情 報 学 基 礎 57－6
自 然 言 語 処 理 136－6
（2000．3．21）

コーパスを用いた概念ベース拡張方式

稲子 希望† 笠原 要† 松澤 和光‡

† {inago,kaname}@cslab.kecl.ntt.co.jp
‡ matsuzawa.kazumitsu@lab.ntt.co.jp

† NTT コミュニケーション科学基礎研究所
‡ NTT システムインテグレーション基盤研究所

概要

単語間の類似性判別を目的とし，単語の意味を表す概念の知識ベースである「概念ベース」の研究を進めている．すでに国語辞典より4万の日常語の概念を自動抽出した概念ベースを構築し，類似性判別や情報検索等へ適用している．この概念ベースの適用範囲を広げるために，扱える概念数を拡張する必要がある．本研究では，概念ベースに含まれない語 (未対応語) の概念を推定する方法を提案する．具体的には，未対応語を含むテキストコーパスにおける単語間の共起情報を利用して，概念ベースとテキストコーパスに共通する語彙の中から未対応語に類似する語を選択し，概念ベース中の類似語の概念を用いて推定する．

A method of representing an unknown concept of a word in the "Gainen-Base" from a corpus

Nozomu Inago† Kaname Kasahara† Kazumitsu Matsuzawa‡

† {inago,kaname}@cslab.kecl.ntt.co.jp
‡ matsuzawa.kazumitsu@lab.ntt.co.jp

† NTT Communication Science Laboratories
‡ NTT Service Integration Laboratories

abstract

We have studied how to make the "Gainen-Base," a knowledge base of word concepts, from dictionaries automatically. It can simulate judgment of the semantic similarity between daily-used words. The Gainen-Base can be applied to word-related information processing such as text retrieval. For the variety of applications of the Gainen-Base, it is important to represent concepts of words even when definitions of the words are not written in dictionaries. In this paper, we propose a method of estimating a word concept which is not included in the Gainen-Base. Our method consists of two steps. First, similar words to the word are retrieved by employing knowledge bases of word co-occurrence which can be extracted from text corpola and one of which are newly proposed. Next, the unknown concept is estimated from the concepts of the retrieved words represented in the Gainen-Base.

1 はじめに

現実世界において，人間は問題を解決するのに十分な知識が与えられなくても，すでに持っている常識的知識を利用して概括的に解決することができる．我々の研究グループでは，実世界の問題に対してこのような概括的判断を計算機上で可能とする推論方式，「アバウト推論」の確立を目指して研究を進めている [1].

アバウト推論を実現するためには，人間の常識に相当する大規模な常識知識のデータベースが必須であると考えた．常識の基本部分は，基本的な単語の表す意味，すなわち「概念」であると考え，その第一歩として，概念の知識ベースである「概念ベース」の検討を行っている．アバウト推論は，欠落した知識を保有する類似した知識で補間することを特徴としているので，常識ベースの基本となる概念ベースでは，概念間の類似性判別が行えることが必要となる．

大規模な概念ベースを実現するためには，人手による知識獲得を介しない自動構築が重要である．そこで，基本的な語の定義が記述されている国語辞典から，約4万の日常語の概念を自動獲得し，3000次元の多次元空間上に表現する方法を提案した [2]. また，日常語の多くは多義性が高く，状況に応じて単語の意味が変化するため，その単語に関する類似性も変化する．そこで，文脈や状況等を表現する単語，「観点」を指定した時に，観点に応じて単語間の類似性を判別する方式を提案し評価を行った．概念ベースはアバウト推論 [3, 4, 5] 以外にも，情報検索 [6] や知識獲得 [7]，テキスト理解などに応用でき，言語を扱う情報処理の汎用的なツールとしても期待されている．

概念ベースでは，常識を構成する語彙と考えられる日常語4万語を対象としたが，アバウト推論が扱う実世界の問題には国語辞典には載っていないような新しい語や専門的な語が含まれている場合がある．これらの語も概念ベースで扱うことにより推論の精度を高めることが期待できる．また，言語処理ツールとしての概念ベースの適用範囲を広げるためにも，概念ベースにその概念が含まれていない語，すなわち未対応語の概念を表現する必要がある．

