TSUBAME 共同利用　平成28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，事態間知識などを対象とする．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 aim to establish robust and accurate knowledge-intensive structural NLP, and to construct a knowledge infrastructure which can relate, retrieve and compare knowledge from various texts. We perform linguistic analysis to a large-scale Web corpus and use the resulting analyses to acquire wide-coverage knowledge, such as inter-event relation knowledge. We accomplished several knowledge acquisition processes quite rapidly using TSUBAME.

Keywords: **5つ程度**
natural Language processing, Web, inter-event knowledge, parsing, similarity learning

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．具体的には事態間知識獲得，論理式と定理証明を用いた文の類似度学習，英語 Wikipedia の構文解析ならびに VecDCS モデルの訓練を行った．

大規模テキストからの事態間知識獲得

背景

自然言語理解のためには様々な言語知識が必要となる．一つには述語と項の関係がある．これは格フレームという形でコーパスから自動獲得され，構文解析などで有効性が示されている．さらに述語項構造の間の知識が重要となる(事態間知識とよぶ)．例えば，「X(人)が Y(財布)を拾う → X が Y を警察に届ける」のような知識である．事態間知識は共参照解析や照応解析などの基礎解析や対話などのアプリケーションで有用である．

方法

本利用課題では[Shibata+ 2014]の手法を用いて事態間知識獲得を行った．手法の概要を図 1 に示す．まず，形態素解析器 JUMAN++，構文・格解析器 KNP で大規模テキストの解析を行い，その解析結果

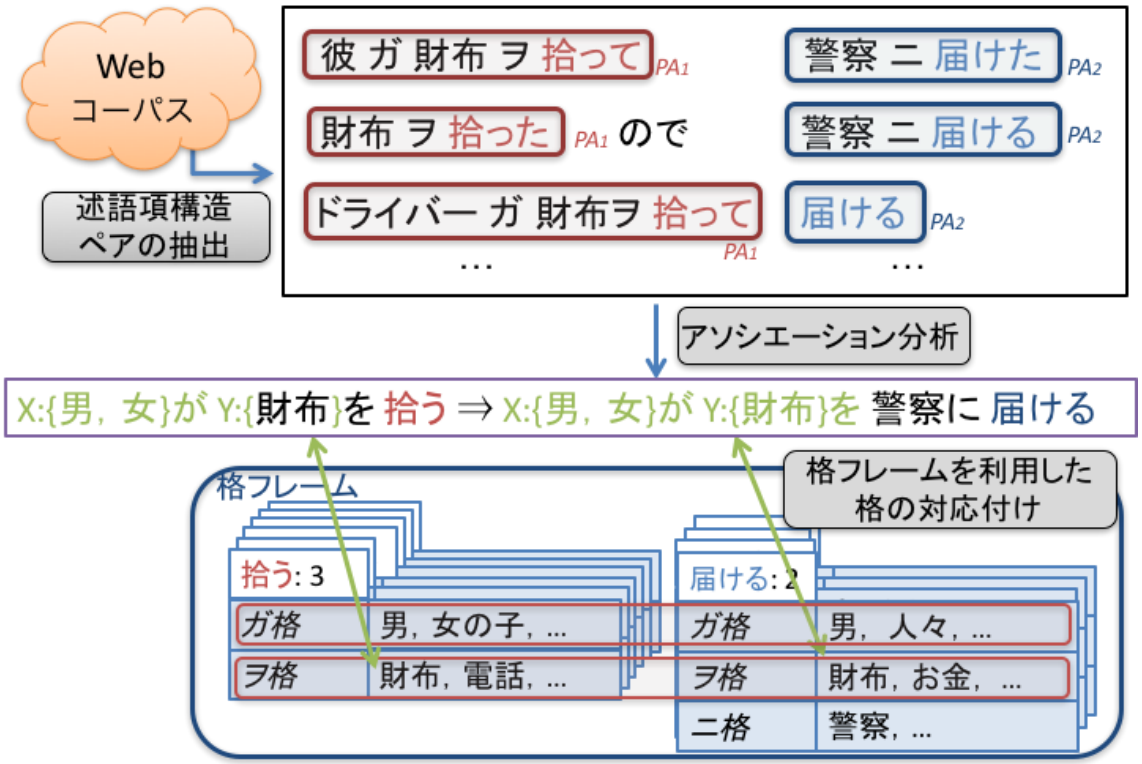

図 1 大規模テキストからの事態間知識獲得

から係り受け関係にある述語項構造を抽出する．

次に，相互情報量の高い述語項構造ペアを得る．ここで任意の述語項構造ペアの組み合わせは膨大となるため，いかにして共起度の高い述語項構造ペアを見つけるかが問題となる．この問題に対してはアソシエーション分析を用いて相互情報量の高い述語項構造ペアを効率的に見つける．

