3B1-4

# 検索エンジンを用いた関連語の自動抽出

Automatic Extraction of Related Terms Using Web Search Engine

渡部 啓吾[*1] Danushka Bollegala[*1] 松尾 豊[*2] 石塚 満[*1]
Keigo Watanabe Yutaka Matsuo Mitsuru Ishizuka

[*1]東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

[*2]東京大学大学院 工学系研究科
School of Engineering, The University of Tokyo

Semantic lexicon such as Roget's Thesaurus or WordNet provide useful knowledge for natural language processing applications, but take long time to build, maintain, and extend. Earlier works have used corpora which are newspaper and magazine articles and can't have dealt with a word newly appeared or whose meaning is changed. Motivated by this problem, we propose method for automatic extraction of related terms using the web as a corpus. The proposed method exploits snippets returned by a web search engine which is essential tool to extract data from the web.

## 1. はじめに

もし日常的にある言葉に関連する語を知りたいと思った場合，専門の辞書を引いて目的の単語を探し，その関連語を調べるというのが普通であろう．そのような辞書はシソーラスと呼ばれ，WordNet や Roget's Thesaurus など Web 上にも多く存在している．しかし，これらは長い年月をかけて人手で作られたものが殆どで，多くの手間がかかっており，それを拡張したり維持していくためには大きなコストがかかってしまう．そこで以前から，新聞記事や学術文書などのコーパスを利用して，自動的に関連語を抽出したり，自動的にシソーラスを構築することを目的とした研究がなされてきた [Lin 98, Hearst 92, Pasca 07, Pantel 06, Snow 04, 榊 07].

Hearst は人手で $NP_0$ *such as* $\{NP_1, NP_2 \ldots, (and \mid or)\} NP_n$ などのパターンをいくつか作り，それを利用して下位語を抽出する手法を提案した [Hearst 92]．また，Snow らは上位下位関係にある少量の正解データが与えられたときに，コーパスからその語間の依存関係を自動で学習し，ある語ペアが上位下位関係かどうかの判定に利用した [Snow 04]．これらの研究において用いられたコーパスは，新聞や雑誌の記事，あるいは百科事典などで，ある程度整形された文章を対象としていたため，高い精度が得られていたが，新出語や固有名詞，意味の変化などに対応することが出来ないという欠点もあった．

この欠点を補うため，近年ではコーパスとして Web 上の文書を利用する手法が提案されている．Web をコーパスとして利用すると，新出語や固有名詞，意味の変化に対応できるだけでなく，自らコーパスを用意する必要がないというメリットがある．そのためには，Web 上の 100 億以上にも上る文書に効率よくアクセスする必要があり，検索エンジンの利用が重要となる．勿論，Web 上には新聞記事や学術文書などのような整形された文書ばかりではないため，得られた情報をどのように処理するかが重要となる．

関連性の評価において Web をコーパスとして用いるとき，一般的に利用されるのは検索エンジンにおけるヒット件数である．ヒット件数を利用すると，語の出現確率を調べることが出来，2 語が共起する確率を利用して関連度を計る研究などが行われてきた [榊 07]．しかし，ヒット件数だけでは語がどのように出現したかという情報は得られないため，Bollegala らはヒット件数に加え，検索結果のスニペット（検索クエリの生じた前後の文脈）を用いることで，語が生じたパターンを抽出し，関連度を計る手法を提案した [Bollegala 07]．関連語や語義の抽出といったタスクにおいても Web をコーパスとして利用する研究が行われている．Cimiano らは検索エンジンのスニペットを利用して語を定義する 4 つの要素（Qualia Structures）を抽出し [Cimiano 07]，Pantel らは関連語を抽出するために新聞記事や雑誌などの整形されたコーパスを主に利用し，精度を高めるために Web をコーパスとして利用した [Pantel 06]．また，Pasca は検索エンジンの検索履歴を利用することで，クラスの属性を抽出する手法を提案している [Pasca 07].

