論文 / 著書情報
Article / Book Information

論題(和文)	統計的手法を用いた質問応答システムの改善
Title(English)	Improvement of a Statistical Question-Answering System
著者(和文)	今井 秀一郎, Edward Whittaker, 古井 貞熙
Authors(English)	Syuuichirou Imai, Edward Whittaker, Sadaoki Furui
出典(和文)	人口知能学会研究会資料, Vol. SIG-SLUD-A703-06, No. , pp. 27-32
Citation(English)	, Vol. SIG-SLUD-A703-06, No. , pp. 27-32
発行日 / Pub. date	2008, 3

統計的手法を用いた質問応答システムの改善

Improvement of a Statistical Question-Answering System

今井 秀一郎[1*]　エドワード ウィッタッカー[1]　古井 貞熙[1]
Shuichiro Imai [1], Edward Whitakker [1], and Sadaoki Furui [1]

[1] 東京工業大学 大学院情報理工学研究科 計算工学専攻
[1] Department of Computer Science, Tokyo Institute of Technology

Abstract: We have previously proposed a language-independent and data-driven approach for statistical question-answering (QA). This method uses word-classes for filtering answer candidates extracted from a large number of documents. In order to improve the QA performance, this paper proposes two new methods for measuring word distances and word clustering, respectively. The proposed methods are evaluated by conducting QA experiments on the NTCIR-3 QAC-1 additional run. Experimental results show that a system using distance calculation after dimension reduction based on singular value decomposition and overlapped word clustering achieves 0.275 MRR and 21.5% Top1 accuracy, which are significantly better than that obtained by our previous methods.

1 はじめに

我々はこれまでにデータ駆動式で，対象言語に依存した情報を用いない「統計的手法を用いた質問応答システム」を提案している [1]．このシステムでは，大量の文書から解答候補を抽出する「検索モデル」と，その中から入力の質問タイプに合致する候補を絞り込む「フィルタモデル」が利用されている．しかし，現状では双方のモデルとも精度が不十分であり [2]，それぞれの性能改善が望まれている．本研究では，このうち「フィルタモデル」の改善に焦点を当てシステムの性能向上を目指す．

従来までのフィルタモデルの精度が低い理由の１つとして，フィルタモデル中で回答候補を絞り込むために用いられる「単語クラス」のクラスタリング精度が低いことが挙げられる．そこで，単語間距離の定義とクラスタリングアルゴリズムを改善することで，単語クラスの精度改善を図る．

2 統計的手法を用いた質問応答システム

2.1 統計的アプローチ

本研究では，l_a 個の単語から成る回答 $A = a_1, \ldots, a_{l_a}$ が l_q 個の単語から成る質問 $Q = q_1, \ldots, q_{l_q}$ 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 $\mathcal{X}$ により取り出される２つの要素 $W = \mathcal{W}(Q), X = \mathcal{X}(Q)$ から構成されると考える．

$W = w_1, \ldots, w_{l_w}$ は質問文の「タイプ」を表す l_w 個の特徴量である．質問文 Q 中に含まれる「いつ」「どこ」「誰」といった質問のタイプを表す単語 (これを「質問タイプワード」と呼ぶ) 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．

一方，$X = x_1, \ldots, x_{l_x}$ 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 中の主題を表すキーワード (これを「質問キーワード」と呼ぶ) から生成される．W と同様に，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度の高い単語と，$\mathcal{W}$ で用いた単語は 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．この事後確率を最大化することで尤もらしい回答 $\hat{A}$ を求める．

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

この式は，複数の仮定の下で，次式のように書き換え

*連絡先：東京工業大学 大学院情報理工学研究科 計算工学専攻
東京都目黒区大岡山 2-12-1-W8-77
imai15@furui.cs.titech.ac.jp

ることができる [1].

$$\hat{A} = \arg\max_{A} \underbrace{P(A|X)^{\alpha}}_{\substack{retrieval\\ model}} \underbrace{P(W|A)}_{\substack{filter\\ model}} \quad (3)$$

