TSUBAME 共同利用　平成 26 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（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 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 種類の知識を獲得した．

1 つ目の知識は，言語を解析する際に基盤となる語彙知識である．自然文を解析するためには，語彙知識として単語と，その品詞，および付随する情報を持った辞書を整備する必要があり，新しく生まれる語彙に対応するためには常に語彙を拡張していく必要がある．このため，WEB テキスト 9653 万文書に対して文分割，形態素解析，係り受け解析を行い[村脇+ 10]の手法により語彙獲得を行った．

2 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,…}

昨年度，[Kawahara+ ACL14]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．
今年度は，その意味フレームをまとめあげることで，動詞の意味クラスの獲得に取り組んだ．また，昨年度は英語を対処に意味フレームの獲得を行ったが，今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．

結果および考察

語彙獲得

9653 万文書を解析したテキストに対して，未知語抽出処理[村脇+ 10]を行い，新規に 5688 語を獲得した．獲得した語の一部を図 1 に示す．

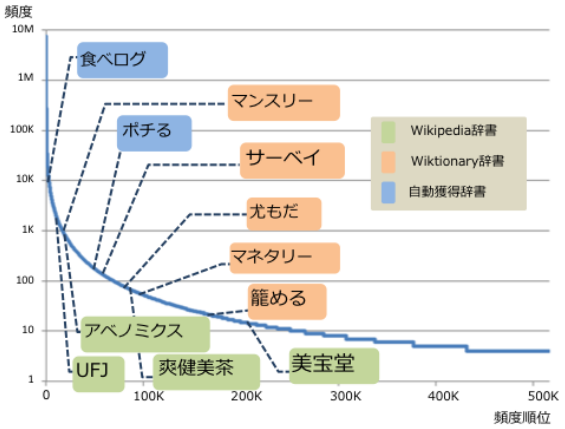


図 1 獲得した未知語例

頻度の高い順に，各単語の頻度をプロットしたグラフ．縦軸は Web1000万文書での単語の出現頻度を表し，横軸は頻度の順位を示す．比較のため Wikipedia 等の他の辞書から得られた語もプロットしている．

新しい WEB テキストから語彙を獲得することにより，”食べログ”や”ポチる”等の近年使われるようになった語彙が名詞にかぎらず獲得できている．

意味フレーム・意味クラスの獲得

昨年度構築した約 1,700 動詞に対するウェブテキスト（約 2.2 億文書，6.9 億文）から構築した約 62,000 個の意味フレーム（例を図 2 に示す）を用いて，[Kawahara+ EACL14]の手法を適用することで英語意味クラスを獲得した．具体的には意味フレーム Chinese Restaurant Process でクラスタリングすることによって 840 個の意味クラスが獲得できた．TSUBAME で並列処理することで，3 日間で構築した．獲得した意味クラスの例を図 3 に示す．

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．

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い，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．

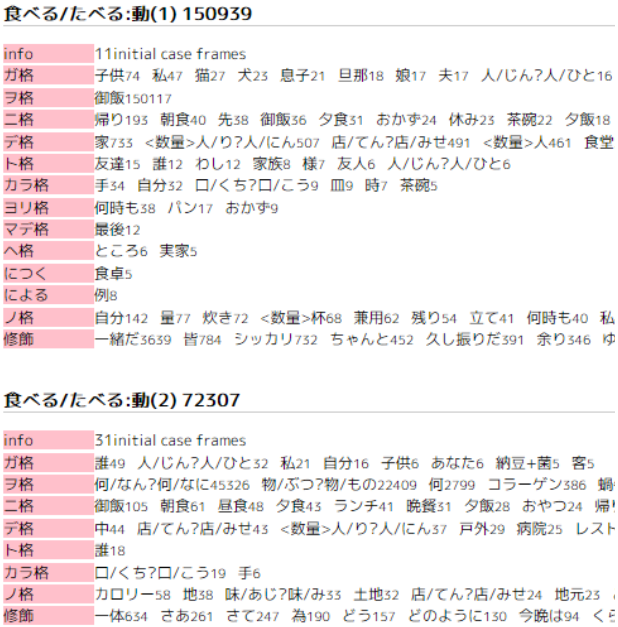
また，web から抽出した日本語 70 億文に対して，[Kawahara+ ACL14]の手法を適用することで，日本語意味フレームの構築も行った． 具体的には 36 万述語に対して Chinese Restaurant Process でクラスタリングすることで，約 1200 万個の意味フレームが獲得できた．獲得した意味フレームの例を図 4 に示す．

