1V-7

WEBを利用した未知語の概念階層情報の獲得

齊藤　弘樹[†]　岸　義樹[‡]

[†]茨城大学大学院理工学研究科情報工学専攻　　[‡]茨城大学工学部情報工学科

1　はじめに

現在，文章の解析に意味資源を利用した研究が盛んに行われている．しかし，多くの意味資源は静的な性質をもつため，新語等の未知語に対応しておらず，その有効性には限界がある．このため，意味資源に登録されていない未知語に対して概念を自動で獲得し，随時拡張を行う必要がある．

そこで本研究では，意味資源としてEDR電子化辞書を利用し，WEB検索から得られるコンテンツの構文情報と辞書における意味情報との共起性，類似性に着目した，未知語における階層関係を特定する手法を提案し，その有効性を検証する．

2　EDR電子化辞書

EDR電子化辞書 (以下EDR) は，日本語単語辞書や概念辞書によって単語を体系づけた辞書である．本研究では，EDRをシソーラスとして利用する．

シソーラスとしてのEDRは202,797の概念からなり，グラフ構造をもつため1つの概念が複数の上位概念を持つことがある．また，本研究では下位に概念を持つ概念のことをノードとして定義する．

3　提案手法

本手法は名詞・固有名詞を対象とし，未知語を入力することで開始される．手法は以下の4つのステップで構成される．各ステップについて詳細を後述する．

1. 未知語をクエリとしてWEBからの文章の収集
2. 構文パターンの解析による上位単語の抽出
3. 上位単語の概念の類似度から基準概念を特定
4. 基準概念からの階層特定

3.1　文章の収集 (ステップ1)

取得するWEB情報について説明する．まず，未知語を検索クエリにして検索したときのHit件数の上位200件分のレスポンスをYahoo! APIを用いて取得する．

さらに，クエリを2-gramで分割しマッチングした際に，一致率が50%以上となる文章をクエリを含んでいる文章として抽出する．

上位下位	同位
Aは(、)B	A，B
Aとは(、)B	Aと(、)B
Aと(いう｜言う)B	Aや(、)B
AなどのB	

表1: 構文パターン

3.2　上位単語の抽出 (ステップ2)

構文解析器 cabocha を用いてステップ1の文章を解析することにより，全体での使用単語 AW と，構文パターンにマッチする単語の出現頻度を測る．構文パターンは表1に示したものを用い，係り受け関係においてA,Bどちらかが未知語と一致する場合を対象する．また，上位下位の関係のAに未知語が相当した場合，Bは上位語が来る可能性が高いと考えられるため，このパターンに当てはまるものを上位語 HW，それ以外を同位語 CW とする．同様に，未知語が形態素に分解可能である場合も，分解された形態素は上位語である可能性が高いため，HW に補正をかける．

これらのスコアをすべて足し合わせたものが単語スコアとなり，スコア上位20件の単語が上位単語となる．ただし，各スコアがおおよそ均一になるようにそれぞれ重み付け (x, y, z) をする必要がある．

$$\text{単語スコア} = xAW + yHW + zCW$$

3.3　類似度計算による基準概念の獲得 (ステップ3)

ステップ2の結果を基に概念の類似度を計測することで，未知語の概念に近いと考えられる基準概念を獲得する．概念a,bにおけるシソーラスを利用した概念間類似度の計算方式として以下の方法を用いる．

$$\text{類似度} = \frac{a, b \text{ の共通段数}}{a \text{ の段数} + b \text{ の段数}}$$

Acquisition of information unknown words using conceptual hierarchy of the WEB

Saito Hiroki[†], Yoshiki Kishi[‡]
[†‡]Ibaraki University
4-12-1 Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan

段数は最上位概念からの距離を表し，1つ下位の概念となるたびに1づつ増加するものとする．

本研究では，上位単語の組み合せを全通り作り，各概念における類似度の合計をスコアとすることで，各概念の絞り込みを行う．スコアはソートされ4節の評価で利用される．

3.4 基準概念からの階層特定 (ステップ 4)

ステップ3で特定した基準概念を基に，以下の手順で概念の階層を特定する．

基準概念の下位にノードを持たない場合は，基準概念のをそのまま未知語の概念が属するノードとして特定する．

基準概念の下位にノードを持つ場合は，複数のノードから一つを選択する必要があるため，概念の日本語説明文の解析とステップ2の未知語と関係性が高いと考えられる単語 CW を基に下位ノードをスコア付けを行う．手順は以下のようになる．

1. 基準概念の下位に属する CW のノードを CWN とする
2. 基準概念の下位に属する全ノードを取得する
3. 取得したノードの日本語概念説明文を形態素解析し，名詞を取得する
4. 各ノードに含まれる名詞をステップ2の単語スコアの値でスコア化し，これをノードスコア NS とする

以上の NS と CWN に各スコアがおおよそ均一になるように重み付け (w) を加えて、すべて足し合わせたものを概念スコアとする．

$$\text{概念スコア} = NS + wCWN$$

このスコア結果を降順にソートし，最終的な上位10件の単語を4節の評価に利用する．

4 手法の評価

4.1 内容

上記手法を実証するためのシステムを構築し，ステップ3とステップ4における結果を検証した．

4.1.1 テストセット

テスト用のテストセットとして，次の2つを使用する．

1. EDR からランダムに抽出した固有名詞 100 件
2. EDR に存在しない未知語 100 件

EDR に存在しない未知語についてはアンケートにより作成した．

4.1.2 評価手法

ステップ3，4のスコア結果上位10件を基に，概念の適合率と平均逆順位で精度を評価する．適合率ではスコア最上位とテストセット全体との正解率を測る．平均逆順位は，上位10件の正解順位の逆数をとって平均した値であり，適切な概念が上位で抽出できていると値が大きくなる評価値である．

また，ステップ3における正解の指標は，正解単語から最上位まで概念に基準概念が含まれているかで判定する．

4.2 結果および考察

結果をまとめたものを表2, 表3に示す．ステップ4において単語数が減っているのは，ステップ3の正解データだけが対象になっているからである．

ステップ3で約70%，ステップ4で約59%の適合率であり，平均逆順位の値も考慮すると，平均で約80%程度の値を出しているので，候補として挙げれる可能性は高いと言える．現状において，ある程度未知語にふさわしい階層関係が抽出できていると考えられる．

データ	単語数	適合率	平均逆順位
テストセット 1	100	0.72	0.82
テストセット 2	100	0.67	0.79

表 2: ステップ 3 の実行結果

データ	単語数	適合率	平均逆順位
テストセット 1	72	0.57	0.81
テストセット 2	68	0.61	0.84

表 3: ステップ 4 の実行結果

5 まとめ

本研究ではWEBから概念を獲得し，EDR電子化辞書を拡張する手法を提案した．評価実験により階層構造の把握の可能性について確認した．

参考文献

[1] 安藤まや, 関根聡, 石崎俊：定型表現を利用した新聞記事からの下位概念単語の自動抽, 情報処理学会研究報告 2003-NL-157, pp. 77.82.

[2] EDR 電子化辞書 http://www2.nict.go.jp/r/r312/EDR/J/_index.html

[3] Yahoo! デベロッパーネットワーク http://developer.yahoo.co.jp/