最適例及び下位範疇化フレームを用いた動詞情報の獲得方式の研究

徐天晟　著

科学出版社

最適例及び下位範疇化フレームを用いた動詞情報の獲得方式の研究

徐天晟　著

科学出版社

北京

要　旨

機械翻訳等の自然言語処理技術の応用が広がるに伴い、言語処理のための言語知識(文法規則及び辞書データ) を格納する知識ベースの構築とその管理方式の確立が大きな課題となっている。これらの膨大な言語知識を人間が分析し、コンヒュータに入力するのは非常に手間がかかり、言語知識獲得システムを用いて効率化することが望まれる。更に、ある時点で必要な全ての知識がコンヒュータに組み込まれたとしても、新単語と新用例が常に生み出されてくる状況を考えれば、言語知識獲得のための有効な手段の開発が自然言語処理システムに必須の要件であることが分かる。本研究では、自然言語テキストから言語知識を自動的に獲得する方式について、特に、自然言語処理システムで重要な位置を占めている動詞情報の自動獲得について研究した。[Brent91]、[Manning93]、[Liu93] 等の研究では、コーパ スの大量なデータに基づき、動詞下位範疇化フレームの種類や文の構成要素に意味機能を付与する方式についての研究を行なっている。本研究ではこれらの意味機能に加えて、形態素情報や統語情報をも含め、未知動詞に関する言語情報の獲得を目的としている。未知動詞情報の獲得は、予め用意した下位範疇化フレーム及び未知動詞の典型的な用例を含む自然言語テキスト(最適例文テキスト) に基づき、副詞との共起関係や下位範疇化フレームと意味機能の関係等のヒューリスティックな手がかりを用いて行う。開発した評価用システムを用いて動詞情報の自動獲得実験を行い、本研究で提案した方法の有効性を確認した。

图书在版编目(CIP)数据

基于最适例及子范畴的动词信息自动学习方式的研究：日文/徐天晟著. —北京: 科学出版社, 2014.9

ISBN 978-7-03-041034-4

Ⅰ. ①基… Ⅱ. ①徐… ②… Ⅲ. ① 自然语言处理-研究 Ⅳ. ①TP391

中国版本图书馆 CIP 数据核字 (2014) 第 125508 号

责任编辑: 赵彦超　李静科 / 责任校对: 郭瑞芝
责任印制: 徐晓晨 / 封面设计: 陈　敬

科学出版社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com

北京彩虹印刷有限公司印刷
科学出版社发行　各地新华书店经销
*
2014 年 9 月第　一　版　开本: 720×1000 1/16
2014 年 9 月第一次印刷　印张: 5 1/2
字数: 110 000

定价: 48.00 元

(如有印装质量问题, 我社负责调换)

目　次

第1章　はじめに

近年コンヒューターの普及に伴い、機械翻訳などの自然言語処理システムが幅広く使用されるようになっている。自然言語は三つの重要な特徴を持っている。一つは、自然言語には単語及び文法に関わる膨大な知識が含まれていることである。例えば、ロングマン現代英英辞典には単語だけでも 56000 語が収録されている。その他に、単語の使用方法など無数の情報も含まれている。もう一つの特徴は、自然言語が変化していることであり、新しい単語と用例が常に生み出されている。このような自然言語を解析・生成するシステムは一般に自然言語処理システムと呼ばれている。従来の自然言語処理システム では、人手で膨大な言語に関する知識の収集・加工・蓄積を行なって来た。しかし、この方法は非常にコストがかかるため、言語知識獲得システムを用いて効率化することが望まれる。また、そのようなシステムは辞書にない単語や用法が出てきた時に、うまく対応出来ないことが多い。このためにも、自然言語処理システムには、新しい言語現象を分析し、その中から必要な言語知識を獲得する機能を備えることが必要となる。

本論文では、自然言語処理システムにおける言語知識自動獲得方式、特に、動詞情報の獲得の原理・方法等について述べる。言語知識というのは通常、文法規則と語葉情報の事を指す。

