TSUBAME 共同利用　平成 23 年度　学術利用　成果報告書

利用課題名　会話エージェント構築のための大規模ウェブコーパスからの知識獲得
英文：Knowledge Acquisition from Large-scale Web Corpora for Constructing Conversational Agents

利用課題責任者　黒橋 禎夫
First name Surname　Sadao Kurohashi

所属　京都大学
Affiliation　Kyoto University
URL　http://www.kyoto-u.ac.jp/

邦文抄録
本利用課題では、人間とインタラクションを行える会話エージェントの実現を目的とし、その基盤となる会話コンテンツを作るために、大規模ウェブコーパスに対して言語解析を行い、解析結果から知識獲得を行う。これにより、人間が常識として持っている膨大な知識を自動的に獲得する。具体的な知識として格フレームに着目し、格フレームを用いて省略解析を行う。省略解析の応用例として、省略解析結果を用いて検索エンジンを構築する。省略解析は他の解析と比べて計算量が膨大であり、省略解析結果を利用した検索エンジンは開発されてこなかったが、本利用課題では、TSUBAME を利用することによってこれを可能とした。

英文抄録
In this project, we aim to construct conversational agents that have the ability to interact with people. This requires conversational contents that are to be used by these agents. In order to achieve this goal, we perform linguistic analysis to a large-scale Web corpus and use the resultant data to acquire wide-coverage linguistic knowledge. As the form of linguistic knowledge, we focus on case frames, and by using them we perform anaphora resolution. On top of anaphora resolution, we build a search engine. Such a search engine has not been built because anaphora resolution requires far more computational resources than conventional analyses, but we realized this using TSUBAME.

Keywords: natural language processing, Web, knowledge acquisition, case frames, anaphora resolution

背景と目的

計算機が人間と自然なインタラクションを行えるようにするには、人間が持つ常識的な知識が様々な局面で必要となる。しかし、こうした知識を人手で記述するのは難しく、「知識獲得のボトルネック」が長年の課題となってきた。近年、ウェブから膨大な量のテキストが得られるようになってきており、こうしたテキストからの自動獲得により、常識的な知識を得る研究が進みつつある。

これまで我々は、日本語ウェブページ数億件を言語解析し、その結果から知識獲得を行ってきた。得られた知識を用いることで、言語解析が向上することを示した。さらに、コーパスの規模と解析精度の関係を調査した結果、さらに規模の大きなコーパスを用いることにより、一層の精度向上が見込まれるとの予想を得た。そこで、数年前にクロールしたウェブページではなく、最近のテキストを用いることで、より大規模で、かつタイムリーな知識の獲得を行う。

本利用課題では、新しくクロールされた日本語ウェブページ 10 億件を解析し、その解析結果から言語知識の獲得を行う。こうした処理を、TSUBAME の大規模並列計算資源を利用して、極めて短期間で達成することを目的とする。

概要

本利用課題では、人間とインタラクションを行える会話エージェントの実現を目的とし、その基盤となる会話コンテンツを作るために、大規模ウェブコーパスに対して言語解析を行い、解析結果から知識獲得を行う。これにより、人間が常識として持っている膨大な知識を自動的に獲得する。人間が持つ知識は膨大であり、知識獲得のボトルネックが長年問題となってきたが、この問

題を解決するための知識源として大規模ウェブコーパスに着目している。TSUBAME を用いた大規模並列処理により、これまで実現不可能であった規模の知識を獲得する。

具体的な知識としては、格フレーム[1]に着目する。格フレームとは、用言とそれに関係する知識を集めたものであり、例えば「積む」という用言の格フレームのひとつとして次のようなものが考えられる。

{ 従業員, 運転手, ・・・} が{ 車, トラック,・・・} に{ 荷物, 物資}を積む

格フレームは述語項構造ともよばれる。述語（積む）といくつかの項（ガ格「従業員」、ニ格「荷物」）との関係を表している。

格フレームは、構文的曖昧性の解決を含む幅広いタスクで利用できる。例えば「弁当は食べて目的地に出発した。」という文において、「弁当は」の係り先としては「食べて」と「出発した。」が考えられる。このとき、「{人, 学生, ・・・} が{ 弁当, パン,・・・} を食べる}」という格フレームを知っていれば、「弁当は」の係り先は「食べて」らしいとわかる。同時に、「弁当は」の提題の「は」の背後には「ヲ格」が隠れていると解析できる。これを格解析とよぶ[2]。

格解析だけでは、文章中の意味関係を充分にとらえることはできない。例えば、「トヨタは 1997 年ハイブリッドカー、プリウスを発売。2000 年からは海外でも販売している。」という文章を考える。2 文目の「販売した」は表層的には「ガ格」も「ヲ格」も取らないが、「トヨタが」「プリウスを」「販売した」ことが文脈から推測できる。こうした省略要素をあてるタスクを省略解析とよぶ[3]。省略解析には、「販売した」の動作主は「トヨタ」などの組織だろうという常識的な知識が必要となるが、こうした知識は格フレームにより実現されている。

