F-001

Webを用いた新概念の自動学習
Automatic Learning of Undefined Concepts in the Concept-Base using Web

岡田信哉† 村上淳哉† 渡部 広一† 河岡 司†
Sinya OKADA Junya MURAKAMI Hirokazu WATABE Tsukasa KAWAOKA

1. はじめに

コンピュータ上で人間のような柔軟で知的な処理を実現するには，人間と同様に常識的な判断を行うことが必要である．そのためには，ある単語から概念を想起し，さらにその概念に関係のある様々な概念を連想できる機能をコンピュータに持たせる必要がある．そこで，語の概念を属性（その語と関連が深い語）の集合で定義する．このように定義された概念を格納する知識ベースを「概念ベース」[1]と呼ぶ．概念ベースは複数の電子化された辞書から構築され，現在約９万語の概念が格納されている．現状の概念ベースの概念の数は我々が日常的に用いている概念の数に比べると，その数もまだまだ不十分である．概念ベースを日常的なものにするためには，新しい概念の追加が必要となってくる．Web 文書には数多くの日常的な概念が存在し，日々情報が更新されているため未定義語(概念ベースに登録されていない語)の自動学習に適した素材である．

本研究では，リアルタイムでの利用を考えているため短期間で，Web 文書から未定義語を取得し，属性を自動的に付加する方式の提案を行う．

2. 概念ベースと関連度

(1)概念ベース

概念ベースは，電子化された複数の辞書から抽出した概念表記や属性によって機械的に構築された知識ベースである．

ある概念 A をその語と関連が強いと考えられる語 a_i と重み w_i の対の集合で定義する．

概念 $A=\{(a_1, w_1), (a_2, w_2), \cdots (a_n, w_n)\}$

ここで，属性 a_i を概念 A の１次属性と呼ぶ．また，属性 a_i も概念ベースに登録されている 1 つの概念である．従って，a_i からも同様に属性を導くことができる．a_i の属性 a_{ij} を概念 A の２次属性と呼ぶ．

(2)関連度

関連度とは概念間の関連の深さを定量化した値である．関連度は，各種提案されているが本研究では「重み付き概念連鎖関連度計算方式」[2]を用いる．関連度は，概念とその属性全てに対して算出した値を用いる．表 1 は概念「自動車」に対する関連度の一例である．

表 1. 関連度の例

基準概念	対象概念	関連度
自動車	車	0.36
	自転車	0.25
	馬	0.09

†同志社大学大学院　工学研究科
Graduate School of Engineering, Doshisha University

3. Webを用いた未定義語の自動学習の手順

未定義語の自動学習手順，即ち概念ベースに Web 文書空間から未定義語と属性を自動的に取得する手法を提案する．

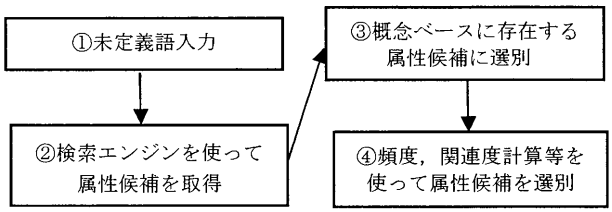

図 1. 未定義語の自動学習手順

4. Web検索エンジンを用いた属性候補取得

属性候補獲得手法は，ロボット検索エンジンの代表である Google を用いて未定義語 C の検索を行い，インターネット上の文書空間から未定義語 C を含んだ 10 件の Web ページを探す．Web ページ 10 件のテキストを形態素解析ソフト茶筌により形態素解析を行う．そこで，得られた名詞を属性候補として，ページにその自立語が出現した頻度情報と頻度順に上位 100 個の属性候補を獲得する．

5. 頻度情報と関連度を用いた属性候補選別手法

Web から獲得した属性候補を未定義語 C の属性とするために，未定義語の意味を表す属性候補だけに選別する必要がある．そこで，属性候補を選別するために，頻度と関連度を用いた属性候補の選別手法(以下，未定義語概念学習手法と呼ぶ)を提案する．

概念に関連の深い属性とは，Web 文書中でその語と関係のある語と共起している可能性が高い．関連度計算を用いて関係のある属性のグループを選別すれば，より概念と関連の深い属性が得られると考えられる．

