テキストアノテーションにおける視線と操作履歴の収集と分析

光田 航 †　　飯田 龍 ‡　　徳永 健伸 ‡
† 東京工業大学 工学部情報工学科
‡ 東京工業大学 大学院情報理工学研究科
mitsuda.k.aa@m.titech.ac.jp, {ryu-i,take}@cl.cs.titech.ac.jp

1 はじめに

現在の自然言語処理では，コーパスと機械学習の枠組みを利用するアプローチが主流である．このアプローチでは，個々の問題に対応する正解（タグ）を人手でアノテーションしたコーパスを用いて学習を行い，その結果得られるモデルを用いて未知の入力に対して同様のタグを自動的に付与する処理を行う．この経験主義的な研究パラダイムに基づく研究は約20年間にわたって発展しており，自然言語処理技術の性能向上に大きく貢献してきた．特に，形態素・統語解析などの比較的表層的な言語処理の問題においては高い性能を得ている．しかし，一方でより深い言語処理（例えば，意味・談話処理）ではこのアプローチを採用しても必ずしも十分な性能が得られるとは限らない．

これは，形態素・統語解析では語や句，品詞などの表層的な情報から求めたい結果が得られるのに対し，意味・談話関係ではそもそもどのような情報が必要となるかが自明ではないことに起因する．このような問題があるにもかかわらず，既存の自然言語処理の典型的なアプローチでは，語や品詞などの表層的な情報を近似的に解析に必要な情報として利用することで問題を解こうしており，この結果，本質的に問題の解決に必要な情報が何であるかについて追求できていない．このような背景から，本研究では意味・談話解析のような深い言語の処理において，実際に人間が利用している知識を解明し，それを解析に利用することで高い性能を得ることを目指す．

この目的を達成するための方法の一つとして，人手によるコーパス構築におけるアノテーションのプロセスを解明することで，その結果を自動解析に利用することを考える．典型的なコーパスの構築では，個々の問題に対してシステムが出力すべき情報を人手でアノテーションするだけだが，ここでアノテーションされたタグは，アノテーション作業者が何らかの言語的な判断を行った結果得られた結果にすぎず，作業者がその判断をするに至るまでの過程は捨象されている．このため，もしこの判断に至る過程を何らかの形式で獲得することができれば，人間の判断の根拠となる有益な情報が得られる可能性がある．

そこで，本稿では，このような作業者のアノテーションの判断の根拠となる情報の獲得を考える．飯田ら [8] が議論しているように，この判断の根拠の情報として，作業者の行為や視線などのような観察可能なデータと，対話的なアノテーションの環境を構築することによって得られる潜在的なデータの二通りが考えられるが，ここでは特に観察可能なデータ[*1]収集に着目し，実際にあるアノテーション課題における作業者の作業の行為，それにともなう視線などを収集し，そもそもそのような客観的に獲得できる情報から作業者の判断に関する情報が得られるのかを調査する．2節で観察可能なデータとして記録する情報の概要を説明し，3節でデータ収集のための実験設定と実際に記録するデータの形式について述べる．4節で実験の結果得られたデータの詳細とその分析結果を示す．5節で関連する既存研究を紹介し，6節でまとめと今後の展望について述べる．

2 観察可能なデータとして記録する情報

アノテーション作業者が実際に作業を行う過程は以下の3つの段階に分けて考えることができる．

1. 提示されたテキストから言語的な情報を読み取る．
2. 読み取った情報からアノテーションに関する判断を行う．
3. 2. の判断に基づきアノテーション作業を行う．

このうち，客観的に観測可能な情報は，1. と 3. が該当する．1. については作業者のテキストへの注視を記録し，3. についてはアノテーション中に作業者が行った操作を記録することで，客観的な情報として記録が可能である．以降で，それぞれの情報をどのように記録するかについて説明する．

2.1 視線情報の記録

アノテーション作業中の作業者の視線データの記録には視線追跡装置 Tobii T60 を利用して記録する．Tobii T60 は 1,280×1,024 のサイズのモニタを持ち，作業者が画面上のどの位置を見ていたかを 1/60 秒で記録することができる．この視線追跡装置を用い，作

