専門用語における階層関係及び関連関係抽出法

○ 頼 静娟†
王 暁晶†
藤原 譲†

Extraction of Hierarchical and Associative Relationships in The Standard Technical Terms

Jingjuan Lai†
Xiaojing Wang†
Yuzuru Fujiwara†

The importance and the necessity of the organization of information have been increasingly recognized. It is reasonable that organization of terms as units for expressing concepts is an important method for organization of information. The method of SS-KWIC extracting hierarchical relationships and associative relationships automatically based on the coining rules in terms is discovered in our reseach. The principle and realization of method of SS-KWIC are shown, and a example of the experiment results is given in this paper. The experimental results that detailed hierarchys and rich associative relationships are obtained, proved reliability and practicality of the method.

1　はじめに

近年情報の高度利用のため意味的関係を体系化する必要性及び重要性がますます認識されるようになってきた。情報ベースシステムの研究開発においては、従来のアクセス重視型から、情報体系化による学習や類推や帰納など高度な機能重視型に変える動向は著しく脚光を浴びている。ところが、情報には様々な形態や類別がある。形態によって、文字、数式、音声、写真、図面、画像などに分けられる。類別としては、例えば、文字データに対して、本、雑誌、新聞、冊子、辞書などがある。相違する形態の媒体によって体系化法も異なる。本文では、文字データのみを議論の対象とする。

本研究において文字情報の体系化は一般的な書誌情報に対する体系化と、概念に対する体系化および、因果関係を主とした論理表現に対する体系化に大別する。この中で、概念を表現する最小単位としての用語の体系化は最も重要視されている。開発中の用語体系自動構築システムにはC-TRAN法（Constrained Transitive Closure）、SS-KWIC法（Semantically Structured Key Word element Index in Terminological Context）、SS-SANS法（Semantically Specified Syntactic Analysis）及び意味的手法のSANS法（Semantic Analysis of Sentences）、INTEGRAL法（Integration of Domain Established Knowledge）による同値関係、階層関係と関連関係などを自動的に抽出する諸機能がある。本研究では、SS-KWIC法の原理及び実現法の高精度化を示し、実験結果の例を挙げる。

†筑波大学 電子・情報工学系
†Institute of Information Science and Elecrtonics, University of Tsukuba

2　SS-KWIC 法と専門用語

SS-KWIC 法は用語の構成規則に基づいて、用語を適当に分解することによって階層関係および関連関係を獲得する方法である。用語は単純語、畳語、擬音語、擬態語および合成語を含む。用語の構成規則はこの合成語に対する考察に由来する。合成語はおもに次のようなものを指す。

合成語::= 複合語 | 派生語
複合語::= 語基＋語基 | 語基＋連結要素＋語基
派生語::= 接辞＋語基 | 語基＋接辞
語基::= 単純語基 | 複合語基
単純語基::= 単純語
複合語基::= 語基＋語基
連結要素::= ・ | ／ | の | な
接辞::= 接頭語 | 接尾語 | 数詞 | 量詞

例えば、「有機化学」や「電子・情報工学」や「軍事同盟／政治同盟」などが複合語である。「全」や「非」や「形」や「一次」などが接辞である。合成語における修飾関係の種類によって、階層関係か関連関係かを判断する。

専門用語は非常に特徴がある用語である：

（1）ほとんど名詞である；

（2）後部分の体言類語基の性質や状態を、前部分の語基が修飾したり、限定したりする修飾関係が最も多い；

（3）用語が長い。専門用語においては、用語によって表される概念が不明確であったり、概念間に混同が生じたりしてはならない。従って、用語は曖昧さを排除し、他の用語との区別を明らかにするために、長い合成語になることが多い。

専門用語の特徴を考察した上で、われわれは専門用語の造語規則に基づく階層関係や関連関係を自動的に抽出する SS-KWIC 法を開発した[1]。この方法はまず専門用語を語基ごとに分解し、語基間の修飾位置関係や品詞などに基づいて、上位語か下位語かあるいは関連語かを決定し、階層構造および関連構造を構築する。FIG.1 に示されたのは SS-KWIC 法によって用語リストから階層関係を得る一例である。

従って、正しい階層関係や関連関係を得るには、基本用語集合の獲得および専門用語における合成語に対する正確な分割が必要である。

3　SS-KWIC 法のアルゴリズム

上述のように基本用語の選択と合成語の分割は意味関係の獲得においては大変重要であるが、困難さも同様に大きい。専門用語において、漢字一文字からなる基本用語は少なく、しかもこれを語基としての用語は遠い関連的な修飾関係がよくあるが、階層的な関

