

# 視線情報を利用した欠損アノテーションの検出

光田 航<sup>1,a)</sup> 飯田 龍<sup>1,b)</sup> 徳永 健伸<sup>1,c)</sup>

**概要:** 本稿では、複数人が行った述語項構造アノテーション作業の不一致を検出する問題について議論する。特に、文章中のある述語-項関係に対して二人の作業者のうち一方の作業者がアノテーションしないというアノテーションの不一致(アノテーション欠損)を検出する問題を考える。アノテーション欠損を検出するために、言語的な情報に加えて、アノテーション作業者の視線情報を利用する。具体的には、アノテーション対象となる述語に関して収集した注視の系列から高頻度の視線のパターンを抽出し、それをアノテーション欠損を検出するための素性として利用する。これまでに収集した視線情報を含むアノテーション結果を用いて評価実験を行い、各素性の有効性を調査した。この結果、視線情報と言語的情報がともにアノテーション欠損検出に有効であり、また、特定の視線パターンが欠損検出の良い指標になることについて報告する。

## 1. はじめに

近年、教師あり学習手法が発展したことにより、正解データを作成するためのテキストアノテーションをどのように実現するかが自然言語処理の分野で重要な課題となっている [18]。テキストアノテーションでは文章に正解として出力したい情報(タグ)を付与するが、アノテーション作業の品質が直接的に教師あり手法の性能に影響するため、多くの研究者が高品質なアノテーションを低いコストで実現することに関心を示している。この目的のために、人間のアノテーションと既存の言語処理のツールを組み合わせた半自動のデータ構築手法 [5], [12], [15], [22] が提案され、また、アノテーションを効率的に行うことができるアノテーションツール [9], [10], [11] が開発されている。

このような背景から、アノテーション品質の評価もコーパス構築を行う上で重要な課題となる。この品質評価に関しては、複数のアノテーション作業者が同一のアノテーション対象に対してタグ付与作業を行った結果の一致率に基づいて評価されることが多く、作業員間の一致率に基づくさまざまな評価尺度が提案されている [1], [3], [8], [13]。このように、既存のコーパスの品質評価などの研究では、作業者が行ったアノテーションの結果のみに依存した観点から研究が進められてきたが、本研究ではこれに対し、作業の結果に加え、アノテーション作業中の作業員の振舞い

に着目し、この振舞いを分析することで自然言語処理の問題を解決するために必要となる情報を明らかにすることを目的とする。本研究で扱う内容は被験者の行為の系列にマイニング技術を適用する行為マイニング [4] で広く研究されている内容と関連するが、人間がアノテーションする過程に着目し、それを分析するという問題の捉え方をした研究は少ない。例外的に、Tomanek ら [20] はアノテーション時のアノテーション作業員の視線情報を利用し、固有名詞のアノテーションの各インスタンスの難易度を推定する問題に取り組んでいるが、彼らのアプローチでは「アノテーション対象となる述語」と「それ以外の文脈」の2つの領域への注視にしか着目していない。しかし、より精密な分析を行うためには、分析対象となる問題に適合した粒度の細かい視線の動きを観察・分析する必要がある。

我々は先行研究において日本語の述語項構造アノテーションに関する作業員の視線とアノテーションツールの操作履歴を収集している [19], [23]。この研究では、収集したアノテーション作業員の行為に関するデータを用いて作業結果の一致率を推定する問題を扱った。視線と操作履歴から各インスタンスに関する作業時間を定義し、作業時間が長くなるほど一致率が低くなる傾向があることを明らかにした。この結果が示すように、アノテーション作業員の行為を分析することで、これまで複数人が作業員した結果に基づき推定していた一致率、つまりアノテーション品質を単一の作業員の結果から推定できる見込みがある。

我々のこれまでの分析の結果、述語項構造のアノテーションの作業員間の不一致は、ある述語のある格に対して、二人の作業員のうち一方の作業員がアノテーションしない

<sup>1</sup> 東京工業大学  
Tokyo Institute of Technology  
<sup>a)</sup> mitsudak@cl.cs.titech.ac.jp  
<sup>b)</sup> ryu-i@cl.cs.titech.ac.jp  
<sup>c)</sup> take@cl.cs.titech.ac.jp

場合に頻出していることがわかった。本研究ではこの種のアノテーションの不一致を**アノテーション欠損**と呼ぶ。アノテーション欠損に関しては過去の我々の研究 [19], [23] では分析の対象外としていたため、このアノテーション欠損を検出することはアノテーション作業者の振舞いに基づくアノテーション品質の評価のために重要な課題となる。そこで、本研究では、アノテーション対象となる述語から得られる言語的な情報とアノテーション作業者の視線情報を利用してアノテーション欠損検出の問題を解く。特に、視線情報を注視の系列に変換し、マイニング技術を利用することでこの注視の系列からアノテーション欠損を検出するためのパタン発見を試みる。本稿では、まず、2節で日本語述語項構造のアノテーションに関する視線情報収集の実験について紹介し、次に3節で本研究で対象とするアノテーション欠損検出の問題設定について説明する。4節で視線データに基づくアノテーション欠損検出モデルを提案し、5節でモデルの性能を調査するための評価実験の結果を報告する。6節で関連研究を概観し、7節でまとめと今後の課題を述べる。

