Vol. 38 No. 5 情報処理学会論文誌 May 1997

コーパスを利用した分類語彙表の未登録語義の発見

新 納 浩 幸†

本論文では，分類語彙表に未登録の語義を，コーパスを利用することで発見する手法を提案する．本手法は，まず慣用表現中の語義の特異性を利用してコーパスから慣用表現を抽出する．その誤抽出から分類語彙表に未登録の語義をもつ名詞を推定し，その未登録の語義と類似の語義をもつ名詞を例示する．最終的な未登録語義の決定はこれらの名詞を利用して手作業により行なう．本手法は分類語彙表中の不適切な語義，あるいは利用したコーパスの分野に固有の語義などの発見も可能である．日経新聞記事 5 年分を利用した実験では，177 種の未登録語義を発見できた．

Finding meanings lacking in a Bunrui-goi-hyou entry by using corpora

Hiroyuki Shinnou†,?

In this paper, we propose a method to find meanings lacking in Bunrui-goi-hyou using a corpus. The proposed method first extracts idioms from the corpus by the lexical peculiarity for the noun in an idiom. By incorrect extractions, the noun with a lacking meaning is deduced. Next, nouns which have meanings similar to the lacking meaning are shown. By observing these nouns, the lacking meaning is manually decided. Further, our method can find also the incorrect meaning in Bunrui-goi-hyou, when deciding the lacking meaning. Moreover, we can find also meanings which are conventionally used in the used corpus domain. We experimented with a corpus consisting of 5 years' worth of articles from a Japanese economic newspaper. As a result, we could find 177 types of lacking meanings.

1. はじめに

分類語彙表 [1] は日本を代表するシソーラスであり，多くの研究に利用されている．分類語彙表を拡充，補強することは今後の日本語自然言語処理にとって重要であることは言うまでもない．分類語彙表の補強には，いくつかの方向があるが，分類語彙表は「多義語の語義をすべて配置したのではない」ために [2]，多義性への対処が補強の有力な方向である．ただし手作業で多義語の語義を追加していくことは，個々の語について多義であるかどうかを内省しなくてはならず，容易な作業ではない．またどの程度まで語義を細かく記述するかという問題もある．

本論文では，分類語彙表によって本来は付与されているはずだが，実際には付与されていない語義をもつ名詞を自動的に探し，その語義を付与するための手法を提案する．これによって実際に必要とする不足分の語義を効率的に分類語彙表に追加できる．

注意として，本手法では，

(1) ある名詞に未登録語義があると予測すること
(2) その未登録語義と類似の語義をもつ名詞を例示すること

までを自動的に行ない，(2) の結果を利用して未登録語義を決定する作業は，手作業により行なう．例えば，「声」という名詞には，少なくとも『発生器官から出す音』という語義と『意見』☆という語義があるが，分類語彙表では「声」に対して，1.3031.1 という分類番号だけが付与されている．これは『発生器官から出す音』に対する分類番号であり，『意見』に対する分類番号 1.3060.5 は付与されていない．本手法では，分類語彙表の「声」という名詞には未登録語義があり，その語義と類似の語義をもつ名詞は「意見」「見解」「主張」などであることを示す．この結果から最終的に「声」には分類語彙表には与えられていない『意見』という語義があると判断することは手作業により行なう．

従来，未知の語彙を動的に獲得する研究としては Wilensky の研究がある [3]．そこでは複数の用法をもつ語のある用法が未知のときに，これを他の類似した語の複数の用法から類推する．この類似の語の振舞い

† 茨城大学工学部システム工学科
Faculty of Engineering, Ibaraki University Dept. of Systems Engineering

☆「読者の声」や「庶民の声」などの声の語義．

から対象とする語の振舞いを推測するというアプローチは一般にコーパスのスパース性に対処する場合にも用いられる [4]．つまり見えない部分を見える類似の部分から推測することが基本的なアイデアである．しかしこのアイデアの単純な適用だけでは，ある語に対する未知の語義を発見することは難しい．直観的に明らかなように，ある単語が多義である場合，その単語と類似の単語が同じ多義性をもつことは稀だからである．例えば，「声」と類似の語として，「笑い声」「喚声」「悲鳴」「さえずり」等があるが，これらの語には『意見』という語義はない．

