様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年　5 月 13 日現在

機関番号：６２６１５
研究種目：基盤研究(C)
研究期間：2009～2012
課題番号：21500131
研究課題名（和文）
超大規模半構造化知識ベースとの融合による高度な言語処理技術の実用化
研究課題名（英文）
Accurate Natural Language Processing with Huge Semi-structured Textbases
研究代表者
松崎　拓也（MATSUZAKI TAKUYA）
国立情報学研究所・社会共有知研究センター・特任准教授
研究者番号：40463872

研究成果の概要（和文）：構文解析など基本的な言語処理を施した大量のテキストデータを用いて、そこから必要な情報を動的に抽出することで種々の言語処理技術を高精度化することを目指し研究を行った。具体的な成果として、大規模半構造化データベースに対する高速な検索システムを開発し、それを応用した知的テキスト検索システムを実現した。また、大量テキストデータから動的に抽出した統計量を従来の解析モデルに統合する枠組みに関する基礎研究を、構文解析および共参照・照応解析を対象として行い、それぞれについて高精度な解析システムを実現するとともにテキストベースとの統合へ向けての知見を得た。

研究成果の概要（英文）：We have developed the technical basis for accurate natural language processing through the integration of a huge textbase with various processing modules. Specifically, we developed an efficient search engine for huge semi-structured databases and an intelligent literature search system based on the search engine. We also investigated how we can enhance the accuracy of syntactic parsers and coreference resolvers by using external knowledge that shall be extracted on-the-fly from the textbase.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,400,000	420,000	1,820,000
2010 年度	1,300,000	390,000	1,690,000
2011 年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：構文解析、共参照解析、データベース

１．研究開始当初の背景

人手による解析済みデータを「お手本」として統計的言語処理モデルの学習を行うという枠組みは、実用レベルの言語処理ソフトウェアを素早く開発するための基本的な手法として 80 年代以降広く定着した。新しい統計モデルや学習手法の開発は現在も活発に行われているものの、処理精度がある程度のところで頭打ちになるという現象が近年様々な言語処理タスクで共通に起こっている。その大きな原因は知識ボトルネック、すなわち、高精度な言語処理を行うのに必要な統語的・意味的知識の詳細さと多様さに比べ、処理モデルの学習源である人手による解析

データが圧倒的に足りないことにある。

大量の知識を得るための一つの方策として、大規模なテキストデータに対して種々の言語処理を行い、そこから様々な統計量をシンボリックな知識の近似物として取りだすことが考えられる。しかし、様々な言語処理タスク、様々な入力テキストに対して、一般に必要となる知識・統計量を事前に列挙し、全て得ておくことは困難であり、処理済みの大規模テキスト（テキストベース）から必要な知識を動的に得るような手法が望ましいと考えられる。

研究代表者および分担者は、本研究以前に半構造化テキストベースに対する知的検索システムの開発を行ってきており、この検索システムのコアである構造検索エンジンをさらに高速化し、言語処理モジュールが知識ベースから動的に情報を抽出するための基盤とすることで様々な言語処理モジュールを高精度化するという着想を得た。

２．研究の目的

本研究の目的は、構文解析を始めとする言語処理基盤技術と超大規模テキストベースに対する構造検索処理を密に結合し、高精度な言語処理のための知識源としてテキストベースを利用することで、種々の言語処理基盤技術の精度を飛躍的に向上させることである。大量のテキストに対し構文解析・固有名認識など多種の自動処理を施すことで情報を付加したテキストベース(TB)を構築し、この TB に対する高速な構造検索処理を従来の統計モデルによる言語処理手法と融合することで、入力テキストを正確に処理するために必要な情報を TB から動的に抽出することができると考えられる。

本研究では、特定の言語処理タスクのための巨大なシステムの開発ではなく、言語処理モジュールと TB を結合するための基礎理論を構築し、統一的な方法で様々な処理タスクを共通の TB に結合することを目指す。この目的は、処理モジュール毎に異なる大規模 TB を構築するコストを避けるという現実的な要求であるとともに、大規模知識ベースに支えられた言語処理という大きな枠組みを、様々なタスクに実際に応用できる具体的な技術として提示するという構想に基づく。

これを実現する技術的な鍵は、研究分担者を中心に開発を進めてきた拡張領域代数に基づく構造検索エンジン(GCL)を TB 検索のための共通基盤として用いることである。GCL は複雑な構造検索を行えると同時に、既存の構造検索エンジンでは対応不可能な超大規模 TB に対しても動作するという特徴を持つ。しかし、言語処理モジュールと検索エンジンを統合するに当たっては、検索エンジンの速度を大幅に向上させる必要がある。