## 2. 視線情報の収集

### 2.1 データと収集の手続き

本研究では既存研究 [19], [23] で収集した日本語述語項構造アノテーションにおける作業者の作業履歴と視線情報を用いる。この収集実験における述語項構造のアノテーションでは、自動解析の結果に基づいて述語と項の候補の範囲(セグメント)はあらかじめ付与しておき、作業者はキーボードで付与するリンクの種類(ガ格, ヲ格, ニ格)を選択し、マウスのドラッグで述語とその項の間に関係(リンク)を付与する。

アノテーションにはセグメントとそのセグメント間のリンクを付与できるアノテーションツール Slate<sup>[9]</sup>を使用した。図1に Slate で提供されるアノテーションの GUI のスクリーンショットを示す。図1では、述語を表すセグメントは背景が青色の長方形でタグ付けされており、項候補が赤の枠線の長方形でタグ付けされている。リンクの色は関係の述語-項関係の種類を表し、赤色がガ格、青色がヲ格、緑色がニ格を表す。

また、現在公開されている版の Slate<sup>\*1</sup>には操作履歴を記録する機能がないため、表1に示す8種類のアノテーション作業者の操作を、操作時刻や操作にともなうセグメントの情報とともに記録できるようツールの修正を行った。

アノテーション時の作業者の視線の記録には視線計測装置 Tobii T60 を利用した。Tobii T60 のディスプレイサイズが17インチで解像度が1,280×1,024であり、ディスプレイと作業者の間の距離は約50cmになるよう調整した。ア

表1 アノテーションツールで記録する作業者の操作

操作の種類	説明
create_link_start	リンク付与を開始 (マウスドラッグ開始)
create_link_end	リンク付与を終了 (マウスドラッグ終了)
select_link	リンクを選択
delete_link	リンクを削除
select_segment	セグメントを選択
select_tag	タグ (付与するリンクの種類) を選択
annotation_start	アノテーションを開始
annotation_end	アノテーションを終了

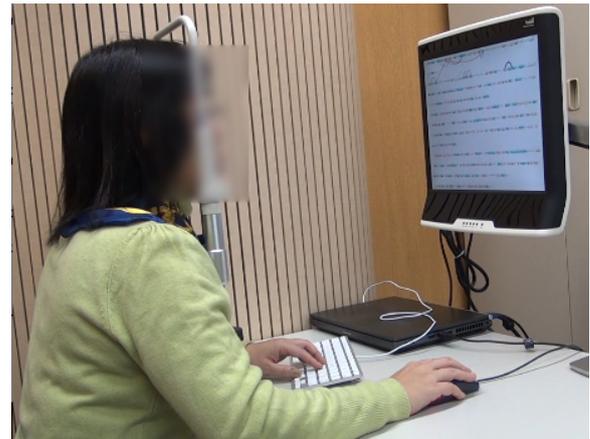


図2 Tobii T60 を使ったアノテーションのスナップショット

ノテーションを行う前にはキャリブレーションを行い、また、アノテーション作業時には視線検出のエラーを抑制するために、図2のように顎台を用い、作業者の頭の動きを固定した状態でデータ収集を行った。

このデータ収集実験では、述語項構造アノテーションの経験のある3人のアノテーション作業者を雇用了。各アノテーション作業者は現代日本語書き言葉均衡コーパス(BCCWJ)の書籍コーパス(PB)から選択した43記事<sup>\*2</sup>を対象にアノテーション作業を行う。また、作業中に画面のスクロールが起こった場合、収集した視線情報と画面中の文字との対応付けが困難になるため、画面に表示される文字数を約1,000文字に制限し、画面のスクロールが起こらないようにした。また、アノテーション作業者は記事単位で作業を行い、記事全体への作業が終了した後は必要に応じていつでも休憩をとることができ、休憩後に作業を始める際は毎回キャリブレーションを行うようにした。

### 2.2 作業結果

3人のアノテーション作業者 A<sub>0</sub>, A<sub>1</sub>, A<sub>2</sub> が述語-項関係をアノテーションした結果、関係の数(リンク数)はそれ

<sup>\*2</sup> 記事の選別では、記事の先頭から1,000文字程度抽出する際に、節見出しなどの文章の断絶を引き起す要因が含まれない記事のみを抽出し、さらにその記事集合の一部に対して第一著者もしくは第二著者があらかじめアノテーションを行い、局所文脈だけを参照することでアノテーション可能な事例が頻出する記事を除外したものをランダムに43記事を選択した。

