2D2-04

助詞で結合された名詞句の意味的な類似判別法

Calculation Method of the Degree of Semantic Similarity between Noun Phases Consisting of Nouns and Particles

佐藤 雅彦　上原 勇樹　石川 勉
Masahiko Satou　Yuuki Uehara　Tsutomu Ishikawa

拓殖大学 工学部 情報工学科
Department of Computer Science, Takushoku University

We propose a method to measure the degree of semantic similarity between noun phases, which consist of some pairs of noun and particle. In this method, the degree of similarity can be calculated accurately, without depending on the numbers of nouns and their order. This method has to be modified when applied to ARSK(Approximate Reasoning with Similar Knowledge), which is an inference mechanism coping with incompleteness of knowledge-base by using similar knowledge. This paper also describes how to modify the proposed method for the application to ARSK.

1. はじめに

インターネットの普及により電子化文書が広く流通する現在, 必要な情報をより効率的に獲得するため, 各種情報検索さらには自動要約[1]が注目されてきている. これらの技術では, 文章間あるいはその構成要素 (主に名詞句) 間の意味的な類似性判定が重要な課題となり, その研究も盛んとなってきている[2].

一方, 我々はこれらへの応用も意識した, 知識の類似性を利用して概略解を得る推論法 (ARSK) を研究している [3].ARSK では知識は拡張型述語論理 [4] で表現され, 述語部や引数部の主構成要素としては単語や複合語だけでなく名詞句も想定しているため, その類似性判定を如何に精度良く行なうかが課題となっている.

本報告では, これらへの適用を想定した名詞句間の類似性判別法について提案する. なお, ここでは“ ～の～の～ ”のように名詞と助詞の繰り返しからなる構造の名詞句を対象とする.

2. 名詞句間の類似度計算法

2.1 基本的な考え方

前述したように, ここでは以下のような名詞と助詞の繰り返しからなる名詞句を対象とする.

$$\text{名詞句：}\underbrace{\text{名詞+助詞+}\cdots\text{+名詞+助詞}}_{\text{修飾部}}+\underbrace{\text{名詞}}_{\text{主部}}$$

このような名詞句では, 全体の意味を決定する主要的な部分は主部であり, 修飾部はこの意味を限定するようにはたらく. 従って,2 つの名詞句 A,B 間の全体の類似度 R(A,B) は, 主部間の類似度 R(Ac,Bc) がベースであり, 修飾部間の類似度 R(As,Bs) はそれを低下させる要素となる. 従って,R(A,B) の算出では以下の条件が必要である.

$$R(A,B) = R(Ac,Bc) \times f(R(As,Bs)) \qquad (1)$$
$$R(A,B) \leqq R(Ac,Bc)$$

上式において,f(R(As,Bs)) としてどのような関数を設定するかが重要となるが, ここでは, 以下のように設定する.

$$f(R(As,Bs)) = 1/(1+(1-R(As,Bs)))$$

こうすると R(A,B) は最大で R(Ac,Bc), 最小でその半分となる. なお(1−R(As,Bs))の意味するところは, 修飾部の意味的な違いの度合い(非類似度)とも言え, 合理性もある考えられる.

連絡先: 拓殖大学 工学部 情報工学科
〒 193-0985 東京都八王子市館町 815-1
E-mail: y3m312@st.takushoku-u.ac.jp

2.2 修飾部間の類似度計算法

修飾部間の類似度 R(As,Bs) の計算については,“ 名詞+助詞 ”を 1 つのセットとして考えるべきで, セット間の類似度をもとめ, それを統合することになる. これは, このセットは文節に相当し意味的に分離すべきでないからである.

また, 対象となる名詞句にこのセットが複数あるときは, セット間の最適な対応を決定し, そのときのセット間の類似度を統合する必要がある. 例えば,

①母から兄への手紙
②先生から兄への手紙
③兄貴宛ての田舎のおふくろからの手紙

において, ①と②では, セットの数, 順番が同じであるが意味的には大きく異なるのに対し, ①と③ではこれらは異なるが意味的には非常に類似している. すなわち,R(As,Bs) の計算では, セットの数や順番の違いに依存しない方法をとる必要がある. このような計算法として, ここでは以下の方法を提案する. なお, 名詞句 A,B のセット数をそれぞれ m,n (n < m) とする.

