社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－FI－73 (4)
2003／11／13

Webディレクトリを用いた2言語オントロジーの構築

木村 文則
奈良先端科学技術大学院大学 情報科学研究科
〒630-0192 奈良県生駒市高山町 8916-5
Tel: 0743-72-5336, Fax: 0743-72-5339, E-Mail: fumino-k@is.aist-nara.ac.jp
前田 亮
立命館大学 理工学部 情報学科
〒525-8577 滋賀県草津市野路東 1 丁目 1-1
Tel: 077-561-5049, Fax: 077-561-2669, E-Mail: amaeda@cs.ritsumei.ac.jp
越田 高志
奈良先端科学技術大学院大学 情報科学研究科
〒630-0192 奈良県生駒市高山町 8916-5
Tel: 0743-72-5336, Fax: 0743-72-5339, E-Mail: takas-ko@is.aist-nara.ac.jp
松江工業高等専門学校 情報工学科
〒690-8518 島根県松江市西生馬町 14-4
Tel: +81-852-36-5248, Fax: +81-852-36-5248, E-Mail: takashi@matsue-ct.ac.jp
宮崎 純, 植村 俊亮
奈良先端科学技術大学院大学 情報科学研究科
〒630-0192 奈良県生駒市高山町 8916-5
Tel: 0743-72-5336, Fax: 0743-72-5339, E-Mail: {miyazaki, uemura}@is.aist-nara.ac.jp

概要

機械に単語や文章の表層的な理解だけでなく、意味的な理解も実現しようという取り組みが盛んになっている。その実現のために重要な役割を担うのがオントロジーである。また、同様の試みをWebにも適用しようという試みも行われている。そのためには、オントロジーが一つの言語だけでなく、複数の言語を対象とすることができなければならない。そこで本研究では、オントロジーを翻訳することにより、2言語を対象としたオントロジーを構築する手法を提案する。

キーワード

オントロジー, 曖昧性解消, WordNet, OWL

※本論文は第24回ディジタル図書館ワークショップの論文です.

Construction of Bilingual Ontology using Web Directory

Fuminori Kimura
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
Phone: +81-743-72-5336, Fax: +81-743-72-5339, E-Mail: fumino-k@is.aist-nara.ac.jp
Akira Maeda
Department of Computer Science, College of Science and Engineering, Ritsumeikan University
1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan
Phone: +81-77-561-5049, Fax: +81-77-561-2669, E-Mail: amaeda@cs.ritsumei.ac.jp
Takashi Koshida
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
Phone: +81-743-72-5336, Fax: +81-743-72-5339, E-Mail: takas-ko@is.aist-nara.ac.jp
Information Technology, Matsue National Colledge of Technology
14-4 Nishiikushima, Matsue, Shimane 690-8518, Japan
Phone: +81-852-36-5248, Fax: +81-852-36-5248, E-mail: takashi@matsue-ct.ac.jp
Jun Miyazaki, Shunsuke Uemura
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
Phone: +81-0743-72-5336, Fax: +81-0743-72-5339, E-Mail: {miyazaki, uemura}@is.aist-nara.ac.jp

Abstract

Researches that not only surface understanding of words or texts but semantic understanding to computer will be promoted. Ontology has an important role for the realization of these trials. Moreover, the trial in which the same trial will be applied to Web is also performed. For this purpose, ontology must be able to treat with not only one language but two or more languages. Then, in this research, we propose the method of constructing the bilingual ontology by translating monolingual ontology.

