情報処理学会論文誌 Vol.54 No.7 1951–1966 (July 2013)

意味的極性と単語クラスを用いたWhy型質問応答の改善

呉 鍾勲[1,a)] 鳥澤 健太郎[1,b)] 橋本 力[1,c)] 川田 拓也[1,d)] デサーガ ステイン[1,e)]
風間 淳一[f)] 王 軼謳[1,g)]

受付日 2012年8月22日, 採録日 2013年4月5日

概要：本稿では，意味的知識を用いて Why 型質問応答システムの精度を向上させるための手法を提案する．具体的には，「ネガティブな（望ましくない）事象の理由はネガティブな（望ましくない）事象であることが多い」，「ポジティブな（望ましい）事象の理由はポジティブな（望ましい）事象であることが多い」という意味的極性に関わる傾向，そして，質問に含まれた単語（たとえば「病名」）とその回答に含まれた単語間（たとえば「有害物質」，「ウィルス」，「体の部位」）の意味的な相関関係を機械学習による回答抽出に用いることにより Why 型質問応答の性能改善を図る．評価実験は人手で作成した 850 個の Why 型質問と 6 億件の Web 文書から抽出したその回答候補（各 20 個からなる）を用いて行った．NTCIR6-QAC4 の non-factoid 型質問応答のタスクにおいて，正しい回答を 1 つ以上得られた質問の数が最も多かった従来手法を実装して提案手法と比較した結果，提案手法によって最上位結果の精度が 15.2%向上したことを確認した．

キーワード：質問応答，意味的極性，評価表現，単語クラス，non-factoid 型質問

Improving Why Question Answering Using Semantic Orientation and Semantic Word Classes

Jong-Hoon Oh[1,a)] Kentaro Torisawa[1,b)] Chikara Hashimoto[1,c)]
Takuya Kawada[1,d)] Stijn De Saeger[1,e)] Jun'ichi Kazama[f)] Yiou Wang[1,g)]

Received: August 22, 2012, Accepted: April 5, 2013

Abstract: In this paper we explore the utility of sentiment analysis and semantic word classes for improving why-question answering on a large-scale web corpus. Our work is motivated by the observation that a why-question and its answer often follow the pattern that *if something undesirable happens, the reason is also often something undesirable*, and *if something desirable happens, the reason is also often something desirable*. To the best of our knowledge, this is the first work that introduces sentiment analysis to non-factoid question answering. We combine this simple idea with semantic word classes for ranking answers to why-questions and show that on a set of 850 why-questions our method gains 15.2% improvement in precision at the top-1 answer over a baseline system that achieved the best performance in the number of questions for which there is at least one correct answer in system results in the shared task of Japanese non-factoid question answering in NTCIR-6 QAC4.

Keywords: question answering, semantic orientation, sentiment analysis, semantic word class, non-factoid question

1. はじめに

質問応答の研究において，factoid 型質問に対する質問応答技術は広く研究されてきたが，Why 型質問，How-to 型

[1] 情報通信研究機構
National Institute of Information and Communications Technology, Souraku, Kyoto 619–0289, Japan
[a)] rovellia@nict.go.jp
[b)] torisawa@nict.go.jp
[c)] ch@nict.go.jp
[d)] tkawada@nict.go.jp
[e)] stijn@nict.go.jp
[f)] junichikazama@gmail.com
[g)] wangyiou@nict.go.jp

質問を含む non-factoid 型質問に対する質問応答技術の研究は比較的少ない．また，最高レベルの non-factoid 型質問応答システム [12], [16], [18] の精度（たとえば，Why 型質問に対する上位 150 個の結果で 34%の MRR [18]）は最高レベルの factoid 型質問応答システムの精度（最上位結果で 85%の精度 [4]）よりも非常に低いというのが実状である．

本稿では，このように困難なタスクであると認識されてきた non-factoid 型質問応答の中でも，特に Why 型質問応答の精度を向上させるための手法を提案する．本研究は，「ネガティブな（望ましくない）事象の理由はネガティブな（望ましくない）事象であることが多い」，そして「ポジティブな（望ましい）事象の理由はポジティブな（望ましい）事象であることが多い」という意味的極性に関わるパターンが Why 型質問とその回答によく現れるという我々の観察結果を出発点とし，このような意味的極性のパターンを機械学習による回答抽出に用いることにより Why 型質問応答の性能改善を図る．たとえば，以下のように「なぜガンになるのですか？」というネガティブな（望ましくない）事象の理由を求める質問 **Q1** に対して，「ガンのリスクを高める」というネガティブな（望ましくない）事象を説明する **A1-1** と，ガンを予防するための望ましい行為を説明する **A1-2** が回答候補として得られたとする．提案手法は，「ネガティブな（望ましくない）事象の原因はネガティブな（望ましくない）事象であることが多い」という意味的極性のパターンから **A1-1** を **Q1** の正しい回答として選ぶことができる．こうした意味的極性のパターンは non-factoid 型質問応答において，我々の知る限りこれまでに検討されたことがない．

Q1	なぜ ガンになるのですか？（ネガティブな事象）
A1-1	ニトロソアミンなどの発がん因は細胞のもつ遺伝子を変化させ，ガンのリスクを高める（ネガティブな事象）．
A1-2	健康的な体重を維持することは ガンのリスクを下げる（ポジティブな事象）．

また，本研究のもう 1 つの基本的アイディアは，Why 型質問が含む単語とその回答が含む単語間の意味的な相関関係を用いて性能の向上を図るということである．たとえば，**Q1** のように「病気」の原因を求める質問の回答は「有害物質」（たとえば，**A1-1** の「ニトロソアミン」），「ウィルス」，「身体の部位」などを表す単語を含む場合が多い．質問と回答からこのような「病気」と「有害物質」の間の相関関係を把握し，質問応答における回答抽出に応用することによって Why 型質問応答の性能向上が期待できる．このため，提案手法では単語クラスタリング手法 [10] を用いて大量の Web 文書から単語の意味的クラス（意味的に類似する単語の集合）を自動獲得し，機械学習の素性として活用する．本稿ではこのように単語クラスタリングにより得られた単語の意味的クラスを単語クラスと呼ぶ．

Why 型質問と回答における意味的極性のパターンを Why 型質問応答に適用する場合，意味的極性を持つ言語表現，すなわち評価表現の内容も考慮する必要がある．これは，回答に意味的極性が異なる複数の表現が存在する[*1]と，意味的極性のみでは Why 型質問応答の性能改善が期待できないためである．たとえば，**A1-2** が以下の例文（**A1-2'**）ように「リスクを下げる」というポジティブな（望ましい）表現と「効果的ではない」というネガティブな表現を持つと仮定すると，**Q1**，**A1-1**，**A1-2** が同様にネガティブな（望ましくない）表現を持つことになり，これらの質問と回答における意味的極性のパターンの効果が期待できない．

> "**A1-2'**：がんに良いとされる食品を食べ過ぎるのは ガンの予防に効果的ではない が，健康的な体重を維持することは ガンのリスクを下げる"

このような問題を解決するため，提案手法では Why 型質問と回答候補における評価表現の意味的極性とともに評価表現の内容（評価表現を構成している単語，係り受け関係など）も考慮した．そして，この際のデータの過疎性を回避するため，評価表現の内容を単語とともに単語クラスで表現した．最終的に以上のようなアイディアは回答候補のランキングを行う教師あり学習で使われた．

評価実験は人手で作成した 850 個の Why 型質問と 6 億件の Web 文書から抽出したその回答候補（各 20 個からなる）を用いて行った．NTCIR6-QAC4 の non-factoid 型質問応答のタスクにおいて，正しい回答を 1 つ以上得られた質問の数が最も多かった従来手法 [12] を実装して提案手法と比較した結果，提案手法によって最上位結果の精度が 15.2%向上したことを確認した．

なお，non-factoid 型質問応答（Why 型質問応答，How-to 型質問応答など）において，教師あり学習による回答リランキングに用いた先行研究 [7], [16], [18] がある．これらの先行研究では，質問に対する回答候補を文書から取り出した後，回答候補から抽出した構文情報（係り受け関係など），意味情報（単語間の因果関係，WordNet 情報など），統計情報（頻度など）などを素性として用いて学習した分類器を回答候補のリランキングに用いた．提案手法は，特に回答候補のリランキングのために，大規模な単語クラスタリングにより得られた単語クラスと意味的極性という意味的知識を用いた新たな素性を提案し，Why 型質問応答におけるその有効性を示した点が先行研究と大きく異なる．

以下，2 章と 3 章で提案手法の詳細を記し，4 章から 6 章までは評価データの作成，実験とその結果を述べる．7 章で関連研究との比較についてより詳細に説明する．8 章で

[*1] 本研究の評価データでは約 33%の正しい回答がこのような特徴を持っていた．

結論を述べる．

2. 提案手法

提案手法は図 1 に示すように**回答候補検索**と**回答リランキング**の 2 ステップからなる．回答候補検索では与えられた Why 型質問を用いて検索した文書から回答候補を抽出し，回答リランキングでは抽出した回答候補に対して機械学習による回答リランキングを行う．本研究の主眼は，単語クラスと意味的極性という意味的知識を用いて回答リランキングの性能を向上させることである．

2.1 回答候補検索

回答候補検索は，NTCIR6-QAC4 の non-factoid 型質問応答のタスクにおいて，正しい回答を 1 つ以上得られた質問の数が最も多かった Murata らの手法 [12] の我々の実装である．回答候補検索のため，まず Why 型質問に含まれた単語を文書検索の入力として与え，6 億件の Web 文書から最大 600 件の文書を検索する．文書検索には，情報検索ツール Solr [*2] を使用した．Solr による文書検索では 2 種類のブール型（*Boolean*）の検索クエリ「t_1 AND $\cdots$ AND t_n」と「t_1 OR $\cdots$ OR t_n」を用いた．ここで，$T = \{t_1, \cdots, t_n\}$ は Why 型質問が含む内容語（名詞，動詞，形容詞）の集合を表す．ここでは，2 種類のブール型の検索クエリごとに最大上位 300 件，合わせて最大 600 件の文書を検索し，これらの検索結果を統合して回答候補の抽出を行った．これは，異なるブール型の検索クエリを用いることで Why 型質問と関連ある多様な文書を検索でき，回答候補の抽出におけるカバレッジの向上が期待できるからである．次に，検索結果の文書から一連の 5 文[*3] を 1 つの回答候補として抽出した．ここで，この 5 文の抽出によって似たような内容を含む回答候補が多く取られることを防ぐため，回答候補が連接する前後の回答候補と 1 文ずつだけ共有するようにした．

