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

令和　元 年　 ６ 月 １１ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)（一般）
研究期間：2016～2018
課題番号：１６Ｈ０２８６５

研究課題名（和文）言語理解における人間の振舞いの分析と言語処理の高精度化への応用

研究課題名（英文）Analysis of human behaviour in language understanding and its application for natural language processing

研究代表者

徳永　健伸（Tokunaga, Takenobu）

東京工業大学・情報理工学院・教授

研究者番号：２０１９７８７５

交付決定額（研究期間全体）：（直接経費）　　10,200,000円

研究成果の概要（和文）：自然言語処理の分野では，単語の品詞など，解析結果として得たい正解を解析対象のテキストに人手で付与(アノテーション)した「コーパス」を構築し，それを機械学習の訓練データとして用いることによって問題を解く手法が主流である．自然言語処理の広範な課題を網羅するために，自然言語処理の課題をセグメント課題，リンク課題，変換課題に分類し，各タイプの課題においてアノテーション作業者の視線，キー入力，マウス操作などの振舞いを記録し，アノテーション結果に加えてこれらの振舞いの情報も機械学習の入力とすることを提案し，実験を通してその有効性を検証した．

研究成果の学術的意義や社会的意義
文章読解や問題解決における視線情報などの振舞いを分析する研究はおこなわれているが，アノテーション中の振舞いを記録し，さらにそれを言語処理に活用する研究は類を見ない．本研究課題の成果であるアノテーション過程の作業者の振舞いから言語処理に有用な情報を抽出する手法は，十分な解析性能が得られていなかった言語処理の分野においても解析性能を向上させることができ，その学術的貢献は大きい．自然言語処理技術は実社会ですでに利用され始めている．今後，意味や文脈に踏み込んだ研究が発展することにより，既存の実用システムがさらに高度化され，使い易くなることが期待できる．

研究成果の概要（英文）：The main stream of natural language processing (NLP) research is the machine learning-based approach in which a model for a specific task is trained by using manually annotated corpora. To cover broad types of NLP tasks, we classify the task into three: segmentation, linking and transforming. For each kind of task, we collected annotator's eye gaze, key inputs and mouse operations during their annotation activity. We proposed to utilise such annotator's behaviour information for machine learning and verified its effectiveness through the experiments.

研究分野：自然言語処理

キーワード：自然言語処理　コーパス　アノテーション　視線情報　振舞い情報　機械学習

１．研究開始当初の背景

自然言語をコンピュータで処理する自然言語処理の分野では，単語の品詞など，解析結果として得たい正解を解析対象のテキストに人手で付与(アノテーション)した「コーパス」を構築し，それを機械学習の訓練データとして用いることによって問題を解く手法が主流である．この「コーパス構築+機械学習」の手法は単語の品詞同定などの基盤技術から機械翻訳などの応用に至るまで多くの課題に適用され一定の成功をおさめている．

我々は種々のコーパス構築やアノテーションの高精度化に関する，構築したコーパスを用いて「コーパス構築+機械学習」の枠組で種々の課題を解く研究で成果をあげてきた．その結果，従来の枠組は，テキストの表層に近い情報を扱う問題では有効だが，言語理解において重要となる意味や意図を扱う意味・文脈解析における性能は未だ不十分な水準にあるという知見を得た．たとえば，文脈を考慮して代名詞の先行詞を同定する照応解析では，英語で70%程度，省略の多い日本語ではわずか40%程度の精度しか得られていない．この理由として以下の2つの問題が考えられる．

問題(1) 意味・文脈解析に有用なテキスト中の情報が明らかではなく，どのような情報が有用かという点から検討する必要がある．

問題(2) 意味・文脈解析のためにアノテーションする情報が複雑なため，人手によっても信頼性の高いアノテーションをされた高品質のコーパスを構築することが難しい．

我々はこれまでの研究成果から，アノテーションによって付与された解析結果だけではなく，人間のアノテーション過程の振舞いから得られる情報を利用すればこれらの問題が解決できるという着想を得た．人間によるアノテーションは解析器が解く課題を人間が解いていると考えられる．この過程で人間がテキスト中のどのような情報を参照しているかを人間の振舞いから同定し，それを解析器の学習に取り込めればより性能の高い解析器を構築できる可能性がある(→問題(1))．また，アノテーション中の人間の振舞いからアノテーションの信頼性を見積る手法を確立することによって，より質の高いコーパスを効率よく構築し，解析器の性能向上に寄与することができる(→問題(2))

我々はすでに予備的な実験により，アノテーション過程における作業者の視線やマウス操作などの振舞いデータを収集・分析しており，テキスト中の述語とその主語や目的語などを同定する課題において有望な結果を得ている．