ここでは新納が慣用表現の自動抽出で利用した手法 [5] を応用する．そこではまずコーパスから動詞 v に共起する名詞の集合を作り，次にその名詞の集合から分類語彙表を利用して，類義語となる名詞どうしを取り除き，残った名詞と動詞 v を組み合わせることで，慣用表現を抽出する．もしある名詞 n が多義であり，その中のある語義に対する分類番号が，その名詞に与えられていないとすれば，上記の処理を行なった場合，類義語どうしを取り除いた後に名詞 n が残ることが多い．つまり上記の処理を行ない，ある特定の名詞だけが特異に出現することを手がかりに未登録語義をもつ名詞 n が推測できる．次にその未登録語義と共起する動詞群から名詞 n と類似の名詞を相互情報量を用いて取り出す．それらの名詞が未登録語義と類似の語義をもつ名詞である．それら名詞から登録されていない語義が推測できる．

本論文ではコーパスとして日経新聞5年分（約785万文）を利用して，上記手法の実験を行なった．この実験結果についても報告する．

2. 慣用表現の誤抽出による未登録語義の推定

本手法は以下の5つのステップからなる．

(1) 語義の特異性を利用して慣用表現を自動抽出する．
(2) 1. の誤抽出を利用して，ある名詞 n に未登録語義 g があると予想する．
(3) g と類似の語義をもつ名詞群 $\{n_{ur_i}\}$ を取り出す．
(4) n の登録語義 g' と類似の語義をもつ名詞群 $\{n_{r_i}\}$ を取り出す．
(5) n_{ur_0} がある n_{r_k} と等しく，しかも n_{r_k} の語義が g' とある程度以上類似する場合，$g = g'$ と判断する☆．逆に $g \neq g'$ と判断される g に対する $\{n_{ur_i}\}$ を表示する．

2.1 語義の特異性による慣用表現の抽出

新納は慣用表現を構成する単語の語義の特異性を利用して慣用表現を抽出する手法を提案した [5]．本論文の第1段階の処理として，上記論文の手法を利用して，慣用表現を抽出する．ここでは上記の慣用表現の抽出方法を概説する (図1参照)．

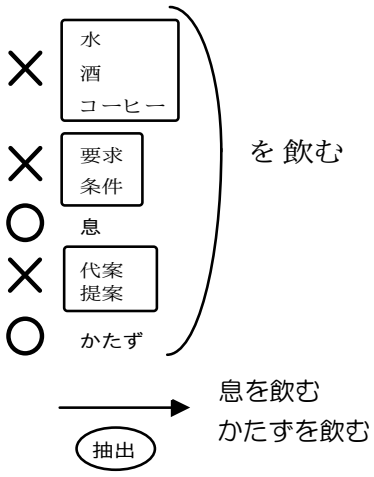

図1 慣用表現の抽出
Fig. 1 Extraction of idioms

まずコーパスから〈名詞〉，を，〈動詞〉の形態の共起データを収集する．例えば「昨夜酒を飲んだ」からは [酒, を, 飲む] を取り出す．次に収集された共起データから動詞を固定し，その動詞と共起する名詞を収集する．例えば固定する動詞が「飲む」ならば，[酒, を, 飲む] や [要求, を, 飲む] などから「酒」や「要求」などが収集できる．こうして集めた名詞の中で類義語どうしを省いてゆく．残った名詞と固定した動詞から慣用表現を抽出する．例えば図1では「かたずを飲む」や「息を飲む」を慣用表現として抽出する．

また類義語の判定として，上記論文では分類語彙表上の分類番号が上位4レベルで一致していれば類義語としたが，ここでは上位5レベルの一致で判定する．

また名詞が k 個の語義をもつ場合，$k/2$ 個以上の語義に対して類義語が見つかれば，その名詞には類義語があると，本論文では表現する．

2.2 誤抽出による未登録語義をもつ名詞の推測

もしある名詞が多義であり，その中の一つの語義に対する分類番号がその名詞に対して，分類語彙表では与えられていないとする．このとき先の慣用表現の抽

☆ g' は厳密には集合であるが，ここでは説明の簡略化のために，点とする．集合と考えても，本手法の処理に変更はない．

出処理を行なった場合，その名詞がその未登録語義の意味で使われていれば，その名詞を構成単語とする表現が慣用表現として誤抽出される可能性が高い．

例えば，固定する動詞として「異にする」をとると，共起する名詞は図 2 で示される．

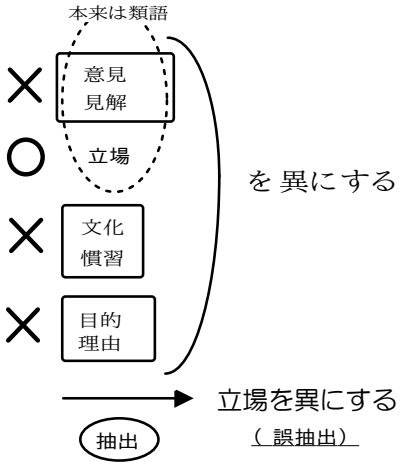

図 2 未登録語義による慣用表現の誤抽出
Fig. 2 Incorrect extraction because of a lacking meaning

ここで本来ならば「立場」には『見地，観点』という語義があるので，「立場」は「意見」や「見解」と類義語となり，「立場を異にする」が慣用表現として抽出されることはない．しかし分類語彙表では「立場」の語義として『立っている場所』しか登録されていないので，先の手法では「立場を異にする」が慣用表現として誤抽出される．

