1V-5

潜在関係検索における Indexing 対象の拡張

信田頼宏 † Nguyen Tuan Duc ††
Danushka Bollegala †† 石塚満††

† 東京大学工学部電子情報工学科 †† 東京大学大学院情報理工学系研究科

1. 目的と背景

潜在関係検索はエンティティペア間の関係類似性に基づくエンティティ検索手法である。例えば、クエリ {(Tokyo, Japan), (?, France)} に対して、“Paris” という結果を出力する。これは入力ペア (Tokyo, Japan) には首都-国の関係があると認識され、それに基づいて “France” の首都となる “Paris” が検索されることである。現在のシステムでは、潜在関係検索用 Index はサイズの点から固有名詞エンティティペアに制限しており、例えば {(Google, search),(Oracle, ?)} のような固有名詞・一般名詞や一般名詞のペアを高速に検索できていない。そこで、本研究では潜在関係の Indexing 範囲を一般名詞まで拡張する手法を提案する。また、有意に関係を持つ名詞ペアだけを選択するフィルタリングアルゴリズムで Index のサイズを減らす手法を考案する。

2. 関 連 研 究

本研究では、膨大なコーパスの中から名詞のペアを抽出することが重要となる。そこで、この章では英語の概念辞書である WordNet と、名詞間の関係を抽出する TextRunner [1, 2] について紹介する。まず概念辞書とは、英単語を同義語のグループにまとめ、それらのグループをさらに上位・下位の関係によって定義される階層構造にまとめたものである。本研究との類似点としては、概念をより細かく分類し、いろいろな関係の上位語・下位語を用いることで類似関係の検索に役立てることができる。しかし、この方法ではより正確に検索するために、よりたくさんの新しい概念の導入が必要となる。また、WordNet に存在しない単語は検索できない。特に現代において、流行の言葉は日々変わり、また次々に新しい言葉が生みだされているので、新しい単語に対応できない。また、名詞間の関係を抽出するモデルとしては TextRunner [1, 2] がある。TextRunner は、2 つの単語を入力したとき、その 2 つの単語間に成り立つ述語を出力することができる。しかし、周辺文脈は関係に取り込まれない。ゆえに、名詞抽出には使えるかもしれないが潜在関係検索に不向きである。

3. 潜在関係検索

ここでは、潜在関係検索 [3, 4] の仕組みを軽く紹介する。

3.1 エンティティペアの抽出

エンティティペアの抽出において、まずテキストドキュメントを文ごとに区切り、各文に形態素解析を行う。形態素解析とは、文の一つ一つの単語に品詞のタグ付けを行うことである。これに固有表現抽出器を用いることでエンティティペアを抽出することができる。固有表現抽出器は、タグが固有名詞の単語を抽出するものである。図 1 はエンティティペアを抽出するための一連の流れを図にしたものである。

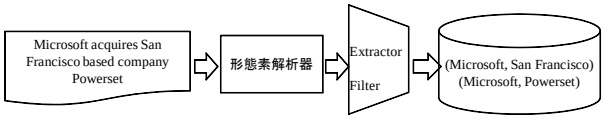

図 1 エンティティペアの抽出

3.2 語彙パターンの抽出

エンティティペアの潜在関係を知るためにまず語彙パターンを抽出する必要がある。具体的な方法としては、それぞれのエンティティで挟まれる語彙パターンとそれらの前後に出現する語彙パターンの頻度でエンティティペアの潜在関係を表現する。ここで類似度の計算は、関係類似度計算アルゴリズム [5, 6] に則る。表 1 を用いて説明する。“It is now official: Microsoft aquries San Francisco based company Powerset for $100M.” においてエンティティペアと成りうるのは (Microsoft, San Francisco), (Microsoft, Powerset), (San Francisco, Powerset) の 3 つである。今回は (Microsoft, Powerset) のエンティティペアの語彙パターン抽出を例に説明する。まず、“Microsoft” の前 3 単語以前とと “Powerset” の後ろ 3 単語以降の単語列を削除する。次にエンティティペアを X,Y に置き換える。これは、語彙パターンにエンティティペアを含まないようにするためである。そして、この文にステミング (stemming) という作業を行う。ステミングは、英語の単語の活用によって変化する部分を取り除く作業である。日本語でいうと動詞における語幹を残すことに類似している。最後に n-grams 生成を行う。n-grams というのは n 個の部分単語列をあらわす。n-grams の中に X,Y を含まなければ、X*(n-grams)*Y(*は 0 語以上の単語列) と書くことにする。

