単語ペア間の潜在的関係を利用する関係検索エンジン

Nguyen Tuan Duc †　Danushka Bollegala †　石塚満†

WWW などのテキストコーパスの中に、エンティティ間の関係が多数潜在的に記述されている。従来のキーワードベース検索エンジンは単語ペア間の関係を利用して検索することが出来なく、潜在的に存在する多くの関係情報を積極的に利用できない。本研究では、単語間の潜在的な関係を積極的に利用し、類似関係による潜在関係検索エンジンを提案する。本検索エンジンは例えば入力クエリー (Japan,Tokyo), (France, ?) に対して、答え「Paris」を出力する。関係抽出手法などを使い、上記の検索エンジンを実装し、評価することにより、上記の検索パラダイムを大規模 WWW 空間での実現可能性を明らかにした。

1. はじめに

潜在関係検索とは、単語ペア間の潜在的な関係を利用することにより入力単語ペアと類似する単語ペアを検索する新しい検索パラダイムである。潜在関係検索エンジンの概要は 図 1 に示している。クエリー「(Tokyo, Japan), (?, France)」が入力されたときに、「Paris」を最初にランキングされた結果リストを返す。その理由は「Tokyo」と「Japan」との関係は「Paris」と「France」との関係に類似するからである (東京が日本の首都、パリもフランスの首都などの関係があるからである)。潜在関係検索のアイディアはアナロジー・シソーラスの研究 [1] や関係類似度測定研究 [2] で検討されてきた。本研究は参考文献 [3] の研究と同時に最初の

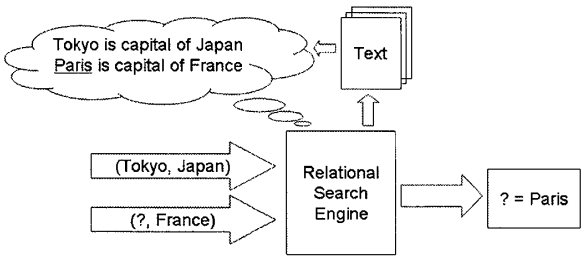

図 1　潜在関係検索の例

潜在関係検索エンジンの実現方法を提案する。参考文献 [3] では、既存のキーワード・ベース検索エンジンを利用し、潜在関係検索を実現したが、単語間における関係を十分に抽出できず、精度や平均逆順位 (MRR) がまだ低い。また、クエリー処理時にキーワード・ベース検索エンジンに数十個のクエリーを投げているので速度が遅い。本研究では関係抽出の手法を使い、潜在関係検索エンジンのインデックスを作成し、高精度の関係類似度測定 [2] の研究成果を応用し、高速かつ精度の高い潜在関係検索を実現する方法を提案する。また、作成した検索エンジンを評価することにより、上記の検索パラダイムの実現可能性や実際に応用する可能性を明らかにする。

本稿は潜在関係検索の実現手法と評価結果をまとめる。以降、第 2 節では、関連研究を紹介し、第 3 節で潜在関係検索の実現手法を述べる。第 4 節は評価結果を示す。最後に、第 5 節でまとめと今後の課題について説明する。

2. 関連研究

テキストデータから単語ペアと単語間の関係を自動的に抽出するシステム TextRunner [4] では、単語間の関係を述語で出力する。TextRunner は正確な述語を関係として抽出するが、この関係表現方法は周辺文脈は関係に取り込まなく、関係類似度を高精度で測定出来ない (文脈が似ているペアと似ていないペアも同じに扱うからである)。

関係類似度測定の研究 [5] [2] では、単語間の関係を周辺文脈の語彙パターンで表し、パターン集合の類似度で関係類似度を定義する。本研究も同様に、関係を語彙パターンで表す。また、Bollegala et al. [2] で述べたパターンクラスタリング手法を使い、似ているパターンを一つのクラスタにまとめ、正確マッチングパターンの低頻度問題を解決する。

潜在関係検索をキーワード・ベース検索エンジンを利用して、実現したシステム [3] では、前述したように、単語ペアの前後のコンテキスト (文脈) を考慮せず、単語間の関係を十分に表現できない。そこで、精度や MRR が低い。また、一つの潜在関係検索クエリーを処理するために、キーワード・ベース検索エンジンに数十個のクエリーを投げ、応答時間がかかるという欠点がある。

