様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年 6 月 5 日現在

機関番号：14301
研究種目：若手研究(B)
研究期間：2009～2012
課題番号：21700163
研究課題名（和文）大規模テキストから自動獲得した知識に基づく言語解析の精度向上
研究課題名（英文）Improvement of Linguistic Analysis Based on Automatically Acquired Knowledge from Large Text
研究代表者
柴田 知秀（SHIBATA TOMOHIDE）
京都大学・大学院情報学研究科・助教
研究者番号：70452315

研究成果の概要（和文）：
大規模テキストから言語知識を自動獲得し、言語解析の精度向上に利用する研究を行った。大規模 Web テキストや Wikipedia から、語と語の類似度計算、事態(イベント)間の関係、同義語・上位語に関する大規模語彙知識などの言語知識を自動獲得し、獲得した言語知識を統一的・整合的に管理した。また、獲得した言語知識をテキスト含意認識システムや固有表現解析システムの精度向上に利用した。
研究成果の概要（英文）：
This research aims to acquire linguistic knowledge from a large text, and utilize it for improving the accuracy of linguistic analysis. The linguistic knowledge acquired from a large Web text and Wikipedia includes the similarity between words, knowledge between events, and a large-scale lexical knowledge such as synonym and hypernym, and it is managed in integrated and consistent manner. Furthermore, the acquired linguistic knowledge was utilized for improving the accuracy of linguistic analysis such as textual entailment recognition and named entity recognition.
交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	900,000	270,000	1,170,000
2010 年度	800,000	240,000	1,040,000
2011 年度	800,000	240,000	1,040,000
2012 年度	800,000	240,000	1,040,000
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
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，大規模テキスト，知識獲得，同義語，分布類似度

１．研究開始当初の背景

10 年ほど前は、言語解析の精度が十分でなかったため、知識獲得と言語解析のデッドロックが生じていた。つまり、知識が少ないため、言語解析を高精度に行なうことができず、また、言語解析の精度が低いため十分な知識を自動獲得することができなかった。しかし、近年の計算機性能の向上と大規模テキストの利用により、自然言語処理の基本的な解析の精度が向上してきている。これに伴い、言語解析に必要な知識を獲得することが徐々に行なえるようになってきた。
新聞テキストに対する基本的な解析は十分な精度をあげられるようになったが、近年のウェブの爆発的普及により、自然言語処理の対象は新聞からウェブテキストに移行してきており、ウェブテキストには人手で整備した辞書には登録されていない未知語・新語な

どが多数存在するため、新聞テキストを対象とした場合に比べて精度が低下する。ウェブテキストを頑健に解析するためには語に関する膨大な知識が必要となる。
近年、大規模コーパスからの知識獲得がさかんに研究されている。例えば、Lin は類似した文脈で現れる語同志は意味も類似しているという考えを用いて、シソーラスを自動構築している[Lin1998]。また、Snow らは同義・上位下位関係を大規模コーパスから獲得し、人手で整備されたシソーラス(Wordnet)のノードに未知語を自動登録している[Snow et al. 2006]。しかし、これらの知識は言語解析に利用されているわけではない。
本研究では、獲得する知識をより大規模・構造化したものとし、様々な知識を統合的に利用することによって、言語解析の精度を向上させる。また、精度向上した解析器を用いて更なる知識獲得を行なう。

２．研究の目的
(1)日本語数億ページや Wikipedia から言語知識を自動獲得する。獲得する言語知識は語と語の類似度計算、事態間の関係、同義語・上位語に関する大規模語彙知識などである。

(2)上記で獲得した言語知識を利用し、言語解析の精度を向上させる。具体的には構文解析や固有表現解析、テキスト含意認識の精度を向上させる。

(3) 高精度化した言語解析器を用いて再度大規模テキストを解析し、更なる知識獲得を行なう。そしてまた、言語解析の精度を向上させる、といったサイクルを回すことにより、言語解析、知識獲得ともに精度を向上させる。その過程において、言語解析結果ならびに知識獲得結果を検討し、検討結果をお互いにフィードバックさせる。