３．研究の方法

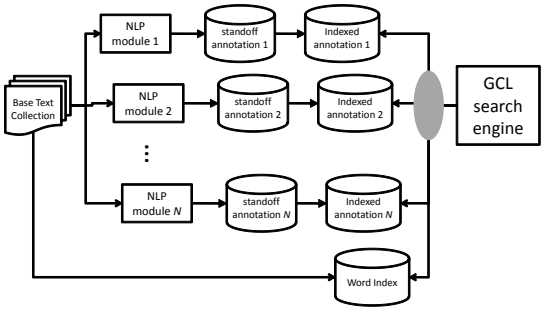
（１）既存の GCL 検索手法は人間が対話的に用いる文献検索システムへの応用を想定して設計されており、数百ギガバイト程度の TB に対して１クエリあたり数十秒の検索時間がかかる。これを、１文に対し多数回の検索を行いながら実用的な速度で動く言語処理モジュールへと応用するにあたっては、さらに高速な構造検索アルゴリズムの開発が必要になる。本研究では、特に、頻出クエリに対してそれを弱く近似するクエリに対する検索結果をあらかじめ索引付けしておくことで検索結果の候補を大幅に絞り込み、検索を高速化する手法に着目し、これを実現した。

（２）構文解析と共参照・照応解析をモデルタスクとし、現在の曖昧性解消モデルに不足している知識要素を究明すると同時に、それらの知識要素を教師付き学習による曖昧性学習モデルと柔軟に統合するための枠組みについて研究する。

特に、構文解析については、 Supertagging と呼ばれる処理ステップに着目して研究を行う。Supertagging は構文解析の一種の前処理だが、後続の解析処理の精度および速度に大きく影響する重要なステップである。しかも、Supertagging は形式的には単純な系列ラベリングタスクと見なせるため、多様な知識を取り込むことが容易であるという特徴をもつ。

４．研究成果

（１）高速な検索が可能で、かつ様々な言語処理の結果を柔軟に取り込める半構造化データ検索システムの開発を行った。このシステムの特徴は、第一に、テキストに情報を付加する言語処理モジュール毎に、それらの出力を索引付けして格納するデータベースを用意し、検索時にそれらを横断的に検索する構造である（下図）:



これにより、新たな言語処理モジュールの追加や、言語処理モジュールの改良に伴う既存の処理結果の置き換えなどの操作がデータベース全体を更新することなく柔軟に行え

るようになった。
第二の特徴は、前節で述べた近似クエリに対する検索結果の事前の索引付けによる検索の高速化の仕組みである。この手法では、まず、頻出する事が予想される部分クエリの集合に対し、それらの検索結果を事前に索引付けしておく。次に、与えられた複雑なクエリに対して、それに論理的に包含される部分クエリを可能な限り列挙する。最後に、列挙された部分クエリに対し、事前に索引付けされた結果を読みだし、それらの共通部分をブーリアンクエリに対する AND 検索と同様に計算する。これにより、検索クエリにヒットする必要条件を全て満たす結果の候補が大幅に絞り込め、時間がかかる詳細な検索条件のチェックを行う候補テキストを減らすことで高速な検索が実現できた。
この検索システムは言語処理モジュールとの統合の基礎技術となるとともに、それ自体を生命・医学文献の検索システム MEDIE へと応用し、サービスを WEB で公開した。

（2）構文解析タスクについて、テキストベースとの統合に向けた解析モデルの開発を中心とした研究を行い、以下の成果を得た。

① Supertagging 処理のモデルとして、最も単純な単語ごとの予測モデルが有効であることを実験的に明らかにした。構造化パーセプトロンなど、より複雑なモデルにも、精度が多少向上する利点はあるものの、その差は微小であり、単語ごとの予測モデルがもつ柔軟性の利点が勝ると結論できた。

