研　究　報　告　書

「計算機・人の知を統合したビッグテキスト解析基盤」

研究タイプ：通常型
研究期間：　平成 26 年 10 月～平成 30 年 3 月
研 究 者：　河原　大輔

1.　研究のねらい

言語およびその記述であるテキストは人類の知の源泉である。現代において、人々はテキストからさまざまな情報を取得するとともに、テキストとして記述、情報を発信し、テキスト情報を利活用している。このように、人々は膨大なテキスト情報にアクセスできるようになったが、逆に自分にとって役に立つ情報を選別することに時間がかかることが問題になっている。このため、計算機によるテキスト情報の検索、抽出、分析の支援が非常に重要になりつつある。これらの処理を高精度に実現するためには、自然言語解析および理解が重要な役割を担う。

新聞記事、ブログ、ツイートなど、さまざまな種類の大規模テキスト集合(本稿においては「ビッグテキスト」と呼ぶ)が検索、分析の対象となりうる。しかし、これまでの自然言語解析・理解技術は主に新聞記事テキストに最適化されており、新聞記事以外のテキストに対しては高精度な解析や理解を行うことができない。その原因の一つとして、新聞記事以外のテキストには、感情や情動などによって事態に関する言外の意味(connotation)を伝えるものが多いことが挙げられる。

人々がビッグテキストを効果的、効率的に利活用することを計算機によって支援するためには、言内・言外の意味の両方について解析・分析できる基盤を構築することが必要である。本研究では、次の二つの方法を組み合わせることによって、この問題を解決することを狙う。

1. クラウド(cloud)コンピューティングによるビッグテキストからの大規模事態知識獲得
2. クラウド(crowd)ソーシングによる事態知識アノテーション

方法 1 を用いてビッグテキストから事態知識を自動獲得することによって、カバレッジの高い事態知識を得ることが可能である。言内の意味については、この方法である程度抽出することができるが、言外の意味については、テキスト中にほとんど記述されないため獲得できないという問題がある。

方法 2 は、クラウドソーシングすなわち人の集合知によって、方法 1 で得た事態知識にアノテーションを行い、修正や制約を記述することによって行う。ここで行うアノテーションは、再利用可能なものにするために、メタレベルの知識として設計する。

本研究は、大規模な言語使用から計算機的並列高速処理によって導出される知と、高度にアブストラクトされた人間の知を、相補的に統合することによって、ビッグテキストに対する自然言語解析・理解を実用的なレベルに昇華させることを狙う。

2.　研究成果

（1）概要

本研究では、まず、テキストからの自動獲得とクラウドソーシングを統合した知識獲得手法を考案し、文・文章の意味の基本的な単位である事態(「誰がどこで何をする」のような述語を

中心とした事柄)に関する知識の獲得を行った。事態知識として、事態の前後における事態参与者の素性変化と定義し、大規模テキスト集合から自動獲得した格フレームに対して、クラウドソーシングによって素性変化に関する知識を付与した。

次に、獲得した事態知識に基づく三つのアプリケーションを構築した。一つ目は日本語 Winograd Schema Challenge 解析であり、これは日本語テキスト中の照応現象の解析である。二つ目は、対話応答判定であり、あるユーザー発話に対して適切な応答を候補から選択するタスクである。三つ目は、Facebook リアクション推定であり、短い文・文章に対するリアクションを推定するタスクである。いずれのタスクにおいても、獲得した事態知識を用いることによって、有意に精度向上を達成しており、獲得した事態知識が有効に働いていることを示している。これらのアプリケーションは今後、SNS の社会問題対策や、雑談・対話ロボットなどのコア技術として利用されていくと考えられる。

これまでの自然言語解析研究は、述語項構造解析のような言内の意味理解に関するものがほとんどであったが、本研究では、言外の意味理解に向けて事態に関する知識を獲得し、その有効性を三つのアプリケーションで示した。チャレンジングな研究課題であるが、良質な事態知識が獲得できたこと、事態知識に基づくテキストからの感情推定が可能になったことが研究成果である。

（2）詳細

本研究は、研究テーマ A「テキストからの自動獲得とクラウドソーシングを統合した事態知識獲得」、研究テーマ B「獲得した事態知識に基づくアプリケーション」の二つからなる。以下では、この二つのテーマについて詳細に説明する。