概念ベースを拡張する方法としてこれまでに，複合語の概念を合成する研究が行われている [8]. 複合語の構成語が概念ベースに含まれている場合，構成語の概念を利用して複合語の概念を表現することを特徴とする．これによって概念ベースに含まれない多数の語の概念を表現することができる．また，漢和辞典や語彙数の多い国語辞典を利用して，階層的に概念ベースを拡張する方式が提案されている [9]. 概念ベースに含まれない単語に対し，国語辞典の類義語情報や近似的な計算によって概念ベース中の類似した概念で表現を試みている．

そこで，国語辞典以外の知識源を用いて従来の概念ベースを拡張する方法を考える．国語辞典以外の言語データとしては，新聞記事などのテキストコーパスが代表的である．テキストコーパスには単語の定義が陽には記述されていない．しかし，テキストコーパスより単語の共起を取得し，その分布を比較することで，単語の類似性判別が可能である．[10, 11, 12]. そこで本研究では，概念ベースの拡張にテキストコーパスを利用する方法を検討し，テキストコーパスに含まれる未対応語の概念を概念ベース上で表現することを目的とする．

2 従来研究

ここでは，本研究の前提となる概念ベースの研究とコーパスによる単語の類似性判別の研究について述べる．

2.1 概念ベース

概念ベース [2] では，大規模な語彙の概念を自動獲得するために，容易な知識表現を用いている．国語辞典の見出し部分にある語の概念として，その語義文中の自立語である属性と，その語義文中での属性の出現頻度である属性値の対の集合で表現する．例えば「馬」は図1のように表現される．国語辞典

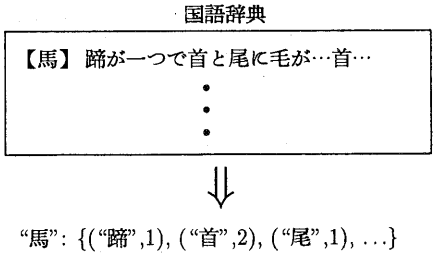


図 1: 国語辞典を利用した “馬” の概念の表現

は人間の利用を前提として作られているので，説明の省略や他の語を使った説明等が含まれ，個々の語義文だけでは十分に属性を取得することができない．そこで，再帰的参照，逆引き参照の線形結合

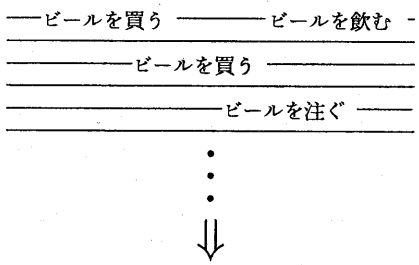
などの操作から成る自己精錬化を行っている．さらに，属性を日本語語彙大系 [13] の約3千種類のカテゴリで一般化することにより，意味的に独立した属性で概念を表現している．

概念に含まれない属性の属性値を0と考えると，それぞれの概念を3千次元のベクトルと見なすことができる．このベクトルを概念ベクトルと呼ぶ．現在の概念ベースには，約4万語に対する概念ベクトルが収録されている．この概念ベクトル同士を比較して2つの語の類似度を計算することができる．本稿ではこれを「概念類似度」と呼び，概念ベクトル同士の余弦を用いる．概念類似度は0～1の値をとり，最も類似するとき1，全く類似しないとき0となる．

2.2 コーパスを用いた単語の類似性判別

コーパスを用いた単語の類似性判別の研究 [10, 11, 12] では，コーパス中で単語毎に同時に現れる語(共起語)の出現分布を比較して類似性を判別している．このとき利用する共起の種類としては，近接共起 [12] や文法的共起 [10, 11] などがある．近接共起とは，コーパス中で単語の前後一定範囲内に出現する語を共起語とするものである．文法的共起として主要なものは，主語 - 述語，目的語 - 述語などを共起関係と見なす述語共起である．文法的共起にはほかにも，修飾語 - 被修飾語 [14] などが考えられている．

