様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１３　日現在

機関番号：　１３９０１
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１６０４９

研究課題名（和文）日本語意味解析のための意味辞書および機能語用例データベースの開発

研究課題名（英文）Semantic dictionary and function word usage database for Japanese semantic analysis

研究代表者

松崎　拓也（Matsuzaki, Takuya）

名古屋大学・工学研究科・准教授

研究者番号：４０４６３８７２

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：数学問題テキストを対象ドメインとし，組合せ範疇文法とよばれる形式文法枠組みに基づく日本語の意味辞書を開発した．辞書は8,000の表層形に対する55,000のエントリを含む．また，この文法に基づく構文・意味解析器やその高精度化のための技術を開発した．さらに，問題全体に対する意味表現を得るため，文間関係解析や照応解析など複数の文の意味から問題テキスト全体の意味を合成するための処理を実現し，問題自動解答を通じた意味表示の精度評価や未知語・未知用法の発見のための基礎を実現した．

研究成果の概要（英文）：A Japanese semantic dictionary was developed. It is based on a formal grammar framework called Combinatory Categorial Grammar. The target domain is math problem texts. The dictionary contains 55,000 lexical entries for 8,000 surface forms. A syntactic/semantic parser based on the grammar was also developed. Furthermore, inter-sentential semantic analysis modules, such as a coreference resolution system and a discourse analysis system, were developed for generating a semantic representation of a problem including several sentences. Out-of-vocabulary words and usages can be found through automatic problem solving based on these analysis systems.

研究分野：　自然言語処理

キーワード：　文法開発　構文解析　意味解析

１．研究開始当初の背景

(1) 構文解析技術は、2000 年代半ばまでに、日本語・英語に対する解析精度が約 90%で頭打ちとなり、本研究の開始時には、基礎研究および応用の両面において、構文解析に基づく意味解析へと踏み込む研究動向が本格化しつつあった。しかし、すでに行われている研究には、自然言語入力によってデータベースを知識源とするファクトイド質問応答を行うようなタスクに関するものが多く、旧来のデータベースに対する自然言語インタフェースを質的に大きく超えるタスクに取り組む例は少なかった。

(2) 形式的推論を始め、さらに高度な知的情報処理技術のためのインタフェースとして自然言語を用いるためには、助詞・助動詞といった機能語およびそれらが複合した機能表現をその意味へと対応付ける意味辞書の整備が必要である。しかし、研究開始当初時に存在した日本語辞書のうち、複雑な推論を可能とするレベルの精緻な意味表現を出力可能なものは存在しなかった。

２．研究の目的

(1) 日本語意味解析のための基礎資源として、助詞・助動詞・接続詞・様々な量化表現など、機能語・機能表現の種々の用法について、統語特性の形式的表示と高階論理による意味表示を対応付けた意味辞書を開発する。対象ドメインとしては、複雑な意味内容を含み、問題解答を通じて意味表現の精度評価が可能である大学入試数学問題を想定する。

(2) 上記(1)の開発のための中間データとして、数学問題テキストから機能語・機能表現の用例を収集し、係り受け構造、述語項構造および論理式による意味表示を付与した言語資源としてまとめる。

(3) 数学問題テキストへの係り受け構造・述語項構造のアノテーションを行い、そこから半自動的に内容語の統語・意味記述を取り出すことで内容語の意味辞書と注釈つき用例集を同時に作成する。また、開発段階の(1)、(2)の意味辞書を用いた自動意味解析（の失敗）によって、未知用例の収集を効率的に行う技術を開発する。

３．研究の方法

(1) 新聞テキストに対する既存のアノテーション付きコーパスに加え、入試数学テキストに対する少量のアノテーションおよび専門用語リストを用いることで、意味解析の前提となる形態素解析・構文解析の精度を向上させる。

(2) 収集した入試数学テキストに対する観察・分析に基づき、基礎的な語彙に対してそれらの統語的性質と意味の形式的な記述を行うとともに用例を集積する。

(3) (2) で開発した基礎的語彙に対する辞書に基づき意味解析を行う構文・意味解析器を開発する。

(4) 問題の自動解答を通じてテキスト解析の精度評価を行うために、テキストに含まれる数式の意味解析や文間関係の解析など、問題全体に対する意味表現を生成するために必要となる解析器を開発する。

