Vol.2011-NL-201 No.8
Vol.2011-SLP-86 No.8
2011/5/17

質問・回答事例を利用した non-factoid 型質問応答に対する確率的言語モデルの導入

吉田恭輔[†1]　上田太郎[†1,*1]
石下円香[†2]　森辰則[†2]

本研究では，non-factoid 型質問に対する質問応答システムの精度向上のために，入力された質問と記述スタイルの類似する質問事例の取得と，取得した質問事例と対応する回答事例と解候補の記述スタイルが類似しているかどうかの判定に言語モデルを用いる手法を検討した．言語モデルとしては，品詞と表層の 2-gram モデルを用いた．評価型ワークショップ NTCIR-6 の QAC-4 タスクにおけるテストセット用いた評価実験により，提案手法を用いたシステムの精度向上を確認した．

Introduction of a Probabilistic Language Model to Non-Factoid Question-Answering Using Examples of Q&A Pairs

Kyosuke Yoshida,[†1] Taro Ueda,[†1,*1]
Madoka Ishioroshi[†2] and Tatsunori Mori[†2]

In this paper, we propose a method which utilizes a probabilistic language model in non-factoid type question-answering system in order to improve its accuracy. The model is a mixture probabilistic language model of part-of-speech and serface expressions. We introduced the model into two sub-processes which calculate similarity in terms of description style. One is for collecting examples of questions similar to an input question. The other one is for measuring similarity of an answer candidate to the answer examples paired with the collected question examples. Experimental results showed that the accuracy of the system was improved by introducing the proposed method.

1. 序　　論

近年，計算機の高性能化やネットワークの発達に伴い電子化された文書が増大しており，大量の文書群から利用者が必要な情報を効率良く取得する為の情報アクセス技術が必須となっている．情報アクセス技術には情報検索をはじめとして，情報抽出，自動要約，質問応答などが存在する．その中の一つである質問応答は，利用者の自然言語による質問に対して情報源となる文書集合から回答そのものを抽出する技術である．

従来の質問応答システムは，人名や地名，数量等を問う事実型 (factoid 型) の質問を対象としたものが一般的であった．factoid 型質問に対する回答としては名詞や名詞句で表されるような短い表現が一般的であり，解抽出には固有表現抽出や数量表現抽出が用いられる．

しかし，実際に利用者が尋ねることが想定される質問の種類は多岐に渡り，定義・理由・方法などを問うような質問においては比較的長い文章表現による記述的な回答が想定される．このような質問は non-factoid 型質問と呼ばれる．non-factoid 型質問を対象とした質問応答に関する研究は近年広まりつつあるが，依然難易度の高いタスクである．本研究では non-factoid 型の質問応答を扱う．

non-factoid 型の質問を処理する際，入力された質問を「定義型」「理由型」「方法型」といった型に分類して個別に処理を行うのか，それとも質問の型を分類せずに統一的な処理を行うのかという問題がある．前者は質問の型ごとに個別の処理を用意するのが非効率的であるし，質問の型分類の精度が解抽出の精度に大きく影響してしまう．そこで, 石下ら[1)] は Q&A コミュニティサービスの質問・回答事例集を Q&A コーパスとして用い，質問の型分類を行わないアプローチを取っている．本研究では，石下ら[1)] のシステムをもとにして，入力された質問と記述スタイルの類似する質問事例の取得を質問事例の言語モデルによって行い，これによって取得した質問事例に対応する回答事例を用いて，回答の言語モデルを作成し，解候補文の生成確率を求めることによって, 解抽出を行う．このような手法を用いることで，解候補から日本語の文法上正しくない文章や無意味な記号などが含まれている文章を取り除くこと，質問事例の文書全体の記述スタイルを考慮することができるのではないかと考えられる．

†1 横浜国立大学大学院環境情報学府
Graduate School of Environment and Information Sciences,Yokohama National University
*1 現在，ヤフー株式会社
Presently with Yahoo Japan Corporation
†2 横浜国立大学大学院環境情報研究院
Graduate School of Environment and Information Sciences,Yokohama National University

Vol.2011-NL-201 No.8
Vol.2011-SLP-86 No.8
2011/5/17

表 1 non-factoid 型質問の分類
Table 1 Types of non-factoid type questions

質問の型	質問の記述スタイル例	回答の記述スタイル例
定義型 (definition)	～とは何 ～って何	～とは…である ～は…のこと
理由型 (why)	なぜ～ ～の理由は何	～ため ～から
方法型 (how)	～にはどうしたらいい ～の方法は何	～するにはまず… ～のやり方は…
その他 (other)	X と Y の違いは何	X は～だが，Y は…
	～したらどうなる	～した場合，…

2. 研究背景

本節では，non-factoid 型質問応答の種類と処理方法，関連研究と提案手法の位置付けについて述べる．

2.1 質問の種類と処理方法

non-factoid 型の質問は定義や理由，方法等を問う質問であり，数文にまたがる比較的長い文章表現による記述的な回答が想定される． non-factoid 型質問には様々な種類があるが，大まかに分類すると表 1に示す通りとなる．

non-factoid 型質問に対しては，キーワード等による検索によって得られた文書中から抽出してきた数文が解候補となるが，このような解候補の適切性は，

【尺度 1】 質問の内容との関連性
【尺度 2】 質問の型に応じた記述スタイルを満たす度合

の組合せで見積もることが多い[2]．質問の「内容」とは，質問の話題 (トピック) を指す．「記述スタイル」とは，表 1に示したような質問や回答の特徴表現を指す．この両者は厳密には独立ではないと考えられるが，本研究においては両者を分離できるものとして話を進める．

上記の二つの尺度で解候補のスコア付けを行なうことを考えた時，【尺度 1】は簡単には質問文と解候補文の類似度で計算できる．【尺度 2】については 2.2節で紹介するように，人手で作成した語彙統語パタンや機械学習により判定する手法が提案されている．