	slot	instances
observe:1	nsubj dobj prep_in ⋮	i:5850, we:5201, he:3796, you:3669, ... what:7091, people:2272, this:2262, ... way:254, world:204, life:194, ...
observe:2	nsubj dobj prep_in ⋮	we:11135, you:1321, i:1317, ... change:5091, difference:2719, ... study:622, case:382, cell:362, ...
observe:3	nsubj dobj prep_in ⋮	student:3921, i:2240, we:2174, ... child:2323, class:2184, student:2025, ... classroom:555, action:509, ...
accept:1	nsubj dobj prep_for ⋮	we:44833, i:6873, order:4051, ... card:28835, payment:22569, ... payment:1166, convenience:1147, ...
accept:2	nsubj dobj prep_as ⋮	i:10568, we:9300, you:5106, ... that:14180, this:12061, it:7756, ... part:1879, fact:1085, truth:926, ...
accept:3	nsubj dobj prep_as ⋮	people:7459, he:6696, we:5515, ... christ:13766, jesus:6528, it:5612, ... savior:5591, lord:597, one:469, ...

図 2 用いた英語意味フレームの例

class	semantic frames (IDs)
Class 1	rave:1, talk:1
Class 2	need:2, say:2
Class 3	smell:1, sound:1
Class 4	concentrate:1, focus:1
Class 5	express:2, inquire:62, voice:1
Class 6	revolve:1, snake:2, wrap:2
Class 7	hand:1, hand:3, hand:4
Class 8	depend:1, rely:1, rely:3
Class 9	collaborate:1, compete:2, ...
Class 10	coach:3, teach:3, teach:4
Class 11	dance:1, react:1, stick:1

図 3 獲得した英語意味クラスの例



食べる/たべる:動(1) 150939

info	11initial case frames
ガ格	子供74 私47 猫27 犬25 息子21 旦那18 娘17 夫17 人/じん?人/ひと16
ヲ格	御飯150117
ニ格	帰り193 朝食40 先38 御飯56 夕食51 おかず24 休み25 茶碗22 夕飯18
デ格	家733 <数量>人/り?人/にん507 店/てん?店/みせ491 <数量>人461 食堂
ト格	友達15 誰12 わし12 家族8 様7 友人6 人/じん?人/ひと6
カラ格	手34 自分32 口/くち?口/こう9 皿9 時7 茶碗5
ヨリ格	何時も38 パン17 おかず9
マデ格	最後12
ヘ格	ところ6 実家5
につく	食卓5
による	例8
ノ格	自分142 量77 炊き72 <数量>杯68 兼用62 残り54 立て41 何時も40 私
修飾	一緒だ3639 皆784 シッカリ732 ちゃんと452 久し振りだ591 余り346 ゆ

食べる/たべる:動(2) 72307

info	31initial case frames
ガ格	誰49 人/じん?人/ひと32 私21 自分16 子供6 あなた6 納豆+菌5 客5
ヲ格	何/なん?何/なに45326 物/ぶつ?物/もの22409 何2799 コラーゲン386 鍋
ニ格	御飯105 朝食61 昼食48 夕食43 ランチ41 晩餐31 夕飯28 おやつ24 帰り
デ格	中44 店/てん?店/みせ43 <数量>人/り?人/にん37 戸外29 病院25 レスト
ト格	誰18
カラ格	口/くち?口/こう19 手6
ノ格	カロリー58 地38 味/あじ?味/み33 土地32 店/てん?店/みせ24 地元23
修飾	一体654 さあ261 さて247 為190 どう157 どのように130 今晩は94 くら

図 4 獲得した日本語意味フレームの例

まとめ，今後の課題

語彙獲得の今後の課題はさらなる語彙の獲得と，質の向上である．また，獲得した意味クラスを用いて，構文解析や意味役割付与，語義曖昧性解消などの応用タスクに取り組むことを考えている．また，構築した日本語意味フレームを分析したところ，フレームの粒度が荒いことがわかったので，さらに精錬したいと考えている．

参考文献

[村脇+ 10] 村脇 有吾，黒橋 禎夫．形態論的制約を用いたオンライン未知語獲得．自然言語処理，Vol. 17，2010

[Kawahara+ EACL14] 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.

[Kawahara+ ACL14] 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.