３．研究の方法
(1)Web テキストから教師なしで言語知識を獲得するには共起関係を利用する。例えば、語と語の類似度は分布類似度という考えのもとに計算される。分布類似度とは、類似した文脈で現れる語は意味も類似しているという考えに基づいて計算される語の類似度である。例えば、「医者」と「医師」という語はどちらも「～に診てもらう」、「～を開業する」といった文脈でよく出現するので、これらの 2 語は類似していると考えることができる。また事態間の関係については、例えば、「財布を拾う」と「交番に届ける」の 2 つの事態は「財布を拾って交番に届ける」などの文においてよく共起することから、これらは関連が強い事態であることを獲得することができる。

(2)Wikipedia は幅広いドメインをカバーした百科事典であり、ここから同義語・上位語に関する知識を獲得する。例えば、「京都大学」と「京大」が同義語で、「エアドゥ」の上位語が「航空会社」などの知識を獲得することができる。

(3)(1)(2)で獲得した知識を統一的・整合的に管理し、形態素解析器・構文解析器で利用可能なようにする。そして、構文解析や固有表現解析、テキスト含意認識などで語彙知識を利用することにより精度を向上させる。

４．研究成果
(1) 大規模テキストを用いた分布類似度計算
各名詞に対して共起する動詞を大規模コーパスから抽出し、例えば「医者」と「医師」がどちらも「～が診察する」「～に診てもらう」などといった動詞と共起することからこれらの 2 語は類似しているといった分布類似度を計算した。また、同様に、各動詞に対して共起する名詞を抽出し、「購入する」と「買う」の分布類似度を計算した。評価セットを用いて、コーパスサイズを大きくすればするほど精度が向上することを確認した。
また、「(景気が) 冷え込む」と「(景気が) 悪化する」のように、述語単体では同義でないが文脈に依存して同義関係となる述語ペアを自動獲得する手法を提案した。格要素と述語を組とした単位に対して、係り受け関係にある述語を要素とした素性ベクトルを構築し、分布類似度を計算することによって類似度の高いペアを同義表現として獲得する。自動生成した評価セットによる実験と人手による評価実験を行なったところ提案手法の有効性を示すことができた。また、コーパスから獲得した述語の同義関係を検索エンジン TSUBAKI に導入することにより、クエリと文書の柔軟マッチングを実現した。

(2)事態間関係知識の自動獲得
「X{人}が Y{財布}を 拾う => X{人}が Y{財布}を Z{警察}に 届ける」のようなよく共起する 2 つの事態（イベント）を大規模テキストから自動獲得した。まず、大規模テキストから係り受け関係にある述語項構造ペアを抽出し、Apriori アルゴリズムにより述語項構造の共起度を効率よく計算した。次に、共起度が高い述語項構造に対して、格フレームを用いることにより、項の対応付けをとった。大規模テキストから約 2 万個の事態ペアを獲得することができた。

(3)Wikipedia からの大規模語彙の自動獲得
Wikipedia の記事から約 80 万語を獲得し、ま

た、語の同義語・上位語などの情報も合わせて獲得した。

(4)語彙知識の統一的・整合的管理のデザイン
形態素解析用辞書、国語辞典、シソーラス、格フレーム、Wikipedia など、様々な語彙知識を統一的に語彙データベースとして管理する枠組みを構築した。日本語の処理で問題となる単語区切りについても、語彙データベース内で管理し、また、各エントリに代表表記を付与することにより、表記の揺れ(例:コンピュータグラフィックス = コンピューターグラフィクスなど)を解消することができる。そして、形態素解析と句認識の結果に語彙知識を埋め込む枠組みをデザインし、構文解析や省略解析などの様々な言語解析時に参照することや、検索・翻訳などのアプリケーションでの利用を可能にした。

