研　究　報　告　書

「知識の自動獲得・構造化に基づく情報の論理構造とリスクの分析」

研究タイプ：通常型
研究期間：　平成23年10月～平成27年3月
研　究　者：　岡崎　直観

1.　研究のねらい

インターネットやモバイル通信などの情報通信技術により、一般の人がいつでも、どこでも、誰とでも繋がり、情報交換を行えるようになった。一方、情報の流通が活発になったからといって、我々は意思決定を正確に下せるようになったのであろうか？　広辞苑によると、情報とは「判断を下したり行動したりするために必要な知識」であるから、多くの情報が得られることによって、意思決定に必要な知識が増えるはずである。しかし、情報の受け手の「情報処理能力」の壁により、実際には新しい情報環境の恩恵を受けられないことが多い。

本研究では、インターネット上に鏤められている知識を獲得・集約し、一般の人の行動決定を支援するシステムを開発する。具体的には、人間の安全・安心・危険・不安に直結するドメインにおいて、ある情報を信じて意思決定を行うことが、良いとされているのか、悪いとされているのかを、計算機がインターネット上の情報を自動的に分析し、その回答の根拠情報と共に提示する技術を構築する。この技術を確立するには、自然言語処理や情報検索の研究分野で、多くの課題に取り組む必要がある。

例えば、「放射線の影響を減らすためにイソジンを摂取する」という行動に関し、その賛否を意味的に集約し、行動に関する是非を根拠情報と共に提示することを考える。この行動に関しては、「放射線の影響を減らすためにイソジンを摂取するのは間違い」など、直接的に行動を否定する言及もあった。また、「イソジンを飲むと嘔吐する」など、危険な事態を提示することで行動を間接的に否定する言及もある。ここで例示した情報提供者のソーシャルな関係が近ければ、発信された情報間の意味的な近さを近似できる。また、先に挙げた行動に対して「イソジンを飲むと嘔吐する」の命題が反論していることを認識するには、「放射線は毒物である」「毒物の影響を減らす→安全である」「嘔吐する→危険である」という知識を組み合わせた上で、前者の行為が安全を目的としているのに、後者の命題が同じ行為の危険性を主張しているため、これらが対立関係にあると認識できる。

本研究では、情報間の意味的な関係の高精度な解析と、意思決定支援システムの開発を目的とし、以下の研究項目に取り組む。

A.　大規模なテキストからの常識的な知識の獲得
B.　情報間の意味的関係の解析
C.　意思決定支援システムの構築

2.　研究成果

（1）概要

研究項目A「大規模なテキストからの常識的な知識の獲得」では、（弱）教師あり学習に基づ

く手法（文献5，6，9）、数量に関する常識的な知識の獲得（文献7）、因果関係の知識獲得を進めた（文献10）。また、分散並列処理、オンライン学習、近似頻度計測などの技術を応用し、数億ページを超える大規模な言語データからの知識獲得に成功し、先行研究との優位性を実証した。これらは、学習に用いるコーパスや訓練事例のドメインに依存しない手法となっており、本研究で用いている東日本大震災後の言語データだけでなく、大規模なWebコーパス（150億文）など、一般ドメインのコーパスからも知識を獲得できる。

研究項目B「情報間の意味的関係の解析」では、東日本大震災後のツイートを用い、信ぴょう性の低い情報の検出や情報の構造化に関する研究を進めた。東日本大震災では、「イソジンを飲むと放射線予防になる」「関東の若者に鼻血症状が見られる」などの誤情報が拡散し、その社会的な影響は今もなお根強い。多くの人から訂正されている情報（＝信憑性の低い情報）を網羅的に収集・集約する手法を開発した。震災時の人々の実際の会話データから流言・誤情報を見つけ出し、その拡散・収束過程を調査する試みは、本研究が初であり、AMT 2014 Best Paper Award および言語処理学会 2013 年度論文賞を受賞した（文献3，4）。さらに、福島の桃の購買に関する風評の実態をツイートから分析するため、評判情報分析とネットワーク分析を組み合わせ、人々の意見の意味的な関係を解析した。誤情報や風評の拡散の温床とされたソーシャル・ネットワークで、風評を減らすためのリスクコミュニケーションに関する知見が得られ、リスク研究学会で優秀発表論文賞を受賞した（文献2）。また、本研究で得られた知見の一部は NHK スペシャル「震災ビッグデータ2」で放送された。

