

(続紙 1)

京都大学	博士 (情報学)	氏名	大村 和正
論文題目	Studies on Data-Driven Discourse Relation Recognition toward Natural Language Understanding (自然言語理解に向けたデータ駆動の談話関係認識に関する研究)		
(論文内容の要旨)			
<p>談話関係とはテキストスパン間の意味的な関係を指す。中でも「お腹が空いたので、ご飯を食べる」といった、一方が他方を引き起こし得る事態間に成立する談話関係を蓋然的関係と呼ぶ。本論文は、自然言語理解において重要であると考えられる蓋然的関係及び談話関係を推論する能力に焦点を当て、これらの改善に向けたデータ構築手法について論じている。また、構築したデータに関連する自然言語理解タスク及び教育分野に応用し、談話関係の有用性の実証も行っている。本論文は、これらの取り組みについて論じた4つの章と序論及び結論の章を合わせた全6章から構成される。</p> <p>第1章は序論である。まず、計算機による自然言語理解の実現に向けて、談話関係を推論する能力に焦点を当てる背景を概説している。特に、深層学習が目覚ましい発展を遂げている現代でも談話関係の自動認識は未だ困難な課題であり、これを改善するためにデータ駆動のアプローチが有効であることを述べている。次に、自然言語理解及び談話関係の既往研究を概観し、前述の背景を補足しつつ本論文の位置付けを論じている。最後に、談話関係を推論する能力の改善並びに談話関係の有用性の実証に向けた各章の取り組みについて概説している。</p> <p>第2章は、基本的な蓋然的関係を推論する能力(常識推論能力)を評価するためのデータセットの構築手法について論じている。蓋然的関係は、因果関係を下位分類に含む主要な談話関係の一つであり、幅広い応用可能性を持つ。しかし、その重要性に反して、蓋然的関係に関する大規模な日本語の常識推論データセットは存在しなかった。この問題を解決するために、テキストからの自動抽出に基づく半自動的な常識推論問題の生成手法を提案している。具体的には、「お腹が空いたので、ご飯を食べる」といった、蓋然的関係を持つ基本的なイベント表現の組をテキストから自動抽出し、クラウドソーシングによる確認を行った後、それらの蓋然的関係を問う多肢選択式問題を自動生成するというものである。本研究では、提案手法に従って日本語ウェブコーパスから10万問規模のデータセットを構築している。また、計算機による解答実験を行い、計算機と人間の解答精度に開きがある、すなわち常識推論能力の評価に有効なデータセットであることを確認している。</p> <p>第3章は、疑似問題による常識推論能力の改善並びに関連する自然言語理解タスクへの応用に取り組んでいる。第2章で評価に有効なデータセットを構築した次は、これを用いたモデルの改善である。前述の常識推論問題の生成手法において、クラウドソーシングによる確認を省略することで常識推論問題を模した疑似問題を自動生成することが可能となる。本研究では、この拡張性の高さを利用して疑似問題を大規模に自動生成し、このデータ拡張による常識推論能力の改善を試みている。また、基本的な蓋然的関係を推論する能力は自然言語理解において重要であるという前提に立つと、生成した問題は関連する自然言語理解タスクにおいても有用であることが期待される。これを確認するために、常識推論問題および疑似問題からの転移学習による関連タスクへの効果を定量的に評価している。</p>			

第4章は、談話関係認識の改善に向けた合成データの生成手法について論じている。第3章で疑似問題によって基本的な蓋然的関係を推論する能力はベンチマーク上で人間と同等の性能を達成したため、研究の対象を蓋然的関係から談話関係全般に拡張している。談話関係認識における大きな問題として、一部の誤りやすい談話関係の訓練データが少ないことが挙げられる。この問題を軽減するために、大規模言語モデルを用いた合成データの生成手法を提案している。具体的には、混同行列をもとに誤りやすい談話関係の組を抽出し、その混同を解消するようなデータを2段階のプロンプティングによって合成するというものである。本研究では、提案手法に従って合成データを生成し、その有効性を英語及び日本語で検証している。

第5章は、研究の過程で構築したデータを応用した小学生を対象とする作文学習アプリの設計・開発について論じている。構築したデータは、その規模や自然言語理解における談話関係（蓋然的関係）の重要性などを考慮すると、計算機だけでなく人間の言語学習にも有用であることが期待される。本研究では、「小学生は作文に苦手意識を持つ傾向がある」という日本語教育における積年の問題に着目し、小学生を対象とした基本的なイベント表現及び蓋然的関係に関する作文学習アプリを設計・開発している。また、小学生の子とその親80組を対象にユーザー調査を行い、作文学習の導入としての有用性を確認している。

第6章は結論である。本論文を総括し、今後の展望について述べている。

(論文審査の結果の要旨)

本論文は、自然言語テキスト中の蓋然的関係及び談話関係を計算機によって推論する能力に焦点を当て、これらの改善に向けたデータ構築手法を論じている。また、関連する自然言語理解タスク及び教育分野への応用を通して談話関係の有用性を実証している。得られた主要な成果は以下のとおりである。

1. 談話関係とはテキストスパン間の意味的な関係を指す。また、一方が他方を引き起こし得る事態間に成立する談話関係を蓋然的関係と呼ぶ。本論文では、基本的な蓋然的関係を問う多肢選択式問題(常識推論問題)の半自動的な生成手法を提案し、過去に類を見ない10万問規模の大規模な日本語の常識推論データセットを構築することに成功した。また、計算機による解答実験を行い、当時の計算機モデルと人間の間で性能の開きがある、すなわち評価に有効なデータセットであることを確認した。

2. 1の提案手法を応用して常識推論問題を模した疑似問題を大規模に自動生成し、このデータ拡張による常識推論能力及び関連する自然言語理解タスクへの効果を検証した。実験の結果、生成した問題を応用することで常識推論能力及び関連タスクの性能の改善を確認し、自然言語理解における蓋然的関係を推論する能力の重要性を実証した。なお、この成果をまとめた論文は自然言語処理の主要国際会議COLING 2022にてOutstanding Paper Awardを受賞し、国際的に高く評価された。

3. 談話関係認識は与えられたテキストスパン間の談話関係を識別するタスクである。談話関係認識における大きな問題として、一部の誤りやすい談話関係の訓練データが少ないことがあった。本論文では、この問題を軽減するために、大規模言語モデルを用いた合成データの生成手法を提案した。実験の結果、合成データが談話関係認識の精度向上に有効であることを英語及び日本語で確認した。

4. 「小学生は作文に苦手意識を持つ傾向がある」という日本語教育における積年の問題がある。本論文では、この現状を改善し、談話関係(蓋然的関係)の有用性を実証するために、構築したデータを応用した小学生を対象とする作文学習アプリを開発した。小学生の子とその親80組を対象に行ったユーザー調査では7割の参加者が好評を示し、作文学習の導入として有用であることを実証した。

よって、本論文は博士(情報学)の学位論文として価値あるものと認める。また、令和6年2月21日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。また、本論文のインターネットでの全文公表についても支障がないことを確認した。