科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１６　日現在

機関番号：　１４３０１
研究種目：　若手研究(A)
研究期間：　2011～2013
課題番号：　２３６８００１５

研究課題名（和文）言語使用の大規模観察に基づく言語知識獲得と言語解析の共深化

研究課題名（英文）Co-deepening Knowledge Acquisition and Linguistic Analysis based on Large-scale Observation of Language Uses

研究代表者
河原　大輔（Kawahara, Daisuke）

京都大学・情報学研究科・准教授

研究者番号：１０４５０６９４

交付決定額（研究期間全体）：（直接経費）　　14,200,000円、（間接経費）　　4,260,000円

研究成果の概要（和文）：本研究では、まず、大規模Webテキスト集合から言語知識を自動獲得した。主な言語知識は、述語とそれがとる項(名詞)を記述した格フレームと呼ばれるものである。次に、獲得した言語知識を言語解析に組み込むことによって言語解析システムの精度向上を達成した。また、この高いカバレッジをもつ言語知識に基づく言語解析器を情報検索に応用し、これまでよりも精度が高い情報検索システムを開発した。

研究成果の概要（英文）：First, we automatically acquired linguistic knowledge from a large-scale web corpus. The acquired linguistic knowledge mainly consists of case frames, which describe the relations between predicates and their arguments. Then, we integrated such linguistic knowledge into a linguistic analyzer to improve the performance of the analyzer. We also developed an information retrieval system based on the knowledge-rich linguistic analyzer, and confirmed that our information retrieval system outperformed conventional information retrieval systems based on bag of words and dependency relations.

研究分野：　自然言語処理

科研費の分科・細目：　情報学・知能情報学

キーワード：　自然言語処理　Web情報処理　言語理解　知識獲得　人工知能　情報検索　言語解析　情報分析

１．研究開始当初の背景

近年の情報通信技術の社会への浸透により、言語の壁や情報格差などが大きな問題となりつつあり、情報検索、機械翻訳、自動要約などの言語処理システムを現実の場面で利用する必要性が増している。しかし、現状では、それらの精度は十分に高いとは言えず、利用者に負担を要求してしまう。たとえば、情報検索技術はサーチエンジンで用いられているが、Yahoo!や Google などの既存サーチエンジンは単語列を基本とした表層的処理を行っており、テキストの多様性、曖昧性をうまく扱うことができない。また、異表記同義語や同表記多義語などに関する知識が十分にないために検索漏れや検索誤りの問題が頻発してしまう。

このように、現状の多くの言語処理システムは、(1) 単語列を基本とする表層的な言語解析に基づいていること、(2) 人間にとっては常識的な言語知識が根本的に不足していることがボトルネックとなっている。

２．研究の目的

本研究では、超大規模 Web テキスト集合から言語知識を自動獲得し、その言語知識に基づく言語解析技術を研究開発する。これによって、テキスト全体をネットワーク構造化できるようになり、言語処理システムの高度化につながることが期待できる。たとえば、情報検索では、自然言語文をクエリとして入力したときに、入力文と対象文書それぞれをネットワーク構造化したものを比較することによって精緻なマッチングが可能になり、高精度な検索が実現できる。

３．研究の方法

(1) ネットワーク構造解析器のプロトタイプを大規模テキスト集合に適用し、そこから言語知識を獲得する。獲得した言語知識を利用して、ネットワーク構造解析器の精度向上を達成する。改良したネットワーク構造解析器をもう一度、大規模テキスト集合に適用し、さらに精度・カバレッジの高い言語知識を獲得する。このように、言語知識の獲得と、それに基づくネットワーク構造解析器の開発を交互に繰り返すことによって、精度の高い言語知識とネットワーク構造解析器を構築する。

(2) ネットワーク構造解析に基づく情報検索システムを開発する。これは、網羅性の高い言語知識に基づいてテキストをネットワーク構造化することによって、検索漏れや不適切な検索結果が少ないシステムとなる。

４．研究成果

まず、超大規模 Web テキスト集合に対して既存の構文解析器を適用し、その結果から格フレームと呼ばれる言語知識を自動獲得した。格フレームは、「誰がいつどこで何をした」のような 5W1H を表現する述語項構造を集約した辞書である。Web テキスト集合としては、Web ページ約 30 億ページから抽出し、重複を除いた約 150 億文を用いた。その結果、約 14 万述語について、1 述語当たり平均 6.7 個の格フレームを獲得した。

