T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	言語に非依存な統計的アプローチによる日本語質問応答システムの構築
Title(English)	
著者(和文)	Julien Hamonic, Edward Whittaker, 古井貞熙
Authors(English)	Julien Hamonic, Edward Whittaker, SADAOKI FURUI
出典(和文)	情報処理学会第68回全国大会講演論文集, Vol. 2, No. , pp. 391-392
Citation(English)	, Vol. 2, No. , pp. 391-392
発行日 / Pub. date	2006, 3
権利情報 / Copyright	ここに掲載した著作物の利用に関する注意: 本著作物の著作権は（社）情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。 The copyright of this material is retained by the Information Processing Society of Japan (IPSJ). This material is published on this web site with the agreement of the author (s) and the IPSJ. Please be complied with Copyright Law of Japan and the Code of Ethics of the IPSJ if any users wish to reproduce, make derivative work, distribute or make available to the public any part or whole thereof.

Powered by T2R2 (Tokyo Institute Research Repository)

言語に非依存な統計的アプローチによる日本語質問応答システムの構築

Julien Hamonic † Edward Whittaker † 古井 貞熙 †
† 東京工業大学 大学院情報理工学研究科 計算工学専攻

1 はじめに

我々はこれまでに，質問応答 (QA) システムの構築方法として，対象言語に依存しない data-driven の統計的アプローチを提案している [1]．これまでの QA システムでは，対象言語に依存した知識に基づいたヒューリスティックなルールや固有名詞情報を利用して，質問文解析と解答の絞りこみが行われていた [2, 3]．それに対し，提案手法は質問文とそれに対する解答 (q-and-a) の多数の実例から統計量を求め，文書データの中から解答抽出を行う．我々の先行研究では，この統計的アプローチを英語 QA システムの構築に利用し，TREC タスクによってそのシステムを評価することで，手法の有効性の検証を行った [1]．そこで本稿では，本手法を日本語 QA システムの構築に適用し，NTCIR-3 の Question Answering Challenge (QAC-1) タスクによってシステムの性能評価を行う．

2 QA のための統計的アプローチ

本研究では l_A 単語から成る解答 $A = a_1, \ldots, a_{l_A}$ が l_Q 単語から成る質問 $Q = q_1, \ldots, q_{l_Q}$ にのみ依存すると仮定する．さらに，質問文 Q は 関数 $\mathcal{W}$ と $\mathcal{X}$ によって取り出される 2 つの要素 ($W = \mathcal{W}(Q), X = \mathcal{X}(Q)$) から構成されるとする．

$W = w_1, \ldots, w_{l_W}$ は質問文の「タイプ」を表す l_W 個の特徴量である．$\mathcal{W}$ では，質問文 Q 中に含まれる「いつ」「どこ」といった質問のタイプを表す単語を取り出し，これらの単語の m 個以下の連鎖パターンを全て抽出して W の要素とする．これらの単語を抽出するためのリストは，予め質問文の実例データから上位の頻出単語を取り出すことで作成しておく．

$X = x_1, \ldots, x_{l_X}$ は質問文の主題に関する情報を表す l_X 個の特徴量であり，質問文 Q 中のキーワードから生成される．$\mathcal{W}$ と同様に，質問文 Q から連続する n 個以下のキーワード連鎖のパターンを全て抽出し，X の要素とする．なお，その際，予めテキストコーパスから求めた出現頻出の高い単語と，$\mathcal{W}$ で用いた質問タイプを表す単語は Q 中から除いておく．

Q に対する A の事後確率 $P(A|Q)$ は

$$P(A|Q) = P(A|W, X), \tag{1}$$

と表すことができる．これを最大化することで尤もらしい解答 $\hat{A}$ を求めることができる．

$$\hat{A} = \arg\max_A P(A|W, X). \tag{2}$$

この式は，複数の仮定の下で，次式のように書き変えることができる [1]．

$$\hat{A} = \arg\max_A P(A|X)^{\alpha} \cdot P(W|A). \tag{3}$$

$P(A|X)$ は X を与えた時の解答候補 A の生起確率であり，X を用いて解答候補となる A を検索するためのモデルである．そこで，この部分を「検索モデル (retrieval model)」と呼ぶことにする．

一方，$P(W|A)$ は解答候補 A と 単語群 W の適合度を表しており，検索モデルによって抽出された複数の解答候補に対し，質問タイプとの適合度を用いてスコアの再計算（フィルタリング）を行うためのものである．そこで，この部分を「フィルタモデル (filter model)」と呼ぶことにする．

