複合語内単語共起による名詞の類似性判別

稲子希望† 笠原 要† 松澤和光†

本稿では，新聞などのテキストコーパスから単語の共起関係を獲得して，単語間の類似性判別を行う手法を提案する．提案手法は，コーパス中の名詞2語から成る複合語に着目し，構成する名詞どうしを共起関係にあると見なす点が新しい．従来の共起関係（主語–述語，目的語–述語）とは異なる特徴に基づく名詞の類似性判別が実現できる．実際に新聞記事1年分をコーパスとして実験を行い，複合語内の共起関係が類似性判別に有効であること，従来の共起関係と併用することによりさらに質の高い類似性判別が行えることを示した．

A Method for Judging the Semantic Similarity between Nouns by Using Noun-Noun Co-occurrence in Compound Words

Nozomu Inago,† Kaname Kasahara† and Kazumitsu Matsuzawa†

We propose a simulation method for judging the similarity between nouns by utilizing co-occurrence of a new type extracted from text corpora that is, noun-noun co-occurence in compound words. The proposed method judges the similarity from a different aspect than ordinal methods, which employ noun-verb co-occurrence. The corpus, we employed to evaluate the proposed method, contains newspapers collected over a period of one year. It is shown that the new co-occurrence are useful for judging the similarity between nouns and that the combination of the new co-occurrence and ordinal co-occurrence results in a better simulation.

1. はじめに

ある単語と単語の意味が似ているかどうかを工学的に判定する技術，すなわち単語間の類似性判別技術が注目を浴びている．たとえば，ユーザが入力したキーワード以外にも類似したキーワードを含む記事を探し出す曖昧情報検索[1)]や，機械翻訳において類似した翻訳例を利用する用例ベース翻訳技術[2)]，あるいは構文解析における係り受けの曖昧性解消に類似性を利用する[3)]など，情報処理や自然言語処理の様々な分野で，この類似性判別技術が必要とされ，また実際に利用されて大きな成功を収めている．

単語の類似性判別を行う手法の1つとして，テキストコーパスを統計的に解析して利用する研究が行われている[3)~5)]．基本的な方法としては，対象とする単語に対してコーパス中で共起する単語を特徴とし，その出現頻度を基にした特徴の重みから成る共起ベクトルを構成する．そして，得られた共起ベクトルどうしの近さに応じて単語の類似性を判別する．共起とは，特定の条件下において2つの単語がコーパス中で同時に現れる関係を指し，ある語と共起する語を共起語と呼ぶ．コーパス中の共起を用いた単語の類似性判別の研究では，どのような共起を利用するかが重要であり，これまでに，近接性に基づく共起[5)]と文法的な共起[3),4)]の2種類の利用が提案されている．

近接性に基づく共起とは，コーパス中で2つの単語が近接する関係を表す．近接性の範囲としては，隣接する単語どうしや固定長単語列内，同一文内など様々に設定されている．それに対し文法的な共起は，文中で一定の文法規則で関係付けられた単語どうしの関係を表す．これまでに，文法的な共起を用いた単語の類似性判別としては，主語に対する述語，目的語に対する述語を共起語として利用する方式が提案されている[4)]．たとえばコーパス中の文を構文解析し，「ワイン」を目的語とする述語「飲む」,「買う」,「注ぐ」などを共起語として獲得し，その出現頻度に基づいて「ワイン」の共起ベクトルを構成する．他の名詞についても同様なベクトルを作った場合，「ワイン」の共起語ベクトルに対し，ほぼ同じ述語共起をとりうる「ウイス

† NTTコミュニケーション科学基礎研究所
NTT Communication Science Laboratories

キー」のような名詞の共起ベクトルとの距離が近くなることが期待され,「ワイン」と「ウイスキー」が類似していると判定することができる．このような手法は，英語だけでなく日本語についても研究が進められている[6]．

主語や目的語に対する述語動詞は，名詞に対する文法的な共起として中心的な存在であり，名詞の類似性に用いられる文法的共起として重要である．しかし，名詞と文法的な結び付きを持つ語は，名詞を修飾する形容詞など様々存在し，類似性判別においては，動詞とともに他の文法的な共起の情報を総合的に利用することで，判別精度の向上が期待される．そこで本稿では，新しい文法的な共起として複合語内の名詞–名詞共起を取り上げ，これを用いた名詞の類似性判別を行う方法を提案する．複合語内の名詞共起による類似性判別は,「類似した 2 つの名詞は，同じ種類の共起語とともに複合語を構成する」という仮定に基づく．そして，提案した複合語内共起と従来の述語動詞に関する共起を併用し，名詞の類似性判別を最適化する方法についても検討を行う．