研究項目C「意思決定支援システムの構築」では、信憑性の低い情報を常時モニタリングするシステムを開発した。このシステムでは、情報間の意味的関係の解析（研究項目B）に基づき、信憑性の低い情報を支持する情報、反論を与える情報に分類して情報を提示できる。さらに、情報間の意味的な関係をブラウザ上で可視化するシステムを開発した。最終的な意思決定を行う利用者に対して、背後にある情報を可視化することで、情報の善し悪しや意思決定を支援する環境を提供した。

（2）詳細

研究テーマ A 「大規模なテキストからの常識的な知識の獲得」

- Wikipedia から抽出したエンティティの知識をシードとして、ウェブページ1千万文書からエンティティの知識インスタンスを自動的に拡充する手法（集合拡張）を開発した。本研究では、 Wikipedia から自動獲得したカテゴリの上位下位関係を制約として活用することで、意味ドリフトを軽減する効果が得られることを示した（文献6）
- リスク情報に多く含まれる数量の意味解析に向けて、数量の大小に関する知識の自動獲得手法、およびその知識を用いた数量の大小判定手法を提案した。本研究は、人間の主観を客観に一般化するという難しいタスクであるが、抽出された知識が人間の感覚と高い相関を示すことを実験的に確認した（文献8）。
- 固有表現抽出器（エンティティの意味クラスの推定）の性能を改善するため、ラベル情報に階層構造を持たせる手法（文献5）、エンティティの周辺文脈をマイニングする手法（文献9）を提案した。
- 因果関係知識の獲得に取り組んだ。新聞記事などでは、「UnitedHealth buys Pacificare.」

という文の後に、「The acquisition gives UnitedHealth operations in Nevada.」という文が出現することがある。このようなイベント共参照を含む文章に、大規模なテキストから自動的に獲得した知識を適用し、「buy(A, B)∧headquartered-in(B,X) ⇒ operate-in(A, X)」という一般化された因果関係知識を獲得する手法を考案した（文献10）。

- 世の中に存在するあらゆる関係（is-a, born-in など）に関し、それらの関係にあるインスタンス（エンティティの組）をテキストから自動的に獲得する手法（教師なし関係抽出）を開発した。近似頻度計算と次元圧縮を利用した手法およびシステムを開発し、数十億文規模の超大規模コーパスから良質なパターンの意味表現を現実的な時間で学習できた。

研究テーマＢ 「情報間の意味的関係の解析」

- 東日本大震災時に拡散した誤情報の網羅的な収集: 「〇〇というのはデマ」「〇〇という事実は無い」など、誤情報を訂正する表現（以下、訂正パターン）に着目し、誤情報を自動的に収集する手法を提案した（図１左）。収集した誤情報のそれぞれに対し、ソーシャルメディア上での拡散・収束状況、訂正情報の拡散・収束状況を分析した（図１右）。評価実験では、まとめサイトから取り出した誤情報のリストを正解データと見なし、提案手法の精度や網羅性に関して議論した（文献1，4）。
- ツイートに対する同意・反論・疑問などの論述関係を整理するため、返信や非公式リツイートの関係にあるツイート（返信・引用ツイート）によって表明される投稿者の「同意」「反論」「疑問」などの態度を推定する手法を提案した（文献7）。
- 福島県産の桃に対する立場（肯定的か否定的か）と、それぞれの立場における議論を分析した。ツイート本文の極性分析とリツイートネットワークの分析により、福島の桃に関する議論は肯定派・否定派に分断されており、立場を途中で変えたり、立場を横断して健全な議論をした形跡が見られなかったことが分かった（図２）。分析から、産地ではなく精密な情報で自ら判断したいという意見を発見し、検査結果のデータを積極的に開示したり、議論の見通しを良くすることで、両派の溝を解消する可能性が見えた（文献2）。

