TSUBAME 共同利用　平成 27 年度　学術利用　成果報告書

利用課題名　知識に基づく構造的言語処理の確立と知識インフラの構築
英文：Establishment of Knowledge-Intensive Structural Natural Language Processing and Construction of Knowledge Infrastructure

利用課題責任者　黒橋 禎夫
First name Surname　Sadao Kurohashi

所属　京都大学 大学院情報学研究科
Affiliation　Graduate School of Informatics, Kyoto University
URL　http://nlp.ist.i.kyoto-u.ac.jp/

邦文抄録（300 字程度）
本利用課題では，知識に基づく頑健で高精度な構造的言語処理を実現し，これによって様々なテキストの横断的な関連付け，検索，比較を可能とする知識インフラを構築することを目的とする．そのために，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知識を獲得する．知識としては，動詞の意味フレームと言語モデルを対象とする．前者は構造的言語処理の基盤的な知識であり，後者はテキストの形態素解析に必要な知識である．この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（100 words 程度）
In this project, we perform linguistic analysis to a large-scale Web corpus and use the resulting analyses to acquire wide-coverage knowledge, such as semantic frames and language model. The knowledge acquired can be used to improve linguistic analysis and morphological analysis. We accomplished these knowledge acquisition processes quite rapidly using TSUBAME.

Keywords: natural language processing, Web, semantic frame, language model

背景と目的

テキストは，専門家によるデータの分析結果や解釈，ステークホルダーの批判・意見，種々の手続きや ノウハウなどが表出されたものであり，人間の知識表現の根幹である．テキストとして表現された知識を計算機によって抽出・関連付けすることができれば，社会における知識循環を円滑化し，異なる分野間での知識の相互関連性の発見や，新しい知識・法則の発見を支援することが可能となる．言語情報処理はウェブをはじめとする大規模テキストの活用によって長足の進歩を遂げつつあるが，本利用課題ではこれをさらに発展させ，知識に基づく頑健で高精度な構造的言語処理を実現し，これによって様々なテキストの横断的な関連付け，検索，比較を可能とする知識インフラを構築する．

本利用課題では，TSUBAMEの大規模計算機環境を用いて，上記の目的を実現するために必要となる知識を大規模テキストコーパスから自動獲得する．

本利用課題では，大規模 Web コーパスに言語解析を適用し，2種類の知識を獲得した．

日本語格フレームの獲得

1つ目の知識は，省略・照応・談話解析の基盤的知識となる意味フレームであり，述語がどのような格をとり，どのような名詞と関係をもつかを記述したものである．次に"observe"という英語の動詞に対する一つの意味フレームを示す．

subj:{child,people, …} observe dobj:{bird, animal, …} prep_at:{range, time,…}

平成 25 年度は，[Kawahara+ 2014]の手法を言語解析結果に対して適用して意味フレームを獲得した．この手法は，言語解析結果から述語項構造を抽出し，まず似た意味をもつものをマージして初期フレームを作る．次に，初期フレームを Chinese Restaurant Process によってクラスタリングすることによって，述語ごとに最適な数の意味フレームを自動的に推定する．このような段階

的プロセスによって，大規模な言語解析結果を入力しても，スケーラブルに意味フレームの獲得を行うことができる．

平成 26 年度は同様の手法を日本語に適用することで，日本語の意味フレームの獲得も行ったが，分析したところ，格パターンの多様性に頑健ではないという問題があった．そこで，本年度は，用例を格の出現の仕方に基づいて 3 種類に分けそれぞれについて格フレームを構築してから最後に統合する手法を提案した．

結果および考察

従来の研究では，用例をクラスタリングして，述語の語義と格パターンの違いの両方をとらえようとしてきたが，異なる格パターンを含む格フレームを作ってしまうという問題があった．例えば「拡大する」は，自動詞用法と他動詞用法の 2 種類の格のとり方がある．自動詞用法は例（1a）のようにガ格に対象物をとりヲ格はとらないパターンで，他動詞用法は例（1b）のようにガ格に動作主体をとりヲ格に対象物をとるパターンである．

