2007－ICS－147 (22)
2007／3／15

コンピュータ日常会話のためのWebからの時事情報獲得技術

藤田 晴樹† 渡部 広一‡ 河岡 司‡
†同志社大学大学院工学研究科 〒610-0394 京都府京田辺市多々羅都谷1-3
E-mail： †dtf0705@mail4.doshisha.ac.jp, ‡{hwatabe, tkawaoka}@mail.doshisha.ac.jp

現在のコンピュータと人間との会話は，コンピュータが人間の質問に対して応えるというものが多く，コンピュータと人間が日常会話を行っているとはいえない．コンピュータが日常会話を行うためには，人間に対して有益な情報を蓄えた知識ベースが必要となる．

本研究では，無数の文書集合の中から人間に対して有益な時事情報を蓄えた知識ベースを構築し，動的に新しい情報に更新する手法を提案する．提案手法では，話題となっているニュースはWebに頻繁に出現している（数多く掲載されている）という考えを元に，Webに存在するニュースの中から，単語の関連性を考慮して獲得した話題語を用いて，ニュースに重要度を付与し，時事情報知識ベースに格納することを行った．

The acquisition method of current events information from web for computer daily conversation

Haruki FUJITA† Hirokazu WATABE‡ Tsukasa KAWAOKA‡
†Graduate School of Engineering, Doshisha University
1-3 Miyakodani Tatara Kyotanabe-shi, Kyoto, 610-0394, Japan
E-mail： †dtf0705@mail4.doshisha.ac.jp, ‡{hwatabe, tkawaoka}@mail.doshisha.ac.jp

The conversation between a computer and a human being is that a computer answers a human being's questions. And, it is hard to say that a computer and a human being have a daily conversation. A knowledge base is necessary so that a computer have a daily conversation.

In this paper, we propose the method that makes and updates Current Topics Information Knowledge Base that is stored useful information for a human being from much information. The proposed method calculates the degree of article's importance with the topic words that got from articles in Web, and stores articles to Current Topics Information Knowledge Base.

1 はじめに

近年，技術の発達により，コンピュータと人間とが会話を行えるような研究が盛んに行われている．しかし，現在のコンピュータと人間との会話は，コンピュータが人間の質問に対して応えるというものが多く，コンピュータと人間が相互に円滑なコミュニケーションを行っているとは言い難い．これは，コンピュータが人間に対して自発的に有益な情報を提供する機能が備わっていないためであると考えられる．コンピュータが自発的に有益な情報を提供するためには，有益な情報を蓄えた知識ベースが必要となる．

本稿では，無数の文書集合の中から人間に対して有益な情報を蓄えるための時事情報知識ベースを構築し，動的に新しい情報に更新する手法を提案する．人間とコンピュータが円滑なコミュニケーションを行うための，人間に対する有益な情報とは，インターネットなどの文書集合の中から特にニュースに関する情報である．すなわち，ニュースは，時事的な事象に関する話題を的確に取得できる情報源であると考えられる．例えば，「松坂大輔」が話題として上がっていると，「松坂大輔」や「大リーグ」に関連するニュースを時事情報知識ベースに格納する．さらに，時事情報は時間と共に更新される特性があるため，時事情報知識ベースは動的に更新する必要がある．このため，長期間に渡り話題に上っている情報は時事情報知識ベース内に残し，一時的に話題になったような情報を削除することが要求される．

提案手法では，話題となっているニュースはインターネット内に頻繁に出現している（数多く掲載されている）という考えを基に，Webに存在するニュースの中から他案後の関連性を考慮して獲得した話題語を用いて，ニュースに重要度を付与する．そして，重要度を付与したニュースを時事情報知識ベースに格納する．この手法により，人間との会話に利用することができる情報を獲得し，時事情報知識ベースの構築やデータの更新を行い，コンピュータと人間とがより充実した日常会話を行うことができると考えられる．