[*1] 飯田ら [8] では「非明示的なデータ」と呼ばれている．

業者の作業開始から終了までの視線をすべて記録することで，作業者がアノテーション中にどこを見ながらアノテーションを行ったのかを分析することが可能になる．

2.2 操作情報の記録

4 節で後述するように，本研究では作業者がアノテーション中に行った判断にどの程度の時間がかかったのかを分析するため，アノテーションに特化した操作を記録する．アノテーション作業に特化した操作は，例えば，述語項構造のアノテーションを考えた場合，項や述語のセグメントをアノテーションする，また項と述語の間の関係ラベルをアノテーションするという操作に対応する．これらの操作は，アノテーションツールで何らかの操作を行うことで達成されるため，ツール上でユーザが行った操作が何に対応するかを把握し，それを記録していくことで操作の履歴を収集できる．

本研究では，後述するようにアノテーションの課題として述語項構造のアノテーションを行うため，その作業が可能なアノテーションツールを利用する必要がある．前述したように，この作業には述語やその項となるセグメントの範囲をアノテーションする機能と，そのセグメント間に関係のラベル付きでリンクを付与する機能が必要となる．この 2 つの機能が実装されたアノテーションツールの一つに Slate[2] がある．Slate ではマウスによる操作で付与するタグの選択，セグメントのアノテーション，セグメント間のリンクのアノテーションが可能である．ただし，公開されている Slate では作業者の操作履歴を記録する機能がないため，操作履歴を記録するための機能を実装したものを本研究のデータ収集で利用する．また，公開版の Slate ではタグ選択のインタフェースやテキストの ID などのメタ情報が画面上に表示されており，それらを作業者が見ることによって，アノテーション作業者がどのように文章を読んでアノテーションを行っていくかという過程を記録するのを阻害する可能性がある．そこで，述語項構造のアノテーションに必要なタグはキーボードのキーを入力することで変更可能なように実装し，作業中の画面上にはアノテーション対象となる文章のみを表示するように実装した．

3 データ収集実験の詳細

2 節で述べた情報を適切に記録するためには，適切なアノテーション課題を選択し，また実験に関して細かな実験条件を設定する必要がある．この節ではその詳細と実験の際に記録する作業者の操作に関するデータのフォーマットについて説明する．

3.1 アノテーション課題

アノテーションの課題として極端に簡単な問題を採用した場合，アノテーション作業が単調になり，そこから得られる客観的情報から作業者の判断に関する分析を行うことが難しい．このため，作業を行う内容には判断が難しい事例や曖昧性を含む事例が適度に含まれることが望ましい．そこで，本研究でまずデータ収集を行う課題として，述語項構造のアノテーションを考える．述語項構造のアノテーションでは項と述語の関係を付与する際，項が省略されている場合にはゼロ照応の関係として前方の文脈全体から先行詞を求める課題となり，比較的複雑な課題が含まれることになる．

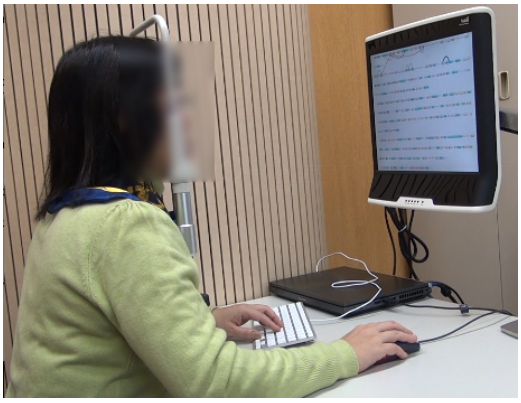

図 1: 実験環境

本研究で行う述語項構造のアノテーションの作業では，NAIST テキストコーパス [9] を構築する際に採用された仕様を単純化して作業を行う．具体的には，述語とその項の候補となる名詞句についてはあらかじめセグメントとして表示しておき，作業者は述語と項のセグメント間にガ格，ヲ格，ニ格のラベルが付いたリンクをアノテーションする．この際，NAIST コーパスにおけるアノテーションと同様に，述語が受動態で出現していても述語の原形に関する格要素をアノテーションする．

