。$\gamma = 0$ならば翻訳確率単体の評価値、$\gamma = 1$ならば Web 上の内容関連度単体の評価値となる。

３．２．Web 上で評価する内容の関連度の計算

石下(2009)と同様の手法を採用する。まず、入力された質問文から内容語(名詞,動詞,形容詞)を取得し、キーワード集合Kとする。次に、Kに含まれる語の三つ組を全通り作り、それぞれの三つ組から論理積検索のクエリを構成し、Web 検索エンジンにおいて検索をする。$|K|<3$の場合は全ての語の論理積検索を構成する。そして、それぞれのクエリに対してスニペット(Web 検索エンジンが出力した、検索結果ページの要約)を最大上位 100 件取得し、このスニペット内の各内容語w_jを関連語とする。各関連語の関連度$T(w_j)$は次式で定義する。

$$T(w_j) = \max_i \frac{freq(w_j,i)}{n_i} \tag{4-6}$$

iはクエリ(内容語の三つ組)の番号、n_iはi番目のクエリを利用して取得したスニペットの件

数、 $freq(w_j, i)$ は i 番目のクエリで取得したスニペットのうち、w_j を含むものの件数である。Web を利用した内容関連度による回答候補の評価値 $Web\_relevance(Q, A)$ を、回答候補に含まれる語の関連度の総和で定義する。

$$Web\_relevance(Q, A) = \sum_{i=1}^{l} T(w_i) \qquad (4\text{-}7)$$

上式において、Qは質問文、A は回答候補文、l は回答候補文の単語数、w_i は回答候補文に含まれる単語である。

３．３．翻訳確率の学習

学習コーパスとして『Yahoo!知恵袋データ』を用い、翻訳確率の学習ツールとして IBM-Model1 の C++実装である GIZA++[Casacuberta07]を用いた。学習時の EM アルゴリズムのイテレーション回数は 5 回に設定した。学習時に、単語数の多すぎる質問応答事例、質問側と回答側の単語数に差がありすぎる質問応答事例は単語アラインメントの学習に悪影響を与えると考え、質問側または回答側の単語数が 60 語を超える事例、および回答側と質問側の単語数に 5 倍以上の差があるものは学習から除外した。結果として、質問応答事例 1,092,144 件を利用して翻訳確率を学習した。

４．評価実験

提案する回答候補評価式を用いた質問応答システムを実装し、質問応答実験を行なった。

４．１．実験内容

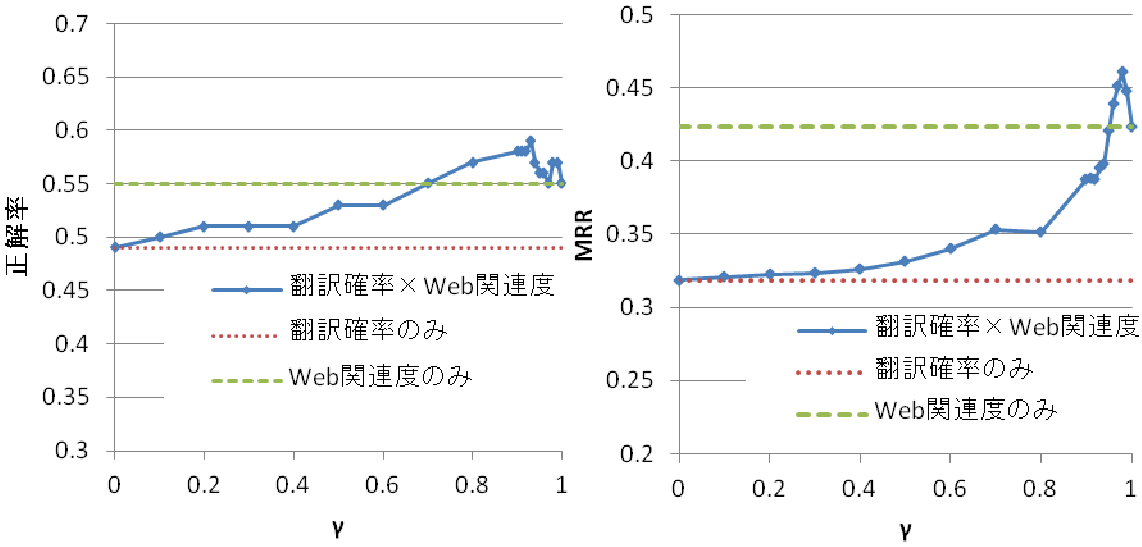
『NTCIR-ACLIA2』(Teruko(2010))の質問応答テストセット 100 問に対し、質問応答実験を行なった。関連文書検索、回答候補抽出は Web 文書を対象として行ない、質問文に含まれる内容語で論理積検索を行ない抽出された文書の出力順位上位 50 件を関連文書とした。抽出された回答候補すべてに対し式(4-4)に基づいた回答候補評価を行ない、スコアの高いものから 5 件を最終的なシステム出力とした。$p(A)$ の計算には通常のバイグラムモデルを単語数で正規化したものを用いた。システムが出力した回答が正解かどうかの判定は人手で行ない、システム性能の評価指標として、Top-5 正解率と MRR を用いた。

４．２．実験結果

γ の値を[0:1]の範囲で変化させていったときの、Top-5 正解率および MRR を図 1 に示す。また、既存手法と提案手法の性能比較を表 1 に示す。γ =0.93 のとき正解率は最大値の 0.59 を示し、γ =0.98 のとき MRR は最大値の 0.461 を示した。ウィルコクソンの符号付順位和検定を行なったところ、MRR については 5%有意水準で既存手法より提案手法が有意に優れていることが示された。

４．３．まとめと展望

本稿では、既存の回答候補評価指標を組み合わせて、「内容の関連度」をより柔軟に評価する手法を提案した。その結果、提案手法は各既存手法よりも優れた質問応答性能を示した。現行のシステムでは「記述の回答らしさ」に関しては十分に評価できていないので、語彙パターンなどを用いて「記述の回答らしさ」を評価する指標を導入することが、今後の課題である。



(a)Top-5 正解率　　(b)MRR

図１：提案手法の質問応答実験結果

表１：提案手法と既存手法の性能比較

	Top-5 正解率	MRR
既存・翻訳確率(γ=0)	0.49	0.318
既存・Web 関連度(γ=1)	0.55	0.423
提案手法(γ=0.93)	**0.59**	0.395
提案手法(γ=0.98)	0.57	**0.461**

謝　辞

本研究を行なうにあたり、ヤフー株式会社が国立情報学研究所に提供した『Yahoo!知恵袋データ』を利用させて頂きました。利用を快諾して下さった各社に感謝いたします。また、評価実験の際に NTCIR の質問応答テストコレクション『NTCIR-8 ACLIA2』を利用させて頂きました。NTCIR の運営にご尽力をいただいている皆様に感謝いたします。

文　献

石下円香、佐藤充、森辰則(2009)「Web 文書を対象とした質問の型に依らない質問応答手法」人工知能学会論文誌,24 巻 4 号, pp.339-350.

Radu Soricut, Eric Brill(2006)”Automatic question answering using the web: Beyond the factoid” Journal of Information Retrieval - Special Issue on Web Information Retrieval, 9,pp.191-206.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer (993) “The Mathematics of Statistical Machine Translation: Parameter Estimation” Computational Linguistics, 19(2), pp.263-311(1993).

Teruko Mitamura, Hideki Shima, Tetsuya Sakai, et al(2010) ”Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access” In Proceedings of 8th NTCIR Workshop Meeting.