教材資料を対象とした用語収集とその分析手法

北村　怜子†　　塚本　享治†
東京工科大学大学院　バイオ・情報メディア研究科†

1. はじめに

世の中は速く変化し、学問の世界で新しい学問分野が次々に作られている。しかしそれらは体系化されておらず、その学問分野が扱う領域が明確でない。そのため関係者が持つイメージの食い違いや、内容が経年変化していく可能性がある。

そこで新しい学問分野を対象にオントロジを構築し、分析する研究を行っている。本稿ではメディア学という分野を対象にオントロジ構築のための分析、設計方法について述べる。

2. 研究の目的とアプローチ

筆者らのメディア学部は、表現系、社会系、技術系の3分野に分けられる。学生は希望科目や進路を考えて科目の選択をする際、科目間の類似度やメディア学全体を知りたくなる。

研究を進めるにあたり、教材資料から用語を手作業で抽出する方法と自動的に抽出する方法を用いた。方法の違いでどの程度抽出した用語に違いがでるか分析した。また、抽出用語をどのように整理すればコンピュータで扱えるのかについて検討した。

3. 教材資料の分析と用語の抽出方法

3.1 科目間の階層構造について

図1のように全科目を3階層に分けて、用語の抽出を行った。また階層別に手動抽出と自動抽出に分けた。

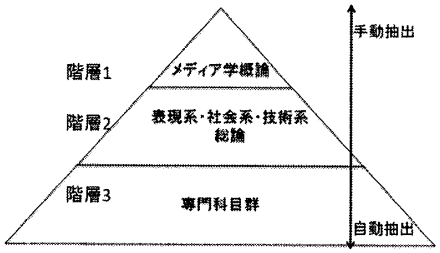

図1　科目間の階層構造

3.2 科目ごとの分析と用語の抽出方法

メディア学概論は学部の最上位科目なので精密な分析が必要と考え手動抽出を行った。また表現系、社会系、技術系の各総論科目もその科目から専門科目が派生するため手動抽出を行った。専門科目は用語数が非常に多いため自動抽出を進めている。

The collection and analysis of vocabularies from teaching materials †Reiko Kitamura, †Michiharu Tsukamoto
†Tokyo University of Technology, a media science major

3.3 手動での用語と用語間の意味関係の抽出

教材資料から用語（キーワード）とその用語間の意味関係（図2）を手作業で抽出しトリプルを作成した。

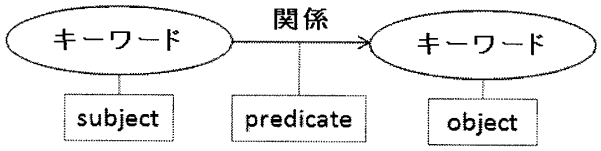

図2　トリプルによるデータ表現

4. 手動抽出した用語について

まず、2科目19ファイルから手作業でトリプルを作成したところ、キーワード（subjectとobject）が2125種類、predicateが463種類あった。predicateは語彙違い、れる/られる（受動関係）、ではない（否定）などにより数が増えていた。predicateを整理、分類し少ない用語数でRDFを記述する必要がある。

4.1 predicateの整理・分類

語彙違いは同じ意味のpredicateの語彙に統一した。またサ変動詞の用語は「する」を除いて動詞的用法の名詞に統一した。

ex: 増大する、倍増する、増加現象→増加

受動関係は、機械的にsubjectとobjectを入れ替えて受動表現をなくした。

ex: 利用される、要求される、減少させる

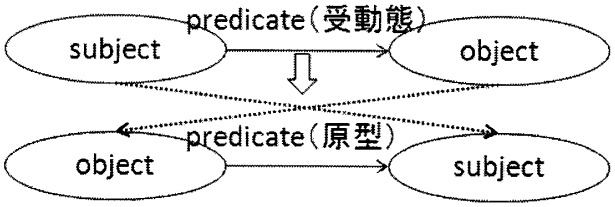

図3　受動関係

ではない（否定）は接頭辞としてnotを付けて否定を表した。また結論として「xxxである」と言えないため対象外とする。

ex: 変わらない→not変わる

4.2 predicateの階層化

predicateは分野が異なれば様々な用語が使われ種類が増えてしまう。predicateを体系化するために上位概念、下位概念によって分類した。

分類できなかったpredicateもあり、補完財、産業化パラダイム、発行部数など科目特有の用語であった。

表1 上位概念と下位概念による分類例

上位	下位	上位	下位
定義、説明	定義	種類	種類
	説明		分野
	意味		技術分野
	である		分類
	つまり		業種
			方向

5. 自動抽出した用語について

用語と用語間の意味関係を全て手作業で抽出するには時間がかかるため、自動抽出を試みた。

5.1 自動抽出の手順

PPTファイルをXML変換し(図4、1、2)用意した辞書にある用語とスライド中の用語で一致しているものを調べた(図4、3)。またPPTファイルとそれに対応するシラバスを結合した(図4、4)。これらの処理はjava、xsltとANTで行った。