アノテーション対象のテキストには，様々な種類の記事が含まれる現代日本語書き言葉均衡コーパス（BCCWJ）[3] の書籍コーパス（可変長）を利用する．文章頭のタイトルは取り除き，それ以降の文章を画面上に表示してアノテーションを行う．コーパス中の文章の中には文章の途中で見出し（title タグで記述される内容）が入ることで，談話的な分断が起こるものが含まれており，そのような文章はゼロ照応関係をアノテーションする述語項構造の課題を行う文章として不適切である．そこで，画面上に表示する文章の中にそのような記述が含まれる記事はアノテーション対象から除外した．

3.2 実験環境

データ収集の実験では，視線追跡装置である Tobii T60，マウスとキーボードを用いて作業を行う．また，視線の計測精度を上げるため，顎台を使用する．作業環境のスナップショットを図 1 に示す．

作業者がアノテーション中にどこを見ているかを分析するためには，記録した視線データをテキストの適

切な区間に対応づける必要がある．この対応付けを簡単にするため，今回はテキストの表示画面がスクロールされないように Slate を修正した．この結果，一回の実験で作業できるテキストの量は一度に画面に表示できる量までに限定されるため，できるだけ多くのテキストを表示することが望ましい．一方でフォントサイズや行間を小さく設定した場合，計測誤差によって視線との対応付けが困難になるため，これらのバランスを考えた設定が必要になる．このため，この調整のための予備実験を行い，経験的にフォントサイズを 16 ピクセル, 行間を 66 ピクセル, 文字間隔を 0 ピクセルに設定した．この設定の結果，画面内に表示される文字数は約 1,000 文字となった．

3.3 作業履歴の出力フォーマット

アノテーションに関連する操作はタイムスタンプとともに記録される．記録される操作の種類はアノテーションの開始や終了, リンク種類選択, リンク作成の開始，終了, リンク削除など 10 種類のアノテーション操作である．以下に記録結果の例を示す[*2]．

```
@ANNOTATION_ACTION
1353460009.604 annotation_start  *,*,*,*,*,*
1353460018.345 select_tagdef     *,ga,*,*,*,*
1353460019.593 create_link_start 1,ga,993,990,*,*
1353460021.029 create_link_end   1,ga,993,990,*,*
...
1353461483.536 annotation_end    *,*,*,*,*,*
```

4 データ収集実験と予備的な分析

3 節で述べた実験設定に基づきデータ収集を行い，その結果得られたデータを利用した予備的な分析を行った．

4.1 収集対象

実験の被験者として述語項構造のアノテーション経験を持つ 3 人の作業者を雇用し， 3 節に示した実験環境で述語項構造のアノテーションを行った． 3.1 に示した基準で BCCWJ の書籍コーパスから選別した文章のうち，ランダムに選択した 43 記事を作業対象とした．実験は数日に分けて行い，実験日の間が空きすぎない様に配慮した．1 つの文章の作業には約 20 分かかり，数記事作業してもらった後には休憩を挟むことで，作業者に過度に負担がかからないよう配慮した．また，作業者にアノテーションツールに慣れてもらうために，データ収集を行う場合と同様の作業を行ってもらい，その後に実際のデータ収集の実験を行った．

4.2 リンクの一致率と作業時間に関する分析

収集したデータを用いてさまざまな分析が可能であるが，本研究ではまず，作業者の作業の一致率と作業者が作業をするためにかけた時間との対応関係を調査した．もし作業者がアノテーションに多く時間をかける場合に作業の一致率が低下するならば，作業時間がかかる事例は作業者の判断の揺れる難しい事例であることがわかる．つまり，作業時間のかかる事例のみを分析することでその問題を解くために本質的に重要な問題を吟味できることになる．