共起関係により獲得した共起語を属性，共起頻度を属性値と見なすことにより，その単語の概念を表現する．例として，目的語 - 述語共起による「ビール」の概念の表現を図2に示す．概念ベクトルと



“ビール”: {(“買う”,2), (“飲む”,1), (“注ぐ”,1), ...}

図 2: コーパス中の共起を利用した“ビール”の概念の表現

同様に，これをベクトルと見なすことができる．このベクトルを共起ベクトルと呼ぶ．コーパスから獲得した共起ベクトルを集めたものを，ここでは共起ベースと呼ぶことにする．共起ベクトルを適切に表現するためには，コーパスとしてはある程度大規模なものを用いる必要がある．

これらの共起ベクトルを比較して単語同士の類似度を計算することができる．本稿ではこれを「共起類似度」と呼ぶことにする．共起ベクトルから共起類似度を計算する方法はいくつかあるが，Hindle の計算法 [10] が代表的である．この計算法では，まず，各属性値を相互情報量に変換する．単語 u における属性(共起語) v の属性値(共起頻度) $c(n,v)$ を，以下のように相互情報量 c' に変換し属性値を補正する．

$$c' = \log_2 \frac{\frac{c(n,v)}{N}}{\frac{U(u)}{N}\frac{V(v)}{N}}$$

ここに，N はコーパスより獲得された共起の総数(延べ数)である．また，$U(u)$ と $V(v)$ は，獲得された N 個の共起(単語対)において，単語 u が出現した回数と共起語 v が出現した回数を表している．2つの単語 u_i と u_j の共起ベクトルの属性値を変換した共起ベクトルを $(c_1^{(i)}, \ldots, c_n^{(i)})$，$(c_1^{(j)}, \ldots, c_n^{(j)})$ とすると，u_i と u_j の共起類似度 $sim_C(u_i, u_j)$ は以下のように定義される．

$$sim_C(u_i, u_j) = \sum_{h=1}^{n} g(c_h^{(i)}, c_h^{(j)}) ,$$

$$g(a,b) = \begin{cases} min\{|a|, |b|\} & (a \cdot b \geq 0) \\ 0 & (a \cdot b < 0) \end{cases} .$$

3 提案方式

ここでは，概念ベースに含まれない語の概念ベクトルを推定する方法と，その方法において利用する共起について説明する．

3.1 未知の概念ベクトルの推定

本研究の目的は，概念ベースに含まれない概念，すなわち概念ベクトルが存在しない語の概念ベクトルを推定することである．図3を例にとると，概念ベースに含まれない「DVD」などの概念ベクトル，すなわち図の右下の斜線部分を推定する．

概念ベースにはこれらの単語の知識は含まれないので，単独で未知の概念ベクトルを推定することはできない．しかし，国語辞典に含まれない新語や専門語であっても，それを用いているコーパスを入手

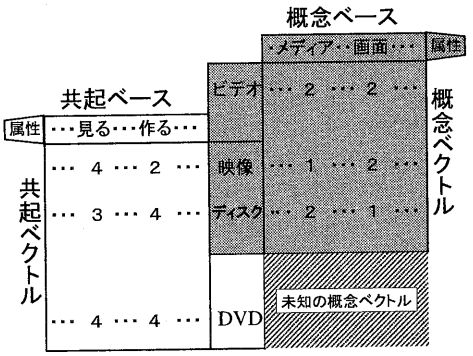


図 3: 概念ベースと共起ベース

することは可能である．そこで，これらの単語を含むコーパスを利用すれば，前述のように共起ベースを構築して，共起ベクトルを獲得できる．ただし，概念ベクトルと共起ベクトルは，構築手法が異なるので，例え同じ名前の属性であっても，その意味合いは異なり，同等なものと考えることはできない．したがって，共起ベクトルをそのまま概念ベクトルと見なすことはできない．

