1V-6

潜在関係検索エンジンにおける根拠となる文のランキング

平嶋一貴 † 河崎翔 †† Nguyen Tuan Duc ††
Danushka Bollegala †† 石塚満††

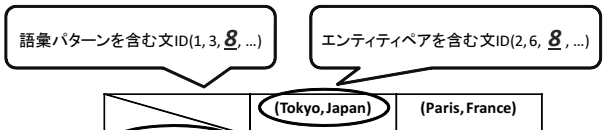
† 東京大学工学部電子情報工学科 †† 東京大学大学院情報理工学系研究科

1. はじめに

潜在関係検索エンジンとは、例えば、{(Tokyo,Japan), (Paris, ?)} というクエリに対して、まず (Tokyo,Japan) から首都-国という関係を抽出し、これを基にして「?」に当たる [France] という単語を検索するものである。既存の潜在関係検索では、それぞれのペア間の潜在関係を得るためにテキストコーパスから参考にした文が根拠の文として出力されているが、根拠の文が意味関係との関連度に基づいてランキングされておらず、数多くの曖昧な根拠の文が出力される。そこで、その例文を見ただけではどんな関係を元にして検索結果を得たのか分かりづらい。例えば “I was born in Tokyo, Japan.” という例文からは [Tokyo, Japan] について「国と国の都市」という関係は見えてくるが「国と国の首都」という関係を見出すことはできないので、証拠文としてあげるには不適切である。そこで本研究では、関係検索エンジンで得られた例文を評価し、より関係性が深い例文を上位にランキングする手法を提案する。例えば “Tokyo is the capital of Japan.” という例文を “I was born is Tokyo, Japan.” という例文よりも上にランキングする。以降、第 2 節では潜在関係検索エンジンの関連研究を紹介する。第 3 節には提案手法をあげる。第 4 節では評価のための実験と結果を説明する。第 5 節にはまとめと今後の課題について述べる。

2. 関連研究

本研究では、Duc らの既存の潜在関係検索エンジン [1, 2] の実装を基にして、実装を行った。既存の検索エンジンはエンティティ間の関係類似度をその周辺の文脈の語彙パターンで判断し高精度なエンティティランキング結果を得ているが、根拠文のランキングの性能は低い。Duc らの根拠文を取得する仕組みを **図 1** に示す。



	(Tokyo, Japan)	(Paris, France)
X is the capital of Y	8	10
Pattern 2	2	5
...	...	...
Pattern n	...	...

図 1 既存の検索エンジンの仕組み

全ての文に ID をつけ、ある語彙パターンとあるエンティティペアが共起する文 ID を記録する。図 1 では注目している語彙パターン (X is the capital of Y) について文 ID8 番の文が証拠文の候補となる。既存の検索エンジンでは根拠文の候補の集合からランダムに文を出力するが、本研究ではここで根拠文として関連度が高いもの順にランキングする。

3. 提案手法

本研究では大きく分けて二つの方法で例文を評価する。まずは語彙パターンのランキングを作り、それをもとに評価する方法である。次に文とエンティティペアとの関連性に注目してそれをもとに評価する方法である。

3.1 語彙パターンのランキング

ここではまず語彙パターンをランキングすることからはじめる。語彙パターンとはエンティティペアの間に現れる表現のパターンのことでペアをそれぞれ X と Y とおき “X is Y” , “X is one of the Y” のように表される。例えば [Tokyo, Japan] の間には数多くの語彙パターンが存在するが、これらを関連度を元にランキングすれば “X is the capital of Y” というパターンが上位に来ると予想できるので、このランキングを使って例文を評価すれば関係をより明確に表した例文を上位に提示できると考える。そこで相互情報量を利用した語彙パターンをランキングする手法を紹介する。エンティティペア (A, B) とパターン P について関連度 r(P,(A,B)) は相互情報量 I を用いて

$$r(P,(A,B)) = I(P,(A,B)) - |I(P,A) - I(P,B)| \quad (1)$$

と表せる。I(P,A) と I(P.B) をそれぞれ考慮するのは A,B の片方のみが多くの情報量をもっている場合は適していないからである。よって r(P,(A,B)) は p(P|A) を用いて

$$r(p,(A,B)) = log(p(P|(A,B))) - |log(p(P|A)) - log(p(P|B))| \quad (2)$$

と表せる。この時、p(P|A) は

$$p(P|A) \approx \frac{\text{count of P in contexts retrieved for A}}{\text{no. of contexts retrieved for A}} \quad (3)$$

