情 報 学 基 礎 42－5
（1996. 7. 25）

類似語検索における観点の自動生成法

笠原 要　　松澤 和光
{kaname, matsu}@nttkb.ntt.jp
NTT(株) コミュニケーション科学研究所
〒238-03 横須賀市武 1-2356

情報検索の手法として、検索のキーワードを意味の類似した単語(類似語)で展開する方法がある。類似語はユーザの意図、すなわち、検索の観点によって変化する。そこで、検索の観点に応じた類似語の検索手法を既に提案したが、観点はユーザが人手で与える必要があった。本稿では、上記検索手法のための観点を、キーワードが含まれる文章から自動生成する方式を提案する。これは、キーワードの近傍の単語の集合から「連想」される単語を観点とする。連想の手段としては、辞書中の見出し語と説明語の関係より連想語に関する知識ベース(連想ベース)を構成し、利用した。実際に4万の日常語について構築した連想ベースを用い、新聞記事から観点を生成する実験を行なって、本方式の有効性について考察した。

A Method for Generating Viewpoints of Similar Words Retrieval

Kaname KASAHARA　Kazumitsu MATSUZAWA
NTT Communication Science Laboratories
1-2356 Take Yokosuka-shi Kanagawa 238-03 Japan

On information retrieval, there is a method of adding similar keywords to source keywords in order to improve the efficiency of the retrieval. Similar words change according to the viewpoint of retrieval. So, we propesed a method for retrieving similar words from a viewpoint word which is given by a user. In this paper we propose a method for generating the viewpoint word by using sentences which include the keyword. The point of this method is that the viewpoint word is associated from words near the keyword in the sentences. Associative knowledge base is constructed from machine-readable dictionaries.

1 はじめに

大規模なマルチメディア・データベースにおいては、データの検索が必須の問題である。データの印象や特徴を表すキーワードをデータに付与し、ユーザが指定した検索語に基づく検索が行なわれている [1]。その中でも検索語を類似語で展開する方法が一般的である。ここで重視すべき点は、単語間の類似関係は固定的なものではなく、データベース検索を行なうユーザの意図によって変化することである。

例えば、検索語「馬」に対して、類語辞典(シソーラス)で「馬」と同じ分類「家畜」に含まれる「豚」や「鶏」を類似語と見なして、これらのキーワードを含むデータを検索する場合を考える。これらの類似語は、ユーザが動物に関するデータが欲しい場合には、適切な類似語であるが、乗り物に関するデータを検索したい場合には不適切であり、類似語「自動車」や「汽車」で展開する必要がある。検索語の展開に用いられているシソーラスは、単語を意味に基づいて分類している知識ベースであるが、単語の分類が固定的であるため、ユーザの多様に変動する意図に沿った類似語を与えることができない。

そこで我々は、意図・状況・文脈を指定する単語(観点)が与えられた時に、単語間の類似度を計算する手法を提案した [2]。また、この手法を用い、検索の観点に応じた類似語を獲得する方式を提案した [3]。しかし、この類似語検索方式は、検索の観点を人手で決定して入力する必要がある。そこで本稿では、類似検索のための観点を自動生成する方法を提案する。

以下、2章では、先に提案した観点に基づく類似語検索手法について簡単に説明する。3章では、観点生成方法について提案する。そして4章では、新聞記事を用いた観点生成の予備実験の結果とその考察について述べる。

2 観点に応じた類似検索方式

2.1 概念ベース

先に提案した類似検索方式では、言葉の意味、すなわち概念についての知識ベースである概念ベースを辞書から自動構築して利用する。概念ベースを用いた観点に基づく類似語検索方式について述べる。

概念ベースでは、人手によらない自動構築を行なうために、単語の特徴を表す属性とその重みの対のリストによる概念の単純な表現を採用している。

$$Word_i = \{(p_{i1}, q_{i1}), \cdots, (p_{in}, q_{in})\}. \quad (1)$$

概念の属性は、辞書の語義文より獲得する。また、語義文中で出現頻度の高い属性は、その単語の概念で特徴的であると仮定し、属性の出現頻度を重みから定義した。現在、日常語4万語の概念知識を4種類の辞書の参照により獲得している。その一例として概念「馬」の内容を表1に示す。