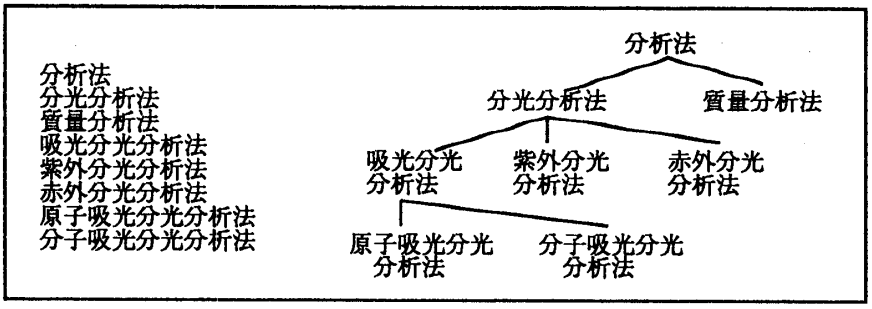

FIG.1: SS-KWIC 法によって階層関係を獲得する例。

係はあまりない。漢字二文字の基本用語は比較的数多く存在する。漢字三、四文字の基本用語もある程度あるが、五文字以上はごく僅かしか存在しない。従って、基本用語の選択基準は漢字二文字以上、四文字以下から構成される用語を基本用語として利用する。ただし、外来語あるいは平仮名のみの用語の場合、長さを 1 と見なす。合成語を分割する際、前処理で得られた基本用語集合と接辞集合に基づき、合成語のキー要素としての基本用語を用いて、まず前方からキー要素まで分解する。終了したら、キー要素から後ろの方へ同様の処理を行う。合成語間の意味関係を判断するにはおもに品詞情報に頼る。

アルゴリズムに使われる集合を先に定義しておく。

$Term_{set} = \{term_i = (e_m e_{m-1} ... e_1 e_0) | e_i$ が一個の文字, $i = 0, ..., m\}$,

$BaseTerm = \{bt_i | bt_i$ は基本用語, $bt_i \in Term_{set}, i = 1, ..., n\}$,

$SS = \{$ 接辞 | 連結要素 $\}$,

$KANA = \{$ 外来語 | 平仮名用語 $\}$,

$MIX = \{$ 漢字 | 漢字片仮名 | 漢字平仮名 $\}$.

三つの関数を次のように定義する。

$$LEN(term) = \begin{cases} 1 & term \in KANA \\ m & term \in MIX \end{cases}$$

$$TEST(term) = \begin{cases} 0 & term \in KANA \\ 1 & term \in MIX \end{cases}$$

$$SORT(x_1, x_2, ..., x_n) = \{x_1^*, x_2^*, ..., x_n^* | LEN(x_1^*) \geq ... \geq LEN(x_n^*)\}.$$

3.1　基本用語獲得アルゴリズム

Algorithm 1.

1. $BaseTerm = \phi$;
2. While $Term_{set} \neq \phi$,
   $\forall term \in Term_{set}$,
   if $(TEST(term) == 0) \| (2 \leq LEN(term) \leq 4)$　　$BaseTerm \Leftarrow term$.
3. end.

3.2　複合語分割アルゴリズム

合成語を分割する際、つぎの点を考慮に入れている。最も長い基本用語を優先的に使って分割を行う。例えば、基本用語の中で、$(x_i), (x_j x_i)$ が同時存在する場合、$(x_j x_i)$ が優先する。よって過剰な分割はしない。例えば、「情報処理アプローチ」を「情報処理|アプローチ」に分割するが、「情報|処理|アプローチ」には分割しない。

Algorithm 2.

(1)$SORT(BaseTerm)$;
(2)While $Term_{set} \neq \phi$,
$\forall term = (e_n e_{n-1} ... e_0) \in Term_{set}$,

$$BT_i = \bigcup_{x_i \in BaseTerm \cup SS} \{(e_n e_{n-1} ... e_t | x_i | e_r ... e_1 e_0)\}. \qquad (i = 1, ..., m)$$

(3) 各 BT_i に対して、
While $s \neq t$

$$BT_i = \bigcup_{x_s \in BaseTerm \cup SS} \{(x_s | ... e_t | x_i | e_r ... e_1 e_0)\}.$$

(4) 各 BT_i に対して、
While $k \neq 0$

$$BT_i = \bigcup_{x_k \in BaseTerm \cup SS} \{(x_s | ... | x_j | x_i | x_k | ... e_0)\}.$$

(5)end.

ここで、接頭辞と後継の語基とは後結合、接尾辞と前行の語基とは前結合によって処理を行なう。

3.3　階層関係判定アルゴリズム

上記の分割アルゴリズムによって、合成語集合は $BT_i = \{(x_l|...|x_0)$ に分割されたとする。

Algorithm 3.

