3P-02　分野オントロジー構築支援システムの設計

廣田 啓一　佐々木 裕　加藤 恒昭

NTT コミュニケーション科学基礎研究所

e-mail: {hirota,sasaki,kato}@cslab.kecl.ntt.co.jp

1　はじめに

我々はこれまで、分野オントロジーを用いて分野に適応した情報抽出を行なう、オントロジー主導型情報抽出ＯＤＩＥ（Ontology-Driven Information Extraction）の提案を行ない、システムとして実装、評価を行ってきた [2][3]。提案手法は従来の情報抽出手法における分野依存性の問題の解決を目的とし、分野依存の知識を分野オントロジーに記述する事により、情報抽出システムに多様な分野への適応性を持たせる事を可能とする。

情報抽出の対象となる分野テキストには特定分野にのみ頻出する専門的な用語や定型表現が数多く存在し、また共通的な用語でも分野によって表現が異なる場合がある。したがって、WordNet など既存の言語オントロジーの利用は適切ではなく、分野毎に必要なオントロジーを構築する必要がある。しかしながら、分野オントロジーの構築の際に必要な用語および関係を網羅的に記述する事は、単純な人手作業では困難である。

そこで、分野オントロジーの構築労力の軽減、および自動構築に必要な技術の検討を目的として、分野オントロジー構築支援システムの設計を行なった。

2　構築支援システムの設計

本構築支援システムは、図 1に示すように、実際の分野テキスト集合から (1) 分野知識の記述に必要な用語の収集および (2) 各用語間の関係の収集を行ない、(3) 視覚的な編集機能をユーザに提供する事で、クラスタリングや用語・関係の追加といった、分野用語の体系化による分野オントロジーの構築を支援する。

以下、本システムが提供する機能の概要を述べる。

2.1　分野用語の収集機能

分野オントロジーの構築に必要な分野用語を対象分野の実テキストから収集する。用語収集の手段は様々な手法があるため、幾つかの手法をモジュール化し、ユーザが選択して呼び出す事により各手法に基づく用語が収集できるようにした。現時点でモジュール化した手法は、複合語抽出と固有表現抽出の二つである。

The Design of Domain Ontology Construction Support System

Keiichi HIROTA, Yutaka SASAKI, Tsuneaki KATO

NTT Communication Science Laboratories

2-4 Hikaridai Seika-cho Soraku-gun KYOTO 619-0237 Japan

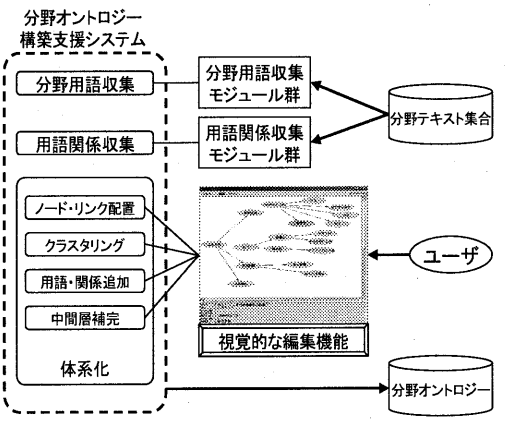

図 1　構築支援システムの概要

(1) 複合語抽出 複合語とは複数の語が組になって新たな意味を持つ単語であり、元の語と上位下位関係を持つ事が多く、分野オントロジーの構築に重要な役割を果たす。包含する語の出現頻度に基づく尺度を計算する、C-value による複合語抽出 [1] を用いた。本モジュールは頻出語の抽出にも用いる事ができる。

(2) 固有表現抽出 人名や地名、組織名、人工物名といった固有の名詞、および金額表現や日付表現を、総称して固有表現という。これら固有表現は、個々の記事において主要な情報を表す要素である事が多く、記事中での個々の表現の記述から意味的な階層の獲得や主要情報の定義などが可能になる。本システムでは、単純置換に基づく固有表現抽出 [5] をモジュール化した。

2.2　用語関係の収集機能

収集した分野用語について、テキスト集合から用語の関係を獲得する。本システムでは、共起の度合いを表す共起強度に基づく獲得と、固有表現の種類に基づく獲得をモジュール化した。

(1) 共起強度に基づく獲得 複合語抽出によって収集した分野用語について、共起頻度と語の出現順に基づいて共起の度合いを表す共起強度と共起順序を表す方向性を計算し、閾値以上の共起強度を持つ関係を獲得する。

(2) 固有表現種類に基づく獲得 同じ固有表現の種類に抽出された単語は同一のカテゴリに入るものとし、固有表現種類を表すタグを上位の語とするインスタンス関係を獲得する。

2.3 視覚的な編集機能

ユーザインターフェースでは、収集した分野用語および用語関係をノードおよびリンクとしてユーザに提示し、視覚的な編集機能を提供する。ユーザは、提示されたノードおよびリンクの取捨選択を行ない、さらに配置位置を決定して、分野オントロジーを構築していく。