$P(A|X)$ は X を与えた時の回答候補 A の生起確率であり，X を用いて回答候補となる 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．一方，$P(W|A)$ は回答候補 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回答候補に対し，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．

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．

2.2 フィルタモデル

$P(W|A)$ は $e_k = (Q^k, A^k) = (q_1^k, \ldots, q_{l_{Q^k}}^k, a_1^k, \ldots, a_{l_{A^k}}^k)$ と表せる K 個の質問-回答の実例 $E_k = (e_1, \ldots, e_K)$ を導入することにより以下のように定義する．

$$P(W|A) = \sum_{k=1}^{K} P(W|Q^k) P(A^k|A) \quad (4)$$

ここで，$P(A^k|A)$ にベイズの定理を適用し，また 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，以下のような式となる．

$$P(W|A) = \sum_{k=1}^{K} P(W|Q^k) \frac{P(A^k) \prod_{j=1}^{l_{A^k}} P(a_j^k|a_j)}{P(A)} \quad (5)$$

ここで，a_j と a_j^k の関係に，事前に作成した T 個の単語クラス $C_A = (c_1, \ldots, c_T)$ を導入することにより，次のように書き換える．

$$P(W|A) = \sum_{k=1}^{K} P(W|Q^k) \frac{P(A^k) \prod_{j=1}^{l_{A^k}} \sum_{t=1}^{T} P(a_j^k|c_t) P(c_t|a_j)}{P(A)} \quad (6)$$

3 従来システムにおける単語クラス生成

従来システム [2] では，フィルタモデルで用いる単語クラス C_A を作成する際に，バイグラム確率に基づく単語間距離と，ランダムに選択した単語に基づくクラスタリングを利用している．以下，それぞれについて説明する．

3.1 バイグラム確率に基づく単語間距離

単語の直前・直後 1 単語を単語の共起範囲とし，クラスタリングの対象単語に対する共起語のバイグラム確率によって対象単語の特徴を表現して，単語間距離を計算する．以下，この手法を「D_BI」と呼ぶ．

クラスタリングの対象となる M 個の単語を $W_c = (w_1, w_2, \ldots, w_M)$ とおくと，単語 w_a と w_b の距離 $D(w_a, w_b)$ は以下の式で表される．

$$D(w_a, w_b) = \sum_{i=1}^{M} |P(w_i|w_a) - P(w_i|w_b)| + \sum_{i=1}^{M} |P_{rev}(w_i|w_a) - P_{rev}(w_i|w_b)| \quad (7)$$

ここで，$P(w_i|w_a)$ は単語 w_a の直後に w_i が出現するバイグラム確率を表し，同様に $P_{rev}(w_i|w_a)$ は w_a の直前に w_i が出現する逆向きのバイグラム確率を表す．

3.2 ランダムに選択した単語に基づくクラスタリング

クラスタリングの対象とする単語 W_c からランダムに T 個の単語を取り出し S とする．S 中の単語それぞれをクラス C_A の初期クラスとおく．そして，以下の処理を繰り返すことで，残りの $W_c - S$ 中の単語を全て振り分け，単語クラスを作成する．以下，この手法を「RAND_INI」と呼ぶ．

1. $W_c - S$ から最も頻度が高い単語 w_h を取り出す．
2. w_h と C_A 中の各クラスの重心との距離を計算し，最も距離が近いクラスに単語 w_h を振り分ける．
3. w_h を割り振られたクラスの重心を再計算する．
4. $S \leftarrow S + w_h$ として，1 に戻る．

4 単語クラスの改良

従来の単語クラス作成手法には，単語間距離の定義とクラスタリングアルゴリズム双方に問題があると考えられている．本章では，従来手法の問題点とそれに対して提案した手法について述べる．