文法規則とは、小さい言語単位からより大きい言語単位を形成する時の規則である。語葉情報とは、単語に関する意味・属性等の情報である。本研究における動詞情報に対する獲得対象は、動詞の文法情報と意味情報とする。動詞の文法情報は動詞の使用パターン、すなわち、下位範疇化フレーム (Subcategorizaion Frame) の事を指す。意味情報は、機能文法 [Dik97(1)] の意味機能 (Semantic Function) のことを指す。意味機能は主題役割 (Thematic Role), 格範疇 (Case Category) などの名称でも用いられている。[Manning93]、[Brent91]などの研究では下位範疇化フレームの獲得について、特に下位範疇フレームの種類・構造について述べられている。本研究では、従来の研究のように下位範疇化フレームの種類・構造の解明を目的とするのではなく、未知動詞の下位範疇化フレームの獲得を目的とする。そのために、まず、ホーンビー [ホーンビー 77] とブレント [Brent91] 等の研究を参考にして、動詞の下位範疇化フレームを 14 個に分類した。下位範疇化フレームの獲得は予め用意したこの 14 個の下位範疇化フレームを未知動詞に適用し、正しく構文解析ができるかどうかによって行う。意味の獲得について、[Liu93]では、人間の介入を前提に曖昧性解消のためのヒューリスティックな情報と文法的な手がかりに基づいて行なう方法を提案している。この方法では、文の全ての構成要素に意味的な格範疇を付与するが、本研究では、動詞下位範疇化フレームの項に成りうるもののみを意味獲得の対象とする。英語では、動詞の意味と副詞の共起関係が良く知られている。例えば、

動作動詞 (Dynamic Verb) は slowly、 quickly などの速度を表す副詞と共起することが多い。機能文法 [Dik97(1)] では、共起関係など多くの手がかりに基づいて、 動詞の意味機能を決定する方式について研究している。本研究では、この機能文法の動詞の意味についての理論に基づいて、意味獲得に最適な例文をシステムに入力する方法を用いて意味機能の獲得を行う。

本研究は、子供の言語学習メカニズムの解明・応用にも重要な示唆を与える。子供は非常に短い期間で母語を学習してしまう。そのメカニズムを解明し、どのように自然言語処理に応用するかは大きな課題である。本研究は、子供が文構造と意味の内在的な繋がりに関する知識と最適な入力文から、未知語の意味を学習するという視点から見ることもできる。

本論文の 2 章で言語知識獲得について紹介する。3 章では、動詞情報の自動獲得の方法について述べ、4 章で動詞情報自動獲得システムの構築について述べる。最後に、5 章で実験方法とその結果につて説明する。

第2章　知識獲得

現代社会は膨大な情報が溢れており、情報社会といわれている。この膨大な情報の収集は従来主として人手に頼って行なわれて来た。この方法は効率が悪く、高いコストがかかる。これを解決するために、研究者達はコンヒュータが情報を自動的に獲得する方法について研究をして来た。本章ではこれらの研究の中の知識獲得及び言語知識の獲得に関する基礎について述べる。

2.1　知識獲得の概要

2.1.1　知識獲得とは

知識獲得とは、システムの開発・改良・拡張において対象ドメインの知識を抽出・分析し、実行可能でかつ汎用な形式にその知識を変換するとともに、獲得された知識と既存の知識ベースとの一貫性を保持・管理することをいう。つまり、新しい情報を学習することだけではなく、その情報を効果的に利用する能力も知識獲得の概念に含まれている。

2.1.2　知識獲得の方法

知識獲得の方法は主に3種類ある。一般的な命題から特殊

な命題を経験に頼らないで理論によって導くという学習方法(演繹学習)、例から知識を獲得する学習方法 (帰納学習) と説明に基づく学習方法である。ここでは、帰納学習を例にとって概説する。

