

階層的語彙体系を利用した産業遺産の記述とその分類への適用 A Description of Industrial Heritage using Hierarchical Japanese Vocabulary System and its Application to Document Classification

澤田 貴行[†] 青野 雅樹[‡]
Atsuyuki Sawada Masaki Aono

1. はじめに

人文社会分野では、研究者は対象事物の状況分析を文章で記述し、他者はその記述を自身の考えと関連付けて新たな知識とする。しかし、記述の理解には、自然言語の多様性等から知見や職人芸的な分析を経なければならず、同じ記述でも人によって理解具合が相違し、共通理解に困難をきたす。実際、産業考古学会関連の中部産業遺産研究会では、研究者相互の共通理解の推進と広く成果を公表するために、書籍等で公表していた遺産記述をデータベース化し、ウェブ公開を試みた。しかし、その構築には、自然言語の多様性から事物の本質的な記述のための枠組みの特定が不十分で、名称や完成年度程度の限定的な内容しか蓄積できず、現在はその利用は滞ってしまっている。

そこで本研究では、産業遺産記述を題材に、ウェブソース記述の枠組み RDF(Resource Description Framework)を利用して多様な記述をすることを提案する。RDF は、記述は主語、述語、目的語の三つ組として、主語と目的語の関係として表現され、W3C により規格化されている。主語と目的語はノード、述語をアークとすれば、有向グラフとして可視化でき利用者の理解向上にも寄与する[1]。しかし、記述を RDF・RDF スキーマで行うには、枠組みではなく意味表現として自然言語との差も懸念されるところでノードへの意味付けによる理解向上を試みた。また、その利用可能性として文書分類を行ったことも報告する。

2. 提案手法

本提案をデータ構築とその利用(分類)という観点で説明する。データ構築では、第一に書籍から記載をスキャンし、テキストを作成する。対象とした書籍は、愛知県の近代化遺産[2]である。これには産業遺産が、建築と土木、農業・工業・金融等の産業業種毎に分類され、業種によっては、さらに細かな業種分類も付与される。所在地等の記載もあり、詳細は自然言語で、図面や写真とともに記述される。次に生成テキストの品詞解析を行い、名詞には意味付けと相互の繋がりを付与する。分類では、構築データの利

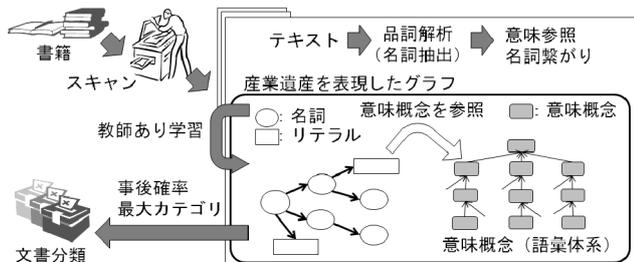


図1 本提案の処理工程

[†]豊橋技術科学大学 大学院 情報・知能工学専攻
[‡]豊橋技術科学大学 情報・知能工学系
Toyohashi University of Technology

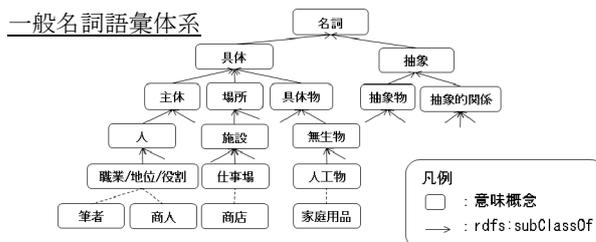
用のひとつとして、産業業種を推定する文章分類を行う。

2.1 データ構築

データ構築では、テキストファイルの本文部分に対し、文毎に分離したうえで、品詞解析を行って名詞を抽出する。名詞はノードになるものであり、解析時には一般名詞・専門名詞(記述の際に利用される建築用語や施設機能等)や地域名等の固有名詞を識別する。そして、識別結果に応じて、それぞれの語彙体系を rdfs:type により参照して意味付けを行う。最後に抽出された名詞群の関係を述語語彙体系により設定する。ここで語彙体系は、意味概念の体系であり、主語と目的語の意味概念を定義した名詞語彙体系と名詞相互の関係を定義した述語語彙体系を構築した。

