言語パターンを用いた検索クエリによる単語間の上位・同位関係の抽出

荻原 由紀恵　山下 達雄　前澤 敏之
ヤフー株式会社
{yogihara,tayamash,tmaezawa}@yahoo-corp.jp

1　はじめに

単語間の関係を集めた語彙データベース (シソーラス) は自然言語処理の基本データとして非常に重要である。このような語彙データベースを構築するためには、単語間の様々な関係性を効率的に取得する技術が必要となる。こうした背景から我々はウェブをコーパスとみなして利用する効率的な関連語抽出手法を開発した。本手法では、シードとなる単語群にウェブ検索 API と複数の言語パターンを適用して関連語候補群を取得したのち、各言語パターンに与えられた重みを元にスコアリングを行い、結果上位の語を関連語として抽出する。

第 2 章ではこれまでの関連語抽出手法の解説を行い、第 3 章で提案手法の詳細を説明する。さらに第 4 章にてこの手法による実験結果を報告する。

2　関連研究

関連語を抽出する方法としては、言語パターンを用いるものがある [2]。近年では、これをウェブ検索と組み合わせて行う手法が数多く提案されている。これは、2005 年から Yahoo!デベロッパーネットワークで提供されているウェブ検索 Web サービス[*1] などのウェブデータをコーパスとして使うための環境が整いつつあることも一因である。これらの手法のうちの一つに、並列を意味する助詞「や」に着目したものがあり、ある単語 A の関連語の候補を得る場合の手順は主に以下のようなものである [6][8]。

1. 「A や」でウェブ検索を行い、「A や B」の形で現れる単語 B を収集
2. 逆の形「B や A」で再度ウェブ検索を行う
3. ２でヒットすれば言語表現として適切である可能性が高いとみなして単語 B を A の関連語とする

「A の C や B の C」の場合、本来得たい A と B ではなく C と B が抽出されてしまう、といった言語構造に由来するノイズ等いくつかの問題があるが、シンプルな手法で十分な抽出性能が得られる。また、文書構造のマークアップ情報から関連語を抽出する手法もある。HTML で記述されたテーブルやリストなどの並列構造を利用して類似語を収集する SEAL[*2] というシステムが知られている [4]。さらに、相互情報量などを用いた統計的手法により関連語を集める方法は、コロケーションの研究として古くからなされている [1][3]。統計的手法は recall は高くても precision が低くなる、つまりはノイズが多くなってしまうという欠点がある。そのため、precision を高めるためには、前述の手法のように言語や文書構造に特化するのが現実的である。

本研究では、日本語特有の言語パターンをベースにウェブ検索 API で出現を調べるという手法で関連語抽出の課題に取り組む。

3　提案手法

語の関連性には多くの種類があるが、本研究では主に「同位関係」及び「上位関係」の二種類を扱う。「上位関係」はその語を含む、より大きな概念を表す語（上位語）への関係を指し、「同位関係」は、共通の上位語を持つような語同士の関係を指す。以降では、ある語の上位関係・同位関係にある語の表す概念を「上位概念」「同位概念」と呼び、それぞれの関係にある語を「上位語」「同位語」と呼ぶ。また、本稿における「関連語」は「上位語」「同位語」を指す意味で使用する。

[*1] Yahoo!デベロッパーネットワーク:ウェブ検索 Web サービス
http://developer.yahoo.co.jp/search/web/V1/webSearch.html

[*2] SEAL - Set Expander for Any Language
http://yeast.ml.cmu.edu/seal/

言語パターンとウェブ検索を用いた典型的な関連語抽出手法 A は下記のような流れとなる。

1. シードとなる単語群を用意する
2. 適用する言語パターンを用意する
3. 各単語（1）に対して、言語パターン（2）を適用し検索クエリを生成する
4. ウェブ検索を行い、検索結果から各単語の関連語候補群を取得する
5. 出現頻度によるランキングを行う
6. ランキング上位の単語を取り出し、関連語とする

これに対し、本研究で提案する関連語抽出手法 B は下記の流れとなる。

1. シードとなる単語群を用意する
2. 適用する言語パターン（複数）を用意する
3. 各単語（1）に対して、各言語パターン（2）を適用し検索クエリを生成する
4. ウェブ検索を行い、検索結果から各単語の関連語候補群を取得する
5. 各言語パターンに与えられた重みを元にランキングを行う
6. ランキング上位の単語を取り出し、関連語とする