そこで本方式では，概念ベースと共起ベースの仲介として，両方に共通する単語のベクトルを利用する．これらの単語は概念ベクトルと共起ベクトルの両方が存在するため，未知の概念ベクトルの推定に有用であると考えられる．この方法は，概念ベースと共起ベースに語彙の重なりがあることを前提としている．どのようなコーパスであっても，国語辞典に含まれる単語を全く使用せずに記述されていることはなく，少なくともある程度は使用しており，妥当な前提と考えられる．

つぎに，未知の概念ベクトルを推定する方法について説明する．人間が日常で意味を知らない語に出会ったときには，その語の意味を理解する方法の1つとして，その語の使われ方を観察し，知っている単語の中で使われ方が似ている単語を探し出し，その意味にもとづいてその語の意味を理解することが考えられる．これをモデル化して，未知の概念ベクトルの推定法を考える．まず，概念ベクトルが未知の単語 (未対応語) に対し，それが含まれるコーパスを用いて共起ベースを構築する．つぎに，この共起ベースを利用し，未対応語に対して共起類似度が高い語を，概念ベースと共起ベースの共通の語彙より検索する．そして，それらの単語の概念ベクトルを用いて，未対応語の概念ベクトルを推定する (図 4).

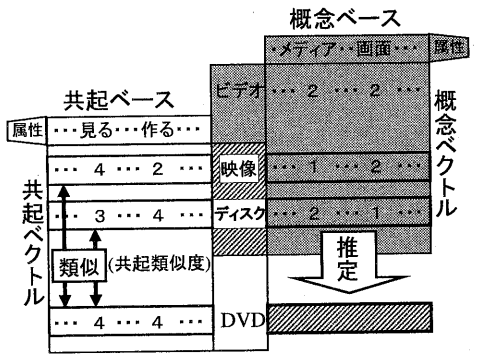


図 4: 未知の概念ベクトルの推定

ここで，各種記号の定義を行う．概念ベースに含まれる単語の集合を V_G，共起ベースに含まれる単語の集合を V_C，両方に含まれる単語の集合を $V_{GC} = V_G \cap V_C$ と書き，単語 $w \in V_G$ の概念ベクトルを $\boldsymbol{g}(w)$ で表す．また，未対応語 $k \in V_C - V_{GC}$ と，V_{GC} 中の全ての単語との共起類似度を計算し，共起類似度が大きい順に単語を並べたものを $\boldsymbol{R}(k) = (u_1, \ldots, u_n)$ で表す．ここに，$u_i \in V_{GC}$ $(i = 1, \ldots, n), n = |V_{GC}|$ である．この $\boldsymbol{R}(k)$ を利用して，k の概念ベクトルを推定する．

本稿では，$\boldsymbol{R}(k)$ から k の概念ベクトルを推定する簡単な方法として，共起類似度が上位である m 個の単語 $u_1, \ldots, u_m (1 \leq m \leq n)$ の概念ベクトルを足し合せて，それを k の概念ベクトルと見なす方法を検討する．

$$\boldsymbol{g}(k) = \sum_{i=1}^{m} \boldsymbol{g}(u_i) \qquad (1)$$

$m = 1$ のときは，k に最も類似する単語 u_1 の概念ベクトルを，そのまま k の概念ベクトルと見なす方法 $(\boldsymbol{g}(k) = \boldsymbol{g}(u_1))$ であり，「画像」と「映像」のように，u_1 が k の適切な類似語となっているならば，有効であると思われる．しかし，そうでないときは，類似していない語の概念ベクトルをそのまま利用するため，未対応語の概念が適切に表現できないと考えられる．これに対して，$m \geq 2$ のときは，たとえ u_1 が k の類似語でなくても，u_2 や u_3 が類似語となっているならば，これらの概念ベクトルによりそれを補うことができる．また，たとえ「類似する」と言える語が1つも入っていない場合でも，「経営」に対する「利益」のように，その語と関連の大きい語が多く含まれていれば，それらで属性値

