TSUBAME 共同利用　平成 29 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．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. We accomplished several knowledge acquisition processes quite rapidly using TSUBAME.

Keywords: **5つ程度**
natural Language processing, Web, predicate-argument structure analysis, 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．具体的には述語項構造解析・共参照解析の同時学習，論理式と定理証明を用いたフレーズ公理生成，関係知識獲得ならびに英語コーパスの構文解析を行った．

Entity-Centric な述語項構造解析・共参照解析の同時学習

背景

述語項構造解析とはテキスト中の「誰が何をどうした」を明らかにする解析である．述語項構造解析は情報抽出や質問応答など様々なアプリケーションにとって重要となる基礎解析である．

述語項構造解析は格解析とゼロ照応解析からなる．格解析は述語と項が係り受け関係にある場合に格関係を明らかにする解析であり，ゼロ照応解析は述語と項が係り受け関係にない場合に項を同定する解析である．ゼロ照応解析はさらに文内ゼロ照応解析（述語と項が同一文にある）と文間ゼロ照応解析（述語の文より前の文に項がある）に大別することができる．述語と項の依存構造上のパスなどの手がかりが使えない，項の候補がたくさん

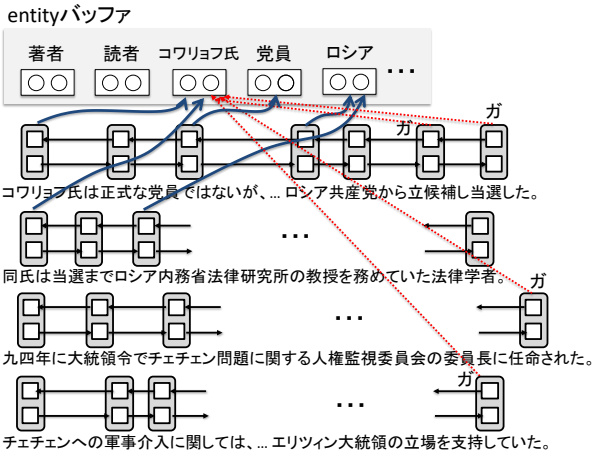

図 1 述語項構造解析と共参照解析の同時学習

あるなどの理由から文間ゼロ照応解析は非常に難しいタスクである．文間ゼロ照応解析の精度を向上させるためには文章中で何が話題の中心であるかなどを捉える必要がある．

本研究では話題の中心を捉えるために entity という概念を導入する[柴田+ 18]．entity を考えるためには共参照解析を行う必要があるので，述語項構造解析と共参照解析を同時に行う．Wiseman らは RNN (Recurrent Neural Network)で entity の embedding を計算しており[Wiseman+ 2016]，本研究でもこの手法にならい entity の embedding を計算し，話題の中心を捉える．

方法

提案手法の概要を図 1 に示す．各 entity には embedding を割り当て，図の上部に示すバッファで管理する．入力テキストが与えられると，まず Bi-directional LSTM を用いて文脈を考慮した句の表現を得る．次に，文章の先頭から対象の句が名詞句の場合は共参照解析を，動詞句の場合は述語項構造解析を行う．共参照解析・述語項構造解析ともにニューラルネットワークを用いてスコアを計算する．その際に mention の embedding(Bi-directional LSTM で得られるもの)だけでなく entity の embedding を考慮する．そして，両解析の解析結果を用いて RNN で embedding を更新し，情報を蓄積する．

結果

評価は京都大学ウェブ文書リードコーパス(ウェブ，約 5,000 文書)と京都大学テキストコーパス(新聞，約 550 文書)の 2 種類のコーパスで行った．entity を考慮しないベースライン手法と提案手法を比較した．

格解析と共参照解析の精度はほぼ変化ないが，ゼロ照応解析の精度はベースラインに対してウェブで 3.8 ポイント(0.523 → 0.561)，新聞で 7.5 ポイント(0.280 → 0.355)の精度向上がみられた．特に文間ゼロ照応解析の精度がウェブで 27.2 ポイント(0.046 → 0.318)，新聞で 15.5 ポイント(0.028 → 0.183)と大幅な精度向上がみられた．