2 関連技術

2.1 概念ベース

概念ベース [1]とは，複数の国語辞書や新聞等から機械的に構築した，語（概念）とその意味特徴を表す単語（属性）の集合からなる知識ベースである（図1）．概念 A に付与される属性 a_n には，その重要性を表す重み w_n が付与されている．概念ベースには，約9万語の概念が収録されており，1つの概念あたり平均30個の属性が付与されている．しかしながら，概念ベースにも登録されていない概念も存在しており，その概念を本稿では未定義語と定義する．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_n,w_n)\} \quad (2.1)$$

各概念に付与されている属性は，概念ベースに概念として登録されている語であるため，各属性を一つの概念表記としてみなした場合，さらにそれを表す属性を導くことができる．このように，概念は概念ベースによりn次の属性連鎖集合として定義する．また，n次

の属性集合を n 次属性と呼ぶ.

概念	属性，重み
雪	(雪，0.61)，(白い，0.30)，…
白い	(雪，0.16)，(白地，0.14)，…
下る	(低い，0.23)，(雪，0.21)，…
…	…

図1 概念ベース

2.2 関連度計算

関連度計算方式[1]は，概念ベースに定義された語と語の関連の強さを，同義性，類似性のみに関わらず定量化する手法である．また，語と語の類似性評価手法として，シソーラスなどを用いて属性の意味的な圧縮を行った概念ベースを前提に各概念をベクトルと見なし，余弦を用いて定量化するベクトル空間モデルが広く利用されている．この方式は，「赤ちゃんと子ども」や「自動車と車」といった，類似性の高い語と語の類似性評価には適しているが，「赤ちゃんと玩具」や「自動車と事故」といった語と語の関連性は，類似性という観点からは関連性評価が困難であると考えられる．そのため，本稿に使用する概念ベースは，より柔軟に語と語の関連の強さを定量化するために関連度計算方式を前提としている．以下，概念間の一致度，並びに一致度に基づき関連度を求める関連度計算方式について述べる．

2.2.1 一致度

概念 A，B の属性を a_i，b_j，対応する重みを u_i，v_j とし，それぞれ属性が L 個，M 個あるとする．($L \leqq M$)．また，各概念の属性の重みを，その総和が 1.0 となるよう正規化している．

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_L, w_{Ln})\} \tag{2.2}$$

$$B = \{(b_1, v_1), (b_2, v_2), \cdots, (b_M, v_M)\} \tag{2.3}$$

このとき，概念 A と概念 B の属性一致度 $MatchWR(A,B)$ を以下のように定義する．ただし，$a_i = b_j$ は属性同士が一致した場合を示している．すなわち，一致した属性の重みのうち，小さい方の重みの和が一致度となる．また，一致度は 0.0~1.0 の値をとる．

$$MatchWR(A,B) = \sum_{a_i = b_j} \min(u_i, v_j)$$
$$\min(\alpha, \beta) = \begin{cases} \alpha\ (\alpha \leq \beta) \\ \beta\ (\alpha > \beta) \end{cases} \tag{2.4}$$

2.2.2 関連度

概念関連度 MR は，対象となる二つの概念において，一次属性の組み合わせについて一致度を求め，これを基に概念を構成する属性集合全体としての一致量を計算することで算出される．

具体的には，見出し語として一致する属性同士($a_i = b_j$)について，まず優先的に対応を決定する．他の属性については，全ての一次属性の組み合わせにおいて属性一致度を算出し，属性一致度の和が最大となるように組み合わせを決定する．一致度を考慮することにより，属性同士の見出し語としての一致だけではなく，一致度合いの近い属性を有効に対応づけることが可能となる．また，概念 A，B 間の見出し語として一致する属性($a_i = b_j$)については，以下の処理により別扱いとする．$a_i = b_j$ なる属性があった場合，それらの属性の重みを参照し，$u_i > v_j$ となる場合は，a_i の重み u_i を $u_i - v_j$ とし，属性 b_j を概念 B から除外する．逆の場合は，同様に b_j の重み v_j を $v_j - u_i$ とし，属性 b_j を概念 B から除外する．見出し語として一致する属性が T 組あった場合，概念 A，B はそれぞれ A'，B' として以下のように定義し直され，これらの属性間には見出し語として一致する属性は存在しなくなる．