研究テーマ A「テキストからの自動獲得とクラウドソーシングを統合した事態知識獲得」

文・文章の意味理解を実現するには、まず基本的単位となる事態を理解する必要がある。事態とは、「誰がどこで何をする」のような述語を中心とした事柄を意味する。本研究では、事態の意味理解を行うために、テキストからの自動獲得とクラウドソーシングを統合することによって、カバレッジが高く、かつ高品質な事態知識を獲得する手法を考案した。本手法では、まず大規模テキスト集合から格フレームを自動獲得する。格フレームとは、事態を表現した言語使用を集約し、用法ごとに整理したもので、言語理解のための基本的な辞書である。次に、獲得した格フレームに対して、クラウドソーシングを用いて事態知識を付与する。このような手法をとることによって、大規模な格フレームについて事態知識に関する情報を付与することができ、その結果、カバレッジが高く、また高品質な事態知識を獲得することができる。

事態知識として、事態の前後における事態参与者の素性変化と定義した。これは、言語理解において、事態が起こることによる効果や影響を把握することが必要条件の一つとなるからである。また、事態の事態参与者の素性変化は、後続する事態が起こる動機となりうるため、文章や対話を理解する上で重要となる事態連鎖の理解につながる。素性変化として、図 1 に示す分類をデザインした [論文 2]。本研究では、言外の意味の中心的な役割を担うと考えられる感情素性と感覚素性を対象とする。従来の多くの研究は、言内の意味を捉えることを目標に行われており、このような言外の意味を捉える研究はチャレンジングである。また、感

カテゴリ	タイプ	素性
物理素性	形	長さ，大きさ，広さ，太さ，厚さ
	色	赤さ，橙色さ，黄色さ，緑色さ，青さ，紫色さ，茶色さ，白さ，黒さ，明るさ
	感触	熱さ，かたさ，粗さ，粘り気
	におい	良さ，悪さ
	音	静かさ
	味	甘さ，酸っぱさ，苦さ，辛さ（からさ），渋さ
	密度	粗密さ
	数量	多さ
心理素性	感情	喜び，信頼，驚き，嫌悪，恐れ，悲しみ，怒り，期待
	評価	極性
感覚素性	感覚	痛み，眠気，疲れ
関係素性	関係	接触，働きかけ，力の有無，所有，社会的関係
	位置	近さ

図 1 素性変化の体系

情や感覚のように人の主観に依存しやすい素性は、専門家がアノテーションするよりも、クラウドソーシングによって多人数から常識的な知識を分散とともに獲得する方が良いと考えられる。

具体的な事態知識獲得手法は次のとおりである。

1. ビッグテキストから格フレームを自動獲得する。
2. それぞれの格フレームから代表文を生成する。
3. 代表文中の事態参与者のそれぞれについて、素性変化をクラウドソーシングで取得する。

ステップ 1 では、Kawahara らが 2014 年に提案した手法に基づいて、日本語 100 億文から大規模格フレームを自動構築した [論文 1, 3]。ステップ 2 では、それぞれの格フレームを構成するガ格・ヲ格・ニ格から代表文を生成した。ステップ 3 では、Yahoo!クラウドソーシングを用いて事態参与者および聞き手の素性変化を収集した。クラウドソーシングにおける問題提示の例を図 2 に示す。1 つの問題について 10 人のクラウドワーカーから回答を収集した。可能なかぎり高品質な回答を得るため、チェック問題および品質管理手法を利用した。チェック問題とは、あらかじめ正解を付与した簡単な問題であり、これに正解しなかったワーカーの回答は質が悪いと考えられるため、それらを削除した。また、各ワーカーの能力と各問題の難しさを考慮した品質管理手法を利用して、10 人の回答を集約して確率化した。

上記の事態知識獲得手法のポイントは二つある。一つは、格フレームを構成するすべての言語表現に対して、代表文について獲得した素性変化が成り立つことを仮定しているため、獲得した事態知識は高いカバレッジをもつようになることである。この仮定は多くの場合に成り立つことを確認している。もう一つのポイントは、格フレーム自体ではなく、格フレームから生

