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　　５　月　２５　日現在

機関番号：　１１３０１
研究種目：　研究活動スタート支援
研究期間：　2016～2017
課題番号：　１６Ｈ０６６１４

研究課題名（和文）断片的に散らばった抽象的世界知識の構造化と集約

研究課題名（英文）Structuralizing and aggregating instances of abstract world knowledge

研究代表者
井之上　直也（Inoue, Naoya）

東北大学・情報科学研究科・助教

研究者番号：８０７７８６０５

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：人工知能をより賢くするために、世の中に存在する大量の言語データから因果関係などの世界知識を自動獲得するための基盤技術を開発した。また、これらの世界知識を用いて、観測された事象から未知の仮説を予測するための「仮説推論」の技術を開発した。特に本研究では、「宇宙開発に投資すべきか」などのトピックについて書かれた論述文を解析対象とし、主張文-根拠文の関係の自動解析器の開発や、「分散表現」と呼ばれる単語のベクトル表現を用いて知識を柔軟に使いこなすための仮説推論の技術開発を行った。

研究成果の概要（英文）：We have studied core technologies for automatically acquiring world knowledge which makes artificial intelligence smarter, such as causality, from large-scale textual data. We also studied an abductive reasoner which predicts a new hypothesis based on observed facts and world knowledge. Focusing on argumentative texts under such topics as "Should we invest in space exploration?", we developed a computational model to identify the relationship between a claim and premise. We also developed a framework that can perform flexible reasoning by leveraging distributed representations.

研究分野：　自然言語処理

キーワード：　世界知識　自然言語処理　推論

１．研究開始当初の背景

我々人間が日常的に行っているような常識的推論の計算機による実現は、自然言語処理の高精度化における重要課題の一つである。近年、こうした課題への取り組みとして、世の中に超大規模に存在するウェブログ、ニュース記事などの文書データから、常識的推論に必要な世界知識を確率統計的に自動獲得する試みが盛んになされている。しかし、これらのうちで実用的な水準に達し知識ベースとして公開・整備が行われているのは、Freebase, DBpedia を始めとした、実在し一意に特定できる具体物（人物、建物など）に関する関係知識であり（例えば、<Barack Obama, is-president-of, US>）、抽象的な関係知識（「アルコールは肝臓ガンを引き起こす」など）の自動獲得の試みには、知識集約に関して解決すべき大きな課題が残っている。

一般に、知識自動獲得における大きな課題として、個々の文から抽出した断片的な知識の同一性をどのように自動判定し集約・整理するか、という問題がある。具体物に関する知識自動獲得の研究では、個々の文から知識を抽出する際、名詞が指し示す具体物を出現文脈から一意に同定し（Entity Linking）、各知識を実世界の実体に結びつけることで知識の集約を可能にしている。しかし、抽象的な世界知識が指し示すものは、文字通り実体のない抽象的なもの（非可算無限集合的なもの）であり、さらに、個々の文から抽出される断片的な知識が相互に談話的な繋がりをもち（例えば、照応関係や条件・対比・例示など）、一つの構造的な知識を構成する場合が多い。

例えば、図１の文章では、“homework has little educational worth”という抽象的な主張がなされ、それに対する根拠として、“homework adds nothing to standardized test scores”であることが、“elementary pupils”という適用範囲と、“Studies”という情報源を伴って提示されている。この文章から抽出されるそれぞれの知識は互いに関連しあって一つの知識を形成しており、別の文章から抽出された知識とのマージをどのように行うべきか、全く自明な問題ではない。例えば、別の文章から抽出した“homework has little educational worth”という知識は、“high school student”という適用範囲を伴った研究結果により支持されたものかもしれない（その場合、まとめあげるべきでない）。このように、抽象的な世界知識に関する知識獲得では、これまでに取り組まれていない知識集約に関する課題が依然として残っており、新しい打開策が必要である。