Keywords

ontology, disambiguation, WordNet, OWL

1. はじめに

Webの急速な発展により、膨大な数のWeb文書が閲覧できるようになった。Webを有効に活用するための研究が行われ、Web文書の検索やWeb文書上に記述された情報の活用ができるようになった。

そして最近では、Web文書の内容を機械が理解できるようにし、さらに高度な処理が可能となることを目指した研究が行われ始めている。セマンティックWebはその代表例である。この目的を達成するのに重要な役割を果たすと考えられるのがオントロジーである。オントロジーとは、対象とする世界に存在するものごとを体系的に分類し、その関係を記述したものである。オントロジーに記述されたこれらの情報を利用することにより、機械にある単語の意味を理解させることが可能となるため、オントロジーの応用例の一つとしてこのような研究が行われている。

また、Webは世界中の多くの地域で利用されるようになったことにより、Web文書は英語だけでなくそれ以外の様々な言語を用いて記述されるようになった。この傾向は今後さらに大きくなることが予想されることを考慮すると、Webを十分に活用するには単一の言語のみが対象では十分とは言えない。そこで我々は、2言語を対象としたオントロジーの構築を行う。2言語オントロジーを利用することにより、言語の制約から解放されたセマンティックWebの実現、異なった言語間にわたるデータの統合などが可能となることが期待される。セマンティックWebを例に挙げると、Web文書の検索などにおいて推論を行うことにより、現在の検索エンジンではできない高度な検索が可能となる。例えば、「世界で液晶ディスプレイを一番安く販売している店」を探すといったようなことも可能となる。

図書館には日本の資料だけではなく、世界各国の資料が蔵書として保管されている。同様に、ディジタル図書館においても、様々な言語で書かれた資料や文書が存在する。このような状況において蔵書を効率的に管理するためには、多言語を扱える必要がある。また、ディジタル図書館が世界的な情報の共有を行うことは、重要な目的の一つである。そのためには言語の壁を乗り越えるための手段が必要となる。これらの目的を実現するために、2言語オントロジーは重要な役割を果たすことができると考えられる。

2. 関連研究

オントロジーはもともと「存在論」を意味する哲学用語であったが、人工知能などの研究においては、「対象とする世界に存在するものごとを体系的に分類し、その関係を記述するもの」として取り込まれた[1]。オントロジーを利用することにより，機械が言葉の意味を理解することが可能となり、文章の表層的な解析だけではできない高度な処理を行うことも可能となる。実際に、ある特定の分野に対するオントロジーを構築し、設計支援に利用することなども行われている[2]。

近年、Web文書の内容を機械が理解できるようにする試みがなされている。その試みとは、Tim Berners-Leeが提唱したセマンティックWeb[3]である。セマンティックWebにおいてもオントロジーは、機械が言葉の意味を理解するための知識を与える役割を期待されている。セマンティックWebにおけるオントロジーの記述言語として、W3C(World Wide Web Consortium)により、OWL(Web Ontology Language)が開発されている[4]。OWLではある一つの概念はクラスにより記述されるが、他のクラスとの関係が記述できるように設計されており、これを推論ルールとして利用することが可能である。

OWLにより記述されたオントロジーのうち公開されているものに、"WordNet.OWL"[5]がある。WordNet.OWLは、英語のシソーラスであるWordNet1.7.1[6]をもとにOWLにより記述することで作られた英語のオントロジーである。現在のところ、公開されている日本語のオントロジーはないが、神崎は日本語の単語を対訳辞書を用いて翻訳しWordNetに対応付けることにより、利用可能な日本語のオントロジーを構築する試みを行っている[7]。しかし、この手法では対訳辞書により得られる訳語を全て対応付けているため、概念の異なる単語が同じ概念として扱われてしまう可能性がある。そこで本研究では、この問題を解決したオントロジーの翻訳手法を提案し、2言語オントロジーの構築を目指す。

3. 曖昧性解消の必要性

神崎の提案した手法では、単語を対訳辞書により翻訳しその訳語を元の単語と対応付けることにより、オントロジーを構築している。対訳辞書を用いる場合、一般に訳語候補は複数存在する。すなわち、"訳語の曖昧性"のことである。これらの表す概念が全て類似している場合は良いが、多義語などのようにそれぞれの訳語候補の表す概念が類似しないこともある。例えば英語の「bank」という単語の訳語候補として「銀行」、「貯蔵所」、「岸」、「堆積」、「土手」、「堤」、「漕ぎ手」など、類似しているとはいえない単語が対訳辞書には登録されている。神崎の手法では、訳語候補を全て元の単語に対応付けている。この場合、表現する概念が異なっているにもかかわらず、それぞれが類似しているとみなしてしまう可能性がある。