本研究では，Bollegala が関連度の評価に用いたパターン生成法を，関連語の抽出タスクに適用する．すなわち，あらかじめ既存の辞書など（例:WordNet，Roget's Thesaurus）を用いて関係 R（例:同義－同義，上位－下位）の正解データを作り，学習させておくことで，ある単語 W（例:dog）が与えられたとき，W と関係 R に当たる語（例:hound，animal）の順位付けされたリストを獲得する手法を提案する．提案手法は Web をコーパスとして用いることで，新出語や固有名詞，意味の変化に対しても対応できることが特徴である．

## 2. 手法

本手法は大きく分けて，パターンの学習，関連語の候補の抽出，候補の順位付けの 3 段階にわけられる．全体の流れを図 1 に示す.

**Step1: $\chi^2$ 値によるパターンの選別** このパターン選別の手法は Bollegala らの手法 [Bollegala 07] と同様である．本研究では以下 Google を検索エンジンとして用いている．まず学習データとして，正解データ（R という関係にある語ペア）と不正解データ（R という関係にない語ペア）のリストを用意し，2 つの語の間に'*'（どのような語でもマッチする）を 1〜3 個入れてクエリとして検索を行う．得られたスニペットから，2 つの語が共起するときに現れ

連絡先: 渡部啓吾, 東京大学大学院 情報理工学系研究科, 東京都文京区本郷 7-3-1,watanabe@mi.ci.i.u-tokyo.ac.jp

表 1: 重み付けされたパターンリスト（F 値 TOP5）

| 同義語抽出 X,Y:同義語 | | | 上位語抽出 X:下位語 | | | 下位語抽出 X:上位語 | | |
|---|---|---|---|---|---|---|---|---|
| パターン | F 値 | $\chi^2$ 値 | パターン | F 値 | $\chi^2$ 値 | パターン | F 値 | $\chi^2$ 値 |
| synset X Y | 0.2736 | 68.4 | X is a * who | 0.2996 | 28.2 | * or other X | 0.1569 | 73.6 |
| X synonyms Y | 0.2051 | 50.8 | X by * of | 0.1377 | 20.2 | a * is a X | 0.1346 | 73.9 |
| syn X Y | 0.1876 | 64.7 | X a * who | 0.1259 | 44.2 | X such as a * | 0.1241 | 42.6 |
| an X or Y | 0.1383 | 29.9 | X or other * | 0.1082 | 73.6 | X or * as | 0.1011 | 15.7 |
| X or Y a | 0.1128 | 20.9 | a * or X | 0.1031 | 43.3 | * n the X of | 0.0933 | 17.6 |

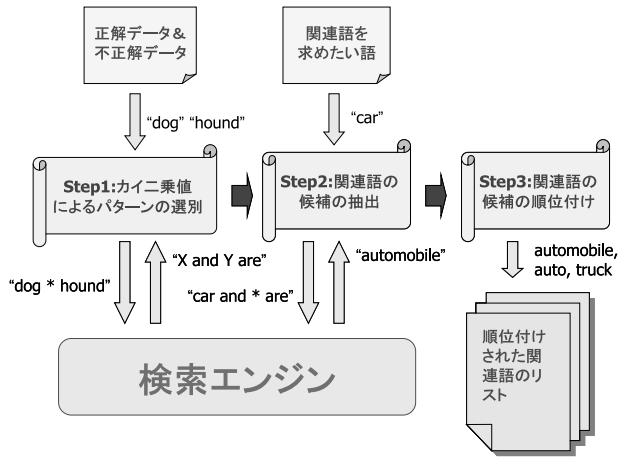


図 1: 全体の流れ

る n-gram（n=2〜10）のパターンを抽出する．そして，抽出したパターンそれぞれについて $\chi^2$ 値を計算し，パターンを選別する．

**Step2: 関連語の候補の抽出** まず Step1 によって得られたパターンの'X' と'Y' の部分を，関連語を求めたい語，'$*$' に置き換えてクエリを作成する．例えば，"dog"の関連語を求めたければ，"dog and $*$ are"，"$*$ and dog are"のような 2 種類のクエリが得られる．そのクエリを用いて検索し，得られたスニペットから，パターンの'$*$' にあたる部分にある 1 語または 2 語からなる語を関連語の候補とする．候補はパターンごとの出現頻度の情報を持つ．

