検索ログによる拡張固有表現辞書の整備

関根聡
ニューヨーク大学

鈴木久美
マイクロソフト・リサーチ

1. はじめに

幅広い分野での応用を目的に 200 種類の階層的カテゴリを持つ拡張固有表現を構築している (Sekine and Nobata 04)。現在までに、様々な知識源を利用して日本語､英語共に約 20～30 万語の辞書を作成し、その辞書とルールを用いたタガーを構築した。タガーは辞書に大きく依存しており、辞書の品質がタガーの精度を左右する。辞書は一般に入手可能な辞書や百科辞典および WEB にある知識源などを利用して構築したが、精度の点でまだ満足できるものとはなっていない。それには 2 点の問題がある。一つは電話帳的な幅広い知識を使うと、本来そのカテゴリの名前として使われることが少ないようなものもそのカテゴリの項目として入ってしまうという適合率的な問題であり、もう一つは、辞書項目の網羅に関する再現率的な問題である。これらの問題を人手で網羅的かつ均一的に見ることは非常に困難である。

この問題に対し、検索エンジンの検索ログを使い、拡張固有表現辞書の整備を行う方法を提案する。基本的な考え方としては、各固有表現のカテゴリには、そのカテゴリを特徴付けるような検索ログのコンテキストが存在するであろうという仮説を前提とする。手法的には、各固有表現カテゴリの項目とよく共起するような特徴的なコンテキストを統計的手法を用いて同定し、そのコンテキストとあまり共起しないような辞書項目はそのカテゴリの項目である可能性が低いであろうという仮定によって適合率の問題を解決し、辞書項目には挙がっていないが、特徴コンテキストと良く共起するような表現があった場合には、そのカテゴリの辞書項目の候補と考えることにより再現率の問題を解決する。

この方法は、検索ログではなく普通のコーパスでも実現可能である。しかし、以下の３つの点から、検索ログを使う長所があると考えられる。１）新聞記事のようなきちんと書かれたコーパスでは雑音は少ないが規模が小さい。逆に WEB コーパスでは規模は大きいが雑音が多い。それに対して、検索ログでは雑音は比較的少なく規模も大きい。２）普通のコーパスではコンテキストとしてどの範囲を抽出すればいいか確定することが困難であるが、検索ログでは表現がシンプルでありコンテキストの同定が容易である。３）検索ログは人が直接検索要求を行った履歴であり、例えば情報要求アプリケーションである質問応答、情報検索、情報抽出などを考えたときには、作成された知識と応用の間に親和性が期待できる。

2. 検索ログ

今回の実験では、MSN の検索エンジンに対して投げられた 14 億の検索ログを基にした。英語の拡張固有表現辞書の整備を目標としているため、14 億の内、英語のアルファベットかシンボルのみで構成された 2 単語以上の検索ログにのみ限った 7.5 億の検索ログを実験に使用した。これは、平均的に新聞の１文から４つの固有表現コンテキストを取れると仮定した場合に約 100 年分の新聞記事に相当する。検索ログは図１のようなものである。

```
bathroom+showers+pictures
miss+mundo
owen+mulligan+tyrone
news+press
free+wedding+planner
marc+arther+glen
gokmenler+agricultural+machinery+co.
```

図１．検索ログの例

本研究は、関根がマイクロソフト・リサーチに訪問研究員として滞在していた期間に行われた。リーダーの Dr. Bill Dolan 他の皆様に感謝する。

3. 拡張固有表現

本実験で利用した拡張固有表現は、関根らによって開発され、一般的な８～９種類の固有表現に対し、細分化、拡張を行い現在は約 200 種類のカテゴリを持った固有表現体系である。例えば、地名は、政治的地名（GPE）、国内地域名等があり、GPE も、国名、都道府県州名、市町村名等がある。また、拡張されたカテゴリとして、製品名、イベント名、自然物名、色名、賞名などがある。英語、日本語とも、20～30 万語の辞書とルールによるタガーを構築している (Sekine and Nobara 2004) (ENE homepage)。例えば、賞名の英語の辞書項目には 641 の項目があり、その例を図 2 に示す。