手順 1 の関連語抽出のためのシードとなる単語は、主に Wikipedia（日本語版）[*3]から取得した約 13 万語を用いた。手順 2 で使用した言語パターンは、安藤ら [5] による上位・下位語に用いられる日本語の言語パターンの分析などを元にしている。今回はよく使用される言語表現として、8 つの同位表現パターン、1 つの優勢表現パターン (参考的な実験として取得)、7 つの上下関係パターンを利用した (表 1)。手順 3 の検索クエリの生成と手順 4 の検索については、実例を用いて手法を解説する。

検索エンジンに投げるクエリは、各パターンと単語を組み合わせて作成する。パターンによって表記が複数ある場合（漢字とひらがな等）は OR 検索とする。また、複数の形態素を含んだクエリをそのまま投げると、検索エンジン側で分割される場合があるため、クエリは全体をダブルクォーテーションで囲ったものを使用する。例えば、「メロン」という語に対して「A(ににた or に似た)B」というパターンを適用した場合、

[*3] Wikipedia　http://ja.wikipedia.org/

表 1　言語パターン

＜同位概念＞ (8 パターン)	
A や B	A だけでなく B
A も B	A か B
A とか B	A(または or 又は)B
A と B	A(および or 及び)B
＜優勢概念＞ (1 パターン)	
A よりも B	
＜上位概念＞ (7 パターン)	
A のような B	A といった B
A(ににた or に似た)B	A という B
A などの B	A(いがいの or 以外の)B
A(とよばれる or と呼ばれる)B	

”メロンににた” OR ”メロンに似た”

というクエリが作成される。このようなクエリを各単語・パターンごとに作成、それぞれ Yahoo!デベロッパーネットワーク ウェブ検索 Web サービスに投げ、返ってきた結果上位 100 件を対象として同位語及び上位語の候補をシード (単語リスト) と照らし合わせて取得する。シードと検索結果のマッチングは高速な検索を実現するため suffix array[7] による検索手法を用いた。

手順 5 が本研究における提案手法である複数パターン使用による精度向上手法である。同位語の抽出で失敗する理由として、「A の C や B の C」といった並列構造に由来するもの、「ヤフーとかウェブを扱う会社は」などの「とか」が「とかの」と代わりに用いられているような口語表現に由来するものなどがある。このような失敗のパターンのバリエーションは多く、一つ一つを分析して対応するのは困難である。本研究では、同位語パターンの重み付け、反例として上位語の利用を行い、前述の問題に対処する。

4　実験・評価

はじめに、第 3 節の典型的な関連語抽出手法 A による言語パターンごとの関連語抽出精度を求める実験を行った。そしてその結果を踏まえ、複数の言語パターンを統合した関連語抽出手法 B による実験を行い、精度を検証した。

4.1　各言語パターンの抽出精度

シード（単語リスト）のうち 1000 件をランダムに取り出し、これらの単語の同位語・上位語をシード全体

表 2　抽出結果の例

例：マリオブラザーズ

パターン	抽出語
A や B	ドンキーコング:4,ff シリーズ:3, ピカチュウ:2, スーパーマリオ:2, マリオカート:1, ゴルフ:1, セサミストリート:1, アイスクライマー:1, バルーンファイト:1, メトロイド:1, 魔界村:1, スーパーマリオ 3:1, エレベーターアクション:1,nintendo:1
A も B	メイドインワリオ:1, ひっぱりだこ:1, オリジナル:1, セーブ:1
A とか B	アイスクライマー:3, ドクターマリオ:2, バルーンファイト:2, ドンキーコング:2, どうぶつの森:1 怪談:1, エキサイトバイク:1, ドラクエ:1, ロマサガ:1, ドラゴンクエスト:1, 歴史:1, スパルタン x:1, ロックマン 2:1 カービィ:1, 魔法先生ネギま!:1, ゼビウス:1, テトリス:1, グラディウス:1, 忍者じゃじゃ丸くん:1
A と B	テニス:2, ニンテンドー ds:1, 本体:1, ステージ:1, ベースボール:1, ドラえもん:1, 神々:1, 画像:1
A か B	テトリス ds:1, ドラクエ 1:1, アイスクライマー:1
A よりも B	ドクターマリオ:3, 操作性:1, アクション:1, リスク:1, ランキング:1
A のような B	rpg:1, スーパーマリオ:1, お子様:1
A もしくは B	おすぎ:1,bis:1,
A(ににた or に似た)B	アクションゲーム:1
A などの B	任天堂:6, コイン:1, キャラクター:1, ゲーム:1, マリオ:1, アーケードゲーム:1
A といった B	ファミコン:2, ソフト:1, ゲームウォッチ:1, 興味:1
A という B	ゲーム:18, タイトル:1, アクションゲーム:1, ファミコン:1, 言葉:1, あだ名:1,gba:1
A(いがいの or 以外の)B	ゲーム:2, 任天堂:2, おまけ:1