質問 q に対して抽出された回答候補 ac は式 (1) によってランク付けられ，上位 20 個の回答候補を次の段階である回答リランキングの入力として与える．また，Murata らの手法 [12] と同様に Why 型質問応答の手がかりとなる単語（「理由」，「原因」，「要因」）を質問に含まれた内容語の

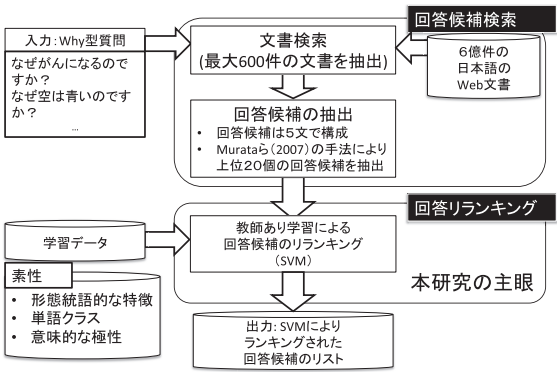

図 1 システムの構成図
Fig. 1 System architecture.

集合 T に追加して Murata ら [12] が提案した式 (1) によるランキングに用いた．

$$S(q, ac) = max_{t_1 \in T} \sum_{t_2 \in T} \phi(t_1, t_2) \times \log(ts(t_1, t_2)) \quad (1)$$

$$ts(t_1, t_2) = \frac{N}{2 \times dist(t_1, t_2) \times df(t_2)}$$

ここで，T は q と ac にともに現れる内容語（名詞，動詞，形容詞）の集合を表す．N は検索対象になった文書の数（6 億件），$dist(t_1, t_2)$ は ac においての t_1 と t_2 の間の距離（t_1 と t_2 の間に含まれる文字数）（$t_1 = t_2$ の場合，$dist(t_1, t_2) = 0.5$），$df(t)$ は t が現れる文書の数，$\phi(t_1, t_2) \in \{0, 1\}$ は $ts(t_1, t_2) > 1$ であるか否かを示す指示関数である（$ts(t_1, t_2) > 1$ であると 1，もしそうでなければ 0）．

回答候補検索における文書検索のため，Murata らの手法 [12] では OKAPI を用いたが，我々の実装ではベクトル空間モデルとコサイン類似度に基づいた文書スコアリング手法を用いたことを承知されたい．

2.2 回答リランキング

回答リランキングは教師あり学習によって作られた分類器（本稿では SVM [9]）によって行われる．分類器の学習には，形態素，文節，係り受け関係などの形態統語的な特徴に関わる素性，大規模な単語クラスタリングにより得られた単語クラスに関わる素性，そして意味的極性に関わる素性が用いられた．これらの素性のうち，形態統語的な特徴に関わる素性は既存研究 [7], [16], [17], [18] においても用いられていたものであるが，単語クラスに関わる素性，そして意味的極性に関わる素性は本研究で新たに提案されたものである．最終的には，学習した分類器による回答候補の分類を行い，回答候補に与えられた分類器のスコアによって回答候補がリランキングされる．実験のため，人手で 850 個の Why 型質問を作成し，これらを回答候補検索の入力とした．そして，上述した Murata らの手法で抽出した上位 20 個の回答候補で評価データを作成した．この

[*2] すべての文書に対して JUMAN による形態素解析を行い，その解析結果を文書索引のために用いた．そして，ベクトル空間モデルとコサイン類似度に基づいた Solr の文書スコアリング手法を用いて文書検索を行った．Solr の詳細については http://lucene.apache.org/solr を参照されたい．

[*3] 回答候補の長さを決めるため，予備実験を行った．予備実験では 20 個の質問に対して 10 文の長さを持つ回答候補を回答候補検索を用いて抽出し（上位 20 個），これらを人手で判定した（合計 400 個の回答候補のうち，35 個が正しい回答と判定された）．そして，正しい回答に現れる回答部分を人手で抜き出しその長さを分析した．その結果，回答部分が 3 文以内のものが約 86%，5 文以内のものが約 97%であった．これらの結果の基づいて本稿の回答候補の長さを 5 文とした．

評価データの作成方法やその他の詳細については 4 章を参照されたい．実験ではこの評価データにおける 10-fold cross validation 方法で提案手法の有効性を検証した．

3. 回答リランキングのための素性

本章では回答リランキングのための素性を説明する．具体的には，質問と回答候補のテキストから MSA 素性（形態統語的な特徴に関わる素性：Features by Morphological and Syntactic Analysis），SWC 素性（単語クラスに関わる素性：Features by Semantic Word Classes），SA 素性（意味的極性に関わる素性：Features by Sentiment Analysis）という 3 種類の素性を抽出し，SVM の教師あり学習の素性として使用する．素性の詳細については以下の節で述べる．

3.1 MSA 素性：形態統語的な特徴に関わる素性

質問と回答候補のテキストに対して JUMAN [*4]による形態素解析と KNP [*5]による構文解析を行い，この解析結果から形態素，文節，係り受けの n-gram（以下，これらを形態統語 n-gram と呼ぶ）を抽出する．これらをもとにした素性を**表 1** の MSA1〜MSA4 とする．ここで，係り受けの n-gram を n 個の文節における一連の係り受け関係と定義する．なお，係り受けの 1-gram は文節 1-gram と同一であるため，素性として使わない．これらの素性は既存研究 [7], [16], [17], [18] においても用いられていたものである．

たとえば，**A1-1** から抽出した形態素 3-gram「ガン/の/リスク」は MSA1 素性として用いられ，かつ，質問 **Q1** の名詞「ガン」を含むため，MSA2 素性として用いられる．そして，「ガン/の/リスク」のうち，ガンを QT（質問中の単語であることを示す記号）に置き換えた「QT/の/リスク」も MSA2 素性として作成される．また，形態素 3-gram「なる/理由/は」は手がかりの単語「理由」が含まれたため，MSA3 素性として用いられる．MSA 素性は「質問に『ガンになる』の表現があると，回答には『ガン（QT）のリスクを高める』，『ガン（QT）の原因は』，『ガン（QT）になる（QT）原因は』のような表現が現れる場合が多い」などの質問と回答候補間の形態統語的な特徴やその相関関係を示すための素性として用いられる．

表 1 回答リランキングに用いられた MSA 素性．n-gram の n は $n \in \{1, 2, 3\}$

Table 1 MSA features used in our proposed method. n in n-gram is $n \in \{1, 2, 3\}$.

MSA1	質問や回答候補のテキストに現れる形態素，文節，係り受けの n-gram．質問の n-gram と回答候補の n-gram は区別される．
MSA2	回答候補から取り出した **MSA1** の n-gram のうち，質問に含まれる単語を含むもの．そして，これらの n-gram に現れる質問中の単語を記号 QT に置き換えた n-gram．
MSA3	**MSA1** の n-gram のうち，手がかりとなる単語（「理由」，「原因」，「要因」）を含むもの．質問の n-gram と回答候補の n-gram は区別される．
MSA4	質問の内容語のうち，回答候補に現れたものの比率．

[*4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[*5] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

3.2 SWC 素性：単語クラスに関わる素性

単語クラスは，6 億件の Web 文書から取り出した名詞間の係り受け関係，名詞と動詞間の係り受け関係を名詞の文脈情報とし，類似する文脈を持つ名詞を式 (2) の隠れクラスモデルを用いる名詞のクラスタリングアルゴリズム [10] でクラスタリングすることにより獲得された．

$$p(n, v, r) = \sum_c p(n|c)p(\langle v, r\rangle|c)p(c) \quad (2)$$

ここで，v は動詞，n は v と係り受け関係 r にある名詞，c は隠れクラスである．名詞 n は，複合語や修飾語の付いた名詞を含む．関係 r は，名詞につづく助詞で表す．$p(n|c)$，$p(\langle v, r\rangle|c)$，$p(c)$ は EM アルゴリズム [8] によって推定される．

名詞 n の単語クラスは $c = argmax_{c^*}\ p(c^*|n)$ により判定され，合計 550 万個の名詞を 500 個の単語クラス[*6]に分類した．本稿の例に用いられた「化学物質」，「栄養素名」，「病名」，「状況」などの意味を表す単語クラスが獲得結果に含まれていることを確認した．以下はこれらの単語クラスに含まれている上位 10 個の名詞（$p(c|n)$ による順位）を示している．以下にあげられた「化学物質」のような単語クラス名は本稿の説明のために人手で付けられており，提案手法の実装では単語クラスタリング手法により自動的に与えられた単語クラスの id を用いて単語クラス素性を作成した．

化学物質：アセチレン類，水素化生成物，リン酸モノエステル，アルカリ土類金属化合物，グリシジルメタクリレート，レボグルコサン，アンモニア塩，ハロゲン化有機化合物，ハロゲン化有機化合物，アルキン類
栄養素名：糖質，炭水化物，水分，ミネラル，食塩，砂糖，糖分，脂肪，カルシウム，栄養素
病名：インフルエンザ肺炎，多発単神経炎，がん，口腔白板症，肥厚性硬膜炎，腎性低尿酸血症，馬原虫性脊髄脳炎，鈍的腹部外傷，心弁膜症，上行結腸癌
状況：習熟，老朽化，不足，汚れ状況，流動，異常，経営危険度，変位，応急危険度，通電状況

このように大規模な単語クラスタリングにより得られた

[*6] 単語クラスの数は経験的に決められた．提案手法における単語クラスの数による影響については 6.2 節で述べる．

表 2 回答リランキングに用いられた SWC 素性．n-gram の n は $n \in \{1,2,3\}$

Table 2 SWC features used in our proposed method. n in n-gram is $n \in \{1,2,3\}$.

