様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２１　日現在

機関番号：　１２６０８
研究種目：　若手研究(A)
研究期間：　2015～2017
課題番号：　１５Ｈ０５３１８

研究課題名（和文）分散表現と構成的意味計算に基づくテキストと知識ベースの頑健なグランディング

研究課題名（英文）Grounding text and knowledge base by using distributed representation and their compositions

研究代表者

岡崎　直観（Okazaki, Naoaki）

東京工業大学・情報理工学院・教授

研究者番号：５０６０１１１８

交付決定額（研究期間全体）：（直接経費）　　12,970,000円

研究成果の概要（和文）：人間の言葉を理解し，推論する計算機の実現には，常識的な知識を計算機が蓄積・活用する機構が不可欠である．このような枠組みで知的な処理を実現するには、大規模な知識ベースの獲得，テキストと知識ベースのグランディング，推論という３つの課題がある．本研究では，知識ベースの事例にグラウンドされたコーパスを構築した．関係パタンを知識ベースにグランディングするため，関係パタンの項となりうる名詞の意味表現（分散表現）に基づき，構成的意味論に基づく新しい意味計算モデルと曖昧性解消の手法を考案した．これらの研究成果は，単体での性能評価に加え，質問応答や賛否分類などの言語処理の実タスクに応用し，その効果を実証した．

研究成果の概要（英文）：In order to realize computers that can understand and infer on natural languages, a mechanism for collecting and utilizing the commonsense knowledge is essential. We need address three research topics, acquisition of a large-scale knowledge base, grounding a text with the knowledge base, and an inference method on the grounded text. In this project, we built corpora where a text is grounded on instances of a knowledge base. In order to associate relational patterns into a knowledge base, we proposed a novel method for composing a vector of a relation pattern from its constituent words and for disambiguating the sense of a relation pattern. We demonstrated the effectiveness of these studies not only by the experiments on individual tasks but also by applying downstream tasks such as question answering and stance detection.

研究分野：　自然言語処理

キーワード：　自然言語処理　人工知能　深層学習　分散表現　知識ベース

１．研究開始当初の背景

人間の言葉を理解し，推論する計算機の実現には，常識的な知識を計算機が蓄積・活用する機構が不可欠である．このような枠組みで知的な処理を実現するには、大規模な知識ベースの獲得，テキストと知識ベースのグランディング，推論という３つの大きな課題がある．知識ベースに関しては，Wikipedia, WordNet, YAGO, DBPedia, Freebase などが登場し，計算機が大規模な知識ベースを利用する環境が整いつつあった．その次のステップとして，推論を行いたいテキストを知識ベースの事例に対応づける問題，すなわちテキストと知識ベースのグランディング問題がある．言葉を理解・推論する計算機の実現に向けて，テキストと知識ベースとのグランディング問題がボトルネックとなっていた．

２．研究の目的

本研究ではテキストを知識ベースのインスタンスにグランディングする解析器を開発し，その実タスクでの応用を検証する．本研究では４つの研究項目に取り組む．

①関係パタンの構成的な意味計算: 知識ベースの関係インスタンスは，テキスト中では「X が監督した映画 Y」のような関係パタンで表現される．テキストの多様な表現に対応するには，「X が監督した映画 Y」という関係パタンが監督―映画関係を表し，「X が監督を務めた映画 Y」「X が監督の映画 Y」なども同一の関係であることを認識する必要がある．このような関係パタンは無数に生成できるため，すべてのバリエーションを事前に列挙し，各パタンの意味を計算しておくことは非現実的である．また，関係パタンが長くなるほどコーパス中の出現頻度が低下するため，データ・スパースネス問題が発生する．そこで，本研究項目では，関係パタンの語の構成性に基づく意味計算モデルを探求し，多様なテキスト表現のグランディングを可能にする．

②関係パタンの曖昧性解消: 「黒澤明の『羅生門』」のように，「X の Y」という関係パタンは監督―映画関係に言及する．一方で，「X の Y」は「マイクロソフトのビル・ゲイツ」のように，組織と創設者の関係を表すなど，曖昧性の高いパタンである．本研究項目では，X や Y を埋める実体・概念（エンティティ）の意味表現を学習し，研究項目①の意味計算モデルを統合することで，関係パタンから知識ベースのインスタンスを頑健に特定する解析器（曖昧性解消器）を構築する．