ある概念 A の頻度順の属性候補 a_1 ～a_{100} があった場合の属性候補のグループ化の様子を図 2 に示す．

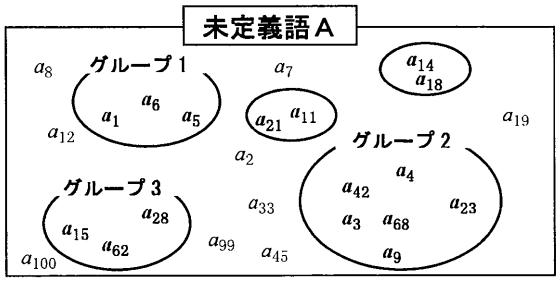

図 2. 属性候補選別の様子

概念と関係のある属性のグループの選び方は，基準の属性候補を選び，その語と他の属性候補とで関連度を算出し，その値が閾値以上となる属性候補を基準となる属性と同一グループとする．基準となる属性候補は，検索

エンジンで取得した属性候補の頻度情報の高い順に選ぶ．人手で作成した未定義語のサンプル 226 個の概念ベースに存在する語だけに選別した属性候補属を頻度順に上位から 10 個，20 個(1~20 位)，30 個(1~30 位)，40 個(1~40 位)，50 個(1~50 位)をそれぞれ目視で適切かどうか比較評価した結果(図 3)より頻度情報が高い方がより概念に関連が深い属性候補と考えられるためである．

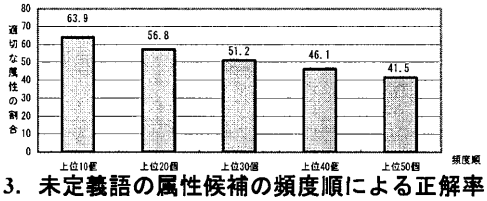

図 3. 未定義語の属性候補の頻度順による正解率

5. 1 未定義語概念学習手法の手順

頻度情報と関連度を用いた属性候補選別の手順は，まず頻度の一番高い属性を基準として，残りの全属性候補と関連度を計算し，関連度が閾値以上となる属性候補と頻度の一番高い属性候補を一つのグループとする．

閾値は，概念ベースおよび関連度計算の精度を機械的に評価する際に用いられる X-ABC 評価セット(人が任意の概念 *X* に対し，同義のように高い関連を示す概念 *A*，ある程度関連のある概念 *B*，関連ない概念 *C* と判断した 4 つの概念よりなるセット)から算出した．評価セットで基準概念となる X と関係のある概念 B の関連度を計算し，全 2370 セットの平均値である 0.183 を閾値として採用した．

ただし，関連度が閾値以上の属性候補が一つも無い場合は，関連度が閾値以上の他の属性候補が出て来るまで，頻度の高い順に基準となる属性を変えていく．取り出したグループ以外の残りの属性候補で，この作業を繰り返して，獲得できるグループを全て取得する．

しかし，これでもまだ未定義語に対してふさわしい属性候補とはいえない．それは，頻度の低い不適切な属性候補が基準の属性となって，その語と関連度閾値以上の属性が選ばれてしまうからである．頻度が低くなれば低くなるほど，適切な属性の数が少なくなるのは図 3 を見れば明らかである．以上の理由から，頻度の低い属性候補を基準として取得した属性のグループは適切な属性である可能性が低いと言える．基準となる属性を頻度順に上位何個までと設定することによってより概念に関連のある属性を選別できると考える．

6. 未定義語概念学習手法の評価

基準となる属性候補数を頻度順に上位 10 個(1~10 件)，15 個(1~15 件),20 個(1~20 件),25 個(1~25 件),30 個(1~30 件)，35 個(1~35 件),40 個(1~40 件)までとし提案手法でそれぞれの平均属性数と適切属性の割合を比較評価した．(図 4)

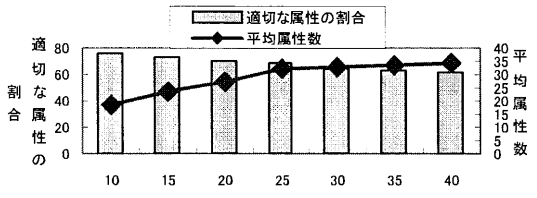

図 4. 基準を設定した未定義語概念学習手法の評価

適切な属性の割合が減少していくに連れて，平均属性数が増加していることが分かる．関連度計算は属性 30 個で打ち切りが行なわれるので平均属性数が 30 個を超えるものが望ましい．基準となる属性を概念ベースに存在する属性の頻度順に上位 25 個までにした場合，グラフの傾き具合から一番適していると言える．平均属性数が約 32.3 個，適切な属性の割合が約 63.2%となった．