SWC1	**MSA1** の n-gram に現れる単語を単語クラスに置き換えた n-gram のうち，単語クラスを含むもの．これらを単語クラス n-gram と呼ぶ．質問からの単語クラス n-gram と回答候補からの単語クラス n-gram は区別される．
SWC2	回答候補からの単語クラス n-gram のうち，単語クラスの元になる単語が質問に含まれた内容語である n-gram．そして，これらの n-gram に現れる質問中の単語に対応する単語クラスを他の単語クラスと区別した n-gram．

単語クラスは Why 型質問が持つ単語とその回答が持つ単語間の意味的な相関関係を示すための素性作成に用いられた（**表 2** の SWC1 と SWC2）．たとえば，以下の質問と回答が回答リランキングのための分類器に正例として与えられると「質問に病名の単語があると，その回答には栄養素名の単語が現れる場合が多い」というパターンを学習することが可能になる．ここで，$W_{condition}$，$W_{disease}$，$W_{nutrients}$ は「状況」，「病名」，「栄養素名」の単語クラスを表すとする．

Q2	なぜ くる病（$W_{disease}$）が起こるのですか？
A2	ビタミン D（$W_{nutrients}$）の 不足（$W_{condition}$）によって くる病（$W_{disease}$）が起こります．

理論的には，学習データの質問と回答における単語間の共起関係（たとえば，「くる病」と「ビタミン D」間の共起関係）を用いることにより単語レベルの相関関係が学習できるが，単語間の共起関係は学習データの単語対に依存するため，小さいサイズの学習データに含まれていない単語対が現れる質問と回答への適用が難しい．一方，単語クラスを用いることによって質問と回答における単語間の意味的な相関関係を小さいサイズの学習データから学習することが可能になり，学習データに含まれていない単語対にもこれらの意味的な相関関係を適用することが可能になる．

単語クラスの素性作成のため，質問と回答候補の形態統語 n-gram に現れる単語を単語のクラスで置き換え得られた n-gram のうち単語クラスを含むもののみを取り出す．これらを単語クラス n-gram と呼び，表 2 に定義されている素性 SWC1 と SWC2 として用いる．

たとえば，**A2** から取り出せる「$W_{condition}$ によって/$W_{disease}$ が/起こる」という文節レベルの単語クラス 3-gram は単語クラス $W_{disease}$ が質問中の単語「くる病」の単語クラスのため，SWC1 と SWC2 の両方で素性として用いられる．また，質問文中の単語を含む単語クラスを他の単語クラスと区別させるため $W_{disease}$ を QT:$W_{disease}$ に置き換えた「$W_{condition}$ によって/QT:$W_{disease}$ が/起こる」を SWC2 素性として用いる．

「なぜ くる病（$W_{disease}$）が起きる？」のような病気の原因に求める質問の回答には「ビタミン D の 不足（$W_{condition}$）は くる病（QT:$W_{disease}$）の原因である」，「ビタミン D の 不足（$W_{condition}$）が起こす くる病（QT:$W_{disease}$）」のような表現が含まれている場合が多い．SWC 素性（特に SWC2 素性）は回答に現れるこのような表現から回答文における質問と回答の単語間の意味的な相関関係（$W_{disease}$ と $W_{condition}$ 間の相関関係）を学習するために用いられる．

3.3 SA 素性：意味的極性に関わる素性

SA 素性は「単語の意味的極性に関わる素性」（SA@W 素性）と「評価表現とその意味的極性に関わる素性」（SA@P 素性）に分類される．SA 素性の作成には一般公開されている「意見（評価表現）抽出ツール v1.2」[*7]とこのツールに含まれている単語の意味的極性辞書を用いる．

3.3.1 意見（評価表現）抽出ツール

「意見（評価表現）抽出ツール」は Nakagawa らの手法 [13] を実装したものである．このツールは，機械学習を使って文章に含まれている「何らかの事象に対する意見や評判を表す言語表現」，（いわゆる，評価表現[*8]）を抽出し，これらの評価表現に対する意味的極性[*9]，すなわちポジティブ（望ましい）とネガティブ（望ましくない），を判定する．たとえば，**Q2** の「くる病が起こりますか」と **A2** の「ビタミン D の不足によってくる病が起こります」はネガティブな意味的極性を持つ評価表現として抽出できる．評価表現の意味的極性の判定では，単語の意味的極性辞書から与えられた単語の意味的極性と構文木の部分木の意味的極性を手がかりとして用いる．

本稿では「高度言語情報融合フォーラム」（ALAGIN）で会員限定で公開されている意見（評価表現）抽出ツール用のモデルデータと単語の意味的極性辞書データ（辞書は約 35,000 語からなる）を用いて単語の意味的極性と評価表現を抽出した．このような設定での「意見（評価表現）抽出ツール」の性能（適合率，再現率，F 値）は**表 3** のように報告されている．評価表現の抽出における性能は，抽出した評価表現の末尾（主辞）が正解データと一致したかどうかに基づいている．また，「意味的極性の判定」の評価結果は評価表現が正しく抽出されたという仮定のうえで得られたものである．ツールの性能についてのより詳細な説明は

*7 http://alaginrc.nict.go.jp/opinion/index.html

*8 本稿で用いられた評価表現は「意見（評価表現）抽出ツール」により抽出されたものであり，「良い」，「悪い」などの 1 語の場合だけではなく，「動詞句」，「文」などの 2 語以上の表現からなるものも含む．

*9 「意見（評価表現）抽出ツール」は評価表現の意味的極性とともに出来事，批評，採否などの評価表現のタイプも出力するが，本研究では評価表現の意味的極性のみを用いた．

表 3 意見（評価表現）抽出ツールの性能
Table 3 The performance of *opinion extraction tool*.

	適合率	再現率	F 値
評価表現の抽出	0.602	0.408	0.486
意味的極性の判定（肯定）	0.873	0.893	0.883
意味的極性の判定（否定）	0.866	0.842	0.854

「意見（評価表現）抽出ツール v1.2」のホームページ[*10]を参照されたい．

3.3.2 単語の意味的極性（SA@W）

SA@W 素性を作成するため，質問と回答候補から抽出した形態統語 n-gram に現れる単語を単語の意味的極性辞書によって意味的極性（ポジティブとネガティブ）に置き換え，置き換えた n-gram のうち単語の意味的極性を含むもののみを取り出す．これらを単語極性 n-gram と呼び，質問と回答における単語の意味的極性間の相関関係を示すための素性である**表 4** の SA@W1 と SA@W2 の作成に用いる．たとえば，回答リランキングのための分類器の学習データに以下の例が正例として含まれていると「質問にネガティブな意味的極性を持つ単語があると，その回答にもネガティブな意味的極性を持つ単語がある場合が多い」というパターンを学習することが可能になる．

Q2	なぜ くる病（W^-）が起こりますか？
A2	ビタミン D の 不足（W^-）によって くる病（W^-）が起こります．

そして，単語クラスと単語極性の対で形態統語 n-gram を置き換え，置き換えた n-gram のうち単語クラスと単語極性の対を含むもののみを取り出す．これらを単語クラス/極性 n-gram と呼び，表 4 の SA@W3 と SA@W4 を作成するために使う．たとえば，**A2** の「不足」は「状況」の意味を表す単語クラスとネガティブな意味的極性を持つため，$W^-_{condition}$ と表現でき，「くる病」は「病気」の意味を表す単語クラスとネガティブな意味的極性を持つため，$W^-_{disease}$ と表現できる．**A2** から取り出せる「$W^-_{condition}$ によって/$W^-_{disease}$ が/起こる」という文節レベルの単語クラス/極性 3-gram は SA@W3 素性になり，$W^-_{disease}$ が質問中の単語「くる病」を含むため SA@W4 素性としても用いられる．そして，この単語クラス/極性 3-gram の $W^-_{disease}$ を QT:$W^-_{disease}$ に置き換えた「$W^-_{condition}$ によって/QT:$W^-_{disease}$ が/起こる」も SA@W4 素性として用いる．これらの SA@W 素性は「質問が『ネガティブな病名』を表す単語を含むと，回答には『ネガティブな状況』を表す単語がある場合が多い」という有意味な相関関係を示すために用いられる．

[*10] http://alaginrc.nict.go.jp/opinion/index.html の「10. 解析精度」に説明されている．

表 4 回答リランキングに用いられた SA@W 素性．n-gram の n は $n \in \{1,2,3\}$
Table 4 SA@W feature sets used in our proposed method. n in n-gram is $n \in \{1,2,3\}$.