なお，α は検索モデルとフィルタモデルの効果のバランスをとるためのパラメータである．

2.1 検索モデル

テキストコーパス S（文書数 $|U|$）からキーワード群 X を用いて解答候補 A を検索するためのモデル $P(A|X)$ は，以下のようにして計算される．

まず，$X_i = x_1 \cdot d_{i,1}, x_2 \cdot d_{i,2}, \ldots, x_{l_X} \cdot d_{i,l_X}$ となる X_i を準備する．ここで，$\vec{d_i} = [d_{i,1}, \ldots, d_{i,l_X}]$ の各要素は，$i = \sum_{j=1}^{l_X} 2^{j-1} d_{i,j}$ の解として求まる 0 か 1 の値である．$P(A|X)$ はこれら $2^{l_X} - 1$ 個の要素に対する事後確率 $P(A|X_i)$ を用いて，次式で推定される．

$$P(A|X) = \sum_{i=1}^{2^{l_X}-1} P(A|X_i)/(2^{l_X} - 1), \tag{4}$$

$P(A|X_i)$ はコーパス S 中の各文における A と X_i の出現頻度から計算される最尤推定値として定義される．

2.2 フィルタモデル

$P(W|A)$ は，K 個の q-and-a の実例 e_k ($k = 1, \cdots, K$) を用いて，以下のように計算される．

$$P(W|A) = \sum_{k=1}^{K} P(W|e_k) \cdot P(e_k|A). \tag{5}$$

$e_k = (Q^k, A^k) = (q_1^k, \ldots, q_{l_{Q^k}}^k, a_1^k, \ldots, a_{l_{A^k}}^k)$ であり，A^k の j 番目の単語が A の j 番目の単語にのみ関連があると仮定すると，

$$P(W|A) = \sum_{k=1}^{K} P(W|e_k) \cdot \prod_{j=1}^{l_{A^k}} P(a_j^k|a_j). \tag{6}$$

となる．ここで，事前に作成した $|T|$ 個の単語クラス c_t ($t = 1, \cdots, T$) を用いて，次のように書き換え，計算を行う [1]．

$$P(W|A) = \sum_{k=1}^{K} P(W|e_k) \prod_{j=1}^{l_{A^k}} \sum_{t=1}^{T} P(a_j^k|c_t) P(c_t|a_j). \tag{7}$$

3 評価実験

3.1 実験条件

利用する全ての文書データを単語単位に区切るために，日本語形態素解析器（Chasen 2.3.3）と辞書（IPADIC 2.7.0）用いた．ただし，品詞などの形態素情報は一切用いていない．

フィルタモデルの学習のために，5TAKU クイズデータ [4] から抽出された $K = 268,531$ 個の q-and-a の実例を利用した．また，毎日新聞データ 2 年分（1998-1999）から抽出した出現頻度の上位 215,000 単語を用いて $T = 5$k 個の単語クラスを作成した．

検索モデルによって解答の抽出を行う対象のデータとして，NTCIR-3 QAC-1 の公式文書データである毎日新聞データ 2 年分（mai）と，Web 検索用に集められた NTCIR-3 WEB タスクの 10GB の文書データ「NW10G-01」（www）の 2 つを用意し，それぞれを用

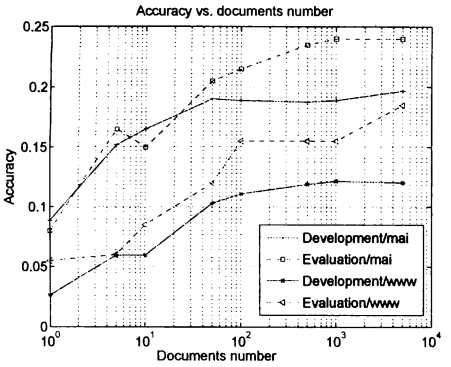

図 1: mai, www のそれぞれを利用した場合の検索対象文書数に対する Top1 accuracy の推移.

表 1: mai (5k 文書), www (5k 文書), mai+www (10k 文書) を使ったときの QA システムの性能.

Data	Accuracy	MRR	F-score
mai	0.240	0.316	0.150
www	0.185	0.237	0.106
mai+www	0.265	0.340	0.159

いた場合についての評価を行った. 実際には, 検索エンジン akechi-2.0.1b [5] に質問文中のキーワードを入力することで, 双方のデータについて, 質問ごとに $|U| = 1, 5, 10, 50, 100, 500, 1k, 5k$ 個の文書を取りだし, それを検索モデルの検索対象データ S とした. また, (mai) と (www) から抽出されたそれぞれ 5k 個の文書を組み合わせて利用した場合の実験 (mai+www) も行った.

なお, 質問文からのキーワード抽出を行うために, 除去する高頻出単語のリストが必要となるが, 本実験では, 毎日新聞データから抽出した出現頻度の上位 75 単語をそれに用いた. 質問タイプを表す単語のリストには, 5TAKU クイズデータの質問文から得られる上位 125 単語を用いた.

