潜在的関係を利用する検索システムの対称性による候補語リランキング法

後藤 友和 † Nguyen Tuan Duc† Danushka Bollegala† 石塚 満 †
東京大学大学院情報理工学系研究科 †

1 はじめに

全文検索エンジンを用いて情報を探すとき，利用者の知識を必要とすることが多い．例えば，lion に対する ostrich と bird の関係と同じものを調べるとき，ostrich と bird の関係を明示出来なければ調べることはできない．そこで，関係を明示しなくても調べることを可能とするのが Relational Search と呼ばれる検索システムである．Relational Search に対して {(ostrich, bird), (lion, ?)} という 2 つの単語対からなるクエリを使ったとき，cat という語が期待できる．何故なら，"?"に cat という語を入れた場合，2 つのワードペア間のそれぞれの関係は「…は大きな…である」となり，関係が似たものになるからである．このように Relational Search は，"?"に語を入れたときに，与えられたワードペア間での関係類似度が高くなるものを返すシステムである．

本論文では，Relational Search を Web 検索エンジンを用いて実装する手法を述べる．さらに，得られた候補語に対して Relational Seach に存在する対称性を用いてリランキングを行った．そして，アメリカの大学入試問題である SAT のアナロジー問題を用いて Relational Search システムの有効性，スコアリングの有効性，およびリランキング手法の有効性を評価した．

2 Relational Search

Relational Search の基礎となる考え方は Veale.T [2] や Danushka ら [3] が述べている．また，最初の実装としては Kato ら [1] の研究がある．Relational Search は 2 つの単語対を入力として持つ．ここでは 2 つの単語対を $\{(A,B),(C,D)\}$ と表す．ここで，A,B,C,D はそれぞれ単語を表す．この単語対において，D が分からないとき，それを"?"としておき，そこに当てはまる語を求めるのが Relational Search である．

Relational Search を行う上では，関係類似度という考え方が大事になる．関係類似度は Danushka ら [3] や Turney[4] によって用いられてきた．関係類似度とは，単語間の類似度ではなく，単語対間の関係の類似度を測ったものである．Relational Search において入力を $\{(A,B),(C,?)\}$ としたとき，"?"に当てはまる語として適しているのは，A,B 間の関係と C,X 間の関係の類似度が高くなるような語 X である．

3 Web 検索エンジンを用いた Relational Search の実装

Relational Search システムを実装するためには，与えられたワードペア間の関係を取得する必要がある．そこでまず，単語対 $\{(A,B),(C,?)\}$ について考える．A,B 間の関係 R_{AB} を抽出するために我々は検索エンジンとして Yahoo! Search BOSS[1] を用い，そこから得られるスニペットを用いた．スニペットとは，検索クエリが含まれている箇所を提示する短い文章である．我々は A と B を含むスニペットを得るために Yahoo! Search BOSS に対して"A * * * B"というクエリを投げた．*は 1 つもしくは 0 の単語にマッチする．そのため，このクエリでは A と B の間に存在する 3 単語以内の単語列を含んだスニペットを得ることができる．*を単語 w_1，w_2，w_3 に置き換えて得られた単語列の集合 (パターン) を P とする．A，B およびパターン $p\,(p \in P)$ を含む文字列 "A p B"に対して A の部分を C に置き換え，B の部分を*に置き換えたものをクエリとする．そして，このクエリを Yahoo! Search BOSS に投げて再びスニペットを取得する．このスニペットから，*に当てはまる語を抽出し，候補語とする．この候補語集合 X に含まれるそれぞれの候補語 $x\,(x \in X)$ に対し，式 1 によるスコア付けを行い，ランキングを作成する．

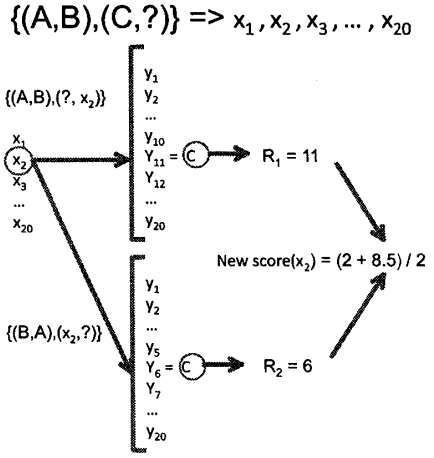

図 1: $\{(A,B),(C,?)\}$ に対するリランキング

$$score(x) = \frac{\sum_{p \subseteq P_x}(freq(\text{"}C\ p\ x\text{"}))}{freq(\text{"}C\ *\ *\ *\ x\text{"})} \quad (1)$$