SA@W1	**MSA1** の n-gram に現れる単語を単語の意味的極性辞書（意見抽出ツールの辞書を使用）によって意味的極性（ポジティブとネガティブ）に置き換えた n-gram のうち，置き換えた意味的極性を持つもの．これらを単語極性 n-gram と呼ぶ．質問からの単語極性 n-gram と回答候補からの単語極性 n-gram は区別される．
SA@W2	回答候補からの単語極性 n-gram のうち，意味的極性の元になる単語が質問の内容語である n-gram．そして，これらの n-gram に現れる質問中の単語に対応する単語極性を他の単語極性と区別した n-gram．
SA@W3	**MSA1** の n-gram に現れる単語を単語クラスと単語の極性の対に置き換えた n-gram のうち，置き換えた対を持つもの．これらを単語クラス/極性 n-gram と呼ぶ．質問からの単語クラス/極性 n-gram と回答候補からの単語クラス/極性 n-gram は区別される．
SA@W4	回答候補からの単語クラス/極性 n-gram のうち，単語クラス/極性の元になる単語が質問の内容語である n-gram．そして，これらの n-gram に現れる質問中の単語に対応する単語クラス/極性を他の単語クラス/極性と区別した n-gram．

3.3.3 評価表現とその意味的極性（SA@P）

「意見（評価表現）抽出ツール」を使い，質問と回答候補のテキストから意味的極性を持つ評価表現を抽出し，これらの評価表現から形態統語 n-gram，単語クラス n-gram，単語クラス/極性 n-gram を取り出す．そして，取り出した n-gram と評価表現の意味的極性を合わせて**表 5** の SA@P1〜SA@P10 を作成する．ここで，SA@P 素性の作成に用いた評価表現は質問の内容語が含まれた回答候補の文から取り出したもののみに限定した．これは，質問の内容語を含んでいない回答候補の文から取り出した評価表現はその数が多いが，そのほとんどが質問との関連性が低いものであったため，回答リランキングに悪影響を与えることが予備実験により明らかになったからである．

SA@P 素性は用いられた情報の種類によって以下の 3 つのカテゴリに分類できる．

意味的極性の一致：SA@P1 と SA@P2．質問に異なる意味的極性を持つ複数の評価表現がある場合，各々の意味的極性に対して SA@P1 と SA@P2 素性を作成する．

評価表現の形態統語的な特徴：SA@P3〜SA@P5．これらの素性は MSA1，MSA2，MSA4 の評価表現版である．

評価表現の単語クラス n-gram と単語クラス/極性 n-gram：SA@P6〜SA@P10．SA@P6 と SA@P7 は SWC1 と SWC2 の評価表現版，SA@P8 と SA@P9 は

表 5 回答リランキングに用いられた SA@P 素性．n-gram の n は $n \in \{1,2,3\}$

Table 5 SA@P feature sets used in our proposed method. n in n-gram is $n \in \{1,2,3\}$.

SA@P1	質問の評価表現の意味的極性と回答候補の評価表現の意味的極性が一致するか否かを示す指示関数．一致する対があると 1 を持つ．評価表現は意見（評価表現）抽出ツールを用いて抽出する．
SA@P2	SA@P1 が 1 になった際の意味的極性，ポジティブとネガティブの二値．
SA@P3	評価表現に現れる形態統語 n-gram と評価表現が持つ意味的極性の対．質問の評価表現からの n-gram と回答候補の評価表現からの n-gram は区別される．
SA@P4	回答候補の評価表現から取り出した **SA@P3** の n-gram のうち，質問に含まれる単語を含むもの．そして，これらの n-gram に現れる質問中の単語を記号 QT に置き換えた n-gram．
SA@P5	質問の内容語のうち，回答候補の評価表現を含む文に現れたものの比率．
SA@P6	評価表現の単語クラス n-gram と評価表現が持つ意味的極性の対．質問の評価表現からのものと回答候補の評価表現からのものは区別される．
SA@P7	回答候補の評価表現からの単語クラス n-gram と評価表現が持つ意味的極性の対のうち，単語クラスの元になる単語が質問に含まれる単語であるもの．そして，これらに対して質問中の単語に対応する単語クラスを他の単語クラスと区別したもの．
SA@P8	評価表現の単語クラス/極性 n-gram と評価表現が持つ意味的極性の対．質問の評価表現からの単語クラス/極性 n-gram と回答候補の評価表現からの単語クラス/極性 n-gram は区別される．
SA@P9	回答候補の評価表現からの単語クラス/極性 n-gram と評価表現が持つ意味的極性の対のうち，単語クラス/極性の元になる単語が質問の内容語である n-gram．そして，これらの n-gram に現れる質問中の単語に対応する単語クラス/極性を他の単語クラス/極性と区別した n-gram．
SA@P10	質問からの **SA@P6** の n-gram と回答候補からの **SA@P6** の n-gram の対．それぞれの n-gram の元になる評価表現の意味的極性が一致する場合のみ（**SA@P1** の指示関数が 1 である評価表現間のみ），n-gram と評価表現の意味的極性が素性として用いられる．

SA@W3 と SA@W4 の評価表現版である．SA@P10 は，意味的極性が一致する質問の評価表現と回答の評価表現から取り出した単語クラス n-gram の対を用いて作成する．たとえば，質問からの評価表現「くる病が起こりますか」とその回答候補からの評価表現「ビタミン D の不足によってくる病が起こります」がネガティブな意味的極性を持ち，質問の評価表現から「$W_{disease}$ が/起こる」，回答候補の評価表現から「$W_{nutrient}$ の/不足」が文節レベルの単語クラス 2-gram として抽出できるとする．質問の評価表現とその回答候補の評価表現の意味的極性が一致するため，上記の文節レベルの単語クラス 2-gram の対〈質問：$W_{disease}$ が/起こる，回答：$W_{nutrient}$ の/不足〉を SA@P10 素性として用いる．この SA@P10 素性は，「質問に『病気が起こる』という意味を表すネガティブな評価表現があると，回答には『栄養素の不足』という意味を表すネガティブな評価表現がある場合が多い」という相関関係を示すために用いられる．

これらの素性は提案手法の性能向上において重要であるが，これは，論文冒頭で述べたように，多くの場合「ポジティブな（望ましい）事象の理由はポジティブな（望ましい）事象である」「ネガティブな（望ましくない）事象の理由はネガティブな（望ましくない）事象である」という傾向があるからである．なお，こうした評価表現の意味的極性の利用は，単語クラスによるデータの過疎性の回避があって（SA@P6〜SA@P10），より有効になると考えられる．

4. 評価データ

評価データは質問作成と回答候補の判定の 2 段階で作成された．

4.1 Why 型質問作成

実験のため，QS1，QS2，QS3 の 3 つの質問集合からなる評価データを用意した．

QS1 は，Yahoo!知恵袋[*11]から質問を自動抽出して作成された．抽出対象になった質問は，1 つの文で構成され，かつ，疑問詞「なぜ」を含んでいる Why 型質問である．そして，抽出した質問に対して追加の文脈情報がなくても理解できるか否かを人手で判定し，追加の文脈情報がなくても理解できるもののみをランダムに選択して QS1 の質問として用いた．たとえば，「野球の WBC になぜボクシングの WBC は抗議しないんでしょうか？」（抗議の対象が漠然としている）と「なぜ，オークションは未成年者の参加不可なのに，参加するのですか？」（どのオークションについての質問なのか不明）は QS1 の質問として選択されなかった．

また，人間が回答することを想定する Yahoo!知恵袋の質問は，機械が回答することを想定する質問応答システム向けの質問と形式的に大きく異なる．Yahoo!知恵袋の質問は，質問自体だけではなく質問の意図，質問した経緯などを含む多くの背景情報が記載されている傾向がある．質問者は他のユーザが自分の質問をより正確に理解できるように質問を作成する傾向があるからである．一方，質問応答

[*11] 本研究では 2004 年 4 月から 2009 年 4 月からの約 1,600 万質問を含む「Yahoo!知恵袋データ第 2 版」を用いて QS1 の質問を抽出した．

表 6 QS1 と QS2 の質問における形式的な差
Table 6 The difference in style between questions in QS1 and QS2.

QS1	海水をバケツですくっても，無色透明な液体なのに，宇宙や宇宙ステーションなど，遠く遠くから見ると青く見えるのはなぜですか？
	毎日学校に行く時は自転車の空気は 1 カ月経っても抜けないのに，休みになり乗らなくなると急に空気が抜けるのはなぜですか？
QS2	佐藤栄作がノーベル平和賞を受賞したのはなぜですか？
	ハリウッドが映画で有名なのはどうしてですか？
	松下電器産業が社名をパナソニックに変更したのはなぜですか？

表 7 QS3 の質問作成に使われたパッセージと作成された質問の例
Table 7 An example of passages used for creating questions in QS3.

パッセージ	昔は加齢に伴う動脈硬化が**原因**となって起こる心筋梗塞が主でした．しかし，最近は食生活の欧米化等に伴う血管内隆起性病変の性状の変化により，コレステロールに富むアテロームプラークが増加し，血栓の形成がスピード化され突然死を招くケースが増加してきているそうです．日常生活の中で心筋梗塞を予防するには，適度な運動を毎日行うこと，休養を十分とること，そしてゆっくり噛んで楽しくいただく一家団らんの食生活を送ることが大切であると強調されました．
作成した質問	食生活の欧米化が心筋梗塞の増加を招くのはなぜですか？

システム向けの質問は，システムに質問の意図などを長く説明せず，Yahoo!知恵袋の質問より簡潔な文で書かれている場合が多い．

このような質問の形式的な差を考慮した評価のため，質問応答システム向けの質問によりふさわしいとされる質問の集合 **QS2** を作成した．**QS2** の質問は著者以外の 6 名のアノテータにより作成された．アノテータは，作成する質問が質問応答システム向けのものであることを念頭に置いて Why 型質問を作成した．具体的には，作成する質問は Yahoo!知恵袋のように人からの回答を求めるものではないため，回答者に質問の意図や背景情報を理解させるための余分な説明は必要ないことを教示した．そして，アノテータは「質問応答システムを用いて自分が求める情報を探すときにどのように質問を作成するか」を意識したうえで質問を作成した．さらに，作成した Why 型質問が実際の事象に関するものかを Web で確認する．たとえば，「木星はなぜ青いのですか？」といった質問は「木星は青い」ということが事実ではないため，QS2 の質問としては採用されない．しかし，QS2 の質問に対する正確な回答が Web 上に存在するかどうかについて，アノテータは確認しないこととした．つまり，Web 上に回答が存在しない質問が QS2 に含まれている可能性もある．

表 6 は QS1 と QS2 における典型的な質問の例を示している．QS1 の質問は，QS2 の質問より長くて質問に関する多くの背景情報を持っていることが分かる．たとえば，「海水をバケツですくっても，無色透明な液体なのに，宇宙や宇宙ステーションなど，遠く遠くから見ると青く見えるのはなぜですか？」という質問は「なぜ海は青いのですか？」という質問に加えてそれに関する背景情報が含まれていると考えられる．