7. 未定義語概念学習の考察

頻度順に上位 30 個の属性を選ぶより，未定義語概念学習手法による選別では約 12%も正しい属性の割合が増加した．

しかし，失敗例を見てみるとひとつの傾向が判明した．それは，人名や組織名，地名などの固有名詞概念の属性候補の選別が，固有名詞以外の一般概念（新語，カタカナ語など）と比べると明らかに正しい属性の割合が低かった．未定義語のサンプル 226 個の結果を一般概念 162 個，固有名詞概念 64 個に分けて比べて見ると図 5 のようになった

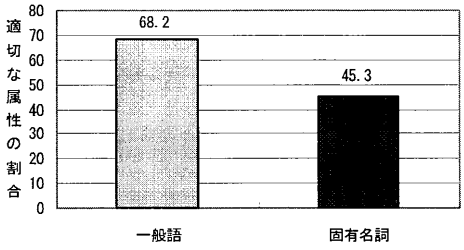

図 5. 一般概念と固有名詞の比較

図 5 を見てもわかるとおり，一般概念に比べ固有名詞概念の正しい属性選別の割合が約 13%低かった．

未定義語概念学習手法は一般概念の属性候補の選別においては有効な手段であるが，固有名詞概念に対してあまり有効であるといえないので，固有名詞概念に対して，新たな手法を考案する必要がある．

8. 固有名詞概念の自動学習

精度が落ちる原因として固有名詞概念特有の問題があると考えられる．Web 空間上に固有名詞概念の意味を表すような文書が一般概念の場合に比べると少ない．そのため概念に関係のない文書から雑多な属性が多く選別されてしまい適切な属性の割合が低くなってしまう．

そこで固有名詞概念の属性特徴を利用した新たな学習手法(固有名詞概念学習手法)を提案する．固有名詞概念には概念を特徴付ける属性(以下特徴語と表す)が存在すると考えられる．固有名詞概念学習手法では特徴語と特徴語に関連ある適切な属性を雑多な属性の中から選別し，より良い属性を獲得することが目標である．

9. 固有名詞概念の属性選別

9. 1 固有名詞概念の属性特徴

固有名詞概念の属性特徴としては概念の特徴を表す語(特徴語)が存在すると考えられる．人物なら職業，肩書きなどといった語である．獲得した属性候補中からこの特徴語と特徴語に関連ある適切な属性を選別し取得することが固有名詞概念を表現するために必要である．本研究では特徴語を持ち，会話において重要な役割をしめる人

物，組織，地名に関する固有名詞概念を対象とした．表 1 に固有名詞概念「松井秀喜」「カネボウ」「今出川」の属性候補中の特徴語を示す．

表 1. 特徴語の例(網掛けのなされた語)

松井秀喜		カネボウ		今出川	
属性	頻度	属性	頻度	属性	頻度
詳細	29	株式会社	18	京都	13
スポーツ	10	化粧	12	休診	9
新人	10	転載	10	当店	7
選手	10	事業	9	上京	7

これら固有名詞概念を特徴付ける属性を獲得することが目的である．が固有名詞概念が人物であるか組織，地名であるかがわからなければ，その概念に適切な属性を選別し獲得することはできない．一方，固有名詞概念がどのカテゴリに属しているかが判別できれば，そのカテゴリに即した属性を獲得することができる．そのため，まずその固有名詞概念がどのようなカテゴリに属するかを判断する必要がある．

今回対象とする固有名詞概念のカテゴリ(分類パターン)を図 6 に示す．大きく人物，組織，地名に分類し，さらに人物の下位に 8 個，組織の下位に 6 個，地名の下位に 5 個のカテゴリ，合計 19 個のカテゴリに分類する．

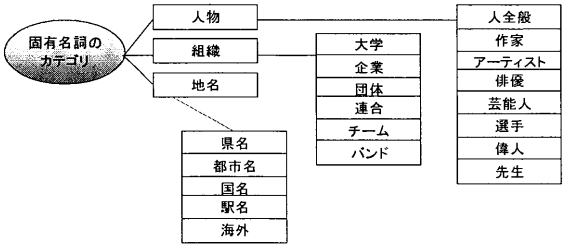

図 6. 固有名詞概念のカテゴリ

9.2 固有名詞概念の分類手法

(1) シソーラスによる固有名詞概念の分類

固有名詞シソーラスとは固有名詞概念の意味的用法を表す意味属性（ノード）の上位－下位関係，全体－部分関係が木構造で示されたものである．このシソーラスを用いて固有名詞概念の分類を行う．例えば固有名詞概念「東芝」の場合，上位ノードを参照すると「企業名」と判断することができる(図 7).