この問題を解決する方法として、言語横断情報検索の分野で研究されている訳語の曖昧性解消の手法 [8] を用いることが考えられる。これは、問合せに用いられた単語の組合せから、どの訳語候補が最も適切であるかを推定する手法である。ただし、これらの曖昧性解消の手法は、問合せに用いられた単語の訳語として適切である訳語を一つ、場合によっては数語を選択するため、類似する概念を表した訳語候補を全て抽出するという本研究の目的には、そのまま適用することはできない。そこで本研究では、Web ディレクトリを利用した曖昧性解消の手法 [9] を二言語オントロジーの構築のために適用することを提案する。

4. 提案手法

図 1 は、本手法の概要を示している。本手法は、英語のオントロジーに登録されている単語を対訳辞書を用いて日本語に翻訳することにより、英語版と同じ構造の日本語のオントロジーを作成する。ただし単語の翻訳は、後述する Web ディレクトリを利用した曖昧性解消の手法を用いて行う。また、そのときに翻訳元となった概念の番号を対応付けておくことで、英語と日本語を対象とした 2 言語オントロジーを構築する。

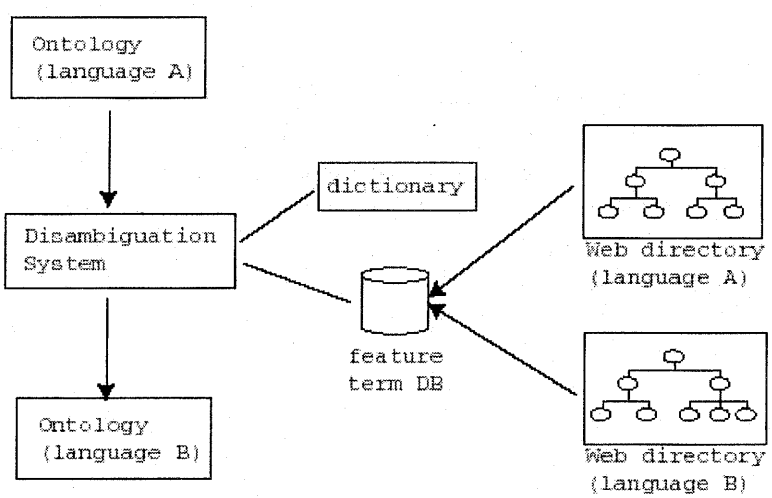

図 1.

Web ディレクトリを利用した曖昧性解消を行うには，例えば Yahoo のように複数の言語版がある Web ディレクトリを利用する．本研究では英語から日本語への翻訳なので、Web ディレクトリも英語版と日本語版を用いる。まず、前処理として次の 2 つの処理をおこなう。