まず，2 章では，複合語内の共起とその獲得法，名詞の類似性判別の方法について述べる．次に 3 章では，新聞記事をコーパスとして用いた実験方法，評価方法について述べる．そして 4 章では，実験結果およびその考察，さらに複数の文法的な共起の併用方法に関する検討を行う．

2. 提 案 手 法

ここでは，複合語を利用して名詞の特徴を表す共起ベクトルを獲得する方法と，これを用いた名詞間の類似度計算法を述べる．

2.1 複合語内単語共起

複合語とは,「複数の語が結合して 1 語となったもの」[7]など，言語学的に様々な定義があるが，ここでは自立語に分割できる語と考える．複合語を構成する語数や品詞の種類は様々であるが，コーパスで出現する複合語は，名詞 2 語から成る場合が多い．たとえば，新聞記事[8]中の 100 語の複合語において，名詞 2 語で構成されるものは 72 語であった．そこで本稿では，複合語を構成する 2 語の名詞間の共起に着目する．

ここでは，複合語中の後側の構成語を後方構成語，前側の構成語を前方構成語と呼ぶ．複合語を成す構成語どうしの関係は,「販売開始」や「企画立案」のように目的語と述語の語幹からなるような文法的に説明ができる関係から,「三角定規」や「観光旅行」のように文法では説明できない関係まで多岐にわたっている[9]．

表 1 2 種類の複合語内単語共起
Table 1 Two kind of cooccerrences in a compound word.

複合語内後方共起	複合語内の前方構成語を基本語，後方構成語を共起語とする共起
複合語内前方共起	複合語内の後方構成語を基本語，前方構成語を共起語とする共起

表 2 「企画」の複合語内共起語の一例
Table 2 Example of cooccurrnce words for ‘企画’ (kikaku).

複合語内後方共起 (企画 ∗)	会社，部長，協力，賞，立案，会議，課長，運営，案，室長
複合語内前方共起 (∗ 企画)	特別，連載，商品，関連，イベント，自主，共同，記念，チャリティー，経営

このため，従来の文法的共起では得られない様々な関係の獲得が期待できる．

複合語中の後方構成語と前方構成語の位置には意味があり，可換性はない．たとえば，複合語中の後方構成語と前方構成語を入れ換えると，別の意味の語や意味のない文字列になる．したがって，共起に基づく単語の類似性判別において，同じ表記であっても後方構成語と前方構成語は別種の共起として扱う必要がある．この考えに基づき，複合語を利用した 2 種類の共起関係を定義する．ここでは，構成語のうちで類似性判別の対象となる語を基本語，基本語に対して共起する語を共起語と呼ぶ．その場合，**表 1** の共起関係が与えられる．

例として，名詞「企画」とともに複合語を構成する名詞，すなわち「企画」の共起語となる名詞を**表 2** に示す．

2.2 共起ベクトル

上記の 2 種類の複合語内共起の個々について，下記のとおり基本語の共起ベクトルを構成する．

コーパス中のすべての 2 語から成る複合語における基本語の集合を $\mathcal{N}$，共起語の集合を $\mathcal{M}$ とする．任意の $m \in \mathcal{M}$ と $n \in \mathcal{N}$ に対して，m が n の共起語として出現する回数を，$c_n(m)$ で表す．基本語 n の特徴を表す共起ベクトル $\overrightarrow{c_n}$ を以下のとおり定義する．

$$\overrightarrow{c_n} = (c_n(m_1), \ldots, c_n(m_{|\mathcal{M}|})),$$

$$(m_i \in \mathcal{M}, i = 1, \ldots, |\mathcal{M}|).$$

以後，共起ベクトルの要素である共起語を属性と呼び，その属性の値を属性値と呼ぶことにする．

n が基本語としてコーパス中に出現する回数および m が共起語としてコーパス中に出現する回数を，それぞれ $N(n)$，$M(m)$ で表す．

$$N(n) = \sum_{m \in \mathcal{M}} c_n(m) \ , \ M(m) = \sum_{n \in \mathcal{N}} c_n(m) \ .$$