データ収集の実験を行った 3 人の作業者 A_0, A_1, A_2 が 43 記事に対して作業した結果，付与された全リンクの総数はそれぞれ 3,353(A_0), 3,764(A_1), 3,462(A_2) となった．以降の分析では，述語を注視した後に項を探す行為を分析するために, ある述語の一つの格について複数の項がアノテーションされるような場合を除いて分析に利用する．上述のリンク総数のうち，この条件に該当する事例を除いたリンク数はそれぞれ 3,054(A_0), 3,251(A_1)，2,996(A_2) となった．

さらに，ここでは作業者間の作業時間を比較するため，そもそも作業者がある述語の格についてアノテーションを行っている必要がある．そこで，3 人すべてがアノテーションしている述語の格のみに対象を限定して分析を行う．この 3 者がすべてアノテーションしたリンク数は 2,209 であり，このうち述語と格のペアに対して 3 人とも同じ項を付与した数は 1,952，2 人以上が同じ項を付与した数は 2,184 であった．つまり，3 者が完全一致した事例の割合は 0.884 であり，2 者以上の一致の割合は 0.988 となる．

また，一致率との相関を調査するために作業者の各事例に関する作業時間を見積る必要があるが，この作業時間を「直前のリンク作成後，初めて注目する述語に注視（停留）が起きてから，リンクの付与が終了するまでにかかった時間」と定義した．視線データから停留を抽出する際には，Dispersion-Threshold Identification (I-DT) アルゴリズム [5] を利用し，各停留がどの行を見ているかは人手で対応付けを行った．上記の 2,209 のリンクのうち，アノテーション時間が定義できたリンク数は 1,751 であり，その割合は 0.793 であった[*3]．以降で，ここで定義した作業時間とある作業時間の区間内に作業された結果の一致率の関係を図 2 に示す．

図 2 の上部に示すグラフは，リンクのアノテーション時間（x 軸：0.5 秒区間）と，一致したリンクのインスタンス数（y 軸）の関係を表している．agree-3, agree-2, noagree はそれぞれ，3 者一致, 2 者一致, 一致なし，のリンクに対応している．アノテーション時間には 3 人の作業者のアノテーション時間の平均を用いている．この図から，リンクの大部分が短いアノテーション時間で付与されていることがわかる．

図 2 の下部に示すグラフは，リンクのアノテーショ

[*2] 'ga' はガ格のリンクを表しており，リンクを作成した際に同時に記録される数，1, 993, 990 はそれぞれ作成されたリンクの ID, リンク元セグメントの ID, リンク先セグメントの ID を表す．

[*3] 述語に一度も停留が起きなかった場合，アノテーション時間の開始点が定義できないため 458 のリンクについては作業時間を求めることができなかった．

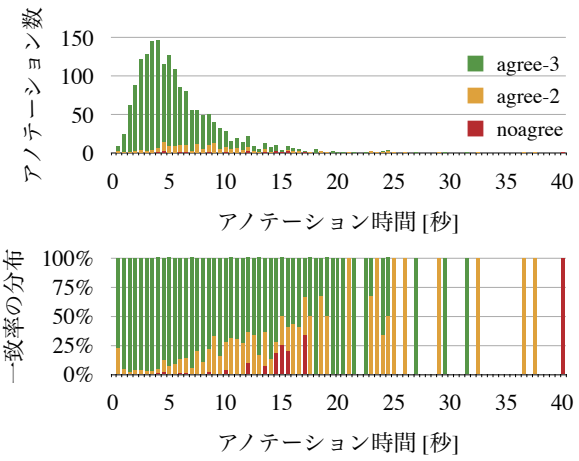

図 2: アノテーション作業時間と一致率の関係

ン時間（x 軸：0.5 秒区間）と，その区間での一致の程度の分布（y 軸）を示している．このグラフよりアノテーション時間が長くなるにつれて，付与されたリンクの一致率が減少していることがわかる．この結果より，アノテーション時間が長くかかる場合にはそのアノテーションは難しく，結果的に作業者間の揺れが大きくなる傾向があると考えられる．この傾向は，アノテーションの信頼性を見積る場合に，ゴールドスタンダードや作業者間の一致率を見ずに推定できる可能性を示唆している．