③グランディングのためのコーパス構築: テキストに知識ベースのインスタンスを付与したコーパスを構築し，研究項目①と②の訓練データ・評価データとして用いる．

④実タスクにおける検証: 研究項目①②③で開発した解析器を質問応答や含意関係認識などタスクに組み込み，本研究の貢献を実証する．

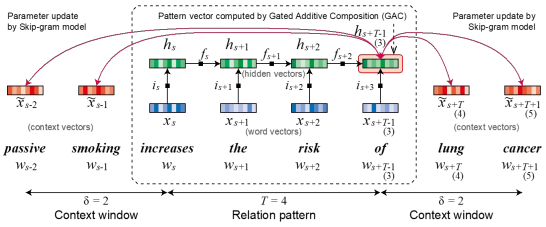
３．研究の方法

知識ベースの事例にグラウンドされたコーパスを構築する（研究項目③）ため，この作業を支援するツールの開発を進め，コーパスを構築した．関係パタンを知識ベースにグランディングするため，関係パタンの項となりうる名詞の意味表現（分散表現）に基づいて曖昧性解消を行う研究（研究項目②）を行った．また，関係を表す言語パタンのデータ・スパースネス問題に対応するため，構成的意味論に基づく新しい意味計算モデルを提案した（研究項目①）．研究項目①②③で開発した解析器単体の評価に加え，質問応答や含意関係認識などの言語処理の実タスクにおける貢献を実験的に調べた（研究項目④）．

４．研究成果

①　関係パタンの構成的な意味計算

「increase the risk of」のような関係パタンの意味を，「increase」「the」「risk」「of」といった構成単語の意味から合成する新手法を提案した．この手法では，構成単語の意味を単語の分散表現（意味ベクトル）として表現し，その合成関数を深層ニューラルネットワークで学習する．深層ニューラルネットワーク構造として，現在位置の単語やその前後の単語の意味の影響を適応的にコントロールする Gated Additive Composition を提案した（図１）[雑誌論文 2，学会発表 6]．



Input gate: $i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1})$

Forget gate: $f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1})$

Gated composition: $h_t = g(f_t \odot h_{t-1} + i_t \odot x_t)$

図１　Gated Additive Composition

この提案手法の評価実験を行ったところ，単語ベクトルの単純平均，長期短期記憶（LSTM）や Gated Recurrent Unit（GRU）による合成よりも高い性能を示した．

②　関係パタンの曖昧性解消

研究項目①の成果として得られた関係パタンの合成関数を用い，任意の関係パタンの意味ベクトルを計算し，SemEval 2010 Task 8 の関係曖昧性解消タスクに適用し，当時としては世界最高性能と同等の性能が得られる

ことを実証した．この研究成果は，自然言語処理分野の最難関国際会議 ACL に採択されるなど，国内外で高い評価を得ている．

③グランディングのためのコーパス構築

実体・概念を知識ベースにグランディングするコーパスとして，BCCWJ の新聞記事コーパスに含まれる固有表現を Wikipedia 記事に対応付けた正解データを構築した（図２）．このコーパスでは，340 件の新聞記事に出現する約 26,000 件の実体･概念への言及の中で，対応する Wikipedia 記事が存在する約 22,000 件に対し，その Wikipedia 記事の ID を付与した．このコーパスを使った評価実験では，約 8 割の正解率でテキスト中の実体・概念の言及から Wikipedia 記事に対応付けることができることを示した [学会発表 7]．

図２　日本語 Wikification コーパス

また，因果関係知識をグランディングするため，Wikipedia 記事 1,494 件に促進・抑制関係を付与したコーパスを構築した（図３）．このコーパスは，社会問題，災害，病気，技術革新，政策，金融，エネルギー技術，生体物質，栄養素 の 5 件のカテゴリと，そのサブカテゴリ，サブ・サブカテゴリに属する Wikipedia 記事を対象として，クラウドソーシングを用いて 1 つの記事につき 10 人のアノテーション結果を収集した．コーパスへの関係知識付与をクラウドソーシングで完結させるため，アノテーションツールである brat を改変した．このコーパスを学習データとして用い，Wikipedia 記事と因果関係知識を結び付ける手法を構築した [学会発表 2]．

