共同利用　成果報告書　平成 21 年度　課題種別

利用課題名　大規模 Web ページコレクションからの言語知識獲得
英文：Linguistic Knowledge Acquisition from Large-scale Web Page Collections

利用課題責任者　木俵 豊
First name Surname　Yutaka Kidawara

所属　独立行政法人 情報通信研究機構
Affiliation　National Institute of Information and Communications Technology
URL　http://www.nict.go.jp/

邦文抄録
本利用課題では、超大規模 Web ページコレクションに対して、言語解析および言語知識獲得処理を適用する。その結果、「アトピーの原因はカビである」「クラドスポリウムはカビの一種である」のような言語知識を自動的に獲得する。このような知識は、言語理解を必要とするさまざまなシステム、たとえば情報分析システム、対話システムなどの改良・精度向上に役立つことが期待できる。TSUBAME Grid Cluster を用いることによって、これまで現実的に計算不可能であった 10 億ページ規模の Web ページコレクションを処理することによって、大規模な語彙に関する言語知識を自動獲得した。

英文抄録
This project applies linguistic analysis and linguistic knowledge acquisition to a large-scale Web page collection. Our objective is to automatically acquire wide-coverage linguistic knowledge to contribute to the improvement of natural language processing systems, such as information analysis systems and dialogue systems. By using the TSUBAME Grid Cluster, we realized the linguistic analysis and linguistic knowledge acquisition on the basis of 1 billion Web pages, which has never been realistic.

Keywords: Web, Linguistic knowledge, Linguistic analysis, Text mining

背景と目的

近年、Web 上の情報は爆発的に増加し続けており、これまでに手に入らなかった規模のテキスト情報が手に入るようになった。その大規模テキスト情報の中から、人間がもっているような常識的な言語知識を自動的に獲得し、それに基づく人工知能、自然言語理解の研究が進みつつある。

これまで我々は、1 億件の日本語 Web ページを収集し、それに対してまず言語解析を適用し、その解析結果から知識獲得を行ってきた。そこから、「アトピーの原因はカビである」「クラドスポリウムはカビの一種である」「防カビ剤の類義語は除菌剤である」というような言語知識が獲得できている。

しかし、1 億件の Web ページでは規模的にまだ小さいため、100 万語程度の語彙に関する知識が獲得できてはいるが、カバレッジはそれほど高くない。例えば、上記規模程度の Web ページに 2 回以上現れている名詞を集めると、新聞中の固有名詞の 80%がカバーできるという調査結果があるが、出現回数 2 回では十分信頼性のある統計量が得られないことは言うまでもない。結果として、例えば、「原因−結果」タイプの自動知識獲得では、80%の精度で獲得しようとすると、獲得数が 3 万程度に減少すると推定されている。

本利用課題では、よりカバレッジの高い言語知識を獲得するために、10 億件規模の日本語 Web ページを解析し、その解析結果から言語知識の獲得を行う。我々の保有する計算機資源では、10 億件規模の Web ページを扱うためには 1 年という期間がかかると推定され現実的ではない。そのため、TSUBAME Grid Cluster の大規模並列計算資源を利用し、これを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、超大規模 Web ページコレクションに対して、言語解析および言語知識獲得処理を適用する。その結果、「アトピーの原因はカビである」「クラドス

ポリウムはカビの一種である」のような言語知識を自動的に獲得する。このような知識は、言語理解を必要とするさまざまなシステム、たとえば情報分析システム、対話システムなどの改良・精度向上に役立つことが期待できる。TSUBAME Grid Cluster を用いることによって、これまで現実的に計算不可能であった規模の Web ページコレクションを処理し、上記の言語知識を自動獲得する。

手順としては、まず、我々の計算機・ネットワーク環境で Web をクロールすることによって得られた 10 億ページ規模の Web ページコレクションを TSUBAME に転送する。この Web ページコレクションに対して TSUBAME を用いて、文抽出、形態素・構文解析を並列に行う。その結果は、XML 形式(標準フォーマットと呼ぶ)に格納し、圧縮する。さらに、この結果を知識獲得器の出発点にし、知識獲得を行う。また、言語解析結果は、今回の知識獲得以外にもさまざまな用途に用いるため、我々の所属機関への転送を行う。

文抽出には、検索エンジン基盤 TSUBAKI[5][1]で開発され、公開されているツールを用いる。形態素解析には JUMAN[2]、構文解析には KNP[3]を用いる。これらのツールは一般公開されている。

言語知識獲得は、関係知識獲得、および語間の類似度計算の 2 つからなる。関係知識獲得のためには、解析結果から「“*アトピー*”、“*カビ*”、“*A が B の原因となる*”」等という、*N* 文節の近接ウィンドウ内で共起する名詞対とその共起コンテキストからなる三つ組を抽出し、共起頻度、相互情報量など様々な統計量を計算し、データベース化する[3]。このように大規模な資源を用いることで獲得対象となりうる名詞が大幅に増加し、多様な共起コンテキストを手がかりにした「ウェブ・スケールの関係獲得」が実現可能となる。知識獲得器の流れとしては、まずユーザが獲得しようとする意味的関係に代表的な言語表現(“シードパターン”と呼ぶ)を入力する。これらのシードパターンは上記の「名詞対の共起コンテキスト」に相当するものである。システムがシードパターンの言い換えを大量に学習し、学習したパターンと共起する名詞対を獲得結果として出力する(**図 1**)。