5.2 自動抽出の問題点

様々な分野の用語に対応し、高速にインターネットで利用できるという理由で「はてな検索」を使用した。しかし、はてな検索は形態素解析しか行っていないため複合語の抽出ができなかった。また研究室で集めた技術用語に特化した用語集も使用したが、現在のところ単語単位の抽出しかできていない。

5.3 複合語辞書の用意

単語単位では意味の扱いに適さないため、構文を利用した用語の抽出が必要であることが分かった。複合語辞書として手動抽出した用語集を用いた。

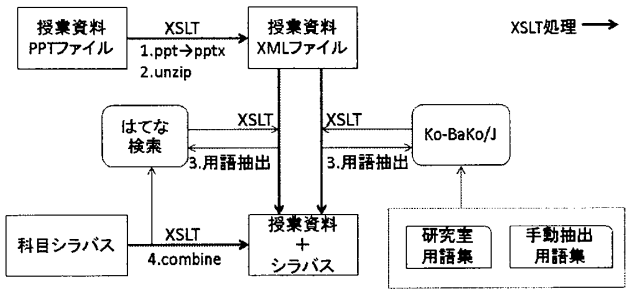

図4 自動抽出の構成

6. 抽出方法による用語の比較と考察

6.1 手動抽出と自動抽出の比較

手動抽出では意味を考えながら用語と用語間の意味関係を抽出したが、自動抽出では単語単位のため抽出結果にギャップがあった。

6.2 構文解析を利用した複合語の抽出

形態素解析では用語がPPTのどの位置に出るか程度のことしかわからない。科目中の用語や科目間の関係は一般に複合語となっており(図5)、構文解析が必要になる。

当初、教材からの用語抽出は全ての階層において単語単位を考えていたが、階層1、階層2を手動抽出した結果、複合語が非常に多く存在していた。そのため階層3でも複合語を扱うことが必要となった。

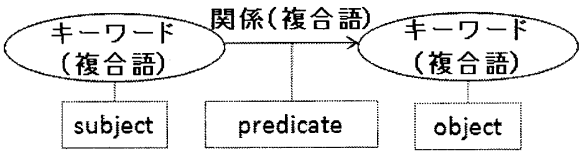

図5 複合語単位のトリプルの表現

試しに多機能日本語処理ライブラリKo-Bako/Jで構文解析を行った。形態素解析と構文解析の違いを表2に示す。

表2 形態素解析と構文解析の例

	形態素解析	構文解析
期待される機能	期待 / さ / れる / 機能	期待される / 機能
メディア技術	メディア / 技術	メディア / 技術
アプリケーション用ソフトウェアの開発	アプリケーション / 用 / ソフトウェア / の / 開発	アプリケーション / 用 / ソフトウェアの / 開発

これからわかるように形態素解析で単語に分け構文解析で係り受けを分析し、句の関係から複合語を作成できると考えられる。

7. おわりに

教材資料から用語と用語間の意味関係の抽出を行ったが、predicateの種類が非常に多いことと複合語抽出が問題であった。predicateをさらに整理、分類し構造化することが必要である。それにはRDFSやOWLを用いて集約することが考えられる。

複合語については、構文解析を利用した複合語抽出ツールの開発に着手した。また複合語と複合語の関係は構文解析を用いても自動抽出が難しいので、会話的に選択することを考えている。

参考文献

[1] 北川 修, 教材資料のRDF化と検索手法に関する研究, 東京工科大学卒業論文, 2007
[2] はてな検索, http://search.hatena.ne.jp/
[3] Ko-BaKo/J, 日本システムアプリケーション, http://www.jsa.co.jp/LANG/ko-bako/index_frame.htm
[4] Dean Allemang, James Hendler, Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL, Morgan Kaufmann Pub, 2008