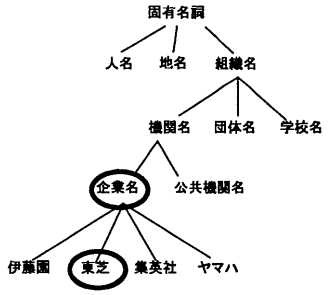

図 7. 固有名詞シソーラス

このように固有名詞概念の上位ノードを参照することによって分類することが可能である．

(2) Web による固有名詞概念の分類

前節ではシソーラスによる固有名詞概念の分類手法について述べた．しかし，シソーラスに格納されている単語数は限りがあり，また日々生まれる新しい固有名詞概念には対応することができない(例えば固有名詞概念「ダスキン」はシソーラスに格納されていないので分類することができない)．そこでもうひとつの方法として Web を用いた固有名詞概念の分類を提案する．ディレクトリ型検索エンジン「Yahoo! Japan」を用い，固有名詞概念に各カテゴリに即したキーワードを追加し，「固有名詞概念 and キーワード」による検索を行い，該当するカテゴリ，ページが存在するか否かで分類する．

概念「ダスキン」では Yahoo 企業カテゴリと一致していることから企業カテゴリに分類することができる(図 8).

Yahoo!カテゴリとの一致 (2件中1～2件目)
- ダスキン (8)
- 害虫駆除 > ダスキンターミニックスカンパニー (2)

図 8．「ダスキン and 企業」で検索した結果

また複数のカテゴリに該当する固有名詞概念も存在する．「和歌山」ように「和歌山県」(県名)や「和歌山大学」(大学)など複数のカテゴリに一致する概念である．そのため一般的な分類を優先して検索し，カテゴリに該当した時点で分類化を終了する．カテゴリに優先順序を設けることによって概念をより一般的なカテゴリに分類する(図 9)．概念「和歌山」での分類化の流れを図 10 に示す(大学に比べ県名の方が一般的であるのでカテゴリ県名に優先して分類).

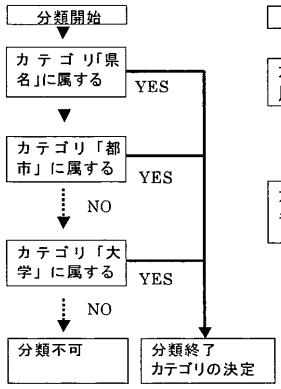

図 9. 分類化の流れ

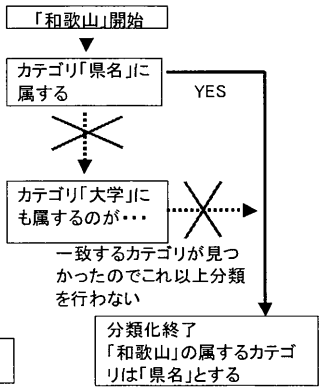

図 10. 和歌山の分類化

10. 固有名詞概念学習手法の実験と評価

10.1. 各カテゴリに与える代表語

固有名詞概念には特徴語(表 1)が存在すると考えられる．この特徴語を獲得するために各カテゴリの中心となり，カテゴリにとってふさわしい単語を人手により知識として与えた(以後代表語と呼ぶ)．この代表語と関連度の高い(関連の深い)属性ほど固有名詞概念にとって重要な語であると判断する．

表 2. 各カテゴリに与える代表語例

カテゴリ名	代表語
先生	人間，作品，功績，絵描・・・
企業	グループ，会社，営業・・・
県，市，駅	市街，敷く，市役所，所在・・・

10.2. TF・IDF[3]と各代表語との関連度を用いた重み付け

情報検索における重み付け手法の1つとして幅広く使われている TF・IDF 手法(TF は属性候補語の頻度，IDF は概念ベース IDF を用いた)に，代表語との関連度を乗じて重み決定する手法を考案した．以下に重み付け手法の流れを示す．

a.Web から固有名詞概念の属性候補を頻度順に獲得．
b.属性候補と分類されたカテゴリの代表語と関連度を計算する．その内最も関連度の高い値を取得する．
c.さらに TF・IDF 値を最大関連度に乗じて重みとする．
"重み=TF × IDF × 最大関連度"
d.この重みを元にソートし，値の高い属性候補から獲得する．
ただし，代表語との関連度が閾値以下の属性は獲得対象外とする．