本研究はウェブからの情報抽出や関係抽出の手法を使い、自動的に単語ペアや単語ペアの関係を特徴づける語彙パターンのインデックスを作成し、潜在関係検索を実現する。

3. 潜在関係検索の実現手法

潜在関係検索を実現するためには、エンティティ(単語) ペアとペアの二つのエンティティ間の関係を抽出すると抽出されたペアの関係類似度を測定する必要がある。

3.1 エンティティ・ペア抽出と関係の表現

本システムは WWW などのテキストコーパスからエンティティ・ペアとペア中の単語間関係を抽出し、検索のデータベースを作成する。まず、テキスト・ドキュメントから文を抽出する。得られた文を単語に分節し、品詞タグを付ける (そのとき、named entity も同時に認識する)。例えば、「東京は日本の首都である」という文は、「東京/固有名詞 は/助詞 日本/固有名詞 の/助詞 首都/名詞 である/動詞」の列に分けられる。分かられた列中に名詞、固有名詞を抽出し、ペアを形成する (現在の実装では、固有名詞しか抽出していない)。上記の文からは、(東京, 日本) のペアが抽出される。実質に関係を持つペアは普通出現頻度が高く、偶然に共起するペアはあまり頻度が高くない。そこで、ある頻度以上のペアだけを検索対象とする (現在の実装では頻度 5 以上)。また、文中の距離が遠ければあまり関係を持たない可能性が高いので、距離がある閾値 D_{max} 以上のペアは検索対象としなく関係を抽出しない。

単語ペアを抽出すると同時に、ペアの関係を表す語彙パターンも抽出する。語彙パターンはペアの出現位置の周辺語彙列を取り、

† 東京大学大学院情報理工学系研究科

列 n-gram として抽出する。例えば、上記の文では、(東京, 日本) ペアの語彙パターン集合は {X は Y の首都, X は Y の首都である, X は Y*首都 } を含む (「*」はワイルドカードで 0 個以上の語彙を意味する)。単語ペア S_j と語彙パターン p_i が一緒に出現する頻度を $h(S_j, p_i)$ とする。そのとき、単語ペア S_j の特徴ベクトルの要素は i の値を $h(S_j, p_i)$ として定義する。また、語彙パターン p_i の特徴ベクトルの要素 j を $h(S_j, p_i)$ とする。抽出された単語ペアと語彙パターンやその頻度をデータベースに保存する。

3.2 関係類似度の測定とクエリー処理

入力されたクエリー (A, B), (C, ?) に対して、まず、ペア (A, B) をデータベースから取得し、ペアの関係を表す語彙パターンを取り出す。ペアの関係を表すパターンを含む他の (C, X) ペアを検索結果の候補ペアとする (つまり、候補ペアの第一要素が C で、入力ペアと一個以上の語彙パターンを共有する)。(A, B) ペアの特徴ベクトル (要素 i が $h((A, B), p_i)$ のベクトル) と (C, X) ペアの特徴ベクトルをデータベースから取り出し、ペアの関係類似度を特徴ベクトルのコサイン類似度により計算する。(A, B) ペアと (C, X) ペアの正確にマッチする語彙パターンは少ないので、上記の語彙パターンの特徴ベクトルを使い、類似するパターンをクラスタリングする (例えば、"X is CEO of Y", "Y's CEO X", "Y chief executive X" を同じクラスタにまとめる)。同じクラスタに入るパターンは同じパターンとして扱い、ペア間のコサイン類似度を計算する。

また、エンティティの複数表現形を吸収するために、似ているエンティティをクラスタリングする必要がある (例えば、"United States", "U.S.", "US" は同じクラスタに入るようにする)。そのために、エンティティとのペアをなすエンティティを次元とするエンティティの特徴ベクトルを定義する。エンティティの類似度はエンティティの特徴ベクトルのコサイン類似度として定義する。類似度の高いエンティティを 1 つのクラスタにまとめ、出力結果で同じクラスタのエンティティであれば、同じグループとして出力する。 図 **2** は潜在関係検索を実現するためのエンティティと語彙パターンの関係モデルである。エンティティをクラスタを形成し、語彙パターンもパターンクラスタをなすことに注意されたい。

