日本語テキストからの用語関連構造の自動抽出

7K－11

服部　真也，峯崎　俊哉，成嶋　弘

東海大学

1. はじめに

学習をどの順序で行っていくかということは非常に重要な問題である．学習要素の順序を把握することは，指導者にとっても，学習者にとっても重要なポイントとなり，そのためには対象教科の学習順序を考慮した教材の構造化（ネットワーク表現）が有効である[1] [2]．

本稿では，文書を読み理解をして知識を得るということを学習としてとらえ，『Aという事柄を理解するためにはBという事柄を理解している必要がある』という文書の内容把握に欠かすことのできない重要な用語の関連構造を自動的に抽出するための一手法を提案する．

2. 牽連チャート

牽連チャートとは，文書の学習順序を考慮した用語のつながりを重み付き有向グラフ（ネットワーク）で表わしたものである（図1）．

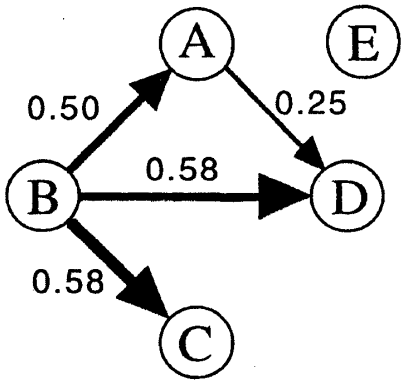

図1．牽連チャート

例えば，図1の牽連チャートからは，「Dを理解するためにはA，Bを理解している必要があり，しかも，Aを理解するためにはBを理解している必要があるので，Dを理解するためにはB，A，Dの順序で学習する必要がある」ということがわかる．また，牽連チャートの各辺の数値（重み）は関連の強さを表わしたものである．

このように，文書の内容を牽連チャートで表わすと各学習要素に対する順序を把握することができる．次節以降では，牽連チャートを自動生成するための具体的な手順を示し，評価及び問題点について述べる．

3. 学習要素の抽出

牽連チャートを作成するにあたって，まずは学習要素を抽出する必要がある．この学習要素は，一般に用いられているキーワードとは異なった基準で抽出する．これは，例えば

S1：Aはオブジェクト指向言語の一種である．

S2：私はAを好んで使っている．

のような2つの文があったとき，S1ではAがなんであるかを説明しているのでAは学習要素と成り得るが，S2ではAがなんであるかを説明していないので，Aは学習要素には成り得ない．つまり，同じAという語でも出現する文によって学習要素と成り得るときと成り得ないときがある．このことを考慮して，**Siの学習要素語**を以下のように定める．

【Siの学習要素語】

S1, S2, ... , Snを入力対象とし，Siがある事柄α, β, γ, ...について説明している文であるならばα, β, γ, ...をSiの学習要素語と呼び，その集合をσ(Si)とする．

The Extraction of Related Terms from Japanese Text
Shinya Hattori,Toshiya Minezaki,Hiroshi Narushima
Tokai University
1117 Hiratsuka , Kanagawa 259-12 , Japan

本手法では，Siの学習要素語の抽出は複雑な構文解析などは行わずに，我々が実験によって得た文型パターンとの照合により抽出するという方法を試みている．

4. 学習要素の関係付け

抽出された学習要素語の間には何の関係付けもなされていないので，直接関係による学習要素語間の階層的関係付けを行う必要がある．ここで，２つの学習要素語α，βがあったときα→βという関係の意味は，『βを理解するためにはαを理解している事が必要である』ということである．このことを**αがβに関連している**という．また，その関連の強さを**αのβに対する関連度**といい，**R(α→β)**で表わす．

【関係付けのアルゴリズム】

S1, S2, ... , Sn を入力対象とする．

$\alpha, \beta \in \bigcup_{i=1}^{n} \sigma(Si)$に対して$R(\alpha \rightarrow \beta)$を0に初期化する．

$\alpha \in Si - \sigma(Si)$ かつ $\beta \in \sigma(Sj)$ならばα→βとなり，$R(\alpha \rightarrow \beta)$を$\frac{1}{|i-j|+3}$加算する．

ただし，i=jでαとβの間に１個以上の助詞のみがあるならば$\frac{1}{2}$を加算する．

5. 実験結果と評価

提案した関係付けのアルゴリズムを用いて実際に文献[4]に対して実験を行った結果を表１に示す．実験に用いたテキストの内容はオブジェクト指向言語に関するもので，文書の量は925文字６段落からなるものであり，抽出した学習要素語は以下のようになる：

A：オブジェクト指向機能，B：Smalltalk，

C：self，D：Dylan，E：Sather

また，表１の結果を牽連チャートとして表わしたものが図１である．

本手法の有効性を確認するために評価を行った．評価は人間が作った牽連チャートと同等な牽連チャートをこのアルゴリズムで作成できるかという点について行い，具体的な評価基準は人間が作った牽連チャートと比較して「適切な学習要素が抽出されているか」と「適切な関係付けが行われているか」の２点とした．その結果，自動生成した牽連チャートは人間が作った牽連チャートと比べ，やや精度の面で劣ることもあるが，中心となる関係付けの再現はほぼ確実に出来るということが確認できた．

このことから，本手法は日本語テキストの用語の関連構造をとらえるのに有効な手段と考えられる．

	A	B	C	D	E
A	0.00	0.00	0.00	0.25	0.00
B	0.50	0.00	0.58	0.58	0.00
C	0.00	0.00	0.00	0.00	0.00
D	0.00	0.00	0.00	0.00	0.00
E	0.00	0.00	0.00	0.00	0.00

表１．実験結果

6. 今後の課題

今後はさらに精度の高い牽連チャートの自動生成を目指し，(1)実験によってさらに多くの文型パターンを取得する，(2)**R(α→β)**の値による構造性質の分析を行う，などについて検討していきたいと考えている．

参考文献

[1] 佐藤隆博：ISM構造学習法，明治図書（1987）

[2] 成嶋弘，西山英樹：順序集合論の学習プログラム構造解析への試み，早稲田大学数学教育学会誌，第7巻，第1号，pp.41-59（1989）

[3] DAVID ELLIS著，細野公男監訳：情報検索論，丸善株式会社（1994）

[4] C MAGAZINE 創刊５周年記念号，p67，ソフトバンク（1994）