<sup>\*1</sup> <https://bitbucket.org/dainkaplan/slate/>

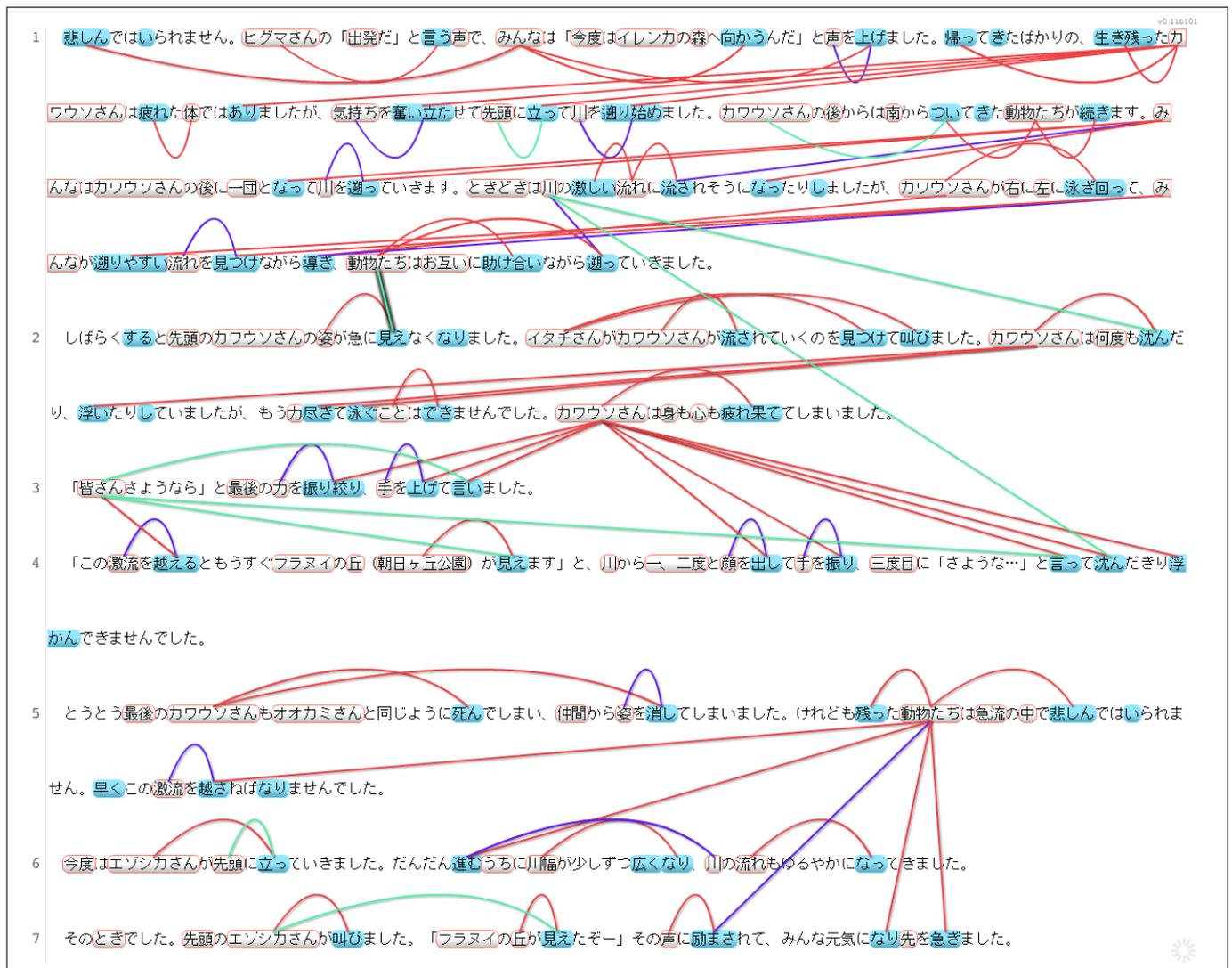


図1 アノテーションツール Slate のスクリーンショット

ぞれ 3,353 ( $A_0$ ), 3,764 ( $A_1$ ), 3,462 ( $A_2$ ) となった。この作業結果のうち、一つの述語のある格に複数の格要素が付与されたものなど、例外的なアノテーション結果を除外した結果、各作業者の関係の数はそれぞれ 3,054 ( $A_0$ ), 3,251 ( $A_1$ ), 2,996 ( $A_2$ ) となった。この3者の作業者の作業内容を確認したところ、作業員  $A_1$  が頻繁に格交替に関するアノテーションを誤っていることがわかったため、本研究で扱う調査対象から除外した。さらに、各作業者のヲ格・ニ格のアノテーションについては本来ならば下位範疇化構造が記述された辞書を用意し、それを参照しながら作業を進めるべきであるが、その場合には辞書と文章の両方を画面に表示する必要があり、視線の分析が困難になるという問題が起こる。このため、実際のアノテーション作業では図1に示すように文章のみを表示してアノテーションしたが、ヲ格・ニ格へのアノテーション結果には誤りが多く含まれることになった。このため、以降の分析をガ格のみに限定して行うこととする。

### 3. アノテーション欠損の検出課題

述語のガ格は例外的な場合を除いて必ず述語の必須格として想定できるため、ガ格を文章全体から網羅的に探索するアノテーション課題は単純な問題に見える。しかし、ガ格のアノテーションでは、項が省略されている場合に前方文脈に適切な項の候補が無い場合述語-項を付与しない場合と、述語を含む表現が機能語相当であるためアノテーションの対象から除外する場合の2つの特殊な状況が存在し、かつそれに対する判断が作業員によって揺れるため、ガ格を付与するか否かの一致率は低くなる。例えば、前者の項の省略に関しては、アノテーションの対象となる述語に対して文章中の全ての項候補を把握し、各項候補がガ格となるか否かを判断する必要があり、文脈が複雑で多くの項候補が出現する場合には確認すべき述語と項の組み合わせが爆発的に増加するために、その判断の漏れが生じて作業結果が一致しづらくなる。また、後者の機能語相当表現にア