2.1.1 名詞語彙体系

名詞語彙体系は、一般名詞と専門・固有名詞の2つを構築した。一般名詞は日本語語彙体系[3]を参考にし、専門・固有名詞には、産業遺産有識者からの聞き取り等から独自作成した語彙体系である。語彙体系は、名詞の具体的・抽象的な意味概念の関係を表すもので、意味概念の汎化・特化の関係を階層構造として表現した。なお、一般名詞語彙体系は、名詞への知識や文章における役割等を表現するのではなく、単純な意味概念の体系である。それに対し、固有・専門名詞語彙体系は、特定領域における有識者による知識を概念化したものである。



固有・専門名詞語彙体系

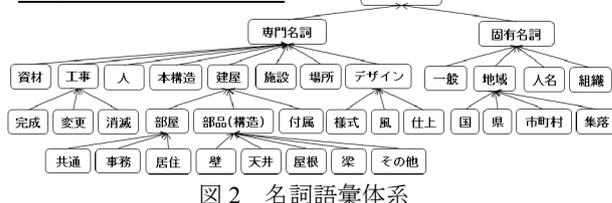


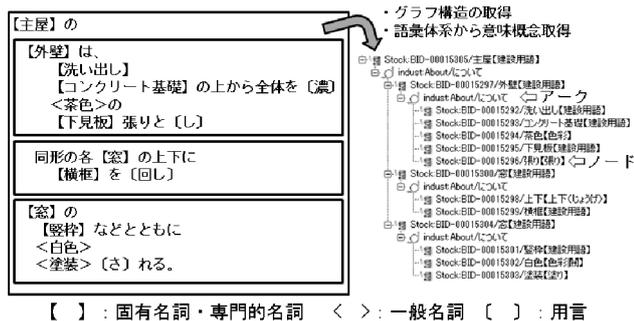
図2 名詞語彙体系

2.1.2 述語語彙体系

述語は、記述された自然言語と比較して、抽出された名詞群を繋げる役割を持つ。ここで述語意味を述語語彙体系への参照で表現することから、名詞体系と同様に形容詞や動詞等の語彙体系を構築しなければならない。しかし、本稿では、名詞語彙体系を重視し、述語は単純な連結を示すもののみとし、意味概念を考慮しない。しかし、その述語意味は文中の言葉を利用してリテラルとして蓄積する。

2.1.3 データ構築事例

自然言語記述から、抽出された名詞群に対し、主語と目的語を選定するには、人による解釈を経て実現した。基本的には、主語となるノードを選定し、その述語は目的語を包含する関係として捉え、主語と述語に対する目的語群という関係を設定することで、同一主語と述語に対する目的語の異なる三つ組群を構築する。あわせて名詞や動詞等は文中の言葉を利用したリテラルとして rdfs:label によりノードの目的語として蓄積する。単純な構造だが、一文に三つ組関係が複数存在する場合も入れ子構造により RDF として変換することができる (図 3 参照)。これらを経ることで、文中の名詞、動詞等の自然言語情報を蓄積する。



【 】: 固有名詞・専門的名詞 < >: 一般名詞 []: 用言

図 3 データ構築の様子

2.2 分類

構築データの利用方法として、RDF による記述が、どの産業業種に属するかを推定する分類を行った。分類には教師あり学習として、ナイーブベイズ分類器[4]を用いる。分類器は、ベイズ定理に基づき文章 (doc) を名詞 (w1..wk) の集合として、名詞の独立性を仮定したうえで、産業業種 (cat) 毎の事後確立を最大にする産業業種を選択する。

$$P(cat | doc) = \frac{P(cat)P(doc | cat)}{P(doc)} \propto P(cat)P(doc | cat)$$

$$= P(cat) \prod_{i=1}^k P(w_i | cat)$$

$$cat_{max} = \arg \max_{cat} P(cat) \prod_{i=1}^k P(w_i | cat)$$

3. データ構築結果

対象は、愛知県の近代化遺産に記載ある建築物とし、産業業種は、農業、工業・金融・商業、インフラの 5 分野に限定してデータ構築した。データ概要を表 1 に示す。

表 1 構築データの概要

分野	記述数	名詞 総計	一般 名詞	専門 名詞	固有 名詞
農業	8	1,239	581	489	169
工業	12	2,382	1,024	965	393
金融	12	2,186	824	974	388
商業	9	1,757	709	777	271
インフラ	10	1,359	622	448	289
計	51	8,923	3,760	3,653	1,510