4.1 単語間距離の改良

従来の単語間距離の定義は，対象の単語と隣接して共起する全ての単語の情報を用いている．しかし，全単語を対象とすると低頻度語が含まれ，そのバイグラ

ムの推定精度が劣化してしまい，単語間距離を適切に計算することが難しくなる．そこで，このような低頻度語・低共起語の悪影響を防ぐため，(1) 各単語について，その隣接単語の統計的信頼度を相互情報量で評価し，その値が上位の単語のみを用いて単語間の類似度を計算する「相互情報量を用いた単語間距離」と，(2) 各単語を高次元の共起単語情報ベクトルとして表現した後，特異値分解によって次元を圧縮することで，重要性の低い共起情報を取り除き，そのベクトルによって単語間距離の定義を行う「次元圧縮した頻度情報を用いた単語間距離」の２種類を提案する．

4.1.1 相互情報量を用いた単語間距離

まず，クラスタリングの対象単語 W_c 中の単語 w と単語 c の間の相互情報量を学習用のテキストコーパス中の w と c の共起情報を用いて以下のように計算する．

$$MI(w,c) = \frac{C(w,c)}{N} \log \left(\frac{\frac{C(w,c)}{N}}{\frac{C(w)}{N} \times \frac{C(c)}{N}} \right) \quad (8)$$

ここで，N はコーパスの全単語数，$C(w)$ は w の出現頻度，$C(c)$ は c の出現頻度，$C(w,c)$ は w と c が隣接して共起する頻度を表す．

そして，W_c 中の単語 w_i に対し，w_i の直後に共起する単語の中から相互情報量が大きい上位 500 語を右特徴語群 R_i として，直前に共起する単語の中から相互情報量が大きい上位 500 語を左特徴語群 L_i として抽出する．これを用いて単語 w_i と w_j の単語間距離 $D(w_i,w_j)$ を以下の式により計算する．

$$D(w_i,w_j) = 1.0 - \frac{match(R_i,R_j) + match(L_i,L_j)}{1000} \quad (9)$$

ここで，$match(R_i,R_j)$，$match(L_i,L_j)$ はそれぞれ，w_i と w_j の両方が共通して持つ右特徴語の数と左特徴語の数を表している．以下，この手法を「D_MI」と呼ぶ．

4.1.2 次元圧縮した頻度情報を用いた単語間距離

まず，単語の共起情報を行列要素とする行列 $A = [A_1, A_2, \ldots, A_M]$ を作成する．行列の構成を図 1 に示す．行列の列成分は対象単語 $W_c = w_1, \ldots, w_M$ を表す．行列作成には対象単語の前後 D 単語分の共起情報を用いており，対象単語 w_i に対応する列ベクトル A_i は以下のように表される．

$$A_i = [a_{(w_1,1),i}, \cdots, a_{(w_M,1),i}, a_{(w_1,-1),i}, \cdots, a_{(w_M,-1),i}, \cdots, a_{(w_1,-D),i}, \cdots, a_{(w_M,-D),i}] \quad (10)$$

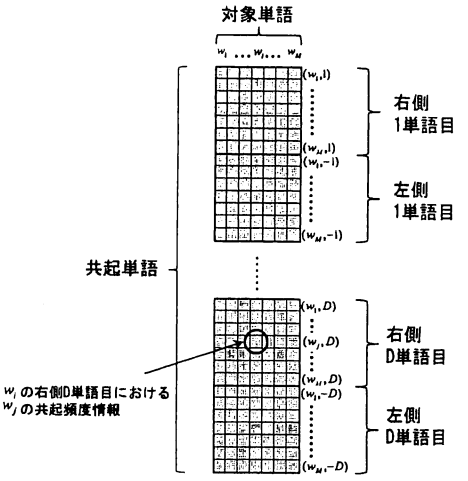

図 1: 共起行列の構成

ここで，要素 $a_{(w_k,d),i}$ は学習コーパスから抽出した単語 w_i に対し位置 d に共起する単語 w_k の共起情報を表す．d は単語 w_i に対する w_k の相対的な位置 $(-D\sim D)$ を表す．$a_{(w_k,d),i}$ の値は，共起語としての重要度を表す $G(w_k)$ と共起情報を表す $L(w_i,w_k,d)$ を用いて以下の式によって計算される [1]．