(5)述語項構造に基づくテキスト含意認識システムの構築
テキストと仮説を述語項構造単位で扱い、テキストと仮説間のマッチングをとる含意関係認識システムを構築し、NTCIR-9 の RITE タスクに参加した。テキスト・仮説の構文構造を解析し、述語項構造の集合として表現し、また、テキスト・仮説間のマッチングには国語辞典や Wikipedia、Web コーパスから得られた語句の同義や上位下位関係を利用した。BC(二値分類)、MC(多値分類)、EXAM(大学入試タスク)、RITE4QA(質問応答タスク)に参加し、それぞれ 0.55, 0.48, 0.66, 0.89 の精度を達成した。

(6)日本語固有表現解析の精度向上
任意の名詞句に対する固有表現の解釈と、ボトムアップに最適な固有表現の解釈を行う 2 段階の機械学習を用いる固有表現解析器を構築した。素性には固有表現解析で標準的に用いられるものに加え、Wikipedia から獲得された語彙知識を利用した。日本語固有表現の評価として広く用いられている CRL コーパスを用いて実験を行ったところ、既存の研究を上回る精度を達成することができた。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 6 件）
①Tomohide Shibata and Sadao Kurohashi: Predicate-argument Structure based Textual Entailment Recognition System exploiting Wide-coverage Lexical Knowledge
Special Issue of ACM TALIP on RITE (Recognizing Inference in TExt), 査読有, Vol.11, No.4, pp.16:1-16:23 (2012.12).
②柴田知秀, 姜ナウン, 黒橋禎夫:
同一文抽出に基づく類似ページの検出と分類
人工知能学会論文誌, 査読有, Vol.25, No.1, pp.224-232 (2010.1).
③船山弘孝, 柴田知秀, 黒橋禎夫:
二段階の機械学習を用いたボトムアップ型の固有表現認識
第 8 回情報科学技術フォーラム (FIT2009), 第 2 分冊, 査読有, pp. 19-26 (2009. 9).

〔学会発表〕（計 17 件）
①黒橋禎夫, 進義治, 柴田知秀, 村脇有吾, 河原大輔:
日本語語彙知識の統一的・整合的管理のデザイン
言語処理学会 第 19 回年次大会, pp.26-29 (2013.3).
②柴田知秀, 村脇有吾, 黒橋禎夫, 河原大輔:
実テキスト解析をささえる語彙知識の自動獲得
言語処理学会 第 18 回年次大会, pp.81-84 (2012.3).
③Tomohide Shibata and Sadao Kurohashi:
Predicate-argument Structure based Textual Entailment Recognition System of KYOTO Team for NTCIR9 RITE
In Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9), Tokyo, Japan (2011.12).
④Tomohide Shibata and Sadao Kurohashi:
Acquiring Strongly-related Events using Predicate-argument Co-occurring Statistics and Case Frames
In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP2011, poster), 査読有, Chiang Mai, Thailand (2011.11).
⑤柴田知秀, 黒橋禎夫:
文脈に依存した述語の同義関係獲得
情報処理学会 自然言語処理研究会 2010-NL-199 (2010.11).
⑥Hirotaka Funayama, Tomohide Shibata, and Sadao Kurohashi:
Bottom-up Named Entity Recognition using a Two-stage Machine Learning Method
In Proceedings of Association for Computational Linguistics/International Joint Conference on Natural Language Processing(ACL/IJCNLP2009): Workshop on Multiword Expressions, 査読有, pp. 55-62, Singapore (2009. 8).

〔図書〕（計０件）

〔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〕
ホームページ等
http://nlp.ist.i.kyoto-u.ac.jp/member/shibata/index-j.html
http://nlp.ist.i.kyoto-u.ac.jp/member/shibata/index.html

６．研究組織
(1)研究代表者
柴田 知秀（SHIBATA TOMOHIDE）
京都大学・大学院情報学研究科・助教
研究者番号：70452315

(2)研究分担者
なし

(3)連携研究者
なし