訓練とテストは 1GPU で半日程度の時間を要する．様々な条件での実行が必要となるが，TSUBAME を用いることで単時間での実行が可能となった．

まとめ，今後の課題

提案手法は文間ゼロ照応解析の精度を大きく向上することができた．今後の課題として橋渡し参照などを行うことや事態間関係知識などの外部知識を取り入れることなどがあげられる．

論理式と定理証明を用いたフレーズ公理生成

背景

含意関係認識(Recognizing Textual Entailment，RTE)は，テキストTが仮説文Hを含意するか否かを自動判定する自然言語処理のタスクであり，意味処理のもっとも重要な基盤技術の一つである．文の意味を論理式で表現し，論理推論によって高度な意味解析を行う手法は，論理式による意味表現と整合性の高い組合せ範疇文(CCG)による頑健な統語解析の発展に伴い，RTE や文間類似度学習のタスク [Yanaka+ 2017]において高精度を達成している．

論理推論によるアプローチでは，否定表現や数量表現など機能語の意味解析を正確かつ効率的に扱うことができるが，内容語の語彙知識が必要な推論については課題が残されている．特に，論理ベースの手法において，WordNet など既存の大規模データベースの語彙知識を利用することで，単語間知識の扱いは一定の有効性が認められているが，フレーズ間知識の扱いは十分に検討されてこなかった．

方法

そこで本研究では，論理ベースの深い意味解析の手法に基づいてフレーズ間の語彙知識を獲得し，論理推論と組み合わせて用いる手法について検討した．フレーズ間知識を論理推論で扱うためには，文中のフレー

ズの適切な対応付け（フレーズアライメント）を特定する必要がある．先行研究では文の表層情報を利用してフレーズアライメントを行う手法が研究されているが，*cut … into pieces* など，非連続なフレーズを扱うことが難しいという問題がある．文の意味を論理式で表現することで，こうした非連続なフレーズを含む広範囲の推論・言い換えが扱えるようになると期待される．

より具体的には， 本研究では，自然演繹に基づく文間の含意関係の証明の実行過程からフレーズアライメントを特定し，論理推論に適用する手法を提案した．提案手法を用いてフレーズ間知識の公理を生成し，RTE の問題に適用することで，単語間知識だけでは推論できなかったパラフレーズを含む文間の含意関係が推論可能となるか，RTE 評価用データセットによる評価を行った．

結果

提案手法の評価には RTE 評価用データセットである SemEval2014 SICK を用いた．SICK は 2 つの文の含意関係（yes, no, unknown）が人手で付与されている．訓練データは 5000 件，テストデータは 4927 件である．CCG 解析によって文を高階論理式に変換し，自動推論を行うシステム ccg2lambda に提案手法であるフレーズアライメント機能，フレーズ間の公理生成機能を拡張し，実験を行った．まず，訓練データ中の RTE 正解ラベルが yes か no の全文例について，提案手法を用いて含意関係の証明を行った上で，フレーズアライメントを特定し，フレーズ間知識の公理を収集した．次に，テストデータ中の全文例で含意関係の証明を行い，評価を行った．

実験の結果，フレーズ間・単語間知識の公理生成を組み合わせた場合の正答率は 84.3%であり，公理生成なし(76.6%)，単語間公理生成のみ(83.1%)の場合と比べて向上が見られた．また，SemEval2014 のトップスコア(83.6%)を上回る結果となった．フレーズ間知識の公理収集時・評価時の証明時間はいずれも 1 文ペアあたり平均 3.0 秒だった．

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と言える．今後，訓練データから収集したフレーズを用いて公理を生成するか否かを判定する分類器を作成することで，さらなる精度向上が見込まれる．

オートエンコーダとの同時学習による関係知識共有

概要

WordNet，YAGO，Freebase などのような大規模知識ベースは，事実を（ヘッドエンティティ，関係，テールエンティティ）の三つ組で表現し，意味解析，情報抽出，質問応答などに広く応用されてきた．しかし事実を完全網羅するのが難しいため，欠損されたエンティティを予測する知識ベース補完が盛んに研究され，これらの研究では，エンティティと関係を含む知識ベースの要素を連続ベクトル空間に埋め込むことで，格納された事実（三つ組）の背後に存在する統計的な規則性をモデル化し，記述漏れした事実を導き出す．