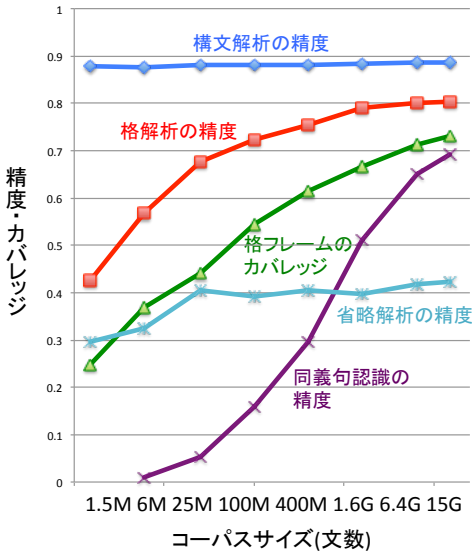

図 1 大規模テキスト集合からの言語獲得による言語解析の高度化

次に、獲得した格フレームを構文解析器に統合し、高カバレッジな言語知識に基づくネットワーク構造解析器を開発した。Web 文書からなる構文・格解析の評価セットで評価実験をしたところ、有意に精度が向上することを確認した。また、言語知識獲得に利用するテキスト集合の量に対して、解析精度がどれくらい向上するかを調査した。その結果を図 1 に示す。このように、より大規模なテキスト集合を用いて言語知識を獲得することによって、カバレッジおよび精度が高くなることがわかる。開発した日本語ネットワーク構造解析器は、KNP4.1 として一般に公開している。

上記で開発したネットワーク構造解析器に基づく情報検索システムを構築した。これは、検索エンジン基盤 TSUBAKI 上で実装し、多言語に対応している。NTCIR 日本語 Web 検索評価セット(1,100 万 Web ページ)を用いて評価実験を行い、単語や係り受け関係を用いた検索と比べて、検索結果の質が有意に向上することを確認した。

このように、係り受け関係より深いテキスト中の関係を捉えることができるネットワーク構造解析は、さまざまな言語処理アプリケーションの基盤技術として今後の利用が期待される。また、日本語の枠組みを利用して、英語、中国語、韓国語など多言語への適応も試みた結果、一部の言語ではすでに state-of-the-art の性能を得ており、今後順

次、多言語の解析システムを公開していく予定である。

５．主な発表論文等

〔雑誌論文〕（計１件）

1. Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara and Sadao Kurohashi. TSUBAKI: An Open Search Engine Infrastructure for Developing Information Access Methodology, Journal of Information Processing, Vol.20, No.1, pp.216-227, 2012, 査読有.
http://www.jstage.jst.go.jp/article/ipsjjip/20/1/216/_pdf

〔学会発表〕（計 13 件）

1. Jungyeul Park, Daisuke Kawahara, Sadao Kurohashi and Key-Sun Choi. Towards Fully Lexicalized Dependency Parsing for Korean, In Proceedings of the 13th International Conference on Parsing Technologies (IWPT2013), short, pp.120-126, Nara, Japan, 2013.11.28.
2. Ryohei Sasano, Daisuke Kawahara, Sadao Kurohashi and Manabu Okumura. Automatic Knowledge Acquisition for Case Alternation between the Passive and Active Voices in Japanese, In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP2013), pp.1213-1223, Seattle, USA, 2013.10.19.
3. Daisuke Kawahara, Keiji Shinzato, Tomohide Shibata and Sadao Kurohashi. Precise Information Retrieval Exploiting Predicate-Argument Structures, In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP2013), pp.37-45, Nagoya, Japan, 2013.10.15.
4. Mo Shen, Daisuke Kawahara, and Sadao Kurohashi. Chinese Word Segmentation by Mining Maximized Substrings, In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP2013), pp.171-179, Nagoya, Japan, 2013.10.15
5. Gongye Jin, Daisuke Kawahara and Sadao Kurohashi. High Quality Dependency Selection from Automatic Parses, In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP2013), poster, pp.947-951, Nagoya, Japan, 2013.10.15.
6. Daisuke Kawahara. Deep Natural Language Processing for Improving a Search Engine using Cloud Computing, Microsoft Research Asia Faculty Summit 2012, Tianjin, China, 2012.10.27.
7. 河原大輔. 構造的言語処理に基づく検索エンジン基盤 TSUBAKI の構築, 第 4 回データ工学と情報マネジメントに関するフォーラム(DEIM2012), 招待ポスター, 神戸, 2012.3.4.
8. 河原大輔. 情報の信頼性判断を支援する言語処理技術, 日本語用論学会年次大会 特別シンポジウム, 京都, 2011.12.3.
9. Daisuke Kawahara and Sadao Kurohashi. Generative Modeling of Coordination by Factoring Parallelism and Selectional Preferences, In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP2011), pp.456-464, Chiang Mai, Thailand, 2011.11.10.
10. Daisuke Kawahara. Deep Natural Language Processing for Improving a Search Engine Infrastructure using Windows Azure, Cloud Futures Workshop 2011, Redmond, USA, 2011.6.3.

〔図書〕（計１件）

1. 河原大輔, 黒橋禎夫. 日本語実テキストの基盤解析技術, ビッグデータ・マネジメント ～データサイエンティストのためのデータ利活用技術と事例～, pp.109-120, NTS 出版, 2014.

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
日本語構文・格・照応解析システム KNP:
http://nlp.ist.i.kyoto-u.ac.jp/?KNP

６．研究組織

(1)研究代表者

河原 大輔（KAWAHARA, Daisuke）
京都大学・大学院情報学研究科・准教授
研究者番号：10450694

(2)研究分担者

なし

(3)連携研究者

なし