を補間し合うことにより，適切な概念ベクトルが推定されることが期待できる．

3.2 共起

本方式において，未知の概念ベクトルの推定に利用する単語は，概念ベクトルが未知である単語に対して類義的であることが望ましい．そのような類似性判別を行える共起ベースを構築する必要がある．また，概念ベースと同様に多数の語彙を共起ベースで扱うため，既存で利用できる技術を前提とする必要がある．

先に述べた通り，共起の種類には近接共起 [12] と文法的共起 [10, 11] がある．近接共起は，前後定数語内といった単純な共起の判定を行っているので，共起の獲得は容易であるが，常に近接する単語同士では似ていなくても共起類似度が必然的に高くなる．このような単語同士は類義的というよりは連想的である場合が多い．

一方文法的共起は，共起する単語間に必ず直接的な関係があるため，共起類似度が高い単語同士は類義的な関係にあることが多いので，本方式に有効であると考えれれる．しかし，文法的共起の獲得には構文解析が必要であり，大規模なコーパスからの共起の獲得は難しい．

そこで本稿では，文法的共起を近似的に判定し，大規模なコーパスから比較的容易に共起を獲得する方法について検討する．また，それを補うため，複数の種類の文法的共起を併用することを考える．

近似的な獲得が可能な文法的共起として，複合語内単語共起 [15] に着目する．複合語内単語共起は，複合語を構成する単語 (構成語) 同士を共起関係と見なす方法である．複合語の構成語間には「共に複合語を構成する」という直接的なつながりがある．しかもその関係は，「情報: 抽出」(情報を抽出する)，「株価: 暴落」(株価が暴落する)，「ノート: パソコン」(ノート型のパソコン) など，実に様々である．ここでは 2 語の名詞で構成される複合語を対象とし，複合語の抽出は，形態素解析済みのテキストに対して「.../ 非名詞 / 名詞 / 名詞 / 非名詞 /...」の単語列パタンの検索により行なう．この複合語内単語共起を利用した「ビール」の概念の表現を図 5 に示す．

また，別種の共起として，同じ文内の名詞と動詞を共起関係とする文内名詞動詞共起 [16] を考える．この共起の中には主語 - 述語や目的語 - 述語の関係にある単語対が多く含まれ，述語共起の近似的獲得法と見なすことができる．したがって，共起類似度

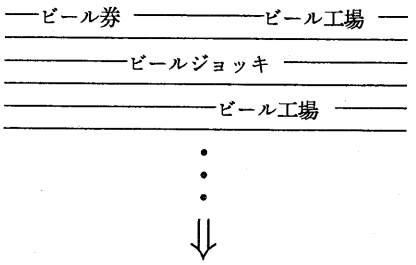


図 5: 複合語内単語共起による "ビール" の意味の表現

が高い単語同士は，近接共起に比べて類義的な関係である場合が多いと思われる．

いずれの共起も，形態素解析という比較的高速で安定した技術のみで獲得が可能であるため，大規模なコーパスに対しても容易に共起を獲得することができる．

さらに，複合語内共起の共起類似度と文内名詞動詞共起の共起類似度を足した値を，新たに共起類似度として定義することにより，2 つの共起を併用する．本稿ではこれを併用法と呼ぶことにする．

4 実験

ここでは本方式の評価法とその実験結果を示す．

4.1 評価法

未知の概念ベクトルを推定するとき，足し合わせる概念ベクトルの数 m に関する最適化を行うため，m を変化させて比較するための評価法を考える．

式 (1) において $m = 1$ のとき，すなわち $\boldsymbol{g}(k) = \boldsymbol{g}(u_1)$ とするときは，u_1 が k の類義語となっているかどうかを評価者が判定することにより，$\boldsymbol{g}(u_1)$ が k の概念ベクトルとして適切であるかを計ることができる．しかし，複数の概念ベクトルを足し合わせるとき ($m \geq 2$) は，概念ベクトル $\sum_{i=1}^{m} \boldsymbol{g}(u_i)$ を明確に表す単語が存在しない．そこで，類似概念検索の結果にもとづく評価を行った．類似概念検索とは，概念ベクトルが未知である単語に対して本方式によって推定した概念ベクトルと，概念ベース中の全ての概念ベクトルとの概念類似度を計算し，類似度が高い単語 (概念ベクトル) を検索するものであ