表 1: 概念「馬」(一部)

属性名	重み	属性名	重み
馬	125	餓	40
駒	95	長い	39
蹄	73	科	38
競馬	64	下馬	36
家畜	57	足	36
乗馬	52	轡	35
牛	51	双六	34
奇	50	下る	34
博労	49	馬術	34
同類	46	草食	33
走る	44	鬣	31
脚立	44	駄馬	30
将棋	42	竜	29
桂馬	41	馬耳東風	29
踏み台	41	馬上	29

2.2 概念の類似性判別法

前記の概念ベースを用い、観点に応じた概念の類似性判別を実現した。2つの単語の意味の近さを表す類似度 S は、概念ベース中の2つの単語の概念 $Word_1$、$Word_2$ と、判別の観点となる概念(「観点」と呼ぶ)$View$ より、次の4つの手順に従って計算する(図1)。(詳しくは文献 [2] 参照)

- **STEP 1: 属性のシソーラス圧縮**
  類似度の計算は同じ属性同士の重みの比較に基づくため、表記が異なるが意味の似た属性は等しく扱う必要がある。そこで、属性をシソーラスのカテゴリーに変換する。

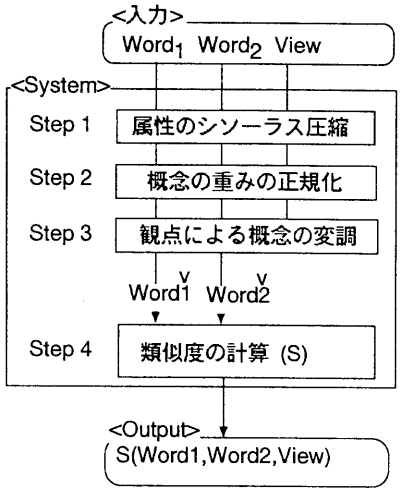

図 1: 類似性判別方式の全体図

- **STEP 2: 重みの正規化**
  概念毎に、重みをノルム1になるように正規化を行なう。
- **STEP 3: 観点に応じた重みの変調**
  観点に応じた類似度の計算とは、観点に含まれる特徴的な属性を重視して行なうものと仮定する。つまり、比較する概念中で観点と共通する属性の重みを大きくする。
- **STEP 4: 類似度の計算 (S)**
  情報検索で一般に行なわれる、ベクトル空間での検索手法の考え [4] を用い、類似度 S を、属性空間上の2つの変調概念がなすベクトルの余弦から計算する。

以上の方式を用いて類似性判別を行なった例を表2に示す。「馬」に対して「豚」と「自動車」のどちらが類似しているかを判別する際に、観点「動物」では、「豚」の方が類似していることを示すが、観点「乗る」では、反対に「自動車」の方が類似していることを示している。これは、人間の感覚に合った判別結果と言える。

表 2: 類似性判別の実例

観点	S('馬', '豚', V)	S('馬', '自動車', V)
動物	0.84	0.29
乗る	0.23	0.60

2.3 観点に基づく類似語検索方式

前記の類似性判別法を用いた、観点を考慮した類似語検索方式について説明する。基本的な考えとしては、観点に関わる属性を持つ概念を類似語と見なし、それらと検索語の類似度に基づいて検索するものである。以下では「赤い」を観点とした「林檎」の類似語検索を例に取って説明する(図2参照)。