これは「異にする」という動詞だけで生じる誤抽出ではなく，「立場を～する」という表現で，「立場」が『見地，観点』という語義で使われるときに生じる誤抽出である．例えば「立場を明言する」「立場を明示する」「立場を力説する」「立場を無視する」「立場を要約する」なども慣用表現として誤抽出される．

そこで先の手法で慣用表現として抽出された表現から，慣用表現でないものを収集し，それらを名詞によって分類すると，ある名詞に対して，多くの表現が存在する場合がある．このときその名詞には，未登録語義があると推測できる．

慣用表現として抽出された表現から，慣用表現であるかどうかを判定するために，ここでは既存の慣用句辞典 [6] を利用する．この辞典に登録されている表現を慣用表現とする．

2.3 未登録語義と類似の語義をもつ単語の選出

ある名詞 n に未登録語義があると推測した後，次のステップとして，その未登録語義と類似の語義をもつ名詞を共起データから取り出す．

先の操作から $[n, を, v_i]$ の形の共起データが，慣用表現として誤抽出されている．ここで動詞 v_i の集合を 名詞 n の未登録語義を支持する動詞群 と呼ぶことにする．まず，動詞 v_i と名詞 n の相互情報量 MI を以下のように定義する [7]．

$$MI(v_i, n) = \log_2 \frac{\frac{f(v_i,n)}{m}}{\frac{f(v_i)}{m}\frac{f(n)}{m}}$$

ここで，m は共起データの総数，$f(v_i), f(n)$ は共起データ中の v_i および n の総数，$f(v_i, n)$ は，共起データ $[n, を, v_i]$ の頻度である．これは動詞 v_i に対する名詞 n の分布を表す．次に，動詞 v_i から見た，名詞 n と名詞 n_j の類似度 sim を以下の式で定義する．

$$sim(v_i, n, n_j) = \begin{cases} min(|MI(v_i, n)|, |MI(v_i, n_j)|) \\ \quad : MI(v_i, n) * MI(v_i, n_j) > 0 \\ 0 \quad : otherwise \end{cases}$$

最終的に，名詞 n の未登録語義を支持する動詞群全体から見た n と n_j の類似度を以下の式で定義する．

$$SIM(n, n_j) = \sum_i sim(v_i, n, n_j)$$

すべての n_j(ただし $n \neq n_j$) について $SIM(n, n_j)$ を計算し，大きな値をとる名詞を名詞 n の未登録語義と類似の語義をもつ名詞として選出する．ここでは大きな値順に 5 つの名詞

$$U_n = \{n_{ur_0}, n_{ur_1}, n_{ur_2}, n_{ur_3}, n_{ur_4}\}$$

を選出する．

2.4 登録語義との類似性からの選別

上記の処理によって，得られた結果には誤りも多い．なぜなら，慣用表現の誤抽出の原因は未登録語義によるものだけではないからである．その他の原因としては，類似の名詞が分類語彙表に未登録，類似の捉え方の違い，コーパスの不備などがある [5]．

ここでは先に選出した名詞 n の未登録語義と類似の語義をもつ名詞 n_{ur_0} と，名詞 n の登録語義と類似の語義をもつ名詞群 $\{n_{r_i}\}$ とを比較することによって，名詞 n が実際に未登録語義があるかどうかを判断する．もしも名詞 n_{ur_0} が名詞群 $\{n_{r_i}\}$ のある要素 n_{r_k} と等しく，しかも n_{r_k} がある程度以上，名詞 n の登録語義と類似性をもったときに，想定した未登録語義は実際には登録されていると考える (図 3 参照)．