100 greatest Britons, 100 worst Britons, aaass/orbis books prize for polish studies, abel prize, academy award, academy awards, acm turing award, agatha award, Agatha awards, air medal, akutagawa prize,

図２　賞名の辞書項目例

4. アルゴリズム

本節では、辞書整備のアルゴリズムについて説明する。アルゴリズムは3つの部分からなる。まず、現在ある各カテゴリの辞書項目を検索ログに当て、コンテキストを抽出する。そのコンテキストの中からそのカテゴリにとって特徴的と考えられるものを計算する。次に、特徴コンテキストを使って、現在の辞書項目中からのノイズの発見、および、現在の辞書項目にない未知語の発見を行う。

4.1. 特徴コンテキストの抽出

まず、現在の辞書項目を検索ログに当てる。例えば、賞名だと 641 の項目を含む検索質問を集め、それをカウントする。結果の一部を図３に示す。

202 academy+awards the+#
86 academy+awards #+winners
76 academy+awards #+history
74 academy+awards #+nominations

図３　賞名のコンテキストの例

ここでは例えば、”the academy awards”という検索式が 202 回、”academy awards winners” が 86 回、検索ログの中にあったということを示している。

このようにして、得られたコンテキストを基に、そのカテゴリに特徴的なコンテキストを同定する。出現したコンテキストを頻度順に見てみると、上位には、そのカテゴリ以外のカテゴリでも頻繁に使われるようなコンテキストが存在していることが分かる。例えば、賞名では、”#+pictures”, “#+photos”, “#+history”といったコンテキストがタイプ頻度において 2,3,4 位に出現する。この問題は頻度情報を基に重要な特徴を見つける一般的な問題であり、例えば TF/IDF、相対頻度比、相互情報量、カイ二乗検定等の統計手法が知られている。しかし、本データではそれらの方法では上手く行かなかったため、以下の式で重要度を求めた。この式はコンテキストのタイプ頻度（共起している項目の種類数）を、全体におけるインスタンス頻度（延べ数）との比で補正したものである。細かく言うと、補正の項もそのカテゴリの（上位 1000 から求めた）平均で正規化している。この式によるスコアで上位９以内の賞名の特徴コンテキストを図４に挙げる。特徴的なコンテキストが抽出できていることがわかる。

$$Score(c) = f_{type}\{c\} * \log(g(c) / C)$$
$$g(c) = f_{type}\{c\}/F_{inst}\{c\}$$
$$C = f_{type}\{ctop1000\} / F_{inst}\{ctop1000\};$$

f: 注目するカテゴリでの頻度
F: 全てのカテゴリでの頻度
ctop1000:トップ 1000 のコンテキスト

#+winners, #+nominees, #+nominations, #+winner, #+award, who+won+#, winners+of+#, list+of+#+winners, winners+of+the+#

図４　賞名の特徴コンテキスト（上位９個）

4.2. ノイズの発見

4.1 の方法で、各カテゴリにおいて特徴コンテキストを同定することができた。実際、賞名以外のカテゴリでも非常にいい特徴コンテキストが同定されている。

次に、このコンテキストを使って、現在の辞書項目の中からノイズと思われる項目をみつ

ける。これは、上記で見つけられた特徴コンテキストはこのカテゴリで重要なものであるはずなので、このコンテキストをあまり含まない辞書項目はノイズではないかという仮説を用いる。実際には、上位 20 個の特徴コンテキストの頻度割合が低いもの(1%以下)を抽出する。以下に賞名および本名での結果の例を図５に挙げる。

賞名：makeup, mvp, short film, directing, iron cross, toni morrison, golden gloves, baseball hall of fame, golden bear, …
本名：it, space, night, we, working, candy, wheels, foundation, jazz, ghost …