2.3 類似度計算法

共起ベクトル間の近さを表す尺度として，以下の3つの**類似度の条件**を満たすような類似度 sim を採用する．

(1) **近接尺度**：2つの語が類似しているほど，類似度の値が大きい．

(2) **可換性**：類似度に方向性がない．すなわち，任意の基本語 $n_1, n_2 \in \mathcal{N}$ に対して，$sim(n_1, n_2) = sim(n_2, n_1)$ となる．

(3) **最大値**：ある基本語に対して最も類似度が高くなるのは，その語自身との類似度である．すなわち，任意の基本語 $n_1, n_2 \in \mathcal{N}$ に対して，$sim(n_1, n_2) \leq min\{sim(n_1, n_1), sim(n_2, n_2)\}$ となる．

このような類似度計算法としては，まず Hindle が主語–述語，目的語–述語共起に対して適用した方法[4]（以下，ヒンドル法）があげられる．しかし，文献 4) においては，この方法が類似度の計算法として最適であるかは述べられていない．そこで本稿では，ベクトルどうしの近さを計る尺度として一般的な，余弦，ダイバージェンスも類似度計算法として考え，ヒンドル法と比較する．

2.3.1 ヒンドル法

ヒンドル法ではまず，基本語と共起語の単語対 $(n, m) \in \mathcal{N} \times \mathcal{M}$ に対して，共起の相互情報量 $I_n(m)$ を次のように定義する．

$$I_n(m) = \log_2 \frac{\frac{c_n(m)}{S}}{\frac{N(n)}{S}\frac{M(m)}{S}} \ , S = \sum_{n \in \mathcal{N}} \sum_{m \in \mathcal{M}} c_n(m) \ .$$

基本語 $n_i, n_j \in \mathcal{N}$ の類似度 $sim_H(n_i, n_j)$ は，以下の式で与えられる．

$$sim_H(n_i, n_j) = \sum_{m \in \mathcal{M}} g(I_{n_i}(m), I_{n_j}(m)) \ ,$$

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

$sim_H(n_i, n_j)$ は，前述の3つの類似度の条件を満たす．また，類似度の上限はなく，共起ベクトルによって異なる．

2.3.2 余　　弦

情報検索におけるベクトルモデル[10]の考え方を適用して，2つのベクトルの成す角の余弦を類似度とする．まず，各共起ベクトルの属性値を，重要性尺度 idf ($inversed\ document\ frequency$)[11]によって補正する．基本語 $n_i, n_j \in \mathcal{N}$ の類似度 $sim_C(n_i, n_j)$ は，補正済のベクトルどうしの余弦として定義される．

2.3.3 ダイバージェンス

確率分布どうしの距離を計るダイバージェンスを利用する．まず，共起ベクトル $\overrightarrow{c_n}$ を，その属性値の和が1となるように，それぞれの属性値を補正する．補正後のベクトルを $\overrightarrow{c'_n}$ で表すことにする．

基本語 $n_i, n_j \in \mathcal{N}$ の類似度 $sim_D(n_i, n_j)$ を以下のように定義する．

$$sim_D(n_i, n_j) = 1 - \frac{D(\overrightarrow{c'_{n_i}}, \overrightarrow{r}) + D(\overrightarrow{c'_{n_i}}, \overrightarrow{r})}{2} \ .$$

ここに，$\overrightarrow{r}$ は，$\overrightarrow{c'_{n_i}}$ と $\overrightarrow{c'_{n_j}}$ の各属性値の平均値を属性値として持つベクトル（重心ベクトル）である．ダイバージェンス関数 D は，以下のように定義される．

$$D(\overrightarrow{p}, \overrightarrow{q}) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} \ .$$

ここに，$\overrightarrow{p}$ と $\overrightarrow{q}$ は，事象の集合 $\mathcal{X}$ 上の確率分布である．関数 D 自体は，前述の3つの類似度の条件を満たさないため，上のように $sim_D(n_i, n_j)$ を定義している．