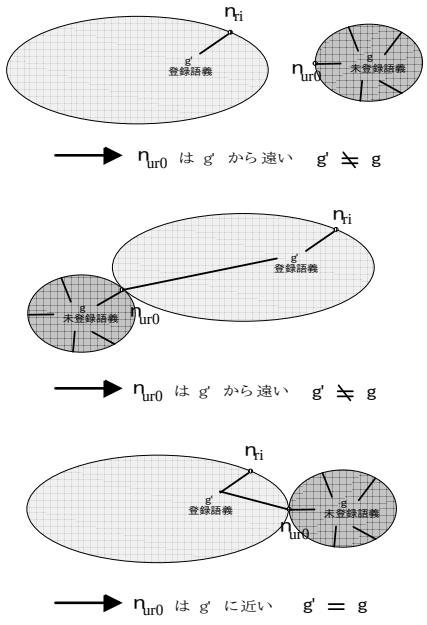

図 3 登録語義と類似との類似性
Fig. 3 Similarity between the entry meaning and the lacking meaning

名詞 n の登録語義と類似の語義をもつ名詞群 $\{n_{r_i}\}$ の作成方法を示す．今，$[n,を,v_i]$ という慣用表現として誤抽出された共起データが得られている．また共起データ全体として $[n,を,V_j]$ も得られている．当然，V_j の集合は先の動詞群 $\{v_i\}$ を完全に含む．そこで V_j の集合から先の動詞群 $\{v_i\}$ を取り除いた動詞の集合 v'_k を作ると，この動詞の集合は名詞 n の登録語義を支持する動詞群となる．

つまり先の相互情報量の定義を用いて，名詞 n の登録語義と類似の語義をもつ名詞の集合を得ることができる．この集合を R_n で表す．

$$R_n = \{n_{r_0}, n_{r_1}, n_{r_2}, \cdots, n_{r_i}\}$$

次に各 n_{r_k} と名詞 n の登録語義との類似性は，

$$\frac{SIM(n, n_{r_k})}{SIM(n, n)}$$

によって測り，この値が 0.2 以上の場合，類似性があるとした．

3. 手作業による未登録語義の決定

上記までの処理により，分類語彙表のある名詞 n には未登録語義があり，その未登録語義と類似の語義をもつ名詞群は U_n であることが導ける．最終的にその未登録語義が何かを決定する作業は手作業により行なう．

この部分が自動的に行なえないのは，n_{ur_0} の語義が名詞 n の未登録語義と類似の関係にあるというだけで，n_{ur_0} の語義が直接，名詞 n の未登録語義となることは少ないからである．これはある動詞の「を」格に位置できる名詞がその動詞の観点から似ているのであって，語義が同じであるとは限らないことに起因する．例えば，「〜を飲む」の名詞部分には「酒」と「水」が入り，「酒」の語義と「水」の語義は類似しているが同じではない．

ただし人間は名詞 n と類似のいくつかの名詞を示されれば，「名詞 n には○○の語義がある」と判断することは比較的容易にできる．例えば，名詞として「チェーン」が選ばれ，その未登録語の語義と類似の語義をもつ名詞として

$U_{チェーン}$ ＝{ アミューズメント施設，ホテル，
レストラン，レジャー施設，食品スーパー }

が例示されたとする☆．この中には「チェーン」に対する未登録語義である『連鎖組織の店舗経営，チェーン店』と同じ語義をもつ名詞はない．しかしこれらの名詞が例示されれば，「チェーン」の語義として『連鎖組織の店舗経営，チェーン店』があることを判断することは容易である．

4. 実験と評価

本手法の有効性を確認するための実験を行なった．コーパスは日本経済新聞'90 年から'94 年の 5 年分（約 785 万文，1 文の平均文字数は 49.0 文字）を利用した．取り出した共起データの総数は約 441 万組である．この中から，頻度が 1 であるもの，動詞の頻度が 20 未満であるものを取り除き，最終的に，総数 3,268,602，種類数 245,951 の共起データを作成した．次に語義の特異性からの慣用表現として抽出した表現は 46,930 種類であった．これらから実際に慣用表現であるものを慣用句辞典 [6] を利用することで取り除き，更に名詞によって分類し，頻度 4 以下の名詞を構成要素としてもつ表現も取り除いた．結果として 2,142 種類の名詞が残った．これらが未登録語義をもつ名詞の候補である．次に各候補の名詞 n に対して，未登録語義と類似の語義をもつ名詞群 U_n と登録語義と類似の語義をもつ名詞群 R_n を作成した．U_n と R_n から実際に未登録語義があると推測できた名詞は 1,110 種類であり，それらに対して U_n を例示した．

評価は U_n の名詞群から，名詞 n にある語義 g があると予想でき，g に対する分類番号が分類語彙表に

☆ この例は次章で述べる実験結果の一部である．また分類語彙表には「チェーン」の語義としては『鎖』だけが登録されている．

登録されておらず，しかも g が広辞苑 [8] には登録されている場合に，有効な抽出とした．例えば，上記の 1,110 種類の名詞には「スキー」という名詞があり，$U_{スキー}$ として以下の名詞が例示された．