２．研究の目的

研究の焦点を明確化するために、知識獲得の対象を論述文に絞り、二つの課題に取り組む。第一に、抽象的な世界知識の集約の第一歩として、従来の関係抽出技術によって論述文から得られる、断片的な抽象的世界知識の間の関連付けを行い、構造的な世界知識を自動獲得する手法を実現する。第二に、構造的知識を文書横断的に集約し、常識的推論に利用できるようにするための基礎的な演算機構を構築し、ディベート構造の自動理解タスクの上で評価を実施する。また、研究成果をオープンソース・ソフトウェアで公開すると同時に、ウェブ上に存在する論述文から得られる実用規模の知識データベースの構築・公開を行う。

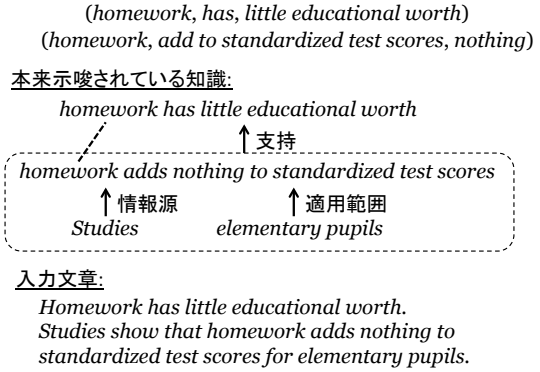

図 1. 文章から得られる構造的な知識

本研究では、文から抽出した断片的な抽象的世界知識を集約・利用するための基盤技術の確立を目的とする。課題の発散を防ぐため、ディベートの自動生成を応用先として想定し、論述文からの知識自動獲得に関して研究する。より具体的には、次の二つの課題に取り組む: (1) 課題 1: 論述文からの構造的知識の抽出技術の開発、(2) 課題 2: 構造的知識に対する基礎的演算の実現。

課題１では、従来の関係抽出技術によって論述文から得られる、断片的な抽象的世界知識の間の関連付けを行い、図１で示すような構造的な世界知識を大規模に自動獲得する手法を実現する。これ以降、構造化前の知識を一次知識と呼び、構造的知識と区別する。

課題２では、課題１により得られる構造的知識を文書横断的に集約し、常識的推論に利用できるようにするための基礎的な枠組みを構築する。より具体的には、構造的知識の間に、同義関係・対比関係等の対応付けや、マージといった基礎的な演算を定義し、これを頑健に実現するための計算モデルを構築する。そのために、各文書で文脈付けられた概念間の同義関係認識、他の事象により条件付けられた事象間の同義性の認識等を行うための方法を明らかにしていく。また、構築した演算モデルの良さを評価するために、応募者がこれまでに従事してきたディベートの自動理解に関するベンチマークを構築し、大規模な評価実験を行うことで、経験的に、かつ人手による分析を通して、残る課題を整理する。

３．研究の方法

ウェブ上の論述文から大規模に一次知識を抽出し、これを土台にして一次知識の自動構造化モデル、構造的知識間の対応付けモデル、文書横断的な構造的知識の集約モデルの実現へとステップアップしていく。既存の資源として、応募者のこれまでの研究成果や、応募者が参画している学術文献解析プロジェクトより提供される既存の技術や言語資源・計算機資源を積極的に利用し、発展的な課題に素早く取り組めるように留意する。それぞれの技術を一つのコーパスの上で串刺し的に評価するために、オンラインディベートウェブサイトから収集した論述文に対して、構造的知識、対立する論述文との構造的知識レベルの対応関係を層状に付与した評価用データセットを新たに構築する。また、各取り組みの節目で、知識の自動獲得技術をソースコード付きで論文発表し、構築された知識データベースを整備・公開し、広く研究者に利用可能にするとともに、外部からのフィードバックを取り入れていく。