最後に，格フレームにおける格要素の分布の類似性に基づいて項の対応付けをとる．

結果

日本語約 15 億文のコーパスから約 40 万ペアの事態間知識を得ることができた．形態素解析や構文・格解析，ならびに事態間知識獲得は時間のかかる処理であるが，TSUBAME を用いることにより，短期間で行うことができた．今後は得られた事態間知識を省略照応解析などでの有効性を確認する予定である．

論理式と定理証明を用いた文の類似度学習

背景

文の持つ意味に基づいて文の類似度を計算できれば，膨大な文書から知りたい情報を含む文書を特定でき，文書検索や文書要約への幅広い応用が期待される．

従来の研究では，文の表層的な特徴に基づいて文の類似度を計算する統計的手法が一定の成果を収めているが，否定やモダリティなど，より深い意味処理を要する現象の扱いに問題が残されている．一方，文全体の意味を論理式で表現し，推論を用いて文の意味を捉えるアプローチは，組合せ範疇文法(CCG)など，現代的な統語解析手法と組み合わせることで，含意関係認識など意味処理の分野で成功を収めつつある．しかし，論理的アプローチを文の類似度の計算へ応用する試みはこれまで十分になされてこなかった．

方法

このような背景のもと，本研究では，統計的手法と論理的手法を組み合わせることで，文全体の意味を論理式で表現し，定理証明の手法を用いて文の類似度を学習する方法を検討した．この方法を実現するために，本研究班がこれまでに開発した，統語解析・意味解析・推論を統合したシステム ccg2lambda[Martinez-Gomez+ 2016]を利用した．ccg2lambda では，組合せ範疇文法に基づく統語解析と意味合成によって文を高階述語論理式に変換し，定理証明系 Coq を用いて自動推論が行われる．

本研究の提案手法の全体像を図 2 に示す．

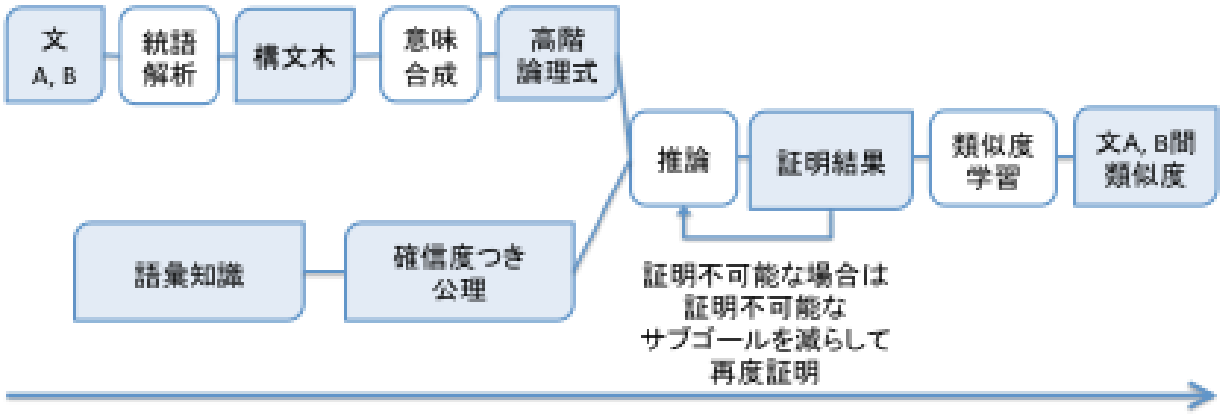

図 2 文の類似度学習

本研究では文A, B 間の意味の類似度を文 A, B の含意関係を介して定義する. これは, A → B, 及び, B → A を証明できる可能性が高いほど, 文 A, B の類似度は高いという仮説に基づく. A, B 間に含意関係が成立するか否かを自動推論によって定量化し, その結果を特徴量に用いることで, 文 A, B 間の類似度学習を実現する. 特に, 自動推論によって推論規則の適用回数といった証明の過程や, 証明の結果に関する情報を取得し, 特徴量を設計する. そして, 決定木学習によって類似度の学習を行う.