表 3　パターンごとの取得語数及び精度評価

パターン		抽出結果（対象 1000 語）			精度評価（対象 30 語）		
		抽出成功語数	抽出関連語総数	抽出語数平均	語数	正解数	精度
同位	A や B	556	6522	6.52	444	407	0.92
	A も B	494	2940	2.94	188	99	0.53
	A とか B	436	3612	3.61	263	239	0.91
	A と B	637	4465	4.47	267	209	0.78
	A だけでなく B	188	1003	1.00	44	31	0.70
	A か B	296	1122	1.12	70	60	0.86
	A(または or 又は)B	195	899	0.90	49	41	0.84
	A(および or 及び)B	258	1535	1.54	75	58	0.77
優勢	A よりも B	242	1165	1.17	66	41	0.62
上位	A(のような)B	116	269	0.27	20	12	0.60
	A(ににた or に似た)B	106	272	0.27	31	24	0.77
	A(などの)B	389	2974	2.97	187	132	0.71
	A(とよばれる or と呼ばれる)B	152	489	0.49	9	2	0.22
	A(といった)B	275	1208	1.21	79	55	0.70
	A(という)B	582	3189	3.19	211	150	0.71
	A(いがいの or 以外の)B	263	1494	1.49	65	42	0.65
合計		-			2068	1602	0.77

から取得した。これは前述した 2 つの手法のうち、関連語抽出手法 A に従った処理である。さらにそのうち 30 件をピックアップし、人手による精度評価を行った。精度評価は、抽出された同位語・上位語がそれぞれ対象となる語と関連しているかどうかを人間が見て判断し、正解もしくは誤りとした。

実際の取得結果の例を表 2 に、1000 件のデータに対する取得結果及び 30 件の精度評価結果を表 3 に表す。結果として全体で約 7 割以上の精度を得られた。

同位語の取得パターンは、全体的に高い精度を得た。特に「A や B」パターンは取得語数も多く、精度も優れている。また、「A とか B」パターンは「A や B」パターンほど取得語数は多くないが精度が高く、同位語取得パターンとしては「A や B」同様に利用価値が高い。逆に「A も B」パターンは精度が低いが、これは並列要素が接続されるとは限らないためであろう。この問題に対しては「A も B も」などさらに続く語を限定するなどすれば、より正確な取得が期待できる。上位語に関してはいずれも精度は 6〜7 割という結果が得られたが、「A と呼ばれる B」パターンの精度が著しく低かった。このパターンは取得語数そのものが少ないことに加え、上位概念を表すパターンとしても（他のパターンに比べて）使用頻度が低いことが伺える。同位語・上位語共にパターンによって取れる語に違いがあり、これらのパターンを組み合わせることでより多くの語を収集できることもわかった。同位語として取得した語の集合と、上位語の集合とを比較すると、それぞれに固有な語が取れていることが多いのも特徴である。

4.2　複数パターン使用による精度向上

同位語集合と上位語集合との比較では、それぞれに固有な語が取れることが多かった。これは、同位語集合と上位語集合は、互いに排他的な関係にあるためであり、逆に両方に含まれる語はノイズとして判断できる。さらに、パターンによって精度に違いが生じることから、このような信頼性の値を重みとしてスコアリングすることで、より信頼性の高いパターンによる結果を優先させ、同位語・上位語を絞ることができる。そこでこれらを踏まえ、関連語抽出手法 B(第 3 節) に従っ