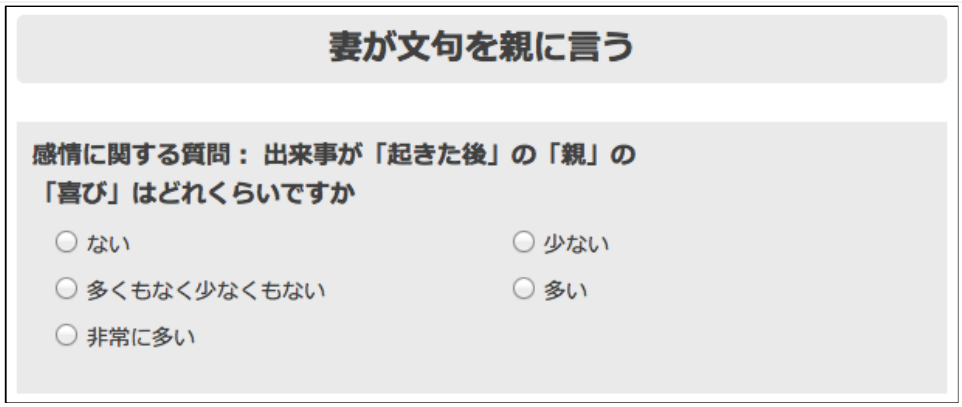

図 2 「妻が文句を親に言う」における「親」の喜びの変化を問う質問

成した代表文を介して事態知識を付与することである。このようにすることによって、格フレームの改良のために再構築を行っても、クラウドソーシングによって獲得した事態知識を適用することができるようになる。

高頻度な約 1,000 動詞・形容詞を対象として上記の手法を適用し、合計約 27 万個の素性に関して知識を獲得した。「妻が文句を親に言う」に対して獲得した素性変化を図 3 に示す。獲得した事態知識をサンプリングして主観評価したところ、おおむね良好な結果が得られた。

研究テーマ B「獲得した事態知識に基づくアプリケーション」

本研究テーマの目的は、獲得した事態知識の有効性を示すこと、および獲得した事態知識に基づくアプリケーションを構築することである。(1)日本語 Winograd Schema Challenge 解析、(2) 対話応答判定、(3) Facebook リアクション推定の三つのアプリケーションを構築した。

(1) 日本語 Winograd Schema Challenge 解析

日本語 Winograd Schema Challenge は、Winograd Schema Challenge という英語の照応解析データセットを日本語に翻訳したものである。たとえば、「赤チームは青チームを負かした。**彼ら**が最後のペナルティキックを成功させたからだ。」という文章において、照応詞「彼ら」の先行詞が「赤チーム」「青チーム」のどちらであるかを選択するタスクであり、この場合の正解は「赤チーム」である。「負かす」のガ格である「赤チーム」と「成功させた」のガ格である「彼ら」の感情変化が一致すると考えられるため、事態知識を利用することによって精度向上が見られると考える。

SVM を用いて本タスクの解析器を構築した。ベースラインとして、照応詞と先行詞候補に関するベクトル表現などを素性とした。提案モデルとしては、照応詞・先行詞が係る述語についての事態知識を用いた。その結果、ベースラインの 49.9%の精度に対して、提案モデルは 52.3%の精度を達成した [論文 4]。本タスクは意味理解を必要とする本質的に難しいチャレンジであるが、有意な精度向上を達成することができた。さらに精度を向上させるためには、本研究で対象とする事態知識だけでなく、他の常識的な知識を統合していく必要があると考えている。

