

意味関係抽出による概念の構造化

† 近藤 雄裕、‡ 藤原 譲

Construction of conceptual structures by extraction of semantic relationships among term

† Takahiro Kondo, ‡ Fujiwara Yuzuru

Abstract

Semantic processing which uses a great deal of knowledge which was structuralized require to be equipped with learning and consideration functions in computers. The conceptual structure according to conceptual relation exists as a construction of structuralized information that is necessary to semantic processing. Systematizing terminology as minimum unit which expresses a concept is most important to semantic processing. It made conceptual structure which combined the C-TRAN method extracting equivalent relationships and the SS-KWEIC method which extracts hierarchy relationships and associative relationships, moreover extracted data by the SS-SANS method extracting associative relationships. Hence the new relationships are found among terminologies and judgment material for semantics of terminology increases. Finally it became possible to get more detailed terminologies to systematize.

1. はじめに

計算機に学習や思考機能を備えるには、構造化された大量の知識を用いた意味処理が必要になる。意味処理を行うのに必要である情報の構造化として、概念関係に対する概念構造が存在する。また、概念を表現する最小の単位としての用語の体系化は、意味処理を行うのにもっとも重要である。

しかしながら用語の意味的な位置づけの確定的な判断は難しい。なぜなら構造化において用語が適切でないと思われる箇所に存在してしまってもまったく可能性がないということは言い切れず、どこかにそのような関係が存在するまたは、これからそのような関係になるなどの可能性があるからである。それ故、用語には様々な関係がリンクされていることが重要となる。

本研究では同値関係を抽出する C-TRAN 法、階層関係と関連関係を抽出する SS-KWEIC 法を組み合わせることにより生成される概念構造に加え、関連関係を抽出する SS-SANS 法により抽出されるデータを組み入れることにより、得られる効果と問題点の考察を試みた。

2. 用語の体系化

本研究に使用する用語の体系化手法を以下に示す。

2.1 C-TRAN 法 (Constrained Transitive Closure)

用語間の同値関係すなわち同義語集合を抽出する具体的実現方法。具体的には同値関係を含む情報資源として日英対訳用語集を使い、記載されている英語に対して日本語、またはその逆の対訳を同値関係として抽出する方法。

2.2 SS-KWEIC 法 (Semantically Structured Key Word Element Index in Terminological Context)

専門用語の構成規則に基づいて、複合用語を基本構成用語に分解し、相互の関係を解することに

2.3 SS-SANS 法 (Semantically Specified Syntactic Analysis of Sentences)

テンプレート構文から出発し、構文解析の結果として得られる階層構造に基づき、同値の新しい構文パターンを探し、その構文により階層構造を拡大する方法。自然科学において重要な論理関係である因果関係を抽出する。

3. 概念構造

前述した用語の体系化手法である、SS-KWEIC 法と C-TRAN 法により生成される概念構造に、SS-SANS 法により抽出されたデータを加え構造化を行った概念構造を以下の図.1 に示す。

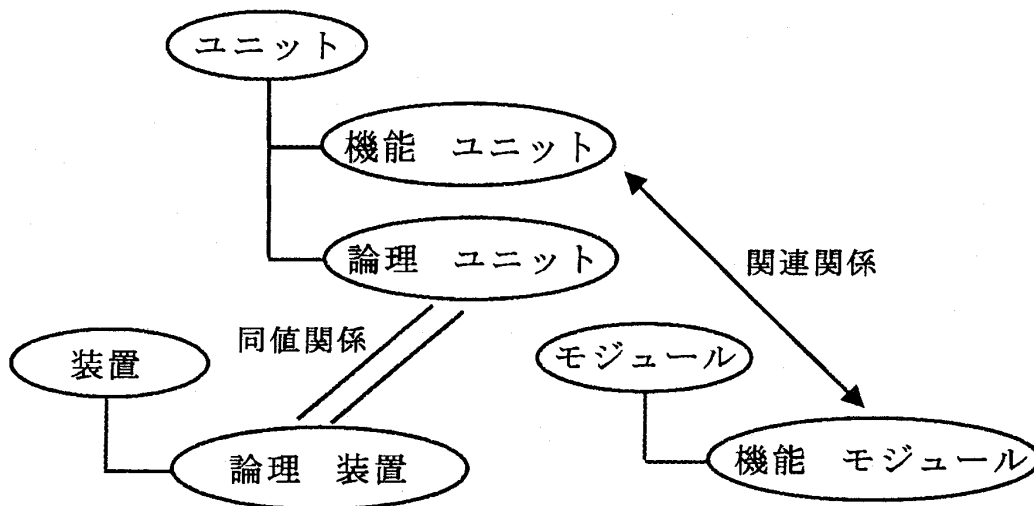


図.1 概念構造

4. 実験環境

4.1 入力データ

入力データ 1…『情報処理用語大事典』(用語数:28130 語)の英和索引を日本語形態素解析システム JUMAN[1]を用いて各用語を複数の語基に分割した用語と、その対訳。