$$A' = \{(a'_1, u'_1), (a'_2, u'_2), \cdots, (a'_{L-T}, u'_{L-T})\} \tag{2.5}$$

$$B' = \{(b'_1, v'_1), (b'_2, v'_2), \cdots, (b'_{M-T}, v'_{M-T})\} \tag{2.6}$$

見出し語として一致した属性の関連度を $MR\_com(A,B)$ とし，以下の式で定義する．

$$MR\_com(A,B) = \sum_{a_i = b_j} \min(u_i, v_j)$$
$$\min(\alpha, \beta) = \begin{cases} \alpha\ (\alpha \leq \beta) \\ \beta\ (\alpha > \beta) \end{cases} \tag{2.7}$$

次に，見出し語として一致する属性を除外した A'，B' の関連度を $MR\_def(A',B')$ とする．$MR\_def(A,B)$ を算出するために，属性数の少ない方の概念 A' の並びを固定し，属性間の属性一致度の和が最大になるように概念 B' の属性を並べ替える．この時，対応にあふれた属性は無視する．概念 A' の属性 a'_i と概念 B' の属性 b'_x が対応したとすると，概念 B' は以下のように並び換えられる．

$$B' = \{(b'_x, v'_x), (b'_{x+1}, v'_{x+1}), \cdots, (b'_{x+L-T}, v'_{x+L-T})\} \tag{2.8}$$

この結果，見出し語として一致する属性を除去した属性間の関連度 $MR\_def(A',B')$ を以下の式によって定義する．

$$MR\_def(A', B') = \sum_{s=1}^{x+L-T} Match(a'_s, b'_s) \times \frac{\min(u'_s, v'_s)}{\max(u'_s, v'_s)} \times \frac{u'_s + v'_s}{2}$$
$$\min(\alpha, \beta) = \begin{cases} a(\alpha \leq \beta) \\ \beta(\alpha > \beta) \end{cases}, \quad \max(\alpha, \beta) = \begin{cases} \alpha(\alpha \geq \beta) \\ \beta(\alpha < \beta) \end{cases} \tag{2.9}$$

このように，見出し語として一致する属性間の関連度 $MR\_com(A,B)$ と，それら以外の属性間の概念関連度 $MR\_def(A',B')$ をそれぞれ算出し，合計を概念 A，B の関連度 $MR(A,B)$ とする．

$$MR(A,B) = MR\_com(A,B) + MR\_def(A', B') \tag{2.10}$$

関連度も，一致度と同様 0.0~1.0 の値をとる．また，実験より属性数が 30 個使用すればよいとの報告がなされているため属性数は 30 個まで使用する．

2.3 TF・IDF

TF・IDF 法[2]とは，語の頻度と網羅性に基づいた重み付け手法である．TF はある文書中 d に出現する索引語 t（文書の内容を表す要素）の頻度 $tf(t,d)$ を表す尺度である．IDF はある索引語が全文書中のどれくらいの文書に出現するかを表す尺度であり，N を検索対象となる文書集合中の全文書数，$df(t)$ を索引語 t が出現する文書数とすると式 2.11 で定義される．

$$idf(t) = \log \frac{N}{df(t)} + 1 \tag{2.11}$$

2.4 Web-IDF

2.3 で説明した IDF は一般的な文書（新聞や書籍など）を用いて索引語の特定性を考慮する手法である．IDF の中でも特に，Web-IDF[3]は Web にある文書のみを用いて索引語の特定性を考慮する手法である．Web-IDF では式 2.3 の N を Google[4]が保有している日本語のページ数（Google は全言語において保有しているページ数は公開されているが，日本語のページとして保有している数は公開されていないため，日本語の文書として最も使われている「は」で検索を行ったヒット件数（416,000,000）を Google が保有している日本語の全ページ数としている），$df(t)$ を索引語 t を Google で検索を行ったときのヒット件数とする．

