

論文内容の要旨

博士論文題目 大規模コーパスの解析と検索に関する研究

氏名 山下 達雄

(論文内容の要旨)

本論文は、大規模言語コーパスに関して、言語情報、特に形態素情報の自動付与技術とそのような情報が付与されたコーパスを解析および検索するために利用できる技術を提案し、その有効性を示すことを目的としている。

言語情報が付与された大規模なコーパス、いわゆるタグ付きコーパスは、近年の自然言語処理において、不可欠なものとなっている。例えば、大量のタグ付きコーパスから統計的な手法によりパラメタを学習し、それを用いた形態素解析器やパーザが開発されており、それらは意味解析などの深い言語処理の研究に役立っているだけでなく、表層的な言語情報を用いて最大限の効果を得ることを目的に情報検索などで利用されている。タグ付きコーパスを大量に作成していくということは、自然言語処理の基礎データの蓄積を意味し、自然言語処理研究に欠かせない重要な課題である。

本研究では、まず、情報付与の対象となる言語に着目し、特定の言語や同系統の数言語の解析のみを念頭に置いて開発されている現状の言語技術を拡張すべく、いくつかのシステムの提案と構築を行った。

最初に、言語非依存の形態素解析の枠組を提案した。同じ解析システムに対して、言語毎の特徴を記述するだけで、様々な言語に対して適用可能な方法を提案して実装し、実験により有効性を示した。

次に、タグ付きコーパスを作成する人間に着目し、タグ付きコーパスを効率的に蓄積するための支援環境(グラフィカル・ユーザ・インターフェース)を提案した。また、タグ付きコーパス作成作業の効率化への寄与という視点に基づき、これまでのコーパスから統計的な手法で学習されたパラメタを用いる解析手法ではなく、人手作成ルールとの融合や事例ベースによる解析手法といった利用法を提案し有効性を示した。

さらに、解析処理以外でのコーパスの利用法として、人間によるタグ付きコーパス蓄積作業の際の「参照」に着目し、そのための、大規模なタグ付きコーパスからの類似文字列検索を高速、高精度で行なう手法を、Suffix array を拡張することによって提案し、その有効性を確認した。

(論文審査結果の要旨)

平成 13 年 12 月 25 日に開催した公聴会の結果を参考に平成 14 年 2 月 12 日に本博士論文の審査を行った。以下のとおり、本博士論文は、提案者が独立した研究者として、研究活動が続けていくための十分な素養を備えていることを示すものと認める。

山下 達雄は、本博士論文において、形態素情報を付与したコーパスを作成するための環境について考察し、様々な解析システムと関連技術の提案と実装を行った。コーパスに基づく言語処理が大きな進展を遂げている現在、精度の高い解析済みコーパスを効率よく作成し、利用することは極めて重要であり、この分野への寄与が大きいと認める。本論文では、以下の研究成果が報告されている。

1. 特定の言語に依存しない形態素解析の手法を提案し、実装した。各言語の特徴を少ないパラメータで指定することにより、様々なタイプの言語への適用が可能であることを示すため、日本語、英語、中国語についての実装を行った。
2. タグ付きコーパスを効率的に蓄積するための支援環境を提案し、実装した。これにより、新しいデータに対して、システムの支援を受けて効率よいタグ付きコーパス作成が行えるとともに、作業者の不注意なエラーがコーパスに入り込む余地を削減した。
3. 解析済みのコーパスを実際の解析システムに活かすためのいくつかの方法を提案した。一つは、人手で作成されたルールとタグ付きコーパスから得られた統計量の融合に関する提案であり、もう一つは、解析済みコーパスを直接利用した事例ベースによる解析手法である。両者について実験によりその有効性を示している。
4. 大規模なタグ付きコーパスからの類似文字列検索を高速、高精度で行なう手法を、Suffix array を拡張することによって提案し、その有効性を確認した。また、これらのプログラムを利用しやすい形でインプリメントし、フリーソフトとして公開した。

このように、タグ付きコーパスに関して、作成、蓄積、利用のすべてに渡る技術を一連の流れとしてとらえてシステムを設計し、実装することにより、極めて効率よく自然言語処理技術の基盤であるタグ付きコーパスを整備することができることを示した。その意味で本研究の成果は顕著なものであると評価する。

よって、本論文は博士(工学)の学位論文として価値あるものと認める。