2.2 関連研究

2.2.1 質問の型ごとに処理を分ける手法

本節では，質問の型ごとに個別の処理を行なうアプローチについて関連研究を紹介する．

Han ら[2] は英語の定義型質問応答において，前述の二つの尺度をコーパスから推定した確率モデルに基づいて計算している．【尺度 1】に関する確率は検索文書から計算し，【尺度 2】の計算には定義文コーパスを用いている．

森本ら[3] は， RST(意味的なまとまりをもつスパン間に成り立つ関係を記述したもの) の例文から理由型質問応答に利用できる関係を探し出し，質問文とそれらの関係が成り立つ部分を回答とする手法を提案している．

三原ら[4] のシステムでは，方法型質問応答において，検索された文書内から，質問者がすべき行動であると思われる，名詞句と動詞からなる「行動表現」を抽出して回答とする．

以上のように質問の型が限定された場合には，型に応じた回答表現のパタンや個別のルールを作成したり，専用のコーパスを学習データとして用いたりすることが有効である．質問の型を限定しない場合でも，入力された質問を予め用意した型に分類し，型ごとに処理を分けることで同様のアプローチを実現できる．

諸岡ら[5] は，質問を「定義型」，「理由型」，「方法型」のどれかに分け，個別に処理を行う手法を提案している．解抽出には，それぞれの型用に用意した人手の回答表現パタンを用いている．型の判定は人手による表層表現のパタンを用いており，想定したどの型にも当てはまらない質問は，「定義型」として処理される．

2.2.2 質問の型分類を行なわない手法

2.2.1節で質問の型ごとに処理を分けるアプローチについて述べたが，実際には質問の型が何種類あるかは不明であるし，型ごとに個別の処理方法を用意するのは非効率である．また，質問の型分類の精度が回答精度に大きく影響してしまう．可能ならば質問の型に依らない統一的な手法が望ましい．

水野ら[6] は日本語 non-factoid 型質問応答において， Q&A コミュニティサービスの質問・回答事例集合を学習データとし，質問と回答の型の一致を判断する分類器を作成することにより，質問の型分類を行なわずに【尺度 2】を判定する手法を提案している．この手法では回答の範囲を一段落と先に決め，得られた解候補を質問と型が一致するかどうかで分類するため，解候補の範囲を質問に応じて動的に変更できない．

Soricut ら[7] は英語 non-factoid 型質問応答において， FAQ サイトの質問・回答事例集合をパラレルコーパスとみなし，回答が質問に「書き換え」られる (*translation*) 確率を計算するという，質問の型に依らない手法を提案している．この手法でも回答の範囲を 3 文に固定しているため，水野ら[6] の手法と同様の問題がある．また，質問の長さから回答の長さ (語数) を推定する必要があるのが難点である．

Vol.2011-NL-201 No.8
Vol.2011-SLP-86 No.8
2011/5/17

石下ら[1] も日本語 non-factoid 型質問応答において，【尺度 2】の見積もりに Q&A コミュニティサービスの質問・回答事例集合をコーパスとして用いる手法を提案している．特に，入力された質問に適合する回答の特徴表現をコーパスから動的に取得するという方法を用いており，この点において水野ら[6] や Soricut ら[7] の手法とは異なる．石下ら[1] の手法は機械学習を基にしておらず，Q&A コーパスの追加が容易である．また，Soricut ら[7] の手法では【尺度 1】と【尺度 2】の両方に関係することをコーパスから学習しているのに対し，この手法では【尺度 2】に関係する特徴表現のみをコーパスから取得するため，【尺度 1】に相当する，コーパス中での質問事例や回答事例の内容の網羅性を考慮する必要がなく，回答の範囲を質問に応じて動的に決めることができる．石下ら[1] のシステムの処理の流れを以下に示す．

(1) 入力された質問と記述スタイルの類似する質問事例を Q&A コーパスから取得する．
(2) 取得した質問事例に対応する回答事例からその特徴表現を取得する．その際，言語表現と回答事例集合との間の相関度を測る尺度として χ^2 値を用いる．
(3) 入力された質問の内容語を用いて，Web 上から情報源となる文書を検索する．
(4) (3) で取得した文書に対して，【尺度 1】と【尺度 2】をどれだけ満たしているかスコアリングを行う．【尺度 1】には，トピック語の網羅性を用いる．【尺度 2】は (2) で求めた χ^2 値に基づき計算する．
(5) スコアリングの結果，上位のものを回答として出力する．

石下ら[1] のシステムの概要を図 1に示す．このシステムでは，Q&A コーパスとして，「Yahoo! 知恵袋」*1 を利用している．このコーパスは，2004 年 4 月から 2005 年 10 月までの間に蓄積された約 311 万件の質問と約 1347 万件の回答が研究利用のために国立情報学研究所より提供されているものである．1 つの質問には複数の回答が存在し，このうち「ベストアンサー」のみを使用している．質問事例の取得の際に，疑問詞を中心とした語の 7-gram を利用するため，コーパスとして用いた Q&A ペアは，質問文に疑問詞を含むもののみに限定し，かつ質問文中の 7-gram の出現頻度が極端に少ないものを除いたものとした．その結果，約 90 万件の Q&A ペアを用いている．これと同じものを本研究においても Q&A コーパスとして利用している．

2.3 提案手法の位置付け

本研究の目的は石下ら[1] のシステムに言語モデルを導入することによって，質問応答の精度を向上させることである．石下ら[1] のシステムでは，質問事例の取得には，疑問詞を中心とする