3. 実　　験

ここでは本手法の具体的な実験方法，提案手法に対する比較手法，および評価法について述べる．

3.1 具体的実験方法

実験では，コーパスとして毎日新聞 95 版[8]を利用し，形態素解析器として ALT-JAWS[13]を用いる．ここでは ALT-JAWS の形態素辞書に含まれる語を単語として，それらの組合せで表現することができる語を複合語とする．複合語の獲得および構成語への分割には構文解析が必要であるが，大規模なコーパスを対象とした場合，構文解析に要する時間的コストは大きい．そこで本稿では近似的な複合語の判定法を提案する．まずテキストに対して形態素解析を行う．次に文内の品詞の並びに注目して，

$$\cdots \text{非名詞} \cdot \text{名詞} \cdot \text{名詞} \cdot \text{非名詞} \cdots$$

のパターンが現れたときに，2つの名詞が共起していると判定する．実際に毎日新聞 95 版に対してこの方法を適用し，のべ 120 万，30 万種類の複合語内共起を獲得した．これにより得られた複合語内後方共起の基本語数は 4 万 9 千語，共起語数は 4 万 3 千語であった．複合語内前方共起は，逆に基本語数が 4 万 3 千語，共起語数が 4 万 9 千語であった．

この近似的判定法の妥当性を確かめるため，コーパス中の 100 文に対する本判定法の精度を調べた．獲得

した名詞対 110 個に対して，人手で複合語と見なせるかどうかを判定した．その結果，複合語と見なせる名詞対は 110 個中 101 個であり，91.8%の精度であった．この判定方法は，十分な精度で複合語を獲得できるといえる．

3.2 比較手法

提案する複合語内共起を用いた類似性判別従来手法と比較するために，従来用いられている文法的な共起を利用する．文献 4) では，主語–述語あるいは目的語–述語の共起を個々に獲得しているが，そのためには構文解析が必要であり，前述のように構文解析に必要な時間的コストは大きい．そこで，ブロック内のすべての名詞を基本語，すべての動詞を共起語とする文内名詞動詞共起を比較手法として用いる．ブロックは句読点ではさまれた区間とする．これを毎日新聞 95 版 1 年分に適用し，のべ 400 万，170 万種類の文内名詞動詞共起を獲得した．これにより得られた基本語数は 5 万 5 千語，共起語数は 1 万 1 千語であった．

同じ毎日新聞 95 版から抜き出した 100 文においては，得られた文内名詞動詞共起はのべ 241 対であった．そのうち，名詞と動詞が文章中で格関係にあると人手で判断されたものは 94 対であり，精度は 39%であった．これは，十分な精度とはいえないが，新聞記事 1 年分を短期間で解析する妥当な方法はなく，ここでは精度よりも大規模なデータに対して比較することが重要であると考え，上記データを用いる．

3.3 評価法

本稿では，類似語検索の結果に基づいて類似性判別の評価を行う．類似語検索とは，ある基本語をキーワードとして，このキーワードとすべての基本語との類似度を計算し，類似度が上位となる基本語を検索するものである．つまり，類似度が上位となる語をキーワードの類似語と見なすわけである．

情報検索結果の評価尺度としては，適合率（精度）が一般的に用いられるが，提案方式で得られる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．したがって，類似語検索においても類似語の範囲を決定できないので，検索順位に着目した評価指標を用いる．

類似語検索のためのキーワードとして，50 個の単語を選択する．ただし，コーパスから得るデータにはばらつきが大きく，属性値が 0 でない属性が 1 つしかない基本語も多い．このような基本語は，その特徴を十分に獲得できていないため，統計的共起情報を利用する本手法においては，評価用のキーワードとして適当でない．逆に 0 でない属性数が極端に多い基本語も，代表的な基本語とはいえないために評価用のキーワードとしては適当でない．そこで，複合語内共起ベクトルを参考にして，キーワードとして適当と思われる基本語を選ぶ．具体的には，複合語内後方共起の基本語 4 万 9 千語および複合語内前方共起の基本語 4 万 3 千語の中から，今回は，0 でない属性数が 51〜100 の基本語の中からランダムに 50 語を選択する．このキーワードの集合を $\mathcal{K}$ ($|\mathcal{K}| = 50$) とする．$\mathcal{K}$ は基本語集合の部分集合である．今回使用するキーワード集合 $\mathcal{K}$ を以下に提示する．

表 3 評価者の判定値
Table 3 Similarity judgement scores.