帰納学習とは、外部教師あるいは環境から得られる事実に基づいて帰納推論を行ない、知識を獲得するプロセスである。帰納学習ではパターン表現という方法がよく用いられている。パターン表現を用いるアプローチとは、概念等の獲得対象をパターンによって表現し、そのパターンに基づいて知識の獲得を行なう方法である。帰納的学習は例からの学習と観察・発見による学習に分類できる。例からの学習は、概念獲得ともいう。つまり、教師に提示された例からクラス概念を獲得したり、限られた少数部分から全体像を構成したりすることである。また、観察・発見による学習は記述一般化ともいう。この方式では、教師の参与がなく、すべての例が与えられた時に、それらの例に関する概念クラスを獲得する。

本研究では、この3種類の知識獲得方法と違って、学習する対象ドメインの知識をいくつかのパターンにまとめ、パターンに基づいて知識の獲得を行なう。次章で説明する。

2.2　言語知識の獲得

第一章で述べたように、自然言語システムには言語に関する膨大な知識が含まれている。これらのデータを人手で入力するのは途方もない仕事であり、学習システムを用いてこの

問題を解決することが望まれる。自然言語は開いたシステムであり、新しい単語と用法が常に生み出されている。このような常に変化しているシステムをうまく運用するには、学習能力が必要となる。本節では言語知識の獲得の基礎理論について述べる。

2.2.1　言語知識と言語獲得の対象

言語知識には、辞書、文法規則、文と文の繋がりの情報、文章と内容の対応に関する知識、文章の言及する内容に関する知識など様々な知識が含まれる。普通、言語知識獲得の対象とする言語知識には下のようなものがある。

・意味情報

(i)　格情報

(ii) 意味素 (シソーラスにおける位置).

・文法情報

(i)　文の構成素の構文的役割

(ii)　文法規則

(iii) 下位範疇化フレーム

・文脈情報

(i)　照応関係

(ii) 因果関係

1. 意味情報

まず、意味情報について簡単に説明する。

意味情報には主に格情報と意味素が含まれている。ここで、

訳語選択等でよく使われている格について説明する。格 (case) とは名詞・代名詞・形容詞が文中で他の語に対してどのような関係にあるかを示す文法形式である。伝統的な理論では、主格 (nominative case)、属格 (genitive case)、与格 (dative case)、対格 (accusative case)、奪格 (ablative case)、所格 (locative case)、助格 (instrument case)、呼格 (vocative case) などがある。例えば、John opened the door with the key という文の場合、John は主格で、 the door は対格で、the key は助格である。フィルモア (C. J. Fillmore) が 1960 年後半に格文法 (Case Grammar) という理論を提唱した。格文法では格範疇 (Case Category) という概念が重要な位置を占めている。格範疇は深層格と名詞句からなる。格範疇には次のようなものがある：行為者 (agent)、経験者 (experiencer)、対象 (object)、道具 (instrument)、起点 (source)、着点 (goal)、場所 (location)、時間 (time) など。このような格を想定することにより、構造的には機能の異なる構成素の意味的関係を明らかにすることが可能になる。例えば、下のような文の場合に、表層上の相違にも関わらず、つねに the door は object, the key は instrument, John は agent である。

The door opened.

The key opened the door.

John opened the door with the key.

The door was opened by John.

格の概念は生成文法 [Haegeman94] 、機能文法 [Dik97(1)] などにも取り入れられている。本研究では、主に [Dik97(1)] の

機能文法に基づいて、格情報の獲得を行なった。詳しくは次章で説明する。

2. 文法情報

次に、文法情報について述べる。

言語知識獲得の対象となる文法情報は大きく三つに分けることが出来る。一つは、文の構成素の構文的役割である。文は、名詞句 (NP)、動詞句 (VP) などの構成素からなり、その構成素の構文的役割には品詞 (名詞、動詞、形容詞など)、動詞句、名詞句などが含まれている。例えば、 I love a girl という文の場合、個々の単語の品詞名、名詞句 (a girl)、動詞句 (love a girl) のような情報が獲得の対象となる。

言語知識獲得の対象となる文法情報のもうひとつは文法規則である。簡単にいえば、文の構成素の並べ方である。例えば、VO 型言語の英語の場合、動詞が目的語の前にくるという規則がある。それに対して、OV 型言語の日本語の場合、動詞が目的語の後ろにくるという規則がある。このような句構造規則が獲得対象となる。その他に、時制関係、共起関係などがある。