a. 規模が 倍に 拡大する

b. 企業が 規模を 倍に 拡大する

従来の手法では，格パターンの違いを考慮せずにクラスタリングしていたため，

{ 規模, . . .} が { 規模, . . .} を { 倍, . . .} に 拡大する

という実際には用いられない格の組み合わせをもつ格フレームができてしまっていた．

そこで，用例を下図のように，

(1)ヲ格を含む用例

(2)ヲ格は含まず無生物であるガ格を含む用例

(3)その他の用例

の3種類に分けそれぞれについて格フレームを構築してから最後に統合する手法を提案した．（図1）これにより，異なる格パターンを含む格フレームの作成が避けられる．

評価実験では，日本語ウェブテキスト 40 億文に対して提案手法を用いた格フレームを構築し，そのうち 200 格フレームについて妥当な格フレームの割合を調べた．結果，従来手法と比較して 72.5%(145/200) から 95.5%(191/200) に大きく改善したことを確認した．なお本研究の成果は，[林部+ 2015]で発表した．

まとめ，今後の課題

今後は，本手法で構築した格フレームを用いて述語項構造解析を行う予定である．

意味的言語知識に基づく形態素解析モデル

2つめの知識は，統計的言語モデルである．統計的言語モデルとは，「雨が」には「降る」や「止んだ」が続きやすいというような単語の出現のしやすさを確率として表す，語彙に関する統計的な知識である．

高次の文脈解析を正確に行おうとすれば，意味の基本単位であり，もちろん照応・省略の基本単位でもある単語の認識（形態素解析）を極めて高い精度で行っておく必要がある．従来の形態素解析は 98%前後の精度が達成されているものの，この観点からまだ十分な精度とはいえない．

従来の形態素解析には，語彙の不足と，解析において意味を考慮していないという，2つの大きな問題点が残されていた．語彙の不足は，その単語を適切な単位で認識できないというだけでなく，その前後の

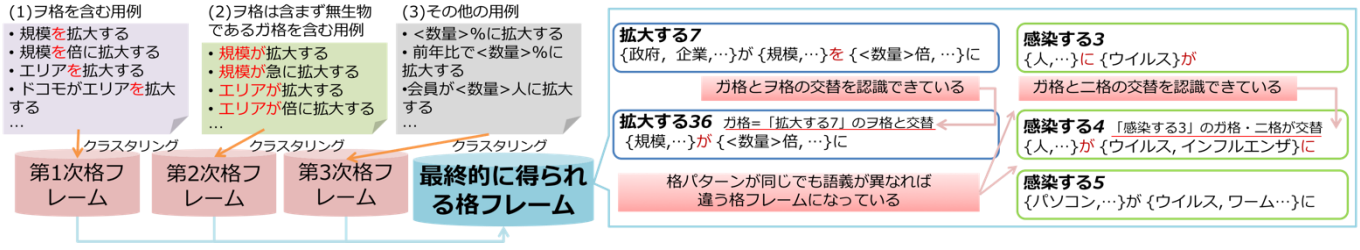

図 1 構築した格フレームの概要

単語の解析にも悪影響を及ぼすことが少なくない．意味を考慮していない問題は，特に複数の解釈がありうる複合名詞の分割などで問題となる．例えば日本語テキストの解析で広く利用されている形態素解析器 JUMAN や MeCab においても，「外国人参政権」という文字列を「外国/人参/政権」のように誤って分割してしまう．

本研究ではまず，Wikipedia，Wiktionary および Web テキストコーパスから，大規模に語彙を獲得し，これによって，未知語による形態素解析誤りを大きく削減した．Wikipedia からは，「橋下」「お茶の水」「東プレ」「アベノミクス」「妊活」等の 844,263 語，Wiktionary からは，「インセンティヴ」「糾す」「肥沃だ」等の動詞・形容詞を含む 2,207 語，Web テキストコーパスからは「逆進税」「政独委」（政策評価・独立行政法人評価委員会の略称)といった 5,688 語を獲得した．