る．

情報検索結果の評価尺度の1つとして，適合率(精度)が一般的に用いられる．しかし，概念ベースにおける概念類似度は相対尺度であり，類似した語を与える類似度の下限を定義できない．従って，類似概念検索においても類似語の範囲を決定できないので，ここでは検索順位に着目し，20位以上の語を評価の対象とする．

足し合わせる概念ベクトル数が m のときの評価はつぎのようにして行う．まず，$V_C - V_{GC}$ の中からいくつかキーワードを選ぶ．このキーワードの集合を $K \subseteq V_C - V_{GC}$ とする．つぎに，全てのキーワード $k \in K$ に対して以下を行う．

1. 共起類似度の順位による単語列 $\boldsymbol{R}(k) = (u_1, \ldots, u_n)$ を計算する．
2. $\boldsymbol{g}(k) = \sum_{i=1}^{m} \boldsymbol{g}(u_i)$ により k の概念ベクトル $\boldsymbol{g}(k)$ を得る．
3. $\boldsymbol{g}(k)$ と全ての $\boldsymbol{g}(w)$ $(w \in V_G)$ との概念類似度を計算し，類似度上位20語 $(w_1^{(k)}, \ldots, w_{20}^{(k)})$ $(w_i^{(k)} \in V_G, i = 1, \ldots, 20)$ を検索する．
4. 評価者が k と $w_i^{(k)}$ を比較して，表1の判定値 $judge(k, w_i^{(k)})$ を決定していく．

表 1: 評価者の判定値

$judge(k, w_i^{(k)})$	k と $w_i^{(k)}$ の関係
2	類似する．
1	関連する．
0	関係ない．

最後に，全ての $(k, w_i^{(k)})$ $(k \in K, i = 1, \ldots, 20)$ に対する $judge(k, w_i^{(k)})$ の平均を計算し，これを概念数 m のときの評価値 $Eval(m)$ とする．

$$Eval(m) = \frac{\sum_{k \in K} \sum_{i=1}^{20} judge(k, w_i)}{|K| \times 20}$$

4.2 実験環境

本実験では，形態素解析器としてALTJAWS[17]，コーパスとして毎日新聞95版の経済欄[18]を用いた．評価に使うキーワードは，コーパス中の単語の出現確率に基づき，$V_C - V_{GC}$ の中から85語を選択した $(|K| = 85)$．

4.3 実験結果

共起の選択

未対応語の概念ベクトルを推定するために用いる共起ベースの評価を行った．具体的には，複合語内共起，文内名詞動詞共起，および併用法を比較した．それぞれにおいて，いくつかの単語に対して共起類似度が高い単語を共起ベース中から検索し，人手で評価したところ，併用法が最も高い評価となった[16]．その一例として，「資産」に対して共起類似度が高くなった単語を表2に示す．

共起	共起類似度が高い単語
複合語内共起	財産，資金，収入，所得，預金
文内名詞動詞共起	債権，土地，円，社，金融機関
併用法	財産，資金，土地，債権，預金

表 2: “資産” と共起類似度が高い単語

概念ベクトル推定法の実験結果

未知の概念ベクトルを推定する本方式の評価を行った．ここでは，併用法による類似度を共起類似度として用いた．まず，式(1)において $m = 1$ のとき，すなわち $\boldsymbol{g}(k) = \boldsymbol{g}(u_1)$ とする方法を評価した．ここでは単純に，それぞれの k に対して，最も共起類似度が高かった u_1 とのペア (k, u_1) を比較して，u_1 が k の類義語となっているかを人手で判定した．85対のペアを比較した結果，類義語となっているものは21対であり，適合率は25%であった．最も共起類似度が高い単語1語だけで未知の概念ベクトルを表現することは難しいことがわかった．

