言語横断型の関係類似性を利用する潜在関係検索

Nguyen Tuan Duc † Danushka Bollegala † 石塚 満†

本研究では，言語横断型の潜在関係検索という検索パラダイムを提案する．本検索エンジンは，言語を跨った類似関係を認識し，入力エンティティペアと類似関係を持つエンティティペアを検索する．例えば，クエリ {(日本, 富士山), (Germany, ?)} に対して，“日本の最も高い山は富士山である” や “Germany’s highest mountain is Zugspitze” という文を根拠として，“Zugspitze” という結果を出力する．本稿では，上記の検索パラダイムを実現するための語彙パターン翻訳手法や，言語横断型の言い換え表現の認識手法について説明する．ウェブコーパスを用いた評価実験では，提案手法がベースラインよりも高い平均逆順位 (MRR) を達成し，英語 - 日本語の言語横断型のクエリセットでは，MRR で 0.430 の値を達成した．

1. はじめに

我々が Web 検索エンジンでドイツで最も高い山の名前を検索したい時，{(Japan, Mt. Fuji), (Germany, ?)} のようなクエリが思い浮かぶが，このクエリに対して既存のキーワードベースの検索エンジンは答えることができない．我々はこのようなクエリに答えられるように，潜在関係検索という関係を利用した検索方法を考案 [1] し，高精度な英語の潜在関係検索エンジンを実現した [2]．潜在関係検索は，与えられたエンティティペア (A, B) (以降，ソースペアという) と与えられたエンティティC (以降，キーエンティティという) に対して，(A, B) と (C, D) が類似関係を持つようなエンティティD を検索できる．ここで，(A, B) ペア中の A と B との関係が (C, D) ペア中の C と D との関係と強く類似するときに，(A, B) と (C, D) の関係類似度が高いという．

潜在関係検索エンジンは “Mt. Fuji is the highest mountain in Japan” などの，エンティティ間の関係を記述する文を根拠として検索を実現する．しかし，これまでに考案 [1]，実現された [3, 4, 2] 潜在関係検索エンジンは単言語の検索エンジンであり，根拠となる文章は同じ言語で書かれている必要がある．これにより検索エンジンの根拠となる文章の範囲が限定され，特に Web 上で日本語でよく書かれているが，英語であまり書かれていないエンティティ(固有名詞) に対しては，精度が低い．また，英語でエンティティ(例えば，“Mt. Fuji”) を記述できないユーザは，英語テキストを利用した検索ができない．そこで本研究では，言語横断型の潜在関係検索という新しい検索パラダイムを提案する．図 1 に言語横断型の潜在関係検索の例を示す．図 1 では，クエリ {(日本，富士山),(Germany, ?)} の入力に対して，“Zugspitze” が最初にランキングされた結果が出力される．その根拠である，“日本で最も高い山は富士山である” や “The highest mountain in Germany is Zugspitze” などの文も出力される．