表 1 語彙パターンの抽出過程

文	It is now official: Microsoft aquries San Francisco based company Powerset for $100M.
変数置き換え	now official: X aquries San Francisco based company Y for $100M.
stemming	now offici: X aquri San Francisco base compani Y for $100M.
語彙パターン	X acquir * Y, X * San Francisco * Y, offici: X acquir * Y, X acquir San Francisco * Y

4. 一般名詞抽出への拡張の提案

既存の研究において、抽出するエンティティペアは全て固有名詞であった。これは、固有名詞の方が潜在関係がはっきりしているためであり、また実用的であるからである。しかし、潜在関係検索は本来固有名詞のみに対応するわけではなく、また固有名詞と

Method for Indexing Broader Range of Entity Types in Latent Relational Search

† Yorihiro Nobuta (Dept. of Information and Communication Engineering, Faculty of Engineering, The Univ. of Tokyo)
†† Nguyen Tuan Duc, Danushka Bollegala, and Mitsuru Ishizuka (Grad. School of Information Science and Technology, The Univ. of Tokyo)

一般名詞のエンティティペアにおいては固有名詞同士のエンティティペアと同じくらいの実用性があると考えられる。例えば、クエリ {(Google, search), (Youtube, ?)} を入力したとき、? のところに video と出力されるのが望ましい。これはエンティティペア同士の関係が会社名 (固有名詞) とそのサービス (一般名詞) となっているからであるが、このように一般名詞まで拡張すれば、この例のように今まで以上に検索の幅が広がる。この章では提案手法である一般名詞の拡張について説明する。

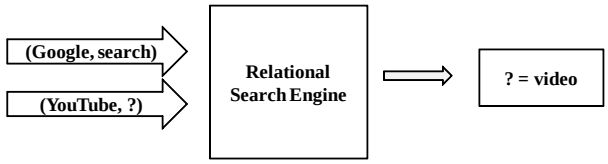

図 2 一般名詞の抽出

4.1 予 備 実 験

あらかじめ既存の潜在関係検索を用いて一般名詞を抽出できるようにして実験を行った。これは、抽出の段階でどんな名詞であろうと抽出できるようにしただけで一切のフィルタリングを行っていない。行った実験は、抽出されるエンティティペアの個数、潜在関係を持つエンティティペアの個数、クエリ自身の品詞（一般名詞か固有名詞）とそのクエリを要求した時の出力との関係、この3つである。また、この結果は表にまとめてある。まず、抽出されるペアの個数は表 2 より莫大な数になる。そして潜在関係をもつペアの割合が少ないことより、余分なエンティティペアをとりすぎていることが分かる。ここで、潜在関係を持たないペアのほとんどが一般名詞同士のペアであることに注意したい。また、実際にクエリを与えて出力を見た時も一般名詞のペアのクエリは満足のいく結果が得られたものが極端に少ないことが表 4 からわかる。よって、本研究の目的として、余分な抽出を減少、一般名詞と固有名詞のペアの出力の精度の上昇、この 2 つを徹底的に改善することにした。