表2 作業員 A<sub>0</sub> と A<sub>2</sub> がガ格をアノテーションした結果

A <sub>0</sub> \ A <sub>2</sub>	付与する	付与しない
付与する	1,534	312
付与しない	281	561

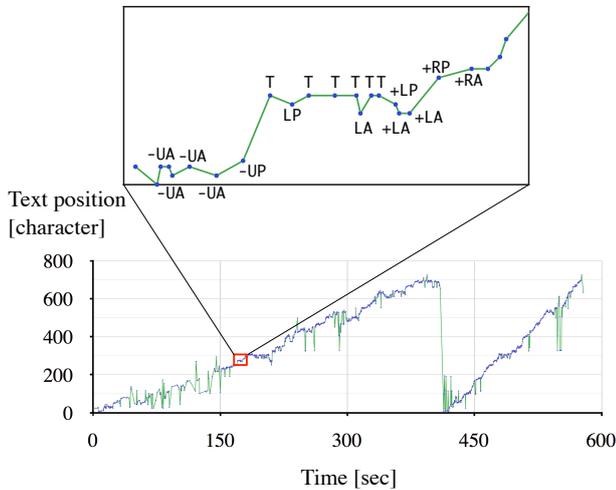


図3 アノテーション時の注視の遷移の例

アノテーション対象となる述語が含まれる場合（例えば、「彼に加えて」など）では、機能語相当表現とすべきか否かの判断が作業員によって揺れるため、作業結果が一致しづらい。この一致率を調査するために、2人の作業員 A<sub>0</sub> と A<sub>2</sub> の間でガ格のみをアノテーション対象にした場合の作業の一致の分布について調査を行った。結果を表2に示す。表2からわかるように、例えば、一方の作業員 (A<sub>0</sub>) の作業結果全体の約 17% (312/(1,534 + 312)) をもう一方の作業員 (A<sub>2</sub>) はアノテーションできておらず、このために作業全体の一致率が低下することになる。そこで、本研究では、一方の作業員の作業結果の中で述語のガ格が付与されていない事例 (561+312 事例もしくは 561+281 事例) のうち、もう一方の作業員がアノテーションした事例 (312 事例もしくは 281 事例) を検出する問題を解く。

#### 4. アノテーション欠損の検出モデル

アノテーション作業員の視線の動きは、例えば、アノテーション対象の述語とその項の近傍を視線が集中したり、また、文章を読み進める際にある述語を読み飛ばすなど、作業員の関心を反映した動きをすると考えられる。このため、視線がどのセグメントに対してどのように遷移するかという特定の視線のパターンがアノテーションの欠損を検出する手がかりになると考えられる。

そこで、本研究では、アノテーション時の視線の特定の動きのパターンを捉えるために、アノテーション中の注視<sup>\*3</sup>の遷移について調査を行った。図3はある文章を対象に作業

<sup>\*3</sup> 注視は Dispersion-Threshold Identification アルゴリズム [17] を用いて求めた。

者がアノテーション作業を行ったときの注視の遷移を表している。図3の横軸はアノテーション開始時点からの時間経過を表し、縦軸は注視の文章中の位置（文章開始位置からの文字単位のオフセット）を表している。図3に示すような時間と注視の位置に関するグラフを複数の作業結果について調査したところ、どの作業結果でも注視の遷移は同じ傾向にあることがわかった。まず、アノテーション対象となる述語に関する注視（例えば、図3の拡大図のTの系列）は狭い期間に集中して出現していることがわかった。このため、アノテーション対象となる述語に関する局所的な注視のみを分析することで、アノテーション欠損を検出するために有益な視線のパターンを抽出できる可能性がある。また、作業員によっては一度文章の最後まで作業を進めた後に、作業結果の確認を行う場合がある。例えば、図3では作業開始から約410秒経過したときに、注視が文章頭に戻っており、これ以降が一度作業した結果の確認に相当する。ただし、この確認作業では、必ずしもアノテーション対象となる述語に関して有益な注視のパターンが存在するとは限らないため、本研究ではアノテーション開始から文章の末尾まで動く最初の注視の遷移（図3の例では0秒から約410秒まで）のみから注視のパターンを抽出する。

本研究では、以下に示す3つの手順に従って注視の系列からその系列に特徴的な視線のパターンを抽出する。

- (1) まず、アノテーション対象となる述語に関する注視の期間（作業期間）を対象述語ごとに同定する。
- (2) 次に、作業期間内の注視の系列を、注視の特徴を表す記号の系列に変換する。
- (3) 最後に、テキストマイニングの技術を適用し、記号の系列の集合から頻出する記号のパターンを抽出する。