２．研究の目的

以上の背景をふまえ，本研究では具体的な目標として以下の3項目を設定する．

1. コーパス・アノテーション過程における作業者の振舞いデータ収集法の確立
   アノテーション中の作業者の動作や視線など計測可能なデータを高精度で効率的に収集する手法を確立する．特に視線情報については，従来の研究が画面を数分割した程度の解像度を扱っていたのに対し，本研究では，単語レベルの解像度で視線情報を収集する手法を確立する．
2. 振舞いデータの分析法および分析結果の言語処理への適用法の確立
   振舞いデータの分析を通して言語理解において人間が利用する知識を明らかにする分析法を確立するとともに，得られた知見を解析器の構築に利用する手法を確立する．さらに，振舞いデータから得られた情報を利用することによって解析性能が向上することを評価実験により確認する．
3. 収集データを利用したアノテーションの質を評価・改善する手法の確立
   従来，複数のアノテーション結果の比較によって評価していたアノテーションの質を，人間の振舞いデータによって評価する手法を確立する．さらに，振舞いデータは作業者が誤りやすい点の検出にも利用できるため，これを利用してアノテーションの質を向上させる手法を確立する．

３．研究の方法

(1) 振舞いデータ収集のための環境構築
アノテーション時のアノテータの視線とキーボード入力，マウス操作を記録できるように既存のツールを改修する．

(2) 振舞いデータの収集と分析
広範な自然言語処理に課題に対応する枠組を考えるために，以下のように課題を抽象化・分類し，各分類の具体的な課題についてデータを収集する．
《セグメント課題》:単語への品詞付与，固有表現認識など，テキスト中の語や句などのセグメントに情報を付加する課題
《リンク課題》:係り受け解析，照応解析，述語項構造解析など，テキスト中の複数のセグメントに関係(リンク)を付与する課題
《変換課題》:翻訳，要約，言い換えなど，テキストを別のテキストに変換する課題

これらの課題についてアノテーション中の作業者の操作履歴，および注視しているテキスト中の文字/単語を時間情報とともに記録する．
収集したデータについて，作業時間やマウス操作のタイミング，さらに視線研究で多用される視線の停留回数，停留時間，スキャンパスなどの作業者の振舞いに関する情報とテキストの長さ，対象語の意味カテゴリ・出現位置，他の語の関係などテキストが持つ性質の関係を相関分析や回帰分析などの統計処理によって分析し，課題に正答するために必要な要因を明らかにする．

４．研究成果

セグメント課題の具体例として固有表現の意味カテゴリ同定課題を取り上げ，16名のアノテータが各々72テキストにアノテーションする際の振舞いデータを収集した．従来の機械学習に基づく固有表現認識では，対象固有表現の周辺文脈を素性として用い学習をおこなう．我々はこの課題をおこなう人間の視線がどのように分布しているかを調査した．まず，各視線停留とテキストとの対応を明らかにするために，テキストを文節に区切り，文節と視線停留との対応付けをした．対象固有表現の両側1文節あるいは2文節上にある停留数/停留時間の割合の平均と標準偏差を分析した結果，異りで8割以上の停留は対象固有表現の前後1文節以外の文節にあることがわかった．範囲を2文節に拡げても局所文脈内の停留は3割程度である．つまり，人間は広範な文脈を参照して固有表現の種別を決めているということが示唆される．
次に各課題の正解者と不正解者の視線が停留した文節にどのような差があるかを分析した．正解者と不正解者の数が拮抗した課題について，両グループの停留傾向の差を見るために，各グループについて各文節に対する停留回数と総停留時間をグループ内人数で正規化した値を計算し，正解者の値から不正解者の値を引いた差が0より大きな文節に注目し，分析をおこなった．その結果，正解者の視線が対象固有表現が項となっている述語やその述語の他の項に停留している傾向が観察できた．対象固有表現に近接する単語だけではなく，対象固有表現との統語関係が重要な手掛りになることがわかった．また，局所的な文脈が特定の意味カテゴリを強く示唆するために不正解の原因になる例もあった．以上のことから固有表現認識の精度を向上させるには，より広範な文脈情報を利用する必要があるという結論を得た．この成果は2017年9月に開催された国際会議Recent Advances in Natural Language Processing (RANLP 2017)で発表した．
リンク課題の具体例として日本語述語項構造解析課題を取り上げ，20名のアノテータが各々184テキストにアノテーションする際の振舞いデータを収集した．収集したデータを分析をおこなった結果，人間のアノテータが統語的な依存情報を有力な手掛りとして利用する傾向にあることを明らかにした．この分析を元に日本語述語項構造解析課題において，解析モデルのパラメータ推定をおこなう際にアノテータの視線情報を利用し，テキスト内の言語的な情報のみだけでなくアノテーション時のアノテータの視線から得られる情報を取り入れることで解析の精度を向上させる手法を提案した．アノテータの特定の述語に対するガ格をアノテーション