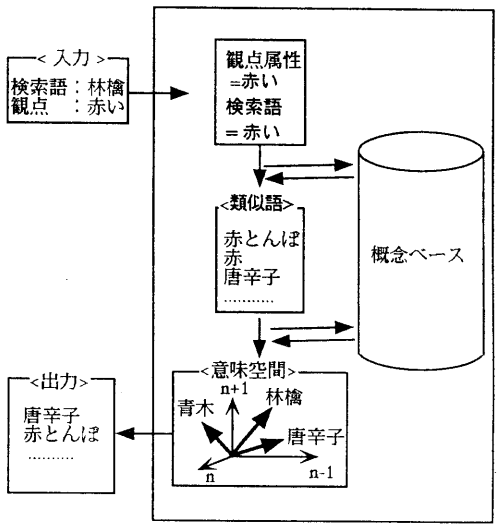

図 2: 類似語検索方式の摸式図

- **観点属性の決定**
  観点属性とは、「類似語検索において重視すべき概念の属性」のことで、通常は観点として指定された単語をそのまま用いるか、あるいは、その同義語を含めて用いる。例えば、「赤い」に対して「赤色」、「赤」等を観点属性とする。
- **類似語候補の選択**
  概念ベース中で属性に観点属性が含まれている概念の単語を類似語の候補と見なす。例えば、「赤い」を観点とした「林檎」の類似語検索で、生物学的に同じ分類に含まれる「梨」であっても、「赤い」属性が含まれないので類似語とせず、「赤とんぼ」のように共通な属性がほとんど含まれない場合でも「赤い」という点では共通するので、類似語とする。
- **類似語の検索**
  上記の候補について検索語と候補の類似度を計

算し、値が大きな順に検索結果として出力する。

様々な観点における「馬」の類似語を上記手法を用いて検索した結果を図3に示す。

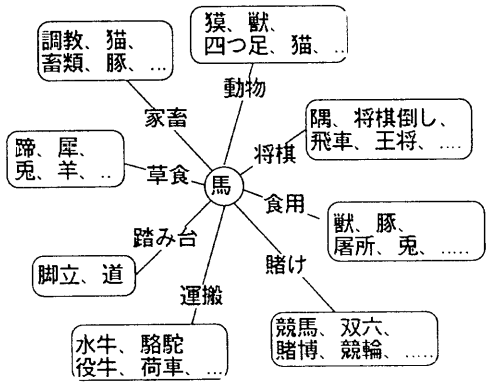

図 3: 「馬」の類似語検索結果

3 観点生成法の提案

前章で述べた類似語検索手法では、ユーザが適切な観点を与えれば適切な類似語を検索することができるが、検索を依頼するユーザ自身でも観点を特定することは難しい場合がある。その為、観点を自動的に生成することが必要となる。

その方法としては一般に、検索語の辞書における語義(sense)を検索語の含まれる文章等から特定することが行なわれる(例えば[5])。しかし辞書では、厳密に異なる意味を語義として分類している場合が多く、語義のどれかを観点とすることができない場合がある。例えば、先に挙げた「馬」の類似語検索結果(図3)では、観点「動物」、「踏み台」、「将棋」は一般的な辞書の語義と対応している。しかし、それ以外は辞書の語義と一意に対応はしていない。また、「運搬」や「家畜」といった観点は、検索語近辺の文章の主題と対応していると考えられるが、語義の曖昧性解消のレベルでは対応できない。

そこで本章はまず、人間が観点を生成するモデルを想定し、そのモデルの中心となる「連想」の実現方法について説明する。最後に、実際に観点を生成するアルゴリズムについて提案する。

3.1 基本的な考え方

観点は、単語が用いられる文脈や状況を特定する単語と規定する。観点を複数の単語の集合やそのネットワーク表現として表現することも考えられるし、さらに文章を観点と見なすことも可能であるが、その場合、その観点が適切であるかの評価が困難であり、利用しにくくなる。そこで、観点が単語であるという前提で以降の説明を行なう。

まず、検索語が含まれる文章から観点を人間が生成するモデルを図4の様に想定する。まず、文章中の検索語の近傍の文を眺め、検索語に関連が深いと考えられる単語に関する情報、すなわち状況を取得する。そして、その状況を構成する単語(状況語)に関連する単語を過去の記憶に基づいて想起(連想)し、連想される単語の内で尤もらしいものを観点として挙げる。