まず、手順(1)では、文章中の各述語に対して固定のウィンドウサイズで注視の系列を走査する。この際、我々の収集したデータを用いた予備的な分析に基づき、ウィンドウサイズを40の連続した注視を必ず含むような区間とした。次に、走査した範囲で最も多く述語に関する注視を含んでいるウィンドウを決定する。ここで最大の注視の個数が含まれるウィンドウが複数存在する場合は、よりアノテーション開始時間に近いウィンドウを選択する。さらにウィンドウ内で最初と最後に述語に注視した箇所を検出し、さらに最初に注視より前の5つの注視と最後の注視から後の5つの注視を含めた時間を作業期間として定義する。図4に作業期間の定義を示す。

次に、手順(2)で作業期間内の各注視をあらかじめ定義しておいた記号に変換する。アノテーション対象となる述語とその近傍のセグメントへの注視は、セグメント間の相対的な位置・時間関係やセグメントの種類によって特徴付けられると考えられる。例えば、対象述語への注視の後に左

表4 アノテーション欠損検出のための素性

タイプ	素性	説明
ling	is_verb	対象述語が動詞の場合は 1, それ以外は 0.
	is_adj	対象述語が形容詞の場合は 1, それ以外は 0.
	lemma	対象述語の見出しの基本形.
gaze	gaze_pat <sub>i</sub>	対象述語に関する注視の系列に 4 節で抽出された視線パターンが含まれる場合は 1, それ以外は 0.

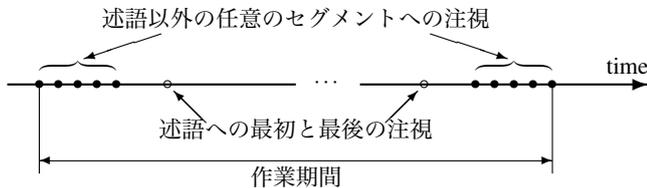


図4 作業期間の定義

表3 視線パターンを表す記号の定義

カテゴリ	記号
位置	上 (U), 下 (B), 右 (R), 左 (L)
セグメントの種類	対象述語 (T), それ以外の述語 (P), 項の候補 (A)
出現した時刻	作業期間の前に出現 (-), 作業期間より後に出現 (+)

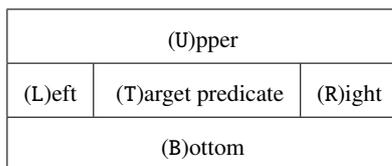


図5 視線の領域の定義

側の項候補へ注視が遷移するという視線の動きは、述語の項を探すための読み返しに相当すると考えられる。このような視線の動きを捉えるために、各注視から表3に示す3つのカテゴリに分類された情報を抽出する。また、注視の位置に関しては図5に定義された位置に従って注視の位置のラベル付けを行う。例えば、対象述語の左側に出現する項候補への注視は記号“LA”で表現される。この結果、作業期間内の注視の系列は、例えば、“-UA -UA -UA -UP T LP T T T LA T T +LP +LA +LA +RP +RA”のような記号の系列に変換される。

最後に、手順(3)で prefixspan アルゴリズム [14] を用い、手順(2)で作成された記号の系列の集合から頻出する記号のパターンを抽出する。prefixspan は可能なパターンの完全集合を効率的に抽出するマイニング手法であり、これを利用することで注視を変換した記号の系列から効率的に頻出する記号のパターンを抽出可能となる。我々のアノテーション欠損検出モデルでは、抽出された頻出パターンはアノテーション欠損検出のための素性として利用する。また、アノテーションの欠損はある特定の動詞や形容詞に偏っていると考えられるため、語彙的な素性は分類に有効であると考

表5 アノテーション欠損検出の結果

	(正解:A <sub>0</sub> , システム:A <sub>2</sub> )			(正解:A <sub>2</sub> , システム:A <sub>0</sub> )		
	R	P	F	R	P	F
ベースライン	1.000	0.358	0.527	1.000	0.333	0.500
ling 素性のみ	0.933	0.402	0.562	0.846	0.467	0.599
eye 素性のみ	0.997	0.358	0.527	0.964	0.342	0.505
全素性を利用	0.750	0.404	0.525	0.829	0.403	0.542

R, P, F はそれぞれ再現率, 精度, F 値を表す。

えられる。そこで、視線パターンの素性に加え、表4に示す品詞や語彙的な情報などの言語的な素性も導入する。

## 5. 評価実験

4節で導入した視線パターンの有効性を調査するために、アノテーション欠損検出の評価実験を行った。表2に示した事例のうち、A<sub>0</sub> と A<sub>2</sub> の両方の作業者がアノテーションを行った事例を除くデータで 10 分割交差検定を用いて評価を行う。評価の際は一方の作業者 (例えば、A<sub>0</sub>) がアノテーションしていない述語のガ格に対し、もう一方の作業者 (A<sub>2</sub>) がアノテーションしたガ格を正解とみなして評価 (正例 281 事例と負例 561 事例を弁別する問題の評価) を行う。また、逆の場合 (A<sub>0</sub> を正解とみなし、A<sub>2</sub> から問題を抽出する) についても同様に評価を行う。

