

学 位 論 文 題 名

帰納的学習によるタグなし日本語文からの
統語・意味解析手法に関する研究

学位論文内容の要旨

日本語や英語など我々人間が日常生活で使用している言語をプログラミング言語などの人工言語に対して自然言語という。この自然言語を計算機上で扱うことを自然言語処理といい、解析、生成などの基礎技術から、かな漢字変換、機械翻訳、文書要約、対話システムなどの応用技術まで多岐に渡っている。かな漢字変換などの処理については既にある程度完成の域に達し一般に普及しているが、それ以外の高度な自然言語処理については未だ改善の余地が残されている。この理由として、高度な処理においては、入力された文章の字面による処理だけでは解決困難な点が多く、語彙や文法の知識から一般常識や背景知識なども含んだ総合的な知識に基づく判断が必要とされていることが挙げられる。このような判断を計算機に行わせることを解析といい、参照する知識のレベルに応じて、形態素解析、構文解析、意味解析、文脈解析の4種類に大きく分けられ、一般的な自然言語処理ではこの順序で解析が行われる。このうち、形態素解析と構文解析は、併せて統語解析とも呼ばれ、入力された文が語彙的、文法的に正しいかどうかを判断する。意味解析と文脈解析では、語彙や文法以外の知識から入力文の正しさを意味的に判断する。意味的な正しさというのは、統語的な正しさと比較して判断することが非常に難しく、また、判断に必要な知識がどのようなものであるかも不明な点が多い。そのため、形態素解析や構文解析に比べて意味解析や文脈解析の研究は進んでいないのが現状である。しかしながら、上に述べたように高度な自然言語処理のためには、意味を含んだ総合的な解析が必要とされている。それゆえ、本研究では形態素解析から文脈解析までを行う総合的な解析システムを作成することを最終的な目的として研究を行った。

総合的な解析システムを作成する際、問題となる点が大きく2点ある。一点目は、解析時に参照する知識を定量的評価ができるほど与えることが困難なことである。二点目は、意味解析や文脈解析などの上位の解析においては、形態素解析や構文解析のように確立した解析アルゴリズムが開発されていないことである。従って、総合的な解析システムを作成するためには、まず、意味解析や文脈解析における解析アルゴリズムを明らかにし、解析時に参照される知識を与える必要がある。それゆえ、本論文では、形態素解析から文脈解析までの統合的解析システムを最終的な目的としながらも、統語解析システムと意味解析システムの2つの独立したシステムを対象とした。統語解析システムでは、形態素解析

と構文解析の両方を対象として小規模ながらも統合的解析システムについて考察する。意味解析システムでは、未だ確立されていない意味解析アルゴリズムを明らかにする。また、どちらのシステムにおいても、解析時に参照される知識は、同じ基本思想に基づいた学習によって獲得する。学習の基本思想は、実例から学習するための帰納推論、ユーザによる正解を与えない教師なし学習、辞書の初期状態が空の状態からの学習、特定の言語に依存しない言語非依存性、解析対象の変化に追従する動的適応の5つから成り立っている。これら全てを備えた学習手法は本手法以前に提案されていない。以上を最終的な目的を達成するための下位目的として、本論文では研究を行った。

本研究では、言語に非依存で白紙の状態からでも教師なしで学習するための指標として頻度を用いた。獲得された知識が正しいかどうかの判断は多数決の原理に従って解決する。すなわち、「現実の文に多く出現する（または適用できる）知識は正しい」という仮説を立て、複数の知識が競合する場合、それらの内で最も多く出現した（適用された）知識が正しいとした。頻度情報ならば、特定の言語に依存した知識ではなく、難型となる知識がなくとも利用することが可能である。本研究の帰納的学習は、複数の実例に共通した要素を識別し共通要素に基づいて類似性を判別する能力（類推能力）と、実例中に出現した要素の頻度を計算し正誤の指標として用いる能力（統計能力）による学習である。

本手法を実装した統語解析システムでは、統語規則のない状態から、外国人のための日本語学習用テキスト 860 文を用いたクローズド実験で 85.3%の解析成功率が得られる規則を獲得できた。また、解析が成功した結果の 42.8%が正解であることを確認した。本手法を実装した意味解析システムでは、意味的知識を持たない状態で統語的情報だけから EDR コーパス 500 文を用いたオープン実験で 19.7%の解析正解率が得られる規則を獲得できた。以上から、本論文の帰納的学習が、統語解析システムと意味解析システムの両方において有効であることを立証した。