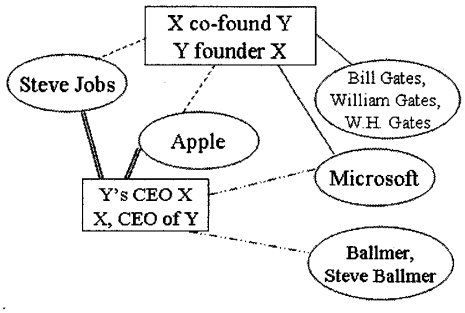

図 **2** エンティティと語彙パターンの関係

4. 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 4 種類の関係が含まれている: 人の生まれ場所 (Einstein - Germany), 会社の本社所在地 (Microsoft, Redmond), 会社の社長 (Eric Schmidt - Google) と会社買収関係 (Google - Youtube)。上記のテキスト・コーパスから 113742 個の単語ペアが抽出された (その内、4103 ペアが頻度 5 以上)。また、抽出された語彙パターン数は 2069121 である。テストのクエリー・セットは 842 クエリーがあり、その内、12 個のクエリーが複数正解 (ある会社が買収した会社集合) を持つ。正解が 1 つしかない場合、トップ 1 結果だけの精度と再現率を評価し、正解が複数の場合、トップ 10 の結果を評価する。精度、再現率と F-score の評価結果を表 **1** で示す。正解が一つしか存在しないクエリーに対しては高い精度と再現率が得られた。正解が複数存在した場合も 65%以上の精度が得られた。

表 **1** 精度、再現率と F-score の評価結果

関係種類	精度	再現率	F-score
人 - 生まれ場所	91.74	87.63	89.64
会社 - 所在地	95.71	95.71	95.71
社長 - 会社	90.8	90.42	90.61
会社買収関係	65.83	65.83	65.83
平均	86.02	84.90	85.45

表 **2** 既存研究との比較結果 (@N はトップ N 結果に正解があるクエリーの割合)

手法	MRR	@1	@5	@10	@20
既存手法 [3]	0.545	43.3	68.3	72.3	76.0
提案手法	0.930	89.3	97.1	97.5	97.5

また、先行研究 [3] との MRR(平均逆順位) や正解がトップ N (N = 1, 5, 10, 20) の中にあるクエリーの割合の比較を 表 **2** に示している。上記のすべての指標で提案手法がより良い結果を出力することが分かる。 本システムが良い結果を出したのはパターン抽出手法がよく関係を表現できることと似ているパターンがクラスタリング手法により同じように扱えることの効果だと考えられる。

5. 終わりに

本稿では、潜在関係検索検索をウェブ上で実現する手法について説明した。テキストからエンティティ(単語) ペアを抽出し、エンティティ・ペアの関係を語彙パターンで表現し、ペア間の類似度測定手法を応用して、潜在関係検索を実現した。提案手法は既存手法と比べて MRR やトップ 1 の正解率が高い。今後は重要なパターンの特定や分離の実現 (例えば、「東京は日本の最大都市で首都である」の文から「X は Y * の首都である」や「X は Y の最大都市*である」というパターンが重要で、2 つの関係を表現) と膨大コーパスからエンティティ・ペアを抽出し、実践に使えるウェブ潜在関係検索エンジンを実現する予定である。

参 考 文 献

1) Veale, T.: The Analogical Thesaurus, *Proceedings of the 15th Conference on Innovative Applications of Artificial Intelligence, IAAI 2003*, AAAI Press, pp. 137–142 (2003).
2) Bollegala, D., Matsuo, Y. and Ishizuka, M.: Measuring the similarity between implicit semantic relations from the web, *Proceedings of the 18th International Conference on World Wide Web, WWW 2009*, ACM, pp.651–660 (2009).
3) Kato, M. P., Ohshima, H., Oyama, S. and Tanaka, K.: Query by analogical example: relational search using web search engine indices, *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009*, pp. 27–36 (2009).
4) Banko, M. and Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction, *Proceedings of ACL-08: HLT*, Association for Computational Linguistics, pp. 28–36 (2008).
5) Turney, P. D.: Measuring Semantic Similarity by Latent Relational Analysis, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, IJCAI 2005*, pp. 1136–1141 (2005).