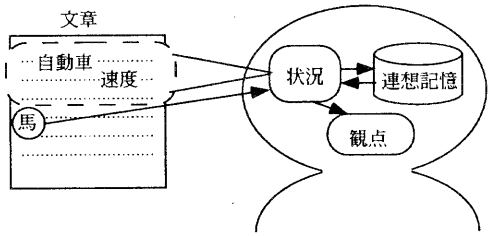

図 4: 観点生成のモデル

検索語が用いられる状況は、検索語の近傍のみではなく文章全体から構成される場合があるが、ここでは、近傍の単語の影響が大きいと想定し、近傍の単語から状況を近似的に表現をする。また、状況とは、単語以外に、単語同士の文法的・意味的結び付きや文同士の結び付き等により構成されるが、ここでは、状況語と、状況語と検索語の距離情報に限定する。そして、状況語から単語を連想する際に、単語の集合から総合して連想する「非線形」な連想が有り得るが、ここでは、個々の単語について連想を行なうものとする。

図4のモデルに従った観点生成法の方式を提案する(図5)。まず、辞書から構成した「連想」に関する知識ベースである連想ベース(連想ベース)を構築する。これは、辞書中の説明語と見出し語の関係から重みつきの連想語を与えるものである。次に、検索語の近傍の単語とその位置情報を

取得し状況語とその重みとする。そして、状況語のそれぞれについて連想ベースを参照して連想語と重みを獲得し、状況語の重みからもっとも良く連想される単語を選択し観点とするものである。以下では、連想を実現する機構と、それを用いた観点生成のアルゴリズムについて説明する。

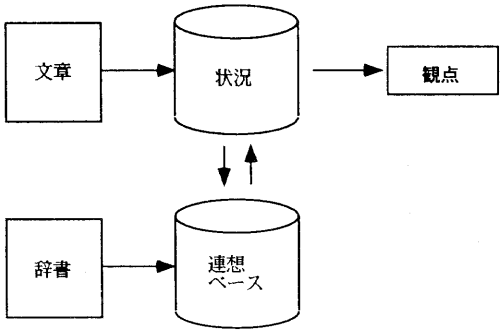

図 5: 観点生成方式

3.2 連想ベース

人間の連想機能に関しては、人間の記憶構造に基づくニューラルネットの研究(例えば[6])や、認知心理学におけるメタファーやアナロジーの研究(例えば[7])などがある。しかし、これらの研究では、連想記憶をどのように獲得するかについては触れられていない。また、情報検索において、ユーザーの意図をネットワークに表現する連想メカニズムの研究[8]があるが、我々の目指す類似語検索では、日常語を対象としており、日常語に関する連想知識をどのように獲得するかは、上記研究では取り上げられていない。そこで、辞書を用いて連想知識の獲得を試みる。

国語辞典や百科辞典等の一般的な辞書は、見出し語とその見出し語の説明文から構成されている。説明語中の自立語(説明語)は、見出し語の概念の特徴であり、この考えに基づいて概念ベースを構築し、その有効性をこれまでに示してきた。ここで辞書は、個々の見出し語について考えられる全ての説明がなされている訳ではなく、その代表的な説明のみが記述されている。従って、説明語は、概念の代表的な特徴と考えられ、個々の説明語に対し見出し語が強く想起されることが予想される(図 6)。

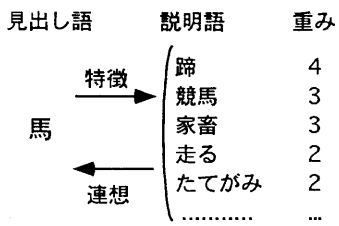

図 6: 見出し語と説明語の関係

そこで、連想に関する知識を表現し、辞書からの連想知識の獲得法について説明する。

まず、単語 $Word_i$ に対する連想知識 $Asc(Word_i)$ を連想語とその重みにより表現する。

$$Asc(Word_i) = \{(Word_{i1}, w_{i1}), \cdots, (Word_{im}, w_{im})\} \quad (2)$$

$Word_{im}$ は、$Word_i$ から連想される単語(連想語)であり、w_{im} は、連想語 $Word_im$ の重みである。重みは、2 乗和を 1 とするように正規化されている。連想知識の獲得には、先に述べた通り、辞書の見出しごとに説明文中から説明語を抽出して出現頻度を求め、説明語に対する見出し語を連想語、出現頻度を重みとし連想ベースを構築する。なお、この様に獲得された連想ベースを用い、3 つのヒント語から連想する単語を求め桝目を埋めるクロスワードパズルに対して、約 8 割の桝目を埋める自動解答に成功しているおり[9]、こうして構築した連想ベースが人間の本質に連想機能を実現していると考えている。