図５　ノイズとして発見された辞書項目

多少の誤りはあるものの、普通には賞や本の名前として使われないものや一般名詞が抽出できていることが分かる。ちなみに、makeup はアカデミー賞のカテゴリ名である。

4.3. 未知語の発見

現在の辞書項目にない未知語を発見する手法も特徴コンテキストを使用する。特徴コンテキストとの共起が多いものは、現在の辞書項目になくとも、そのカテゴリの辞書項目ではないかという仮説を利用する。ここでは、いくつの特徴コンテキストと共起しているかというタイプ頻度によって項目を並べる。図６に、賞名、鳥名で見つかった未知語の例を示す。鳥名では、cardinal（コウカンチョウ族）のような一般的な鳥の名前が挙げられているが、実際の辞書項目には、Northern Cardinal や Red-capped Cardinal といった個として存在する鳥の名前は挙がっていたが、族の名前は挙がっていなかったために未知語として見つかった。辞書整備の観点からすると、非常によい結果である。

賞名：golden+globes, grammys, golden globe, kentucky derby, daytime emmy, sag, sag awards, American idol, daytime, emmys
鳥名：cardinal, eagle, bird, penguin, hawk,

図６　未知語として発見された項目

5. 評価結果

本節では、発見されたノイズおよび未知語の評価結果を報告する。発見されたノイズと未知語は非常に多いため、賞名、本名、鳥名においてのみ評価を行った。評価は、Wikipedia、Google サーチ、本の場合には Amazon.com の順に検索を行い、項目として上位に見つかったものとカテゴリとのマッチを見て判断した。Wikipedia は約 150 万項目の百科事典、Google 検索は Wikipedia よりも幅広い検索が可能で、Amazon.com は本ならばほぼ全てカバーしているという特徴があり、一般には後になるほどよりカバレージが広い順番に並んでいる。従って、前の方の評価で見つからなかった場合のみ、後ろの方の評価も行うというやり方をとった。

5.1. ノイズの発見

表１にノイズ発見における評価結果を示す。今回の実験では、20 個の特徴コンテキストとの共起頻度の割合が 1%以下の場合をノイズと仮定している。例えば、ノイズと判断された本名で、Wikipedia に本名としてあったものが 22、それ以外で Google で見つかったものが 1、Amazon.com で見つかった物が９個あったということである。この結果を見るとあまり良い結果とは言えない。しかしながら、たとえ Wikipedia に載っていようとも、例えば本名では”It”, “Space”, “Night”, “We”など普通名詞であり、Web 検索では本として探される頻度が低いであろうというものが多く含まれている。また、鳥名では、Wikipedia の網羅率は非常に高く、非常に稀な名前の鳥でも載っている。例えば、ニュージーランドに生息する”Tui”という鳥が抽出されたが、Google 検索で”tui”とすると、結果から見て、欧州の旅行会社、アイルランドの教員組合などが主な検索対象であるようである。（ただし、Wikipedia の”tui”の項目も Google 検索の 7 位に挙がる）したがって、今回の手法でノイズと判断されたものは、検索時にはその単語がそのカテゴリとして使われていない可能性が高いといったものであろう。また逆に、拡張固有表現辞書自身にはノイズは少ないということが言える。ただし、賞のカテゴリ名や明らかな間違えは多少見つかっている。

