視線と操作情報を利用した誤りアノテーションの検出

光田 航　　飯田 龍　　徳永 健伸
東京工業大学 大学院情報理工学研究科
{mitsudak,ryu-i,take}@cl.cs.titech.ac.jp

1 はじめに

近年の自然言語処理分野では，ある与えられた問題に対し，正解となる情報をコーパスにアノテーションしたものを正解データとして学習する手法が主流となっている．この手法では，コーパスの品質が直接的に解析モデルに影響を与えるため，高品質なアノテーションを効率良く行うことは自然言語処理のさまざまな問題において重要な課題となる．このため，アノテーションされた結果の品質を推定する問題もコーパス作成の過程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作業を行った結果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, 7]．ただし，これらの評価尺度では，アノテーションの品質を推定する際に，複数人の作業者が同じデータに対してアノテーションを行うことが前提とされているため，一致率を見積るために冗長な作業が必要となる．また，アノテーション作業者が 3 名以上の場合はその組み合わせで一致率が異なるため，必ずしも作業結果全体に対する評価尺度としては適していない．これに対し，本研究では，アノテーション作業者の作業中の振舞い（視線・操作履歴）などを参照することで，単一の作業者の結果に基づいて作業の品質の推定を行う．

作業者の振舞いに基づく品質推定を行うために，我々のこれまでの研究では，日本語の述語項構造アノテーションを例にアノテーション作業者の作業中の視線情報と操作履歴を収集し [8]，その結果を利用して，各事例に対するアノテーション時間と作業結果の揺れの相関を示した [10]．また，二人の作業者がある述語の格に対してアノテーションした結果のうち，一方がアノテーションをしなかった事例を検出する欠損アノテーション検出タスクを設定し，その自動検出で言語情報と視線情報がともに検出に役立つことを示した [6]．

本研究は，これまでの研究の自然な拡張として，ある作業者が述語の格に対してアノテーションを行った際の誤りを自動検出する問題を扱う．先行研究 [6] では欠損アノテーション，つまり，アノテーション作業者がアノテーションを行わなかった場合に，対象となる述語の格をアノテーションすべきか否かという問題を述語のみの情報を利用して解こうとしていたのに対し，本研究で対象とするアノテーション誤りの検出では作業者がアノテーションした結果が利用できるため，述語とアノテーションされたその述語の項の情報が利用できるという違いがある．このため，先行研究で使用した素性と比較して，より多様な言語情報を利用できるため，特に項の言語情報が検出性能にどう影響するかを調査可能となる．また，視線などの非言語情報に関しても，明確にアノテーション時間を規定し，その期間中の視線情報を利用できるという利点もある．この調査のため，本研究では特に言語情報と視線情報の利用方法に関する考察も行う．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操作と視線情報の収集実験と収集結果について説明する．次に 3 節でアノテーション誤り検出モデルを提案し，4 節でモデルの評価実験の結果を報告する．最後に，5 節でまとめと今後の展望を述べる．

2 視線情報と操作情報の収集

2.1 データと収集の手続き

我々の先行研究 [8] では日本語の述語項構造アノテーションの問題を対象に，アノテーション作業者のアノテーションツールの操作履歴とアノテーション時の視線情報を収集した．このデータ収集時には，アノテーションツール Slate [4][*1] を修正したツールを利用し，文章中の述語に対し，ガ格，ヲ格，ニ格のアノテーションを行った．述語やその項となる名詞句はあらかじめアノテーションされており，アノテーション作業者はマウスのドラッグで述語とその項の関係をアノテーションした．この際，作業者がどのような過程で作業を行ったかに関するツールの操作履歴に加え，視線計測装置 Tobii T60 を利用し，作業者の視線情報を記録した（図 1）．

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 5 人のアノテーション作業者を雇用し，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）[5] の書籍コーパス（PB）中の 43 記事を対象に記事単位でアノテーション作業を行った．アノテーション仕様は NAIST テキストコーパス [11] を構築する際に利用されたアノテーション仕様を単純化したものを利用した．

2.2 アノテーションの作業結果

我々のこれまでの研究 [6, 8, 10] では，作業者間の一致を基準として揺れの自動検出の問題を設定していたが，作業者間のアノテーションが一致してもそれが正しいとは限らないので，あらかじめ正解データを作成し，その正解