5 関連研究

ここでは本研究がアノテーション課題と視線情報を扱うことから，作業者が何らかの問題を解決する際の視線情報の分析を行った研究について述べる．

Rosengrant[4] は物理の問題を解く際の玄人と素人の振舞いについて，解答プロセスの発話内容と視線情報を組み合わせた gaze scribing と呼ばれる分析方法を提案した上で，両者の間に特徴的な違いがあることを実験から示している．

Bednarik ら [1] はプログラムのデバッグ実験において玄人と素人が注視する画面を 3 つの領域に分け，両者の傾向の違いを報告している．しかし被験者の着目対象を扱う粒度が粗いため，被験者の傾向分析には効果的に分析を行える一方，デバッグが成功した原因やその際の思考過程に関する分析を行うことは難しい．

Tomanek ら [6] はテキストアノテーションのコスト推定のために固有名詞を分類する課題設定において，テキスト領域を 5 つに分けて作業者の注視時間や注視パターンを分析し，コスト推定モデルを構築している．

本研究は作業者が行った判断の根拠を獲得することを目的としているため，これらの研究と比べて注視の対象をより細かい語や句の単位で扱う必要があり，収集や分析においても異なる研究課題を解決しなければならない．

6 おわりに

本稿では，アノテーションに関する新たな試みとして作業者の判断に関する根拠を記録する方向性を示し，客観的に収集可能なデータを収集し分析した結果について述べた．述語項構造のアノテーションを対象に 3 人の被験者に作業を行ってもらい，その作業時の視線と作業時の操作を記録し，その結果を分析することで作業にかける時間と作業の難しさについて相関があることを示した．

ここで提案している方法論は，現在のアプローチが抱える問題を解決する大きな一つの指針になると考えている．今回は予備的に，客観的に観測が可能な情報に限定して小規模なデータ収集を行ったが，将来的には作業者が言語の処理を行う際に着目している点についてプロトコル分析などを用いて口述させることで主観的な情報も記録することを考えている．このようにして得られる客観的もしくは主観的な人間の言語に関する判断の情報を利用し，さまざまな言語処理へと活用していく予定である．プロジェクトの詳細は飯田ら [8]，徳永ら [7] で議論されている．

参考文献

[1] Roman Bednarik and Markku Tukiainen. Temporal eye-tracking data: Evolution of debugging strategies with multiple representations. In *Proceedings of the 2008 symposium on Eye tracking research & applications (ETRA '08)*, pp. 99–102, 2008.

[2] Dain Kaplan, Ryu Iida, and Takenobu Tokunaga. Slate - a tool for creating and maintaining annotated corpora. In *Journal for Language Technology and Computational Linguistics*, Vol. 26, pp. 89–101, 2012.

[3] Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. Design, compilation, and preliminary analyses of balanced corpus of contemporary written japanese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pp. 1483–1486, 2010.

[4] David Rosengrant. Gaze scribing in physics problem solving. In *Proceedings of the 2010 symposium on Eye tracking research & applications (ETRA '10)*, pp. 45–48, 2010.

[5] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications (ETRA '00)*, pp. 71–78, 2000.

[6] Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Ju rgen Ziegler. A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1158–1167, 2010.

[7] 徳永健伸, 飯田龍. アノテーションのためのアノテーション. 言語処理学会第 19 回年次大会発表論文集, 2013.

[8] 飯田龍, 徳永健伸. アノテーション作業者の内省を顕在化するためのデータ収集. テキストアノテーションワークショップ・コンテスト, 2012.

[9] 飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治. 述語項構造と照応関係のアノテーション: Naist テキストコーパス構築の経験から. 自然言語処理, Vol. 17, No. 2, pp. 25–50, 2010.