つぎに，足し合わせる概念の数を $m = 1, 3, 5, 7, 9$ と変化させて，前述の評価値 $Eval(m)$ を計算し，比較した．結果を表3に示す．

表 3: 足す概念数による評価値の変化

足す概念数 m	1	3	5	7	9
評価値	.454	.469	.472	.448	.437

$m = 5$ を頂点とする山型となり，未知の概念ベクトルを推定するときに，1語の概念ベクトルのみを使うより，3 ～ 5個の概念ベクトルを足し合わせた方が適切な表現となっていると言える．

例として，$m = 1, 5$ のときの「小売店」$\in V_C - V_{GC}$ の類似概念検索結果を表4に示す．表4において，$m = 1$ よりも $m = 5$ の方が適切な類似性判別となっている．“小売店”に対する，共起類似度の順位による単語列は以下のようであった．

表4: “小売店”の類似概念検索結果

順位	$m = 1$	$m = 5$
1	大規模	業者
2	小規模	商品
3	壮大	スーパー
4	絶大	大規模
5	大掛かり	デパート
6	過大	万屋
7	大仕掛け	特売
8	特大	見切り品
9	甚大	棚浚え
10	大いなる	売り子

$\boldsymbol{R}$(“小売店”) =
(“大規模”, “商品”, “業者”, “スーパー”, “店舗”, ...)

$m = 1$ のときは，$\boldsymbol{g}$(“小売店”) = $\boldsymbol{g}$(“大規模”) であり，“小売店”と“大規模”は類似していないため，このような結果になった．しかし，$m = 5$ のときは，$\boldsymbol{g}$(“小売店”) = $\boldsymbol{g}$(“大規模”) + $\boldsymbol{g}$(“商品”) + $\boldsymbol{g}$(“業者”) + $\boldsymbol{g}$(“スーパー”) + $\boldsymbol{g}$(“店舗”) であり，“小売店”と類似する単語や関係が深い単語の概念ベクトルを含んでいるため，適切な類似概念検索ができていると思われる．

本方式における共起の種類の最適化

前述のように共起類似度による共起の評価では，複合語内共起，文内名詞動詞共起，併用法(両共起の併用)の3つのうち，併用法が最も高い評価となった[16]．この結果が未知の概念ベクトルの推定に反映されていることを確かめた．ここではそれぞれの共起による本方式の実験を行い，前節と同様の方法で評価した．このとき，足す概念ベクトルの数は，$m = 5$ とした．結果を表5に示す．

表5: 共起を用いた類似性判別法による評価値の変化

	複合語内共起	文内名詞動詞共起	併用法
評価値	.419	.429	.472

併用法が，それぞれの共起を単独で利用したときよりも良い結果となった．共起類似度による類似性判別の精度が本方式の精度に反映されていることがわかる．

5 おわりに

本稿では，概念ベースに含まれない語(未対応語)の概念を推定する概念ベースの拡張研究の一つとして，未対応語を含むテキストコーパスを知識源として利用する方法を検討した．具体的には，テキストコーパス中の共起情報を利用して，未対応語に類似した単語の概念ベクトルを用いて，未対応語の概念ベクトルを推定した．

未対応語の概念ベクトルを推定するときに用いる概念ベクトルの数を変化させて，利用する概念ベクトルの数に関する最適化を行った．今回の実験においては，5個の概念ベクトルを利用したときに，最も良い結果が得られ，複数の概念ベクトルから未対応語の概念ベクトルを推定することが適当であることがわかった．

また，共起を用いた類似性判別法として，複合語内共起，文内名詞動詞共起，両方を併用した方法を用いて実験を行い比較した結果，両方を併用した方法が最も良い結果となった．