４．研究成果

３．で述べた２つの課題それぞれに対して、次のような成果が得られた。

(1) 課題 1: 論述文からの構造的知識の抽出技術の開発

①一次知識の構造化の前段階として、大規模コーパスからの一次知識の抽出、一次知識の表現方法についての基礎的検討を行った。具体的には、約２万件の自然言語処理の論文データを集めた ACL Anthology Reference Corpus より、数百万のオーダで関係知識（一次知識）の抽出を行い、その中に（TFIDF, is used for, feature weighting）のような有用な関係が多く含まれることがわかった。

②知識の表現に関する基礎検討として、大規模 Web コーパス ClueWeb2012 より、20 億個の因果関係知識を抽出し、知識表現として“figure out”などの複単語表現を考慮することの重要性を、因果関係認識タスク COPA の上で経験的に確認した（学会発表⑦）。図２は、複単語表現を考慮するか否かで COPA の精度がどのように変わるかを示しており、複単語表現による(CS w/ MWP) 精度の向上が見られる。

Results of COPA evaluation

Method	Corpus	Accuracy (%)
Random		50.0
PMI (Roemmele +, 2011)	Project Gutenberg	58.8
PMI-EX (Gordon +, 2011)	Personal stories	65.4
CS w/o MWP$_{\lambda=1.0}$ (Luo +, 2016)	Causal Net	70.2
CS w/o MWP$_{\lambda=0.8}$	ClueWeb12	69.9
CS w/ MWP$_{\lambda=0.7}$	ClueWeb12	**71.2**

図 2. 複単語表現を考慮した因果関係表現の評価結果

③因果関係知識の表現方法として、ニューラルネットワークに基づくエンコーダ・デコーダモデルにより得られる分散表現的なアプローチについても基礎検討を行い、従来の記号的アプローチに対する有用性を実験により確認した（学会発表⑥）。

④一次知識の自動獲得のために、主張-根拠の関係を持つ文の自動同定モデルの構築を行った。このとき、解析対象の文の周辺から、より大域的な論述構造を抽出・利用して解析を行うことにより、その解析精度を向上させられることを経験的に示した（学会発表②）。例えば、図３に示すような大域的な論述パタンを考慮することで、精度が向上していると考えられる。

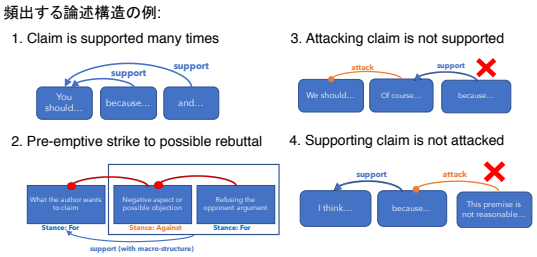

図 3. 頻出する大域的な論述構造の例

⑤一次知識の獲得に必要な「主張-根拠」関係の解析に向けて、論述文において、あるエンティティに対する感情極性情報を根拠付きで自動抽出する課題の基礎検討を行った。例えば、「掃除機 A はハウスダストも吸ってくれます」という文は、掃除機 A に対するポジティブな評価をしていることを解析するものである。アノテーションスキームの設計、および試験的なコーパスアノテーションを行い、その実現性を検討した（学会発表③）。

(2) 課題 2: 構造的知識に対する基礎的演算の実現

①構造的知識間の関係認識の枠組みとして、論理推論に基づく枠組みについて検討し、論理に基づく仮説推論の一種である Etcetera Abduction の推論効率を改善することに成功した（学会発表⑤）。図 4 は、先行研究と開発した推論エンジンの比較を行った結果で

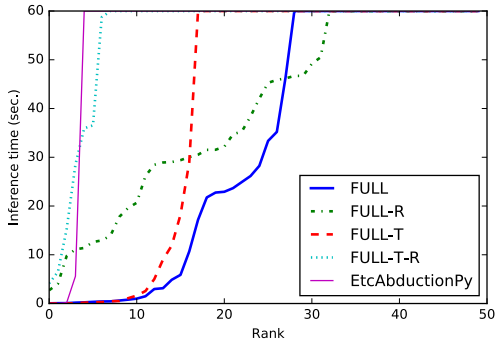

図 4. 推論効率の評価結果

ある。先行研究（EtcAbductionPy）は、入力の問題のサイズに対してすぐに推論時間が爆発的に増えてしまうのに対して、提案手法（FULL-*）はより頑健であることがわかる。

