（ 続紙 １ ）

京都大学	博士（情報学）	氏名	橋本　力
論文題目	Knowledge Acquisition from the Web for Text Understanding （テキスト理解のためのＷｅｂからの知識獲得）		

（論文内容の要旨）

本論文は，計算機によるテキスト理解にとって重要な知識を獲得する手法について論じたものであり，内容語に対するドメイン知識，動詞間含意知識，フレーズ間言い換え知識の３種類の知識を対象に，Ｗｅｂから知識を自動獲得する手法を提案している．さらに，実際に獲得した知識を人手でチェックすることで，研究コミュニティ等で再利用可能な言語資源を構築し，広く配布している．つまり，高精度な知識獲得技術の開発と，大規模な言語資源の構築／公開が本研究の貢献である．論文は，３種類の知識の獲得手法についてそれぞれ１章を割り当て，序論と結論の章をあわせて，全５章で構成されている．

第１章は序論であり，テキスト理解研究の歴史を振り返った上で，テキスト含意認識と呼ばれる近年活発に研究されるようになったタスクが，テキスト理解をプログラムとして実装する上での見通しの良さ，テキスト理解に必要な知識の判定のしやすさの点で，テキスト理解という漠然とした課題の定式化として優れていることを論じている．それを受けて，テキスト含意認識という問題を解く上でどういう知識が必要なのか，それらのうち獲得技術が発展途上にあるものはどれかを考察し，本研究で対象とする３種類の知識の獲得の重要性を指摘している．また，Ｗｅｂが新語を多数含む，多種多様なドメインの文書から成る世界最大のテキストデータベースであるため，Ｗｅｂからの自動知識獲得が，計算機によるテキスト理解実現における難問の一つである知識獲得ボトルネックを打破するための鍵であることを論じている．

第２章は内容語ドメイン知識獲得の提案手法について述べている．ドメイン知識とは，ある単語に対してその単語が属するドメインをラベルとして与えたものであり，例えば「教科書」という単語に対しては＜教育・学習＞ドメインが与えられる．従来のドメイン知識獲得手法が，人手で整備された高品質のシソーラス，あるいは知識ベースを前提とするのに対し，提案手法は，Ｗｅｂ検索エンジンへのアクセスのみを前提とする．単語へのドメイン付与精度は８０％以上と高い値を示している．ドメイン体系の設計に関しても，従来手法が既存のシソーラスあるいは知識ベースのドメイン体系に強く依存せざるを得ないのに対して，提案手法は，ユーザが自由にドメイン体系を設計することが可能である．さらに本論文では，３万語程度の基本的な語彙に対して与えたドメイン知識を元にして，未知語に対しても高速でドメインを推定する手法を開発した．この未知語ドメイン推定手法も前提としているのは，基本語に対するドメイン知識とＷｅｂ検索エンジンへのアクセスのみである．獲得したドメイン知識は，ブログ分類と慣用句検出の２つのタスクに用いることでその有効性が検証されている．両タスクにおいて高い精度（ブログ分類精度９４％，慣用句検出精度８９％）が得られており，本論文のドメイン知識獲得手法が高精度であることが確認された．本論文の手法によって獲得したドメイン知識は，人手によるチェックを経て，オープンソースの形態素解析器ＪＵＭＡＮの辞書に組み込まれる形で公開され，広く利用可能になっている．

第３章は動詞間含意知識獲得の提案手法について述べている．動詞間含意知識とは，例えば「がぶ飲みする　→　飲む」のように，一方の動詞の事態が成立するなら必然的にもう一方の動詞の事態も成立するような動詞ペアについての知識である．本論文では，従来の含意知識獲得手法には，低頻度語に対する頑健性とコーパスの偏りに対する頑健性の２つの点で問題があると指摘している．提案手法は，多くの先行研究で用いられているが低頻度語を過大に評価する傾向がある相互情報量ではなく，動詞とその文脈単語の出現のしやすさに関する条件付き確率に基づいて文脈単語を重み付けることで低頻度語への頑健性を向上させている．また，動詞間で共有されている，重みが最大の文脈単語を動詞含意スコアの計算から一様に除外するというトリックを用いることでコーパスの偏りに対する頑健性を向上させている．代表的な先行研究の手法３つと比較した結果，提案手法が最も高い精度を示すことが確認された．比較的低頻度な動詞を対象にした実験により，提案手法の低頻度語への頑健性の高さが確認された．また，重みが最大の共有文脈単語を除外するというトリックを用いた場合と用いない場合とを比較した結果，当該トリックがコーパスの偏りへの頑健性を高める上で有効であることが確認された．提案手法で獲得した動詞間含意知識は，人手によるチェックを経て，高度言語情報フォーラムＡＬＡＧＩＮから配布中である．