QS3 の質問は，本研究の対象文書に正しい回答があると保証されるものとした．質問作成のため，前述した 6 億件の Web ページ中の連続する 3 文で構成されたパッセージから「ある事象の理由や原因を説明する部分」を探し，その部分が回答になる Why 型質問を著者以外の 3 名のアノテータが人手で作成した．しかし，もととなるパッセージを対象文書からランダムに選択すると，そこに「ある事象の理由や原因を表す説明」が含まれる可能性は非常に低いため，パッセージを「理由」,「原因」,「要因」という手がかりとなる単語を含む連続 3 文に限定した．たとえば，**表 7** のように与えられたパッセージからその下線部が回答になる質問「食生活の欧米化が心筋梗塞の増加を招くのはなぜですか？」が作成できる．

当然ながら，このような設定には「実世界のユーザは検索対象になる文書に求める回答があるか否かを気にせず，自分が知りたい事象の理由や原因に関わる Why 型質問をする」という現実とはズレがある．しかし，QS3 の質問作成に使われたパッセージはその質問の正しい回答であるため，「回答候補検索の理想的な設定上」での回答リランキングの性能を評価するために用いることができる．つまり，QS3 の質問に対する回答候補検索の結果にその質問作成のもとになったパッセージを 1 つの回答候補として加えることにより「いつも 1 つ以上の正しい回答が含まれた回答候補を出力する理想的な回答候補検索モジュール」が作られ，この条件での回答リランキングの評価実験が可能になる．これが QS3 を作成した主な目的である．実験の詳細は 5.3 節で述べる．なお，QS3 は上述したような意味で「現実にユーザが発する可能性が低い質問」（QS3）を用いて作成した学習データにより「実世界のユーザによる質問」（QS1 と QS2）に対する質問応答の性能改善が可能であるかを検証するためにも用いた．

最後に QS1，QS2，QS3 に含まれているすべての質問に対してトピック単語を人手で抽出し，同一のトピック単語を持つ質問からは 1 つのみをランダムに選択した．これは，作成した質問集合に特定のトピックを持つ質問が多く含まれることを防ぐためである．トピック単語は，質問を「X に関するある事象の理由を求める質問」と解釈したと

き，X に相当する質問中の単語と定義される．たとえば，「Twitter の投稿文字数が 140 字に限定されているのはどうしてですか？」は「Twitter について投稿文字数が 140 字に限定されている理由を求める質問」と解釈でき，Twitter をこの質問のトピック単語と抽出できる．最終的には，合計 850 個の質問（QS1：250 質問，QS2：250 質問，QS3：350 質問）が作成され，これらを用いて評価データを作成した．

4.2 回答候補の判定

次に，作成した 850 個の質問を提案手法の回答候補検索の入力として得られた上位 20 個の結果を 3 名のアノテータ（著者以外）が判定した．与えられた回答候補が質問の正しい回答か否かについて 3 名が判定を行い，3 名の判定結果における多数決によって最終判定結果を得た．3 名の判定結果は相当な一致率（Fleiss の kappa 値で 0.634）を示した．判定結果を見ると 519 個の質問（850 個の 61.1%）に対して上位 20 個の結果に正しい回答が含まれており，これらの質問に対する正しい回答の個数は平均 4.1 個であった．

4.3 評価データの作成

QS1，QS2，QS3 とこれらの回答候補の判定結果を用いて以下のように評価データ Set1 と Set2 を用意した．Set1 と Set2 における質問の数と評価データの量[*12]を**表 8** に示す．

- **Set1**：QS1 と QS2 の質問とこれらの上位 20 個の回答候補で構成される．実験では 10-fold crosss validation のための評価データとして用いた．
- **Set2**：QS3 の質問とその上位 20 個の回答候補で構成され，以下の 2 つの目的で使われた．

(1) 現実にユーザが発する可能性の低いと考えられる人工的に作られた質問が実世界の質問応答システムの性能改善に寄与するかを試すための学習データ
(2) 回答候補検索の理想的な設定上で，回答リランキングの性能を評価するための評価データ（QS3 の質問作成に使われたパッセージが回答候補として追加される）

なお，質問作成と回答判定を含む一連の評価データの作成作業においてアノテータは意味的極性をいっさい考慮していない．そして，評価データにおける意味的極性の情報は 3.3.1 項に述べた「意見（評価表現）抽出ツール」と単語の意味的極性辞書によって自動的に与えられた．評価データに意味的極性を自動付与した結果，約 35%の質問と約 40%の回答候補が 1 つ以上の評価表現を持ち，約 45%の質問と約 85%の回答候補では 1 つ以上の単語が意味的極性を持つことを確認した．

[*12] 我々の回答候補検索では 2 種類のブール型の検索クエリ（AND 検索と OR 検索）を用いて文書検索を行い，文書検索部は結果として大量の文書を返す傾向があった．評価データ作成時にも各々の質問の対して文書検索により大量の文書が得られ，これらから 20 個以上の回答候補が抽出することが可能であった（文書検索による上位 300 個の文書から平均約 2,231 個の回答候補が得られた）．

表 8 Set1 と Set2 における質問の数と評価データの量：各々の質問は 20 個の回答候補を持っている

Table 8 Statistics on Set1 and Set2: There are 20 answer candidates for each question.

	対応する質問集合	質問の数	質問–回答候補対の数
Set1	QS1 と QS2	500	10,000
Set2	QS3	350	7,000

Set1 と Set2 に含まれている質問とその正しい回答の対の例を**表 9** と**表 10** に示す．ここで，質問に評価表現がない質問–回答対を表 9 に，質問に評価表現がある質問–回答対を表 10 に示す．「意見（評価表現）抽出ツール」を用いて抽出した評価表現は下線で表示し，その意味的極性は肩文字の *positive* と *negative* で示した．そして，回答に現れる質問の内容語は太字で表示した．なお，「意見（評価表現）抽出ツール」によって抽出された評価表現の意味的極性が誤って判定された場合もある．たとえば，表 10 の Q4 の質問にある評価表現「コアラの数が減ってきているのは」にはポジティブな意味的極性が与えられたが，その正しい意味的極性はネガティブと考えられる．

5. 実験

提案手法の回答候補のリランキングには TinySVM[*13] の線形カーネルで学習した SVM を用いた．正しい回答を「1」，正しくない回答を「−1」と設定した学習データを用いて回答候補が正しい回答か否かを分類するための SVM を学習し，各々の質問に対する 20 個の回答候補を SVM 出力値（分類結果とその分類スコア[*14]）で下降順にソートすることにより回答候補のリランキングを行った．評価実験では，以下に述べるように 2 つの異なる設定での 10-fold cross validation を行い，提案手法の有効性を検証した．

- **CV(Set1)**：Set1 における 10-fold cross validation を表す．まず，10,000 個の質問–回答対を持っている Set1 を 1 つの質問に対する回答候補が同一部分に含まれるように 10 等分する．そして，その 9 つの部分（9,000 個の質問–回答対を含む）を回答候補のリランキングの学習データとして，残りの 1 つの部分（1,000 個の質問–回答対を含む）を提案手法の評価のためのテストデータとして用いて 10-fold cross validation を行う．この設定による実験は，実世界の質問とその回答候補を学習データとテストデータとして用いた場合の提案手法の性能を示すためである．
- **CV(Set1)+Set2**：**CV(Set1)** と同様な設定のうえ

[*13] http://chasen.org/~taku/software/TinySVM/
[*14] サンプルと超平面との距離を分類スコアとして用いた．

表 9 評価データにある質問と正しい回答の対：質問に評価表現がない場合
Table 9 Correct question-answer pairs in our test set, where questions have no sentiment phrase.

Q1	台風とか渦巻きの方向は決まってて北半球と南半球で逆だそうですがなぜこのような現象がおきるんですか？
A1	大規模なサイクロンの地上付近の回転**方向**は，**北半球**では反時計回り，**南半球**では時計回りと常に**決まっ**ています．[...] このような回転**方向**は，「コリオリの力」とよばれる，地球上の気体に働く見かけの力によって**決まっ**ています．
Q2	ガムを噛みながらチョコ食べちゃうときがあるのですが，なぜガムが溶けてしまうのでしょうか？
A2	[...] ガムベースは，チクルなどの植物性樹脂やドイツで開発された酢酸ビニル樹脂に，弾力性を出すポリイソブチレンなどを加えて作られる．酢酸ビニル樹脂は脂溶性．一方チョコレートは油脂を含んでいる．一緒に**食べる**ことで，**ガム**の組織がバラバラになり**溶けて**いくということになる^{positive}．[...]

表 10 評価データにある質問と正しい回答の対：質問に評価表現がある場合
Table 10 Correct question-answer pairs in our test set, where questions have at least one sentiment phrase.

Q3	なぜ輸入品の値段が上昇すると インフレ懸念が強まる^{negative} のですか？
A3	[...] コスト・プッシュ・**インフレ**は賃金・原材料費・地代等の コスト**上昇**が生産価格や販売価格に影響を与えて生じる**インフレ**である^{negative}．**輸入品**の高騰や海外**インフレ**が原因となる場合もあります．[...]
Q4	オーストラリアで コアラの数が減ってきているのは^{positive} なぜですか？
A4	[...] 20 世紀になって，**オーストラリア**の**コアラ**の**数**は 300 万匹から 8 万匹にまで減少しています^{negative}．ハンティングや山火事，都市開発などが，このような悲しい結果を招いているのです．[...]