各 BT_i に対して、$\forall x_0, ..., x_i, ..., x_s (\in BT_i)$ は名詞ならば、

(1)x_i は

$$\bigcup_{x_i \in BT_i} \{(x_s|...|x_j|x_i)\}.$$

の上位語；

(2)x_i は

$$\bigcup_{x_i \in BT_i} \{(x_i|x_k|...|x_0)\}.$$

の関連語；

(3)$\forall x_a, x_b \in BT_i, x_a \equiv x_b$（同値）ならば、$x_a x_i, x_b x_i$ は $\{...|x_a|x_i\}, \{...|x_b|x_i\}$ の上位語。

(3) は語基の位置が交換可能の場合を対処するための条件である。たとえば、「同期 | 誘導 | 電動機」と「| 誘導 | 同期 | 電動機」は、C-TRAN 法[2]によって両者が同義語であると分かったので、「同期 | 誘導 | 電動機」は「かご形 | 誘導 | 同期 | 電動機」の上位語と判断する。

「研究」のような多品詞の用語が存在するので、Algorithm 3. を使うと、「調査研究」のようなノイズを階層関係として抽出されてしまうので、次の補足アルゴリズムを与える。

Algorithm 3.1(completment)

各 BT_i に対して、$x_a, x_b \in BT_i$,

if ($x_a \in$ { サ変動詞 }) & ($x_b \in$ { 名詞 })

x_a は $(x_b|x_a)$ の上位語と認める.

4　実験結果及び議論

英日工業用語集を SS-KWIC 法の評価実験に用いた。大部分が専門用語によって構成されるほか、対訳関係をもつのがこの用語集の特徴である。上記のアルゴリズムを工業標準用語集の日本語部分に対して実験した結果の一例を Table.1 に示す。任意 1000 個抽出された階層関係に対する抽出正解率は 91% にも達していることが示された。

Table.1: Numbers of Word Elements in The Standard Technical Terms

	Total	Number of elements in terms					
		2	3	4	5	6	More than 7
Number of terms	45464	19330	4092	1007	228	50	9

マスク
ダ｜マスク
送気｜マスク
酸素｜マスク
防毒｜マスク
隔離式｜防毒｜マスク
直結式｜防毒｜マスク
直結式｜小形｜防毒｜マスク
防護｜マスク
送風｜マスク
送風｜機形｜ホース｜マスク
ホース｜マスク
肺力吸引形｜ホース｜マスク
防じん｜マスク
簡易｜防じん｜マスク
隔離式｜防じん｜マスク
直結式｜防じん｜マスク
シャドウ｜マスク
エアライン｜マスク
複合式｜エアライン｜マスク
一定流量形｜エアライン｜マスク

FIG.2: 実験結果の一例。

複写機
感熱｜複写機
写真｜複写機
青｜写真｜複写機
静電｜複写機
直接｜静電｜複写機
乾式｜直接｜静電｜複写機
湿式｜直接｜静電｜複写機
間接｜静電｜複写機
乾式｜間接｜静電｜複写機
湿式｜間接｜静電｜複写機
カラー｜複写機
ジアゾ｜複写機
乾式｜ジアゾ｜複写機
熱式｜ジアゾ｜複写機
湿式｜ジアゾ｜複写機
加圧式｜ジアゾ｜複写機
拡散転写｜複写機
スタビライズ｜複写機
ドライフォト｜複写機
ダイトランスファ｜複写機

FIG.3: 抽出した意味関係の一例。

分光分析法
赤外｜分光分析法
紫外｜分光分析法
発光｜分光分析法
吸光｜分光分析法
分子｜吸光｜分光分析法
原子｜吸光｜分光分析法
炎光｜分光分析法
原子螢光｜分光分析法
X線螢光｜分光分析法
核磁気共鳴｜分光分析法
電子スピン共鳴｜分光分析法