ここで，$freq(\text{"}C\ p\ x\text{"})$ は "$C\ p\ x$" で検索したときのヒット件数を表す．p は x と共に出てきたパターンの集合 P_x に含まれるパターンである．$freq(\text{"}C\ p\ x\text{"})$ の値が大きい時，C と p と x の関係が強いことを意味する．ここでは，$p\,(p \in P_x)$ に対し，$freq(\text{"}C\ p\ x\text{"})$ を取得し，その合計を取ることで C,p,x 間の関係の強さを表している．また，正規化を行うために C と x が 3 つ以内の語で区切られている場合全ての頻度で割っている．

4 対称性を用いたリランキング法

Relational Search において，単語対の構成に対称性を見出すことができる．一例として {(ostrich, bird),(lion,?)} という Relational Search への入力を考える．そして，"?"に当てはまる語として cat が得られたとする．このとき，Relational Search への入力として {(ostrich, bird),(?,cat)} を与える．そうした場合，ostrich と bird の関係と lion と得られた語 cat の関係が類似しているときには，"?"には lion が来ることが予想できる．もし，lion と cat が ostrich と bird の関係を満たしていなければ，"?"には lion が来ることはない．本論文ではこの対称性を用いることで検索結果のリランキングを行った．

クエリペア $\{(A,B),(C,?)\}$ を Relational Search に投げて得られた候補語集合を X とする．この X に含まれる候補語 x に対して式 1 を用いて得られたスコアを昇順に並べたものを

Exploiting Relational Symmetries for Re-ranking Latent Relational Search Results

Tomokazu Goto, Nguyen Tuan Duc, Danushka Bollegala, and Mitsuru Ishizuka
Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Tokyo 113-8656, Japan

[1] http://developer.yahoo.com/search/boss/

初期ランキングとする．本論文では，このランキング上位20件に対してリランキングを行った．

$\{(A,B),(C,?)\}$ に対するリランキングを図1に示す．まず，上位20件に含まれる x を用いて，クエリペア $\{(A,B),(?,x)\}$ を作成し，Relational Search に投げる．この時，得られた候補語の集合を Y とする．次に，各候補語 $y(y \in Y)$ に対して，式2によるスコア付けを用いてランク付けを行う．

$$score(y) = \frac{\sum_{p \subseteq P_y}(freq(\text{"}y\ p\ x\text{"}))}{freq(\text{"}y\ *\ *\ *\ x\text{"})} \quad (2)$$

そして，このランキング内で C が現れる順位を記録する．この順位を R_1 とする．同様にして $\{(B,A),(x,?)\}$ から候補語を取得し，式2を用いて得られたランキングに対しても C が現れる順位を取得し，順位を R_2 とする．そして，R_1 と R_2 を平均したものと，x の順位との平均を取ったものを x の新たなスコアとする．同様の手順をクエリペア $\{(A,B),(?,D)\}$ に対しても行い，D が現れる順位の平均と，候補語の順位を平均したものを新たな候補語のスコアとする．そして最後に，$\{(A,B),(C,?)\}$ の時に D が現れる平均順位と，$\{(A,B),(?,D)\}$ の時に C が現れる時の平均順位をさらに平均して，最終スコアとした．この最終スコアを昇順に並べたものをリランキング後のランキングとした．

5 実験

本手法の精度を測るために，本研究では SAT データセットを用いた．SAT データセットは SAT の問題のうち，アナロジー問題のみを374件集めたものである．Danushka ら [3] や Turney[4] の手法においても関係類似度を測るためのベンチマークとして使われてきた．正解率だけを見れば良いため，容易に手法の比較ができる．この SAT データセットには問題となるワードペアがあり，その関係と同じものを探すための選択肢となるペアが各問題に対して5件出現する．例えば，問題として，(ostrich, bird)，選択肢として，1.(lion, cat), 2.(goose, flock), 3.(ewe, sheep), 4.(cub, bear), 5.(primate, monkey) があり，この場合は 1.(lion, cat) が正解となる．ostrich は大きな bird であり，lion もまた大きな cat であるからである．本論文では，この SAT データセットを用いて達成できる正解率を見ることで，システムの有効性，スコアリングの有効性，およびリランキング手法の評価を行った．

