5P-5

Webからの関係情報の抽出

辻下 卓見†, 森 純一郎‡, 石塚 満‡

東京大学工学部電子情報学科†, 東京大学院情報理工学系研究科‡

１．はじめに

近年のWeb利用者の急激な増加により、Webを利用することで取得できる情報は爆発的に増加している。Googleにインデックスされているページだけでも８０億ページを遙かに超える。またWeb上の情報処理において関係に関する情報の重要度があがってきている。

そこで本稿では、Webを用いて地名、組織、といったエンティティ間の関係を自動で抽出する手法を提案する。提案手法では非教師アプローチの学習手法を用いて、Web上から様々なエンティティ間の関係を抽出する。

提案手法によって抽出された関係を用いて、メタデータの自動生成や質問応答システムへの応用が考えられる。本稿では、まず第２章で関連研究についてのべ、第３章で関係情報の抽出手法を提案し、第４章で実験結果を報告し、第５章でまとめを述べる。

２．関連研究

従来の関係抽出のタスクでは、新聞記事などを対象に、教師付き学習のアプローチが多く行われてきた。しかしこの方法は学習用に事前にタグ付けされたコーパスを用意しなければならない。また抽出される関係も、タグ付きコーパスを作成するときに定義した関係に限定される。このため、多量のタグ付きコーパスを必要としない半教師付き学習のアプローチも行われている。[Brin 98]の手法では、本のタイトルと著者の少数のサンプルをブートストラッピングに必要な種として用意して、共通に出てくるパターンを抽出している。この手法を改良したものとして[Agichtein and Gravano 2000]などがある。しかし半教師付き学習アプローチも種の選び方が不明瞭なことなど問題点がある。

一方、これらの問題点を解決するために、非教師学習を用いて関係を抽出する手法もある[Hasegawa 04]。

３．手法

本稿では新聞記事をコーパスとしている[Hasegawa 04]の手法をWebに適用する。

まずContextual hypothesis[Miller and Charles 91]より、出現する文脈が類似しているエンティティペアは類似した関係にあり、後述する文脈モデルをもとにクラスタリングすることで同一のクラスタにまとめることができる、と仮説を立てた。この仮説の元に行う提案手法の基本的な流れは以下の通りである。

１．Webからエンティティペアを抽出する。

２．エンティティペアの文脈モデルを作成する。

３．文脈モデルをもとにエンティティペア間の類似度を計算する。

４．エンティティペアをクラスタリングする。

５．できたクラスタごとにラベルを作成する。

[Hasegawa 04]のように新聞記事をコーパスとしているときは、コーパス全体を対象にタスクを行えばいいが、Web全体をコーパスとする場合は、当然その全てを対象とすることはできない。そのため、検索エンジンを利用してコーパスを得る必要がある。

また、新聞記事と異なり、Web上の文章は定型的ではないため文脈ベクトルを得るときは、ある程度余裕を持って行う必要がある。

３．１ 文脈モデル

文脈モデルとは、Web上でエンティティペアが出現するときの文脈をモデル化した者である。

まず固有表現(Named Entity)抽出ツールを用いて抽出したエンティティのペアから、文脈モデルを作成する。エンティティペアの文脈モデルはベクトルで表す。これを文脈ベクトルと呼ぶ。文脈ベクトルの生成は以下のように行う。

関係性の低いペアを省くため、エンティティペアの共起値を求め、共起値が閾値以上である関係性の高いペアを求める。共起とは、同じWebページに二つ語のが出現することであり、共起値とはその共起の強さを示す指標である。ここでは共起値としてSimpson係数を利用した。語Wの検索ヒット件数を$|W|$と表すとき、AとBのSimpson係数は$|A\cap B|/\min(|A|,|B|)$で表される。共起値を用いてフィルタリングしたエンティティペアを検索エンジンに投げ、得られたURLのうちスコアの高い上位100件のページを抽出する。

抽出したページを形態素解析ソフトをつかって分かち書きし、エンティティペアが出現する前後の十語と、ペア間に現れる語を抽出する。ここで、エンティティペアが出現するとは二つのエンティティの固有表現が前後十語以内に共起して現れるのときのことを言う。

次に抽出した語を用いて、エンティティペアの文脈ベクトルを生成する。

ベクトルの各要素の値にはTfidfを用いた。

Tfは、各エンティティペアの文脈中の単語頻度、DFを単語の文章頻度、Nを文章数とすると、Idfは$\log(N/DF)+1$である。TfidfはTf*Idfであり、文脈中の語の網羅性と特定性の両方を考慮した値である。

３．２ エンティティペア間の類似度

クラスタリングをするために、前節で求めたエンティティペアの文脈ベクトルを用いて、エンティティペア間の類似度を求め、距離行列を作る。

本稿では[Hasegawa 04]に従い、最も簡単に求まるコサイン尺度を用いた。コサイン尺度は、通常のベクトル同士のコサインであり、文脈ベクトル同士が同じ時１になり、全く異なるとき０になる。

類似度を求める手法としては、ほかにユークリッド距離

Extracting Relations from Web

†Takumi Tsujishita, Faculty of Engineering, The University of Tokyo

‡Mori Junichiro, Ishizuka Mitsuru , Graduate School of Information Science and Technology, The University of Tokyo