3.3 観点生成アルゴリズム

上記の連想ベースを用いた観点生成法について図 7に従って説明する。図7の様な天気予報に関する文章において、注目語「前線」の観点の生成を行なう。「前線」には、戦争における「前線」と天候における「前線」の語義レベルでの曖昧性があり、天候に関する単語を生成することにより、多義の曖昧性を解消することが可能となる。

- STEP 1: **状況の生成**
  最初に、文章を形態素解析し、意味の最小単位と考えられる自立語を抽出する。そして、注目語よりも前に現われる一定数の単語を状況を形成する単語(状況語)と見なし抽出する。状況

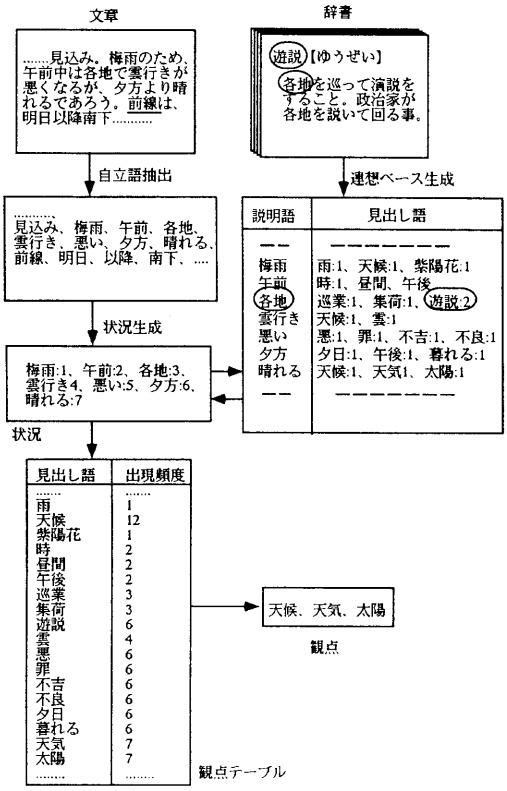

図 7: 観点生成例

語として、注目語を含む文中の単語や、注目語を含む段落中の単語を考えることができる。また、文章中で状況が大きく変化する単位である場面 [10] 中の単語を状況語とすることもできる。図 7では、注目語の前の 7 語までの自立語を状況語としている。

そして、状況語に対して状況への重みを与える。注目語に近い状況語程状況への寄与が高くなると考え、状況語の並び順に対し単調増加な関数の値を重みとする。ここでは、最も遠い状況語の重みを 1 とし、順番に応じて 1 ずつ加算した値を状況語への重みとしている。

このようにして、注目語 W_c の用いられる状況 S を以下のように定義することができる。

$$S(W_c) = \{(W_1, wt(W_c, W_1)), (W_2, wt(W_c, W_2)) , \cdots (W_1, wt(W_c, W_n))\} \quad (3)$$

$W_1, \ldots W_n$ は、注目語 W_c に対する状況語、$wt(W_c, W_n)$ は、注目語 W_c と状況語 W_n の位置関係より定まる重みを表す。

- STEP 2: **連想ベース**
  前章で述べた連想ベースを観点生成に先だって辞書より構築する。例えば、「各地」から連想される単語とその重みは、巡業:1, 集荷:1, 遊説:2 となる。
- STEP 3: **観点生成**
  状況を構成する個々の状況語に対して連想ベース参照し、連想語とその重みを獲得する。そして、連想語に対して連想語の重みと状況語の重みの積を加算し、図 7の様な観点テーブルを生成する。重み閾値をあらかじめ定めておき、全ての状況語の参照後、最も大きな重みを持つ観点テーブル中の連想語を観点とすることもできる。

