

英文科学技術抄録文からのオントロジー自動作成の試み

6 E-4

柴田 誠[†] 辻 龍彦[‡] 竹田 正幸[†] 松尾 文碩[†][†]九州大学工学部 [‡]大村市役所

1. まえがき

自然言語処理において、単語の分類を必要とすることが多い。しかし、人手によって分類を行うことは大変困難な作業である。特に、科学技術分野では、語彙はたえず増加していることと、科学技術用語を理解できる者は専門家に限られるため、更に困難である。本稿では、英文科学分野のオントロジー自動作成の試みについて述べる。

2. オントロジー作成手続き

オントロジーを語彙の順序集合と考え、語彙の類似性を位相数学の開集合によって規定した場合の数学的性質は文献3であたえた。

そこでは、共通の性質をもつ集合を開集合であると仮定する。ある世界についてのすべての開集合が観測されたならば、上記の性質に基づき、その観測からオントロジーをつくることができる。次の手続きは、ボトムアップ的にオントロジーをつくるものである。

<開集合族からオントロジーをつくる手続き>

すべての開集合の族を Ω と書く。集合 A の要素数を $|A|$ で表わす。 $\max_{A \in \Omega} |A| = m$ とする。また、 $O_k(A) \iff |A| = k \wedge A \in \Omega$ 。 Y を集合の集合とし、 Y の要素から和集合をとる演算によってつくられるすべて集合の集合を Y の和集合閉包といい、 $\bigcup^+(Y)$ で表わす。

- ステップ1 $X = \emptyset$, $Y = \emptyset$, $\Gamma = \emptyset$, $k = 1$.
- ステップ2 $O_k(A)$ のとき、 $A = \{a\}$ ならば、 $X \leftarrow X \cup \{a\}$, $Y \leftarrow Y \cup \{\{a\}\}$ とせよ。 $O_k(A)$ であるすべての A について、この手順を繰り返せ。

- ステップ3 $k \leftarrow k + 1$. $k > m$ ならば、終了せよ。

- ステップ4 $Y \leftarrow \bigcup^+(Y)$ とせよ。 $O_k(A)$ のとき、ある $B \in Y$ が存在して、 $A - B = \{a\}$ ならば、 $X \leftarrow X \cup \{a\}$, $Y \leftarrow Y \cup A$ とせよ。 $b \in B \cap X \wedge \neg \exists x (b \triangleleft x \notin \Gamma)$ であるすべての b に対して、 $\Gamma \leftarrow \Gamma \cup \{b \triangleleft a\}$ とせよ。 $O_k(A)$ であるすべての A について、この手順を繰り返し、ステップ3に行け。□

この手続きによって、 Ω からオントロジー $(X, \triangleleft)$ が得られる。関係 $\triangleleft$ は、 Γ によってあたえられる。

3. 英文科学技術文からのオントロジー自動作成の試み

2節の手続きは、二つの仮定に基づいていて、ステップ4のように差集合がシングルトンかどうかをみるような神経質な判定を含んでいるため、現実的な作成法とはみなしがたい。そこで、前節の手続きを基に現実的なオントロジー自動作成法を開発するための指針を得るためにINSPECテープの抄録文から、オントロジーをつくることを試みた。方針は次のとおりである。

「主語 + be 動詞 + 属性形容詞」の文型をとる場合には、同じ属性形容詞をとる主語の主語の集合を開集合と考え、オントロジー作成を試みた。また、「主語 + be 動詞 + 冠詞 + 名詞」の文型の場合には、主語と補語がISA関係にあると考え、その抽出を行った。

資料には英文科学技術抄録文(INSPEC)の25年分を用いた。ここで、対象が広がりすぎるのを防ぐために文献の分野を「太陽系」に関する161,770文献のデータだけに限定した。

まず、本研究室で開発した名詞決定法と動詞決定法を用いて文章中の主語 - 補語関係を特定した。この結果、37,017文献のデータを得ることができた。

4. 主語 - 属性形容詞からのオントロジー

まず、文章から主語・属性形容詞の対を抜き出した。ここで、以下の形容詞は「属性形容詞」とみなさ

Toward Automatic Construction of Ontology from Scientific and Technical Documents

Makoto Shibata[†], Tatsuhiko Tsuji[‡], Masayuki Takeda[†] and Fumihiro Matsuo[†]

[†]Kyushu University 36, Hakozaki, Fukuoka, 812 Japan

[‡]Omura City Office, 1-25, Kushima, Omura, Nagasaki, 856 Japan

なかった。

- 形容詞の比較級・最上級
- 他の単語との比較を前提とした形容詞 (same, other, etc.)
- 明確な属性を表さない形容詞 (such, any, etc.)
- 「動詞 + 形容詞 + 前置詞」などの形で熟語をつくるもの (be able to, be possible to, etc.)