＜修飾部間の類似度計算法＞

Step.1：2 つの名詞句の修飾部から,i 個ずつのセットを選択する. 選択したセットを 1 対 1 で対応させてセット間の類似度を求め, 平均値をそれらセット間の総合の類似度とする.

この操作をセット間の対応を変えて, 可能な組合せの数だけ行い, その中で最大となった組合せの値を, 選択されたセット間の類似度とする.

Step.2：次に,Step.1 で選択されなかったセット間の可能なすべての組合せの類似度を計算しその平均値を, 選択されなかったセット間の類似度とする.

Step.3：Step.1 で選択されたセット間の類似度と Step.2 で求めた類似度の平均値を修飾部間の類似度とする.

Step.4：Step1～3 を可能な選択ｎ Ci ×ｍ Ci 通りについて計算する. ｉは１からｎ－１まで変えて行い, その最大値を最終的な修飾部間の類似度とする.

例えば, 名詞句として以下の 2 つがあり (a～k はセット),

名詞句 A の修飾部　：　a, b, c, d
名詞句 B の修飾部　：　e, f, g, h, k

i=2 で A からは b,c, B からは g,h が選ばれた場合,Step.3 までで得られる修飾部間の類似度 R(bc,gh) は以下のようになる.

$$R(bc,gh) = ave(max(ave(Rbg,Rch),ave(Rbh,Rcg)), ave(Rae,Raf,Rak,Rde,Rdf,Rdk))$$

2.3 一方の名詞句に修飾部がない場合の計算法

この場合の類似度をどう定義するかは難しい問題である. 例えば,“ 机の上のテレビ ”と“ テレビ ”の関係であり, 前節の計算法をそのまま適用すると類似度は 0.5 となる. ここでは, このような場合の類似度は 1, すなわち主部間の類似度を名詞句全体の類似度とすることにした.

2.4 セット間の類似度計算法

修飾部のセット間の類似度計算では, 対応するセットの名詞間, 助詞間で類似度を求める. 名詞間については概念ベース [5] を利用する. 概念ベースでは各概念 (単語) が 2,715 次元のベクトルで表されており, 類似度は比較する概念のベクトル間での内積で求める. 助詞間については, 国語辞書に掲載されている助詞の意味 (働き) を参考にして, 類似度を,5 段階で人為的に決定した表 1 の値を用いる. セット間の類似度は名詞間の類似度と助詞間の類似度の積とする.

表 1　助詞間の類似度

	の	についての	に関する	への	からの	による	としての	までの	での
の	1	0.8	0.8	0.2	0.4	0.4	0.6	0.2	0.6
についての	0.8	1	1	0	0	0	0	0	0
に関する	0.8	1	1	0	0	0	0	0	0
への	0.2	0	0	1	0	0	0	0	0
からの	0.4	0	0	0	1	0.2	0	0	0.2
による	0.4	0	0	0	0.2	1	0.2	0	0
としての	0.6	0	0	0	0	0.2	1	0	0
までの	0.2	0	0	0	0	0	0	1	0.2
での	0.6	0	0	0	0.2	0	0	0.2	1

3. ARSK への適用法

前章の類似度計算法は文章間の意味的な類似度を計算するような場合を想定している. すなわち,ARSK のような SLD 法に基づく推論処理の中で, 名詞句間の類似度を計算する場合には若干修正する必要がある. 以下, これについて述べる.

ARSK では“ ～A ”と“ A’ ∨ B ”から, 導出節として“ B ”を導くが, このとき A と A ’の類似判定が必要となる (そのときの状況からみて A A’ なら導出を行う). この A や A’ の素式の主構成要素は名詞句であり [4], 名詞句間の類似度計算が必要となるが, この場合文章の比較等と異なり, 方向性があることに注意しなくてはならない. すなわち, 導出は～A と A’ の両立が矛盾することに基づくのであるから,A の方が A’ より抽象的 (上位概念) なら, それら名詞句は意味的に同一として扱っても問題は生じないこととなる.

