未知語の概念カテゴリ推定法の検討

3B-3

山階正樹　加納英文　小橋史彦

(NTT ヒューマンインタフェース研究所)

1．まえがき

近年，商用データベースばかりでなく，社内データベースの利用が拡大してきている[1]．これらの検索システムにおいて，品質のよい検索結果を得るためには，システム管理者あるいは利用者のニーズを反映してシソーラスを効率よく更新できる技術が重要となる．

本報告では，電子ファイリングシステムのように新たなデータ投入の際にキーワードを付与する場合を考え，そこで概念が未知な用語をシソーラスに登録する作業を支援するため，未知語の概念カテゴリを推定する方法を検討する．

本方法は，未知語とすでにシソーラスに登録されている用語（シソーラス語）の部分一致関係と類似概念の共起関係に着目して未知語の概念を推定するものである．

2．未知概念の推定

ここでは前記の想定に合致する場合として，日経Ｎｅｅｄｓ－Ｉｒデータベースにおけるシソーラス（１６，３３９語）[2]と記事データに付与されたキーワードを用いた検討を行う．各記事に付与されているキーワード（記事キーワード）の中で，シソーラスに登録されていないキーワード（補助キーワード）を未知語とし，それらの用語が属する概念カテゴリをシソーラス語との部分一致関係，キーワードの共起関係を用いて推定する．ここで概念カテゴリとはシソーラス内の独立した一つのトリー構造をいい，その総数は１，４０６である．

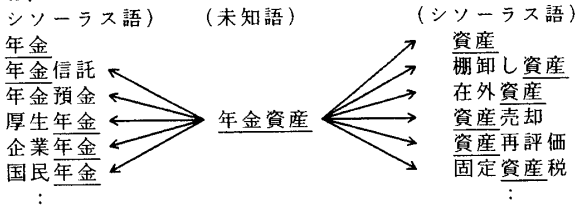
以下では，未知語とシソーラス語の一致パタンを２種類に分類した推定の方法，および，一致要素を持たない場合の処理法について検討する．

2．1　第１種部分一致における概念推定

第１種部分一致とは未知語の部分要素とシソーラス語が完全一致あるいは部分一致する場合を言い，以下の３段階で候補となるカテゴリを絞り込む．

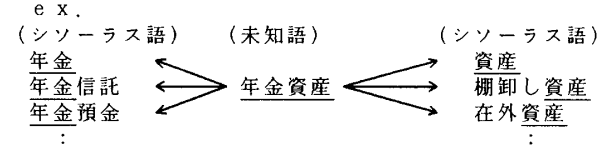
Ｓｔｅｐ１．：未知語の部分要素と完全一致あるいは部分一致するシソーラス語のカテゴリをすべて抽出し，それらを全て候補カテゴリとする．下記の例の場合，未知語の構成要素「年金」，「資産」に完全あるいは部分一致する下記のようなシソーラス語が抽出される．

ｅｘ．



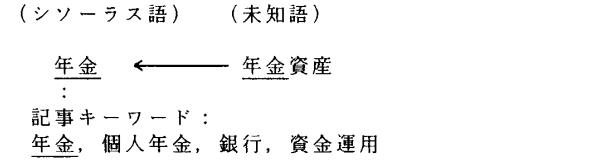
Ｓｔｅｐ２．：Ｓｔｅｐ１で得られたシソーラス語の中で未知語の部分要素と完全一致する，あるいは未知語と語構成パタンが一致する語のみを採用し，それらのカテゴリを候補カテゴリとする．ここでは，「年金」と「年金＋・・」および「資産」と「・・＋資産」のパタンを持つシソーラス語が採用される．

ｅｘ．



Ｓｔｅｐ３．：一つのまとまった文書中では類似した概念の用語が共起する傾向に着目し，未知語に複数のシソーラス語が対応する場合は，シソーラスに登録されている記事キーワードとの共起関係を利用して候補カテゴリを選択する．この例では「年金」のみがこの条件に合致し，「年金資産」の概念として「年金」のカテゴリが選ばれる．

ｅｘ．

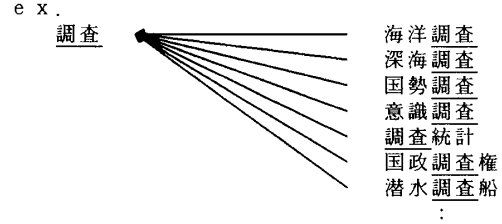


2．2　第２種部分一致における概念推定

第２種部分一致とは未知語がシソーラス語の部分要素と完全一致する場合を言う．この場合には，未知語がシソーラス語よりも大きな概念を示していると考えられるため，以下の処理を行う．

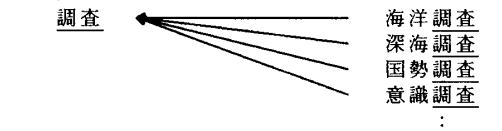
Ｓｔｅｐ１．：未知語のカテゴリとしてそれを部分要素とするシソーラス語の概念カテゴリを与える．

ｅｘ．



Ｓｔｅｐ２．：複合語の部分要素の中では末尾の部分要素がその語の概念を代表する傾向が強いため，Ｓｔｅｐ１．で抽出されたシソーラス語の中で，未知語を末尾の部分要素とする語のカテゴリのみを与える．

ｅｘ．



2．3　一致要素を持たない未知語の処理

シソーラス語と一致要素を持たない未知語については，シソーラスに登録されていない新たな概念である可能性が高いため，孤立語（他の概念と関係を持たない語）として新しいカテゴリを与える．