- カテゴリごとに属する Web 文書から特徴語を抽出する。
- それぞれのカテゴリにおいて，言語間でのカテゴリの対応付けを行う。

これらの前処理が完了すると、訳語の曖昧性解消を行うことができる。

4.1. 前処理

図 2 は、前処理の流れを示している。

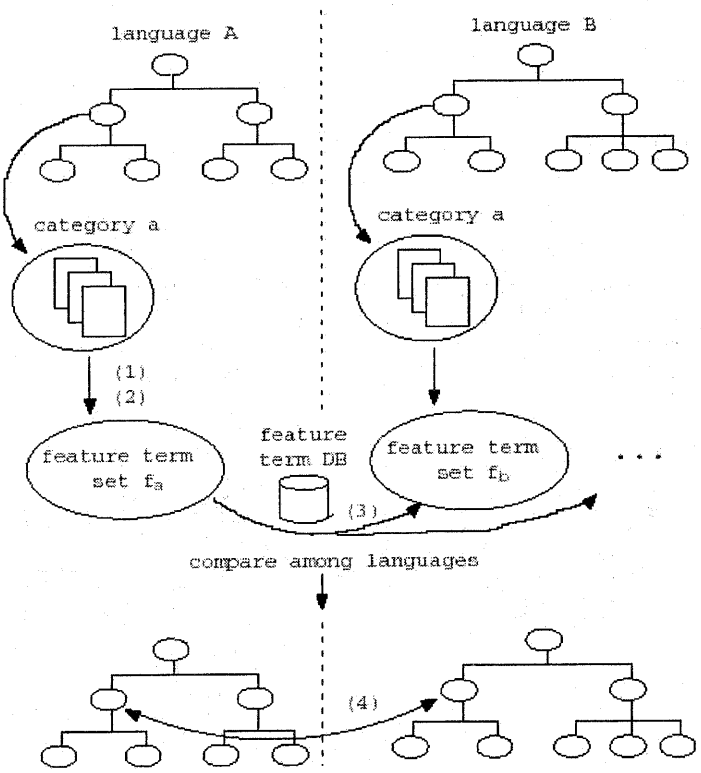

図 2.

前処理では最初に行うのは、特徴語の抽出である。各カテゴリは特徴語集合によりその特徴を表現される。特徴語集合は、そのカテゴリの特徴を表現していると思われる単語の集合である。特徴語を抽出するために、まず各カテゴリに属する Web 文書から単語を抽出する。次に、抽出された単語をカテゴリごとに集計し、その単語がカテゴリの内容を表現する程度を表す重みを計算する。抽出された単語のうち、重みが大きいものから上位 n 語をそのカテゴリの特徴語として抽出する。

Web 文書から抽出された単語の重みは、TF・ICF （term frequency ・ inverse category frequency）により計算する。これは、一般に良く知られた単語の重み付けの手法の 1 つである TF・IDF （term frequency ・ inverse document frequency）を発展させたものである。TF・IDF は単語の出現頻度（TF）と文書頻度の逆数との積により求められる。TF は単語の網羅性を表し、IDF は単語の特定性を表しており、これらの積である TF・IDF は網羅性と特定性がともに高い単語の重みが大きくなるようになっている。TF・IDF では文書を単位として重みを計算するが、文書のかわりにカテゴリを単位として重みを計算したのが TF・ICF である。TF・ICF により重みを計算することで、文書単位で計算する TF・IDF より、カテゴリの内容をより反映した重み付けを行うことができる。

前処理で行うもう一つの処理は言語間でのカテゴリの対応付けである。あるカテゴリに対して、異言語の一つのカテゴリが対応するカテゴリとして選択される。それぞれの言語版の Web ディレクトリのカテゴリ数が同数であるとは限らないため、一つのカテゴリが複数の異言語のカテゴリから選択されることはあるが、一つのカテゴリが選択するカテゴリは一つだけである。