2.5 未定義語の属性獲得手法

未定義語の属性獲得手法[3]とは，未定義語 X（概念ベースに定義されていない概念）の意味的特徴を表す属性（単語）とその重要性を表す重みの組を Web を用いて自動的に構成する手法である（図 2）．まず，ロボット型検索エンジン[4]を用いて検索を行って獲得したテキスト情報から形態素解析を行い自立語を出現単語とし抽出する．その後，獲得したテキスト情報空間内での出現単語の出現頻度と Web-IDF の算出を行い，TF・Web-IDF 重み付けを行う．重み順に上位から自立語とその重みの対の集合を X の属性とする．この手法を用いて未定義語 X の属性とその重みの組を構成する．未定義語 X の属性は式 2.12 のように構成される．

$$X = \{(x_1, w_1), (x_2, w_2) \cdots, (x_n, w_n)\} \qquad (2.12)$$

本稿では，この未定義語の属性獲得手法をオートフィードバック（Auto Feedback : AF）と呼ぶことにする．

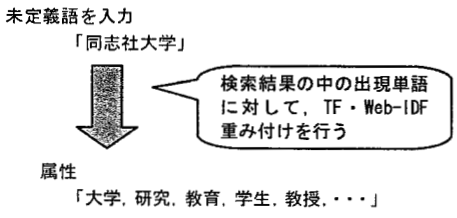

図 2 未定義語の属性獲得手法

3 時事情報獲得技術

本稿では，コンピュータと人間との会話に利用することができる有用な情報を集めた時事情報知識ベース（時事情報 KB）を構築し更新することを目的としている．ここで会話に利用することができる情報として，天気に関する情報と話題情報の 2 つの情報に着目し，この 2 つの情報の獲得を行い時事情報 KB の構築および新しい情報の更新を行う．また，時事情報 KB は，天気に関する情報を格納した天気情報知識ベース（天気情報 KB）と話題に関する情報を格納した話題情報知識ベース（話題情報 KB）の 2 つの知識ベースで構成されている．3.1 に天気に関する情報の獲得について，3.2 に話題に関する情報の獲得について述べる．

3.1 天気に関する情報の獲得

天気に関する情報は，天候，気温，降水確率，その日の天気に対する概要など，毎日同じ情報が更新される．そのため，天気に関する情報については，場所・天気・最高気温・最低気温・降水確率・概要の 6 つのデータの獲得を行い天気情報 KB に格納する（表 1）．

表 1 天気情報 KB

場所	天気	最高	最低	降水確率	概要
京都	晴れ	10	2	0	今日…
大阪	晴れ	11	4	0	今日…

3.2 話題に関する情報の獲得

話題情報とはニュースに記載されているような時事的な情報を指す．また，話題に関する情報は，天気とは異なり，毎日いろいろな内容の記事が更新される．そのため，話題に関する情報は，話題語（注目度が高い情報のキーワード）を用いて記事に重要度付けを行い，話題情報 KB に格納し獲得する．

話題情報 KB の構築・更新方法については，獲得処理と精錬処理の大きく 2 つの処理に分けることができる．獲得処理は，Web から獲得してきたニュースに対して，重要度を求め，人間との会話に利用することができる話題情報はどれであるのかを調べる処理である．精錬処理は，話題情報 KB 内に存在する古いニュースに対して，再び重要度を求め，再び会話に利用される可能性が低いと考えられる情報を削除する処理である．

