TSUBAME 共同利用　平成 24 年度　学術利用　成果報告書

利用課題名　大規模 Web コーパスからの世界知識の獲得

英文：Acquisition of World Knowledge from Large-scale Web Corpora

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、大規模 Web コーパスに対して言語解析を適用し、その解析結果から膨大な言語知識を獲得する。次に、獲得した知識に基づく、より高度な解析を情報検索の標準的なテストセットに適用し、その有効性を評価する。その結果、獲得した知識に基づく解析を利用することによって、特に検索結果上位の適合率が有意に向上することを確認した。これらの言語解析および知識獲得の処理には膨大な計算量が必要になるが、TSUBAME の大規模並列計算資源を利用することによって、極めて短期間で達成することができた。

英文抄録

In this project, we perform linguistic analysis to a large-scale Japanese Web corpus and use the resulting analyses to acquire wide-coverage linguistic knowledge. The knowledge acquired can be used to improve linguistic analysis and realizes advanced applications including information retrieval. To confirm the effectiveness of the deep analysis based on the acquired knowledge, we apply it to the standard test set of information retrieval and evaluate it. The result showed that the precision of top-ranked search results was significantly improved. We accomplished these processes quite rapidly using TSUBAME.

Keywords: natural language processing, Web, automatic knowledge extraction, predicate-argument structures, information retrieval

背景と目的

2011 年に Watson と呼ばれる計算機システムがクイズ番組で人間のチャンピオンに勝利したことが象徴的に示すように、自然言語に基づく高度な計算機処理が次第に現実のものとなりつつある。こうした処理の実現には、人間がもっているような膨大な世界知識が必要となる。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、Web 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 Web ページ数億件に言語解析を適用し、その解析結果から知識獲得を行ってきた。その知識の一つは、格フレームと呼ばれるもので、述語とそれが関係をもつ語(項)を集めたものである[1]。たとえば、「積む」という述語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

さらに、獲得した格フレームを用いることで、構文解析のような基本的な言語解析の精度が向上することを示している[2]。また、このような解析は、情報検索、自動要約、自動翻訳などのさまざまな言語処理アプリケーションで利用している。

しかし、構文解析より高度な解析、たとえば省略・照応解析を含む述語項構造解析(「誰がいつどこで何をした」の同定処理)や談話構造解析(文間の関係を認識する処理)については、格フレームを用いるだけではいまだ精度が低く、実際のアプリケーションで利用することが難しい。これらの解析の精度を向上させるには、格フレーム以外に語句のカテゴリや名詞格フレームのような知識が必要と考えられる。

本利用課題では、大規模 Web コーパスを解析し、その解析結果からこれらの知識を獲得する。獲得した知

識を述語項構造解析に統合し、それを情報検索の標準的なテストセットに適用、その有効性を評価する。これらの処理を、TSUBAME 2.0 の大規模並列計算資源を利用して行い、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 Web コーパスを解析し、その解析結果から知識を獲得する。次に、獲得した知識を統合した述語項構造解析を情報検索の標準的なテストセットに適用し、その有効性を評価する。

述語項構造解析に必要な知識として、格フレーム以外に、名詞格フレームおよび語句のカテゴリがある。名詞格フレームとは、名詞に対して必須となる項を集めたものであり、たとえば「コーチ」に対する「スポーツ:{野球, ラグビー, ...}」のようなものである[3]。格フレームは述語に対するものであったのに対して、名詞格フレームは名詞に対する格フレームといえる。語句のカテゴリとは、たとえば「学生」に対するカテゴリ「人」、「リンゴ」に対するカテゴリ「植物」または「人工物−食べ物」のような分類である。

名詞格フレームは、これまで 1 億文程度の比較的小規模な Web コーパスからの自動構築については実現していた。今回は、70 億文の大規模 Web コーパスを利用して自動構築を行った。まず、この大規模 Web コーパスに対して構文解析を適用し、その解析結果から「X の Y」というパターンで出現する名詞ペアを抽出した。これを名詞 Y ごとに整理し、名詞 X の集合をクラスタリングすることによって名詞格フレームを構築した。

カテゴリは、形態素解析システム JUMAN[1]において 22 種類を設定し、基本語 3 万語について人手で付与している。これは基本語のみをカバーしており、「京都大学」のような句(複合名詞)はカバーしていない。述語項構造解析を高精度化するには、このカテゴリを複合名詞にも付与する必要がある。複合名詞のカテゴリは、基本的には、その主辞(末尾の名詞)のカテゴリとみなせばよい(「京都+大学」[2]のカテゴリは「大学」のカテゴリとみなす)が、末尾が漢字一文字の場合は誤ってしまうことが多い。たとえば、「受験+生」のカテゴリを「生」のカテゴリである「抽象物」とみなすと誤りである。