4.2. 曖昧性解消

曖昧性解消の流れを図 3 に示す。

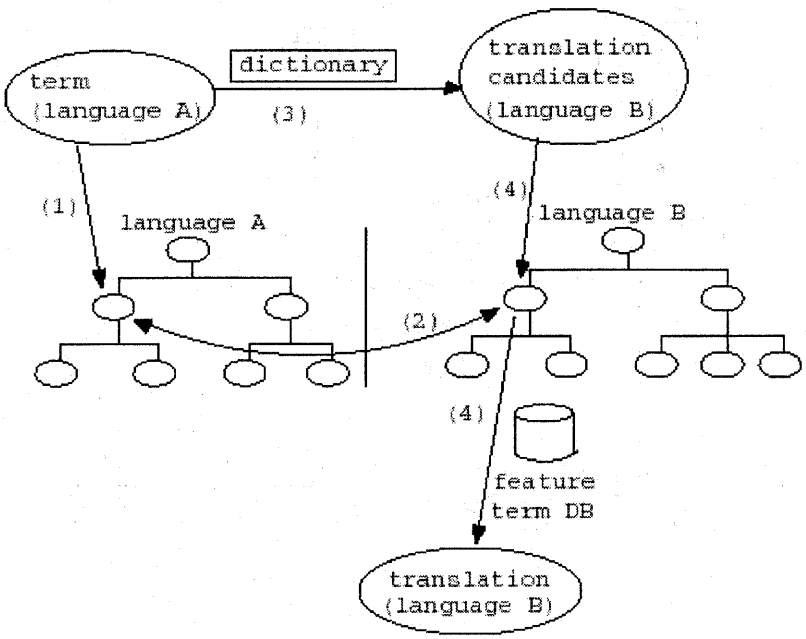

図 3.

- 1) 翻訳したい単語が英語の Web ディレクトリのどのカテゴリに適合するか推定する。
- 2) 推定された英語のカテゴリに対応付けられている日本語のカテゴリを選択する。

- 3) 翻訳したい単語の訳語候補を、対訳辞書をから全て抽出する。
- 4) 抽出されたそれぞれの訳語候補が、選択した日本語のカテゴリの特徴語として存在するか調べ、存在していたものを訳語として用いる。

1) において、一つの単語だけから適合するカテゴリを推定することは困難であるため、複数の単語を用いる必要がある。WordNet.OWL ではある一つの概念に複数の単語が登録されていることも多い。そこで、一つの概念に登録されている全ての単語を用いることにより、適合するカテゴリの推定を行う。

Web ディレクトリの同じカテゴリには類似した分野を主題とした Web 文書が属しているため、訳語候補のうちで同じカテゴリに属しているものは、類似した概念を表している可能性が高いと考えられる。よって、上記手法により類似した概念を表す単語を抽出することが可能である。

5. 実験

提案手法の有効性を検証するため、ある単語の訳語の中から類似する概念の訳語を抽出する予備実験を行った。今回の実験では、英語の"bank"という単語の訳語の中から"銀行"という日本語の単語の概念に類似するものを抽出した。

今回の実験において、Web ディレクトリは Yahoo の英語版と日本語版を用いた。尚、下位の階層のカテゴリを上位のカテゴリに統合することにより、それぞれの言語版の最上位のインデックスページに登録されている 13 カテゴリに統合した。ただし、英語版の"Regional"および日本語版の"地域情報"のカテゴリは除いている。また、最上位のインデックスページに登録されているカテゴリはどの言語版においても同じ構造であるため、その対応をそのまま言語間におけるカテゴリの対応付けに用いた。

Web 文書から単語を抽出する際に，英語版では単語の活用形を原形にしたのち，ストップワードを取り除いた．日本語版では，"茶筌"[10] を用いて名詞，動詞，形容詞，未知語を抽出した．特徴語の翻訳には，EDR 電子化辞書の対訳辞書 [11] を用いた．また、各カテゴリの特徴語数は、特徴語の重みの上位 5,000 語とした。

本手法ではまず、抽出したい概念がどのカテゴリに適合しているかを推定する。今回の実験では、上述の抽出したい概念が、英語版のカテゴリ"Business_and_Economy" に適合するものと仮定した。よって、このカテゴリに対応する日本語のカテゴリ"ビジネスと経済"が選択される。