以上の 3 つのステップにより、注目語に対する観点が生成される。

4 実験と考察

実験で用いた連想ベースは、概念ベース [2] と同じように辞書から自動的に構築した。見出し語は日常語約 4 万あり、連想語数は、総数で約 5 万 5 千語、平均して 16 語ある。表 3に「馬」の連想語とその重みを示す。

表 3: 「馬」の連想知識 (一部)

連想語	重み	連想語	重み
博労	11	馬糞	5
馬	11	白眉	5
乗馬	10	拍車	5
駒	10	種馬	5
轡	10	竹馬	5
餞	9	じゃじゃ馬	5
下馬	9	競馬	5
駄馬	8	騎馬	5
手綱	8	騎手	5
馬上	7	駆ける	5
駅	7	愛馬	5
駄	6	土竜	4
絵馬	6	奔馬	4
馬乗り	6	風馬牛	4
名馬	5	馬丁	4

形態素解析は、当研究所にて開発された日英翻訳システムALT-J/E[11]の形態素解析パッケージを用いた。評価用として、1993年の日経新聞の記事600件[†1]について観点を生成するシステムをCとPerl、tcl-tkを用いて作成した(図8)。

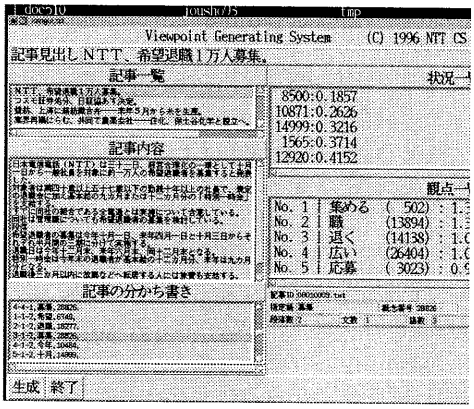

図8: 観点生成デモシステム

以下では、下記の新聞記事について行なった実験結果である。

ＮＴＴ、希望退職１万人募集

日本電信電話（ＮＴＴ）は三十一日、経営合理化の一環として十月一日から一般社員を対象に約一万人の希望退職者を募集すると発表した。対象者は満四十歳以上五十七歳以下の勤続十年以上の社員で、規定の退職金に加え基本給の九カ月または十二カ月分の「特別一時金」を支給する。すでに同社の組合である全電通とは実施について合意している。同社は管理職についても希望退職者の募集を検討している。［日経新聞93年9月1日］

最初に、観点生成の効果を調べるために、先頭から注目語の直前までの全ての単語を状況語として観点生成を行なった。表4は文章中の注目語と生成された観点の対応を示している。

先頭から10語目あたりまでの注目語では、注目語の変化に応じて適切な観点が現われている

[†1]株式会社日本経済新聞の協力によって、社団法人 情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワー キンググループが、１９９３年９月１日から１２月３１日の日本経済新聞記事を基に構築した情報検索評価用データベース（テスト版）を利用

表4: 実験結果

No	注目語	観点	No	注目語	観点
1	経営	-	16	九	会社
2	合理化	営み	17	特別	数
3	十月	組織	18	一時金	数
4	一般	月	19	支給	数
5	社員	月	20	組合	数
6	対象	会社	21	全	数
7	希望	会社	22	実施	数
8	退職	会社	23	合意	数
9	募集	会社	24	管理職	数
10	発表	集める	25	希望	数
11	勤続	広い	26	退職	数
12	十	広い	27	募集	職
13	社員	広い	28	検討	職
14	規定	会社	-	-	職
15	基本	会社			

(例えば、「希望」に対する「会社」)。しかし、新聞記事には数字に関する記述が多いので、17語目の観点から種々の数字を一般化した「数」が現われている。数字に限らずに、文章中には状況の生成と関わらず一定の出現頻度が多くあり、状況語の単語数を増やした時に、それらが、状況語の中心となる恐れがあることがわかる。