$$a_{(w_k,d),i} = G(w_k) \cdot L(w_i,w_k,d) \quad (11)$$

$$G(w_k) = 1.0 - \sum_{l=1}^{M} \frac{P(w_l|w_k) \cdot \log_2 P(w_l|w_k)}{-\log_2(M)} \quad (12)$$

$$L(w_i,w_k,d) = \log(1 + C(w_i,w_k,d)) \quad (13)$$

ここで，$P(w_l|w_k)$ は，w_k の直後に w_l が出現する確率を表し，$C(w_i,w_k,d)$ は w_i に対し位置 d に共起する w_k の頻度を表す．

以上のように作成した共起行列 A に特異値分解を適用し，重要度の高い情報のみを抽出して，対象単語 W_c の特徴を表す単語特徴ベクトルを作成する．具体的には，以下の手順に従う．

1. 対象単語 W_c から頻出単語上位 H 個を取り出して上位頻度語 W_H とし，残りの単語を W_L とする．行列 A から W_H 中の単語に対応する列成分を全て取り出し，行列 A_H を作成する．同様に，W_L に対して A_L を作成する．

2. 行列 A_H に特異値分解を適用して 3 つの行列に分解し，それぞれから特異値の上位 r 個に対応する成分を取り出したものを，$\hat{U}_H$，$\hat{S}_H$，$\hat{V}_H^T$ とする．

$$A_H \approx \hat{A}_H = \hat{U}_H \hat{S}_H \hat{V}_H^T \quad (14)$$

3. 共起行列 A_L の左から $\hat{U}_H\hat{S}_H$ の逆行列をかけることで行列 $\hat{V}_L$ を求める．

このように作成した $\hat{V}_H$，$\hat{V}_L$ の列成分は，それぞれ A_H，A_L の列成分の特徴を表す行列となっており，対応する単語の共起情報の特徴を表している．よって，$\hat{V}_H$，$\hat{V}_L$ の列成分を，対応する単語の単語特徴ベクトル群 $F = f_1, \ldots, f_M$ として利用し，単語間距離を計算する．

単語 w_i と w_j の距離 $D(w_i, w_j)$ はそれぞれの単語特徴ベクトル f_i，f_j を用いて以下のように計算される．以下，この手法を「D_RDF」と呼ぶ．

$$D(w_i, w_j) = 1.0 - \cos(f_i, f_j) \qquad (15)$$

なお，上記のように共起行列 A を W_H と W_L に分別しなくとも A 全体に特異値分解を適用することで特徴ベクトルは作成できる．しかしこの場合，高頻度語と低頻度語の情報を同時に次元圧縮することになり，特異値分解の時点で信頼性の低い共起情報を除去できない可能性がある．そこで，上記のような方式を採用した．

4.2 クラスタリングアルゴリズムの改良

従来のクラスタリングアルゴリズムは，初期クラスに強く依存する手法であり，初期クラスとして適当でない単語が選択された場合，クラスの精度を低下させてしまう．そこで，第 1 段階において高頻出語についてボトムアップクラスタリングを行い，精度の高い単語クラスを生成した後，第 2 段階でその単語クラスに対して全対象単語の割り振りを行う「2 段階方式に基づくクラスタリング」を提案する．さらに，単語が複数のクラスに属することを考慮した，「オーバーラップを許したクラスタリング」による性能改善について検討を行う．

