TSUBAME 共同利用　平成 25 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
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事象間因果関係知識を対象とする。前者は構造的言語処理の基盤的な知識であり、後者はテキストの関連付けに必須の知識である。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
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 causal knowledge between events. The knowledge acquired can be used to improve linguistic analysis and further realize cross-document statement linking, search and comparison. We accomplished these knowledge acquisition processes quite rapidly using TSUBAME.

Keywords: natural language processing, Web, knowledge acquisition, semantic frame, causal knowledge

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。

概要

本利用課題では、大規模 Web コーパスに言語解析を適用し、その解析結果から 2 種類の知識を獲得する。一つ目の知識は、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,…}″

もう一つの知識は、″shoot X → X die″のような事象間因果関係知識である。さらに、獲得知識のアプリケーションにおける評価として、獲得した事象間因果関係知識を Winograd Schema Challenge という照応解析タスクに適用し、その有効性を評価する。なお、すべての実験の対象言語は英語とした。

意味フレームおよび事象間因果関係知識を大規模

に獲得するために、英語の大規模 Web コーパスである ClueWeb 2012 [1] の一部に対して、言語解析処理を適用した。より具体的には、約 2.2 億文書、6.9 億文を含む Web 文書の集合に対して、Stanford CoreNLP [2] による言語解析を行った。言語解析の内容は、トークン化 (tokenization)、文分割、品詞タグ付け、原型化 (lemmatization)、固有名詞認識、句構造解析、依存関係解析、共参照解析の 8 種類である。共参照解析については、信頼性の高い解析結果を得るために、同一文内、および隣接文間に存在する共参照関係のみを抽出した。

意味フレームの獲得に関しては、Kawahara らの手法 [6]を言語解析結果に対して適用した。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。

事象間因果関係知識の獲得に関しては、これらの言語解析結果に対して、Chambers ら[3]の手法に基づく知識獲得手法を適用し、事象間の関係知識データベースを獲得した。この知識獲得処理の流れを、**図 1** に示す。処理は、大きく次の二つである。(1) まず、大規模文書集合の各文書について、共参照関係となる名詞とその述語の集合 (例えば、{(her, dobj, shoot), (she, nsubj, die)}) を抽出する; (2) 次に、集合の各要素を pairwise に組み合わせ、事象間の因果関係を得る (例えば、shoot X → X die, X=(her, she))。

結果および考察

言語解析に約 3.4 万 CPU コア・時を要した。知識獲得に関しては、意味フレーム獲得に 1 万 CPU コア・時、事象間因果関係知識データベースの構築に、0.4 万 CPU コア・時を要した。この結果、約 1,700 動詞に対して約 62,000 個の意味フレーム、および約 2.3 億事例からなる事象間因果関係知識データベースが構築された。英語の知識獲得に関する研究においては、ここまで大規模な解析済みコーパスから意味フレームおよび因果関係知識を獲得した例はなく、TSUBAME の分散並列処理の利用により極めて短期間で重要な成果を上げられたといえる。

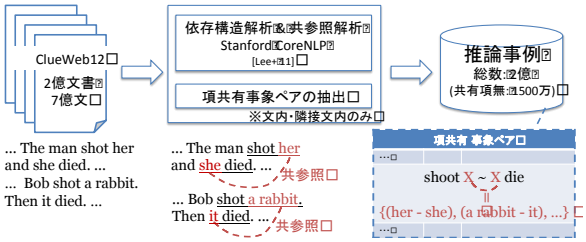

図 1 事象間因果関係知識獲得の流れ

獲得した意味フレームを評価するために、人手で PropBank のフレームを付与した SemLink コーパスを用いて評価実験を行った。その結果は F-measure で 93.9%であり、いくつかのベースライン手法を有意に上回ることを確認した。また、獲得した事象間因果関係知識の品質を評価するために、データベースから 100 事例をランダムにサンプリングし、人手による共参照解析・構文解析ミスの調査を行った。この結果、約 9 割の事例で正しく解析が行われており、比較的高品質な知識が獲得できたことを確認できた。

さらに、事象間因果関係知識の外生的な評価を行うために、知識の有効性を確認するための照応解析タスクである Winograd Schema Challenge [6] に、本知識を適用した。この結果、獲得した知識を適切な枠組みで利用することにより、解析性能が向上することを確認できた [5]。英語の言語処理の研究において、この規模の因果関係知識を利用して照応解析の研究を行い、その効果を確認した事例はほとんどなく、これも TSUBAME の恩恵を受けて可能になった成果といえる。

まとめ、今後の課題

意味フレームに関しては、述語あたり平均 36 個の意味フレームが獲得されている。獲得された意味フレームを分析したところ、PropBank のような人手で構築された意味フレームと比べて、高精細に分割しすぎている傾向があった。これに対処するために、Chinese Restaurant Process のハイパーパラメータの自動調整手法を適用することを検討している。

事象間因果関係知識に関しては、現状では、獲得した知識の一般化を一切行わず、事例レベルで知識を保

持している。しかし、この枠組は、知識適用のたびに巨大な事例集合に対して検索処理を行う必要があるため、知識の規模に対してスケーラブルでないという問題がある。これを解決するために、事象の類似性に基づいて知識をある程度一般化し、数百万程度の知識集合に落としこむことを予定している。また、引き続き ClueWeb 2012 の未解析部分への言語解析処理の適用を行い、さらに大規模な知識の獲得・言語処理への応用を行う予定である。

参考文献

[1] http://lemurproject.org/clueweb12/

[2] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu and Dan Jurafsky. Stanford's Multi-pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, 2011.

[3] Nathanael Chambers and Daniel Jurafsky. Unsupervised Learning of Narrative Event Chains. In Proceedings of ACL2008, 2008.

[4] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema Challenge. AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, 2011.

[5] 井之上直也, 杉浦純, 乾健太郎. 共参照解析のための事象間関係知識の文脈化. 言語処理学会第 20 回年次大会, pp.717-720, 2014.

[6] Daisuke Kawahara, Daniel W. Peterson, Octavian Popescu, and Martha Palmer. Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses. In Proceedings of EACL2014, 2014.