本稿では，見出しやタイトルなどの単文をニュースとして扱う．そして，話題語 KB を用いてニュースに重要度を付与し，話題情報 KB に格納する．話題語 KB は話題語（会話を行っている時点において話題となっている話題を示すキーワード）を格納した知識ベース（表 2）であり，話題情報 KB はロボットが日常会話を行うために用いる話題情報を格納した知識ベース（表 3）のことである．

表 2 話題語 KB

話題語
松坂大輔，レッドソックス，大リーグ，サッカー，Jリーグ，テニス，…

表 3 話題情報 KB

ニュース
松坂は「日本の宝」　レッドソックス首脳、繰り返す
日本勢、金 50 個　前回より 6 個増える　アジア大会
…

4 獲得処理

獲得処理の流れは，まず Web からニュースの獲得を行う．次に獲得してきたニュース群の中から話題語候補の抽出を行い，話題語候補を用いて話題語の獲得を行う．そして，獲得した話題語を用いてニュースに重要度付けを行う（図 3）．

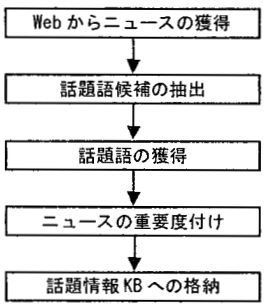

図 3 獲得処理の流れ

4.1 Webからニュースの獲得

ニュースは，Web にあるニュースサイトやブログ，掲示板などから，定期的に獲得を行う．

4.2 話題語候補の抽出

話題語とは，Web から情報を獲得してきた日時に話題となっていた（注目を集めている）物事に関するキーワードのことを指す．話題語は，当日に話題となっているキーワードだけでは，以前から話題となっていたキーワードを的確に考慮することができない．そのため，話題語は 1 日分の話題を表した話題語候補を数日分用いて獲得する．1 日分のニュースの中から獲得するのではなく，連続した数日分のニュースの中から獲得を行う．

話題語候補の抽出の流れは，まず，Web から獲得したニュースの中から話題語候補を抽出する．そして，抽出された話題語候補に対してグループ化を行う（図 4）

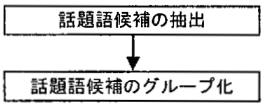

図 4　話題語獲得方法の流れ

4.2.1 話題語候補の抽出

話題語候補の抽出方法としては，まず，Web から獲得してきた 1 日分のニュースに対して形態素解析ソフト「茶筌」[5]を用いて形態素解析を行い，ニュースの文中に含まれている自立語を抽出する（図 5）．このとき，「茶筌」で形態素解析を行うと最小単位で意味を持つ自立語ごとに区切られるので，名詞の連続や，「ソフトバンクの松中」のような「名詞」と「名詞」の間に「の」を挟んだ連続の場合は自立語を接続し話題語候補を抽出する．また，抽出した話題語候補の出現頻度を話題語候補の初期重みとする．また，「ブログ」や「リンク」などといった不要語を取り除くために，Web-IDF を用いて閾値が 3.0 未満（実験により求めた）のものを削除している．

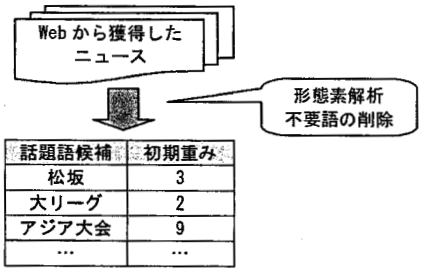

図 5　話題語候補の抽出

4.2.2 話題語候補のグループ化

話題語候補のグループ化とは，抽出した話題語候補に対して，関連がある語同士を集める処理である．このとき，関係がある語同士の初期重みの変調を行い変調重みを求める（図 6）．また，話題語候補同士の関連性は，概念ベースの属性を用いた関連度計算により調べる．