入力データ 2…NACSIS テストコレクション[2](1999 年度版)の NTCIR1 を SS-SANS 法により処理を行い、そこから抽出される関連関係になる対の用語を日本語形態素解析システム JUMAN[1]を用いて各用語を複数の語基に分割を行った対の用語。用語総数は 51302 語。

4.2 実験手順

入力データ 1 を SS-KWEIC 法により階層化を行いつつ同時に入力データ 1 の対訳から C-TRAN 法によって同値関係を抽出し構造化を行う。次に入力データ 2 を SS-KWEIC 法により階層化を行いつつ同時に入力データ 2 における対の用語を関連関係として構造化を行う。

5. 結果

結果として「ユニット」、「装置」、「処理」を語基とした結果を一部抜粋し示す。[]の中の数字は SS-KWEIC 法により階層化された階層の深さ、=は C-TRAN 法により抽出された同値関係を表し、<->は SS-SANS 法により抽出された関連関係を表す。

ユニット [1]

メモリ 管理 ユニット [2]

= memory management unit = MMU

メッセージ 通信 ユニット [2]

<-> メッセージ 処理 [関連関係]

論理 ユニット [2]

= logical unit = 論理 装置

= logical device = ロジカルデバイス

演算 論理 ユニット [3]

= arithmetic and logic unit = 算術 論理 演算 装置

結果.1 用語「ユニット」を語基とする場合の構造

装置 [1]

= device = デバイス

論理 装置 [2]

= logical device = ロジカルデバイス

= logical unit = LU

= logical unit = 論理 ユニット

ネットワーク 型 論理 装置 [3]

<-> 論理 検証 [関連関係]

結果.2 用語「装置」を語基とする場合の構造

処理 [1]

= process = プロセス

= process = 処理 過程

テキスト 処理 [2]

= text processing = 文章 処理

メッセージ 処理 [2]

<-> メッセージ 通信 ユニット [関連関係]

<-> 可変 構造型 並列 計算機 [関連関係]

自動 メッセージ 処理 [3]

結果.3 用語「処理」を語基とする場合の構造

6. 考察

結果.1 に示す通り、「ユニット」を語基としそこから階層関係となる用語「論理ユニット」を通じ同値関係にあたる「論理装置」「ロジカルデバイス」という情報を得ている。また、

この用語は「ユニット」の階層関係にある用語の同値関係ということで、「ユニット」の階層関係にもなり得るといことが推測できる。さらに、同値関係として抽出された「論理装置」について、その語基である「装置」からの構造を示したものが結果.2である。結果.2に示されるように「論理装置」にはその階層関係として「ネットワーク型論理装置」が存在する。よってこの用語も「ユニット」の階層関係と推測することが可能である。

また、「ユニット」の階層関係となる用語「メッセージ通信ユニット」と「メッセージ処理」は関連関係で結ばれている。「メッセージ処理」は「処理」を語基とする結果.3のような構造となる。「メッセージ処理」には階層関係として「自動メッセージ処理」が存在し、このことは「メッセージ通信ユニット」と「自動メッセージ処理」の間にも関連関係が存在する可能性を示している。また、「メッセージ処理」は関連関係として「可変構造型並列計算機」にもリンクされており、それをたどることによりさらに用語間の関係を得ることが可能となる。

このようにSS-KWEIC法により抽出される階層構造を中心に考え、それらにC-TRAN法、SS-SANS法、により得られる同値関係、関連関係を組み合わせることにより、用語対概念、概念対概念間の比較が可能となり、用語や概念に対し、より精密に意味的な処理を可能とする構造が作成できた。

7. 終わりに

SS-KWEIC法、C-TRAN法、SS-SANS法により抽出される意味関係を組み合わせ、概念構造を作成することにより、用語の意味的な位置づけのための判断材料が増え、より詳細な用語の体系化を行うことが可能となった。

今後、それぞれの意味関係抽出手法の精度の向上や、さらに多くの用語の意味関係を構造化することにより、計算機に意味処理に適した知識構造を持たせることが可能となると考える。

参考文献

- [1] 日本語形態素解析システム JUMAN:
<http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html>
- [2] NACSIS テストコレクション:
<http://research.nii.ac.jp/ntcir/index-ja.html>
- [3] 藤原 譲, 情報学基礎論の現状と展望-学習・思考機構と超脳計算機への応用-, 情報知識学会誌, Vol1.9, No.1, pp13-29, 1998.

近藤 雄裕 神奈川大学大学院理学研究科 (平塚市土屋 2946 番地)

藤原 譲 独立行政法人工業所有権総合情報館 (東京都千代田区霞が関 3-4-3)

Takahiro Kondo (s965521@educ.info.kanagawa-u.ac.jp) Graduate School of Science, Kanagawa University

Yuzuru Fujiwara (fujiwara-yuzuru@ncipi.jpo.go.jp) National Center for Industrial Property Information.