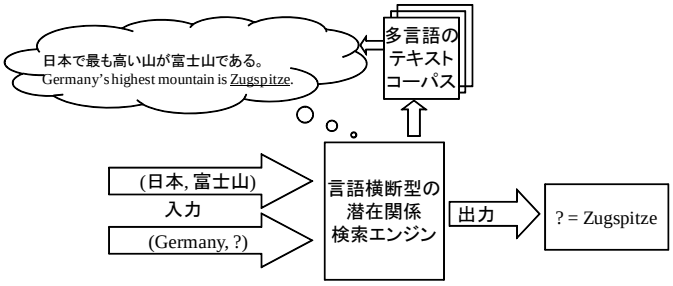

図 1 言語横断型の潜在関係検索の例

本稿では言語横断型の潜在関係検索の実現手法と評価結果について述べ，単言語の潜在関係検索の既存研究との比較を行う．

2. 言語横断型のエンティティと関係の Indexing

潜在関係検索を行うために，Web などのテキストコーパスからエンティティを発見し，エンティティ間の関係を抽出する必要がある．また，高速な検索を実現するには，予めエンティティと関係を Indexing する必要がある．そこで，本研究は既存研究 [2] で用いられている，エンティティとエンティティ間の関係を Indexing する手法を利用する．この手法ではまず，テキストコーパス中の文書を文に分割し，各文に対して，固有表現抽出器 ☆ と形態素解析器 ☆☆ を適用する．これらの解析は高速かつ高精度に行えるので，大規模なコーパスを Indexing することが出来る．解析された文中に，2 つ以上のエンティティがあれば，エンティティの順序を保ったままエンティティペアを抽出し，Index に入れる．次に，エンティティ間の関係を認識するために，エンティティペアの 2 つのエンティティが出現した位置の文脈から，関係を特徴づける語彙パターンを抽出する．例えば，“大都市東京は日本の首都である” という文から，エンティティペア (東京, 日本) が抽出される．このペアにおける 2 つのエンティティ間の関係を表現する特徴量として，“大都市 X は Y の首都”，“X は Y の首都”，“X は Y の首都である” などの語彙パターンが抽出される (ここで，語彙パターンが特定のエンティティペアに依存しないように，ペアの第 1 のエンティティを X に，第 2 のエンティティを Y に置き換える)．ここで注意すべき点は，語彙パターンはエンティティペアの真ん中の語彙だけではなく，前後の語彙も含んでいることである．これにより，エンティティペアが出現した文脈情報が関係の情報に含まれる．言語横断型の潜在関係検索を行うために本研究では，言語横断型のエンティティと関係の Indexing 手法を提案する．この手法では，エンティティペアや語彙パターンが異なる言語で書かれても同じ Index に保存し，同じように扱うことができる．図 2 に言語横断型の Index の例を示す． Index 中には，エンティティペアと語彙パターンの共起頻度の情報が入っている．異なる言語のエンティティペアと語彙パターンが同じように扱えるので，エンティティペア (東京，日本) と (Birkirkara, Malta) の特徴ベクトル間のコサイン類似度が 0 よりも大きくなる．これは日本語の文章中で，“マルタ” という固有名詞が “Malta” と記述されることがあるため，エンティティペア (Birkirkara，Malta) と語彙パターン “X は Y の首都” の共起頻度が 0 よりも大きいからである．故に，エンティティペア (東京，日本) と (Birkirkara，Malta) の関係類似度をコサイン類似度で定義すると，両者がある程度類似することが分かる．

語彙パターン \ エンティティペア	(東京, 日本)	(ビルキルカラ, マルタ)	(Tokyo, Japan)	(Birkirkara, Malta)
XはYの首都	80	95	0	5
XはYの最大都市	70	60	0	3
X is the capital of Y	0	0	50	40
Y's capital is X	0	0	30	35
X is the largest city in Y	0	0	20	15

0ではない

共起頻度

図 2 言語横断型の Index

3. 対訳語彙パターンと言語横断型の言い換え表現の発見

言語横断型の Indexing 手法を適用しても，異なる言語間で類似するエンティティペア間の共通の語彙パターンは依然として少ないため，関係類似度は正確に計算できない．そこで，語彙パターンの対訳を特定し，対訳関係にあるパターンを同じパターンとして扱う必要がある．本研究では，対訳パターンを特定するためにまず，統計翻訳システム ☆☆☆ を用い，各語彙パターンを英語に翻訳する．統計翻訳システムの出力語彙パターンが既に Index の中にあれば，翻訳システムがよい結果を出した可能性が高い．その理由は既に Index の中にあるパターンは，コーパス中の文章に出現したパターンなので，実際に用いられるパターンである可能性が高いからである．故にこの場合，翻訳システムの入力パターンと出力パターンが，実際に対訳関係にあるパターンであると仮定する．

翻訳システムで対訳関係にあるパターンを特定しても，類似エンティティペア間で共通のパターンが少ない可能性が高い．なぜなら同じ関係を表す言い方は複数あるからである．例えば，首都の関係を表現するものとして，“X は Y の首都” や “Y の首都が X” などの表現が存在する．更に，言語を跨った言い換え表現もあるが，統計翻訳システムは 1 つの対訳結果しか出力しない．そこで本研究では，語彙パターンのクラスタリングにより，言語横断型の言い換え表現発見の手法を提案する．語彙パターンのクラスタ

† 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.
{duc@mi.ci.i, danushka@iba.t, ishizuka@i}.u-tokyo.ac.jp

☆ http://nlp.stanford.edu/software/CRF-NER.shtml

☆☆ http://mecab.sourceforge.net/, http://nlp.stanford.edu/software/tagger.shtml

☆☆☆ http://translate.google.com

リングは先行研究 [1, 5] において，同じ言語での言い換え表現や意味的に類似する語彙パターンの認識に使われている．本研究では言語を跨った類似の語彙パターンを認識するために，ハイブリッドなクラスタリング手法を提案する．