で Set2 を追加の学習データとして用いた 10-fold cross validation を表す．Set1 における設定は **CV(Set1)** と同一であるが，7,000 個の質問–回答対を持つ Set2 を追加学習データとして用いた．つまり，各 fold における分類器の学習データは 16,000 個の質問–回答対（Set1 から 9,000 個，Set2 から 7,000 個）を持つことになり，テストデータが含む質問–回答対の数は **CV(Set1)** と同様に 1,000 個となる．この設定は，Set2 を作成した 1 つの目的である「現実にユーザが発する可能性の低いと考えられる人工的に作られた質問が実世界の質問応答システムの性能改善に寄与するか」を検証するために用いられた．

そして，最上位回答の精度を示すための P@1（Precision at the top-1 answer）と上位 n 個の回答候補における全体的な精度を示すための MAP（Mean Average Precision）で性能評価を行った．

5.1 比較実験

実験では以下の 7 つの手法を比較する．B-QA は提案手法の回答候補検索そのものであり，その他の 6 つの手法は B-QA の結果である上位 20 個の回答候補を各手法が持つ SVM でリランキングするものである．特に B-Rank+Causal-Relation, B-Ranker+WordNet, Proposed(WordNet), Proposed は SVM の学習時に用いられた意味的素性が異るため，これらにおける比較実験により提案手法で利用した意味的素性の有効性を示すことが可能である．

- **B-QA**：提案手法の回答候補検索のみを用いる手法．これは Murata らの手法 [12] の実装したものである．ただ，Murata らの手法では OKAPI を用いて文書検索を行ったが，この実装ではベクトル空間モデルとコサイン類似度に基づいた Solr の文書スコアリング手法を用いて文書検索を行った．
- **B-Ranker**：表 1 の MSA 素性（MSA1〜MSA4）のみで学習した SVM を回答リランキングに用いる手法．
- **B-Ranker+Causal-Relation**：表 1 の MSA 素性（MSA1〜MSA4）と Higashinaka ら [7] の因果関係（Causal-Relation）素性を用いて学習した SVM [9] で回答リランキングを行う手法．用いられた因果関係素性は，1) 回答候補が因果関係のパターンを含んでいるかを表す素性，2) 回答候補が含む因果関係のパターン，そして 3) 質問と回答候補が因果関係の名詞対を持っているかを表す素性（より正確には，質問に結果の名詞，回答候補に原因の名詞を持っているかを表す指示関数）を含む．De Saeger らの手法 [3] を用いて本研究の対象文書から獲得した因果関係の名詞対（上位 100,000 個）とこれらの名詞対の獲得に用いられた 490 個のパターン（10 個以上の名詞対の獲得に用いられたもの）を因果関係素性作成に使用した．なお，この因果関係素性の作成に用いられた因果関係のパターンと因果関係の名詞対は Higashinaka らの手法 [7] と異なる手法によって得られたものである．
- **B-Ranker+WordNet**：表 1 の MSA 素性（MSA1〜MSA4）と Verberne ら [18] の WordNet 素性を用いて学習した SVM で回答リランキングを行う手法．用いられた WordNet 素性は，質問の内容語とこれらの類似語（WordNet synset によるもの）のうち，回答候補に

表 11 各手法による評価結果の比較
Table 11 Comparison of different systems.

手法	CV(Set1)		CV(Set1)+Set2	
	P@1	MAP	P@1	MAP
B-QA	0.222 (0.368)	0.270 (0.447)	0.222 (0.368)	0.270 (0.447)
B-Ranker	0.256 (0.424)	0.319 (0.528)	0.274 (0.454)	0.323 (0.535)
B-Ranker+Causal-Relation	0.262 (0.434)	0.319 (0.528)	0.278 (0.460)	0.325 (0.538)
B-Ranker+WordNet	0.257 (0.425)	0.320 (0.530)	0.275 (0.455)	0.325 (0.538)
Proposed(WordNet)	0.292 (0.483)	0.344 (0.570)	0.312 (0.517)	0.358 (0.593)
Proposed	**0.336 (0.56)**	**0.377 (0.624)**	**0.374 (0.619)**	**0.391 (0.647)**
UpperBound	0.604 (1)	0.604 (1)	0.604 (1)	0.604 (1)

あるものの比率，そして *WordNet::Similarity* [14] を用いて計算された質問と回答候補間の意味的な連関度を含む．*WordNet::Similarity* は単語間の意味的な類似度を計算するものであるため，質問と回答候補間の意味的な連関度は質問の内容語と回答候補の単語間の意味的な類似度を平均して得られる．実験では日本語 WordNet 1.1 [1] を用いて WordNet 素性を作成した．日本語 WordNet のカバレッジが英語 WordNet より低いため[*15]，日本語の WordNet 素性は Verberne ら [18] で用いられた英語の WordNet 素性より回答リランキングへの寄与度が低い可能性がある．

- **Proposed(WordNet)**：提案手法と同様に MSA 素性，SWC 素性，SA 素性のすべての素性を用いて学習した SVM で回答リランキングを行うが，単語クラスの代わりに日本語 WordNet の synset を単語クラスの情報として用いた手法．素性作成に用いられた WordNet の synset は名詞の synset のみに限定した．そして，1 つの単語が複数の synset を持つ場合，各々の synset に対して素性作成を行った（具体的には各々の synset による単語クラス n-gram と単語クラス/極性 n-gram を作成し，SWC 素性と SA 素性のために用いた）．Proposed(WordNet) は，Why 型質問応答において大規模な単語クラスタリングにより得られた単語クラスの有効性を人手で作成した比較的小規模な単語クラスを用いた場合と比較することで示すために用いられる．
- **Proposed**：提案手法．MSA 素性，SWC 素性，SA 素性のすべての素性で学習した SVM で回答リランキングを行う．
- **UpperBound**：回答候補検索の結果である回答候補に l 個の正しい回答がある場合，これらをいつも上位 l 個にランク付けする手法．この実験における回答リランキングの性能の上限値を表す．**表 11** では，この上限値に対する各手法の相対 P@1 値と相対 MAP 値を括弧で示した．

B-Ranker+Causal-Relation と B-Ranker+WordNet に用いられた因果関係素性と WordNet 素性は，提案手法の意味的素性（SWC 素性と SA 素性）と従来の手法 [6], [18] で用いられた意味的素性を比較するためのものであり，提案手法には用いられていないことを承知されたい．

7 つの手法の評価結果（P@1 値と MAP 値）を表 11 に示す．UpperBound を除いて 6 つの手法を比較すると，提案手法が CV(Set1) と CV(Set1)+Set2 の両者で最高性能[*16]を示している．提案手法と B-QA 間の 11.4%～15.2%の性能差（P@1）は回答候補検索に対する性能向上を，提案手法と B-Ranker 間の 8%～10%の性能差（P@1）は本稿で提案した素性の有効性を示している．そして，B-Ranker+Causal-Relation や B-Ranker+WordNet との 7.4%～9.9%の性能差（P@1）は従来手法の意味的素性に対する提案手法の意味的素性の有効性を示している．また，B-Ranker と B-Ranker+Causal-Relation 間の比較，そして B-Ranker と B-Ranker+WordNet 間の比較により，因果関係素性と WordNet 素性がある程度の性能向上に寄与するが，その向上値は P@1 で 0.1%～0.6%にすぎないことが分かった．少なくとも本実験の設定では，因果関係素性と WordNet 素性が有効であるといえない[*17]．なお，Proposed(WordNet) と Proposed 間の 4.4%～6.2%の性能差（P@1）は，大規模な単語クラスタリングにより得られた単語クラスが Why 型質問応答において有効であることを示している．この原因は，自動獲得した単語クラスの対象語（合計 550 万個の名詞）が WordNet の名詞の synset による対象語（約 66,000 個の名詞）と比べて非常に大きいことであると考えている．SVM を回答のランク付けに用いた手法（B-Ranker, B-Ranker+Causal-Relation, B-Ranker+WordNet, Proposed(WordNet), Pro-

[*15] 日本語 WordNet 1.1 では 93,834 個の日本語の単語が 57,238 個の WordNet synset にリンクされているが，英語の WordNet 3.0 は 155,287 個の英語の単語が 117,659 個の WordNet synset にリンクされている．

[*16] 提案手法と他の手法間の性能差（P@1）は McNemar 検定により統計的に有意な差であった（$p < 0.001$）．

[*17] これらをより明らかにするため，B-Ranker+Causal-Relation と B-Ranker+WordNet に用いられた因果関係素性と WordNet 素性，そして提案手法の素性のすべてを用いて学習した SVM を回答リランキングに適用した評価実験を行った．その結果，P@1 が表 11 の Proposed より 0.2%～0.4%下落したことを確認した．

表 12 素性ごとの評価結果
Table 12 The performance of the proposed method with different feature sets.

素性の組合せ	CV(Set1)		CV(Set1)+Set2	
	P@1	MAP	P@1	MAP
Ranker(MSA)	0.256	0.319	0.274	0.323
Ranker(SWC+SA)	0.302	0.324	0.314	0.332
Ranker(MSA+SWC)	0.308	0.349	0.318	0.358
Ranker(MSA+SA)	0.300	0.352	0.314	0.364
Ranker(MSA+SWC+SA@W)	0.312	0.358	0.325	0.365
Ranker(MSA+SWC+SA@P)	0.323	0.369	0.358	0.384
Ranker(MSA+SWC+SA)	**0.336**	**0.377**	**0.374**	**0.391**
UpperBound	0.604	0.604	0.604	0.604

posed）は，CV(Set1) より CV(Set1)+Set2 で高い P@1 値と MAP 値を示している．この結果は，Set2 に含まれている質問のような「現実にユーザが発する可能性の低いと考えられる人工的に作られた質問」が実世界の質問に対する Why 型質問応答システムの性能改善に有効であることを示唆している．

5.2 素性ごとの評価

どのような素性が質問応答の性能改善に寄与したかを明らかにするため，提案手法に用いられた MSA 素性，SWC 素性，SA 素性のうち 1 つの素性を取り除いて SVM を学習し，この SVM で回答候補をランク付けした結果を評価した．そして，SA 素性を SA@W 素性（単語の意味的極性）と SA@P 素性（評価表現とその意味的極性）に分けて，SA@W 素性と SA@P 素性についても同様な評価実験を行った．この評価結果を**表 12** に示す．ここで，本稿で提案した単語クラスと意味的極性に関わる素性の効果を示すため，形態統語的な特徴に関わる素性，MSA のみを用いて SVM を学習した場合（Ranker(MSA)：表 11 の B-Ranker と同一のものである）の評価結果を示した．また，Ranker(MSA+SWC+SA) は提案手法を表す．