$U_{スキー}$ = { 靴下，長靴，靴，運動靴，ワラジ }

これらの名詞から「スキー」には『スキーを行なうときに履く板』という語義があると予想できる．次に広辞苑を調べると，「スキー」の語義として以下の 2 つが登録されている．

(1) 雪の上を歩き，滑べって進むために，それぞれ両足にはく細長い板状の具．

(2) スキーを使用する雪上のスポーツ．

(1) より「スキー」には『スキーを行なうときに履く板』という語義が確かにあることが確認できる．次に，分類語彙表を調べると「スキー」の分類番号は 1 つであり，同じ分類番号をもつ名詞として「ゴルフ」「スケート」「ローラースケート」が挙げられている．つまり分類語彙表には「スキー」の語義として広辞苑での (2) に対応する語義『スポーツとしてのスキー』しか登録されていない．以上より，「スキー」の抽出は有効と判断できる．

上記の作業を行ない，結果として，177 種類の未登録語義を発見できた．その一部を付録の表に示す．参考として付録の表中には分類語彙表に登録されている語義も付記する．

5. 考　　察

5.1 知識の拡大と修正

近年，自然言語処理システムで必要とする知識を，大規模なコーパスから自動的に抽出する研究が盛んに行なわれている [9]．しかし解析で生じる曖昧性（特に多義性）の問題や，必要とする言語現象がコーパス中にはほとんど出現しないというスパース性の問題などから，実際のシステムで利用できる知識を自動獲得することは難しい．

知識を獲得する場合，完全自動は理想であるが，コーパスだけからそれを行なうことは困難である．また必要な知識を完全にゼロの状態から作り出すことは現実的ではない．既に構築した知識が存在するのであれば，それを利用目的に応じて修正，拡大してゆく方が効率的に必要な知識を獲得できるはずである．また実際にシステムで利用されている既存の知識であれば，それを修正，拡大した知識もまた実際のシステムで利用でき，より現実的なアプローチと言える．

例えば Hearst は文書の主題付けを目的に，WordNet の階層構造をカテゴリーの集合に分解し，その集合をコーパスにより作成した Word Space から修正，拡大している [10]．また金田は，英語動詞選択ルールを自動獲得する際にコーパスだけでは十分な量の翻訳事例を収集できないことを指摘し，人手で作成したルールをコーパスによって修正するアプローチを試みている [11]．また Shinnou は一般的な広い範囲の名詞に対するシソーラスをコーパスから得られた共起データだけでは作成できないことを示し，既存のシソーラスをコーパスによって再構築している [12]．

本手法もコーパスを利用して既存知識を修正，拡大する手法の 1 つと位置付けることができる．このために獲得した知識が実用的であるという長所がある．

またここでは未登録語義の発見という既存知識の拡大について述べたが，本手法を応用して既存知識の修正も可能である．本来，既存知識を拡大し人間の記入洩れを補うだけでは，コーパスを使う意味は薄い．利用するコーパスの分野に依存した知識を発見したり，その分野に適した知識に修正する方が好ましい [13]．

本手法では，登録語義を支持する動詞群を調べているが，その際にそのような動詞群の要素が 0 あるいは非常に少ない名詞を発見できる．そのような名詞は分類語彙表に登録されている語義が，そのコーパスの分野ではおおむね利用されないことを示しており，そのような語義の優先順位を低くするといった修正が可能である．例えば，本実験では「在庫」という名詞には登録語義を支持する動詞群がない．分類語彙表では「在庫」の類義語として「在住」「在宅」「在京」「在外」「在天」があげられており，『倉庫にあること』の語義で登録されていることがわかるが，利用したコーパスの分野ではこの語義で利用されることは少なく，分類語彙表では未登録の『保有している商品』として利用される．同様の例として，「ベスト」の場合では，「ベストをつくす」という慣用表現を除けば，分類語彙表に登録されている『最善』の語義で利用されることはなく，分類語彙表では未登録の『チョッキ』の語義で利用される．ただしこれらは「AをBする」という表現に限定した話であり，すべての用法で上記の性質が成立するかどうかは更に調べる必要がある．

また語義の与え方が不適切な部分も，手作業による未登録語義の決定の際に発見でき，この点でも修正が可能である．例えば，「離脱」の未登録語義と類似の名詞として，以下の名詞が例示される．