を用いる手法や、ベクトル要素同士の共起値を求める手法、分散などを用いる方法が考えられる。

３．３ クラスタリング

作成した距離行列を元に階層クラスタリングを行う。クラスタリングは大きく分けて、階層クラスタリングと、非階層クラスタリングの二つに分けられる。階層クラスタリングは、さらに分枝型と凝集型に分けられる。

凝縮型の階層クラスタリングは、1 個の対象だけを含む N 個のクラスタがある初期状態から、クラスタ間の距離関数に基づき、最も距離の近い二つのクラスタを順に併合する。そしてこの併合を、全ての対象が一つのクラスタに併合されるまで繰り返すことで階層構造を獲得する。クラスタ間の距離関数の違いにより、最短距離法、単連結法、最長距離法、完全連結法、群平均法等といった手法がある

３．４ ラベルの抽出

クラスタ内に含まれるエンティティのペアの文脈ベクトルに共通に含まれている語は、そのクラスタの特徴を表すと考えることができる。つまり、共通語がクラスタが表す関係を端的に示す語であると考えることができる。

[Hasegawa 04]ではこの考えに基づき、単純にクラスタ内の共通語の頻度を数えることで、ラベルを作成している。しかしこの手法では全てのクラスタの全てのベクトルに含まれるような語がラベルになってしまう可能性が存在する。そのため提案手法では各クラスタごとに重心ベクトルを求め、そのベクトル同士で Tf*Entoropy 値を求めることで特定性を考慮してラベルを抽出した。Entoropy は情報理論から、語の特定性を表す値である。

４．実験

本稿ではエンティティの固有表現は固有表現抽出ツールを用いて判別できることを仮説とする。また、本稿の実験ではエンティティのペアは抽出できたものとして、あらかじめ用意したものを用いた。また、同姓同名問題については、解決されているものとする。

まず、実験対象として、人(PERSON) -地名(GPE)のエンティティペアを用意した。ここでは人エンティティとして政治家を用いる。エンティティペアの総数は１４４ペア用意した。また、各エンティティペアに対して正解データとして、あらかじめ、大統領、首相、議員、知事、市長というラベル付けを行った。

そしてそれぞれのペアについて、検索エンジン、ここでは Google を使って上位 100 件のページをダウンロードした。次に形態素解析ソフトとして Chasen を用いて形態素解析を行い、エンティティペアの固有表現の間に挟まれている語と、前後十語うち、名詞および未知語を文脈ベクトルの要素として抽出した。例えば「小泉純一郎」 「日本」というエンティティペアの文脈ベクトルを作った場合、値の大きな要素順位７件は図１ のようになった。なお文脈ベクトルを作るときにはストップワードの除去と、低頻度語の除去を行っている。

クラスタリングでは、階層クラスタリングのうち完全連結法を用いた。クラスタリング数は、あらかじめ用意した正解ラベルの数＋１である６をあらかじめ与えた。

病理	0.401
藤原	0.111
光文社	0.110
首相	0.101
肇	0.089
構造	0.062
大臣	0.054

図１　小泉純一郎 日本 ベクトル

５．実験結果

できたクラスタの評価には Precision(P)と Recall(R)および F 値(F = 2RP/R+P)を用いた。Precision、Recall、F 尺度それぞれ、エンティティペア全体で見たときの正解ペアの割合、各クラスタ中の正解ペアの割合、Precision と Recall の組み合わせである。

各評価値の値は図２のようになった。いずれもかなり精度の高い値になっていることがわかる。

Precision	0.923
Recall	0.764
F 尺度	2.769

図２ 実験結果

また、大統領クラスタのラベルとして図３が得られた。ラベルを手動で評価したところ、精度は約八割ほどになった。しかし、各クラスタも二位以下の重要語に関してはあまりいい精度は出ていない。

大統領	ページ	関連	政権	首相

図３ 大統領クラスタのラベル

６．まとめ

以上の実験により[Hasegawa 04]の手法が Web を対象としたときにも有効であることが示された。

しかし、今回実験した研究者ドメインは、比較的定型的な文章を拾いやすいと言う特徴があるため、関係ドメインへ同手法を適用することを考えると、ネット上の文章の非定型性をより考慮するよう文脈モデルを改良する必要が考えられる。

また現状ではクラスタ数はあらかじめ与える必要があるが、これも自動で決定できるようにする必要がある

ほかにも、エンティティペア間の類似尺度の決め方などクラスタリング手法についても改善の余地があると考えられる。

参考文献

[1]Eugene Agichtein and Luis Gravano. 2000. Snowball:Extracting relations from large plain-text collections. In *Proc. of the 5th ACM InternationalConference on Digital Libraries (ACMDL'00)*, pages 85–94.

[2]Hasegawa,Sekine,Grishman 2004 Discovering Relations among Named Entities from Large Corpora , ACL 2004

[3]長谷川,関根:教師なし学習による関係抽出に基づくパラフレーズの獲得, 言語処理学会第11回年次大会発表論文集, pp. 1145-1148

[4]松尾豊、友部　博教、橋田　浩一、石塚満：Web からの人間関係ネットワークの抽出，人工知能学会論文誌 Vol.20 No.1 p46 -56 (2005)

[5]森　純一郎、松尾豊、石塚満：Web からの人物に関するキーワード抽出，人工知能学会論文誌，Vol.20，No.5，pp.337 -345