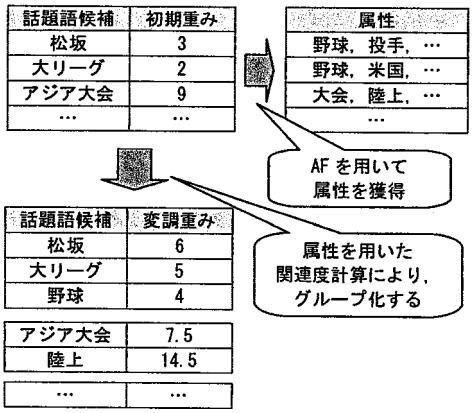

図 6　グループ化

グループ化の処理の流れは，まず，抽出した話題語候補に対して AF を用いて，概念ベース内に存在する属性を獲得する．次に，抽出した話題語候補同士の関連の深さを属性を用いた関連度計算により求める．このとき，関連度の値が 0.015 以上（実験により求めた）なら同じグループとして話題語候補をまとめる．最後に，グループごとに初期重みを調整し精錬重みを求める．精錬重み RW は初期重み IW と同じグループの話題語候補の初期重みの平均 $averageGroupIW$ を足して求める（式 4.1，4.2）．

$$averageGroupIW = \frac{1}{l}\sum_{i=1}^{l} IW_i \qquad (4.1)$$

（l：グループ内の話題語候補の数）

$$RW = IW + averageGroupIW \qquad (4.2)$$

重みを変調した話題語候補をその日の話題語候補として，4.3 の話題語の獲得で用いる．

4.3 話題語の獲得

数日分の話題語候補を合わせて話題語を獲得する（図 7）．話題語の重み $TopicWordWeight$ は，3 日分の話題語候補を用いて，式 4.3 で求めることができる．

$$TopicWordWeight = \sum_{i=1}^{3} RW_i \times e^{(-i+1)} \qquad (4.3)$$

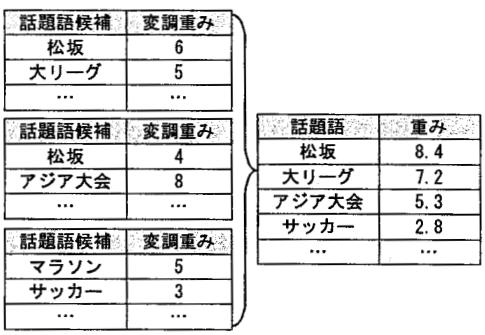

図 7　話題語の獲得

4.4 ニュースの重要度付け

ニュースは，話題語を用いて個別に記事の重要度を付与する（図 8）．重要度はニュースに話題語が存在すれば（表記一致），その話題語の重みを付与する．重要度 *ArticleWeight* は，式 4.4 で求められる．

$$ArticleWeight = \sum_{i=1}^{n} TopicWordWeight_i \qquad (4.4)$$

（n：表記一致した話題語の数）

ニュース

・松坂大輔：レッドソックス入団決定
重要度 = 8.4 + 4.5 = 12.9

・競馬：ディープは１．６倍　ジャパンカップ
重要度 = 2.0

・チーム青森、決勝進出逃す　パシフィック選手権
重要度 = 0.0

話題語	重み
松坂	8.4
大リーグ	7.2
アジア大会	5.3
競馬	2.0
…	…

ニュース中の単語と話題語との表記一致により，重要度を求める

図 8　ニュースの重要度付け

4.5 話題情報知識 KB への格納

重要度を付与したニュースに対して，重要度の閾値を設定し話題情報 KB に格納する．

5 精錬処理

精錬処理は，話題情報 KB 内の過去のニュースの中から不要なニュースを削除する．これは，過去のニュースの中から必要な（話題となっている）情報を獲得すると考えることができる．そこで，4.3 の話題語の獲得で，獲得処理よりも長い期間の話題語候補を用いて獲得すれば，獲得処理と同じ方法で精錬処理を行うことができると考えられる．