Relational Search を用いてアナロジー問題を解くために，問題単語対と，選択肢ペアをクエリとした．そして，選択肢ペアのうち一つを"?"とし，Relational Search に投げた．例えば問題：(ostrich, bird)，選択肢の一つ：(lion, cat) の場合，$\{(ostrich,bird),(lion,?)\}$ と $\{(ostrich,bird),(?,cat)\}$ がクエリとなる．この場合，得られたランキングにおける *cat* と *lion* の順位の平均がこの選択肢のスコアとなる．このスコアを他の選択肢と比較し，最も低いものを正解候補した．そして，この正解候補と実際の正解を比較し，正解数を記録した．

5.1 実験結果

SAT データセット全374件に対して実験を行った結果，リランキング前の正解数は105件で，リランキング後は114件の正解となった．それぞれ 28.1%と 30.5%の正解率である．しかし，"?"と置き換えた語（上記の例の場合，*cat* または *lion*）が一つも見つからない場合，本手法では正解か不正解の判断が付かない．それらの問題を除くと，243件あり，その場合の正解率はそれぞれ 43.0%と 46.9%となった．374件全てに対して答えの候補が見つかったとすると，全体の正解率も 46.9%である．さらに，リランキング対象となったものとして，正解を含み，候補が2つ以上あるものを見ると，216件あり，その中でリランキング前の正解は78件，リランキング後の正解は87件あり，正解率はそれぞれ 36.1%と 40.3%となり，リランキングによって 4.2 %正解率が上がった．ここで，正解候補が2つ以上あるものを選んでいるのは，候補が1つしかないものに関してはリランキングを行っても正解となるペアが変わらないためである．

表 1: SAT データセットにおける正解率

Algorithm	score
ランダム	20.0%
提案手法	46.9%
MSW [3]	51.1%
LRA [4]	56.1%
人間	57.0%

他の関連研究との比較を表1に示す．MSW と LRA はそれぞれ Danushka ら [3] と Turney[4] が提案した手法である．ランダムに SAT の問題を解いたときは，選択肢が各問いに対して5問あるので，20%の確率で正解でき，この問題のベースラインとなる．また，実際に SAT を解いた大学受験者の正解率が 57.0%である．本手法では，ベースラインよりも上位に位置したが，他の手法よりは低かった．

6 考察

実験においては，提案手法は他の研究に比べて高い正解率を得ることはできなかった．しかし，Relational Search を用いて SAT 問題を解く試みは本研究が最初である．また，本研究で提案した対称性を用いたリランキング手法に関しては 4.2%の正解率の向上を見ることができた．対称性は Relational Search を行うときに特に有効な手段である．単に (A,B),(C,D) 間の関係類似度を測る際には (B,A),(D,C) の対称性のみ使用することができる．一方，Relational Search では $\{(A,B),(C,?)\}$ から得られた候補語集合 X に含まれる x $(x \in X)$ に対して $\{(A,B),(?,x)\}$, $\{(B,A),(x,?)\}$, $\{(x,C),(A,?)\}$, $\{(x,C),(?,B)\}$, $\{(C,x),(?,A)\}$, $\{(C,x),(B,?)\}$ の6種類をリランキングに用いることができる．このことから，対称性が有効に活用できるのは Relational Search を行った場合だといえる．本論文においては，A と B が最初の単語対となる2つの対称性しか用いていないが，A,B 間のパターンはあらかじめキャッシュすることで高速化を行っていたためである．もし，x を用いた場合，キャッシュ量が膨大になり，キャッシュを貯めるために多くの時間が必要となる．

7 おわりに

本論文では，Relational Search システムの実装方法，評価方法，システムの特性，システムの特性を利用したシステムの改善方法について述べた．SAT データセット全体においては先行研究より低い結果であったが，対称性を用いた候補語のリランキングを行うことによって正解率が向上することが確認できた．

参考文献

[1] Kato, M. P., Ohshima, H., Oyama, S. and Tanaka, K. Query by analogical example: relational search using web search engine indices. In *Proc. of CIKM '09*, pages 27–36, 2009.

[2] T. Veale. The analogical thesaurus. In *Proc. of 15th IAAI' 03*, pages 137–142, 2003.

[3] D. Bollegala, Y. Matsuo and M. Ishizuka. Measuringthe similarity between implicit semantic relations from the web. In *Proc. of WWW'09*, pages 651–660, 2009.

[4] P. Turney. Similarity of semantic relations. *Computational Linguistics*, 32(3):pages 379–416, 2006.