トピック	スコア	命題
田尻智さん	18.7	海外で、ポケモンの生みの親の田尻智さんが亡くなった
コスモ石油千葉製油所	11.8	@市原市のコスモ石油千葉製油所LPGタンクの爆発により、千葉県、近隣圏に在住の方に有害物質が雨などと一緒に飛散する
女性暴行	10.4	@「阪神大震災の際には女性暴行が増えた」
有毒物質	8.6	コンビナート火災に関し『有毒物質が発生し、雨に混じって降ってくるので肌をさらさないように』
尾田栄一郎先生	5.3	尾田栄一郎先生が15億円を寄付
円資産	5.2	池田経済産業副大臣は、「日本企業が円資産を買い戻している
空中投下	3.0	『物資の空中投下が日本で認められていない』
サーバルーム	1.9	サーバルーム、ラックで腹部を怪我した
イソジン	1.6	被爆予防にイソジンを飲むと良い
茨城県知事	1.2	茨城県知事災害派遣要請出してない
食塩	1.0	「食塩にはヨウ素が含まれており被曝の防止に効果がある」

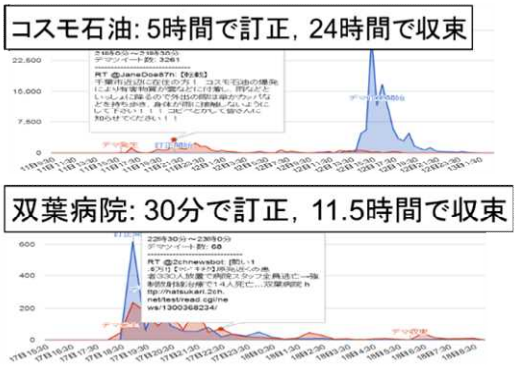


図１． ソーシャルメディアから抽出した誤情報（左）と拡散収束過程（右）

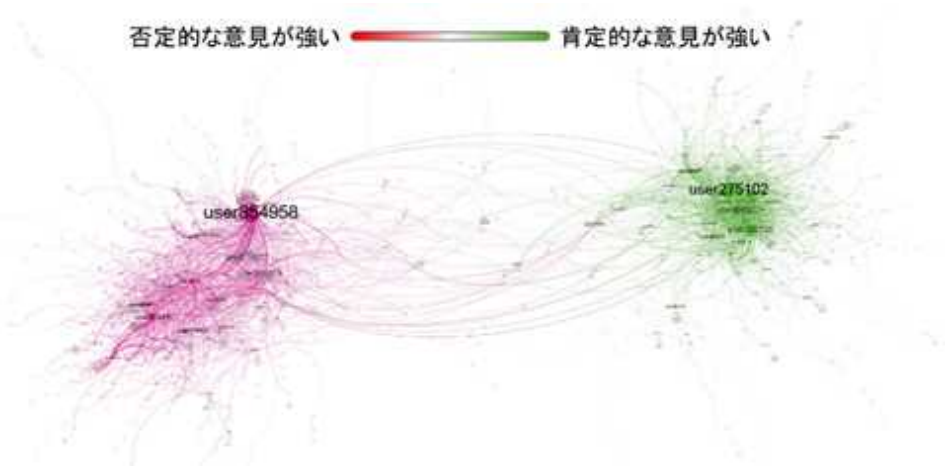
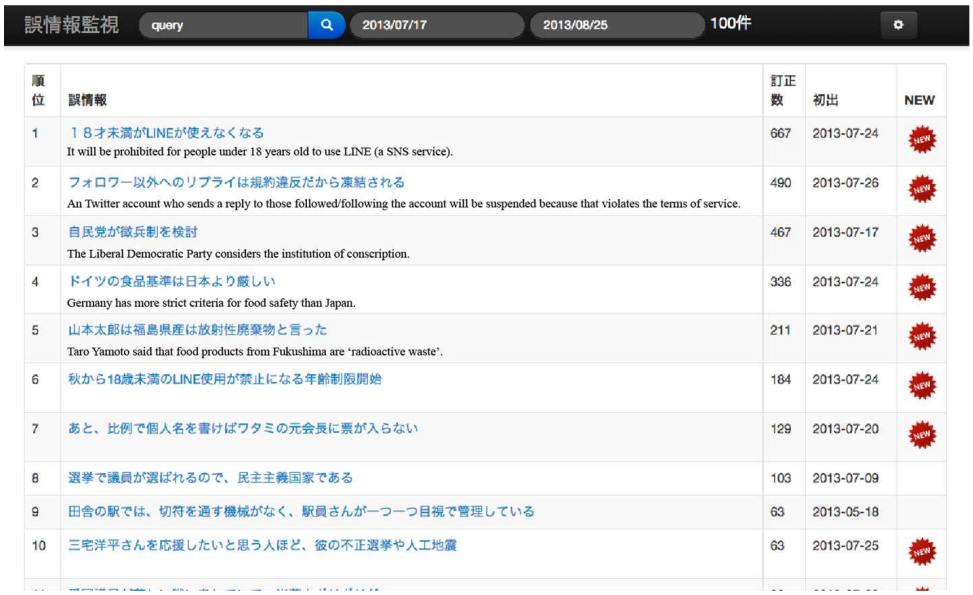


図2. 福島の桃に関するツイートの意見分析およびソーシャル・ネットワーク分析

研究テーマ C 「意思決定支援システムの構築」

- 研究テーマBの成果に基づき、信憑性の低い情報をリアルタイムでモニタリングするシステムを開発した（図3）。このシステムでは、信憑性の低い情報を支持する情報、反論を与える情報に分類して情報を提示できる。
- 情報間の意味的な関係をブラウザ上で可視化するシステムを開発した。



順位	誤情報	訂正数	初出	NEW