意味的極性に関わる素性（SA），もしくは単語クラスに関わる素性（SWC）を取り除いた場合，いずれも性能低下（P@1 で 2.8%～6%の性能低下）があり，これら全部を取り除いた場合（Ranker(MSA)）は P@1 で 8%～10%の性能低下があることを確認した．これらの結果は，意味的極性，大規模な単語クラスタリングにより得られた単語クラスの各々が独立に性能向上に貢献しているが，両者の組合せがより有効であることを示している．また，MSA 素性を取り除いた場合も性能低下（P@1 で 3.4%～6%の性能低下）があることを確認した．以上の結果からいずれのタイプの素性も提案手法による性能向上に貢献していることが分かった．

なお，SA@W 素性を取り除いた場合は P@1 で 1.3%～1.6%の性能低下があるため，SA@W が性能向上に有効であるといえるが，Ranker(MSA+SWC) に SA@W を加えることにより P@1 で 0.4%～0.7%の性能向上しか得られないことが分かった．これは，SA@W と SWC が単語レベルの意味的素性であり，似たような意味的情報を用いるためであると考えられる．たとえば，病名を表す単語クラスの単語はネガティブな意味的極性を持つ場合が多い．このような SA@W と SWC 間の類似性から，Ranker(MSA+SWC) に対する Ranker(MSA+SWC+SA@W) の性能向上が限定的であったといえる．

5.3 回答リランキングのみの評価

さらに，提案手法が理想的な回答候補検索モジュールを持つ場合の性能を推定するための評価実験を行った．この実験では，Set1 を回答候補のランク付け用の SVM を学習するために用い，Set2 を提案手法の性能を評価するためのテストデータとして用いた．本実験の回答候補検索モジュールは，Set2 の質問に対する上位 20 個の回答候補に加えてその質問の作成に用いられたパッセージ（正しい回答）を回答候補として出力する．つまり，各質問に対する 21 個の回答候補はいつも 1 つ以上の正しい回答を持つことになる．このような設定で行われた回答リランキングの評価実験の結果，提案手法は P@1 で 64.8%，MAP で 66.6%を示すことを確認した．これは回答候補に 1 つ以上の正しい回答が含まれていると，提案手法により比較的高い精度の結果が得られるという可能性を示唆している．

6. 考察

回答リランキングの効果，単語クラスの数による影響，質問と回答における評価表現極性の一致による影響，回答結果の数による影響を明らかにするため，各々に対して評価結果を分析した．

6.1 回答リランキングの効果

回答リランキングの効果をより明らかにするため，回答候補検索と提案手法による最上位結果を以下の 4 つの条件で比較分析した．

(1) 提案手法の最上位結果のみが正解
(2) 回答候補検索の最上位結果のみが正解
(3) 両者の最上位結果が正解
(4) 両者の最上位結果が不正解

ここで，条件 (1) と条件 (2) は，回答候補検索に比べ提案手法の回答リランキングにより最上位結果の精度が良くなった場合と悪くなった場合を表す．そして，条件 (1) に該当する結果が条件 (2) に該当するものより多いと，提案手法による回答リランキングが有効であるといえる．**表 13** は CV(Set1)+Set2 において以上の条件を満たす質問の数を示す．条件 (1) を満たす質問の数は 106 個，条件 (2) を満たす質問の数は 30 個であることから，提案手法による回答リランキングの有効性が確認できる．

表 13 提案手法による回答リランキングの効果：条件 (1)〜(4) を満たす質問の数

Table 13 The effect of re-ranking by our proposed method: the number of questions satisfying condition (1)–(4).

(1) 提案手法の最上位結果のみが正解	106
(2) 回答候補検索の最上位結果のみが正解	30
(3) 両者の最上位結果が正解	81
(4) 両者の最上位結果が不正解	283
合計	500

表 14 単語クラスの数による影響

Table 14 The performance of the proposed method with different number of semantic word classes.

単語クラスの数	CV(Set1)		CV(Set1)+Set2	
	P@1	MAP	P@1	MAP
100	0.330	0.372	0.366	0.386
500	**0.336**	**0.377**	**0.374**	**0.391**
2,000	0.334	0.376	0.370	0.386

6.2 単語クラスの数による影響

大規模な単語クラスタリングにより得られた単語クラスの数が提案手法に与える影響を明らかにするため，単語クラスタリングにより作られる単語クラスの数を 100，500，2,000 にした場合の提案手法の評価実験を行った．**表 14** にその結果を示した．単語クラスの数を 500 にした場合が，単語クラスの数を 100 と 2,000 にした場合より高い性能を示す．そして，単語クラスの数 500 と 2,000 の場合と比べ単語クラスの数が 100 の場合がより低い性能を示すが，その差は大きくないことが分かる．少なくとも本実験の設定では，用いられた単語クラスの数によらず提案手法による性能向上が可能であることが分かった．なお，本稿の実験で使われた単語クラスは，6 億件の Web 文書に現れる単語とその係り受け関係を用いて単語クラスタリングした結果であり，各々の単語クラスを得るためには数カ月の計算時間が要する．そのため，より多様な数の単語クラスを用いた検証は困難であった．

6.3 質問と回答における評価表現極性の一致による影響

「ポジティブな（望ましい）事象の理由はポジティブな（望ましい）事象である」，「ネガティブな（望ましくない）事象の理由はネガティブな（望ましくない）事象である」という意味的極性に関わるパターンが提案手法の性能にどのような影響を与えるかを明らかにするため，提案手法の素性のうち，質問と回答における評価表現極性の一致を表している SA@P1（質問の評価表現と回答の評価表現が持つ意味的極性が一致するか否かを表す素性），SA@P2（質問の評価表現と回答の評価表現の意味的極性が一致した際，その意味的極性を表す素性），SA@P10（質問の評価表現と回答の評価表現の意味的極性が一致する場合，質問の評価表現から取り出した単語クラス n-gram と回答の評価表現から取り出した単語クラス n-gram の組合せを表す素性）の各々を取り除いた場合とこれらのすべてを取り除いた場合の評価実験を行った．**表 15** にその結果を示す．

表 15 質問と回答における評価表現極性の一致による影響

Table 15 The performance of the proposed method with/without features related to the agreement between polarities of sentiment expressions in a question and its answer.

極性一致の素性	CV(Set1)		CV(Set1)+Set2	
	P@1	MAP	P@1	MAP
All-SA@P1	0.332	0.372	0.370	0.384
All-SA@P2	0.330	0.371	0.366	0.384
All-SA@P10	0.326	0.370	0.360	0.384
All-{SA@P1,SA@P2,SA@P10}	0.321	0.368	0.352	0.381
All	**0.336**	**0.377**	**0.374**	**0.391**

SA@P1，SA@P2，SA@P10 のすべてを取り除いた場合（表 15 の「All-{SA@P1,SA@P2,SA@P10}」），極性一致の素性を含むすべての素性を用いた場合（表 15 の「All」）より P@1 で 1.5%〜2.5%の性能低下があることを確認した．これらの結果は，質問と回答における評価表現極性の一致という意味的極性に関わるパターンが提案手法による性能向上に貢献していることを示している．

また，SA@P1，SA@P2，SA@P10 のうち，1 つの素性のみを取り除いた場合（「All-SA@P1」，「All-SA@P2」，「All-SA@P10」）は「All」より P@1 で 0.4%〜1.4%の性能低下があることを確認し，各々の素性が提案手法による性能向上にある程度貢献していることが分かった．

6.4 回答候補の数による影響

回答候補の数による影響を調べるため，回答候補の数を 20 から 400 までの範囲で増やし，この条件での提案手法の評価実験を行った（この実験での回答候補の数 n は集合 $\{n|n = k \times 20, 1 \leq k \leq 20\}$ に属するもの）．実験では，Set1 と Set2 の両者を SVM の学習データとして用いて学習し，200 個の新たな Why 型質問に対して提案手法の最上位結果の精度（P@1）を評価した．新たな Why 型質問は，Yahoo!知恵袋から 100 個の質問，人手による 100 個の質問で構成されており，これらは 4 章に述べた QS1 と QS2 の質問の作成方法に従って作られた[*18]．**図 2** はその結果を示している．ここで x 軸は回答候補の数を y 軸は P@1 を示す．回答候補の数が増えると，提案手法による最上位の精度が低下することが分かった．これらに対する我々の仮説は，SVM の学習データに含まれている回答候補が各質問に対する上位 20 個の結果であるため，上位 20 個以降の低順位にある正しくない回答候補における特徴が SVM

[*18] これらの質問を回答候補検索の入力として得られた上位 20 個の回答候補を 3 名のアノテータが判定し，多数決によって最終判定結果を得た．3 名の判定結果は相当な一致率（Fleiss の kappa 値で 0.651）を示した．

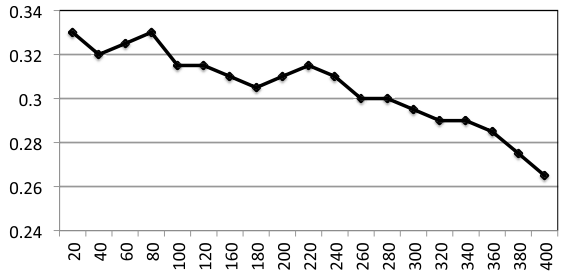

図 2 回答候補の数の変化による提案手法の評価結果
Fig. 2 The performance of our proposed method when changing the number of answer candidates.