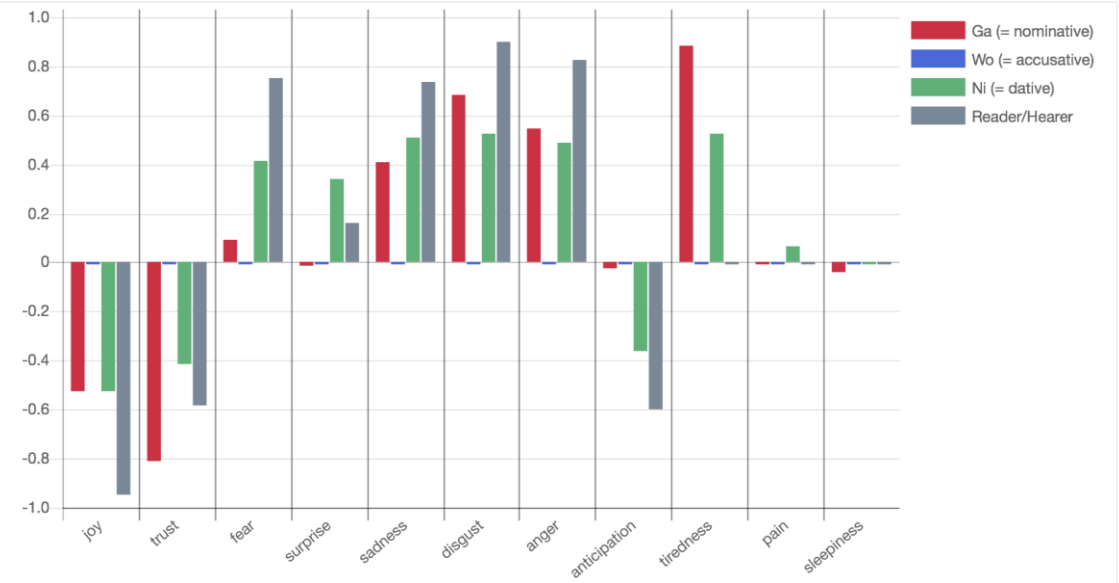

図 3 「妻が文句を親に言う」に対する素性変化

(2) 対話応答判定

獲得した事態知識を用いて、対話応答判定タスクを解くシステムを開発した。このタスクは、たとえば「選手がゴールを決めたよ」という発話に対して、応答候補 1「やったね」、応答候補 2「可哀相ですね」から、より適切な応答候補 1 を選択するタスクである。このタスクを高精度に解くことができるようになると、別途生成した応答候補をランキングして、より適切な対話応答ができるようになると考えられる。

学習に用いる対話データとして、感情変化をもつ事態文約 2,500 文に対する自然な応答をクラウドソーシングを用いて獲得した。まず、1 つの事態文に対する自然な応答の記述を 10 人のワーカーに依頼し、次に、それらの文ペアが自然かどうかを 10 人に判定してもらうという二段階の手法をとった。その結果、約 22,400 ペアの自然な対話応答を獲得した。

ニューラルネットワークを用いて本タスクの判定器を構築した。正例は上記の自然な対話応答ペアの応答文、負例はランダムサンプリングして抽出した。ベースラインとして BiLSTM (Bidirectional Long Short-Term Memory)によるモデルを構築し、提案モデルはベースラインに事態知識を統合した。ベースラインの精度が 64.2%のところ、事態知識を用いることによって、精度が 71.0%に向上することを確認した [論文 5]。

(3) Facebook リアクション推定

獲得した事態知識を用いて、短い文・文章に対するリアクションを推定するシステムを開発した。リアクションは、Facebook におけるリアクションである「いいね」「超いいね」「うけるね」「すごいね」「悲しいね」「ひどいね」の 6 つに、「リアクションしない」という選択肢を加えた計 7 つであり、この中から適切なリアクションを選択する。たとえば「娘が高熱を出した」という文に対しては「悲しいね」を選択するのが適切である。

クラウドソーシングを用いて、感情変化をもつ事態文約 2,500 文に対するリアクションを収集し、文・リアクションのペアを獲得した。1 つの文に対するリアクションは 10 人のクラウドワーカ

一から収集し、研究テーマ A と同じ品質管理手法を用いて確率化した。確率値がもっとも高いリアクションをその文に対する正解リアクションとして採用した。

ニューラルネットワークを用いて本タスクの推定器を構築した。BiLSTM を用いたベースラインモデルの精度が 66.5%のところ、事態知識を用いることによって、精度が 75.7%に大きく向上することを確認した。事態知識を用いた推定結果を詳細に分析したところ、感情に強く関係する「悲しいね」「ひどいね」についてはほとんど正答しており、事態知識が有効に機能していることを確認した。判定が難しかったのは、「いいね」と「すごいね」の区別、また「いいね」と「リアクションしない」の区別であった。さらなる精度向上には、学習データを増やすこと、リアクションしやすいシステムかどうかのパラメータ化が必要と考えられる。学習データの増加については、Facebook API を用いたリアクションデータの収集を始めており、今後このデータを利用して改良していく予定である。

3. 今後の展開

本研究で獲得した事態知識およびそのアプリケーションは、今後、社会に展開して行くことができる。たとえば、Facebook リアクション推定を発展させることによって、LINE などの SNS 上で適切なレスポンスを生成するとともに、LINE いじめのような社会問題を検知・アラートすることができる。これによって、SNS の社会問題を軽減し、コミュニケーションの好循環をうながすことを目指す。さらには、対話応答判定を発展させることによって、感情を理解することができる雑談ロボットや対話ロボットを開発することができる。ただし、対話履歴の利用や音声、画像、映像などのマルチモーダル情報の利用が必須となるため、それらの領域の研究者と協働していく必要があると考える。