$judge(k, n_i^{(k)})$	k と $n_i^{(k)}$ の関係
1	類似している，関連している．
0	判定できない．
−1	関係ない．

$\mathcal{K}$ = {航空，保険，証券，設立，反対，企画，通信，救済，モデル，ビデオ，練習，エネルギー，映像，金利，警察，住民，芸術，治療，発行，メディア，美術，道路，希望，青年，監視，審査，投票，輸送，保護，犯罪，継続，民族，連絡，製造，制作，試合，採用，鉄道，推薦，ホテル，援助，労働，資産，車両，責任，取材，科学，ソフト，指定，演劇}

ある類似性判別法 sim の評価は次のようにして行う．

(1) 類似性判別法 sim により，キーワード $k \in \mathcal{K}$ に対する類似語検索を行い，上位 20 位までの単語 $n_i^{(k)} (i = 1, \ldots, 20)$ を sim による k の類似語とする．i は順位を表している．

(2) 評価者が k と $n_i^{(k)}$ を比較して，**表 3** の判定値 $judge(k, n_i^{(k)})$ を決定していく．

(3) $judge(k, n_i^{(k)})$ を順位の逆数で重み付けした和を，キーワード k における類似性判別法 sim の評価値 $eval(k, sim)$ とする．ただし，いずれの類似性判別法においても $n_1^{(k)} = k$ となるので，必ず $judge(k, n_1^{(k)}) = 1$ である．よって，1 位は評価の対象外とし，2〜20 位を対象に評価値を計算する．

$$eval(k, sim) = \sum_{i=2}^{20} \frac{judge(k, n_i^{(k)})}{i} .$$

$eval(k, sim)$ のとる値の範囲は，およそ −2.6〜2.6 である．

(4) 任意のキーワード $k \in \mathcal{K}$ の評価値 $eval(k, sim)$ の平均を，類似性判別法 sim の評価値

表 4 各類似度計算法による類似語検索の評価
Table 4 Result of evaluation for methods of calculating a dgree of similarity.

類似度計算法	評価値 $Eval(sim_*)$
ヒンドル法 sim_H	0.784
余弦 sim_C	−0.634
ダイバージェンス sim_D	−0.079

$Eval(sim)$ とする．

$$Eval(sim) = \frac{\sum_{k \in \mathcal{K}} eval(k, sim)}{|\mathcal{K}|} .$$

本実験では，さらに 3 人の評価者による評価値の平均をとり，最終的な評価値とする．

4. 実験結果と考察

ここでは，前章で述べた評価方法に基づいた提案方式の最適化，および比較手法との比較結果を示し，さらにその統合方法について考察を行う．

4.1 最適な類似度計算法の選択

本手法における類似度計算法として最適な方法を決定するために，前述の 3 つの計算法を比較した．共起としては，複合語内後方共起を使用した．結果を**表 4** に示す．

余弦やダイバージェンスに比べて，ヒンドル法の評価値が高くなった．余弦やダイバージェンスによる類似語検索結果で，ふさわしくないと人手で判定された語の多くは，その共起ベクトルのほとんどの属性値が 0 であった．獲得された共起語が数語程度の基本語では，コーパス中で頻出する共起語が含まれる場合，その共起ベクトルへの寄与は大きくなる．同じ共起語がキーワードの共起語に含まれる場合，その基本語とキーワードの類似度は 2 つの方法では必然的に高くなってしまう．つまり，獲得された共起語が数語程度の基本語は，その特徴を十分に獲得できていないために，関係のないキーワードとの類似度が高くなってしまっている．しかしヒンドル法では，共起ベクトルの属性値を相互情報量に変換しているために，コーパス全般で頻出する共起語の影響を軽減できる．さらに，2 つの共起ベクトルに対する類似度の算出法が，属性ごとに属性値を比較し，値が小さい方の値の和を類似度とする方法となっているため，数語程度の属性一致により類似度が上位となる検索結果の類似語は少なくなったと考えられる．

この結果により，複合語内共起を用いた名詞類似性判別においては，他の文法的な共起での類似性判別と同様に，ヒンドル法が有効であることが明らかになった．以降，類似度計算法としてはヒンドル法 sim_H を用いる．

表 5 類似語検索における共起種の効果
Table 5 Effect of species of cooccurrence on similar word retrieval.