10.3.実験結果と評価

実験に伴い新たに固有名詞概念のサンプル数を拡大し，概念 302 個(人物 99 個，組織 93 個，地名 110 個)をアンケートで収集し評価を行った．

10.3.1.固有名詞概念分類の結果

適切に分類できた概念，適切なカテゴリに分類できなかったあるいは全く分類できなかった概念の一例を表 3，表 4 に示す．またそれぞれの概念数を人手によって集計した結果が図 11 である．

表 3. 適切分類例

固有名詞概念	分類カテゴリ
桂三枝	人物
渡辺謙	俳優
エプソン	企業
芦屋	都市名

表 4. 不適切分類例

固有名詞概念	分類カテゴリ
ジーコ	企業
ジョーダン	駅名
ローソン	駅名
朝青龍	分類不可

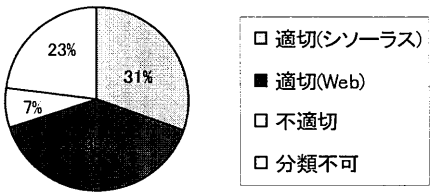

図 11. 固有名詞概念分類の結果

約 7 割の固有名詞概念については適切に分類できている．しかし残りの約 3 割の概念は適切に分類することができなかったか，まったく分類できなかった．

10.3.2 閾値評価実験

代表語と関連度が閾値以下の属性候補は獲得対象外とした．閾値を設けず属性を選別した場合，代表語と全く関係無い語や形態素解析の失敗による雑音属性が多く含まれる．閾値決定のため，関連度 0.10～0.20 の範囲，0.01 刻みで閾値実験を行った．正解率(適切な属性の割合)を図 12 に示す．合わせて獲得した 1 概念あたりの適切属性数の平均値を示す．

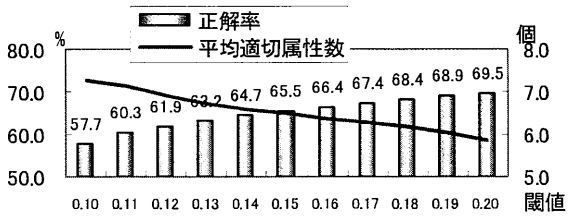

図 12. 各閾値の正解率と平均適切属性数

平均適切属性数は若干ではあるが閾値 0.12～0.13 付近で落ち込んでいる．一方閾値 0.10～0.11 付近では平均適切属性数の落ち込みが比較的穏やかである．正解率に関しては閾値 0.10(57.7%)より閾値 0.11(60.3%)の方が優れている．よって関連度の閾値は 0.11 とする．

10.3.3. 各手法との属性選別との比較

今回何らかのカテゴリに分類できた 229 セットの固有名詞概念について，10.2 節で示した重み付け手法による属性選別を行った．比較のため頻度学習による属性選別(頻度の高いものから 20 件取得)，未定義語概念学習による選別も併せて行った．それぞれの手法によって選別し獲得した属性が適切であるか不適切であるかを目視によって判断し評価した．以下の図 13 に結果を示す．

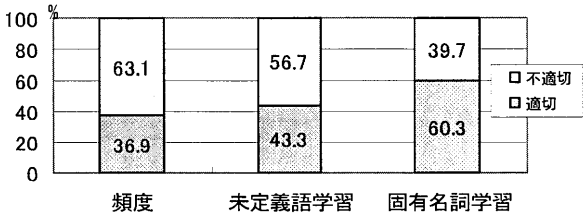

図 13. 属性選別の評価

TF 学習(36.9%)，未定義語学習(43.4%)による属性の選別に対し，固有名詞学習(60.3%)の方がより適切な語を獲得できた．

11. おわりに

本研究では Web を用いた新概念の概念学習手法の提案を目標とした．一般概念については未定義語概念学習手法，固有名詞概念については固有名詞概念学習手法を使い分けることによって全ての語に対応した新概念自動学習が可能となった．今後概念ベースに新概念を追加するためには，適切な重み付け手法を考案することが課題となる．

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った．

参考文献

[1] 広瀬幹規，渡部広一，河岡司：概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法，信学技報，NLC2001-93，pp.109-116，2002
[2]渡部広一，河岡司：常識的判断のための概念間の関連度評価モデル，自然言語処理，Vol.8，No.2, pp.39-54, 2001.
[3]徳永健伸:"言語と計算 5 情報検索と言語処理",東京大学出版会，1999