下位範疇化フレームとは、動詞、或は、形容詞や名詞がどんなような項をもっているかを表す構造である。例えば、John opened the door.という文の場合, 動詞[open]の下位範疇化フレームは下のようになる。

open[項 1 (John), 項 2 (the door)]

上の open の下位範疇化フレームは open という動詞が三

つの項を持っていることを表している。一番目の項は John, 三番目の項は the door である。文の構成素の構文的役割 (NP、Adj など) と下位範疇化フレームを用いて文の構造を表すこともできる。上記の文の構造は下のようになる。

open[項 1 (John, NP), 項 2 (the door, NP)]

英語の動詞の下位範疇化フレームは細かく分類しても数 10 個のパターンに限られている。本研究では、下位範疇化フレームに基づいて、個々の動詞の下位範疇化フレームの学習を行なう。

3. 文脈情報

文脈情報の中で学習の対象となるのは、照応関係と文脈因果関係がある。照応関係とは it のような指示詞の照応の同定規則のことをさす。文脈因果関係とは文の意味から、事実などを導くことを指す。

2.2.2　言語知識獲得の方法

本節では、節 2.2.1 の各種の言語知識の獲得方法について述べる。意味情報の獲得は格情報の獲得を例に取って述べる。また、文法情報の獲得は主に文構成素の構文的役割の同定と下位範疇化フレームについて述べる。

1. 意味情報獲得の諸方法

i) 対訳コーパスによる格フレームの獲得

対訳コーパスを用いて、格フレームの獲得を行なう方法がよ

く使われている [長尾 96]。例えば、コーパスに下のような文がある場合に、

(1) 私は上着をかぎにかけた。

(2) I hung my coat on the hook.

(1) の文の格フレームは

(3) 私 (主格、は) 上着 (目的格、を) かぎ (場所格、に) かけた。

というふうになる。

(2) は (1) の対訳例である。この文の 'on the hook' は動詞 'hung' にも係り得るし、'my coat'にも係り得るので、通常の構文解析では統語的曖昧性が残ることになる。ここで、日本語文 (1) の構文解析結果 (3) を見てみると、 (3) を使って、 (2) の英語の文の曖昧性を解消できることが解る。英文の格フレームをも獲得できる。下のようになる。

(4) I (主格) hung my coat(目的格) on the hook(場所格).

このように、対訳コーパスから格フレームの獲得がある程度できる。

(ii) 表層的手がかりによる深層格の獲得

表層的手がかりによる深層格の獲得という方法を用いる研究の内に、[Liu93] がある。この研究では、例えば、行為主体を表す意味的役割'agent' について、

(1) 有意志体である。

(2) 主語になり得る。

(3) 目的語にならない。

(4) 前置句になる場合の前置詞は'by' である。

(5) 意図を表す句とともに現れやすい。

(6) 命令文で省略される主語である。

といったヒューリスティックスを用意し、これらを用いて,意味役割 agent の判断を行なう。すべての意味的役割にたいしてこのようなヒューリスティックス規則を用意しておき、意味的役割の割り当ての曖昧性を解消し、同定を行なう。本研究では、この考え方を参考している。詳しくは 3 章で述べる。

2. 文法情報獲得の諸方法

i) 文の構成素の構文的役割の同定

文の構成素の構文的役割 (品詞名など) の同定では、ヒューリスティックスに基づく方法がよく用いられている。ヒューリスティックスに基づく方法の基本的な考え方について説明する。以下のような 2 つの例文を考える。

He gave the book to her yesterday.

He gave the flower to her yesterday.

この二つの文の唯一の違いは、'book' と'flower' である。このように 1 単語だけ異なる二つの文が与えられる場合、異なるそれぞれの単語は同じ品詞に属するという予測が成り立つ。これは同一品詞に属する単語は置換え可能であるという規則に基づいている。

ii) 下位範疇化フレームの獲得

下位範疇化フレームの獲得のうち、自然言語コーパスからの獲得が比較的よく研究されている。