結果

本研究の提案手法について, SemEval2014 ワークショップで文の意味的類似度評価のデータセットとして導入された SICK データセットを用いて評価実験を行った. 訓練データとして 5000 文ペア, テストデータとして 4927 文ペアを用いた. 正解スコアと予測スコアとの Pearson 相関係数 γ によって評価した結果, γ =0.833 と従来の論理的手法に基づくアプローチ[Bjerva+ 2014]を上回る結果を達成した[谷中+ 2017]. TSUBAME の分散並列処理の利用により, CCG による統語解析, ccg2lambda による意味解析と自動証明, 決定木学習の各処理に関して, きわめて短期間に成果を得ることができた. 今後改良を重ねることで, さらなる精度向上が見込まれる.

英語 Wikipedia の構文解析

概要

構文解析は, 例えば “Alice drove down the street in her car.” のような文に対して, Alice が主語であり, drove が Alice の行った動作で, もっと具体的には “drove in her car” -- 彼女の車を運転したことを認識するプロセスである. このような解析は, 言語の意味を理解するために極めて重要だと考えられる.

従来英語の構文解析には, Stanford Parser が広く使われてきたが, 2016 年に公開された構文解析器 SyntaxNet がこれを精度と速度の両面で大きく上回った. 本プロジェクトの目的は, 英語 Wikipedia の文書に対して, SyntaxNet による構文解析を行うことである.

結果および考察

英語 Wikipedia は, およそ 1 億文を含む. 従来 Stanford Parser を使った場合, 6コアのクラスター・ノード 20 個でおよそ一週間の解析時間を必要とした. 今回, SyntaxNet はニューラルネットを使ったモデルであるため, GPU による計算が効率的である. 3つの GPU を有効に利用するため, SyntaxNet のソースコードを一部変更した. 結果, TSUBAME の G キュー18 ノードを使い, 9 時間で解析が終了した.

まとめ, 今後の課題

今後は, より大規模なコーパス, 例えば ClueWeb などに対する解析を検討している.

VecDCS モデルの訓練

概要

本プロジェクトは, [Tian+ 2016]の手法を使い, 前項で解析した Wikipedia の構文データから, 意味解析に適した分散表現を学習する目的にある. [Tian+ 2016]の手法は, 例えば “drive” の主語に人が, 目的語に車がより多く現れる傾向にあることを学習し, かつ簡単なベクトル

の足し算によって意味の合成（例えば “take a drive” と “go out” のような，複数語からなるフレーズの意味的類似性を捉える）ができるモデルである．

SyntaxNet による高精度な構文解析を使用することで，より高精度な分散表現の学習が期待できる．今回は，より洗練したモデルの改良を目指すため，訓練のハイパー・パラメータ・チューニングを行った．

結果および考察

S キューの計算ノードを使い，一回およそ 24 時間で訓練したモデルを評価し，収束の早いハイパー・パラメータの傾向がわかった．

まとめ，今後の課題

今後は，訓練したモデルを意味解析タスクへの応用や，訓練で得られた知見をモデルの改良に使う予定である．

参考文献

[Shibata+ 2014] Tomohide Shibata, Shotaro Kohama and Sadao Kurohashi: A Large Scale Database of Strongly-related Events in Japanese, Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC2014), pp.3283-3288, Reykjavik, Iceland, 2014.

[Martinez-Gomez+ 2016] Pascual Martinez-Gomez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. ccg2lambda: A compositional semantics system. ACL2016 System Demonstrations, pp.85-90, 2016.

[Bjerva+ 2014] Johannes Bjerva, Johan Bos, Rob Goot, and Malvina Nissim. The Meaning Factory: Formal semantics for recognizing textual entailment and determining semantic similarity. SemEval 2014: International Workshop on Semantic Evaluation, pp. 642-646, 2014.

[谷中+ 2017] 谷中瞳，峯島宏次，Pascual Martinez-Gomez，戸次大介，論理式による意味表現と証明プロセスに着目した文の類似度学習方法の提案，言語処理学会全国大会，2017.

[Tian+ 2016] Ran Tian, Naoaki Okazaki, Kentaro Inui: Learning Semantically and Additively Compositional Distributional Representations; In Proceedings of ACL, 2016.