学習・分類には線形カーネルを使った SVM[21] を用いた。SVM の実装としては svm.light<sup>\*4</sup> を用い、人手で-j と-c オプションを変更しながら、F 値最大になる点を求めた。また、prefixspan を用い頻出パターンを求める際は、計算効率を考え、パターンの最大長を 5, 最小長を 3 として訓練事例の正例と負例それぞれから頻出パターンを求めた。その結果得られた視線パターンのうち、正例と負例それぞれ上位 50 の頻出パターンを視線に関する素性として利用した。

### 5.1 ベースラインモデル

ベースラインとして全ての事例を正例と判断するモデルを採用した。このモデルの出力結果は、アノテーション作業者がガ格が付与されていない事例をすべて見直し、アノテーションの漏れを検出するという典型的な戦略に相当する。

### 5.2 実験結果

アノテーション欠損検出に関する 2 値分類の結果を表 5

\*4 <http://svmlight.joachims.org/>

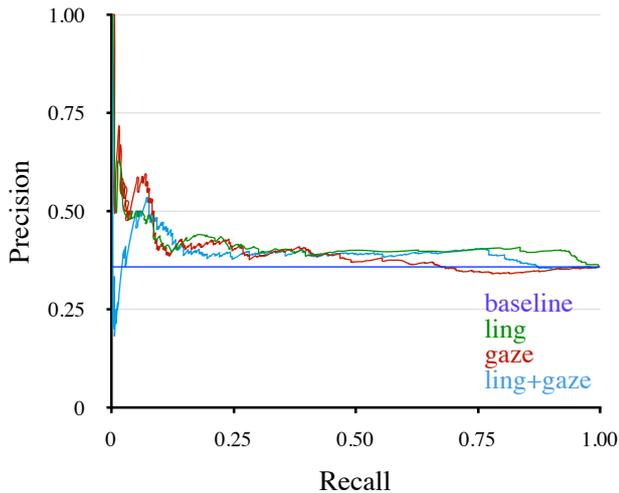


図6 再現率-精度曲線 (正解:A<sub>0</sub>, システム:A<sub>2</sub>)

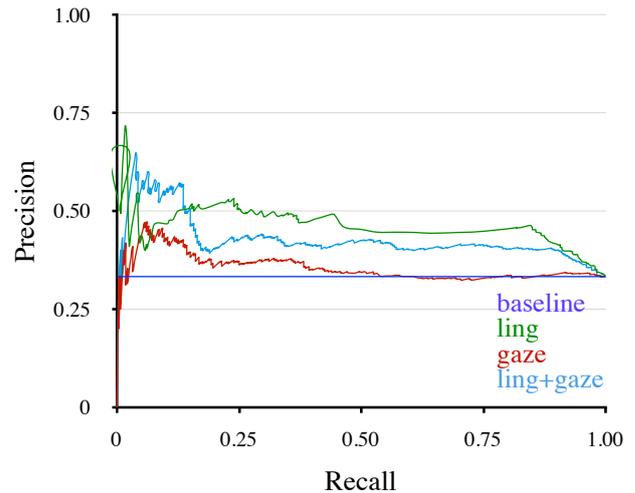


図7 再現率-精度曲線 (正解:A<sub>2</sub>, システム:A<sub>0</sub>)

に示す。表5の左側がA<sub>0</sub>を正解とし、A<sub>2</sub>のアノテーション結果を問題とした場合の結果であり、右側はその逆の場合の結果である。表5に示された各モデルのF値を比較すると、どの機械学習に基づくモデルもベースラインモデルと同等かそれ以上の性能でアノテーション欠損を検出できていることがわかる。この結果から、言語的な素性と視線情報に基づく素性の両方が分類に有効であることがわかる。しかし、両方の素性を組み合わせた場合よりも言語的な素性のみを利用したモデルが最も良い結果を得ており、視線情報が欠損検出に役立っているとは言いがたい。ただし、一般的な文章を想定した場合、必ずしも学習データに出現した述語が出現するとは限らないため、言語的な特徴に依存した欠損検出だけでは良い性能が得られない可能性がある。本研究で採用した視線パタンの利用方法は視線情報を扱うための一例でしかないが、今後はアノテーション欠損検出における言語的な特徴の欠落を補うために、視線情報を有効に利用する方法を再検討する必要があると考えられる。

本研究で出力する結果は、アノテーションの欠損を効率的に検出し、その結果の修正に利用することが考えられる。そのため、2値で厳密に分類するよりも事例を信頼度に基づいてランキングし、その結果を順位の高い事例から順に修正するという作業のほうが望ましい。そこで、この観点から評価するために、再現率-精度曲線を描き、出力結果を再評価する。表5に示した各モデルの再現率-精度曲線を図6と図7に示す。図6では各モデルの再現率-精度曲線は競合しているが、図7では言語的な素性だけを使ったモデルと比較して、言語的な素性と視線の素性を利用したモデルが再現率が低い領域において最も精度が高い結果を得ていることがわかる。そこで、言語的な素性と視線の素性を利用したモデルに関して、低い再現率で頻出する視線のパターンを調査した。再現率が0から0.15までの領域に出

表6 上位20の頻出視線パターン (正解:A<sub>2</sub>, システム:A<sub>0</sub>)