4.2.1 2 段階方式に基づくクラスタリング

本手法では，まず第 1 段階で，対象単語 W_c から頻出単語上位 H 個を取り出してこれを W_H とし，残りの単語を W_L とする．この W_H に対し，最長距離法を用いたボトムアップクラスタリングを適用して T 個のクラスを作り，これを単語クラス C_A の初期クラスとする．そして，第 2 段階で W_L 中の単語 w_l それぞれについて，W_H 中の全ての単語との類似度を計算し，最も類似度が高い単語が属するクラスに w_l を振り分ける．以下，この手法を「C_TWOPS」とおく．

4.2.2 オーバーラップを許したクラスタリング

節 3.2，節 4.2.1 で述べた方法で単語クラスを作成する場合，各単語は常に 1 つのクラスにのみ属するため，複数の概念を持つ単語を適切にクラスタリングできない可能性がある．そこで，本手法では今までに提案されている Clustering by Committee [3] に基づき，単語が複数のクラスに属するようにクラスタリングを行い，単語が持つ複数の概念を単語クラスで適切に表現する．

アルゴリズムは大きく 3 つの段階に分かれている．まず第 1 段階では，クラスタリングの対象単語 W_c から上位頻度語 H 語を取り出しこれを W_H とする．そして，W_H の各単語について類似度の高い単語群を W_H から抽出する．第 2 段階ではそれらの単語群から「committee」と呼ばれるクラスの核となる単語群を複数個作成する．最後に，第 3 段階において，W_c 中の全ての単語を類似度の高い複数の committee に振り分け，単語クラスを作成する．なお，第 2 段階においては committee の数を調節するためのパラメータを，第 3 段階においては単語のオーバーラップの度合を調節するためのパラメータを設定する．以下，この手法を「C_OVER」と呼ぶ．

5 評価実験

5.1 実験条件

各手法により作成した単語クラスを，統計的手法を用いた質問応答システムに適用し，NTCIR-3 QAC-1 Additional Run の 757 問の質問文で評価を行う．評価基準として，「Top1 Accuracy」と「MRR(Mean Reciprocal Rank)」の 2 つを用いた．

利用する文書データはあらかじめ Chasen 2.3.3 [4] と IPADIC 2.7.0 [4] を用いて単語単位に分割しておく．

クラスタリングの対象単語 W_c には，NTCIR-3 WEB タスクの 10GB の文書データ「NW10G-01」から出現頻度上位 M =300k 語を使用した．また，単語の共起情報の作成にも NW10G-01 を用いた．

検索対象とする文書は，毎日新聞データ 2 年分 (1998-99) から Additional Run の質問毎に検索エンジン akechi-2.0.1b [5] を用いて，$|U|$=5k 文書ずつ抽出した．

フィルタモデルで用いる質問-回答実例データには 5TAKU クイズデータ [6] から抽出された $K = 257,019$ 個のデータを用いた．また，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上位 200 単語をそれに用いた．質問タイプワードのリストには，5TAKU クイズデータの質問文から得られる出現頻度上位 125 単語を用いた．

なお，関数 $\mathcal{W}$, $\mathcal{X}$ 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調整するパラメータ α はそれぞれ，$m = 4$, $n = 3$, $\alpha = 2.0$ とした．

5.2 従来手法と提案手法の比較

単語間距離として，従来手法である D_BI と提案する D_MI，D_RDF の 3 種類を，クラスタリングアルゴリズムとして，従来手法である C_RAND と提案する C_TWOPS の 2 種類を適用して作成した，計 6 種類の単語クラスを用いて，システムの性能比較を行った．

D_RDF では，共起行列を作成する際の，共起範囲を対象単語の前後 3 単語 ($D = 3$) とし，単語特徴ベクトルを作成する際の上位頻度語 W_H の個数 H は 25k とした．また，特異値分解の適用により作成する単語特徴ベクトルの次元数 r は 200 とした．

C_RAND では，クラス数 T を文献 [2] において最適とされている 500 とした．また，C_TWOPS では，上位頻度語 W_H の個数 H を 25k とした．また，C_TWOPS でのクラス数 T も 500 とした．