6 評価方法

獲得処理を行い話題情報 KB に格納したニュースが，どれくらい当日の話題と一致するかの評価を行う．

今回は，12 月 26 日の朝日新聞[6]と毎日新聞[7]，読売新聞[8]Web サイトのニュースに対して獲得処理を行った．また，ニュースは，5 つのカテゴリ（社会，経済，政治，国際，スポーツ）に分類して，獲得処理を行った．

評価方法は，読売新聞が話題であると考えて提供しているニュースを正解（基準）として，朝日新聞と毎日新聞から獲得したニュースに対して，獲得処理を行い，話題情報 KB に格納したニュースの中にどれくらい正解が存在しているか（精度）と，話題情報 KB の中にどれくらい全体のニュースの中にある正解を再現できたか（再現率）を目視により○，×で判断をする．（図 9）

表 4 に獲得処理を行ったニュースの例を，表 5 に正解の例を示す．

獲得処理を行ったニュース
・朝日新聞と毎日新聞の全てのニュース
正解（話題）
・読売新聞が話題として提供しているニュース

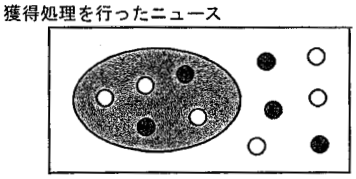

図 9　評価方法

表 4　獲得処理を行ったニュースの例

カテゴリ	ニュース
社会	母親への殺人未遂容疑で４１歳女性を逮捕 小学校長がセクハラ行為　愛知県教委が…
経済	日興人事：新社長に桑島氏　有村社長と… 東京株式市場・前場＝小反落、日経平均は…
政治	沖縄県が「頭越し合意」と反発、政府が謝罪 安倍首相「高収益企業は家計に分配を」
国際	６カ国協議：「核施設の廃棄、北朝鮮が示唆」 北朝鮮：外務次官、金融制裁関連の米朝協…
スポーツ	ＮＦＬ：ペイトリオッツが東地区優勝 西武：和田、２度目の交渉も保留

表 5　正解例

カテゴリ	正解
社会	ネベス容疑者、犯行のひも事前入手か…焼… 大阪府の裏金問題再調査、新たに１６８０…
経済	東京円、４７銭円安の１ドル＝１１８円８… 日興の会長・社長が引責辞任、新社長に桑…
政治	普天間移設、２５日に安倍政権発足後初の… 政府税調の後任会長、伊藤元重氏ら軸に調…
国際	次回６か国協議、金融制裁解除が先決…北… エチオピア、ソマリアのイスラム原理主義…
スポーツ	ペイトリオッツ、ＡＦＣ東地区で４季連続… ディープが北海道へ出発、２６日朝に到着…

7 評価結果

5 つのカテゴリ全体での精度と再現率を図 10 に，各カテゴリの精度と再現率を図 11 に示す．また，重要度を付与したニュースを表 6 に示す．

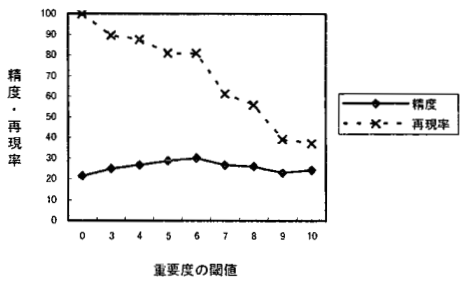

図 10　全体での精度と再現率

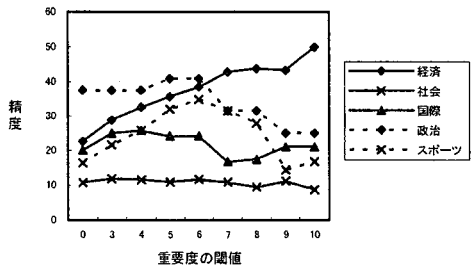

図 11　各カテゴリの精度

表 6　重要度を付与したニュース
（社会：重要度上位 10 件）

重要度	ニュース
18.75	元日本冷蔵社長の浅原英夫さん死去
17.95	強制わいせつ：財務省職員を逮捕、電車で…
17.79	電車内で女性にさわった容疑で財務省職…
16.38	ワールド創業者の木口衛さん死去
15.38	元公明党衆院議員の大野潔さん死去
14.48	小学校長がセクハラ行為　愛知県教委が…
14.10	質店強殺事件：遺族が情報に２００万円懸…
13.95	セクハラ行為：小学校長を停職６カ月懲戒…
13.76	父親逮捕：１歳二男殴られ重体　千葉
13.64	４人死刑執行：日弁連、アムネスティ日本…