知識獲得の具体的な手順は以下の通りである。まず、我々の計算機・ネットワーク環境でクロールして得られたウェブページを TSUBAME に転送する。各ウェブページから日本語文を抽出し、テキストコーパスを作成する。従来の格フレーム構築では、重複文を除去してテキストコーパスを作成していたが、本利用課題ではこれを行わない。省略解析で文章内の関係を解析するため、元の文の並びを保持したまま知識獲得を行う。

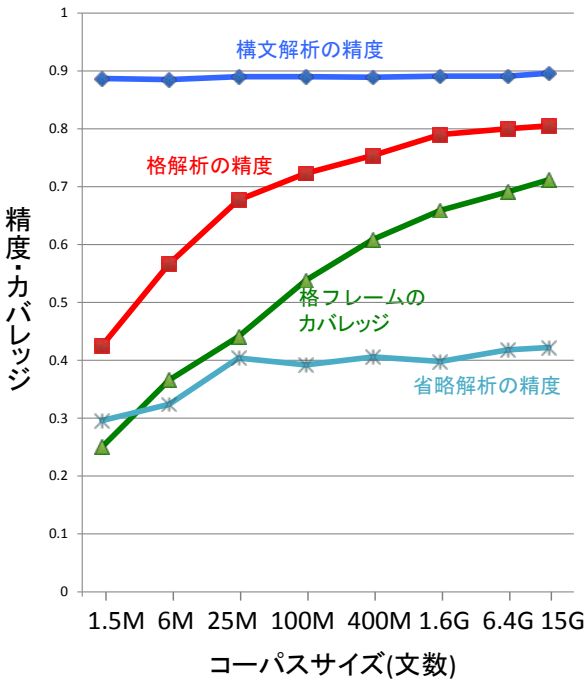

図 1 精度・カバレッジのコーパスサイズによる変化

次に、作成されたテキストコーパスに対して、既に構築済みの格フレームを用いて格解析・省略解析を行う。解析はページごとに独立で依存関係がないため、embarrassingly parallel に実行できる。

最後に、省略解析結果を用いて検索エンジンを構築する。各ウェブページの解析結果に対してインデクシングを適用する。

結果および考察

当初の研究計画では、大規模ウェブテキストからの語彙獲得、格フレーム構築、含意関係知識の構築も計画していたが、これらの処理は同じ研究グループによる別の TSUBAME の利用課題により順調に進んだため、本利用課題では省略解析を中心に取り組んだ。

まずは省略解析の精度と知識獲得に用いるコーパスのサイズとの関係[4]を調べた。同じ研究グループによる別の TSUBAME の利用課題によって得られたウェブコーパス 150 億文をベースとし、サンプリングによって、150 万文、600 万文、2500 万文、1 億文、4 億文、16 億文、64 億文のコーパスを作成した。それぞれのコーパスからの格フレーム構築とそれに基づく解析器の作成を行った。省略解析については、当初の予想を下回ったものの、コーパスのサイズが大きくなるにつれて、

図 2 格解析・省略解析が検索において有効な例

解析精度が向上することを確認した(図1)。

日本語ウェブページ約 1 億件に対して、省略解析を適用し、その解析結果を用いた検索エンジンを構築した。当初の研究計画では、約 10 億件のウェブページを対象としていたが、省略解析は他の解析と比べて膨大な計算量が必要となるため、約 1 億件に縮小して処理を行った。これまでも、言語解析結果を利用した検索エンジンは開発されてきたが、計算量の膨大さゆえに、省略解析のような意味解析は適用されてこなかった。本利用課題では、TSUBAME を利用することによってこれを可能とした。計算量としては、省略解析の適用とそれ解析結果を用いたインデクシングに対して約 70 万 CPU コア・時を要した。

省略解析結果を用いた検索エンジンについては、その定量的評価は今後の検討課題である。人手による調査により、図 2 のように、省略解析の効果が見られる例があることを確認している。

まとめ、今後の課題

本利用課題では、日本語ウェブページ 1 億件に対して省略解析を行い、その解析結果から検索エンジンを構築した。こうした処理を、TSUBAME の大規模並列計算資源を利用して、極めて短期間で達成することができた。

今後は検索エンジンの定量的評価を行いたい。また、解析精度については、コーパス規模の拡大によりさらなる改善が予想されることから、さらに規模を拡大して知識獲得に取り組みたい。

参考文献

[1] Daisuke Kawahara and Sadao Kurohashi. Acquiring Reliable Predicate-argument Structures from Raw Corpora for Case Frame Compilation, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pp.1389-1393, Valletta, Malta, 2010.

[2] Daisuke Kawahara and Sadao Kurohashi. A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis, Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pp.176-183, 2006.

[3] 笹野遼平，黒橋禎夫．大規模格フレームを用いた識別モデルに基づく日本語ゼロ照応解析，情報処理学会論文誌，Vol.52，No.12，pp.3328-3337, 2011.
[4] Ryohei Sasano, Daisuke Kawahara and Sadao Kurohashi. The Effect of Corpus Size on Case Frame Acquisition for Predicate-Argument Structure Analysis, IEICE TRANSACTIONS on Information and Systems, Vol.E93-D, No.6, pp.1361-1368, 2010.