以上の条件で作成した各単語クラスを用いた場合の Top1 Accuracy を表 1 に，MRR を表 2 に示す．

単語間距離の定義について見ると，従来手法である D_BI よりも，D_MI や D_RDF を利用したクラスの方が Top1 Accuracy, MRR がともに高い．これは，D_MI や D_RDF は，重要度の高い共起情報のみを抽出し計算する手法を取ることにより，信頼性の低い低頻度語・低共起語の悪影響を抑制できたためと考えられる．

一方，クラスタリングアルゴリズムについて見ると，全ての単語間距離において，従来手法である C_RAND より，C_TWOPS の方が Top1 Accuracy, MRR がともに高い．C_TWOPS では，第 1 段階で上位頻出語によるボトムアップクラスタリングで精度の高い初期クラスを作成した上で，第 2 段階で全単語の振り分けを行っていることから，C_RAND よりも高い精度の単語クラスが得られたものと考えられる．

表 1: 各単語クラスを用いた時のシステム性能：Top1 Accuracy(%)

		単語間距離の定義		
		D_BI	D_MI	D_RDF
クラスタリング	C_RAND	16.7	17.4	17.7
アルゴリズム	C_TWOPS	16.8	18.6	19.5

表 2: 各単語クラスを用いた時のシステム性能：MRR

		単語間距離の定義		
		D_BI	D_MI	D_RDF
クラスタリング	C_RAND	0.233	0.237	0.242
アルゴリズム	C_TWOPS	0.234	0.244	0.254

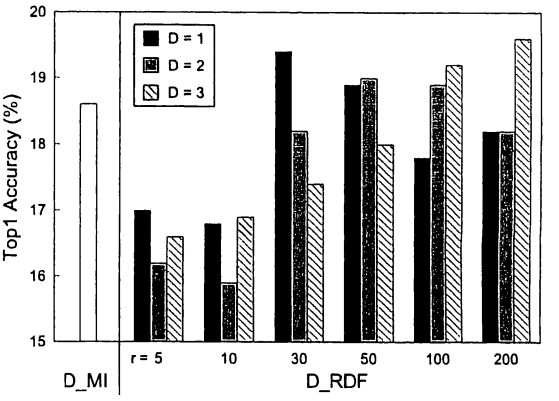

図 2: D_MI と圧縮次元数 r と共起範囲 D を変化させた D_RDF の Top1 Accuracy(%)

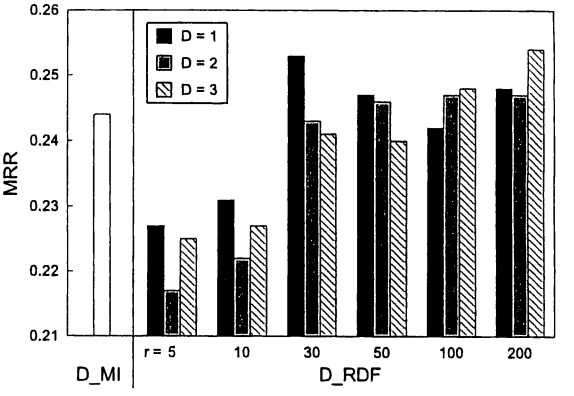

図 3: D_MI と圧縮次元数 r と共起範囲 D を変化させた D_RDF の MRR

5.3 相互情報量を用いた距離と次元圧縮をした頻度情報を用いた距離の比較

クラスタリングアルゴリズム C_TWOPS において D_MI と D_RDF それぞれを用いて単語クラスを作成し，比較実験を行った．D_RDF では共起行列の圧縮の際の次元数 r と共起範囲とする対象単語の前後の単語数 D をそれぞれ $r = 5, 10, 30, 100, 200$，$D = 1, 2, 3$ と変化させ，それぞれの組みで単語クラスを作成した．クラス数 T は 500 とした．