FIG.4: 抽出した意味関係の一例。

記数法
基数｜記数法
混合｜基数｜記数法
固定｜基数｜記数法
１０進｜記数法
純２進｜記数法
混合基底｜記数法

FIG.5: 抽出した意味関係の一例。

れんが
泡｜れんが
鋳造｜れんが
せき｜れんが
焼成｜れんが
不焼成｜れんが
中心｜れんが
湯道｜れんが
耐酸｜れんが
く形｜れんが
並形｜れんが
固形｜れんが
扇形｜れんが
炭素｜れんが
漏斗｜れんが
耐火｜れんが
耐火｜断熱｜れんが
異形｜耐火｜れんが
粘土質｜耐火｜れんが
高アルミナ質｜耐火｜れんが
技術｜れんが
化粧｜れんが
ハブ｜れんが
負荷｜れんが
攻め｜れんが
半枚｜れんが
取鍋｜れんが
つり｜れんが
電鋳｜れんが
中空｜れんが
穴あき｜れんが
ばち形｜れんが
標準形｜れんが
注入管｜れんが
けい石｜れんが
半けい石｜れんが
ろう石｜れんが
換気用｜れんが
造塊用｜れんが
ドーム｜れんが
クロム｜れんが
クロム質｜れんが
クロム｜マグネシア質｜れんが
裏張り｜れんが
内張り｜れんが
横ぜり｜れんが
縦ぜり｜れんが
ノズル｜れんが
ノズル受け｜れんが
チェッカ｜れんが
チェッカー｜れんが
耐火｜断熱｜れんが
せり受け｜れんが
羊かん形｜れんが
スリーブ｜れんが
マグネシア｜れんが
クロ・マグ｜れんが
ブルノーズ｜れんが
プレス成形｜れんが
ジルコン質｜れんが
ドロマイト｜れんが
安定化｜ドロマイト｜れんが
タール｜ドロマイト｜れんが
特殊な形の｜れんが
メタル｜ケース｜れんが
ジルコニア質｜れんが
炭化けい素質｜れんが
コンクリート｜れんが
ワイヤーカット｜れんが
フリン｜トライム｜れんが
珪酸カルシウム｜れんが
フォルステライト｜れんが
ロータリー｜キルン用｜れんが

FIG.6: 抽出した階層関係および関連関係の一例。

実験に使用した接頭辞および接尾辞はそれぞれ{非、不、再、若、安、真、各、第、約、全、半、副、助、主、無}と、{上、下、左、右、別、時、回、章、化、機、類、制、性、製、前、後、長、機形、質、式、形、種、系、重、次、元、型、部、法、系列、層、進、用、受け}である。

FIG.2～FIG.8に実験によって抽出された階層関係と関連関係の実例を示している。ここではごく一部の結果しか示せないが、SS-KWIC法は用語を適切に分割し、多くの意味関係が抽出できることから、簡単ながらも大変有効な方法であることが実証さ

れた。FIG.1 はSS-KWIC 法を説明するためのシミュレーションだが、実際の実験結果（FIG.4）のほうがより優れている。なお、SS-KWIC 法によって、C-TRAN 法で抽出された同義語集合のノイズを検出することができる。敢えて漢字一文字の用語をも基本用語として実験を行なってみたところ、FIG.7 のような階層関係も抽出されたが、多くはやはり FIG.8 のようなノイズであった。

雨
雷｜雨
除｜雨
暴風｜雨
熱帯性｜暴風｜雨

FIG.7: 抽出した意味関係の一例。

法
司｜法
熱量｜法
質量｜法
重量｜法
反射｜法
零｜位｜法
パルス｜反射｜法
（以下省略）

FIG.8: 抽出した意味関係の一例。

しかし、SS-KWIC 法はまだ改善の必要がある。たとえば、基本用語獲得方法の強化、品詞による判断精度の向上、省略語、簡易語の取り扱い方など。現段階の基本用語の獲得は専門用語集のみ利用しているが、これから、単純語辞書の取り入れによってもっといい結果が期待できる。

5　むすび

本研究では合成語の造語規則に基づいて専門用語における階層関係および関連関係を抽出する方法を提案した。通常シソーラスに含まれない細かい階層と豊富な関連関係が得られたことによってこの方法が有効かつ実用的であることが実証された。この方法においてまだ充実すべき部分が残っているが、専門知識に全面的に依存していた過去のやり方に対して、簡単明快かつコンピュータ処理向きの方法としては大いに期待できると言えよう。今後の課題としては、一般用語の合成語についての解析および処理条件の改善を行ない、一般用語へ応用することである。

参考文献

[1] Y.Fujiwara, J.J.Lai, and T.Makino: Management and Advvanced Utilization of Semantically Organized Terminology and Knowledge. Proceedings of TKE'93. 1993.

[2] J.J.Lai, H.Kitagawa and Y.Fujiwara: Structuralization of Information by the Automatically Constructed Thesaurus. Information Media, 7(4), 25-32, 1992.

[3] 西野哲朗, 藤崎哲之助: 漢字複合語の確率的構造解析. 情報処理学会論文誌, Vol.29 No. 11 Nov. 1988 pp.1034-43.

[4] 水谷静夫編修, 朝倉日本語新講座 1 : 文字・表記と語構成. 朝倉書店 1987.

[5] 岩波講座日本語 6 : 語彙と意味. 岩波書店, 1977 年.

[6] 宮崎正弘: 係受け解析を用いた複合語の自動分割法. Vol.25 No.6, Nov 1984.