②単語ごとの予測モデルと同型の Supertagging モデルを用いながら、その学習方法を変更することで、解析処理時におけるモデルの単純さの利点を保ったまま、精度を向上させる手法を開発した。この手法は、単語ごとの予測モデルと全く同じ素性を用いるモデルだが、(i) 学習を、CRF などの系列ラベリングモデルと同様に、文中の全単語に対して一斉に割り当てる設定で行う；(ii) 学習時の解候補の探索空間を、文法的に可能な supertag の組合せだけに限定する、という２つの改良を加えることで、より解析精度の高いパラメータセットを得るというものである。学習は通常の系列ラベリングのような設定で行うものの、全ての素性が単語ごとの予測モデルと同一であることから、Supertagging 以降の処理は単語ごとの supertagging モデルの場合と全く同様に行うことができる。このため、モデル・処理の複雑化を招くことなく、解析精度を大きく向上させることが可能になった。

③ 係り受け解析器の出力を Supertagging 処理の曖昧性解消における素性のひとつとして用いることで、高精度な Supertagging を行う手法を開発した。これは、表層的に離れた位置にあるが、統語的に関係をもつ単語の情報を直接 Supertagging 処理へ反映させることで、精度を向上させたものである。この結果は、近年、係り受け解析における研究において明らかになった大規模テキストデータを用いた self-training の手法の有効性を考え併せると、係り受け解析結果を含むテキストベースの検索と Supertagging 処理との統合の有効性を示唆する。

④ 統語的曖昧性のなかでも、特に正確な解決が難しい事が知られている並列句の構造曖昧性について研究を行った。並列句の曖昧性の正しい解消には、意味情報や構造的な類似性など多様な要素を考慮することが必要である。これを解決するため、CKY アルゴリズムに基づく従来型の構文解析手法と、近年提案された並列された句どうしの表層的な類似性を用いる並列句解析手法を、双対分解を用いて統合するモデルを提案し、その有効性を示した。このモデルにおける双対分解の利用法は、従来の教師付き学習に基づく解析アルゴリズムとテキストベースから得た統計量を用いた解析の統合にも直接応用可能であると考えられる。

（3）共参照・照応解析（先行詞の同定）のタスクにおいて、正確な解析に必要となる知識のタイプについて詳細な分析を行い、分野特有の知識の重要性を明らかにした。また、多種の知識・素性を考慮した解析モデルを提案し、その有効性を実験的に示した。

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 3 件）

① Ngan Nguyen, Jin-Dong Kim, Makoto Miwa, Takuya Matsuzaki and Junichi Tsujii. Improving protein coreference resolution by simple semantic classification. BMC Bioinformatics. 査読有. 13:pp.304-315. 2012.
doi: 10.1186/1471-2105-13-304

② Yao-zhong Zhang, Takuya Matsuzaki, Jun'ichi Tsujii. Structure-guided supertagger learning. Natural Language Engineering. 査読有. 18(2): pp205-234. 2012.
doi: 10.1017/S1351324912000034

③ Katsuya Masuda, Takuya Matsuzaki and

Jun'ichi Tsujii. Semantic Search based on the Online Integration of NLP Techniques. Procedia - Social and Behavioral Sciences. 査読有. 27:pp281-290. 2011.
doi: 10.1016/j.sbspro.2011.10.609

〔学会発表〕(計4件)

① Atsushi Hanamoto, Takuya Matsuzaki, Jun'ichi Tsujii. Coordination Structure Analysis using Dual Decomposition. The 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012). 2012年4月23日-27日. Avignon, フランス. (査読有)

② Yao-zhong Zhang, Takuya Matsuzaki, Jun'ichi Tsujii. A Simple Approach for HPSG Supertagging Using Dependency Information. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010). 2010年6月3日. ロサンジェルス, アメリカ. (査読有)

③ Yao-zhong Zhang, Takuya Matsuzaki, Jun'ichi Tsujii. Forest-guided Supertagger Training. International Conference on Computational Linguistics (COLING-2010). 2010年8月26日. 北京, 中国. (査読有)

④ Yao-zhong Zhang, Takuya Matsuzaki, Jun'ichi Tsujii. HPSG Supertagging: A Sequence Labeling View. The 11th International Conference on Parsing Technologies (IWPT-2009). 2009年10月7日-9日. パリ、フランス. (査読有)

〔その他〕
研究代表者ホームページ
http://researchmap.jp/mtzk/

検索システム MEDIE デモページ
http://www-tsujii.is.s.u-tokyo.ac.jp/medie/

６．研究組織
(1)研究代表者
松崎　拓也 (MATSUZAKI TAKUYA)
国立情報学研究所・社会共有知研究センター・特任准教授
研究者番号：40463872

(2)研究分担者
増田　勝也 (MASUDA KATSUYA)
東京大学・工学（系）研究科（研究院）・特任研究員
研究者番号：20512114

(3)連携研究者
なし