1	１８才未満がLINEが使えなくなる It will be prohibited for people under 18 years old to use LINE (a SNS service).	667	2013-07-24	NEW
2	フォロワー以外へのリプライは規約違反だから凍結される An Twitter account who sends a reply to those followed/following the account will be suspended because that violates the terms of service.	490	2013-07-26	NEW
3	自民党が徴兵制を検討 The Liberal Democratic Party considers the institution of conscription.	467	2013-07-17	NEW
4	ドイツの食品基準は日本より厳しい Germany has more strict criteria for food safety than Japan.	336	2013-07-24	NEW
5	山本太郎は福島県産は放射性廃棄物と言った Taro Yamoto said that food products from Fukushima are 'radioactive waste'.	211	2013-07-21	NEW
6	秋から18歳未満のLINE使用が禁止になる年齢制限開始	184	2013-07-24	NEW
7	あと、比例で個人名を書けばワタミの元会長に票が入らない	129	2013-07-20	NEW
8	選挙で議員が選ばれるので、民主主義国家である	103	2013-07-09	
9	田舎の駅では、切符を通す機械がなく、駅員さんが一つ一つ目視で管理している	63	2013-05-18	
10	三宅洋平さんを応援したいと思う人ほど、彼の不正選挙や人工地震	63	2013-07-25	NEW