共起	評価値 $Eval(sim^*_H)$
後方 sim^r_H	0.784
前方 sim^f_H	0.885
平均 sim^c_H	1.093
文内 sim^s_H	1.018

4.2 共起種の比較

2 章で述べたとおり，複合語内の構成語の位置（後方/前方）によって共起語の役割や性質が異なる．そこで，表 1 で定義した複合語内後方共起と複合語内前方共起を用いた類似性判別を比較した（**表 5**）．sim^r_H は複合語内後方共起による類似性判別法，sim^f_H は複合語内前方共起を用いた類似性判別法を表す．その結果，sim^f_H の方が類似性判別の能力が高いことが明らかとなった．これは，2 語から成る複合語において前方構成語が後方構成語に従属することが多く，後方構成語に対する前方構成語が，前方構成語に対する後方構成語に比べて，語の特徴をより適切に表しているためと考えられる．

文献 4) では，主語としての名詞に共起する動詞を属性とした共起ベクトルと，目的語としての名詞に共起する動詞を属性とした共起ベクトルを個々に構成し，個々に得られる類似度を平均した値を名詞に対する類似度として定義している．これは，類似性判別において，異なる種類の共起を併用して扱う方法と見なすことができる．そこで，本稿で提案した複合語内後方共起と複合語内前方共起について適用し，評価を行った．2 種類の複合語共起を併用した類似性判別法 sim^c_H は，以下のように表される．

$$sim^c_H(n_i, n_j) = \frac{sim^r_H(n_i, n_j) + sim^f_H(n_i, n_j)}{2} .$$

その結果，表 5 に示すとおり，個々の種類の複合語内共起を単独に用いた場合よりも，両者の共起に基づく類似度を平均する類似性判別法 sim^c_H の評価が高くなった．名詞の後方構成語としての特徴と前方構成語としての特徴を別々にとらえ，併用することにより類似性判別の精度が高くなったためであり，さらに，複数種類の共起を併用して類似性判別を行うこの方法は，汎用的であると考えられる．

比較手法として，文内名詞動詞共起による類似性判別法 sim^s_H を評価した（表 5）．この評価結果に対し，個々の種類の複合語内共起を用いた類似性判別の評価

表 6 類似語検索結果
Table 6 Example of slimilar words retrival.

	複合語内 sim_H^c	文内 sim_H^s
1	資産	資産
2	財産	債権
3	資金	土地
4	収入	円
5	所得	社
6	預金	金融機関
7	株	預金
8	土地	事業
9	企業	財産
10	株式	資金
11	債権	企業
12	担保	一部
13	投資	金
14	融資	融資
15	情報	現在
16	取引	ドル
17	債券	会社
18	事業	法人
19	資料	業者
20	債務	銀行

値の方が値が低いが，両者を併用した類似性判別の評価値は比較手法よりも高くなり，複合語内共起の有効性を示している．

今回は，構文解析を使わない簡略な文内名詞動詞共起を比較手法として採用し，厳密な主語–述語共起，目的語–述語共起との比較は行わなかった．これらの共起の獲得に必要な時間的コストは大きいが，提案手法との類似性判別の相違をみるために，今後，これらの共起を用いた実験を行いたい．

基本語「資産」に対する，複合語内単語共起（平均）と文内名詞動詞共起の類似語検索結果の例を**表 6** に示す．また，比較のために，検索対象である基本語集合からランダムに選択した単語の例を以下に提示する．

イレギュラー，麻奈美，総合雑誌，ほうせんか，一直，林，屈伸，豊能，堀，天引き，逝去，主語，ピースト，強国，玉英，ブッダ，瑞穂，クルーガー，兼二，雄介，光岡，ウェルギリウス，年克，命中，権現，ろうばい，鐘淵，めった，口前，高潔さ，キログラム，滋賀，享史，内観，小岩井，待ち合い，登録名，ｊａｐａｎ，色紙，素姓，徳三郎，暗中，ムジール，WDR，足柄，積極さ，マフラー，角野，党派，全貌，...

このように，基本語の多くは意味的には「資産」とかけ離れているが，表 6 では，その中から「資産」に対して関連の深い語が上位に検索されていることが分かる．それにもかかわらず，2 つ手法においては異なる検索結果となっており，次に述べる，さらなる共起種の併用の効果が期待される．

表 7 基本語集合，共起語集合，共起ベクトル
Table 7 Definition of terms.