さらに，大規模な Web テキストコーパスを解析することにより語彙の意味的な振る舞いに関する知識を，統計的言語モデルとして獲得した．空白で区切られた英語などと違い，日本語や中国語のように空白で区切られていない言語においては，言語モデルを構築するために形態素解析システムの出力が必要となるが，このシステムの出力には，「外国/人参/政権」のような誤った解析が含まれてしまう．そのため，このような形態素解析の誤りから学習された言語モデルをそのまま形態素解析に利用したとしても，単に誤りの再生産が起こるだけである．しかし，意味的に汎化したレベルからこの単語の並びを見ると，
〈国〉-〈野菜〉-〈政治〉のような特異的な並びになっており，このような誤った解析は特定の単語の並びに対してのみ起こるため，Web テキストコーパスを解析して得られる意味的に妥当な大多数の解析により，意味的に不自然な単語並びの悪影響を打ち消すことができる．

本項目では，RNNLM（Recurrent Neural Network Language Model）と呼ばれる意味的に汎化された言語モデルを用いることで，意味的に不自然な解析を抑える形態素解析手法[Morita+ 2015]を提案した．TSUBAME の大規模計算機環境を用いることにより，JUMAN により解析した大規模な Web テキストコーパスを用いて RNNLM を構築し，さらに，中規模の人手でラベル付けされたコーパスで言語モデルの再学習を行った(図 2)．Web 上のリソースから獲得した語彙知識による辞書の更新と合わせ，新聞テキスト（京都大学テキストコーパス）および，Web テキスト一般（京都大学 Web 文書リードコーパス）の各 2,000 文で評価したところ，大幅な精度の改善を達成した(図 3)．図において，新聞テキスト，Web テキスト一般のいずれにおいても，デフォルトの辞書の JUMAN，MeCab

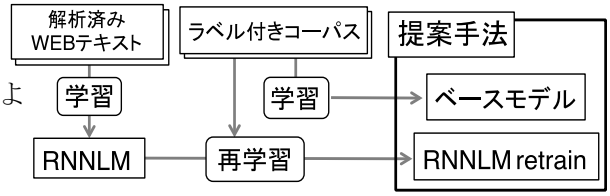

図 2: 形態素解析器の学習フロー

よりも，大規模語彙を補強した JUMAN，MeCab の方が精度が上昇している．さらに，ここに言語モデルを加えた場合，一般の単語ベースの言語モデルである SRILM を用いる場合にはかえって精度が悪化するのに対して，提案手法の意味的汎化による言語モデル RNNLMretrain を利用する方法では大きく精度が向上していることが分かる．

本手法を用いて，TSUBAME の計算機環境を用いることで38億文の解析を行った．今後はこの解析

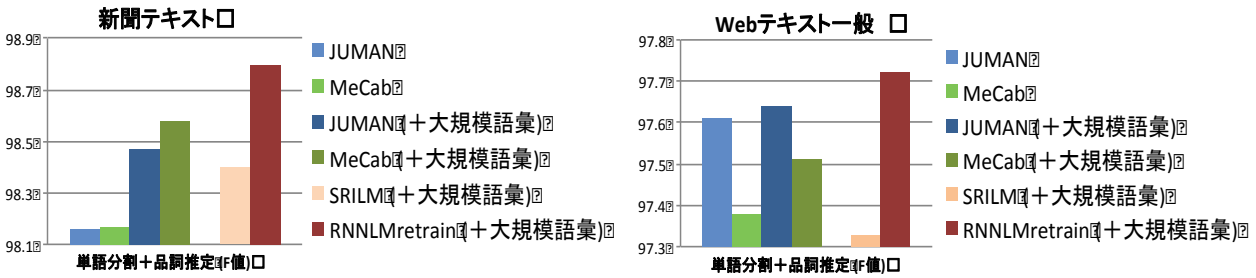

図 3: 語彙的言語知識および意味的言語知識に基づく形態素解析モデルによる精度向上

結果を元に，より高精度な知識の獲得を行う．また，より多くの語彙知識を利用する方法について検討を進めていく．

参考文献

[Kawahara+ 2014] Daisuke Kawahara, Daniel W. Peterson, Octavian Popescu and Martha Palmer. A Step-wise Usage-based Method for Inducing Polysemy-aware Verb Classes,. In Proceedings of ACL2014, 2014.

[林部+ 2015] 林部祐太，河原大輔，黒橋禎夫 .「格パターンの多様性に頑健な日本語格フレーム構築」， 情報処理学会第 224 回自然言語処理研究会，2015

[Morita+ 2015] Hajime Morita, Daisuke Kawahara and Sadao Kurohashi: Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model, Proceedings of EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 2015