図3. リアルタイムで誤情報を監視・閲覧するシステム

3. 今後の展開

計算機が利用できる大規模な知識ベースの獲得に関しては、人手で整備したもの

（WordNet やUMLSなど）、Wikipediaのような半構造言語資源から抽出したもの（YAGO、DBPedia、Freebase など）、研究テーマAのように計算機が自動的に獲得したものが使われるようになってきた。大規模な知識ベースが整備されつつあるためか、近年では知識ベースに基づいて意味的な推論を行う手法が再興の兆しを見せている。したがって、テキスト中で言及されたエンティティや関係の曖昧性を解消しながら、知識ベースに対応付ける「グラウンディング問題」が、今後の重要な研究テーマになると考えている。本プロジェクトで構築した手法や言語資源は、テキストと大規模知識ベースとのグラウンディング問題の研究を進める上での基盤技術となるであろう。

研究テーマBの成果は、デモシステム等で社会に還元できる形態となっている。さらに、本さきがけプロジェクトの成果の出口として、データ駆動型ジャーナリズムに着目し、新聞社やテレビ局への技術提供・取材協力を行った（全国紙での紙面掲載 5 件、地方紙での紙面掲載 1 件、テレビ局への取材協力 1 件）。このように、研究テーマBの社会実装への展開はすでに始まっている。今後は、個別の分析案件ごとに開発していた分析手法の中で共通性の高い部分（例えば事態の成立・不成立を推定する問題や事象の主体を推定する問題）を切り出し、応用の幅を広げながら学術研究としての体系化していきたい。

4. 評価

（1）自己評価

（研究者）

研究項目Aでは、ウェブ数億ページを入手可能な最大規模のコーパスと想定し、その規模に耐えうる知識獲得手法を探求した。（弱）教師あり学習、教師なし学習のそれぞれの設定において、他に類を見ない超大規模コーパスから知識を自動獲得するための計算基盤インフラと、その規模に耐えうる手法・ツールを構築し、数千万規模の関係インスタンスを含む知識ベースを現実的な処理時間（数日程度）で構築するという目標を達成できた。東日本大震災時の言語データにも本手法を適用し、関係インスタンスを獲得し、研究項目Bの研究の精度向上に貢献した。本研究項目は、2013 年頃から流行している分散表現の学習や構成的意味論の研究と関連が深く、エンティティや関係知識の意味計算モデルへの発展を見込んでいる。

研究項目Bでは、情報の信憑性や論述関係を実データで分析し、計算機で情報を整理する技術の開発しながら、本さきがけプロジェクトの出口を探った。大規模災害時に情報の混乱が引き起こされた事例と、情報の混乱が発生するメカニズムを、震災後1週間の全量（約2億）ツイートデータから解明するという、従来の言語処理の枠にとらわれないフィールドに挑んだ。データ利用規約上の制約から、国内ジャーナルへの投稿となったが、本研究成果は言語処理学会の論文賞を受賞した。また、本成果を発展させた研究も、Active Media Technology に関する国際会議においてBest Paper Award を受賞した。本研究の成果をリスクコミュニケーションに応用した研究では、自分の専門外であるリスク研究学会において研究の価値が認められ、優秀発表論文賞を受賞した。当初の目標以上の研究成果を生み出したと考えている。

研究項目Cで本さきがけ研究の成果をデモシステムとして世の中に発信したことをきっかけとして、マスメディアと共同研究の体制を組めたことも大きな収穫であった。言語というメディアを通じて個人や社会を分析し、一般向けのサービスを開発するという貴重な経験を積むことができた。一方で、言語よりも言語を利用する「人」に一歩踏み込んだ解析が必要であり、現状の言語処理で達成できている部分、実社会から自然言語処理に向けられている期待とのギャ

