TSUBAME 共同利用　平成 23 年度　学術利用　成果報告書

利用課題名　大規模ウェブコーパスからの知識獲得およびその応用
英文：Knowledge Acquisition from Large-scale Web Corpora and its Applications

利用課題責任者　　河原 大輔
First name Surname　Daisuke Kawahara

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、大規模日本語ウェブコーパスに対して言語解析を行い、その解析結果から高被覆な言語的知識を自動獲得する。得られた知識を用いることにより、言語解析の精度向上が図れるほか、テキスト間の含意関係の推論などの高度な応用処理の実現が期待できる。言語知識としては、具体的には、格フレームと含意関係知識に着目し、TSUBAME の大規模並列計算資源を利用してこれらを自動構築する。データサイズと解析精度との関係を調べたところ、コーパスのサイズが大きくなるにつれて、解析精度が向上することが確認できた。

英文抄録（100 words 程度）
In this project, we perform linguistic analysis to a large-scale Japanese Web corpus and use the resultant data to acquire wide-coverage linguistic knowledge. The knowledge acquired can be used to improve linguistic analysis and realizes advanced applications including the recognition of textual entailment. As the forms of linguistic knowledge, we focus on case frames and textual entailment. We are able to automatically construct them using TSUBAME's large-scale parallel computing resources. We investigated the relationship between the data size and the accuracy of linguistic analysis, and found that the accuracy of linguistic analysis did become higher as the data size increased.

Keywords: natural language processing, Web, knowledge acquisition, case frames, recognizing textual entailment

背景と目的

自然言語を解析し、また解析結果を利用した応用処理を実現するためには、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、膨大な量のテキストがウェブから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、日本語ページ数億件からなる大規模ウェブコーパスを言語解析し、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。また、応用面においても、人間の実利用者を想定すると、従来用いてきた数年前にクロールしたウェブページではなく、最近のテキストを解析し、応用処理で利用可能とする必要がある。

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、大規模日本語ウェブコーパスに対して、言語解析を行い、その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。得られた知識を用いることにより、言語解析の精度向上が図れるほか、テキスト間の含意関係の推論などの高度な応用処理の実現が期待できる。

常識的な知識としては、具体的には、格フレーム[1]

と含意関係知識[2]を構築する。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[3]。

含意関係知識として、具体的には事態間の関係を獲得する。例えば、「人が財布を拾って警察に届ける」という文から、「人が財布を拾う」という事態の次には「警察に届ける」という事態が発生だろうという知識が得られる。ここで、事態は述語項構造で表現される。すなわち、含意関係知識は、格フレームに基づいている。

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、重複を除去してテキストコーパスを作成する。このテキストコーパスに対して、形態素解析・構文解析を行う。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。

次に格フレーム構築を行う。まず、構文解析結果から曖昧性のない述語項構造を並列に抽出する。次に、抽出された述語項構造を用言ごとにクラスタリングする。こうして得られた格フレームを用いて再度構文解析を並列に行い、構文解析結果から新たな情報を格フレームに加える。

こうして得られた格フレームを用いてウェブコーパスに対して並列に構文・格解析を行う。得られた格解析結果から、述語項構造の共起関係を抽出することにより含意関係知識を構築する。

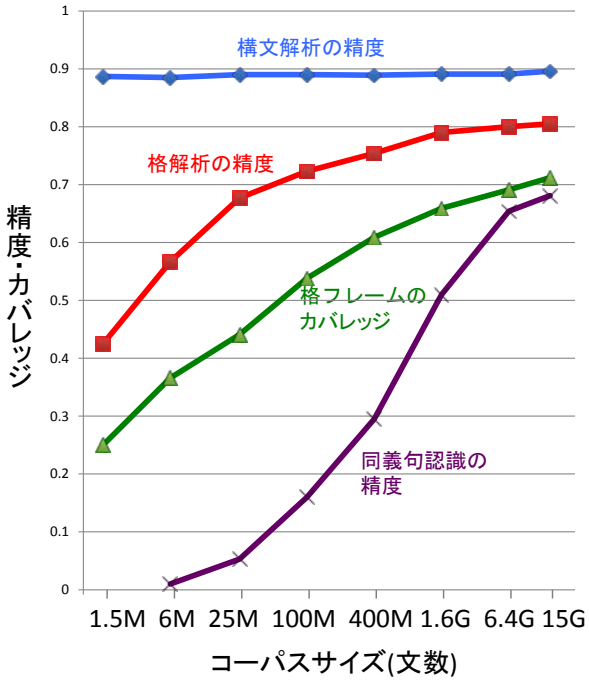

図 1 精度・カバレッジのコーパスサイズによる変化

結果および考察

日本語ウェブページとして、当初の計画よりも大幅に多い約 30 億件が利用可能となった。そのため、計画に含まれていたテキストからの語彙獲得は取りやめ、格フレームと含意関係知識の構築に取り組んだ。

約 30 億件のウェブページから日本語文を抽出、重複を除去し、日本語約 150 億文からなるテキストコーパスを作成した。このコーパスに対して構文解析を適用し、その解析結果から格フレームを構築した。この一連の処理を行うために、約 20 万 CPU コア・時の計算を要した。また、この結果得られた格フレームを解析器に組み込むことによって、より高精度の解析器が得られた。

さらに、コーパスのサイズと解析器の精度との関係[4]を調べるために、上記 150 億文コーパスからサンプリングすることによって、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い、コーパスのサイズが大きくなるにつれて、解析器の精度が向上することを確認

した(図1)。この処理に対して、約 10 万 CPU コア・時の計算を要した。

含意関係知識の構築は、ウェブページ約 1 億件を対象に行なった。約 300CPU コアを用いることにより約 20,000 事態ペアを約 3 日で獲得することができた。

まとめ、今後の課題

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 30 億件を解析し、その解析結果から言語知識の獲得を行った。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。

今後は得られた言語知識を含意関係認識タスクなど具体的な応用処理に用いて、その効果を検証したい。また、精度・知識のカバレッジについて、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] 柴田知秀, 黒橋禎夫. 述語項構造の共起情報と格フレームを用いた事態間知識の自動獲得, 情報処理学会 第 203 回自然言語処理研究会, 2011.

[3] 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.

[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.