評価結果を図 2，3 に示す．図 2 は各クラスにおける Top1 Accuracy を示し，図 3 は MRR を示している．

まず，D_RDF のパラメータの変化による性能への影響を見ると，圧縮次元数が $r = 5, 10$ の場合，Top1 Accuracy, MRR が共に非常に低いことが分かる．これ

表 3: 各クラスタリング手法の比較 (単語間距離は D_RDF を使用)

	単語間距離の定義		
	C_RAND	C_TWOPS	C_OVER
Top1 (%)	17.7	19.5	21.5
MRR	0.242	0.254	0.275

は，情報を過度に圧縮したため単語の特徴を表現できず，単語クラス精度が劣化したためと考えられる．また，$r = 30$〜200 では，次元数によるスコアの変化には，はっきりとした傾向が見られない．共起範囲 D について見てみると，D の違いによるスコアの変化にも，はっきりとした傾向が見られない．共起情報が多くなれば性能が改善されると予想していたが，この実験では，共起範囲 D はシステムの性能に大きな影響を与えないということが分かった．

D_RDF と D_MI を比較すると，$r = 30$〜200 においては，両者がほぼ同等の性能となることが分かる．

5.4 オーバーラップを許したクラスタリング手法の検証

C_OVER を適用して作成した単語クラスを評価し，有効性について検証した．

W_c から取り出す上位頻出語 W_H の単語数 H は 25k とした．パラメータは，予備実験においてシステム性能が最大となった時の値を設定した．作成した単語クラスは，クラス数が 1,043 で，のべ単語数が約 440 万となった．よって，$W_c = 300$k であることから，1 単語が平均 14.8 クラスに振り分けられていたことになる．

評価結果を表 3 に示す．比較のため，C_RAND, C_TWOPS で作成した単語クラスによる結果も示す．なお，単語間距離の定義には D_RDF($D = 3$, $r = 200$) を選択した．C_RAND, C_TWOPS で使用するパラメータは節 5.2 で設定したものと同様である．

表より，C_OVER の方が，C_RAND や C_TWOPS よりも，Top1 Accuracy, MRR がともに高いことが分かる．これは，C_OVER により，単語が持つ複数の概念をクラスで適切に表現することができ，回答候補のフィルタとしての効果が向上したためと考えられる．

最終的に，次元圧縮した頻度情報を用いた距離とオーバーラップを許したクラスタリングによって作成した単語クラスにより，Top1 Accuracy が 21.5%，MRR が 0.275 となった．

6 おわりに

本論文では，統計的手法を用いた質問応答システムにおいて，回答候補の絞り込みに用いられる単語クラスの作成手法の改良を検討した．単語間距離の定義として相互情報量を用いた距離と次元圧縮した頻度情報を用いた距離の 2 つを，クラスタリングアルゴリズムとして 2 段階方式に基づくクラスタリングとオーバーラップを許したクラスタリングの 2 つを提案した．

評価実験により，提案手法が従来手法に比べ有効であることが確認され，最終的に，次元圧縮した頻度情報を用いた距離とオーバーラップを許したクラスタリングによって作成した単語クラスにより，Top1 Accuracy が 21.5%，MRR が 0.275 となった．また，相互情報量を用いた距離と次元圧縮した頻度情報を用いた距離の比較実験では，両手法による性能がほぼ同等となることが分かった．

今後の課題としては，オーバラップを許したクラスタリングにおいて，パラメータの最適化によりクラス精度をより向上させることや，コーパスデータの拡充により単語クラス作成に用いる共起情報の信頼性を向上させること等が挙げられる．

謝辞

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 and S. Furui : TREC2005 Question Answering Experiments at Tokyo Institute of Technology, Proc. of the TREC 2005 Conference, 2005.

[2] J. Hamonic, E. Whittaker, 古井貞熙 : 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, 情報処理学会第 68 回全国大会, 2006.

[3] P. Patrick and L. Dekang : Document *Clustering* with Committees, Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 199-206, 2002.

[4] http://chasen.naist.jp/hiki/ChaSen/.

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

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