$U_{離脱}$ = { 停止，脱退，解除，解散，制裁 }

「離脱」には『所属から抜けること』の語義があることはわかるが，これが $U_{離脱}$ 内の「脱退」と類似にならないのはおかしいことに気がつく．分類語彙表の「離

脱」と類似の名詞をみると「離れ」「隔離」「遠心」「離陸」「離水」「離京」「離日」があり，分類語彙表では「離脱」の語義として『離れること』を与えていることがわかる．また広辞苑では「離脱」の語義として『所属から抜けること』だけが与えられている．『離れること』は『所属から抜けること』と類似しているとも考えられるが，筆者には「離脱」の分類語彙表での位置が不適切であると感じる．

また，手作業による未登録語の決定の際に，広辞苑にも登録されていない語義を発見できる．これは，その名詞が利用したコーパスの分野で慣用的にその語義をもつことに対応している．例えば「トップ」の未登録語義と類似の名詞として，以下の名詞が例示される．

$U_{トップ}$ = { 社長，首相，大統領，幹部，代表 }

この「トップ」の語義は『役職や地位が最も高い人』であるが，この語義は分類語彙表にも広辞苑にも登録されておらず，利用したコーパスの分野で慣用的に利用される語義であることがわかる．同様に「組」には『暴力団』，「活字」には『文字』という慣用的な語義があることも発見できる．

5.2 問題点と課題

本手法の本来の目的は，上記したように，知識の拡大と修正である．しかし本手法を抽出手法として捉えることもできる．この場合，正解率と再現率が問題になる．この点から本手法の問題点，課題などを考察する．

まず正解率であるが，最終的に 1,110 種類の名詞 n と集合 U_n の組を抽出し，その中で 177 種類の組が有効な抽出であったので，正解率は 15.9 % である．あまり高い数値とは言えない．これは誤抽出の中から未登録語義をもつ名詞の候補を選出する際の条件が緩いことに起因する．この部分は意図的にそうしている．本手法は収集のツールとしての側面があるため，正解率よりも再現率の方を重視したためである．

次に再現率であるが，本手法は正解を予めすべて取り出しておくことが不可能なために，単純に再現率を調べることができない．またランダムに名詞を一定数取り出し，それらで近似することも難しい．なぜなら，ほとんどの名詞は不正解である（未登録語義をもたない）ため，サンプル数を多くとる必要がある．しかしある名詞が未登録語義をもつかどうかは，単純には，判定できない．例えばその名詞に対して広辞苑に複数の語義が記載されていたとしても，その中でどの語義が分類語彙表に記載されていないかを調べ，しかもその未登録語義が実際にコーパス中で利用されていることを確認しなければならない．またその名詞に対して広辞苑に 1 つの語義しか記載されていないとしても，まさにその 1 つの語義が分類語彙表では未登録語義である可能性もあるため，この場合も上記の調査が必要になる．つまりある一定数以上の正解の名詞を収集することは，非常に困難な作業であることがわかる．そこでここでは 1 つの工夫として，カタカナ表記の名詞に注目する．カタカナ表記の名詞は未登録語義をもつ場合が多いからである．その原因としては，カタカナ表記の名詞は外来語である場合が多く，その概念を日本語化すると，複数に分けられること，また sheet と seat のように異なる単語でもカタカナに直すと同じ表記「シート」になることなどが考えられる．実際に本実験の有効な抽出の 177 種類中 34 種類がカタカナ表記の名詞であった．そこで共起データ中のカタカナ表記の名詞（この数は少ない）だけを対象に，それら各名詞が本論文で設定した未登録語義をもつかどうかを調べ，未登録語義をもつものをあらいだす．次にそれら名詞が抽出されたか，あるいはどの段階で取りこぼしたかを調べることで，本手法の再現率，問題点を考察する．

本実験が対象とした約 24.6 万種類の共起データ中に，分類語彙表に記載されているカタカナ表記の名詞は 493 種類存在した．次にそれら各名詞が未登録語義をもつかどうかを調査した．結果，102 種類の名詞が未登録語義をもっていた．その中で本手法が抽出できた名詞は 34 種類であり，この点から本手法の再現率は 33.3% と考えられる．68 種類の取りこぼしは，以下の処理段階で生じている．

(1) 慣用表現として誤抽出されるという条件で候補として残る名詞は 102 種類中 89 種類．つまり，この処理で 13 種類の名詞を取りこぼしている．
(2) その名詞に対して誤抽出が 5 種以上という条件で候補として残る名詞は先に残された候補の 89 種類中 61 種類．つまりこの処理で 28 種類の名詞を取りこぼしている．
(3) 登録語義との類似性が認められないという条件で最終的に抽出される名詞は先に残された候補の 61 種類中 34 種類．つまりこの処理で 27 種類の名詞を取りこぼしている．