[*1] https://bitbucket.org/dainkaplan/slate/

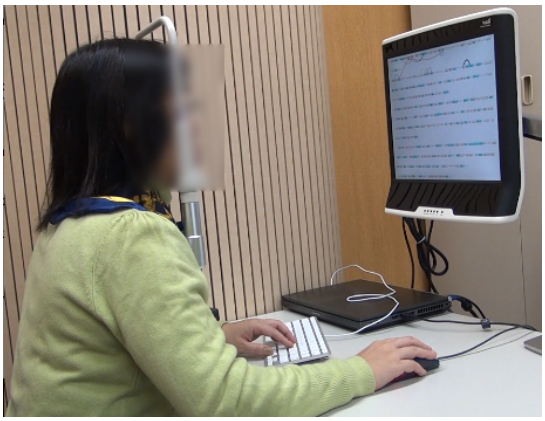

図 1: 視線情報の収集実験のスナップショット

データと各作業者の作業結果の一致しない箇所を誤りとして自動検出の対象とする．

正解データを作成するために，分析対象となる 43 記事に対し，まず第 1 著者が一通りアノテーションを行い，その結果を第 2 著者と第 3 著者が確認し，判断が揺れた箇所は協議を行い，正解を一意に決定した．この結果，アノテーションされた述語の数は 2,460 事例，また，その述語の格にアノテーションされた個数は，ガ格が 2,313 事例，ヲ格が 1,063 事例，ニ格が 614 事例となった．

次に，5 人のアノテーション作業者 A_0〜A_4 がアノテーションした結果と正解データとの差分を，作業者のアノテーション結果をシステムの出力とみなして，再現率，精度，F 値で評価した結果を表 1 に示す[*2]．表より，A_1 を

表 1: 正解データと作業者が付与した結果との違い

作業者	再現率	精度	F 値
A_0	0.72 (2,855/3,990)	0.86 (2,855/3,302)	0.78
A_1	0.65 (2,611/3,990)	0.74 (2,611/3,521)	0.70
A_2	0.67 (2,680/3,990)	0.82 (2,680/3,255)	0.74
A_3	0.76 (3,048/3,990)	0.79 (3,048/3,858)	0.78
A_4	0.74 (2,964/3,990)	0.77 (2,964/3,847)	0.76
平均	0.71	0.80	0.75

除いて，各作業者と正解データとの差分は F 値では大きく違わないことがわかるが，一方，5 名中の任意の 2 名の作業者のアノテーション誤りの重複率の平均を調べたところ，その重複率は 0.49 であることがわかった．つまり，作業者間で約半分の事例しかアノテーションを誤る箇所が重なっていないため，言語的な特徴のみを用いてその誤りを検出することは難しいことがわかる．そこで，作業者ごとに異なる作業者の視線情報も利用して，アノテーションの誤りを検出する．

3 誤りアノテーションの検出モデル

アノテーション作業者の視線の動きは，例えば，作業者がアノテーション対象の述語とその項の近傍を集中して注視したり，また，文章を読み進める際にある述語を読み飛ばすなど，作業者の関心を反映した動きをすると考えられる．また，操作履歴に関しても，個々の事例のアノテーションにかけた時間や，どの事例を先にアノテーションするかといったアノテーション作業の順序を表わしており，視線同様に作業者の特徴を反映すると考えられる．

[*2] 作業者はアノテーション時に述語のある格の項を文章中から 1 つアノテーションするが，正解データには共参照関係も付与されているため，正解となる項が共参照関係のために文章中に複数ある場合はそのいずれかをアノテーションしている場合に正解とみなした．

表 2: アノテーション時間に基づく誤り検出結果

作業者	P	R	F
A_0	0.18	0.51	0.27
A_1	0.26	0.51	0.35
A_2	0.18	0.56	0.27
A_3	0.20	0.59	0.30
A_4	0.24	0.52	0.33

表 3: 視線の特徴を表す記号の一覧

カテゴリ	記号
位置	対象の述語と同一文内（述語より前：W，後：E）， それ以外（述語が出現する文より前：N，後：S）
種類	対象の述語（P），それ以外の述語（Q）， 対象の述語に付与された項（A），それ以外の項（B）
時刻	アノテーション時間の前（−），後（+）

例えば，我々の先行研究 [10] では，視線情報と操作履歴の情報に基づいて各事例に対するアノテーション時間を定義したが，このアノテーション時間は作業者間の作業の揺れと相関があることがわかっている．このため，この作業時間に基づき，ある作業時間 t 以上かかるアノテーションは間違いであるとみなし，それ以外は正しくアノテーションされたと仮定するモデルを考えることができる．そこで，まず，このパラメタ t を人手で動かし，検出結果に関する F 値が最大になる点を作業者ごとに調査した．この結果を表 2 に示す．この結果より，アノテーション時間に基づき誤りを検出することを考えた場合，上限値として F 値で約 3 割程度の検出性能が見込まれる．

また，一方で，先行研究 [6] のようにより詳細な視線情報の利用も考えることができる．先行研究では，アノテーション時間中の注視の系列を抽出し，そこに頻出する注視のパタンをテキストマイニングの技術を利用することで抽出し，それを素性として利用することで視線情報を検出の問題に導入している．ただし，マイニングを行った結果を利用した場合，頻出する系列は利用可能であるが，それ以外の注視の系列の情報は捨象されることになる．そこで，本研究では抽出した注視の系列の情報を直接的に利用することを考える．