末尾が漢字一文字の複合名詞に対して、基本語との分布類似度を計算し、K 近傍法にてカテゴリ推定を行った。たとえば、「受験+生」は「学生」「生徒」「若者」などと分布類似度が高いのでカテゴリが「人」と推定され、「出張+先」は「島」「旅先」「出先」など分布類似度が高いのでカテゴリが「場所−その他」と推定される。末尾が漢字一文字である複合名詞 3 万語(Web コーパス中で頻度上位のもの)に対して、このようなカテゴリ推定処理を行った。

これらの獲得知識を述語項構造解析に統合し、これを NTCIR−4 WEB テストコレクション(NW−100G−01)[4]に適用した。このテストコレクションは、jp ドメインからクロールした Web ページ約 1100 万件からなる。これに述語項構造解析を適用し、その解析結果を用いた検索エンジンを構築した。ベースとなる検索エンジンとしては、開放型検索エンジン基盤 TSUBAKI[5]を用いた。これまでも、形態素解析などの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、述語項構造解析のような意味解析は莫大な計算量が必要となるために適用されてこなかった。本研究では、TSUBAME を利用することによってこれが可能となった。

上記の述語項構造解析の他に、高度な解析の一つに談話構造解析があり、自然言語の意味理解のためには、高精度な談話構造解析を実現する必要がある。談話構造解析に必須となるのは、大規模な談話構造の注釈付きコーパスであるが、これまでは存在しなかった。そこで、これを構築するために、注釈付けの対象となる文書集合を Web コーパスから抽出した。まず、各 Web ページの冒頭部分を抽出し、それが自然な日本語の文章になっているかどうかを自動判定した。さらに、ミラーページやスパムによって生み出された重複、類似した文書を除去するために、編集距離によるフィルタリングを行った。

結果および考察

大規模 Web コーパスに対する構文解析に約 3 万 CPU コア・時、名詞格フレームの自動構築に約 1 万 CPU コア・時、カテゴリ推定処理に約 1.5 万 CPU コア・

[1] http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN
[2] 複合名詞中の単語区切りを“+”で表す。

	MAP	P@3	P@10	nDCG@10
word	0.167	0.423	0.371	0.232
dep	0.170	0.423	0.373	0.231
P-A	0.173	0.442	0.379	0.237

表 1 検索エンジンの精度

時を要した。NTCIR-4 WEB テストコレクションに対する述語項構造解析の適用と検索エンジンのインデクシングには、約 8 万 CPU コア・時を要した。

検索エンジンの評価は、NTCIR-3 WEB と NTCIR-4 WEB の検索課題(計 127 件)を用いて行った。その結果を表 1 に示す。word は、形態素解析のみを用いる手法、dep は形態素解析と構文解析を用いる手法、P-A は形態素解析と述語項構造解析を用いる手法を意味する。検索結果は、4 つの指標を用いて評価した。MAP は、平均適合率、P@3 は検索結果上位 3 件の適合率、P@10 は検索結果上位 10 件の適合率、nDCG@10 は文書適合度を考慮した上位 10 件のスコアを表す。特に、P@3 指標で、述語項構造解析の効果が見られ、精度の高い検索エンジンを実現することができたと言える。

談話構造解析のための文書集合抽出については、Web ページ 1500 万件から 30 万文書を抽出した。この処理に約 1 万 CPU コア・時を要した。この結果得たコーパスは、現在、クラウドソーシングによって、談話構造の注釈付けを行っている最中である。

まとめ、今後の課題

本利用課題では、大規模 Web コーパスに対して構文解析を行い、その解析結果から名詞格フレームと語句のカテゴリという 2 種類の知識を獲得した。次に、獲得した知識を統合した述語項構造解析を情報検索の標準的なテストセットに適用し、その有効性を評価した。その結果、述語項構造解析を利用することによって、特に検索結果上位の適合率が向上することを確認した。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。

今後の課題として、談話構造注釈付きコーパスを完成させ、これを用いて高精度な談話構造解析を実現することが挙げられる。また、述語項構造解析と談話構造解析を統合することによって、双方の精度を向上させることを考えている。

参考文献

[1] 河原大輔，黒橋禎夫．格フレーム辞書の漸次的自動構築，自然言語処理，Vol.12，No.2，pp.109-131，2005.

[2] Daisuke Kawahara and Sadao Kurohashi. A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis, In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL2006), pp.176-183, 2006.

[3] Ryohei Sasano, Daisuke Kawahara and Sadao Kurohashi. Automatic Construction of Nominal Case Frames and its Application to Indirect Anaphora Resolution, In Proceedings of the 20th International Conference on Computational Linguistics (COLING2004), pp.1201-1207, 2004.

[4] Koji Eguchi, Keizo Oyama, Akiko Aizawa and Haruko Ishikawa. Overview of WEB Task at the Fourth NTCIR Workshop, Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, 2004.

[5] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto and Sadao Kurohashi. TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology, Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP-08), pp.189-196, 2008.