表示画面上のノードは収集した分野用語を表し、ユーザはノードのコピーや削除などの他、ノード名称や同義語情報の編集を行なう事ができる。リンクは分野用語間の関係を表し、そのリンクの向きと関係子の種類によって階層関係などを定義する。リンクに対しては、削除の他に上位／下位ノードの変更やリンク向・関係子変更などを行なって、体系を詳細化する事ができる。また、収集した分野用語や用語関係が適切でない場合はこれを削除し、逆に必要に応じてユーザが新たなノードやリンクを作成する事もできる。

ユーザは、これらの編集機能を通して、適切な分野用語や用語関係を追加し、同一カテゴリに属する分野用語をクラスタリングして、分野オントロジーの体系を詳細化する。本システムでは、必要な処理機能をノードとリンクに対する編集操作として提供する事で、視覚的かつ直観的な構築を実現する。

2.4 その他の機能

2.4.1 関連テキストの表示

オントロジー中のノードやリンクに関連するテキストの表示を行なう。視覚的な編集機能は分野オントロジーの構築を容易にするが、個々のノードやリンクの情報だけでは適切な編集が難しい場合がある。関連テキストを検索、表示する機能を設け、用語の意味や用法、他の用語との関係を随時確認できるようにした。

2.4.2 オントロジーの評価

構築中のオントロジーについて評価を行なう。他ノードとの関係によりノード種類を頂点、中間、末端、孤立に分類してノード数を示し、全ノード数を提示する。また、リンクについても関係子毎にリンク数を示し、全リンク数を提示する。これらの情報は構築中のオントロジーの規模と構成に関する指標となる。

2.4.3 オントロジーの検査

構築したオントロジーの整合性に関する検査を行なう。具体的には、同一名称を持つ重複ノードや他ノードとのリンクを持たない孤立ノード、および上位／下位ノードを持たない途切れたリンクや複数ノード間で上位／下位関係が一巡する循環リンクの検出を行なう。

3 システムの実装と評価

構築支援システムの実装を行ない、提携記事に関する分野オントロジーの構築を行なった [4]。複数のユーザによる構築では、クラスタリングや体系化において構築手順の違いが見られたが、構築した分野オントロジーは階層関係の細部を除けばそのカテゴリ構成においてほぼ同じであった。

構築したオントロジーがその分野の知識をどれだけ適切に表現しているかを明確に評価する事は難しいが、例えば一つの方法として、分野オントロジーに記述した用語や関係が実際のテキスト上に占める割合を算出する事などを考えている。また、構築した分野オントロジーを用いた情報抽出の精度は、オントロジーの運用性を示す指標の一つと考える事ができる。

4 まとめ

分野オントロジーの構築方法について検討し、構築支援システムを設計した。本システムは構築に必要な分野用語や用語関係を自動的に収集し、視覚的な編集機能を提供する事で、ユーザによる構築を支援する。

システムはオントロジーの自動構築に関わる各種技術を実装評価するための実験システムとしての色合いが強い。例えば、重要語抽出や n-gram 抽出、複合語抽出などの手法により収集した分野用語に対し、実際にユーザがどの語を配置したかという情報は、ランキングの評価は別として、用語収集手段の有効な指標となる。より精度の良い手法をモジュール化し、複数組み合わせる事で、より効率的なオントロジー構築が可能になる。

また、収集した用語のクラスタリングやテキストに出現しない用語と関係の補完といった技術の導入は、オントロジー構築の自動化に向けた重要な今後の課題の一つである。複数ユーザによる提携記事からの構築はカテゴリ構成がほぼ同じとなった事から、収集する分野用語や関係によってクラスタリング結果もほぼ一意に定まると考えられる。今後、実用的な分野オントロジーの構築を基に、これら技術の検討を深めたい。

参考文献

[1] Katerina T. Frantzi, Sophia Ananidou, 辻井潤一, 専門用語の自動抽出, 情報処理学会研究報告,NL-112-12,1996

[2] 廣田啓一, 佐々木裕, 加藤恒昭, オントロジ主導による情報抽出の検討, 情報処理学会研究報告,NL-133-12,1999.

[3] 廣田啓一, 佐々木裕, 加藤恒昭, オントロジー主導による情報抽出, 人工知能学会誌,Vol.14,No.6,1999.

[4] 廣田啓一, 佐々木裕, 加藤恒昭, 分野オントロジー構築支援システムの実装, 情報処理学会第 61 回全国大会講演論文集, デ-13,2000.

[5] 佐々木裕, 廣田啓一, 加藤恒昭, ProCreator: 単純置換に基づく日本語固有表現抽出ツール, IREX ワークショップ予稿集,1999.