で定義される。本手法は**図 2** に図示したように まず、全パターンの内 X と Y の間に in, on, of といった stop words のみを含むパターンを除いて、パターン P_i についてこの $r_1(P_i,(A,B))$, $r_2(P_i,(C,D))$ を取得しそれぞれのパターンについて r_1 と r_2 の平均を最終的なスコアとし高い順からそのパターンとエンティティペアを含む文を取得する。これによりエンティティペア (A,B) とエンティティペア (C,D) の両方と関連度が深いパターンが抽出でき、より重要な例文が取得できる。

3.2 文とエンティティペアとの関連性

続いて文においてエンティティペアがどのような役割を担っているかに焦点を当てた考え方である。

3.2.1 Tf - Idf

まず文章においてその単語自体がどれほど重要であるかの指標

Ranking Evidence Sentences in Latent Relational Search
† Kazuki Hirashima (Dept. of Information and Communication Engineering, Faculty of Engineering, The Univ. of Tokyo)
†† Sho Kawasaki, Nguyen Tuan Duc, Danushka Bollegala, and Mitsuru Ishizuka (Grad. School of Information Science and Technology, The Univ. of Tokyo)

Lexical Pattern	(Tokyo,Japan)	(Paris, France)	Score
p 1= x * largest city * y	r_{AB_p1}	r_{CD_p1}	S(p1)
p2 = x * south of * y	r_{AB_p2}	r_{CD_p2}	S(p2)
p3 = x * capital of * y	r_{AB_p3}	r_{CD_p3}	S(p3)
p4 = x * area of * y	r_{AB_p4}	r_{CD_4}	S(p4)
:	:	:	:

⇩

スコアが高い順にランキング

Score	Lexical Pattern	Sentence
S(p3)	p3 = x * capital of * y	Tokyo is the capital of Japan.
S(p1)	p 1= x * largest city * y	Paris is the largest city in France.
S(p2)	p4 = x * south of * y	I traveled from Tokyo to the south of Japan.
:	:	:

図 2 語彙パターンランキングでの文の取得

として使用するのが Tf - Idf 方式である。これは、Tf (Term Frequency:単語の出現頻度) と Idf (Inverse Document Frequency: 逆文書頻度) という二つの指標にもとづいて計算され、Tf だけで評価するのではなく、Idf が一種の一般語フィルタの役割を果たしているので多くのドキュメントに出現する一般的な語は重要度が下がり、特定のドキュメントにしか出現しない単語の重要度を上げ、総合的に単語の重要度を評価するというものである。

3.2.2 文の主語修飾語

文における単語の重要度をはかる際、その単語が文構成の中心にいるほど重要であると考えられる。つまり調べたエンティティペアが文の主語修飾語であるような文章ではそのエンティティペアについて重要なことが書かれてあると考えられる。この手法を**図 3** に図示した。 [Tokyo, Japan] の場合も、 “Tokyo is the capital of Japan.” という文章の方が、 “I was born in Tokyo, Japan” という文よりも、ペアにとってより重要なことが書いてあるようにこのフィルタを使えばペア間の関連度が高い例文が拾えると考えられる。実装面では構文解析として stanford parser を組み込み、A,B のいずれかが文の一番上の名詞句に含まれているものに、より高いスコアを与える仕様である。