上記 (1) の結果から未登録語義をもつ名詞は慣用表現として誤抽出される可能性が高いことが確認できる．未登録語義をもつが慣用表現として誤抽出されない主な原因は，登録語義と未登録語義が同じ動詞と共起することである．例えば「テスト」は『試験』と『試作機などを実際に試してみること』の 2 つの語義があり，後者は分類語彙表で未登録であるが，この 2 つの

語義はほぼ同じ動詞(例えば「を始める」「を繰り返す」「をパスする」など) と共起する．つまり，用意したコーパスから得た共起データからはこれらの語義を区別できないために抽出できていない．これは広辞苑の語義が分類語彙表の語義よりも詳細化されているからだという捉え方もできる．(2),(3) の部分の取りこぼしは，慣用表現の誤抽出によって得られた候補から無駄な抽出を省くために設定したフィルタが不十分であることを示している．この部分の再現率は，正解の 89 種類中，34 種類を抽出できているので，38.2% である．(3) の処理で取りこぼす主な原因としては，本来，登録語義を支持する動詞であるのに未登録語義を支持する動詞群に入る動詞が存在することである☆．そのような動詞との共起頻度が高い場合，n_{ur_0} が正しく設定できず，適切な取り除きが困難になる．(2),(3) のフィルタ部分を改良していくことが今後の課題といえる．

本手法は抽出システムとして考えると，正解率，再現率に問題があるが，既存知識の拡大や修正のためのツールとして考えれば，利用価値はある．本来，未登録語義を発見すること，あるいは特定の分野で慣用的に使われる語義や，通常の語義としては存在するがその分野では使われない語義を発見することなどは，困難な作業だと思われる．それをある程度自動化できることには，意味があると考える．また本論文が注目した「慣用表現からの誤抽出により，未登録語義をもつ名詞を推定できる」という点は，再現率が 87.2%と考えられ，比較的有効なヒューリスティクスと思われる．

また別件の問題としては，利用するコーパスの大きさがある．本手法は大規模なコーパスの方が効果が現れると思われる．しかしコーパスからの知識獲得では小さなコーパスに応用できる方が望ましい．小さなコーパスでの実現方法も今後の課題である．

6. おわりに

本論文では分類語彙表に未登録の語義をコーパスを利用することで発見する手法を提案した．本手法は，まず慣用表現中の語義の特異性を利用して慣用表現を抽出する．その誤抽出から分類語彙表に未登録の語義をもつ名詞を推定し，その未登録の語義と類似の語義をもつ名詞を例示する．最終的な未登録語義の決定は手作業により行なう．日経新聞記事 5 年分を利用した実験では，177 種の未登録語義を発見できた．

本手法は既存知識をコーパスを利用して修正，拡大する手法の一種である．コーパスによる知識獲得は重要であるが，それを完全自動で行なうことは困難であり，コーパスを利用して知識を修正，拡大するアプローチは有力である．本手法の課題としては無駄な抽出の排除の工夫，小さいコーパスでの実施方法などがあげられる．

謝辞 本実験で利用したコーパスは，日本経済新聞 CD-ROM '90 ～ '94 版から得ています．コーパスの利用を許可して頂いた日本経済新聞社，及び，このコーパスの利用に関して尽力された方々に深く感謝します．

参考文献

1) 国立国語研究所: 分類語彙表, 秀英出版 (1994).
2) 中野洋: 分類語彙表の増補とその利用, 言語処理学会第 1 回年次大会, pp. 141–144 (1995).
3) Wilensky, R.: Extending the Lexicon by Exploiting Subregularities, *COLING-90*, pp. 407–412 (1990).
4) Dagan, I., Marcus, S. and Markovitch, S.: Contextual word similarity and estimation from sparse data, *31th Annual Meeting of the Association for Computational Linguistics*, pp. 164–171 (1993).
5) 新納浩幸, 井佐原均: 語義の特異性を利用した慣用表現の自動抽出, 情報処理学会論文誌, Vol. 36, No. 8, pp. 1845–1854 (1995).
6) 井上宗雄: 例解慣用句辞典, 創拓社 (1994).
7) Hindle, D.: Noun classification from predicate-argument structures, *29th Annual Meeting of the Association for Computational Linguistics*, pp. 268–275 (1990).
8) 新村出編: 広辞苑 第四版, 岩波書店 (1993).
9) 宇津呂武仁, 松本裕治: コーパスを用いた言語知識の獲得, 人工知能学会誌 – 特集「コーパスに基づく音声・自然言語処理」, Vol. 10, No. 2, pp. 197–204 (1995).
10) Hearst, M. and Schutze, H.: Customizing a Lexicon to Better Suit a Computational Task, *Corpus Processing for Lexical Acquisition* (Boguraev, B. and Pustejovsky, J.(eds.)), MIT press, pp. 77–96 (1996).
11) 金田重郎, 秋葉泰弘, 石井恵: 事例に基づく英語動詞選択ルールの修正型学習手法, 言語処理学会第 1 回年次大会, pp. 333–336 (1995).
12) Shinnou, H.: Redefining similarity in a thesaurus by using corpora, *COLING-96*, pp. 1131–1135 (1996).
13) 辻井潤一: 視点の変換 – 言語の理論から設計の理論へ –, 人工知能学会誌 – 特集「自然言語の再構築」, Vol. 11, No. 4, pp. 530–541 (1996).

