TSUBAME 共同利用　平成 30 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, knowledge acquisition, predicate-argument structure analysis, text generation

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と LSTM を用いた論理式からの文生成を行った．

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．

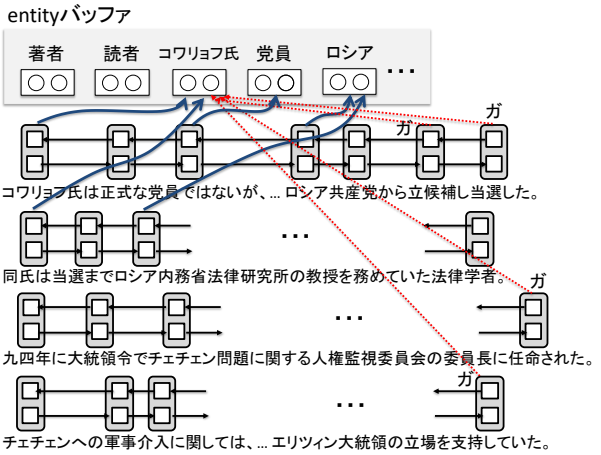

図 1 述語項構造解析と共参照解析の同時学習

述語項構造解析のサブタスクにゼロ照応解析と呼ばれる解析があり，これ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．文間ゼロ照応解析は非常に難しいタスクであり，そ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．

話題の中心を捉えるために entity という概念を導入するモデルを提案した[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．

今年度は昨年度に提案したモデル[柴田+ 18]の上で橋渡し指示解析(名詞に関する関係解析)も行うなどの拡張を試みた．

結果

訓練とテストは 1GPU で半日程度の時間を要する．TSUBAME を用いることで様々な条件での実行が短時間で可能となった．

LSTM を用いた論理式からの文生成

背景

近年の構文解析と意味解析の技術の発展によって，文の意味を論理式で表して高度な推論を行うシステムの構築が可能となった．このようなシステムは，含意関係認識[Mineshima+ 2015] や文間類似度計算のタスク[Yanaka+ 2017]で高精度を達成しており，今後さらなる自然言語処理タスクへの応用が期待されている．

文からその論理式への変換が高精度に行われる一方で，論理式を自然言語文に戻す方法については自明ではない．しかし，論理式から自然言語文に逆変換することができれば，推論システムの改善や，様々な自然言語処理タスクへの応用が期待できる．推論システムにおいては，実社会への応用を考えると，推論に失敗した場合において，なぜ推論に失敗したのかという解釈性が求められる．そこで，推論において証明不可能と判定された論理式を文に変換することができれば，どのような知識が推論に必要であったかを言語化することができる．

また，論理式から自然言語文に変換する方法は，パラフレーズ抽出，テキスト平易化等への応用も可能である．パラフレーズ抽出については，例えば，二文を論理式に変換した上で，それらの論理式に共通する部分を自然言語に変換することにより，二文の意味の重複を言語化する，といった応用が考えられる．また，テキスト平易化については，統計的機械翻訳を用いた手法が研究されているが，統語構造の差異による意味の違いを抽象化する論理式の性質を利用すれば，難しい文を論理式に変換し，論理式から同じ意味を持つ平易な文を生成

することが考えられる.

方法

そこで本研究では，機械翻訳等の系列変換において高い精度を示しているニューラルネットによる系列変換モデル(Sequence-to-Sequence model)を用いて高階論理式から文を生成する手法を提案する．エンコーダの入力となる論理式をどのようにシーケンス化するかについて，複数の方法を提案・比較した．具体的には，文字単位での埋め込みや論理記号単位での埋め込みの比較，加えて論理記号単位の場合においては，論理式の埋め込み順の比較を行なった．さらに，デコーダ部には，論理式に述語として出現しない名詞，動詞を出力させないようにする masking ベクトルを追加した．

また，文生成において，表層の n-gram 一致率を計算する BLEU 等が定量指標として用いられることが多いが，これは類似度以上の意味を考慮した評価を行うことができない．本研究班がこれまでに開発した，統語解析・意味解析・推論の統合的システム ccg2lambda [Mineshima+2015]を用いて，元の文と生成された文との双方向含意関係の証明を行うことで，深い意味的な関係を考慮して生成文を評価する方法を提案する．

結果

実験では，大規模な含意関係認識用データセットである SNLI を用いた．データセット中の開発データ 50,000 件を ccg2lambda を用いて論理式と文のペアを作成した．ルールベースによる手法との比較を行ったところ，BLEU 値，含意関係認識評価のどちらも既存手法を上回った．論理式の埋め込みは，文字単位でなく論理記号単位で埋め込みを行い，論理式のグラフ構造に基づいて論理式中の項を系列に落とし込むことで，非文の生成が抑制された．また，デコーダ部に masking ベクトルを追加することで，BLEU 値，含意関係認識評価のどちらも高くなり，精度向上が見られた．含意関係認識を用いた評価は，文が非文かどうかの判断を行える点でも非常に有用であった．TSUBAME の利用により，大規模データを用いた系列変換モデルの学習が可能となり，本研究の成果を得ることができたと言える．

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.

[Mineshima+ 2015]

Koji Mineshima, Pascual Martínez Gómez, Yusuke Miyao, and Daisuke Bekki. Higher-order logical inference with compositional semantics. In Proc. of EMNLP, 2015.

[Yanaka+ 2017]

Hitomi Yanaka, Koji Mineshima, Pascual Martínez Gómez, and Daisuke

Bekki. Determining Semantic Textual Similarity using Natural Deduction Proofs. In Proc. of EMNLP, 2017.

[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(NLP2018)，岡山，Mar.2018.

[Manome+ 2018]

Kana Manome, Masashi Yoshikawa, Hitomi Yanaka, Pascual Martínez-Gómez, Koji Mineshima and Daisuke Bekki. Neural sentence generation from formal semantics. In Proc. of INLG, 2018.