によって学習できなかったというものである．この結果，SVM は低順位の正しくない回答候補を正しい回答と分類してしまったと考えられる．この仮説を検証するためには低順位の正しくない回答候補を学習データに追加して回答リランキングを行うなどの実験が必要であるが，これらは今後の課題としたい．

7. 関連研究

non-factoid 型質問応答において，教師あり学習による回答リランキングに用いた先行研究 [7], [16], [18] がある．これらの先行研究は，質問と回答のテキストから抽出した単語の n-gram [7], [16], [18]，構文情報 [16], [18] などの形態統合的な情報と以下に述べる意味的情報と統計情報を用いて作成した素性を回答リランキングのために使用した．Higashinaka ら [7] は，因果関係に関わる意味的情報（自動獲得した因果関係のパターン，因果関係の名詞対，そして人手で作成した因果関係のパターン）とコサイン類似度（cosine similarity）やシソーラスを用いて計算した質問と回答候補間の類似度を素性として用いて分類器を学習し，Why 型質問応答の回答リランキングを行った．Verberne ら [18] は，Wikipedia を対象文書とした Why 型質問応答に回答リランキングのための分類器を用いた．分類器の学習では，回答候補の検索に用いられた TF-IDF などの統計的な情報，そして WordNet の類似語を用いて計算した質問と回答候補間の類似度などの意味的な情報が素性として用いられた．また，Surdeanu ら [16] は，Yahoo! Answers から How-to 型質問とその回答を自動抽出し，これらを学習データとして用いて与えられた How-to 型質問の正しい回答を Yahoo! Answers から検索する手法を提案した．このため，Yahoo! Answer の How-to 型質問とその質問のベストアンサの対を正例とし，How-to 型質問とその質問のベストアンサではない回答の対を負例として作られた学習データを用いて学習した分類器で回答リランキングを行った．分類器学習用の素性作成には，回答候補の検索に用いられた BM25 スコア，質問と回答候補間の類似度，質問と回答候補に含まれている単語間の相関関係などの情報が用いられた．特に，Why 型質問応答の従来研究 [7], [18] において，因果関係と WordNet が意味的素性の作成に用いられており，これらの意味的素性は 5 章で行った評価実験の B-Ranker+Causal-Relation と B-Ranker+WordNet に実装されたものである．

また，機能語集合を用いて Why 型質問応答のための Why テキストセグメントを識別する先行研究 [19] がある．田中ら [19] は，Bag-of-word といったすべての品詞からなる単語集合に加え，「～なので」，「～は～です」などの文法情報が異なる機能語集合で定義される Bag-of-Grammar を機械学習の素性として用いて，因果関係の表現を含む Why テキストセグメントを識別した．

提案手法は，意味的極性と大量の Web 文書から自動獲得した単語の意味的クラスという意味的知識を用いた新たな素性を提案し，Why 型質問応答におけるその有効性を示した点が先行研究と大きく異なる．また，意味的極性のパターンは主観的な意見を尋ねる質問に対して回答する opinion 質問応答のタスク [2], [11], [15] に用いられてきたが，我々の知る限りこうした意味的極性のパターンは non-factoid 型質問応答においてこれまでに検討されたことがない．

8. まとめ

本稿では，「ネガティブな（望ましくない）事象の原因はネガティブな（望ましくない）事象であることが多い」，そして「ポジティブな（望ましい）事象の原因はポジティブな（望ましい）事象であることが多い」という意味的極性に関わるパターンと質問に含まれた単語とその回答に含まれた単語間の意味的な相関関係を意味的知識として用いて Why 型質問に対する質問応答システムの精度を向上させるための手法を提案した．850 個の Why 型質問に対する評価実験の結果，提案手法による 15.2%の性能改善を確認し，Why 型質問応答において意味的極性と大量の Web 文書から自動獲得した単語の意味的クラスという意味的知識の有効性を示した．今後，矛盾の意味的関係や「活性/不活性」の意味的極性 [5] などの新たな意味的知識を Why 型質問応答に適用し，その有効性を検証する予定である．

参考文献

[1] Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T. and Kanzaki, K.: Enhancing the Japanese WordNet, *Proc. 7th Workshop on Asian Language Resources*, pp.1–8 (2009).
[2] Dang, H.T.: Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks, *Proc. TAC 2008* (2008).
[3] De Saeger, S., Torisawa, K., Kazama, J., Kuroda, K. and Murata, M.: Large Scale Relation Acquisition Using Class Dependent Patterns, *Proc. ICDM 2009*, pp.764–769 (2009).
[4] Ferrucci, D.A., Brown, E.W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J.W.,

Nyberg, E., Prager, J.M., Schlaefer, N. and Welty, C.A.: Building Watson: An Overview of the DeepQA Project, *AI Magazine*, Vol.31, No.3, pp.59–79 (2010).
[5] Hashimoto, C., Torisawa, K., De Saeger, S., Oh, J.-H. and Kazama, J.: Excitatory or Inhibitory: A New Semantic Orientation Extracts Contradiction and Causality from the Web, *Proc. EMNLP-CoNLL 2012*, pp.619–630, Association for Computational Linguistics (2012).
[6] Higashinaka, R. and Isozaki, H.: Automatically Acquiring Causal Expression Patterns from Relation-annotated Corpora to Improve Question Answering for why-Questions, *ACM Trans. Asian Language Information Processing*, Vol.7, No.2, pp.1–29 (2008).
[7] Higashinaka, R. and Isozaki, H.: Corpus-based question answering for why-questions, *Proc. IJCNLP 2008*, pp.418–425 (2008).
[8] Hofmann, T.: Probabilistic latent semantic indexing, *Proc. SIGIR 1999*, pp.50–57 (1999).
[9] Joachims, T.: *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*, Kluwer Academic Publishers (2002).
[10] Kazama, J. and Torisawa, K.: Inducing Gazetteers for Named Entity Recognition by Large-Scale Clustering of Dependency Relations, *Proc. ACL-08: HLT*, pp.407–415 (2008).
[11] Li, F., Tang, Y., Huang, M. and Zhu, X.: Answering opinion questions with random walks on graphs, *Proc. ACL-IJCNLP 2009*, pp.737–745 (2009).
[12] Murata, M., Tsukawaki, S., Kanamaru, T., Ma, Q. and Isahara, H.: A System for Answering Non-Factoid Japanese Questions by Using Passage Retrieval Weighted Based on Type of Answer, *Proc. NTCIR-6* (2007).
[13] Nakagawa, T., Inui, K. and Kurohashi, S.: Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables, *Proc. NAACL-HLT 2010*, pp.786–794 (2010).
[14] Pedersen, T., Patwardhan, S. and Michelizzi, J.: WordNet::Similarity: Measuring the relatedness of concepts, *Demonstration Papers at HLT-NAACL 2004*, pp.38–41 (2004).
[15] Stoyanov, V., Cardie, C. and Wiebe, J.: Multi-perspective question answering using the OpQA corpus, *Proc. HLT/EMNLP 2005*, pp.923–930 (2005).
[16] Surdeanu, M., Ciaramita, M. and Zaragoza, H.: Learning to Rank Answers to Non-Factoid Questions from Web Collections, *Computational Linguistics*, Vol.37, No.2, pp.351–383 (2011).
[17] Verberne, S., Boves, L., Oostdijk, N. and Coppen, P.-A.: Evaluating discourse-based answer extraction for *why*-question answering, *Proc. SIGIR 2007*, pp.735–736 (2007).
[18] Verberne, S., Boves, L., Oostdijk, N. and Coppen, P.-A.: What is not in the bag of words for why-QA?, *Computational Linguistics*, Vol.36, pp.229–245 (2010).
[19] 田中克幸，滝口哲也，有木康雄：Bag of Grammar を用いたドメイン依存性の少ない Why テキストセグメント識別器の自動構築法，電子情報通信学会論文誌 D，情報・システム，Vol.94, No.12, pp.2047–2057 (2011).

呉 鍾勲

2000 年韓国科学技術院（KAIST）電子電算学科電算学専攻修士課程修了．2005 年同大学院博士課程修了．韓国科学技術院研究員，独立行政法人情報通信研究機構専攻研究員，同機構研究員を経て，現在，同機構主任研究員．博士（工学）．自然言語処理の研究に従事．言語処理学会，ACL 各会員．

鳥澤 健太郎 （正会員）

1992 年東京大学理学部卒業．1994 年同大学院修士課程修了．1995 年同大学院博士課程中退．同年同大学院助手．1998 年科学技術振興事業団さきがけ研究 21 研究員兼任（2002 年まで）．北陸先端科学技術大学院大学助教授を経て，2008 年より独立行政法人情報通信研究機構言語基盤グループ，グループリーダー．現在，同機構情報分析研究室室長および情報配信基盤研究室室長．博士（理学）．自然言語処理の研究に従事．日本学術振興会賞等受賞．言語処理学会，人工知能学会各会員．

橋本 力 （正会員）

京都大学情報学研究科産学官連携研究員，山形大学大学院理工学研究科助教，独立行政法人情報通信研究機構専攻研究員，同機構研究員を経て，現在，同機構主任研究員．自然言語処理の研究に従事．博士（情報学，言語科学）．言語処理学会，ACL 各会員．情報処理学会論文賞，言語処理学会論文賞，言語処理学会優秀発表賞等受賞．

川田 拓也

2005 年京都大学大学院文学研究科博士前期課程修了．2010 年同大学院博士後期課程修了．現在，独立行政法人情報通信研究機構情報分析研究室研究員．博士（文学）．言語資源の設計と構築に従事．

デサーガ ステイン

2006 年に北陸先端科学技術大学院大学博士課程修了後，2007 年に情報通信研究機構専攻研究員を経て，現在，同機構主任研究員．知識の自動獲得の研究に従事，言語処理学会第 16 回年次大会優秀発表賞等受賞．博士（知識科学）．

風間 淳一

2004 年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程修了．博士（情報理工学）．同年北陸先端科学技術大学院大学情報科学研究科助教．2008 年から 2013 年まで情報通信研究機構．

王 軼謳

2004 年中国大連理工大学情報学修士課程修了，2008 年国立岐阜大学大学院工学研究科博士後期課程電子情報システム工学専攻修了．同年独立行政法人情報通信研究機構に入所．情報分析研究室研究員，現在に至る．博士（工学）．意見分析，機械翻訳，形態素解析の研究に従事．