する最中の視線を観察してみると，最終的な判断をするまでに様々な候補を見ていることがわかる．最終的な判断において選ばれなかった候補は，そのテキストにおける対象述語の項ではないものの，頻繁に注視していた候補については他のテキストにおいて同じ述語の項となる可能性がある．そこでそのような候補をニアミス候補と考え，ランキング学習の枠組みを利用することでニアミス候補を活用できると考えた．ランキングの生成にのみ視線データを利用することで，パラメータ推定に視線が必要となるが視線データのない新規のテキストに対して項を推定することが可能になる．評価実験の結果，視線情報を利用することで述語と同一文内に現われるガ格項の同定精度が最大で0.07ポイント向上し，視線情報が述語項構造解析に有効であることを示した．しかしながら文を越えた述語項構造の同定の精度は依然として低く，さらに視線情報の使い方に改善の余地がある．この研究成果は2016年12月に開催された国際会議 The 26th International Conference on Computational Linguistics (COLING 2016)で発表した．

変換課題として文章要約と文書の推敲を取りあげ，アノテーション中の視線やツール操作などの行動履歴を対象として分析をおこなった．文章要約については，英語試験IELTSの読解問題から800語程度の文章を3つ選び，それぞれを1/4程度に要約する課題を10名の博士課程学生に与え，要約作成過程の視線とキー入力を記録した．10作業者が3文章について作成した30の要約とその要約過程の行動履歴を収集した．人間の要約過程は元文書を読む段階と情報を選択する段階の2つの段階からなると仮定した要約モデルを提案し，要約に使われた語と使われなかった語への視線の停留の比率がこれらの2つの段階で逆転するという知見を得た．この知見は自動要約において要約にどのような語を使うべきかの指針を与えるものである．この成果については論文を準備中である．文書の推敲については，同一命題についてアジア圏の大学生が作成した議論的エッセイを推敲し，エッセイを改善する課題を扱った．推敲対象のエッセイは神戸大学で構築されたコーパスICNALEからエッセイの評価点が中位のものを150選択した．作業者がガイドラインにしたがって，これらのエッセイに議論構造を付与し，それを参考にしてより説得力のあるエッセイに修正する過程のツール操作を行動履歴として収集した．データ分析の結果，議論構造を構成する文間の関係のうち，作業者はdetailとsupport関係の決定に迷うことが多く，この傾向は作業者間の一致度の分析におけるこれらのラベルの不一致率が高いことと整合する．これはアノテーションの信頼性を単一作業者の行動履歴からも推定できることを示している．この成果については，2019年3月に開催された言語処理学会第25回年次大会でデータの収集を中心に論文を発表した．データの分析結果については論文を準備中である．

５．主な発表論文等

〔雑誌論文〕（計0件）

〔学会発表〕（計7件）

1. Jan Wira Gotama Putra, Simone Teufel and Tokunaga Takenobu, An argument annotation scheme for the repair of student essays by sentence reordering, 言語処理学会第25回年次大会(NLP2019) 発表論文集, pp. 546-549, 2019/3/13, 名古屋.
2. 山城颯太，西川仁，徳永健伸，分散表現による大規模格フレームの汎化を利用した統合的ニューラルゼロ照応解析，言語処理学会 第24回年次大会 発表論文集，pp. 588-591, 2018/3/13，岡山.
3. Takenobu Tokunaga, Hitoshi Nishikawa and Tomoya Iwakura, An eye-tracking study of named entity annotation, Proceedings of Recent Advances in Natural Language Processing (RANLP 2017), pp. 758–764, 2017/9/4, Varna (Bulgaria).
4. 牧諒亮，西川仁，徳永健伸，視線情報を用いた述語項構造解析モデルへの単語分散表現の導入, 言語処理学会 第23回年次大会 発表論文集, pp. 605-608, 2017/3/13, 筑波.
5. 山城颯太，西川仁，徳永健伸，分散表現による格フレームの格要素の汎化を利用したゼロ照応解析，言語処理学会 第23回年次大会 発表論文集, pp. 206-209, 2017/3/13, 筑波.
6. 牧諒亮，西川仁，徳永健伸，視線情報を用いた日本語述語項構造解析モデルのパラメータ推定，情報処理学会自然言語処理研究会，Vol.2016-NL-229，No.8，pp.1-8,

2016/12/21，東京.

7. Ryosuke Maki, Hitoshi Nishikawa and Takenobu Tokunaga, Parameter estimation of Japanese predicate argument structure analysis model using eye gaze information, Proceedings the 26th International Conference on Computational Linguistics (COLING 2016), pp. 2861-2869, 2016/12/13，大阪.

〔図書〕（計1件）

1. Takenobu Tokunaga, Ongoing efforts: Toward behaviour-based corpus evaluation, in Nancy Ide and James Pustejovsky Ed. Handbook of Linguistic Annotation, Springer, 2017. June.

〔産業財産権〕

○出願状況（計0件）
○取得状況（計0件）

〔その他〕

なし

６．研究組織

(1)研究分担者

研究分担者氏名：相澤 彰子
ローマ字氏名：Aizawa Akiko
所属研究機関名：国立情報学研究所
部局名：コンンツ科学研究系
職名：教授
研究者番号（8桁）：90222447

研究分担者氏名：西川 仁
ローマ字氏名：ニシカワ ヒトシ
所属研究機関名：東京工業大学
部局名：情報理工学院
職名：助教
研究者番号（8桁）：00765026

(2)研究協力者

なし