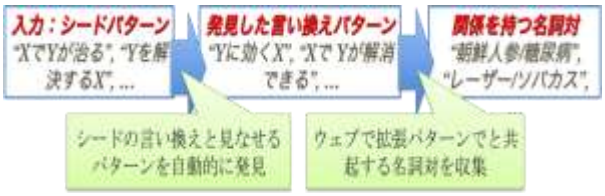

図 1 関係知識獲得の手順

語間の類似度計算には、言語解析結果に対して大規模単語クラスタリングを適用する[1][2]。これによって、1000 万語規模の単語間類似度を計算することを目標とする。

結果および考察

10 億ページ規模の Web ページコレクションから 6.3 億件の日本語 Web ページを抽出した。10 億ページに満たなかったのは、アダルトページフィルタや非日本語ページフィルタによって排除されたページがあったからである。得られた日本語 Web ページ集合の容量は、圧縮して約 4TB であった。

この 6.3 億ページに対して言語解析を適用することに成功した。これには、平均的に 512CPU コアを用い、約 6 週間で行うことができた。得られた標準フォーマットの容量は、圧縮して約 20TB であった。

次に、言語解析結果を我々の所属研究機関に転送した。当初予定していたネットワークによる転送に速度面での困難が生じたため、HDD を直接 TSUBAME サイトに持ち込んでデータをコピーすることとした。USB 2.0 の容量 4TB の HDD を 6 台とノート PC3 台を用意し、各ノート PC に 2 台の HDD を接続した。各ノート PC は 1GbE にてハブに接続し、そのハブが TSUBAME へと接続された。この状態で、各 PC 上で 2 並列で scp を実行した(つまり、計 6 並列)。各々のコピーが 15MB/s 程度であったため、計 90MB/s 程度の転送を行えたことになる。ネットワーク越しの場合、10MB/s 程度であったので、十分意味のある試みであった。このようにして、約 4 日かけてデータを転送し、HDD を輸送して受け取り側で逆のことを行い、我々の計算機環境にデータを無事配置することができた。

次に、6.3 億ページの言語解析結果を処理し、概要で述べたような関係知識獲得器を作成した。これにより、

[1] http://tsubaki.ixnlp.nii.ac.jp/
[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
[3] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

約 400 億個の 3 つ組の関係を抽出した。この関係知識獲得器を用いて獲得した知識の質は、これまでの 1 億ページのデータと比べて向上しており、予備的な評価結果から、例えば、「原因-結果」では、80％以上の精度 5 万組の関係を発見することができていると考えられ、精度が向上したことを意味する。今後、さらに詳細な分析を行いたいと考えている。

言語知識獲得における計画の一つであった 1000 万語規模の単語クラスタリングについては、残念ながら TSUBAME 環境で完了させることができなかった。これは、提案者環境において OpenMPI（バージョン 1.4.1）で動作していたプログラムが TSUBAME にインストールされている OpenMPI では正常に動作しなかったためである。不具合の症状としては、マスターノードにデータを転送・集約し、ファイルへと書き出すルーチンにおいてブロックが起きるというものであった。TSUBAME 担当の方のアドバイスに従い、こまめに flush を挿入することや、一度に転送するデータ量を制限するなどプログラムの修正を試みたが、TSUBAME 利用期限までに解決することができなかった。

単語クラスタリングの目的は、その結果を使用して単語の類似度を計算することであった。提案者は、クラスタリングを用いない類似度計算法をベースラインとして使用している。上述のように、単語クラスタリングが実行できなかったため、今回の TSUBAME 利用により得られた 6.3 億ページの言語解析結果を用いてベースライン手法で類似度を計算することを行った。これには、6.3 億ページ分の言語解析結果から約 33 億種類の係り受け関係（例：｛みかん，を食べる｝）を抽出して利用した。まず、従来の 100 万語の語彙について、1 億ページからのデータと、今回利用可能となった 6.3 億ページからのデータによってどの程度類似度の精度の差が出るかを調査した。その結果、我々の使用している評価データについて、例えば MTP@40 という指標が 0.0639 から 0.0688 へと改善することが分かった。これは、元となるデータ量を大幅に増大させることで、得られる言語知識の質の改善が実現できたということである。次に、このベースライン法で 400 万語の類似度データを生成した。結果の詳細な分析が今後の課題である。

まとめ、今後の課題

本利用課題では、TSUBAME Grid Cluster を用いることによって、これまで現実的に扱うことが難しかった 10 億ページ規模の Web ページコレクションに対して言語解析と言語知識獲得を適用した。

本利用課題で得られた言語知識は、言語解析技術の精度向上に役立つだけではなく、さまざまなアプリケーション、たとえば情報分析技術、機械翻訳技術、対話システム研究において効果的に利用することができる。これらのシステムにおける利点の他に、人間の発想支援などにも用いることができるという利点がある。

知識獲得は当初目標の規模では実行できなかったという問題があった。TSUBAME において、小規模（数ノード）でもよいので予約なしで排他的に並列プログラムのデバッグに使用できる環境があれば、より効率的にデバッグができたのではないかと思われる。

参考文献

[1] 風間淳一，Stijn De Saeger，鳥澤健太郎，村田真樹．係り受けの確率的クラスタリングを用いた大規模類似語リストの作成．言語処理学会第 15 回年次大会，pp. 84-87，2009.

[2] Jun'ichi Kazama and Kentaro Torisawa. Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations. In Proc. of ACL-08: HLT (full poster paper), pp.407-415, 2008.

[3] Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata. Large Scale Relation Acquisition Using Class Dependent Patterns. In Proc. of the 9th ICDM, pp. 764-769, 2009.

[4] Stijn De Saeger，鳥澤健太郎，風間淳一，黒田航，村田真樹．単語の意味クラスを用いたパターン学習による大規模な意味的関係獲得．言語処理学会第 16 回年次大会，pp.932-935，2010.

[5] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto and Sadao Kurohashi. TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology, In Proc. of IJCNLP-08, pp.189-196, 2008.