- 社会問題，災害，病気や症状，イノベーション，政策，金融，エネルギー技術，生体分子，栄養の9つのカテゴリに属する1,494記事について，促進・抑制などの因果関係知識を付与したコーパスをクラウドソーシングで構築
- コーパスを深層学習の訓練事例として用い，因果関係の自動抽出器を構築
- 構築したコーパスを公開中: http://www.cl.ecei.tohoku.ac.jp/wikipedia_pro_sup/

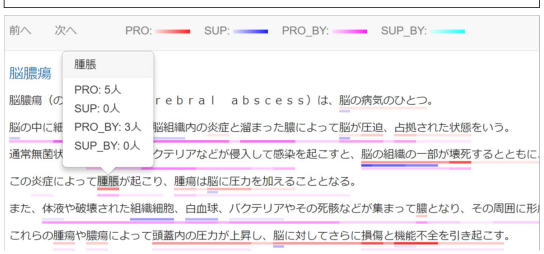


図３　Wikipedia からの因果関係知識獲得

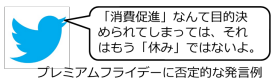
④実タスクにおける検証

実タスクとして、質問応答と賛否分析に取り組んだ。質問応答では、与えられた質問文に対して、その回答の候補を含む Wikipedia 記事を対応付け、Wikipedia 記事の読解により答えを出力するアプローチに取り組んだ。早押しクイズの約 12,000 件の質問に対し、約 46,000 件の Wikipedia 記事との対応を明らかにしたデータセットを作成した。このデータセットを用いることで、質問の答えが Wikipedia に見当たらない場合、という新しい問題設定を提起し、その状況を考慮することの効果を検証した。

研究項目②と③の研究成果である Wikipedia の記事の中で言及される促進・抑制などの因果関係を活用し，賛否分類のタスクに取り組んだ。14,000 件のソーシャルメディアの投稿を収集し、大阪都構想やプレミアムフライデーなどの 7 話題に対する賛否を付与したデータを構築した（図４）。このデータを分析し、テキストと関係知識のグランディングの必要性を調査したところ、約 1/3 の投稿において対応付けが必要であることが分かり、本研究の重要性が示された。

- 7トピックに言及しているツイートを収集し，トピックに賛成／反対／中立なのかを人手で付与した
- 40.2%の意見表明の賛否を推定するには，トピックに関する知識が必要であることが分かった

トピック	賛成	反対	中立
大阪都構想	239	259	380
安保法案	168	352	262
プレミアムフライデー	153	744	218
TPP	53	802	230
原発	47	783	202
集団的自衛権	160	468	196
共謀罪	86	592	308
合計	906	4000	1795



プレミアムフライデーに否定的な発言例

賛否分類において必要な知識	%	意見の例
一般的な賛否表明	56.3	原発は絶対必要（原発: 賛成）
トピックの因果関係 (∈ Wikipedia)	26.3	関税が機能するべき（TPP: 反対）
トピックの因果関係 (∉ Wikipedia)	13.9	遺伝子組み換え食品とか心配（TPP: 反対）
その他の知識	2.5	治安維持法を復活させたいのか（安保法案: 反対）

図４　賛否分類コーパスの構築

ソーシャルメディアのテキストから賛否を分類するため、「○○というトピックに賛成する人は××というトピックにも賛成する」という形の知識（トピック間選好知識）を抽出する手法を提案した（図５）。この手法は，情報推薦の分野でよく用いられる行列因子分解に基づくものである [学会発表 3]．