A stady of Semantic Category Inference for Unknown Words

Masaki YAMASHINA, Hidefumi KANOH, Fumihiko OBASHI

NTT Human Interface Laboratories

3．未知語の概念推定実験

前記の方法を用いて日経新聞記事に付与されている約２５０語の未知語（補助キーワード）をシソーラスに位置付ける実験を行った．表１に示すように第１種の部分一致関係を持つ未知語は６７％，第２種の関係を持つ語は約１７％，一致関係を持たない語が約１６％ある．

表１　部分一致パタンの分布

種別	第１種	第２種	一致要素なし
比率（％）	67．0	16．8	16．2

表２に各々の場合におけるカバー率と平均候補カテゴリ数を処理レベル別に示す．ここで，カバー率とは各処理で得られる候補カテゴリの中に正解カテゴリ（未知語が属することのできるカテゴリ）を含む率を言う．なお，ここでのカテゴリ数は多義語ではそれぞれを別概念として計数するとともに，同一の用語に海外，地域の区別を表すために‘F’，‘L’が付加されたもの（約２０％）も別概念として計数している．

表２　カバー率と平均候補数

種別		第１種	第２種	一致要素なし	総合
Step1	カバー率（％）	89．5	69．7	53．1	80．7
	候補数	19．4	13．5	1．0	15．5
Step2	カバー率（％）	78．2	69．7		73．1
	候補数	12．9	1．3	——	9．0
Step3	カバー率（％）	74．2	——		70．1
	候補数	5．9			4．3

＊第１種，第２種および一致要素なしでは各場合におけるカバー率を，総合ではテストデータ全体に対するカバー率を示す．

（ｉ）第１種部分一致の場合

Ｓｔｅｐ１でのカバー率は８９．５％，平均候補数は１９．４であり，部分一致するシソーラス語を全て候補としても正しい概念を抽出できるのは９０％程度である．ここで正解カテゴリを抽出できないのは以下のような場合である．

①未知語と部分一致した要素がその語の代表概念を示さないシソーラス語のみが存在する場合．
②部分要素が一致しても，他の部分要素によって用語の概念が類似しないシソーラス語のみが存在する場合．
③部分一致した要素が未知語の代表概念を示さない場合．

未知語「故障診断」の場合，シソーラス語「超音波診断装置」，「健康診断」の２語と部分一致する．前者の代表概念は上位語「ＭＥ機器」に対して「装置」であり，後者では「健康診断」の上位語が「健康管理」であるため，「故障診断」はこれらの概念に含まれず，誤ったカテゴリが推定される．

Ｓｔｅｐ２では候補数を約２／３に減少できるが，カバー率が１０％程度低下する．ここで誤りとなるのは未知語「開発競争」に対してシソーラス語「技術開発」が部分一致するが，語構成パタンが異なるために候補から除外するような場合である．この例に示すように，概念の関連性の有無を語構成パタンで捉えようとすることには限界があり，この点を解決するには，複合語での部分要素間の概念間係パタンを記述した知識が必要である．

Ｓｔｅｐ３の処理を行うことによりカバー率の低下を４％に抑えて，候補数を１／２以下に減少できる．ここで誤りとなるのは，記事キーワードに「民間活力」が含まれていると，未知語「民間企業」のカテゴリに「民間活力」のカテゴリを選択するような場合である．

（ii）第２種部分一致の場合

Ｓｔｅｐ１の処理で，約７０％の未知語について正解カテゴリを得ることができ，この場合の平均候補カテゴリ数は１３．５である．ここで正解カテゴリを得られないのは未知語「業種」に対して，「異業種進出」，「異業種間交流」等，未知語がシソーラス語の代表概念に対応していない用語のみがシソーラスに登録されている場合である．これらの未知語を処理することは，ここで述べた簡易な方法では困難であり，他のシーソラスを参照する等の処理が必要である．

Ｓｔｅｐ２の処理を行うことによりカバー率を低下させずに候補を１．３とすることができ，多義を考慮しなければ，未知語のカテゴリを１つの候補に絞ることができる．

（iii）一致要素を持たない場合

一致要素を持たない未知語を孤立語とした場合，約半数が孤立語として新しいカテゴリを与えることができる．孤立語として扱えないものには，シソーラス登録語「住宅」に対する「持ち家」，「貸家」等類義関係を持つ場合，シソーラス登録語「税制」に対する「滞納」，「納税」等あるカテゴリと関連を持ち孤立語とはできない場合などがある．これらの未知語を処理するためには，類義語，関連語の知識を持つことが必要である．

（iv）総合カバー率

Ｓｔｅｐ１でのテストデータ全体に対するカバー率は約８０％であり，平均候補数は１５．５である．候補カテゴリ数を減少させるために，Ｓｔｅｐ２，３の処理を行うことによって，最終的には未知語の約７０％程度について平均４．３個の候補の中に正解カテゴリを絞り込むことができた．

4．むすび

本報告ではシソーラスの更新作業を効率化するため，未知語の概念をシソーラスとの部分一致，さらに，シソーラス語との共起関係に着目して判別する方法を検討し，７０％程度の未知語について平均４．３個の候補の中に正解カテゴリを絞り込むことができた．また，ここで用いた種々の処理の効果と問題点を明らかにした．

今後，さらにカバー率の向上，候補カテゴリ数の減少を図るためには，複合語における概念関係パタンや，類義語・関連語等の知識が必要である．また，第２種部分一致，シソーラス語と一致要素を持たない未知語については異種シソーラスを用いた概念変換[3]等の方法を検討し，本手法との融合を図る．

参考文献
1）データベース白書：データベス振興センター（1987）
2）日経ＮＥＥＤＳ－ＩＲシソーラス：日経新聞社(1985)
3）加納，山階，小橋：シソーラス統合支援法の検討，第３７回情処学会全国大会予稿集