ップの深さを体験した。本さきがけ研究の経験を活かし、文全体や文を超えた文脈の理解や、言語を発する人や社会の理解に向けた研究を継続していきたい。

（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）。

（研究総括）

本研究では、震災後の膨大な Web データを対象に、自然言語処理技術を発展的に用いて、大模なデータから因果関係知識を獲得する基礎技術を生み出した。そうした技術を適用し、ツイートデータを対象にデマなどの誤情報を収集し、投稿者の「同意」「反論」「疑問」などの態度を推定し、福島県の農作物に対する風評分析を行っている。さらに、信憑性の低い情報をリアルタイムでモニタリングするシステムを開発するなど、研究の技術的・社会的意義は極めて大きい。その成果は、国際会議や論文誌で受賞しているだけでなく、NHK スペシャルで放映されるなど、学術的にも社会的にも高く評価されている。岡崎氏は、震災直後に東北大学に採用された。さきがけの申請時には、研究の対象を震災データに絞ってはいなかったが、異動後、社会的要請に応え、研究の方向を見直し、自らの技術を用いて全力で震災データに取り組んでいった。若手研究者であるにも関わらず、多くの会議から講演を招待されていることも、岡崎氏の研究成果や研究姿勢に対する高い評価を表している。

5. 主な研究成果リスト

（1）論文（原著論文）発表

1. Naoaki Okazaki, Keita Nabeshima, Kento Watanabe, Junta Mizuno, Kentaro Inui. Extracting and Aggregating False Information from Microblogs. Proceedings of the Workshop on Language Processing and Crisis Information 2013, IJCNLP 2013, pp. 36—43, Nagoya, Japan, October 2013.
2. 岡崎直観, 佐々木彬, 乾健太郎, 阿部 博史, 石田 望. ツイッター分析に基づく福島県産桃に対する風評の実態解明とその対策. 第 26 回日本リスク研究学会年次大会, 2013 年 11 月. （優秀発表論文賞）
3. Keita Nabeshima, Junta Mizuno, Naoaki Okazaki and Kentaro Inui. Mining False Information on Twitter for a Major Disaster Situation. In Proceedings of the 2014 International Conference on Active Media Technology (AMT2014), pp.99-106, August, 2014. (Best Paper Award)
4. 鍋島啓太, 渡邉研斗, 水野淳太, 岡崎直観, 乾健太郎. 訂正パターンに基づく誤情報の収集と拡散状況の分析. 自然言語処理, Vol. 20, No. 3, pp. 461—484, 2013 年 6 月. （**言語処理学会 2013 年度論文賞**）
5. Han-Cheol Cho, Naoaki Okazaki, Makoto Miwa, Jun'ichi Tsujii. Named entity recognition with multiple segment representations. Information Processing & Management, Vol. 49, No. 4, pp. 954—965, July 2013.

（2）特許出願
研究期間累積件数：0件

（2）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）
受賞

- 石田實記念財団研究奨励賞．岡崎直観．大規模データに基づく自然言語処理とその応用に関する研究．2014 年 11 月 14 日
- AMT2014 Best Paper Award．Keita Nabeshima, Junta Mizuno, Naoaki Okazaki and Kentaro Inui. Mining False Information on Twitter for a Major Disaster Situation. 13 August 2014.
- 言語処理学会第 20 回年次大会 優秀賞．島岡聖世，村岡雅康，山本風人，渡邉陽太郎，岡崎直観，乾健太郎．ガウス分布による単語と句の意味の分布的表現．2014 年 3 月 20 日．
- 言語処理学会 2013 年度論文賞．鍋島啓太，渡邉研斗，水野淳太，岡崎直観，乾健太郎．訂正パターンに基づく誤情報の収集と拡散状況の分析．2014 年 3 月 19 日
- 第 26 回日本リスク研究学会年次大会 優秀発表論文賞．岡崎直観，佐々木彬，乾健太郎，阿部博史，石田望．ツイッター分析に基づく福島県産桃に対する風評の実態解明とその対策．2013 年 11 月 16 日．