例えば,～A 内,A’ 内の名詞句としてそれぞれ,“ 鳥の巣 ”,“ カラスの巣 ”があったとする. この場合は,“ 鳥” は“ カラス” の上位概念なので, この名詞間の類似度は“ 1” とする必要がある. なお, これが逆の場合には ARSK の思想に基づいてそれらの間 (この場合は“ 鳥” と“ カラス”) の類似度を計算することになる. これらは比較する名詞が上位/下位の関係だけでなく, 普通名詞/固有名詞間の比較においても同様である. すなわち,A の方が普通名詞,A’ の方が固有名詞の場合には, 類似度は“ 1” とする. この操作は, 比較する名詞句に含まれるすべての名詞間について行う必要がある.

4. 評価

2 章で述べた類似度計算法の評価を行った. 評価方法は, まず新聞・インターネット等から集めた検索用の名詞句を 20 個, 被検索用名詞句を 250 個用意した. 次に, 検索用名詞句から, 意味的に類似する名詞句を 1 つずつ作成し, これを被検索用名詞句の集合に加えた. これは, 被検索集合の中に類似する名詞句が存在しない場合があると考えたからである. 図 1 に検索イメージの例を示す. また, 図 2 に検索用名詞句を“ 酒の人体への影響 ”としたときの検索結果を示す. 1 位には主部が異なるが, 名詞句全体としての意味が類似している名詞句があがってきていることが分かる.2 位,3 位には主部が同一の名詞句が上がっているが, 本計算法では修飾部間が全く非類似の場合, 全体の類似度が 0.5 となることから, あがってきたと考えられる.

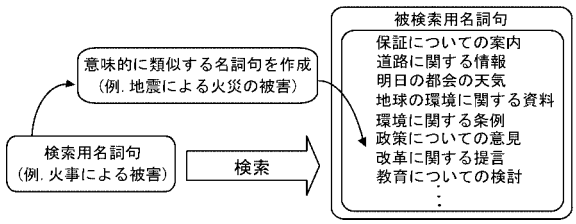

図 1　検索イメージ

検索文：酒の人体への影響
検索順位：1．『人間の健康へのアルコールの効果』の名詞句間の類似度＝0.535199
2．『戦争による観光への影響』の名詞句間の類似度＝0.501045
3．『生活への影響』の名詞句間の類似度＝0.500816
4．『今日までの成果』の名詞句間の類似度＝0.104115

図 2　検索結果

表 2 に検索用名詞句 20 個について行った検索実験の結果を示す. これらに意味的に最も類似する名詞句が検索された順位毎の数を示したものである.20 個のうち 9 個が 1 位に検索され,4 位以内に 15 個が検索されている.

表 2　検索実験の集計結果

検索順位	検索された名詞句の個数
1	9
2	4
3	2
4	0
5位以下	5

5. さいごに

名詞と助詞の繰り返し構造の名詞句を対象とした名詞句間類似度計算法について提案した. さらにこの方法を概略推論法“ ARSK” へ適用する場合の方法を示した. また, この方法をインターネット上の情報から収集した名詞句を利用した検索実験を通し評価し, その有効性を示した. なお, 本計算法は形容詞や動詞の連体形を含む一般的な名詞句にも若干の変更で適用可能であり, これについては別途報告する.

参考文献

[1] 難波英嗣・奥村学：“ここまで来たテキスト自動要約”, 情報処理学会誌 Vol.43 No.12-002 pp.1287-1294(2002)

[2] Wakiyama, M., Noda, H., Nozaki, K.,Kawaguchi, E., : “Computation Algorithm of Semantic Difference Measure in the SD-Form Semantics Model”, IPSJ Journal Vol.40 No.03-032 pp.1065-1079(2001)

[3] NguyenVietHa, 石川勉, 阿部明典: “知識の類似性を利用した概略推論法の研究”, 電子情報通信学会論文誌 D-I Vol.J84-D-I No.4 pp.389-400(2001)

[4] 石川勉, 佐々木智彦, 佐藤雅彦：“言葉をベースとする拡張型述語論理形式の知識表現法”, 第 16 回ことば工学研究会資料 P25～32 (2004)

[5] NguyenVietHa, 帆苅譲, 石川勉, 笠原要: “単語の意味の類似性判別のための大規模概念ベース ”, 情報処理学会論文誌 vol.43 No.10 pp.3127-3136 (2002)