頻度	重み	視線パターン
35	0.2349	T T T
34	0.0258	T LA LA
30	-0.0510	LA LA T
25	0.1220	-LP -LP -LP
25	0.0554	+RP +RP +RP
24	0.0265	-LA -LA T
22	0.1390	-LA -LA -LA -LA
21	-0.1239	LA T T
20	0.0164	T T T T
20	0.1381	+RA +RA +RA
18	0.0180	+RA +RP +RP
17	0.0267	-LA -LP -LP
16	0.1023	-LA -LA -LA -LA -LA
14	0.1242	LA LA LA T
14	0.0045	-LP -LP -LA
13	0.1891	+RA +RP +RP +RP
12	0.1566	RA RP RP
11	0.1543	LA LA T T
10	0.0387	T LA LA LA
10	-0.0629	-LA -LA -LA T

現した事例を抽出し、その区間で頻出する上位20の視線パターンとその素性の重みを表6に示す。表6に示されたパターンからアノテーション時のアノテーション作業の典型的な振舞いがわかる。例えば、正例の注視の系列で頻出している視線パターン“T T T”は連続して対象となる述語に注視が続く視線の動きであり、このパターンが出現しているときには、例えば、アノテーション対象を見続けながらアノテーションすべきかを吟味していると考えられる。一方、頻出の視線パターン“T LA LA”と“LA LA T T”は項を探すために文頭の方へ動く視線の動きに対応しており、項の候補を探すという振舞いがアノテーションすべきか否かを判断する上で重要な要因になっていることがわかる。

上記の分析のように、視線パタンの一部は特定のアンノテーション欠損を検出するために有効であることがわかる。本研究では視線のパタンを得るための `prefixspan` のパラメータや記号の定義は直観に基づいて経験的に定義したが、この定義については改良の余地があるため、今後さらに調査を進めたい。

## 6. 関連研究

近年の視線追跡に関する技術が発展したことによって、視線データはさまざまな研究領域で利用されている [6]。例えば、Bednarik ら [2] はプログラマがデバッグを行う際の視線データを収集している。彼らはプログラムの統合開発環境 (IDE) の領域に基づき、プログラマの3つの関心領域 (ソースコードの領域、クラス関係が可視化された領域、プログラムの出力の領域) を定義し、初心者のプログラマと玄人のプログラマが関心領域に関してどのように視線を遷移させるかを比較している。ただし、彼らが定義した関心領域は画面を3分割するだけの荒い分割であるため、プログラマの技能の推定には利用可能かもしれないが、得られた玄人の遷移パタンに基づき、ある特定の遷移パタンとプログラミングの技能があることの関連性を説明することは難しいと考えられる。同様に、言語に関する視線の動きから分析を行う際も、より細かい関心領域に関する調査が必要となるため、本研究では文字レベルのより細かい関心領域を採用し、それに従ってどのセグメントを注視したかの調査を行った。

Rosengrant [16] は、被験者の視線データとプロトコル分析 [7] で導入されている被験者の口述説明を統合する `gaze scribing` と呼ばれる新しい分析手法を提案している。この研究では、素人と玄人の問題解決時の戦略の違いを発見するために、事例研究としてディスプレイ上に表示された物理の電気回路の問題を解く際の被験者の振舞いを分析している。彼は他の問題へ `gaze scribing` を適用し分析することの重要性を強調しているが、プロトコル分析を単純に用いることは被験者の認知負荷が高まり、想定していた作業への影響が懸念されるため、もともと達成すべき目標を阻害しないよう課題を設計することが重要となる。

Tomanek ら [20] は能動学習の効率的な事例選択のために、固有名のアンノテーションの難易度を推定するために視線データを利用している。彼らは固有名の特徴を制御することでさまざまな設定におけるアンノテーション時の視線データを収集している。彼らの固有名に関する視線データの収集では、アンノテーション対象となる表現を見ているか、もしくはその周りの文脈を見ているかという荒い関心領域を扱っているが、本研究で扱う述語項構造アンノテーションの場合、正解となる項に対して競合する項の候補が出現しているか否かという分析が必要になるため、より細かい関心領域を扱う必要性があるという違いが存在する。

我々の先行研究では2節で紹介したデータを利用して述語項構造アンノテーションの難易度推定について調査を行った [19], [23]。人手による分析結果に基づき、経験的に各インスタンスに関する作業時間を定義し、その作業時間が短いほど複数人が作業した結果の一致率が高くなることを示した。

## 7. おわりに

本稿では、複数人のアンノテーション作業者のアンノテーション結果の不一致の検出の問題、特に、二人の作業者のうち一方だけがアンノテーションするアンノテーション欠損をどのように検出するかという問題について議論した。これを実現するために、機械学習ベースの手法において、言語的な特徴に加え、アンノテーション作業者の視線情報を用いる方法の一例を示した。視線の情報から問題に適用するための有効な情報を抽出するために、テキストマイニングの技術を用いて注視の系列から頻出する視線パタンを抽出し、それを素性として利用する手法を提案した。評価実験では、我々がこれまでに収集した日本語述語項構造アンノテーション時の視線情報を用いて実験を行い、視線から得られる素性と言語的な素性のそれぞれがアンノテーション欠損検出に貢献することを示した。また、追加で行った調査の結果から、特定の視線パタンがアンノテーション欠損検出のための良い指標となることも明らかにした。