また、先頭が大文字の主語は固有名詞であると考えられるために除外した。こうして主語と属性形容詞の対 3,838 組が抜き出された。

これらの対から、同じ属性形容詞をもつ主語の集合をつくる。方針は次の通りである。

- (1) Swallow is swift と Airplane is swift のような文があれば, swallow と airplane は似ているとする。すなわち, $\{x|x \text{ is/are swift}\}$ を一つの開集合とみなす。そこで, 形容詞の個数だけ開集合が存在することになる。
- (2) Swallow is black and swift などのように形容詞が接続詞で連結されている文では最初の形容詞だけをとり, Swallow is black とみなす。
- (3) Swallow is not yellow における not は無視した。not だけではなく, すべての副詞を無視した。

(2) と (3) の方針をとったのは, まだ連言/選言句の範囲決定ができないこと, scarcely などの準否定の副詞の意味をとらえることができていないことなどによる。

結果として, 形容詞の異なり語数は 586 であり, x となる名詞の異なり語数は 1,340 であった。そこで, 各形容詞には 1,340 からなる名詞の集合の部分集合が対応する。これらの部分集合の包含関係を調査したところ, 1,542 個の包含関係が存在していることがわかった。この関係から開集合の順序構造を求めることができる。その結果, $\prec$ 関係のある名詞の対は 27,986 組であった。

5. ISA 関係

次に, ISA 関係を抽出した。抽出法は, 名詞修飾語や副詞を無視し, ‘名詞 1 is a/an 名詞 2’ の形の文があれば, 名詞 1 と名詞 2 には ISA 関係があるとした。連言/選言句の処理は形容詞の場合と同じである。また, ここでも先頭が大文字の名詞は除外した。その結果, ISA 関係にあると考えられる名詞の対が 1,136 組抜き出され, 異なりは 996 であった。

これらの対から, 名詞の階層構造の作成を試みた結果の一部を図 1 に示す。系列の最後に括弧で囲まれた名詞が出現しているのは, 既に出現した名詞が再び出現したことを示す。すなわち, サイクルが生じたことを意味している。図 1 において比較的長い系列が存在することは, 注目に値する。

absorption $\succ$ line $\succ$ asymmetry $\succ$ feature $\succ$ time
 $\succ$ energy $\succ$ flare $\succ$ component $\succ$ velocity $\succ$ flux $\succ$
 effect $\succ$ anomaly $\succ$ (line)

absorption $\succ$ line $\succ$ asymmetry $\succ$ feature $\succ$ time
 $\succ$ energy $\succ$ flare $\succ$ component $\succ$ velocity $\succ$ flux $\succ$
 effect $\succ$ circulation

図 1 名詞の階層構造

また, $\prec$ 関係があり, かつ ISA 関係があるものは 75 組で, 異なり名詞数は 94 であり, いずれも 1/10 程度に減少した。

6. むすび

オントロジーの自動作成の実験を行ったが, 良質なオントロジーを作成するには至らなかった。集合の包含関係では, 包含する集合 A に比べて, 包含される集合 B の要素数が小さく, 上位概念となるべき単語が出てこない, という問題がある。また, ISA 関係として取り出した中にも実際には ISA 関係にないものも含まれていた。

原因としては, 連言/選言句や副詞の問題などにより, ‘主語-補語’の正確な関係がとらえられていないことがあげられる。今後, これらの問題を解決し, 良質なオントロジーを作成する試行を行う予定である。

なお, 本研究は, 一部文部省科学研究費補助金 (#07408007) の援助により行った。

参考文献

- 1) Dahlgren, K. and McDowell, J.: Knowledge Representation for Commonsense Reasoning with Text, Computational Linguistics, 15, 3, 149-170 (1989).
- 2) Rosch, E., Mervis, C., Gray, W. D., Johnson, D. M., and Boyes-Braem, P.: Basic Objects in Natural Categories, Cognitive Psychology, 8, 382-439 (1976).
- 3) Matsuo, F.: Topology on Ontology, Memoirs of the Faculty of Eng., Kyushu Univ., 54, 1, 25-29 (1994).
- 4) 久富健治: オントロジー自動作成についての研究, 九州大学大学院工学研究科電気工学専攻修士論文, 1994