☆ この原因は慣用表現の誤抽出が未登録語義によるものだけでないことに起因する．

付録　有効な抽出例

名詞	分類語彙表登録の語義	未登録語義
キー	鍵	キーボードのキー
紙面	紙の表紙	新聞などの記事面
定期	一定の期間	定期預金の略
仏	ほとけ	フランスの略
ゴール	目標，終着点	サッカーなどで得点すること
ネット	網	ネットワークの略
覚悟	あきらめること	心構え
大勢	おおよその形勢	多数の人
両国	両方の国	東京都墨田句の両国橋付近の地名
パンチ	なぐること	切符などに穴を開けるハサミの類
ホール	会場，集会所	穴
レバー	食料品としての肝臓	飛行機などの操縦かん
手間	手間賃をもらってする仕事	あることのために費やす時間，労力
出血	体外に血が流れること	犠牲
助け	助けること，助力	必要とするもの
リード	指導，先導	野球などの競技で先行していること
騒動	みだれ騒ぐこと	非常の事態，事変
大手	肩から手先まで	大手筋の略
膜	生物体内の器官を包みへだてる薄い膜	物の表面を覆う薄い物
クラブ	共通目的で集まった人たちの団体	ゴルフなどでボールを打つ棒
コード	太ひも	情報を表現する記号･符合の体系
ハンドル	とって	自動者，自転車の方向操縦用の握り
獲物	うばい取ったもの	漁猟でとった魚など
頭脳	脳	中心となっている人物
マイナス	負数の符合	不足，不利益
火種	火をおこすもとの火	ことを起こすもののきっかけ
待遇	人をあしらいもてなすこと	職場などの地位，給与などのとりあつかい
ダイヤ	ダイヤモンドの略	ダイヤグラムの略
ボタン	衣類等の合わせ目を留めるもの	機械を作動させるための指で押す突起
電源	電気を得るもと	電池やコンセント
土俵	土をつめたたわら	あることが行なわれる場
無理	道理のないこと	強いて行なうこと
涙	涙せんから分泌される液体	人情，思いやり
潮流	潮の流れ	時勢の動き，世間のなりゆき
案内	その場所に未知な人を導いて連れ歩くこと	官庁の先例，内規を書き写した文書
機	組み立ててできた道具	物事を起こすきっかけ
悲劇	死，破滅などに終る劇	不幸なできごと
チップ	こころづけ，祝儀	集積回路の乗った半導体基盤の単位
シート	席，座席	薄板や紙などの1 枚
腕	手の部分	技量，腕前
ブレーキ	物事の進展を，進行をさまたげるもの	車両の速度を抑える装置
土壌	地殻の最上層	物事を生ずる環境，条件
便	たより	あるところまでの交通運輸機関
いす	こしかけ	かん職などの地位
センター	中央，中心	その分野の専門的，総合的施設
胸	体前面の首と腹との間	心
口	動物が体内に食物を摂取する穴状の器官	言葉
頭	動物の脳がある部分	思考力，考え方
柱	直立して上の荷重を支える材	たよりにとなる人，根幹となるもの
壁	家の四方の囲い，または室と室の隔て	障害物
立場	立っているところ	見地，観点，考え方
夢	睡眠中に持つ非現実的な錯覚，幻覚	将来実現したい願い
声	発生器官から出す音	意見

(平成 8 年 9 月 19 日受付)
(平成 9 年 3 月 7 日採録)

新納 浩幸(正会員)

昭和 36 年生. 昭和 60 年東京工業大学理学部情報科学科卒業. 昭和 62 年同大学大学院理工学研究科情報科学専攻修士課程修了. 同年富士ゼロックス, 翌年松下電器を経て, 平成 5 年 4 月より茨城大学工学部システム工学科助手, 現在に至る. 自然言語処理研究に従事. 平成 9 年 4 月東京工業大学より学位取得. 人工知能学会, 言語処理学会, ACL 各会員.