*1 http://chiebukuro.yahoo.co.jp/

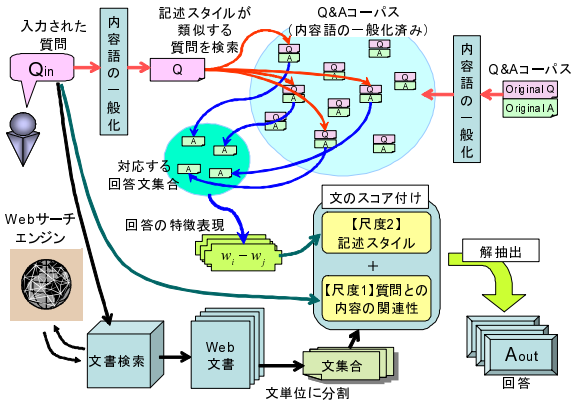

図 1 石下ら[1] のシステムの概要
Fig. 1 Outline of the system proposed by Ishioroshi[1]

語の 7-gram の一致の度合いを類似度としている．この手法では，7-gram の類似度は高いが，質問事例の文全体を見ると，記述スタイルが必ずしも類似しているとは言えないものや，対応する回答事例の内容が回答としては不適切な表現となっているものが含まれていることがあり，質問事例の選別がうまくできていないということが考えられる．また，回答の特徴表現の取得の際には，単語 2-gram による χ^2 値を求めてそれをスコアリングに用いている．この手法では，χ^2 値の計算上，単に 2-gram の頻度だけを見る形になってしまい，特徴表現が上手く取れていない可能性がある．そこで，本研究では，入力された質問と記述スタイルの類似する質問事例の取得，取得した質問事例と対応する回答事例と解候補の記述スタイルが類似しているかどうかの判定を言語モデルを用いて行った．これにより，より記述スタイルの類似した質問事例を取得することができ，また，解候補から日本語の文法上正しくない文章や無意味な記号などが含まれている文章を取り除くことができるのではないかと考えられる．

3. 言語モデルを利用した non-factoid 型質問応答

本節では，まず石下ら[1] の手法 (以下，従来手法) における上記問題点をさらに詳しく説明し，その後に，その問題を解決するために本研究において提案した言語モデル，それを導入した質問応答システムの処理の流れについて説明する．

Vol.2011-NL-201 No.8
Vol.2011-SLP-86 No.8
2011/5/17

3.1 従来手法の問題点

従来手法では，2.1節で説明した【尺度 2】の「質問の型に応じた記述スタイルを満たすか」を見積もる際に収集した質問事例に対応する回答事例から χ^2 値の高い単語 2-gram を特徴表現として取得している．しかし，χ^2 値の計算上，特徴表現として取得している単語 2-gram については単純にその頻度を見ているだけであり，語順やその表現の前後関係などを全く考慮していない．

また，記述スタイルの類似する質問事例集合をコーパスから取得する際に，入力された質問と質問事例の「疑問詞を中心とする語の 7-gram」の一致の度合を類似度としている．この手法の問題点として，同 7-gram 以外の要素を考慮していないということ，同 7-gram を用いる際に一般化する内容語をあらかじめ決定する必要があるということがあげられる．そのため，同 7-gram の類似度は高いが，文全体を見ると必ずしも記述スタイルの類似していないものや，対応する回答事例が入力された質問に対する回答を抽出するための記述スタイルを取得するもととして不適切なものが，質問事例として取得される場合がある．

記述スタイルの類似していないものの例として次のようなものがある．
入力された質問が

【質問 (入力)】ＢＳＥ（狂牛病）が人に感染するとどうなりますか。

である場合に，次のような質問事例を取得した場合である．

【質問 (事例)】「百合の花咲く場所で」を英語にするとどうなりますか。
【回答 (事例)】 At the place where lilies bloom です。

この例では，疑問詞を中心とする語の 7-gram は「感染するとどうなりますか」と「英語にするとどうなりますか」で一致の度合いは高いが，前者は「名詞 (感染) __動詞 (する) __助詞 (と)」であるが，後者は「名詞 (英語) __助詞 (に) __動詞 (する) __助詞 (と)」となっており，記述スタイルの点では異なっている．質問の対象についても，前者は「どのような症状なのか」，後者は「英訳の内容」となっており，異なっている．回答事例の内容についても，入力された質問に対する回答として「○○が△△する」といった記述スタイルのものが想定されるが，回答事例はそのような記述スタイルを満たしていない．これらの理由からこの質問・回答事例は入力された質問に対して不適切であると言える．

また，取得した質問事例に対応する回答事例が，入力された質問に対する回答としては記述スタイルが不適切なものとなっているものの例としては次のようなものがある．入力された質問が

【質問 (入力)】米国が京都議定書を批准しない理由は何ですか。

である場合に，次のような質問事例を取得したとする．

【質問 (事例)】キャンピングカーを買った理由は何ですか。
【回答 (事例)】トレーラーを買って７年になります。買って良かったです。

この例では，両者ともある行動に対する理由についての質問となっており，質問事例の内容としては問題ない．しかし，回答事例の内容は，理由を語っていないため不適切なものとなっている．

本研究では，上記の問題点を解消した上で，次のような質問事例を取得することを目指す．
入力された質問が

【質問 (入力)】米国がコソボの独立を認めない理由は何ですか。

であるときに，次のような，質問事例の内容として適切である上に，入力した質問に対する回答を抽出するための記述スタイルを取得する元として適切なものを取得することを目指す．

【質問 (事例)】中国が台湾の独立を認めない理由は何ですか。
【回答 (事例)】台湾の独立を認めない理由のひとつとして、大陸中国のチベット自治区、新疆ウイグル自治区、内モンゴル自治区などにも独立機運が波及し、中央政府のコントロールが及ばない恐れがあるからで、国家の分裂の誘因になりかねないようなことを見過ごすことができないからです。

3.2 言語モデルを用いた質問応答システムの概要

言語モデルを用いた質問応答システムの概要を図 2に示す．本システムは，石下ら[1)] のシステムに 3.3節で説明した品詞と表層を混合させた言語モデルを導入したものである． 2.3節で述べたように，本システムでは，入力された質問と記述スタイルの類似する質問事例の取得，取得した質問事例と対応する回答事例と解候補の記述スタイルが類似しているかどうかの判定を言語モデルを用いて行った．つまり， 2.1節における【尺度 2】の見積もりに言語モデルを導入したことになる．この手法の利点として，以下のものが挙げられる．