表 4　スコアリングの例（括弧内はランキング順位）

対象語	抽出語	同位語頻度	同位語スコア	同位語スコア 2	上位語頻度	上位語スコア	上位語スコア 2
ドングリ	木の実	16 件 (1)	13.3075(1)	-11.4774(75)	52 件 (1)	24.7849(1)	11.4774(1)
アプリコット	プルーン	4 件 (7)	3.3307(8)	2.6107(9)	2 件 (10)	0.72(14)	-2.617(97)
サマンサタバサ	財布	3 件 (9)	2.3012(9)	-0.2193(48)	5 件 (3)	2.5205(3)	0.2193(23)

たスコアリング計算を行い、効果を調査した。

スコアリングにはまず、先の実験結果による精度の２乗値を重みに利用し、同位語・上位語それぞれのスコアを単語ごとに算出した結果を同位語スコア・上位語スコアとする。さらに、同位語・上位語の影響をそれぞれに反映させるため、お互いのスコアを引いた値を同位語スコア 2・上位語スコア 2 とする (表 5)。

表 5　スコアリング計算

同位語スコア =
　Σ(同位語パターン i の重み w_i × 頻度 x_i)
上位語スコア =
　Σ(上位語パターン i の重み w_i × 頻度 x_i)
（＊パターン i の重み w_i = 手作業で集計した精度^2）

同位語スコア 2 = 同位語スコア − 上位語スコア
上位語スコア 2 = 上位語スコア − 同位語スコア

先の精度評価に用いた 30 件を使用し、同位語頻度によるランキング結果と、同位語スコア 2 によるランキング結果のそれぞれ上位 10 件を取り出した結果について、同位語である場合を正解として目視評価した。その結果、頻度による精度 76.0 %、スコア 2 による精度 77.1 %と全体的な評価については大きな変化は見られなかったものの、各スコアをそれぞれに比較することで精度向上を確認できた例が幾つかあった (表 4)。「ドングリ」と「アプリコット」の例は、同位語・上位語頻度による順位と同位語・上位語スコア 2 の順位で比較すると、ランキング変動の小さい方を正しい集合と判断できる。「サマンサタバサ」の例ではいずれのスコアも大きく変動するため、同位語・上位語という位置づけでは分類し難い語をどちらでもないと扱える。このように、同位語・上位語パターンを同時に用いたスコアリングを行い、ランキング上位の単語を取得することによって精度向上を見込めることが確認できた。

5　まとめ

本研究では、同位語・上位語を Web テキストから抽出する実験を行った。その結果、大方のパターンで一定の精度を確認できたが、パターンによってその信頼性が異なることがわかった。さらに、それらの信頼性を基にしたスコアリングを行い同位語集合と上位語集合を同時に利用することで、精度向上を図る手法を提案、実験した。今回は比較的アドホックな形での実験であったため、今後は抽出データの実サービスへの展開を通じて、スコアリング手法の見直しなどにより抽出手法の精度向上をはかっていきたい。

参考文献

[1] K. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics, 16(1),22-29*, 1990.

[2] M. Hearst. Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th International Conference on Computational Linguistics*, 1992.

[3] Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. Mit Pr, 1999.

[4] Richard C. Wang and William W. Cohen. Language-independent set expansion of named entities using the web. *IEEE International Conference on Data Mining (ICDM 2007)*, 2007.

[5] 安藤まや, 関根聡. 上位語・下位語を含む連体修飾表現の言語的分析. 言語処理学会第 10 回年次大会発表論文集, pp.205-208, 2004.

[6] 相澤彰子, 中渡瀬秀一. 係り受け関係を利用した類語・例文辞書構築法と大規模コーパスへの適用. 第 20 回人工知能学会全国大会 2E1-5, 2006.

[7] 山下達雄. 用語解説: Suffix array. 人工知能学会誌, 15(6):1142, 2000.

[8] 大島裕明, 小山聡, 田中克己. サーチエンジンのインデックスを利用した同位語検索と同位語コンテキストの発見. 情報処理学会研究報告 pp.161-168, 2006.