8　考察

図 10 の結果より，話題情報 KB に格納されたニュースは，重要度の閾値を 6 とした時が，精度が最も高い結果（30.0%）を得ることができ，再現率も高い結果を（81.0%）を得ることができた．これは，ニュースを全て獲得したときの結果（精度 21.5%）と比べると 8.5%の精度を高くすることができ，話題情報 KB 内に人間との会話に利用することができる情報をより高い精度で格納することができたと考えられる．

図 11 の結果を見ると，カテゴリ「社会」，「国際」，「政治」，「スポーツ」の精度は，重要度の閾値が 6 を境に下がっているが，カテゴリが「経済」の精度は，重要度の閾値が上がるたびに，精度も上がっていた．これは，「経済」では，正解の中に為替に関する情報があり，また，獲得処理を行ったニュースの中に，為替に関する情報が数多く存在していたため，重要度が高く付与され，話題情報 KB に格納されたためだと考えられる．また，カテゴリが「社会」の精度が他のカテゴリの精度と比較すると低い値になっていた（10%程度）．これは，「社会」の中には訃報に関するニュースが，連日数多く出現していたため，そういった話題にならないニュースの重要度が高くなってしまい，精度が低くなったと考えられる．

全体的に精度を低くした原因としては，話題語候補のグループ化をした際，グループの中に不要な語が入ったり，本来は 1 つのグループになって欲しかったが複数のグループに分かれたり，本来は複数のグループであるものが 1 つのグループにまとまってしまったといった失敗が考えられる（図 12）．

・グループの中に不要な語が入った

安倍首相，政府，自民党，談合

・2 つのグループに分かれた

沖縄，沖縄県知事

普天間協議

・複数のグループが 1 つにまとまった

サッカー，浦和レッズ

NFL，ペイトリオッツ

野球，阪神，藤川

図 12　グループ化の失敗例

謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行ったものである．

9　おわりに

本稿では，コンピュータと人間が充実した日常会話を行うことができるようになるために，Web から時事情報を獲得し時事情報 KB を構築・更新することを提案した．具体的には，時事情報として，天気に関する情報と話題に関する情報を獲得し時事情報 KB の構築・更新を行った．また，話題に関する情報は，Web に存在するニュースの中から，単語の関連性を考慮して獲得した話題語を用いて，ニュースに重要度を付与する方法で獲得処理と精錬処理を提案した．この手法を用いて，構築・更新された時事情報 KB を用いることで，コンピュータと人間がより充実した日常会話を行うことができると考えられる．

参考文献

1) 渡部広一，河岡司，“常識判断のための概念間の関連度評価モデル”，自然言語処理，Vol.8，No.2，pp.39-54，2001
2) 徳永健伸，“言語処理と計算 5 情報検索と言語処理”，東京大学出版会，1999
3) 辻泰希，渡部広一，河岡司，“www を用いた概念ベースにない新概念およびその属性獲得手法”，人工知能学会全国大会，2D1-01，2003
4) http://www.google.co.jp/
5) 松本裕治，北内啓，山下達雄，平野善隆，今一修，今村友明，“日本語形態素解析システム『茶筅』version1.0 使用説明書”，NAIST TechnicalReport，NAIST-IS-TR97007，1997
6) http://www.asahi.com/
7) http://www.mainichi-msn.co.jp/
8) http://www.yomiuri.co.jp