まず語彙パターンのクラスタリングを行うために，各パターンの特徴ベクトルはその語彙パターンと共起するエンティティペアの共起頻度ベクトルとする．また，語彙パターン間の類似度を特徴ベクトルのコサイン類似度として定義する．提案するクラスタリングアルゴリズムでは，まずステップ 1 として，先行研究 [1] で用いられた逐次クラスタリング手法を使い，同じ言語において意味的に類似する語彙パターンを 1 つのクラスタにまとめる．この手法では各語彙パターンについて，既に作られた各語彙パタークラスタの重心とこの語彙パターンとの類似度を計算し，類似度が最大となる語彙パタークラスタを求める．最大となった類似度がある閾値 θ_1 よりも大きければ，該当クラスタにこの語彙パターンを追加する．どのクラスタにも追加されなかった語彙パターンは，新たな語彙パタークラスタとする．言語横断型の検索では対訳を持つ語彙パターンが重要であるが，対訳パターンを持つ語彙パターンの多くは類似度が θ_1 未満であるため，既存のクラスタに追加されないことが多い．これは，対訳を持つ語彙パターンは複数の言語のエンティティペアと共起すると見なされるので，ある特定の言語のエンティティペアとしか共起していないパターンとのコサイン類似度が低いからである．故に，検索の性能を上げるために，対訳や対訳に近い語彙パターンを，ステップ 1 で作られた適切なクラスタに入れる必要がある．例えば，語彙パターン “X is the capital of Y” を語彙パタークラスタ {X は Y の首都である; Y の首都が X; Y の首都 X} に入れると，首都の関係の検索結果が向上する．そのために，クラスタリングアルゴリズムのステップ 2 では，対訳を持つパターンだけを考え，ステップ 1 で作られたクラスタとの類似度を計算し，類似度がある閾値 θ_2 以上なら，該当するクラスタにこのパターンを入れる．ステップ 2 のクラスタリングはハードクラスタリングであるステップ 1 とは異なり，ソフトクラスタリングである．即ち，類似度が閾値 θ_2 以上のすべてのクラスタに，対訳のパターンを入れる．これは言語横断型の検索の再現率を上げるためである．また，上記の理由でステップ 2 での類似度閾値 θ_2 は，θ_1 よりも低く設定する必要がある．

4. 候補検索とランキング

上記のプロセスで検索のための Index が作られる．クエリ {(A, B), (C, ?)} が入力された時にまず，この Index から入力されたソースペア (A, B) を検索し，語彙パターンを取得する．次に，キーエンティティ(C) を含むペアで，(A, B) の語彙パターンと同じ語彙パターンか，同じクラスタに入る語彙パターンを 1 つ以上持つペアを検索する．この操作は，語彙パターンからそのパターンと一緒に出現したエンティティペアへの転置 Index を用いて，高速に行われる．これらのペアを候補ペアとする．候補ペアをランキングするために，エンティティペアの特徴ベクトル (語彙パターンとの共起頻度) のコサイン類似度を用いる．ただし，コサインを計算する際に，同じ語彙パタークラスタに入る語彙パターンは同じパターンと見なす．これで類似するエンティティペアが異なる言語にあっても，ステップ 2 のパタークラスタリングで同じクラスタに入る語彙パターンが存在するので，候補ペアになり，かつ，関係類似度が高いと期待できる．

5. 実験による評価と既存研究との比較

5.1 パラメータ調整

先行研究 [2] に従い，単言語の検索性能に最適なパラメータ θ_1 の値を 0.4 に設定して，パラメータ θ_2 の調整を行った．実験では，まず，12000 個の英語のウェブページと 6000 個の日本語のウェブページを含むパラメータ調整用のコーパスを解析し，Index を作成した．コーパス中には，既存研究 [2] で用いられた 4 種の関係 (CEO-会社， 会社-本部地，人-出身地，会社買収) のエンティティペアが入っている．これらの 4 種の関係は関係抽出，関係類似度計算や関係検索でよく用いられる評価データである [1, 6, 2]. また，評価用の指標として，各クエリセットの平均逆順位 (MRR) を用いる．クエリセット Q におけるクエリ q の最初正解のランクが r_q なら，$\mathrm{MRR}(Q) = \frac{1}{|Q|}\sum_{q\in Q} 1/r_q$ である．評価クエリセットは 8 つあり，その内 4 つが英語-日本語 (ソースペアが英語で，キーエンティティが日本語) のクエリセットであり，残りが日本語-英語のクエリセットである．1 つのクエリセットには 50 個のクエリがある．評価指標はこの 8 個のクエリセットの平均性能である． 図 3 に θ_2 を変化させながら性能を測定した結果を示す．$\theta_2 = 0.15$ で最大の MRR が得られた．故に，以降の実験では $\theta_1 = 0.4$, $\theta_2 = 0.15$ に設定する．また，θ_2 が 1.0 に近づくと MRR が低くなり，ステップ 2 のクラスタリングがない場合の MRR に近い．従って，提案手法のステップ 2 が言語横断型検索の性能に大きく寄与することが分かる．