第４章はフレーズ間言い換え知識獲得の提案手法について述べている．本知識は，「角質を取り除く⇔角質をはがす」「派遣先企業の社員になる⇔派遣先に直接雇用される」等のような表現のバリエーションを吸収するための知識であり，テキスト理解にとって中心的な役割を果たす．従来手法は分布類似度によるものとパラレルコーパスによるものとがあるが，前者は低頻度表現とコーパスの偏りに対する脆弱性に加えて，同義と反義の区別がつかないという致命的弱点がある．後者からは高品質な言い換え知識が得られやすいが，パラレルコーパスを大量に収集することが困難であるという弱点がある．本論文では，Ｗｅｂ上の大量の定義文を自動獲得して同一概念の定義文をペアにすることで，大規模なパラレルコーパス（実際には定義文ペア）を自動的に構築することに成功した．これにより，従来手法では成し得なかった，言い換え知識の品質と規模を両立させることに成功した．評価実験で，代表的な先行研究の手法３つと本論文独自のベースライン手法の計４手法を提案手法と比較した結果，提案手法が獲得数においても適合率においても最も優れていることが確認された．また，類似度の高いあらゆる文をペアにして得られたパラレルコーパスと定義文ペアのみから成るパラレルコーパスを提案手法の言い換え知識獲得源として比較した結果，定義文ペアを用いた方が，獲得数においても適合率においても優れていることが確認された．本論文の手法で獲得した言い換え知識は，現在人手でチェック中であり，完成後，高度言語情報フォーラムＡＬＡＧＩＮから配布する予定である．

第５章は結論であり，本論文を総括している．

（続紙　２　）

（論文審査の結果の要旨）

本論文は，計算機によるテキスト理解を実現する上で不可欠だが，獲得技術が発展途上にある知識，すなわち，内容語に対するドメイン知識，動詞間含意知識，フレーズ間言い換え知識をＷｅｂから自動獲得する手法を研究し，その成果をまとめたものである．得られた主要な成果は以下の通りである．

１．従来のドメイン知識獲得手法は高度に構造化された既存の知識ベースを前提としており，そのような知識ベースの無い言語には適用できず，また，ドメイン体系も既存の知識ベースに強く依存したものとならざるを得なかった．一方，提案手法は，Ｗｅｂ検索エンジンへのアクセスのみを前提としており，多くの言語に対して適用できる上，ドメイン体系も自由に設計できる．加えて本論文では，基本語のドメイン知識とＷｅｂ検索エンジンへのアクセスのみを前提とする，未知語のドメインを高速で自動推定する手法の開発に成功した．さらに本論文では，基本語に対するドメイン知識獲得結果を元に言語資源を整備，公開しており，現在広く利用可能となっている．

２．従来の動詞間含意知識獲得手法は，低頻度語への頑健性とコーパスの偏りに対する頑健性が大きく欠けていた．一方，提案手法は，多くの先行研究で用いられているが低頻度語を過大に評価する傾向がある相互情報量ではなく，動詞とその文脈単語の出現のしやすさに関する条件付き確率に基づいて文脈単語を重み付けることで低頻度語への頑健性を向上させている．また，動詞間で共有されている，重みが最大の文脈単語を動詞含意スコアの計算から一様に除外するというトリックを用いることでコーパスの偏りに対する頑健性を大幅に向上させている．さらに本論文では，動詞間含意知識獲得結果を元に言語資源を整備，公開しており，現在広く利用可能となっている．

３．従来のフレーズ間言い換え知識獲得手法は，分布類似度によるものとパラレルコーパスによるものとがあるが，前者は低頻度表現とコーパスの偏りに対する脆弱性に加えて，同義と反義の区別がつかないという致命的弱点がある．後者からは高品質な言い換え知識が得られやすいが，パラレルコーパスを大量に収集することが困難であるという弱点がある．本論文では，Ｗｅｂ上の大量の定義文を自動獲得して同一概念の定義文をペアにすることで，大規模なパラレルコーパス（実際には定義文ペア）を自動的に構築することに成功した．これにより，従来手法では成し得なかった，言い換え知識の品質と規模を両立させることに成功した．

よって，本論文は博士（情報学）の学位論文として価値あるものと認める．また，平成２３年８月２４日実施した論文内容とそれに関連した試問の結果合格と認めた．