Exploiting Symmetry in Relational Similarity for Ranking Relational Search Results

Tomokazu Goto, Nguyen Tuan Duc, Danushka Bollegala, and Mitsuru Ishizuka

The University of Tokyo, Japan

Relational Search

Relational search is a novel search paradigm. For the query {(A,B), (C,?)}, in which A, B, and C are input words, a relational search engine finds the words D such that the relation between A and B is also held between C and D.

{(ostrich, bird),(lion,?)}


⇔ Bird

⇔ Cat ?

Realization

(ostrich, bird), (lion, ?)

"ostrich * * * bird"

↓ YAHOO!

the ***ostrich is the largest bird*** in size and weight on earth. **...**

↓

"α is the largest β" ← (lion, ?)

↓

"lion is the largest *"

↓ YAHOO!

the ***lion is the largest cat***, while … → **cat**

Scoring

$$score_{init}(D) = \frac{\sum_{p \subseteq P_D} (freq(``p[C/\alpha, D/\beta"))}{freq(``C * * * D")}$$

P_D are the patterns that appeared with D, the formula p[C/α] represents the substitution of α by C in the pattern p, freq("p[C/α, D/β]") is the frequency of co-occurrences of the word D with the word C and other words in the patterns.

the ***ostrich is the largest bird*** in **...** → "α is the largest β"

the ***ostrich is the big bird***, → "α is the big β"

Taking **ostrich pictures of this bird**, → "α pictures of this β"

the ***lion is the largest cat***, while … → "α is the largest β"

the ***lion is the big cat***, … → "α is the big β"

$P_{lion,cat}$ → {"α is the largest β", "α is the big β"}

freq("lion is the largest cat") = 200
freq("lion is the big cat") = 100
freq("lion * * * cat") = 100000
score(cat) = (200 + 100) / 100000

Symmetries

Let us denote the relational similarity between (A, B) and (C, D) by R((A, B), (C, D)). Relational similarity will remain unchanged under certain permutations of the four words (e.g., R((A, B), (C, D)) = R((B, A), (D, C))). Therefore, the candidates that are ranked at the top by one form of the query (e.g., (A,B),(C,?)) must also be ranked at the top by the other (alternative) forms of the query (e.g., (B,A),(?,C)). In other words, if D is an incorrect candidate, then it will be ranked at the top only in a small number of alternative forms of the query and it will receive bad ranks in almost all alternative forms.

{(ostrich, bird),(lion,?)}

↓ Searching and sorting by the scoring formula

D = {human, cat, animal...}

Is really human better than cat?

{(ostrich, bird),(?, cat)} ... lion is ranked 1
Average rank of cat is (2 + 1) / 2 = 1.5

{(ostrich, bird),(?, human)} ... lion is ranked 20
Average rank of human is (1 + 20) / 2 = 10.5

→ **Cat is better!**

Evaluation

SAT word analogy questions

The SAT dataset contains 374 word analogy questions selected from the Scholastic Aptitude Test. Each questions has a question word pair (stem pair) and five choices for answer word pairs. In the five candidate answers, only one answer is correct. The correct answer word pair has highest similarity with the stem pair. In the below example, the correct answer is (1) (lion, cat) because the relation between lion and cat is most similar to that between ostrich and bird (ostrich is a large bird, whereas lion is a large cat).

Stem pair	Ostrich	Bird
Q.1	Lion	Cat
Q.2	Goose	Flock
Q.3	Ewe	Sheep
Q.4	Cub	Bear
Q.5	Primate	monkey

Results

Criterion	Initial	Using symmetry
# correct answers / questions(recall)	28.1%	30.5%
# correct answers / # questions that **we can get D**(precision)	43.0%	46.9%
# correct answers / # questions that we can retrieve **the correct choice** and **at least one other choice**	36.1%	40.3%