招待講演

- 岡崎直観．EMNLP 2012 参加報告．NLP 若手の会 第 7 回シンポジウム，2012 年 9 月．
- 岡崎直観，吉永直樹，工藤拓．言語処理研究におけるソフトウェアの開発と公開．言語処理学会第 19 回年次大会(NLP2013)・チュートリアル，2013 年 3 月．
- 岡崎直観．大規模データを利活用する自然言語処理．東京大学大学院情報理工学研究科コンピュータ科学専攻講演会，2013 年 10 月．
- 岡崎直観．ACL 2013 参加報告 ～ 会議概要と応用技術 ～．第3回テキストマイニング・シンポジウム（信学技報），2013 年 10 月．
- 岡崎直観，津田大介，奥山晶二郎．ジャーナリズム再考：ビッグデータとテクノロジー．ソーシャルメディアウィーク東京 2014，浜離宮朝日ホール（東京都），2014 年 2 月．
- 岡崎直観．言葉を操るコンピュータの最前線 － 自然言語処理の基礎からビッグデータへの応用まで．QuickSolution イノベーション・フォーラム 2014，2014 年 6 月．
- 岡崎直観．言語処理がつなぐビッグデータと社会．言語処理学会 20 周年記念シンポジウム，2014 年 10 月．
- 岡崎直観．大規模言語データに基づく自然言語処理とその応用．第 17 回情報論的学習理論ワークショップ(IBIS2014)，企画セッション 4: 機械学習のウェブデータおよびマルチメディア活用，2014 年 11 月．
- 岡崎直観．災害と言語処理（仮題）．情報処理学会コンピュータビジョンとイメージメディア研究会（CVIM），2015 年 3 月（予定）．
- 岡崎直観．単語や句の分散表現の学習とその応用（仮題）．2015 年度 人工知能学会全

国大会（第 29 回），OS-1 意味と理解のコンピューティング，2015 年 6 月（予定）．

原著論文

6. 高瀬翔，岡崎直観，乾健太郎．カテゴリ間の兄弟関係を活用した集合拡張．自然言語処理，Vol. 20, No. 2, pp. 273—296, 2013 年 6 月．
7. 大和田裕亮，水野淳太，岡崎直観，乾健太郎，石塚満．返信・非公式リツイートに基づくツイート空間の論述構造解析．自然言語処理，Vol. 20, No. 3, pp. 423—460, 2013 年 6 月．

国際会議論文

8. Katsuma Narisawa, Yotaro Watanabe, Junta Mizuno, Naoaki Okazaki, Kentaro Inui. Is a 204 cm Man Tall or Small? Acquisition of Numerical Common Sense from the Web. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 382—391, Sofia, Bulgaria, August 2013.
9. Han-Cheol Cho, Naoaki Okazaki, Kentaro Inui. Inducing Context Gazetteers from Encyclopedic Database for Named Entity Recognition. Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2013), pp. 378—389, Gold Coast, Australia, April 2013.
10. Shohei Tanaka, Naoaki Okazaki, Mitsuru Ishizuka. Acquiring and Generalizing Causal Inference Rules from Deverbal Noun Constructions. Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Posters, pp. 1209—1218, Mumbai, India, December 2012.

国内口頭発表

11. 岡崎直観，成澤克麻，乾健太郎．Web 文書からの人の安全・危険に関わる情報の抽出．言語処理学会第 18 回年次大会(NLP2012), pp. 895—898, 2012 年 3 月．
12. 岡崎直観，乾健太郎，水野淳太，鍋島啓太，渡邉研人，大和田裕亮．ツイッターデータの意味的解析による災害情報拡散の分析．東日本ビッグデータワークショップ最終成果報告会，2012 年 10 月．
13. 岡崎直観，乾健太郎．Word2vec の並列実行時の学習速度の改善．情報処理学会第 217 回自然言語処理研究会，2014-NL-217(8)，2014 年 6 月．

著書

14. 岡崎直観，鍋島啓太，乾健太郎．言語処理による分析－日本栄養士会活動報告の分析．日本栄養士会雑誌，Vol.55, No.12, pp.6—8, 2012 年 12 月．