また、本研究によって、テキストからの感情推定に端緒が開けたと考える。今後は、問題を明確化し、応用の幅を広げつつ、新たな研究分野として体系化していきたい。

4. 評価

（1）自己評価

チャレンジングな研究課題であったが、テキストからの自動獲得とクラウドソーシングを統合することによる事態知識の獲得、また事態知識に基づくアプリケーションを開発し、研究目的を達成することができたと考えている。これまで、言内の意味理解を中心に取り組んでいたが、本さきがけ研究において言外の意味理解にも着手することができた。今後、本さきがけ研究の経験を生かして、言内・言外の両面から、より深化した言語理解研究を進めていく予定である。

本研究成果は、SNS の社会問題への対策や雑談・対話ロボットなど、幅広い応用に利活用できると考える。2017 年現在、チャットボットやスマートスピーカーの普及が急激に進んでいるが、決められたタスクの実行や質問応答はできても、感情のような言外の意味をとらえることはほとんどできていない。今後、事態知識を強化するとともに、関連領域の研究者と連携しつつ、研究成果の社会実装を進めていきたい。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

現在までに人類が生み出した知識の多くは、自然言語のテキストの形で流通し、また蓄積されている。一方、テキストが表す知識や意味を、機械的かつ完全に読み取ることは容易ではない。本質的に難しい問題ではあるが、機械的な解析の精度を向上することで、さまざまな応用が生まれる可能性がある。

本研究では、100 億文の大規模テキストデータから大規模格フレームを構築し、個々の格フレームから代表文を生成してクラウドソーシングによる人手での処理を行い、その結果を集約することで、コンピュータと人間が力を合わせた知識獲得の試みを行った。クラウドソーシングの対象としたのは、格フレームが表す事態の前後での、事態参与者の感情を含む素性変化等である。テキストを機械的に解析するだけでは読み取ることが難しい人間の常識的な解釈の情報を加えることで、従来手法より高い精度での解析が可能なことを 3 種類のタスクで示した。国際会議等で優れた学術的成果を発表しており、その中にはトップカンファレンスでのチュートリアル論文も含まれる。

今後、クラウドソーシングの可能性のさらなる追求、感情分析を含む意味解釈の深化などの学術的な活動と具体的な応用分野の開拓を進めることを期待したい。

5. 主な研究成果リスト

(1)論文(原著論文)発表

1. Patrick Hanks, Elisabetta Jezek, Daisuke Kawahara and Octavian Popescu. Corpus Patterns for Semantic Processing. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP2015) (Tutorials), pp.12-15, 2015.
2. Tetsuaki Nakamura and Daisuke Kawahara. Constructing a Dictionary Describing Feature Changes of Arguments in Event Sentences. In Proceedings of the 4th Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pp.46-50, 2016.
3. Daniel Peterson, Jordan Boyd-Graber, Martha Palmer and Daisuke Kawahara. Leveraging VerbNet to Build Corpus-Specific Verb Clusters. In Proceedings of *SEM 2016: The Fifth Joint Conference on Lexical and Computational Semantics, pp.102-107, 2016.
4. Tetsuaki Nakamura and Daisuke Kawahara. JFCKB: Japanese Feature Change Knowledge Base. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC2018), 2018 (to appear).
5. Tetsuaki Nakamura and Daisuke Kawahara. JDCFC: A Japanese Dialogue Corpus with Feature Changes. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC2018), 2018 (to appear).

(2)特許出願

なし

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

学会発表

1. 仲村哲明, 河原大輔. 集合知を用いた事態参与者の特徴変化に関する知識の獲得. 言

語処理学会第 22 回年次大会, pp.901-904, 2016.

招待講演

1. 河原大輔. 超大規模テキスト集合からの知識獲得とそれを用いた言語理解. 東京大学大学院情報理工学系研究科コンピュータ科学専攻講演会, 2016.

受賞

1. 言語処理学会 20 周年記念論文賞「格フレーム辞書の漸次的自動構築」(河原大輔, 黒橋禎夫), 2014.
2. 平成 29 年度 科学技術分野の文部科学大臣表彰 (科学技術賞・研究部門)「日本語テキスト解析のための統合的言語資源構築に関する研究」(受賞者:黒橋禎夫, 河原大輔), 2017.

著作物

1. 李在鎬, 石黒圭, 伊集院郁子, 河原大輔, 久保圭, 小林雄一郎, 長谷部陽一郎, 樋口耕一. 文章を科学する. ひつじ書房, 2017.