3.2 パラメータの最適化

関数 $\mathcal{X}, \mathcal{W}$ で用いる m, n, 検索モデルとフィルタモデルのバランスを決定する α などのパラメータを最適化するため, まず, QAC-1 の Additional Run に含まれる 757 問を用いた. ここでは, QAC-1 のタスク 1 による評価のみを行い, 得られる Top1 accuracy が最大になるように最適化を行った. パラメータの最適化を文書数 $|U| = 5k$ の条件で行った結果, mai, www どちらの場合にも, $m, n = 3, \alpha = 2.0$ となった.

これらの最適パラメータを用いて, 他の検索文書数における性能評価を行った結果を, 図 1 の実線で表す. どちらのデータを利用した場合でも, 文書数の増加に伴って性能が向上している様子が分かる.

3.3 実験結果

得られた最適パラメータを用いて, Formal Run の 200 問について, QAC-1 のタスク 1 と 2 による評価実験を行った. この実験では, フィルタモデルの学習に用いる q-and-a セットとして, 5TAKU クイズデータの 268,531 個の実例に加え, Additional Run の 757 個の実例も利用した. 図 1 の点線は, この 200 問で評価した場合の文書数に対する Top1 accuracy の推移の様子を示している. パラメータ最適化に用いたセットで評価した場合と同様に, 文書数の増加に伴って性能が向上することが分かる.

表 2: mai (5k 文書) を用いたときの誤りの分析結果.

各モデルによるエラーの割り合い			年の欠落
R	F	R&F	
42.8%	21.7%	32.9%	2.6%

表 1 には, mai (5k 文書), www (5k 文書), mai+www (10k 文書) を利用したときの Top1 accuracy, Mean Reciprocal Rank (MRR), F-score を示す. NTCIR-3 QAC-1 の公式データである mai を用いた評価結果が, 公式の評価結果となるが, この結果 (MRR: 0.316, F-score: 0.150) は, NTCIR-3 QAC-1 の参加システムの中位の成績であった. また, www を用いた結果は, mai の結果に及ばないが, 両方のデータを組み合わせることで性能の改善が見られた.

なお, 本手法によって構築された英語 QA システムは, TREC2005 での評価において, Top1 accuracy が AQUAINT コーパスを利用した場合に 0.143, Web データを使った場合に 0.177 であった [1]. したがって, 今回構築した日本語 QA システムは, 英語 QA システムと同程度の性能を有していることが分かる.

3.4 誤りの分析

表 2 に, mai (5k 文書) を用いたときの解答の誤りがどの部分から引き起こされているかを分析した結果を示す. R は検索モデル, F はフィルタモデルが原因で引き起こされたものを示し, R&F は両方の影響によるものを示す. なお「年の欠落」は, 解答が日付である場合に, 末尾に「年」が欠けていたことから誤りと見なされたものである. この結果から, 誤りが主に検索モデルによって引き起こされていることが分かる.

4 おわりに

本稿では言語に非依存な統計的アプローチによって構築した日本語 QA システムについて, NTCIR-3 の QAC-1 タスクを用いて評価を行った. その結果, 現在の一般的な日本語 QA システムとして平均的な性能を有し, 同じ手法で構築された英語 QA システムと同程度の性能を有することが確認された. 今後は本アプローチを他の言語に適用し, 様々な言語の QA システムの構築を目指す. なお, 本システムのデモ (英語・日本語・ロシア語・スウェーデン語・中国語) は http://asked.jp にて公開されている.

謝辞

検索システムをご提供頂いた, 筑波大学 大学院図書館情報メディア研究科 藤井敦助教授に深謝致します. 本研究は, 文部科学省 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の支援を受けて行われた.

参考文献

[1] E. Whittaker, P. Chatain, S. Furui, and D. Klakow. TREC2005 Question Answering Experiments at Tokyo Institute of Technology. In *Proceedings of the 14th Text Retrieval Conference*, 2005.

[2] S.-H. Na, I.-S. Kang, and J.-H. Lee. POSTECH Question-Answering Experiments at NTCIR-4 QAC. In *Proc. of NTCIR-4 Workshop*, 2004.

[3] M. Fuchigami, H. Ohnuma, and A. Ikeno. Oki QA System for QAC-2. In *Proc. of NTCIR-4 Workshop*, 2004.

[4] Vector Software Library, http://www.vector.co.jp/

[5] A. Fujii and K. Itou. Evaluating Speech-Driven IR in the NTCIR-3 Web Retrieval Task. In *Proc. of NTCIR-3 Workshop*, 2002.