本論文は 8 章からなる。1 章では本研究の背景と目的を述べる。2 章から 4 章にかけては、研究成果が多く報告されている形態素解析と構文解析を対象とし、両者を統合した統語解析に関する研究を行う。2 章では形態素解析と構文解析の基礎知識を解説し、3 章では本論文の統語解析システムを説明する。4 章では実際に作成したシステムの評価実験を行い、その結果から統合システムを作成する際の課題を明らかにし、その対処方法について考察する。5 章から 7 章にかけては、比較的研究成果の少ない意味解析を対象とした研究を行う。5 章では意味解析の基礎知識を解説し、6 章では本論文の意味解析システムを説明する。7 章では実際に作成したシステムの評価実験を行い、その結果から意味解析における課題を明らかにし、その対処方法について考察を行う。8 章では結論を述べ、さらに、形態素解析から文脈解析までを含んだ総合的な解析システム作成のための課題についても述べる。

学位論文審査の要旨

主査	教授	栃内香次
副査	教授	北島秀夫
副査	教授	青木由直
副査	助教授	荒木健治

学位論文題名

帰納的学習によるタグなし日本語文からの 統語・意味解析手法に関する研究

自然言語処理の基本は、処理システム内に蓄積された種々の統語・意味規則を用いて入力文を解析することである。したがって、これらの統語・意味規則をいかにして構築するかが自然言語処理手法の中心課題となる。これまで用いられてきたのは、これらの規則集合を多数の例文などより人手で収集し、システムに蓄積する手法である。しかしながら、ますます大量、かつ変化の速度が大きい多種多様な文書のコンピュータ処理が必須となっている今日、人手で収集することはすでに不可能になりつつあり、これを自動的に収集・獲得する手法の確立が急務となっている。

このような観点から、近年大量の文書からそれらに内在する規則集合を学習により獲得する学習型の手法、中でも品詞付等の前処理を行っていない生の文、すなわちタグなしコーパスからの学習型規則獲得手法が注目されている。

上記の観点に立って、著者は本論文において、帰納的学習を用いてタグなし日本語文から統語・意味解析を行なうために必要な規則を獲得する手法を提案し、実験的にその有効性を確認した。

著者は、統語解析システムと意味解析システムの2つのシステムを対象とし、文の解析時に参照される知識を同じ基本的枠組に基づいた学習によって獲得する手法を提案した。基本枠組は、実例から学習するための帰納推論、ユーザによる正解を与えない教師なし学習、辞書の初期状態が空の状態からの学習、特定の言語に依存しない言語非依存性、解析対象の変化に追従する動的適応の5つから成り立っている。これら全てを備えた学習手法は本手法以前に提案されていない。本研究の帰納的学習は、複数の実例に共通した要素を識別し共通要素に基づいて類似性を判別する能力（類推能力）と、実例中出现した要素の頻度を計算し正誤の指標として用いる能力（統計能力）による学習である。

著者は、提案手法に基づく統語解析実験システムを構築し、統語規則の全く存在しない

状態から、外国人のための日本語学習用テキスト 860 文を用いたクローズド実験を行い、85.3%の解析成功率が得られることを確認した。また、解析が成功した結果の 42.8%が正解であることを確認した。ついで同様に提案手法に基づく意味解析実験システムを構築し、意味的知識を持たない状態で統語的情報だけから、EDR コーパス 500 文を用いたオープン実験で 19.7%の解析正解率が得られることを確認した。以上から、著者が提案した帰納的学習が、統語解析規則獲得と意味解析規則獲得の両方において有効であることが立証された。

各章の概要は以下の通りである。1 章では本研究の背景と目的を述べた。2 章から 4 章にかけては、研究成果が多く報告されている形態素解析と構文解析を対象とし、両者を統合した統語解析に関する研究を行った。2 章では形態素解析と構文解析の基礎知識を解説し、3 章では本論文の統語解析システムを説明した。4 章では実際に作成したシステムの評価実験を行い、その結果から統合システムを作成する際の課題を明らかにし、その対処方法について考察した。5 章から 7 章にかけては、比較的研究成果の少ない意味解析を対象とした研究を行った。5 章では意味解析の基礎知識を解説し、6 章では本論文の意味解析システムを説明した。7 章では実際に作成したシステムの評価実験を行い、その結果から意味解析における課題を明らかにし、その対処方法について考察を行った。8 章では結論を述べ、さらに、形態素解析から文脈解析までを含んだ総合的な解析システム作成のための課題についても述べた。

以上を要約すると、著者は類推と統計情報に基づく帰納的学習を用いた統語・意味解析手法を提案し、教師なしで自動的に知識を獲得し、かつ、対象に動的に適応できることを示した。また、研究領域の現状の分析と対比して新規提案内容の記述、有効性の主張、研究領域における位置付けを行っており、自然言語処理工学、情報メディア工学の発展に貢献するところ大である。

よって著者は北海道大学博士（工学）の学位を授与される資格あるものと認める。