この連続空間埋め込みを行う際の研究課題の一つに，関係間の知識共有がある．直感的に，例えば「（人物が）（作品を）監督した」と「（人物が）（作品を）製作した」という関係はどちらも「人物」と「作品」を結びつく概念であるように，多くの関係はいくつかの概念を共有するので，このような直感を知識ベースのモデル化に取り入れることが望ましい．本研究は，知識ベースの埋め込みを訓練すると同時に，関係に対するオートエンコーダを合わせて学習することで関係間の知識共有を促す柔軟な枠組みを提案し，評価実験を行なった．この手法は，関係のモデル化自体に何の制約も与えずに，知識共有の強さ度合いをオートエンコーダのコード長で柔軟に調整できる利点を持つ．

方法

具体的には，これまでに開発した VecDCS モデル[Tian+ 2016]をベースにした学習手法を知識ベースの埋め込みに適用し，その学習と同時に関係を表す行列を低次元コードから復号化するオートエンコーダを加えた．

結果および考察

知識ベース補間の標準的な評価データである WN18，FB15k，WN18RR，FB15k-237，YAGO3-10 などを用いてモデルを訓練し，評価を行った．訓練は TSUBAME の f_node を利用し，CPU 28 並列で行なった．その結果，VecDCS モデルの学習手法を利用したベースモデルは，既に多くのデータセットで state-of-the-art を達

成しており，今回提案のオートエンコーダを加えると更なる精度向上が見られた．これら結果の報告は[高橋+ 2018]で行なった．

まとめ，今後の課題

今後は，パラメータ・チューニングによる更なる精度向上と，知識ベース補間と含意関係認識の統合を目指すよう研究を進めて行きたい．

英語コーパスの構文解析

概要

本研究は，単語の分散表現に感情極性や文書トピックなど，単語生成に関わるいろんな因子を取り込むことを目指す．そのために用いたデータセットは Multi-domain Sentiment Dataset で，25 カテゴリの商品に対する 140 万レビューを集めたコーパスである．このコーパスに対して，SyntaxNet を用いた品詞タグ推定，及び依存構造解析を行なった．

結果および考察

構文解析は TSUBAME の q_node を使い，GPU を用いて行なった．24時間以内に全コーパスの解析が終わった．この解析結果を用いて訓練した単語分散表現のモデルから，各因子だけに関わる部分の特徴が現れた．これらの結果は[田+ 2018]で報告した．

まとめ，今後の課題

今後は，訓練した単語表現を他分野への転移学習や含意関係認識に応用する予定である．

参考文献

[柴田+ 2018] 柴田知秀，黒橋禎夫．Entity-Centric な述語項構造解析・共参照解析の同時学習．言語処理学会第 24 回年次大会，2018.

[Wiseman+ 2016] Wiseman, Sam, Rush, Alexander M. and Shieber, Stuart M. Learning Global Features for Coreference Resolution. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[Yanaka+ 2017] Hitomi Yanaka, Koji Mineshima, Pascual Martínez-Goméz and Daisuke Bekki. Determining Semantic Textual Similarity using Natural Deduction Proofs. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP2017), Copenhagen, Denmark, September 7-11, 2017.

[谷中+ 2018] 谷中瞳，峯島宏次，Pascual Martinez-Gomez，戸次大介．自然演繹に基づく文間の含意関係の証明を用いたフレーズアライメントの試み．言語処理学会第 24 回年次大会，2018.

[馬目+ 2018] 馬目華奈，谷中瞳，吉川将司，峯島宏次，戸次大介．RNN 系列変換モデルを用いた高階論理式からの文生成，言語処理学会第 24 回年次大会，2018.

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

[高橋+ 2018] 高橋諒，田然，乾健太郎，オートエンコーダとの同時学習による知識共有，言語処理学会第 24 回年次大会，2018.

[田+ 2018] 田然，渡邉研斗，乾健太郎，多因子単語埋め込みを用いる複数分野感情極性の転移学習，言語処理学会第 24 回年次大会，2018.