Vol.2011-NL-201 No.8
Vol.2011-SLP-86 No.8
2011/5/17

- 質問事例の取得の際に，質問事例の文書全体の記述スタイルを考慮することができ，疑問詞を中心とした 7-gram の一致の度合いを類似度とした場合に比べ，入力された質問と記述スタイルの類似した質問事例を多く取得できると期待できる．
- 解候補文章の文単位の生成確率を求め，それを用いて解抽出を行うことで，解候補から日本語の文法上正しくない文や無意味な記号などが含まれている文を取り除く事ができると期待できる．

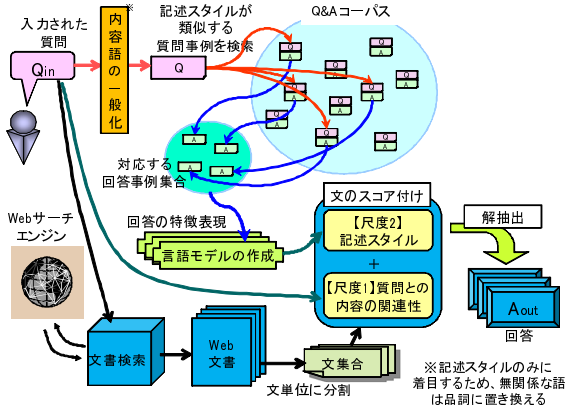

図 2 提案手法の概要
Fig. 2 Outline of the proposed method

3.3 品詞と表層を混合させた 2-gram モデル

3.3.1 概　　要

石下ら[1] のシステムでは，回答事例の特徴表現の際に用いた 2-gram は，疑問詞や助詞・助動詞等の機能語と，予め焦点となりやすい単語としてリストに登録されている一部の内容語を表層表現として用い，その他の語は品詞に置き換える一般化を行っている．しかし，実際は特徴表現取得の際にどの単語が焦点となるかは不明であり，あらゆる質問に対して一定に決まっているわけではない．どの場合に表層表現を用い、またどの場合に品詞を用いるのかという傾向を言語モデルに反映させるために，品詞と表層 (単語) を混合させた 2-gram モデルを用いた．図 3にそのモデルを示す．

モデルの確率推定式を式 (1) に示す．但し，C は品詞を表し，E は表層を表す．

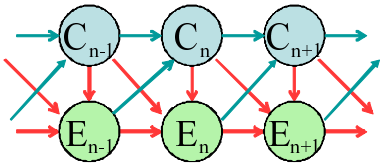

図 3 品詞と表層の混合モデル
Fig. 3 A mixture probabilistic language model of part-of-speech and surface expressions

$$P(E_1E_2...E_n) \approx P(E_n|C_nE_{n-1}C_{n-1})P(C_n|E_{n-1}C_{n-1})P(E_1E_2...E_{n-1})$$
$$= \prod_{i=1}^{n}\{P(E_i|C_iE_{i-1}C_{i-1})P(C_i|E_{i-1}C_{i-1})\} \quad (1)$$

可能性のあるすべての状況において品詞と表層の重要視する割合を柔軟的に決定するために，線形補間法[8] に基づいたスムージングにより 2-gram の近似を行う．近似式，及びパラメータを以下に示す．パラメータは削除補間法[8] を用いて求める．

$$\begin{aligned}P(E_i|C_iE_{i-1}C_{i-1}) &= \alpha_1 P_{ML}(E_i|C_{i-1}E_{i-1}C_{i-1})\\ &+ \alpha_2 P_{ML}(E_i|C_iE_{i-1})\\ &+ \alpha_3 P_{ML}(E_i|E_{i-1}C_{i-1})\\ &+ \alpha_4 P_{ML}(E_i|C_iC_{i-1})\\ &+ \alpha_5 P_{ML}(E_i|E_{i-1})\\ &+ \alpha_6 P_{ML}(E_i|C_i)\\ &+ \alpha_7 P_{ML}(E_i|C_{i-1})\\ &+ \alpha_8 P_{ML}(E_i) \quad (2)\end{aligned}$$

$$\begin{aligned}P(C_i|E_{i-1C_{i-1}}) &= \beta_1 P_{ML}(C_i|E_{i-1}C_{i-1})\\ &+ \beta_2 P_{ML}(C_i|C_{i-1})\\ &+ \beta_3 P_{ML}(C_i|E_{i-1})\\ &+ \beta_4 P_{ML}(C_i) \quad (3)\end{aligned}$$

P_{ML} は最尤推定で求められた確率値であることを意味する． α_i, β_i は補間係数であり，それぞれの合計は 1 になるように設定される．

3.3.2 パラメータの決定

パラメータの決定は削除補間法を用いて行う．ここでは再推定式のみを示す．また，分割数

Vol.2011-NL-201 No.8
Vol.2011-SLP-86 No.8
2011/5/17

m は 20 とした．

$$\bar{\alpha}_1 = \frac{1}{N}\sum_{n=1}^{N}\frac{\alpha_1 P_{ML}(E_n|C_nE_{n-1}C_{n-1})}{P(E_n|C_nE_{n-1}C_{n-1})}$$
$$= \frac{1}{N}\sum_{n=1}^{N}\frac{\alpha_1 P_{ML}(E_n|C_nE_{n-1}C_{n-1})}{\alpha_1 P_{ML}(E_n|C_nE_{n-1}C_{n-1}) + ... + \alpha_8 P_{ML}(E_n)} \quad (4)$$

これを同様に $\alpha_2 ... \alpha_8, \beta_1 ... \beta_4$ についても行う．

3.3.3 文生成確率の導出

決定されたパラメータを用いて文生成確率を求める．対象となる文に形態素解析を適用し 2-gram に分割し，式 (1) を用いて確率推定を行う．式 (1) によって求められた値 $P(E_1E_2...E_n)$ が，言語モデルにおけるその文の生成確率である．