4.2 エンティティペア抽出の際のフィルタリング

前章で述べたとおり問題は、余分なエンティティペアの抽出と出力の精度である。ここで、余分なエンティティペアの抽出を減らすということは、無関係なエンティティペアを抽出しないということになる。つまりそれは、潜在関係を持つペアを抽出するということになるので、そのまま精度の向上につながる。よって、余分なエンティティペア抽出の削減に重点を置くことにした。既存の潜在関係検索では、潜在関係を持たないペアの抽出を減らすために、ペア間の距離が離れているものを抽出しないという手法で行っていた。しかし、これでは距離が近ければ潜在関係を持たないペアも抽出され、逆に距離が離れていれば潜在関係を持つものも抽出されない。ゆえに、距離が近くても潜在関係がないペアはとらない、逆に距離が離れていても潜在関係を持つペアをとるようなフィルタリングを必要とした。そこで構文解析を用いたフィルタリング手法提案する。構文解析は、関係抽出対象となる文に対して行い、得られた構文木で距離が 10 以上のエンティティペアを抽出対象から外す。これは、たとえ 2 単語間の距離がそこそこ近くても構文木上で階層の差が大きいほど潜在関係はなくなるからであり、構文木上の距離 10 というのは大抵階層の差が大きいからである。

5. 評　　価

評価データとして、12000 ウェブページのテキストコーパスを使った。これらのテキストには主に、会社とその会社の特徴的なサービス (Google - search) などの関係がある。1000 個のコーパスに対して、抽出された単語ペアの数が 36234 である。表 2 では 100 個の文書を含むコーパスの解析であるが、提案するフィルタリング手法を適用した後に抽出された単語ペアの数は減少している。また、クエリとその出力のデータを表 5 に示す。表 5 の判定から精度を算出すると、75%以上の精度というよい結果を得られた。本システムが良い結果を出せたのは、構文解析によるエンティティペアの抽出が効果的であったと考えられる。

表 2 ワードペア抽出比較

抽出	ワードペアの個数
固有名詞のみ	231
一般名詞も含む	3213
フィルタリング後	2978

表 3 エンティティペアの関係性 (100 個のペアを判別)

ワードペア	例	フィルタリング前	フィルタリング後
関係性あり	(Google, Youtube)	37	39
関係性なし	(search, giant)	63	61

表 4 クエリを要求したときの出力 (フィルタリング前)

クエリ	例	正答
固有名詞同士	(Tokyo, Japan),(?, Paris)	91
固有名詞と一般名詞	(Google,search), (Youtube, ?)	20
一般名詞同士	(apple, fruit), (carrot, ?)	3

表 5 クエリと結果の例

クエリ	答	判定
{(Google, search),(Android, ?)}	? = phone	正
{(Google, search), (Youtube, ?)}	? = business	否

6. む　す　び

本稿では潜在関係検索における Indexing の一般名詞への拡張手法を説明した。実際に、テキストから一般名詞も含むエンティティペアを抽出し、構文解析を行うことで一般名詞も含む潜在関係検索を実現した。提案手法は、フィルタリングされていない既存手法と比べると、抽出するエンティティペアも減り、さらに精度も高くなった。

参　考　文　献

1) Etzioni, O. et al.: Open Information Extraction from the Web, *Communications of the ACM*, Vol. 51, No. 12, pp. 68–74 (2008).
2) Banko, M. and Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction, *Proc. of ACL'08*, pp. 28–36 (2008).
3) Duc, N. T., Bollegala, D. and Ishizuka, M.: Using Relational Similarity between Word Pairs for Latent Relational Search on the Web, *Proc. of WI'10*, pp. 196 – 199 (2010).
4) Duc, N. T., Bollegala, D. and Ishizuka, M.: Cross-Language Latent Relational Search: Mapping Knowledge across Languages, *Proc. of AAAI'11* (2011).
5) Bollegala, D. et al.: Measuring the Similarity between Implicit Semantic Relations from the Web, *Proc. of WWW'09*, ACM, pp. 651–660 (2009).
6) Bollegala, D. et al.: Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web, *Proc. of WWW'10*, ACM, pp. 151–160 (2010).