共起	基本語集合	共起語集合	n の共起ベクトル
複合語内後方	$\mathcal{N}_r$	$\mathcal{M}_r\,(=\mathcal{N}_f)$	$\overrightarrow{c_n^r}$
複合語内前方	$\mathcal{N}_f$	$\mathcal{M}_f\,(=\mathcal{N}_r)$	$\overrightarrow{c_n^f}$
文内名詞動詞	$\mathcal{N}_s$	$\mathcal{M}_s$	$\overrightarrow{c_n^s}$

4.3 異なる文法的共起種の併用

文法的な共起を用いた類似性判別において，従来用いられてきた名詞–動詞の共起（文内名詞動詞共起）と，本稿で提案した複合語内名詞–名詞共起を併用して，名詞の類似性判別を行う方法について考察する．併用の方法としては，文献 4) で主語–述語，目的語–述語の共起を併用し，本稿で提案した 2 種類の複合語内共起を併用する際に用いた，類似度の平均による方法が有望である．一方，単純な方法として，異なる種類の共起も 1 つの共起ベクトル中で表現する方法も考えられる．そこでこの 2 つの併用方法について比較を行った．

以後の説明のため，基本語集合，共起語集合，基本語 n に対する共起ベクトルを**表 7** のように表す．

併用の仕方としては，以下の 2 種類の方法で行う．

- ベクトル結合：まず，複合語内後方共起ベクトル $\overrightarrow{c_n^r}$ と文内名詞動詞共起ベクトル $\overrightarrow{c_n^s}$ を結合する．得られたベクトルを結合後方ベクトルと呼ぶことにする．

$$\left.\begin{aligned}\overrightarrow{c_n^r} &= (r_1, \ldots, r_{|\mathcal{M}_r|})\\ \overrightarrow{c_n^s} &= (s_1, \ldots, s_{|\mathcal{M}_s|})\end{aligned}\right\}$$

$$\Downarrow$$

$$(r_1, \ldots, r_{|\mathcal{M}_r|}, s_1, \ldots, s_{|\mathcal{M}_s|})$$

結合後方ベクトル．

すべての基本語 $n \in \mathcal{N}_r \cup \mathcal{N}_s$ に対して結合後方ベクトルを計算した後，各属性値の相互情報量をもとめ，類似度を計算する．複合語内前方共起ベクトル $\overrightarrow{c_n^f}$ と $\overrightarrow{c_n^s}$ も結合して，同様に類似度を計算する．今得られた 2 種類の類似度の平均値 SIM_H^{join} を最終的な類似度とする．

- 類似度加算：n_i，n_j に対して，$sim_H^c(n_i, n_j)$ と $sim_H^s(n_i, n_j)$ をそれぞれ計算した後，2 つの類似度を加算する．ただし，ヒンドル法は最大値が決まっていないため，単純に加算すると，どちらか一方の影響が過分に現れてしまう可能性がある．今回の実験では，多くの基本語対 (n_i, n_j) に対し

表 8 複数共起併用による「資産」の類似語検索結果
Table 8 Result of slimilar words retrival.

順位	結合 SIM_H^{join}	加算 SIM_H^{sum}
1	資産	資産
2	土地	財産
3	債権	資金
4	円	土地
5	社	債権
6	財産	預金
7	金融機関	企業
8	預金	事業
9	資金	株
10	企業	収入
11	現在	融資
12	事業	金融機関
13	融資	担保
14	ドル	株式
15	金	円
16	一部	所得
17	収入	経営
18	場合	金
19	負担	一部
20	総額	市場

表 9 複数共起併用による類似語検索の評価
Table 9 Evaluation for each method of using together.

併用法	評価値 $Eval(SIM_H^*)$
結合 SIM_H^{join}	1.180
加算 SIM_H^{sum}	1.312

て，$sim_H^c(n_i,n_j) < sim_H^s(n_i,n_j)$ であった．そこで，次のように重み付けして加算する．

$$SIM_H^{sum}(n_i,n_j) = \frac{1}{\alpha}\cdot sim_H^c(n_i,n_j)+\frac{1}{\beta}\cdot sim_H^s(n_i,n_j)\ ,$$

$$\alpha = \frac{\sum_{n\in\mathcal{N}_r\cup\mathcal{N}_f} sim_H^c(n,n)}{|\mathcal{N}_r\cup\mathcal{N}_f|}\ ,$$

$$\beta = \frac{\sum_{n\in\mathcal{N}_s} sim_H^s(n,n)}{|\mathcal{N}_s|}\ .$$

α，β は，それぞれの共起による類似度の最大値の平均である．

それぞれの類似語検索結果の例，および評価結果を**表 8**，**表 9** に示す．

表 9 と表 5 を比較すると，複合語内単語共起と文内名詞動詞共起を単独で使うよりも，併用方法のどちらもが評価が高くなっていることが分かる．また，共起ベクトルを結合するよりも，それぞれの類似度を計算してから加算した方が評価が高くなっている．両共起は性質の異なるものであるから，それぞれの共起で得られた特徴を同じベクトルに組み込むよりも，別々のベクトルとして扱った方がよいと考えられ，この結果はそれを示している．