②本研究の応用先の一つであるディベート理解を見据え、ディベート理解のタスクを設計した。Argumentation Mining コミュニティで構築された中規模エッセイコーパスに対して実際にアノテーションを行い、アノテーション結果の分析を行った（学会発表⑧）。このアノテーション結果は、https://github.com/preisert/deep-arg-structure-corpus において一般公開している。また、知識評価の環境構築として、構造化パーセプトロンに基づくディベート理解モデルを構築した（論文投稿中）。

③知識演算の枠組みとして、分散表現に基づく推論器の検討を行った。モデルは図５に示すように、ニューラルネットに基づくエンコーダ・デコーダモデルを用いたものである。ウェブから獲得した大規模知識を用いて推論器を主観的に評価し、ある程度の精度で柔軟な推論ができることを示した（学会発表④）。

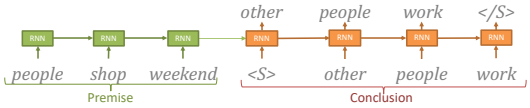

図 5. 分散表現に基づく推論器

④③で構築した後ろ向き推論器をディベート理解タスクに応用するため、McCarthy らの自然論理を用いた知識演算の機構に関する基礎的検討を行った（学会発表①）。

本プロジェクトでは、主に論述文を解析するための基礎技術、特に知識獲得・集約について、基礎的な検討を行った。今後は、残された課題である、より大規模な知識獲得、開発した推論器に基づく知識の集約・実応用の上での評価を行っていく予定である。また、研究の過程で得られた各種技術についても、一般公開のためのコード整備を行い、広く利用可能にする予定である。

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）

〔学会発表〕（計　８　件）

① Naoya Inoue, Pontus Stenetorp, Sebastian Riedel, and Kentaro Inui. Towards Interpretation as Natural Logic Abduction. 人工知能学会全国大会（第 32 回）. 2018.
② Tatsuki Kuribayashi, Paul Reisert, Naoya Inoue, and Kentaro Inui. Towards Exploiting Argumentative Context for Argumentative Relation Identification. 言語処理学会第 24 回年次大会. 2018.
③ 白井 穂乃, 井之上 直也. 情報科学論文における問題解決手法と評価表現の付与仕様の検討. 人工知能学会全国大会（第 32 回）. 2018.
④ Naoya Inoue. Encoder-Decoder Abduction: Flexible Abductive Reasoning with Distributed Representations. 第 4 回 Language & Robotics 研究会. 2018.
⑤ Naoya Inoue and Andrew S. Gordon. A Scalable Weighted Max-SAT Implementation of Propositional Etcetera Abduction. Proceedings of the 30th International Florida Artificial Intelligence Research Society Conference. 2017.
⑥ Melissa Roemmele, Sosuke Kobayashi, Naoya Inoue and Andrew Gordon. An RNN-based Binary Classifier for the Story Cloze Test. Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics. 2017.
⑦ 佐々木翔大, 高瀬翔, 井之上直也, 岡崎直観, 乾健太郎. 複単語表現を利用した因果関係推定モデルの改善. 情報処理学会 自然言語処理研究会報告. 2017.
⑧ Paul Reisert, Naoya Inoue, Naoaki Okazaki, Kentaro Inui. Deep Argumentative Structure Analysis as an Explanation to Argumentative Relations. 言語処理学会第 23 回年次大会論文集. 2017.

６．研究組織
(1)研究代表者

井之上　直也（INOUE, Naoya）
東北大学・大学院情報科学研究科・助教
研究者番号：80778605