(5) 問題の自動解析および自動解答・採点の結果を通じて辞書に未収録の用法の発見を行う。

４．研究成果

(1) 2015 年度：組合せ範疇文法（Combinatory Categorial Grammar、 CCG）による日本語文法の記述および内容語・機能語の辞書の拡充を進め、構文解析器の開発を行った。文法記述に関しては、特に、動作とその結果の表現、複数のモノを表す名詞句にまつわる種々の意味現象の分析、項を取る名詞（不飽和名詞）を含むさまざまな構文の分析などを重点的に行った。これらの現象はテキストジャンルを問わず普遍的に現れるものであり、かつ、文の意味内容の論理的把握のためにはその精緻な解析が必須となる。

　また、文内の論理構造解析のための文法開発に加え、文と文の間の意味関係がなす談話構造の表示に関して、実テキストを基にした分析を行い、比較的少数の文結合オペレータの組合せで多くの構造が表示可能であるとの予測を得た。

　構文解析系は実際に開発した文法を用いる言語処理システムの基盤であるとともに、文法開発の過程でも文法の整合性テストや未収録の語彙・用法を収集するツールとして必須のソフトウェアである。2015 年度はこれまで多くの研究がある文節単位の係り受け解析を組合せ範疇文法による解析の前処理として用いる２ステップ構文解析の実装およびその改良を行った。

　その他、辞書検索・テキストベース検索・構文解析結果の表示など文法開発のための基本的なツール群の開発・整備を進め、効率的な開発を可能とする環境づくりを行った。

　また、組合せ範疇文法による文解析の結果得られるものを想定した形式的意味表示データを人手により作成し、現実的な時間で意味処理が可能であるか検証する予備的実験を行った。その結果、1 階の実閉体（Real Closed Field）の言語で記述可能な大学入試問題のおよそ 6 割に対して、形式的な推論によって正解が得られることを確かめた。さらに、その副産物として、自然言語からの機械的翻訳結果を想定した形式的意味表現を推論の入力とする場合、人にとっての問題の難しさと形式的な推論による正解率との間に相関があることが分かった（表１、表２）。

これらの結果を自動演繹に関する最高峰の国際会議の１つである International Joint Conference on Automated Reasoning および認知科学に関するトップクラスの国際会議である CogSci を含む国際学会３件で発表した。

表１：練習問題（Chart）、国立大学２次試験問題（Univ）、国際数学オリンピック問題（IMO）に対する正答率・計算時間の比較

		Succeeded	
		Success %	Time (sec) Min/Med/Avg/Max
Chart	RCF	63.8% (111/174)	13/18.0/ 37.4/ 343
	PA	57.1% (48/ 84)	12/17.0/ 20.3/ 172
	Other	10.0% (3/ 30)	13/14.0/ 17.7/ 26
	All	56.3% (162/288)	12/17.0/ 32.0/ 343
Univ	All (RCF only)	58.0% (142/245)	12/26.5/ 85.5/1417
IMO	RCF	16.5% (19/115)	14/25.0/ 51.8/ 197
	PA	4.8% (2/ 42)	25/29.5/ 29.5/ 34
	Other	3.6% (2/ 55)	17/24.5/ 24.5/ 32
	All	10.8% (23/212)	14/25.0/ 47.5/ 197

表２：IMO 問題に対する IMO 参加者の正解率（Human Efficiency）と自動推論による正解率（Machine Efficiency）との比較

Years	Human Efficency	Machine Efficiency	Succeeded
1959-69	58.23%	21.11%	26.3% (15/57)
1970-79	46.57%	7.00%	13.3% (4/30)
1980-89	44.35%	1.85%	3.1% (1/32)
1990-99	38.27%	3.33%	5.7% (2/35)
2000-13	34.31%	1.19%	1.9% (1/54)

表３：テキスト中の数式・変数の解釈の一貫性を考慮することによる解析精度の向上

Dataset	Parsing w/ type env.	Typing failure (%)	Correct answer (%)
DEV	no	9.8%	21.8%
	yes	0.6%	27.6%
TEST	no	8.6%	8.6%
	yes	0.0%	11.4%

(2) 2016 年度：組合せ範疇文法による日本語文法の記述および内容語・機能語の辞書の拡充をさらに進めるとともに、構文解析系の精度を向上させるための研究を行った。