今後は，共起を用いた類似性判別法の精度向上を検討する予定である．また，未対応語の概念ベクトルを推定するとき，今回は単純に足し合わせたが，順位や類似度によって重み付けして足す方法についても検討する予定である．さらに，概念ベースの特徴の一つである「観点」を利用して，未対応語の概念ベクトルを推定するとき，観点変調をかけてから概念ベクトルを足し合わせる方法なども考えている．

今回は人手で評価を行ったが，人間による評価は時間がかかるため，より大規模な評価は難しい．そこで，評価法として，キーワードを概念ベースと共起ベースの両方に含まれる単語の中から選び，そのオリジナルの概念ベクトルにどれだけ近い概念ベクトルを推定できるかで，自動的に評価する方法を検討している．

参考文献

[1] 松澤和光, 石川勉. アバウト推論とその類似性判別機構. 人工知能学会研究会資料, SIG-J-9401, pp. 103–110, 1994.

[2] 笠原要, 松澤和光, 石川勉. 国語辞書を利用した日常語の類似性判別. 情報処理学会論文誌, Vol.

38(7), pp. 1272–1284, 1997.

[3] A. Abe. Two-sided hypotheses generation for abductive analogical. In *Tools with Artificial Intelligence*, pp. 145–152, 1999.

[4] K. Fujimoto and K. Matsuzawa. Intelligent systems using web-pages as knowledge base for statistical decision making. In *New Genaration Computing*, Vol. 17, pp. 349–358. Springer-Verlag, Ohmsha Ltd, 1999.

[5] N. V. Ha, T. Ishikawa, and A. Abe. A Mechanism for Inferring Approxiamate Solutions under Incomplete Knowledge based on Rule Similarity, *Proc. of AISTA2000* (2000)

[6] 熊本睦, 島田茂生, 加藤恒昭. 概念ベースの情報検索への適応 - 概念ベースを用いた検索の特性評価 -. 情報処理学会研究報告, SIG-ICS 115, pp. 9–16, 1999.

[7] 賀沢秀人, 藤本和則, 松澤和光. Web テキストを知識ベースとして用いる推論システムの提案 - テキストからの知識獲得方式を中心に -. 人工知能学会研究会資料, SIG-KBS-9803-9, pp. 49–54, 1999.

[8] 永森千晴, 金杉友子, 笠原要, 松澤和光. 概念ベースを用いた複合語概念の合成. 第 55 回情報処理学会全国大会講演論文集, 2, pp. 210–211, 1997.

[9] 帆刈譲, 石川勉, 笠原要. 言葉の意味に関する階層型大規模概念ベースの構築. 情報処理学会研究報告, SIG-ICS-115, pp. 25–32, 1999.

[10] D. Hindle. Noun classification from predicate-argument structures. In *Preoceedings of ACL*, pp. 268–275, 1990.

[11] 平岡冠二, 松本裕治. コーパスからの動詞の格フレーム獲得と名詞のクラスタリング. 情報処理学会研究報告, 94-NL-104, pp. 79–86, 1994.

[12] H. Schütze. Dimensions of meaning. In *Proceedings of Supercomputing 92*, pp. 787–796, 1992.

[13] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林. 日本語語彙体系. 岩波書店, 1997.

[14] V. Hatzivassiloglou and K. McKeown. Towards the automatic identification of adjectival scales: clustering adjectives according to meaning. In *Preoceedings of ACL*, pp. 172–182, 1993.

[15] 稲子希望, 笠原要, 松澤和光. 複合語内の単語共起を用いた単語の類似性判別方式. 「言語資源の共有と再利用」シンポジウム, http://www.etl.go.jp/etl/nl/sympo99/index.html, 1999.

[16] 稲子希望, 笠原要, 松澤和光. 複合語内単語共起による名詞の類似性判別. 情報処理学会論文誌, (投稿中).

[17] S. Ikehara, S. Shirai, A. Yokoo, and H. Nakaiwa. Toward an mt system without pre-editing-effects of new methods in alt-j/e. In *MT Summit '91*, pp. 101–106, 1991.

[18] 毎日新聞社. CD- 毎日新聞 95 版.