3.3.4 本研究における言語モデルの適用

本研究では，記述スタイルが類似した質問事例の取得と，解候補の【尺度 2】の見積もりに言語モデルを導入する．記述スタイルの類似した質問事例の取得では，まず式 (4) より取得の対象となる質問事例集合のモデルのパラメータを推定し，式 (1) よりそのモデルが入力された質問を生成する確率を求めることで，入力された質問と質問事例集合の記述スタイルが類似しているかを判定する．取得した質問事例集合と対をなす回答事例集合と解候補の記述スタイルが類似しているかの判定も同様に，式 (4) より取得した回答事例集合のモデルのパラメータを推定し，式 (1) よりそのモデルが解候補の文を生成する確率を求めることで，解候補と回答事例集合の記述スタイルが類似しているかを判定する．

3.4 言語モデルを用いた質問応答システムの処理の流れ

3.4.1 キーワード抽出と関連語の取得

入力された質問文から内容語を取得し，キーワード集合 K とする． K を名詞キーワード集合 K_n と 動詞・形容詞キーワード集合 K_p に分ける．また， K_n 中で複合語を作れる場合は複合語にした場合の集合を K_c とする．

【尺度 1】の見積りに使用する質問内容の関連語を Web から収集する． K_c から 3 つの語の組を取り出し，それぞれの組合せについて Web 検索エンジンにクエリを入力する[⋆1]．クエリは 3 つの語の AND 検索とする．それぞれのクエリに対して，検索結果の要約である snippet の集合が得られる．この snippet 集合中の各語の snippet 頻度を求める．クエリ q_i に対して得られた snippet の件数[⋆2]を n_i，語 w_j のクエリ q_i に対して得られた snippet 集合中での snippet 頻度を

$freq(w_j, i)$ とする時，語 w_j の内容関連度 $T(w_j)$ を式 (5) で定義する．

$$T(w_j) = \max_i \frac{freq(w_j, i)}{n_i} \quad (5)$$

また，キーワード $k \in K$ に関しては，他の語よりも高い重みを与えるために $T(k) = \max_j T(w_j)$ とする．

3.4.2 質問事例の取得

本研究では，記述スタイルの類似する質問事例の取得を 7-gram の情報のみを使って行うのではなく，言語モデルを利用して，質問事例集合から入力された質問が生成される確率を計算し，それが最大となる質問事例集合を取得する．つまり，回答として最良のモデルを与える質問事例の部分集合を求める．その結果として，取得した質問事例集合と対応する回答事例集合が回答の記述スタイルを取得するための集合として相応しいものとなることが予想される．それらの回答事例を用いて回答の言語モデルを作成し，解候補の抽出を行うことで，出力する回答の精度を向上させることができると考えられる．

これを実現する方法としては， Q&A コーパス中の全質問事例集合の中から，質問事例のすべての部分集合に対して，言語モデルを求め，そのモデルにおいて入力された質問を生成する確率を計算していき，それが最大となった質問事例の部分集合を取得することが確実な方法であり，理想的であると言える．しかし，本研究で用いる Q&A コーパスには，質問・回答事例のペアが約 90 万件含まれており，部分集合候補が膨大になり，確率計算に膨大な時間がかかるという点で，この手法は実用的ではない．そこで，処理時間を削減するために，質問事例の取得を以下の手順で行った．まず，質問事例集合全体から，簡便な類似度計算により，質問事例を多めに取得する．次に，多めに取得した質問事例集合にクラスタリングを適用することで，小部分集合に分ける．そして，小部分集合の組み合わせうち，入力された質問を生成する確率が最大となった組み合わせを取得する．クラスタリングを行うのは，すべての質問事例の部分集合について確率計算を行う場合よりも処理時間を短縮するためである．具体的には以下のような処理を行う．

- 疑問詞を中心とする語の 7-gram を使って，候補となる質問事例の数をある程度絞り込む．
- 上記候補について，すべての部分集合を作るのではなく，まずクラスタリングし，それぞれ

⋆1 検索エンジンに入力する語数が多いとヒット件数が少なくなるため，3 語ずつ組を作る． $|K_c| < 3$ の時は全語を入力する．

⋆2 取得件数には上限を設ける．

Vol.2011-NL-201 No.8
Vol.2011-SLP-86 No.8
2011/5/17

のクラスタの組に対して言語モデルを求める．

質問事例を取得する処理の概要を図 4に示す．

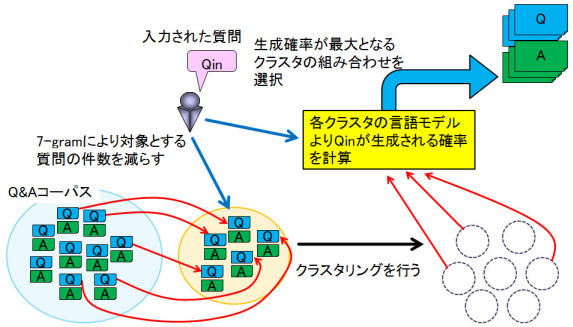

図 4 記述スタイルの類似する質問事例の取得
Fig. 4 Collecting question examples similar to an input question in terms of description style

図 4の処理の流れについて説明する．

(1) 最終的に取得する質問事例の件数を決定する．以下，これを目標取得件数とよぶ．実験では 500 件とした

(2) 質問事例 (実験では 90 万件) の中から，7-gram の一致の度合を類似度として，類似度の大きいものから質問事例を取得する．従来手法では，この段階で類似度の大きいものから目標取得件数分取得しているが，提案手法では，第一次近似の絞り込みとしてこの段階で 7-gram の類似度を用いた．実験では，この時に取得した質問事例の数を目標取得件数の 3 倍とした．その後，次の手順に移り，さらなる質問事例の選別を行う．