5. おわりに

コーパスを用いた単語の類似性判別において，新しい文法的共起と考えられる複合語内単語共起を用いる方法を提案した．コーパスより複合語を抽出し，その構成語どうしを共起関係と見なして単語間の類似性判別を行う方法を提案し，実験を行った．具体的には，名詞–名詞で構成される複合語を形態素解析のみで近似的に抽出し，類似性判別における有効性を示した．

また，従来用いられている文法的な共起を近似的に獲得し，類似語検索において評価し，提案した共起を利用した方が評価が高くなるという結果が得られた．さらに，両手法を併用すると，それぞれを単独で利用したときよりも類似性判別に有効であることも示すことができた．

今回の実験では，属性値が 0 でない属性が 1 つしかないような共起ベクトルも少なくなかった．このような単語は，その特徴を十分に獲得できていないといえる．この問題は，コーパスサイズを大きくすることによってある程度は解消できると思われるので，さらに大規模な実験が今後の課題としてあげられる．また，今回は 2 つの共起を併用した手法も試みたが，さらに近接性の共起種との併用についても検討を行う予定である．

謝辞 ダイバージェンスに関して有益な助言をいただいた奈良先端科学技術大学院大学の松本裕治教授，本稿を完成させるにあたり貴重なアドバイスをいただいた NTT コミュニケーション科学基礎研究所の加藤主幹研究員に深く感謝いたします．

参考文献

1) 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，Vol.SIG-ICS 115, pp.9–16 (1999).
2) 佐藤理史：実例に基づく翻訳，情報処理学会研究報告，Vol.33, No.6, pp.673–681 (1992).
3) Grishman, R. and Sterling, J.: Acquisition of Selectional Patterns, *COLING-92*, Vol.2, pp.658–664 (1992).
4) Hindle, D.: Noun Classification from Predicate-Argument Structures, *Proc. ACL*, pp. 268–275 (1990).
5) Schütze, H.: Dimensions of Meaning, *Proc. Supercomputing 92*, pp.787–796 (1992).

6) 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，Vol.94-NL-104, pp.79–86 (1994).
7) 益岡隆志，田窪行則：基礎日本語文法（改訂版），くろしお出版 (1992).
8) 毎日新聞社：CD（毎日新聞 95 版），毎日新聞社.
9) 宮崎正弘：係り受け解析を用いた複合語の自動分割法，情報処理学会論文誌，Vol.25, No.6, pp.970–979 (1984).
10) Salton, G. and McGill, M.: *Introduction to modern information retrieval*, McGraw-Hill (1983).
11) Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, Vol.28, No.1, pp.11–21 (1972).
12) 広中平祐：現代数理科学事典，大阪書籍 (1991).
13) Ikehara, S., Shirai, S., Yokoo, A. and H., N.: Toward an MT System without Pre-Editing-Effects of New Methods in ALT-J/E, *MT Summit '91*, pp.101–106 (1991).

(平成 11 年 8 月 9 日受付)
(平成 12 年 6 月 1 日採録)

稲子　希望（正会員）

昭和 48 年生．平成 10 年九州大学大学院システム情報科学研究科情報理学専攻修士課程修了．同年日本電信電話（株）入社．大規模知識ベースの研究に従事．現在，NTT コミュニケーション科学基礎研究所所員．

笠原　　要（正会員）

昭和 39 年生．平成 3 年東京工業大学大学院総合理工学研究科電子化学専攻修士課程修了．同年日本電信電話（株）入社．知識処理技術，特に大規模知識ベースの研究に従事．現在，コミュニケーション科学基礎研究所研究主任．平成 10 年 11 月より平成 11 年 11 月までスタンフォード大学 CSLI 滞在．平成 10 年人工知能学会奨励賞受賞．

松澤　和光（正会員）

昭和 28 年生．昭和 52 年東京工業大学大学院工学研究科電子工学専攻修士課程修了．同年電電公社入社．以来，フルウェーハシステム，大規模 ROM，ヒューマンインタフェース，知識処理技術の研究に従事．本研究遂行時は，NTT コミュニケーション科学基礎研究所主幹研究員．現在，NTT サービスインテグレーション基盤研究所主幹研究員．平成 10 年人工知能学会奨励賞受賞．電子情報通信学会，人工知能学会，言語処理学会，日本ファジィ学会各会員．