本研究では、注視の系列を表現するために表3や図5に示した注視の場所や時間的な側面に基づくヒューリスティックな表現形式を採用した。しかし、その表現形式がこの問題の特徴を捉えるための最良のものであるという保証はない。そこで、今後はアンノテーション欠損検出の問題を適切に捉える表現形式についてさらに吟味する必要がある。また、本研究では2名の作業者の欠損を捉える問題を扱ったが、この2名の作業結果が必ずしも正しいアンノテーションの結果とは限らない。そこで、今回分析対象とした43記事に対してあらためて正解となるアンノテーション結果を作成し、その正解と各アンノテーション作業者の作業結果を比較することで、正解と異なる箇所でのどのような視線の動きが起こっているのかを分析したい。

## 参考文献

- [1] Artstein, R. and Poesio, M.: Inter-Coder Agreement for Computational Linguistics, *Computational Linguistics*, Vol. 34, No. 4, pp. 555–596 (2008).
- [2] Bednarik, R. and Tukiainen, M.: Temporal eye-tracking data: Evolution of debugging strategies with multiple representations, *Proceedings of the 2008 symposium on Eye tracking research & applications (ETRA '08)*, pp. 99–102 (2008).
- [3] Carletta, J.: Assessing Agreement on Classification Tasks: The Kappa Statistic, *Computational Linguistics*, Vol. 22, No. 2, pp. 249–254 (1996).
- [4] Chen, Z.: From data mining to behavior mining, *Internation-*

- tional Journal of Information Technology & Decision Making*, Vol. 5, No. 4, pp. 703–711 (2006).
- [5] Chou, W.-C., Tsai, R. T.-H., Su, Y.-S., Ku, W., Sung, T.-Y. and Hsu, W.-L.: A semi-automatic method for annotating a biomedical proposition bank, *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora*, pp. 5–12 (2006).
- [6] Duchowski, A. T.: A breadth-first survey of eye-tracking applications, *Behavior Research Methods, Instruments, and Computers*, Vol. 34, No. 4, pp. 455–470 (2002).
- [7] Ericsson, K. and Simon, H. A.: *Protocol Analysis – Verbal Reports as Data* –, The MIT Press (1984).
- [8] Fort, K., François, C., Galibert, O. and Ghribi, M.: Analyzing the Impact of Prevalence on the Evaluation of a Manual Annotation Campaign, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 1474–1480 (2012).
- [9] Kaplan, D., Iida, R., Nishina, K. and Tokunaga, T.: Slate – A tool for creating and maintaining annotated corpora, *Journal for Language Technology and Computational Linguistics*, Vol. 26, No. 2, pp. 89–101 (2012).
- [10] Lenzi, V. B., Moretti, G. and Sprugnoli, R.: CAT: the CELCT Annotation Tool, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 333–338 (2012).
- [11] Marcińczuk, M., Kocoń, J. and Broda, B.: Inforex – a web-based tool for text corpus management and semantic annotation, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 224–230 (2012).
- [12] Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A.: Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, Vol. 19, No. 2, pp. 313–330 (1993).
- [13] Passonneau, R.: Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 831–836 (2006).
- [14] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U. and Hsu, M.-C.: PrefixSpan: Mining Sequential Patterns Efficiently by PrefixProjected Pattern Growth, *Proceedings 2001 International Conference Data Engineering (ICDE'01)*, pp. 215–224 (2001).
- [15] Rehbein, I., Ruppenhofer, J. and Sporleder, C.: Is it worth the effort? Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation, *Language Resources and Evaluation*, Vol. 46, No. 1, pp. 1–23 (2012).
- [16] Rosengrant, D.: Gaze scribing in physics problem solving, *Proceedings of the 2010 symposium on Eye tracking research & applications (ETRA '10)*, pp. 45–48 (2010).
- [17] Salvucci, D. D. and Goldberg, J. H.: Identifying fixations and saccades in eye-tracking protocols, *Proceedings of the 2000 symposium on Eye tracking research & applications (ETRA '00)*, pp. 71–78 (2000).
- [18] Stede, M. and Huang, C.-R.: Inter-operability and reusability: the science of annotation, *Language Resources and Evaluation*, Vol. 46, No. 1, pp. 91–94 (2012).
- [19] Tokunaga, T., Iida, R. and Mitsuda, K.: Annotation for annotation - Toward eliciting implicit linguistic knowledge through annotation -, *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9)*, pp. 79–83 (2013).
- [20] Tomanek, K., Hahn, U., Lohmann, S. and Ziegler, J.: A Cognitive Cost Model of Annotations Based on Eye-Tracking Data, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 1158–1167 (2010).
- [21] Vapnik, V. N.: *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing Communications, and control, John Wiley & Sons (1998).
- [22] Voutilainen, A.: Improving corpus annotation productivity: a method and experiment with interactive tagging, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 2097–2102 (2012).
- [23] 光田 航, 飯田 龍, 徳永健伸: 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, 言語処理学会第19回年次大会, pp. 449–452 (2013).