5.2 ベースライン手法との比較

ベースライン手法ではまず，コーパス内の文書をすべて英語に翻訳する．そしてクエリ処理の際，クエリ全体を英語に翻訳してから，翻訳されたコーパスで単言語の潜在関係検索 [2] を行う．目的の言語が日本語であれば，検索結果をまた日本語に翻訳する．評価用のコーパスとしては，6000 個の英語の文書と 6000 個の日本語の文書を含むコーパスを用いる．クエリセットは前節で説明した 8 つのクエリセットと同じ関係の種類を含むが，エンティティペアが異なる．実験の結果，ベースライン手法の MRR が 0.345 であったのに対し，提案手法は平均で 0.430 の値を得た．従って，提案手法はベースラインよりもよい性能を出すことが分かった．

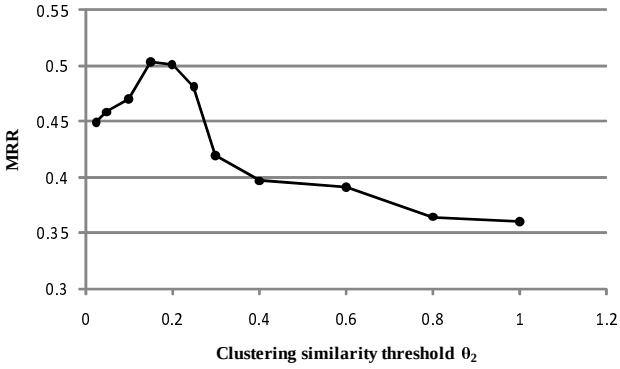

図 3 ステップ 2 の類似度閾値 θ_2 と検索エンジンの性能との関係

表 1 提案手法と既存手法の性能比較
(TopN は トップ N 個の結果に正解があるクエリの割合)

Method	MRR	Top1	Top5	Top10	Top20
KatoTC-日本語 [3]	0.379	25.0	55.3	60.3	67.3
KatoCNJ-日本語 [3]	0.545	43.3	68.3	72.3	76.0
KatoTC-英語	0.332	20.0	50.0	65.6	71.1
MonoRS-英語 [2]	0.963	95.0	97.8	97.8	97.8
CLRS (言語横断)	0.430	31.0	60.0	68.8	73.3

5.3 平均性能と既存研究との比較

提案手法の有効性を評価するために，単言語の潜在関係検索の既存研究 [3, 2] との性能比較を行った． 表 1 に結果を示す．表の最初の 2 行に，Kato ら [3] による，単語共起の情報 (Term Co-occurrence) を利用する単言語検索エンジンの性能 (KatoTC-日本語, KatoCNJ-日本語) を示す．また，我々が Kato らの手法を再実装し，本研究における単言語クエリセット (英語) で実験した結果 (KatoTC-英語) が 3 行目に示されている．この時使ったのと同じ英語クエリセットで提案手法を評価した結果が，MonoRS-英語 [2] である．最後に，CLRS は提案手法の言語横断型の検索の実験結果である．表から分かるように，提案手法の言語横断型の性能は単言語の性能よりも低いが，単語の共起を用いる手法 (TC) における単言語検索の性能よりも高い．

6. 終わりに

本稿では，言語横断型の潜在関係検索という検索手法を提案し，その実現手法を述べた．言語横断型の潜在関係検索は目的の言語でのエンティティの表記を知らない場合や，入力エンティティペアが目的の言語で書かれていない場合に有効である．また，根拠となる文は互いに対訳関係にある可能性が高いので，言語横断型の潜在関係検索によりユーザの翻訳作業を支援することが出来る．今後は潜在関係検索の人間の翻訳作業やパラレルコーパス作成への応用可能性を調べ，大規模なコーパスにおける検索エンジンを実現する予定である．

参 考 文 献

1) Bollegala, D., Matsuo, Y. and Ishizuka, M.: Measuring the Similarity between Implicit Semantic Relations from the Web, *Proc. of WWW'09*, ACM, pp. 651–660 (2009).
2) Duc, N., 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).
3) Kato, M. et al.: Query by Analogical Example: Relational Search using Web Search Engine Indices, *Proc. of CIKM'09*, pp. 27–36 (2009).
4) Goto, T., Duc, N., Bollegala, D. and Ishizuka, M.: Exploiting Symmetry in Relational Similarity for Ranking Relational Search Results, *Proc. of PRICAI'10*, pp. 595–600 (2010).
5) Bollegala, D., Matsuo, Y. and Ishizuka, M.: Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web, *Proc. of WWW'10*, ACM, pp. 151–160 (2010).
6) Banko, M. and Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction, *Proc. of ACL'08*, pp. 28–36 (2008).