**Step3: 関連語の候補の順位付け** Step2 で得られた関連語の候補は膨大な量であり，不要な語も含まれているため，Step2 で得られた候補語ごとのパターン別の出現頻度を利用して，その評価値を求め順位付けを行い，順位付けされた関連語のリストを獲得する．

## 3. 評価

本実験では名詞を対象として関連語の抽出を行う．名詞の関係にはいくつか種類があるが，一般的な，同義－同義（意味がほぼ同じ．例:dog - hound）と上位－下位（包括的．例:animal - dog）の関係を対象として関連語の抽出を行う．ここで WordNet を用いて，それぞれの関係にある語ペアを，それぞれ 1000 個ずつ抽出して正解データとし，それと同時にランダムに抽出した 1000 個の語ペア（関係が登録されていない語ペア）を不正解データとした．これらの正解・不正解データを元に Step1 を実行し，得られたパターンをその $\chi^2$ 値によって上位 100 件を選別する．そのうちの一部を表 1 に示す．

### 3.1 パターンの重み付け

実際に関連語が得られたかどうかをパターンの重みとしてフィードバックさせるため，パターンを F 値によって重み付けする．ここでは {asylum, boy, car, coast, gem, journey, magician, noon, stove, tool} という 10 語に対して関連語の抽出を行い，WordNet に登録されている正解データと比較した．このようにして得られた F 値によって重み付けされたパターン上位 5 件を表 1 に示す．

### 3.2 関連語の候補の順位付けに用いる関連度の指標

Step3 の順位付けにおいて，Step2 で得た候補語（$c$）ごとのパターンの出現頻度（$c_v$）を利用して，候補となる語が出現したパターンの種類数（式 1），パターンの F 値による重みと出現頻度の積の総和（式 2）という 2 つの指標を利用する．ここで $v$ はあるパターンとする．

$$\mathrm{NumofPat}(c) = |\{v|c_v > 0\}| \tag{1}$$

$$\mathrm{RankScore}(c) = \sum_v Fweight_v \times c_v \tag{2}$$

提案した 2 つの指標を評価するために，Bollegala らが用いた検索エンジンのヒット件数を利用した関連度の指標（WebJaccard，WebOverlap，WebDice，WebPMI）による順位付けも行う [Bollegala 07]．この 6 つの指標を評価するため，同義語を豊富に持つ一般的な語として"magician"を選び，Roget's Thesaurus に登録されている 88 語を正解データとし，指標別に取得したデータの個数に対する F 値の推移を調べた（図 2）．この図を見ると，パターンの出現頻度を用いた関連度の指標である NumofPat，RankScore が，単純なヒット件数を利用した他の 4 つの指標よりも常に有効であることが確認できる．以降の実験ではこの 2 つの指標を用いる．

### 3.3 既存のシソーラスとの比較

提案手法を評価するために既存のシソーラスとの比較を行う．ここでは {cord, forest, fruit, glass, slave} という一般的な 5 語を対象とし，正解データは Roget's Thesaurus（同義語），WordNet（上位語，下位語）に関連語として登録されているものとした．図 3〜5 にそれぞれの抽出方法での，取得した語数に対する F 値の推移を示す．

### 3.4 関連研究との比較

Lin は，構文解析された新聞記事などの文章を利用して，自動的にシソーラスを構築する研究を行っている [Lin 98]．この研究は，ある語に対して，その語の依存関係が似ている度合いを評価して順位付けしたものである．ここでは節 3.3 の 5 単語