次に、9番目に現われる「募集」に対して単語数を変化させ、その効果を調べた。表5にその結果を示す。それぞれの単語数の状況語において観点

表5: 単語数の効果(注目語「募集」)

窓	1st	2nd	3rd
1	退く(0.44)	職(0.42)	現職(0.28)
2	退く(0.44)	職(0.42)	現職(0.42)
3	退く(0.44)	職(0.42)	見通し(0.28)
4	退く(0.44)	職(0.42)	見通し(0.30)
5	退く(0.44)	職(0.42)	会社(0.36)
6	退く(0.44)	職(0.42)	会社(0.40)
7	退く(0.44)	会社(0.43)	職(0.42)
8	会社(0.45)	退く(0.44)	職(0.43)

を生成し、重みの順に上位3語を表示している。状況語を構成する単語数が少ない時には、「募集」の直前の「退職」から連想される「退く」、「職」が優勢であるが、単語数を増やすに連れて

集」の直前の「退職」から連想される「退く」、「職」が優勢であるが、単語数を増やすに連れてて会社に関連する単語「社員」、「経営」、「合理化」が状況語として含まれるために、大きな状況から見た観点「会社」を獲得することが出来ている。この様に状況語を構成する単語数が観点の生成に大きく関わることが明らかとなった。

5 おわりに

本稿では、観点に基づく類似語検索において、検索語を含む文章から観点を抽出する方式を提案した。観点を生成する際に、文章より注目する単語が用いられる状況を生成し、その状況に対して連想ベースを用いて連想し、連想結果を総合して観点を獲得するものである。日常語4万語に関する連想ベースを自動構築し、新聞記事を対象とした予備実験を行なった。そして、提案方式のパラメータに関する検討を行ない、状況語の範囲が観点生成に影響することを明らかにした。今後は、より定量的な評価を行ない、提案方式の評価及び方式改善を行ない、観点に基づく類似判別を含めた総合的な評価を行なう予定である。

参考文献

[1] 端山貴也, 清木康: 分散型マルチメディア・データベースシステムの編集・統合系の実現と評価, 情報処理学会全国大会（第50回）, Vol. 4(3G-3), pp. 101–102 (1995).

[2] 笠原要, 松澤和光, 石川勉, 河岡司: 観点に基づく概念間の類似性判別, 情報処理学会論文誌, Vol. 35, No. 3, pp. 505–509 (1994).

[3] 笠原要, 松澤和光: 概念ベースを用いた常識語の類似検索, 信学技報, Vol. AI95-25, pp. 23–30 (1995).

[4] Salton, G. and McGill, M.: *Introduction to modern information retrieval*, McGraw-Hill (1983).

[5] Schutze, H.: Dimensions of Meaning, *Proceedings of Supercomputing 92*, pp. 787–796 (1992).

[6] Nakano, K.: Associatron: A Model of Associative Memory, *IEEE Trans. Syst., Man, Cybern*, Vol. SMC-2, pp. 380–388 (1972).

[7] 楠見孝: 類似性に基づく事例検索の認知的分析. アナロジー、メタファ、デジャビュ, 人工知能学会研究会資料, Vol. SIG-KBS-9502-3, pp. 17–23 (1995).

[8] 巌寺俊哲, 木本晴夫: 動的シソーラスを用いた連想検索〜リンク重みの導入〜, 第44回情全大, Vol. 4(3G-10), pp. 105–106 (1992).

[9] 松澤和光, 金杉友子, 石川勉: 概念ベースによる類似性判別の「クロスワードパズル」への応用, *11th Fuzzy System Symposium*, Vol. 1, pp. 319–320 (1995).

[10] 小嶋秀樹, 伊藤昭: 概念間の連想とエピソード間の連想に基づく言語的記憶のモデル化, 情報処理学会研究報告, Vol. NL110-1, pp. 1–6 (1995).

[11] Ikehara, S., Ikehara, M. and Yokoo, A.: Classification of Language Knowledge for Meaning Analysis in Machine Translations, *Transactions of Information Processing Society of Japan*, Vol. 34, No. 8, pp. 1692–1704 (1993).