○(high score) ***Tokyo*** is the capital of ***Japan***.
S V

× (low score) I was born in ***Tokyo***, ***Japan***.
S V

図 3 文とエンティティペアとの関連性によるスコアリング

4. 評　　価

まず、第 3.1 節で述べたパターンランキング手法を使い、良い証拠文は上位にランキングできるかを検証する。評価のために、[3] を参考にして 5 種の関係種類のクエリセットを利用する。5 種の関係とは、会社買収関係 (Google - Youtube), 国の首都 (Paris - France), 会社の社長 (Eric Schmidt - Google), 国の大統領 (Barack Obama, United States) と 惑星の衛星 (Ganymede, Jupiter) である。これらの関係種類は関係抽出システムの評価 [4], 関係類似度計算アルゴリズムの評価 [5, 6] や既存の潜在関係検索エンジンの評価 [1, 2, 3, 7] によく用いられる関係種類である。評価データとして、Wikipedia のテキストコーパスを 7,000,000 個使い、そこから 213,731,476 つの文が得られ、30,778,223 個の単語ペアが抽出された。抽出された語彙パターン数は 945,857,689 つであった。これらに関して既存のエンジンと提案手法を実装したエンジンにそれぞれ上記のクエリセットを投げかけて返ってきた例文を調べ、求めている関係性を含む例文が何番目に表示されているかの平均逆順位 (MRR) を**表 1** に示した。既存の検索エンジンでは 2 で書いたように文 ID を取得した順に例文を取得しているので語彙パターンは言わばランダムに並んでいる。既存手法では 0.096 程であった MRR が提案した手法では 0.613 と大幅に改善された。本システムが良い結果を出したのはエンティティペア間の関係性を示したパターンのランキングの精度がよい効果だと考えられる。従って本提案手法の一部だけを使ってもかなり良い結果を出すことが出来た。第 3.2 節で提案している文の関連度に着目した手法については、今回の実験ではまだ利用していないが、今後このスコアを組み込んで、より性能をあげられると考える。

表 1 評価結果：既存手法と提案手法の MRR

関係種類	既存手法	提案手法
会社買収関係	0.087	0.575
首都 - 国	0.048	0.55
社長 - 会社	0.064	0.628
大統領 - 国	0.127	0.649
衛星 - 惑星	0.156	0.664
平均	0.096	0.613

5. まとめと今後の課題

本稿では、潜在関係検索エンジンで取得できる証拠文のランキング手法について説明し、手法として「語彙パターンのスコアリング」と「エンティティペアの文における役割に注目する」という主に 2 つの方法を提案した。前者では与えられたエンティティペアとよく共起するパターンを相互情報量によりスコアリングしてランキングを作りそれをもとに例文を取得するという方法である。後者はエンティティペアが文の重要な部分であるような例文に高いスコアを付けるという方法である。今回は前者のみの実装であったが、結果として正しい根拠文の MRR が既存の検索エンジンに比べかなり改善された。今後は後者の手法も実装し更に評価を改善していきたい。

参　考　文　献

1) 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).
2) グェン トアンドゥク, ボレガラダヌシカ, 石塚満: エンティティペア間類似性を利用した潜在関係検索, 情報処理学会論文誌, Vol. 52, No. 4, pp. 1790–1802 (2011).
3) Duc, N. T., Bollegala, D. and Ishizuka, M.: Cross-Language Latent Relational Search: Mapping Knowledge across Languages, *Proc. of AAAI'11*, pp. 1237–1242 (2011).
4) Banko, M. and Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction, *Proc. of ACL'08*, pp. 28–36 (2008).
5) 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).
6) Bollegala, D., Matsuo, Y. and Ishizuka, M.: Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web, *Proc. of WWW'10*, pp. 151–160 (2010).
7) Kato, M. P., Ohshima, H., Oyama, S. and Tanaka, K.: Query by Analogical Example: Relational Search Using Web Search Engine Indices, *Proc. of CIKM'09*, pp. 27–36 (2009).