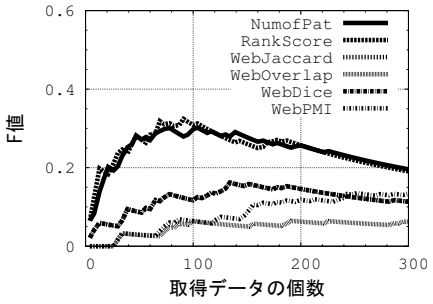


図 2: Step3 の指標の比較

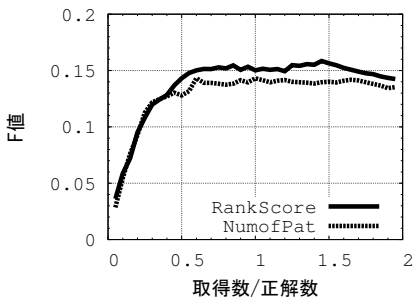


図 3: 同義語抽出

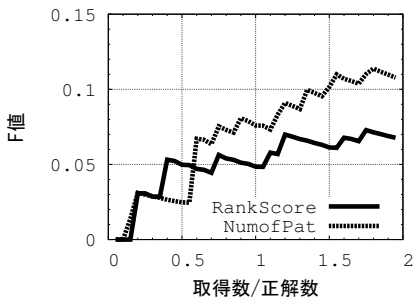


図 4: 上位語抽出

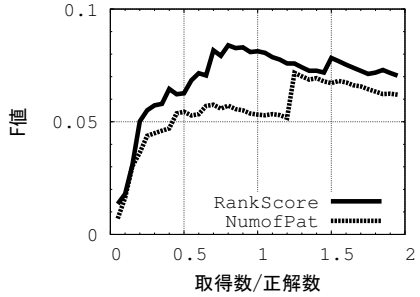


図 5: 下位語抽出

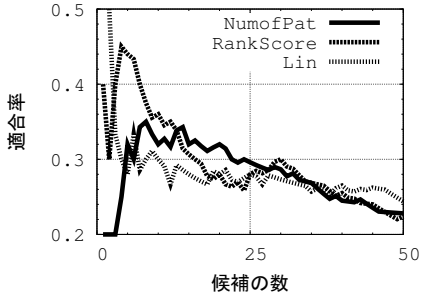


図 6: Lin との比較

に対して，提案手法で抽出した同義語上位 50 件と，Lin のシソーラスに登録されている同義語上位 50 件について平均の適合率で比較した．その結果（図 6）を見るとほぼ同程度の適合率となっており，提案手法では Lin の手法のように，予めコーパスを用意したり構文解析をする必要が無いので有意な結果と言える．

## 4. まとめ

本研究では，検索エンジンのスニペットを利用することで，ある関係にある語ペアの正解データさえ用意すれば，求めたい語に対して様々な関係にある語の順位付けされたリストを獲得する手法を提案した．また，抽出した関連語をまとめることで，シソーラスの自動構築への応用も可能だと考えられる．ただし，既存のシソーラスとの比較で，特に上位下位語抽出において F 値が低くなってしまっているなど改善の余地がある．これは特に Step3 が原因であると考えられるため，順位付けの方法を工夫するため，タグ付けを行ったり，スニペットだけではなく検索にヒットしたページも利用してパターンを抽出する手法が考えられる．精度を上げる方法としては，様々な特徴を SVM などを用いて学習する，正解データの質・量を改善する，などが考えられる．

## 参考文献

[Bollegala 07] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th International World Wide Web Conference (WWW-07)*, pages 757-766, 2007

[Lin 98] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 19th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-98)*, pages 768-774, 1998

[Hearst 92] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 539-545, 1992

[Pasca 07] M. Pasca. Organizing and searching the World Wide Web of facts - step two: harnessing the wisdom of the crowds. In *Proceedings of the 16th International World Wide Web Conference (WWW-07)*, pages 101-110, 2007

[Cimiano 07] P. Cimiano and J. Wenderoth. Automatic acquisition of ranked qualia structures from the web. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 888-895, 2007

[Pantel 06] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-06)*, pages 113-120, 2006

[Snow 04] R. Snow, D. Jurafsky, and A. Ng. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*, 2004.

[榊 07] 榊剛史, 松尾豊, 内山幸樹, 石塚満. Web 上の情報を用いた関連語のシソーラス構築について. 自然言語処理, Vol. 14, Number 2, pages 3-31, 2007.