(3) (2) で取得した質問事例に対してクラスタリングを行い，複数のクラスタに分ける．

(4) (3) で作成した各クラスタを用いて入力された質問が生成される確率を計算し，それが最大となったクラスタの組み合わせを取得し，それらのクラスタに含まれる質問事例集合を，記述スタイルが類似する質問事例集合として取得する．

質問事例集合のクラスタリング

7-gram の類似度を用いて取得の対象となる質問事例の件数を減少させた上で，クラスタリングを行った．このとき，クラスタの数を多くするとクラスタ 1 つあたりにわずかな質問事例しか含まれていないクラスタが作成され，計算の回数が多くなってしまったり，言語モデルを作成するには質問事例数が足りないといった問題が生じることがある．一方で，クラスタの数を少なくするとクラスタ 1 つあたりに多くの質問事例が含まれてしまう場合があり，少ない質問事例集合の組み合わせでしか確率計算を行うことができない．以上のことを考慮して，クラスタ数を決定する必要がある．

文中の語の共起をある程度考慮し柔軟に類似度を計算するために，本研究では語のスキップ 2-gram を素性に用いた．その際，疑問詞や助詞，助動詞等の機能語と一部の内容語を表層表現 (読み) のまま用い，その他の語は品詞で代表させることにより，質問と回答の双方を一般化した．表層表現のまま用いた内容語とは，質問の焦点になりやすい名詞 (「理由」，「方法」，「意味」，「違い」等) やコーパス中で出現頻度が高い動詞と形容詞である．質問の焦点にやりやすい名詞は，コーパス中で「X はなんですか」や「X を教えて」のような現れ方をする名詞 X で頻度の高いものから選択した．スキップ 2-gram とは，数語の間隔を許した n-gram のことである．

質問事例，回答事例をクラスタリングする際，以下の 3 つの場合が考えられる．

- 質問・回答事例双方のスキップ 2-gram の類似度を用いてクラスタリングした場合
- 質問事例のスキップ 2-gram の類似度を用いてクラスタリングした場合
- 回答事例のスキップ 2-gram の類似度を用いてクラスタリングした場合

質問・回答事例双方のスキップ 2-gram を素性とする場合は，質問事例と回答事例双方の特徴を考慮してクラスタリングするため，質問事例の類似度と回答事例の類似度をそれぞれ別の空間で計算し，両者を合わせたものをペアの類似度とした．クラスタリングの手法は，非階層的クラスタリングの手法である k-means 法[9] を用いた．

クラスタの選別

前節の処理で作成されたクラスタについて，どのクラスタの組み合わせで言語モデルを作成した場合に，入力された質問を生成する確率が最大となるかを，山登り法で求めた．

確率計算には，3.3.1節で説明した式 (1) を用いることとし，次のような手順で行った．

(1) すべてのクラスタの中から，入力された質問の生成確率が最大となる言語モデルを生成するクラスタを 1 つ求め，候補とする．

(2) すでに候補となっているクラスタ以外の別のクラスタを 1 つ加えて、新しいクラスタの組み合わせを作成し，そのクラスタの組み合わせから，改めて言語モデルを計算し，入力された質問の生成確率を計算する．生成確率が上昇したもののうち，生成確率が最大となったクラスタを新たな候補として加える．どのクラスタを加えても確率の上昇が見られなくなった場合場合は，確率の減少が最も小さいクラスタを

Vol.2011-NL-201 No.8
Vol.2011-SLP-86 No.8
2011/5/17

候補とする．

(3) ステップ (2) を繰り返し行い，候補としたクラスタに含まれる質問事例の件数の合計が目標取得件数を超えた時点で終了．

3.4.3 文書検索と整形

キーワード及びその複合語から 3 つの集合 K， K_c， $K_c \cup K_p$ を用意する．各集合ごとに，集合内の語の AND 検索をクエリとして Web 検索エンジンに入力する．得られた検索結果から各文書の URL を取得し，HTML 文書をダウンロードする．タグの除去等を行ない，HTML 文書をプレーンテキストに変換する．

3.4.4 解候補の抽出

3.3節で説明した，式 (1) の言語モデルの確率推定式より，3.4.2節で取得した質問事例と対をなす回答事例集合の言語モデルを求める．3.4.3節で取得した検索文書中の各文について，その言語モデルにおける生成確率を求め，その文が回答事例に対して記述スタイルという観点でどれだけ近いかを判定する．

ここで，式 (1) による計算上，文が長ければ長いほど当然確率は低くなる．よって 3.3.1節で求めた生成確率の値をそのまま用いると，スコアリングの際に，長い文ほど不利になってしまう問題が挙がる．この問題を解消するために，生成確率の文の長さによる正規化を行う．正規化の式は以下の通りである．

$$\bar{P}(E_1E_2...E_n) = \frac{1}{n}\log\{P(E_1E_2...E_n)\} \tag{6}$$

正規化を行った後，式 (5) と式 (6) を用いて検索文書中の文 S_i のスコアを式 (7) で見積もる．式 (7) の第一項は【尺度 1】を満たす度合い，第二項は【尺度 2】を満たす度合いを表わしている．

$$\mathrm{Score}(S_i) = \frac{\left\{\sum_{j=1}^{n} T(w_{ij})\right\}^{\alpha}}{\log(1+|S_i|)} \cdot \left\{\bar{P}(E_1E_2...E_m)\right\}^{1-\alpha} \tag{7}$$

ここで，n は文 S_i 中の語 w_{ij} の異なり数，m は文 S_i 中の 2-gram b_{ik} の異なり数，α は，2.1節で説明した【尺度 1】と【尺度 2】の混合比を決めるパラメータである．

式 (7) で計算される値は，文が【尺度 1】を満たす語 (キーワード及び関連語) を多く含んでいたり，【尺度 2】を満たすように，回答事例と記述スタイルがよく似ているときほど高くなる．文内でのこれらの語や文全体での記述スタイルの類似度を見るために文の長さで割っているが，短い文 (回答として不適切な場合が多い) を優遇し過ぎないために文の長さの対数をとっている．