4. 分類実験と結果

構築データは、自然言語を対象とすることから、切妻造り・切り妻造り・切妻造のような使用字句の表記違いや教室や講義室のような同一意味概念の表記違いを考慮すれば、

汎化によりその違いを吸収できると考えられる。このため、名詞をリテラルと意味概念への ID と見立てた場合、かつ名詞、専門的名詞、固有名詞の分類精度へ寄与を考察するため、それぞれの利用有無の場合として、計 14 パターンで実験を行った。評価は、52-分割交差検定で行い、すなわち自身以外のすべてをトレーニングデータとし、自身を正しく分類できたかで評価する。表 2 に正答率を示す。

表 2 リテラルと意味概念 ID による分類結果

条件	名詞総数	リテラル	意味概念 ID
一般名詞のみ	3,760	66.67%	41.18%
専門名詞のみ	3,653	60.78%	17.65%
固有名詞のみ	1,510	45.10%	23.53%
一般+専門	7,413	66.67%	33.33%
専門+固有	5,163	64.71%	23.53%
固有+一般	5,270	64.71%	43.14%
すべて	8,923	68.63%	31.37%

実験結果は、より多くの名詞をリテラルとして、すなわち繊細に利用した場合が 68.63% で最良であり、リテラルでは多くの名詞を利用したほうが良くなる傾向がみられた。また、名詞組み合わせ条件のすべてにおいて意味概念 ID よりリテラルを利用したほうが良い。

「一般名詞のみ」を考えると、一般名詞語彙体系では名詞によっては一般性から多義に定義されることもあり、産業遺産における本来意味以外の意味概念も付与され、それがノイズとなり正答率が下がったと考えられる。また、網掛部分の正答率が低いことは、現状は専門・固有名詞語彙体系での不適切な意味概念 ID 付与したことが原因である。

5. まとめと今後の課題

本稿では、産業遺産を記述した文章から半自動的に RDF による記述へ変換する手法を提案した。加えて、理解向上のため、ノードに意味付けをし、そのための意味概念を階層構造とした語彙体系を構築した。意味概念を階層化することは、利用者が汎化によって抽象的理解を可能とするからデータの理解向上に役立つであろう。今後の課題として、述語語彙体系の構築が挙げられる。詳細な述語語彙体系はもちろん、ノード間の役割概念に関する述語を設けること、他データとの関係付ける述語を設けることで、新たな知識発見や他分野も含めた連携にも期待できると考えている。

構築データの利用として、文書分類実験を行ったが、現状では期待した意味概念への汎化の効果を確認できなかった。自然言語には表記違いがあるのは事実であるが、詳細な語彙体系でないと意味概念が曖昧になってしまう。今後の課題として、リテラルと意味概念の使い分け、及び意味概念の詳細定義と繊細な参照をするように改善する必要性が示唆された。

参考文献

[1] 神崎正英, “セマンティック・ウェブのための RDF/OWL 入門”, 森北出版(2005).
 [2] 愛知県教育委員会, “愛知県の近代化遺産・愛知県近代化遺産(建造物等) 総合調査報告書”, (2005).
 [3] 白井諭, 大山芳史, 池原悟, 宮崎正弘, 横尾昭男, “日本語語彙体系について”, 情報処理学会研究報告 IM, Vol.98, No.106 (1998).
 [4] 阿部倫子, 田中久美子, 中川裕志, “コメントを用いた映画の分類”, 情報処理学会研究報告. 自然言語処理研究会報告, NL-150-16(2002).