次に、対訳辞書から英語の"bank"という単語の訳語候補を抽出する。"bank"という単語の訳語候補は以下のとおりであった。

峰, 灰をかぶせる, 銀行, 貯蔵所, 傾斜する, バンクさせる, 漕ぎ手席, 銀行に預ける, 副見出し, 親元, 岸, 外側を高くさせる, 傾いて飛行する, 堤で囲む, 積み上げる, ゴム縁, 袖見出し, 堆積, 湯銭, 土手, 堤, 配列する, 坑口, 銀行業を営む, 銀行へ預金する, 横に傾ける, 胴元の用意金, 漕ぎ手, バンク, ～バンク, 列に並べる, 土手を築く, 左右に傾ける, 堆積する, 片勾配, 堆, 縦坑口, テーブル, カウンター, 小見出し, 胴元になる, 浅瀬, 層をなす, クッション, 縦坑の入口付近, 貯金箱, 堤防, 堤のようになる, 傾斜させる, 横に傾く, 古代船のオールの列, 胴元, バンクする, 横傾斜, 隆起する, 積み重なる, タイプライターのキーの列, 砂州, 親元になる

そして、上記の訳語候補が日本語のカテゴリ"ビジネスと経済"の特徴語として含まれているかを調べ、含まれていた訳語のみを抽出する。その結果抽出された単語は、"峰", "銀行", "バンク", "カウンター" であった。

実験の結果、十分とはいえないが、全ての訳語候補のうちから目的の概念を表した単語に絞り込むことができた。今回の実験で目的以外の概念を表した単語も抽出された。原因は、カテゴリの統合によりカテゴ

リ数が少なくなったため、一つのカテゴリが対象とする分野の範囲が大きくなったからであると考えられる。また、茶筌では、"銀行に預ける"という語は"銀行"と"預ける"という二語に分割されるため、"銀行に預ける"という特徴語は存在しない。よって、こういった複数の単語に分割される訳語は抽出されないということも問題である。さらに、WordNet.OWLでは類似の概念であっても品詞が異なる場合は別の概念としているため、名詞だけを用いることで抽出精度が向上する可能性がある。

6. おわりに

本稿では、オントロジーを翻訳することにより、二言語を対象としたオントロジーを構築する手法を提案した。本手法では単語を翻訳する際に、Webディレクトリを利用した訳語の曖昧性解消の手法を用いることで、異なった概念を表した単語が、類似した概念を表した単語として抽出されることを回避する。また、本手法の有効性を検証するために予備実験を行った。今後は提案手法により実際に2言語オントロジーを構築する予定である。

参考文献

[1] 溝口理一郎. "オントロジー研究の基礎と応用" 人工知能学会誌, Vol.14 No.6, pp.45-56, 1999.

[2] 来村徳信, 笠井俊信, 吉川真理子, 高橋賢, 吉崎晃司, 溝口理一郎. "オントロジーに基づく機能的知識の体系的記述とその機能構造設計支援における利用" 人工知能学会論文誌, Vol.17 No.1 pp.73-84, 2002.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. "The Semantic Web" Scientific American, 2001.

[4] S. Bechhofer, F. V. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. "OWL Web Ontology Language Reference" W3C Candidate Recommendation, 2003.

[5] Knowledge, Information and Data Processing Group. "WordNet.OWL" http://taurus.unine.ch/GroupHome/knowler/wordnet.html

[6] "WordNet" http://www.cogsci.princeton.edu/~wn/

[7] 神崎正英. "日本語ウェブ・オントロジーの試み" http://www.kanzaki.com/docs/sw/jwebont.html, 2003.

[8] 酒井哲也, 梶浦正浩, 住田一男, Gareth Jones, Nigel Collier. "機械翻訳を用いた英日・日英言語横断検索に関する一考察" 情報処理学会論文誌, Vol.40, No.11, pp.4075-4086, 1999.

[9] 木村文則, 前田亮, 吉川正俊, 植村俊亮. "Webディレクトリの階層構造を利用した言語横断情報検索" 日本データベース学会 Letters, Vol.2, No.1, pp.71-74, 2003.

[10] "茶筌" http://chasen.aist-nara.ac.jp/

[11] "EDR電子化辞書, 対訳辞書" http://www.jsa.co.jp/EDR/J_index.html