文書中でスコアが高い文が連続した場合，極大値の 1/2 以上のスコアを持つ文をまとめて一つの解候補とし，解候補のスコアは極大値のスコアで代表する．他の文は 1 文で解候補とする．こうして得られた解候補の集合を単一リンク法でクラスタリングし，冗長性を制御する．各クラスタ内でスコアが最大の解候補を取り出し，出力する．

4. 評価実験

提案手法の有効性を検証するために，以下の評価実験を行った．本研究では，提案手法を質問応答システムに取り入れた場合に，質問応答システムの精度が向上したかどうかで評価を行うことによって，提案手法の有効性の検証を行った．評価実験には，評価型ワークショップ NTCIR-6 の QAC タスク[10] における Formal Run のテストセットに含まれる 100 問のうち，質問番号の後半 50 問を用いて提案手法を用いたシステムの性能評価を行った．

4.1 実験方法

ベースライン手法として，石下ら[1] のシステムを用意し，提案手法を用いたシステムとの精度の比較を行った．その際，【尺度 1】と【尺度 2】の混合比を決めるパラメータ α の値を変化させ性能評価した結果，最も精度が高かった値 ($\alpha = 0.5$) を用いた．提案手法として，同様に α の値を変化させかつクラスタリングを 3.4.2節のように次の 3 通りの場合に分けて質問応答を行った．

- 質問・回答事例双方のスキップ 2-gram の類似度を用いてクラスタリングした場合
- 質問事例のスキップ 2-gram の類似度を用いてクラスタリングした場合
- 回答事例のスキップ 2-gram の類似度を用いてクラスタリングした場合

Web 検索エンジンは Yahoo! JAPAN[*1] を利用した．質問内容の関連語を収集する際の snippet の取得件数は 1 クエリにつき n_i=100 件，入力された質問と類似する質問事例の Q&A コーパスからの目標取得件数は約 500 件とした．システムはスコアの降順に 5 件，解を出力する．正解判定は人手で行い，出力した回答の一部にでも正解が含まれていれば，正解とした．また，評価尺度として MRR(最上位の正解順位の逆数の，全質問平均) を用いた．更に，正解を 1 件以上出力できた質問数 (正解質問数) もそれぞれ調査した．

4.2 評価結果

評価結果を表 2〜表 4に示す．ベースラインに関しては，最も精度の高かった $\alpha = 0.5$ の時の結果を表記している．提案手法に関しては，α の値を変化させて実験を行った結果，特に精度

*1 http://search.yahoo.co.jp/

Vol.2011-NL-201 No.8
Vol.2011-SLP-86 No.8
2011/5/17

の高かった 3 つの値 ($\alpha = 0.7, 0.8, 0.9$) の結果を示す．
なお，システムは質問の型分類を行わないが，参考までに人手で分類した型ごとに分けて結果を表記している．その他に分類される質問の例として，「どうなりますか」「違いは何」「どのような」「どういう場合に」等がある．

表 2 質問・回答事例のスキップ 2-gram の類似度を用いてクラスタリングした場合
Table 2 Case of clustering using similarity with the skip 2-gram of Q&A-examples

	提案手法 ($\alpha = 0.7$)		提案手法 ($\alpha = 0.8$)		提案手法 ($\alpha = 0.9$)		ベースライン ($\alpha = 0.5$)	
質問の型	MRR	正解質問数	MRR	正解質問数	MRR	正解質問数	MRR	正解質問数
定義型	0.433	5/10	0.475	6/10	0.570	7/10	0.425	6/10
理由型	0.377	9/17	0.345	9/17	0.435	10/17	0.240	6/17
方法型	0.222	2/3	0.261	3/3	0.317	3/3	0.111	1/3
その他	0.350	9/20	0.374	13/20	0.502	14/20	0.412	14/20
全体	0.372	25/50	0.378	31/50	0.482	34/50	0.338	27/50

表 3 質問事例のスキップ 2-gram の類似度を用いてクラスタリングした場合
Table 3 Case of clustering using similarity with the skip 2-gram of question-examples

	提案手法 ($\alpha = 0.7$)		提案手法 ($\alpha = 0.8$)		提案手法 ($\alpha = 0.9$)		ベースライン ($\alpha = 0.5$)	
質問の型	MRR	正解質問数	MRR	正解質問数	MRR	正解質問数	MRR	正解質問数
定義型	0.458	6/10	0.475	6/10	0.550	6/10	0.425	6/10
理由型	0.325	8/17	0.355	8/17	0.422	9/17	0.240	6/17
方法型	0.511	3/3	0.178	2/3	0.4	2/3	0.111	1/3
その他	0.329	10/20	0.385	11/20	0.514	14/20	0.412	14/20
全体	0.365	27/50	0.380	27/50	0.483	31/50	0.338	27/50

4.3 考　　察

表 2〜表 4 より，提案手法によって，システムの精度が向上したことが分かる．

正解質問数については，表 2 の $\alpha = 0.7$ の時に 25 となり，ベースラインを下回ったが，その他の場合はすべてベースラインと同等，もしくはそれ以上の結果を得た． MRR の値は，質問の型ごとの値，総合の値ともにベースラインを上回る結果となった．

このような結果となった要因の一つとして，ベースラインではたまに存在した，日本語として相応しくない文章[⋆1]が出力されているケースが，提案手法における出力にはあまり存在しなかったことが挙げられる．このことから，解候補のスコアリングを言語モデルによる確率推定により行ったことによって，日本語の文法として正しくない文章を概ね淘汰できたからであると考えられる．

表 4 回答事例のスキップ 2-gram の類似度を用いてクラスタリングした場合
Table 4 Case of clustering using similarity with the skip 2-gram of answer-examples