- 個人の意見表明から政治的スペクトル（イデオロギー）を抽出・分析
- 商品推薦の手法に着想を得て「○○に賛成する人は××にも賛成する」を予測c

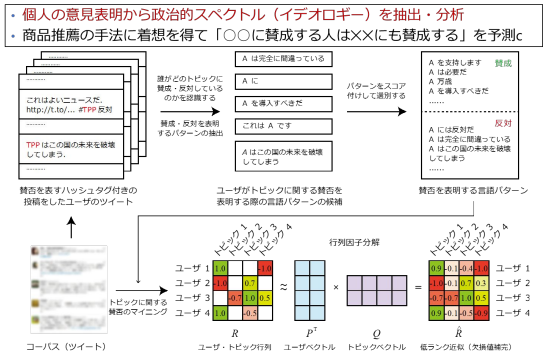


図５トピック間選好知識の獲得

この研究を拡張し、ユーザの日々の投稿と獲得した知識を対応付ける手法を考案した。

この手法は，行列因子分解から Factorization Machine への自然な拡張によるものである（図６）．実験結果より、テキストと知識を対応付けることで意見分析の精度が向上することが示された［学会発表1］.

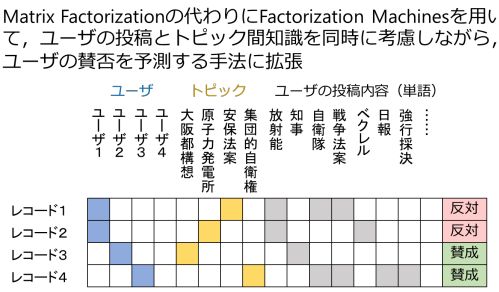


図６　トピック間選好知識による賛否分類

研究項目③で構築したコーパスを用い，Wikipedia で言及されている因果関係知識と，与えられたテキストを照合しながら賛否分類を行う手法を開発した（図７）．与えられたテキストに促進・抑制関係知識を照合するだけでも，賛否分類の正解率が向上すること，注意機構を使って柔軟に知識の照合を行うことで，さらに正解率が向上することが確認され，本研究課題の仮説が実証された.

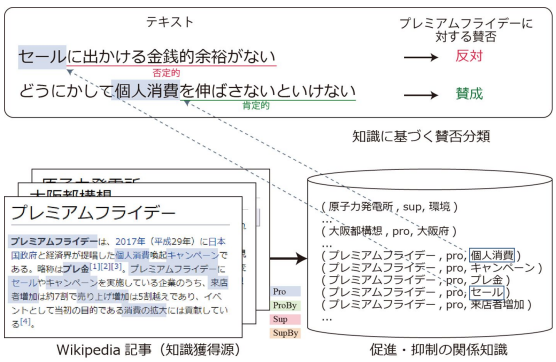


図７　因果関係知識を活用した賛否分類

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　４件）

※すべて査読あり

1. Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, Kentaro Inui. A Joint Neural Model for Fine-Grained Named Entity Classification of Wikipedia Articles. IEICE Transactions on Information and Systems, Special Section on Semantic Web and Linked Data, E101.D(1):73-81, 2018 年 1 月. (ISSN: 0916-8532) [10.1587/transinf.2017SWP0005]
2. 高瀬 翔, 岡崎 直観, 乾 健太郎. 関係パタンの分散表現の計算. 人工知能学会論文誌, 32(4):D-G96_1-11, 2017 年 7 月. (ISSN: 1346-8030) [10.1527/tjsai.D-G96]
3. Shuangshuang Zhou, Naoaki Okazaki, Koji Matsuda, Ran Tian, Kentaro Inui. Supervised Approaches for Japanese Wikification. Journal of Information Processing, 25:341-350, 2017 年 4 月. (ISSN: 1882-6652) [10.2197/ipsjjip.25.341]
4. Sho Takase, Naoaki Okazaki, Kentaro Inui. Modeling semantic compositionality of relational patterns. Engineering Applications of Artificial Intelligence, 50:256 - 264, 2016 年 4 月. [10.1016/j.engappai.2016.01.027]

〔学会発表〕（計　44 件）

※以下のうち，英語のものは査読付き国際会議，日本語のものは招待講演.

1. Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki and Kentaro Inui. Predicting Stances from Social Media Posts using Factorization Machines. In Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), 掲載決定, August 2018.
2. Kazuaki Hanawa, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. A Crowdsourcing Approach for Annotating Causal Relation Instances in Wikipedia. In Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (PACLIC 2017), 10 pages, November 2017.
3. Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki, and Kentaro Inui. Other Topics You May Also Agree or Disagree: Modeling Inter-Topic Preferences using Tweets and Matrix Factorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 398-408, July 2017.
4. Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. Neural Joint Learning for Classifying Wikipedia Articles into Fine-grained Named Entity Types. In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30), pages 535-544, October 2016.
5. Masatoshi Suzuki, Koji Matuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. Multi-label Classification of Wikipedia Articles into Fine-grained Named Entity Types. In IEEE/WIC/ACM International Conference on Web Intelligence (WI 2016), pages 483-486, Omaha, USA, October

2016.
6. Sho Takase, Naoaki Okazaki, and Kentaro Inui. Composing Distributed Representations of Relational Patterns. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2276-2286, Berlin, Germany, August 2016.
7. Davaajav Jargalsaikhan, Naoaki Okazaki, Koji Matsuda, and Kentaro Inui. Building a Corpus for Japanese Wikification with Fine-Grained Entity Classes. In Proceedings of the ACL 2016 Student Research Workshop, pages 138-144, Berlin, Germany, August 2016.
8. Sho Takase, Naoaki Okazaki, and Kentaro Inui. Fast and Large-scale Unsupervised Relation Extraction. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29), pages 96-105, Shanghai, China, October 2015.
9. 岡崎 直観. 深層学習の自然言語処理への応用. 情報処理学会連続セミナー2017 第 4 回:ディープラーニングの活用と基盤, 中央大学(東京都), 2017 年 10 月.
10. 岡崎 直観. 自然言語処理における Deep Learning. 電子情報通信学会総合大会 2017 企画セッション 「もっと知りたい! Deep Learning ～基礎から活用まで～」, 名城大学(愛知県), 2017 年 3 月.
11. 岡崎 直観. 言語処理における常識的知識の獲得・活用. 第 32 回ファジィシステムシンポジウム, 佐賀大学 本庄キャンパス(佐賀県), 2016 年 9 月.
12. 岡崎 直観. 自然言語処理における深層ニューラルネットワーク研究のフロンティア. 電子情報通信学会 バイオメトリクス研究専門委員会(BioX) 2016 年度 8 月研究会, 182, p. 13, 東北大学 片平キャンパス(宮城県), 2016 年 8 月.
13. 岡崎 直観. 深層ニューラルネットワークによる知識の自動獲得や推論. 第 71 回人工知能セミナー「Deep Learning 技術の仕組みと自然言語処理への応用」, 慶應義塾大学 日吉キャンパス(神奈川県), 2016 年 6 月.
14. 岡崎 直観. 単語・句の分散表現の学習. WebDB Forum 2015 特別セッション 3: Deep Learning と自然言語処理, 芝浦工業大学(東京都), 2015 年 11 月.
15. 岡崎 直観. 単語の分散表現と構成性の計算モデルの発展. 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 (2), はこだて未来大学(北海道), 2015 年 5 月.

〔図書〕(計 4 件)

1. 高瀬 翔, 岡崎 直観. 自然言語文からの関係知識ベースの構築. 知能と情報(日本知能情報ファジィ学会誌), 29(2):55-64, 2017 年 4 月.
2. 岡崎 直観. 言語処理における分散表現学習のフロンティア. 人工知能, 31(2):189-201, 2016 年 3 月.
3. 岡崎 直観. 単語の意味をコンピュータに教える.岩波データサイエンス Vol. 2. 岩波書店, 2016 年 2 月.

〔産業財産権〕

○出願状況(計 0件)

○取得状況(計 0件)

〔その他〕

1. 日本語 Wikification コーパス ver 0.1 (2016/03/10): http://www.cl.ecei.tohoku.ac.jp/jawikify/
2. Wikipedia 記事への促進・抑制関係付与コーパス: http://www.cl.ecei.tohoku.ac.jp/wikipedia_pro_sup/

6. 研究組織

(1)研究代表者

岡崎 直観(OKAZAKI, Naoaki)
東京工業大学・情報理工学院・教授
研究者番号:50601118