文法記述に関しては、内包的表現の分析とその辞書項目としての実装を重点的に行った。その他さまざまな言語現象に対応する辞書項目を記述した結果、辞書は表層形約 8,000 に対し 55,000 エントリを実装した。

構文解析器に関しては、冗長な解析結果を排除する処理や係り受け解析レベルでのエラーを自動修正する処理などを実現した結果、解析効率が大幅に向上した。さらに、テキスト中に含まれる数式や変数の解釈に関する一貫性を構文解析中に考慮することで問題全体に対する解析精度を向上させる技術を開発した（表３）。

さらに、文と文の間の意味関係がなす談話構造の分析を行うソフトウェアを開発し、数学問題テキストの解析による評価を行った。これらの結果を自然言語処理分野の最高峰の国際学会である Annual Meeting of the Association for Computational Linguistics（ACL）において発表した。

(3) 2017 年度：組合せ範疇文法に基づいて、形式的な推論を可能とするレベルの精緻な意味表示を導出できる日本語文法・辞書の記述および、辞書の記述と対応する用例の収集を進めた。また、語彙および語の用法を半自動的に収集するための基礎技術でもある構文解析器の高性能化を行った。

特に 2017 年度は、数学テキストの解析のために、数式を含むテキストに対する構文解析技術の高度化に取り組んだ。具体的には、LaTeX あるいは MathML の形式で表現されたテキスト中の数式を解析して、その意味構造および統語カテゴリを推定したのち、数式の周囲の語の用法との整合性をとりながら文全体の構文・意味解析を行う技術の高度化を行った。これにより、LaTeX による数式記述を含む大量の数学テキストを自動的に解析することが可能になった。

また、文境界を越え、テキスト全体の整合的な解析を行うために、照応表現およびゼロ照応の指示対象を同定する処理についての研究を行った。テキスト全体に対する正確な意味表示を得るためには、これまで広く研究が行われてきたモノ（entity）を指す照応表現の解析に加え、これまでの研究の蓄積が少ないコト（命題）を指示する表現の解析が必要になる。この照応表現解析に関する研究では、前者に加え後者についても研究を行い、文書構造（例：「(1)の条件」）や後方参照（例：「以下の条件」）などいくつかのタイプのコトを指す指示表現の解析技術を開発した。以上の成果について、国際学会における口頭発表および招待講演、国内学会におけるチュートリアル講演などを通じ報告した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① Takuya Matsuzaki、 Hidenao Iwane、 Munehiro Kobayashi、 Yiyang Zhan、 Ryoya Fukasaku、 Jumma Kudo、 Hirokazu Anai、 Noriko H. Arai、Can an A.I. win a medal in the mathematical olympiad? - Benchmarking mechanized mathematics on pre-university problems、AI Communications、査読有、 Vol. 31 、 pp.251-266 、 DOI 10.3233/AIC-180762、 2018.

〔学会発表〕（計６件）

① 松崎拓也、計算機が大学入試数学問題を解く、日本数学会 2018 年度年会特別講

演、2018.
② Takuya Matsuzaki、 Machine comprehension of math problem text、3rd Conference on Artificial Intelligence and Theorem Proving、2018.
③ Takumi Ito 、 Takuya Matsuzaki 、 Satoshi Sato、Coreference Resolution on Math Problem Text in Japanese、International Joint Conference on Natural Language Processing、2017.
④ Takuya Matsuzaki、Semantic Parsing of Pre-university Math Problems 、Annual Meeting of the Association for Computational Linguistics、2017.
⑤ Takuya Matsuzaki 、 An Information-Processing Account of Representation Change: International Mathematical Olympiad Problems are Hard not only for Humans but also for Machines、the 38th Annual Meeting of the Cognitive Science Society、2016.
⑥ Takuya Matsuzaki、Race against the Teens – Benchmarking Mechanized Math on Pre-university Problems 、International Joint Conference on Automated Reasoning、2016.
⑦ Takuya Matsuzaki、Solving Natural Language Math Problems、 1st Conference on Artificial Intelligence and Theorem Proving、2016.

〔その他〕
ホームページ等
https://researchmap.jp/mtzk/

６．研究組織
(1)研究代表者
松崎　拓也（MATSUZAKI、 Takuya）
名古屋大学・大学院工学研究科・准教授
研究者番号：40463872