	提案手法 ($\alpha = 0.7$)		提案手法 ($\alpha = 0.8$)		提案手法 ($\alpha = 0.9$)		ベースライン ($\alpha = 0.5$)	
質問の型	MRR	正解質問数	MRR	正解質問数	MRR	正解質問数	MRR	正解質問数
定義型	0.458	6/10	0.483	6/10	0.500	6/10	0.425	6/10
理由型	0.332	9/17	0.345	8/17	0.345	9/17	0.240	6/17
方法型	0.400	3/3	0.611	3/3	0.511	3/3	0.111	1/3
その他	0.527	13/20	0.543	14/20	0.502	14/20	0.412	14/20
全体	0.439	31/50	0.464	31/50	0.437	32/50	0.338	27/50

また，別の要因として，ベースラインでは多く取得されていた，入力された質問に対する回答としては不適切な表現となっている質問・回答事例の取得が少なく抑えられていたことが挙げられる．このことから，質問事例の取得をクラスタリングと言語モデルによる確率推定によって行ったことで，それと対応した回答として相応しい表現を含む回答事例が多く取得でき，それによって回答の言語モデルが洗練されたということが考えられる．

ベースラインでは正解の回答を出力できなかったが，提案手法を用いることにより正解の回答を出力できた例を以下に示す．

【質問 (入力)】京都議定書の発効には何が必要ですか。
【回答 (ベースライン)】2004 年 11 月 18 日にロシアが京都議定書の批准書を寄託したことにより発効要件が満たされ、 90 日後の 2005 年 2 月 16 日に議定書は発効されました。

⋆1 見出し語と本文がくっついているものや，文章中に無意味な記号などが現れる文など

Vol.2011-NL-201 No.8
Vol.2011-SLP-86 No.8
2011/5/17

【質問 (入力)】京都議定書の発効には何が必要ですか。
【回答 (提案手法)】京都議定書の発効には、 55 カ国以上の締約国と、先進工業国の 55% 以上の二酸化炭素排出量をもつ国の批准が必要であり、アメリカ等が交渉から離脱してしまったため、先進工業国全体の 17.4% の二酸化炭素排出量を持つロシアの批准が議定書発効の条件になっていました。

このことは表 2，表 3，表 4のように $\alpha = 0.8$ の場合よりも $\alpha = 0.9$ の場合のほうが精度が向上したことからも読み取ることができる．

クラスタリングに用いる素性については，回答事例のスキップ 2-gram を用いた場合 (表 4) が正解質問数， MRR とも安定し，なおかつ精度の高い結果となっている．このような結果となった要因としては以下のことが考えられる．取得した Q&A ペアのうち，質問事例集合については，文の長さが短く，また 1 段階目の絞り込みで 7-gram を利用しているため記述スタイルに大きな違いがないため，クラスタリングを行う際の素性としては適さない．一方で，回答事例は文の長さが質問事例に比べて長いものが多く，さまざまな記述スタイルで書かれているため，クラスタリングを行うことで，記述スタイルごとに小部分集合を作り出すことができた．このような理由から，解候補を抽出するための言語モデルの作成に用いられる回答事例のみに素性として着目することで，精度向上につながったのではないかと考えられる．

5. 結　　論

本研究では，先行研究である石下ら[1)]のシステムの精度向上を目指し，品詞と表層の混合モデルによる言語モデルをシステムに導入する手法を検討した．

提案手法として，入力された質問応答と記述スタイルの類似する質問事例の取得と，取得した質問事例と対応する回答事例と解候補の記述スタイルが類似しているかどうかの判定に言語モデルを導入した．評価型ワークショップ NTCIR-6 の QAC-4 タスクにおけるテストセット用いた評価実験を行った結果，提案手法を用いることで質問応答システムの精度が向上することを確認した．

今後の課題としては，解候補のスコアリング方法の改善，文書検索の改善，確率計算のアルゴリズムの改善，人手に依らない評価方法の検討などが挙げられる．

謝辞　本研究を遂行するにあたり，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．使用許諾を頂いた各社ならびに同データの研究利用に対してご尽力頂いた皆様に感謝いたします．また， NTCIR の運営に御尽力をいただいている皆様にも感謝いたします．

なお，本研究の一部は科学研究費補助金 (課題番号 22500124) によるものである．

参 考 文 献

1) 石下円香，佐藤　充，森　辰則： 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， pp. 339–350 (2009).
2) Han, K.-S., Song, Y.-I. and Rim, H.-C.: Probabilistic model for definitional question answering, SIGIR, pp. 212–219 (2006).
3) 森本格行，福本淳一： Why 型質問に対する回答抽出，言語処理学会第 10 回年次大会発表論文集， pp. 293–296 (2004).
4) 三原英理，藤井　淳，石川徹也：行動表現に着目したヘルプデスク指向の質問応答，言語処理学会第 11 回年次大会， pp. 1096–1099 (2005).
5) 諸岡　心，福本淳一：非 factoid 型質問に対応した質問応答システム，言語処理学会第 13 回年次大会発表論文集， pp. 958–961 (2007).
6) 水野淳太，秋葉友良：任意の回答を対象とする質問応答のための実世界質問の分析と回答タイプ判定法の検討，言語処理学会第 13 回年次大会発表論文集， pp. 1002–1005 (2007).
7) Soricut, R. and Brill, E.: Automatic Question Answering Using the Web: Beyond the Factoid, Journal of Information Retrieval - Special Issue on Web Information Retrieval, Vol.9, pp. 191–206 (2006).
8) 北　研二：確率的言語モデル，東京大学出版会 (1999).
9) 神蔦敏弘：データマイニング分野のクラスタリング手法 (1) - クラスタリングを使ってみよう !-，人工知能学会誌， pp. 59–65 (2003).
10) Fukumoto, J., Kato, T., Masui, F. and Mori, T.: An Overview of the 4th Question Answering Challenge (QAC-4) at NTCIR Workshop 6, Proceedings of the Sixth NTCIR Workshop Meeting, pp. 433–440 (2007).