表１．ノイズの発見の評価結果

	本名	賞名	鳥名
Wikipedia	22	12	19
Google	1	0	0
Amazon	9	-	-
なし	0	16	3

5.2 未知語の発見

現在の拡張固有表現辞書には含まれていない未知語の発見の評価結果を図7に示す。それぞれ、評価データは信頼度の違う3つの異なったグループから取った。20 個の特徴コンテキストとの重複の度合いを信頼度とし、重複の多い方から 20 個(TOP)、重複が 6 である所から 20 個(MIDDLE)、重複が 3 である所から 20 個(BOTTOM)を評価サンプルとし、それぞれに対して評価を行った。本名では、TOP において Wikipedia との重複が 19 個で、残りの 1 つも Google で本であると見つかっており非常に良い結果を得ている。また、信頼度が下がるにつれ、Wikipedia との重複も減り、信頼度の高い物ほど正確に本名を発見できている。また、信頼度が低くとも Google や Amazon で見つかっていることから、有名ではない本も本手法により見つかることが分かる。賞名においては Wikipedia に載っていれば Google でも見つかるという関係があるが、ここでも信頼度と見つかる数の相関が見られる。鳥名の場合には、TOP でしか未知語が見つかっていないが、これは特徴コンテキストに他の種類の単語でも使われるようなコンテキストが入ってしまい、信頼度を低くすると他の種類の物が取れてしまっている。ただし、多くの特徴コンテキストを使えばきちんとした鳥名が抽出できている。

6. 考察

ここで提案した手法には、あるカテゴリに属する単語リストを抽出する Bootstrapping の手法との親和性が認められる(Brin 98) (Collins 99)。Bootstrapping の手法では、まずあるカテゴリに属する数個の例をシードとして与え、そのシードが現れるコンテキストを使って新たな項目を探していくという手法である。この手法に比べると、本手法では予め構築された比較的規模が大きくほとんどが正しい辞書項目を使える利点がある。実際のインプリメントを考えると、全く新しいカテゴリが次々に現れるような場合でない限り、ある程度の辞書項目を集めることは難無くできることが多いため、本手法の方が現実的であると考えられる。どれだけ性能に違いが出るかについては今後実験していく計画である。

7. まとめ

本稿では、検索ログを使って拡張固有表現の辞書項目の整備をする方法を提案した。各カテゴリの特徴コンテキストを同定し、それによってノイズと未知語を発見する手法である。拡張固有表現は辞書項目の精度が重要であり、その整備に検索ログが非常に有用であることが分かった。

参考文献

ENEhomepage: http://nlp.cs.nyu.edu/ene

S. Brin, “Extracting Patterns and Relations from the World Wide Web”. *Workshop on the Web and Databases 1998.*

M. Collins and Y. Singer, “Unsupervised Models for Named Entity Classification”. EMNLP and VLC 1999.

S. Sekine, C. Nobata, “Definition, Dictionary and Tagger for Extended Named Entities”. LREC 2004

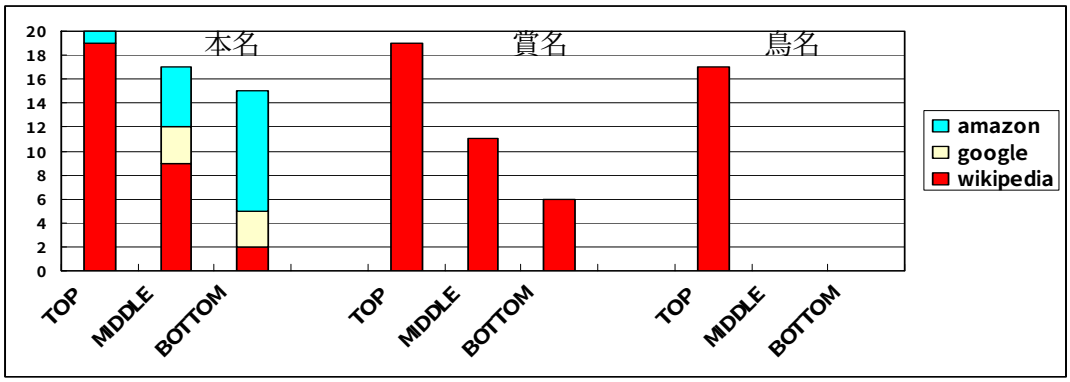

図7．未知語の発見の評価結果