提案手法では，まず，下記の手順にしたがって，先行研究と同様に注視の系列を抽出する．

(1) 分類対象の述語と格に対して作業を行った時間（アノテーション時間）にマージンを加えた期間を抽出する．
(2) 抽出した期間に含まれる注視の系列を，表 3 にしたがって注視の特徴を表す記号の系列に変換する[*3]．

次に，記号の系列からユニグラムとバイグラムの頻度を求め，その結果を素性として利用する．

例えば，“−WA −WQ −WB −WQ −WA P WA P +SB +SB +SQ +SB +SQ +SB” という記号の系列が得られた場合は，‘−WA’ というユニグラム素性や ‘−WA P’

[*3] 変換の詳細は文献 [6] を参照されたい．

表 4: 誤りアノテーション検出のための素性

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 かどうか
	is_adj	対象述語が形容詞の場合は 1，それ以外は 0
	lemma	対象述語と付与された項の見出しの基本形
gaze	uni_gaze	アノテーション時間中の注視系列のユニグラム
	bi_gaze	アノテーション時間中の注視系列のバイグラム

表 5: ling と gaze が出力した上位 N 件の事例の重複率

上位 N 件	10	50	100	500	1000	all
重複率	0.02	0.02	0.03	0.14	0.28	1.00

のようなバイグラム素性をこの系列から抽出して利用する．このような局所的な注視の種類に関する情報を利用することで，少なくともある注視が出現したという情報やどのように注視が移り変わったかという情報が素性としてエンコードされる．さらに，これらの視線情報に関する素性に加え，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して，アノテーション誤り検出のモデルを学習する．

4 評価実験

3 節で導入した視線情報の有効性を調査するために，アノテーション誤り検出の評価実験を行った．学習・分類には線形カーネルを使った SVM [9][*4]を用いた．評価の際は，各作業者が作成した事例集合内で 10 分割交差検定を行い，この際，8/10 を訓練用データに，1/10 をデベロップメントデータに，残りの 1/10 を評価用データとして利用する．デベロップメントデータは F 値が最大となる SVM のパラメタ c を決定するために利用した．

4.1 比較するモデル

表 4 に示した素性の有効性を調査するために，言語情報（表 4 の ling 素性）のみを利用した ling モデル，視線情報（表 4 の gaze 素性）のみを利用した gaze モデル，両方の情報を利用した ling+gaze モデルを比較する．

また，SVM の分離平面からの距離を出力結果のスコアとみなし，このスコアに基づいて ling モデルと gaze モデルの出力した結果を調査した結果を表 5 に示す．この結果からわかるように，ling モデルと gaze モデルのスコア上位の結果はほとんど重複していないことがわかる．これは，言語的な特徴に基づきアノテーション誤りを検出する場合と視線情報に基づいて誤りを検出する際の特徴の捉え方が根本的に違うことを表している．例えば，頻繁にアノテーションを誤る述語がもし存在する場合にはその特徴は言語的な素性で捉えられるのに対し，訓練事例に出現しない述語に関しては注視の系列が似ていることでそのアノテーション誤りを検出できる可能性がある．そこで，ling モデルの出力結果のうち，スコアが閾値 r_1 を越える事例をアノテーション誤りとし，また独立に gaze モデルの出力結果のうち，スコアが閾値 r_2 を越える事例をアノテーション誤りとして出力するモデルを ling_or_gaze モデルとし，比較対象に加える．この際，閾値 r_1 と r_2 はデベロップメントデータを利用して F 値が最大となる値を求めた．

[*4] http://svmlight.joachims.org/

表 7: ling モデルにおける素性重み

作業者	ガ格	ヲ格	ニ格	動詞	形容詞
A_0	0.06	-0.73	0.67	0.06	0.47
A_1	0.36	-1.26	0.90	-0.33	0.41
A_2	-0.10	-1.00	1.10	1.03	1.41
A_3	-0.35	-0.66	1.01	0.17	1.11
A_4	0.06	-0.33	0.27	0.24	0.53
平均	0.01	-0.80	0.79	0.23	0.79

4.2 実験結果

4.1 に示した 4 つのモデルを評価した結果を表 6 に示す．表 6 に示された結果のうち，まず ling モデルと gaze モデルの結果を比較した場合，ling モデルの精度がすべての作業者に関して良いことがわかる．これは，注視の系列のような漠然とした情報よりも，問題の難易度を特定するために言語的な情報が役立っているためだと考えられる．ただし，ling+gaze モデルのように両方の素性を利用することで，作業者によっては検出精度が向上することがわかる．例えば，ling と ling+gaze の結果を比較した場合，A_0 と A_3 で F 値が向上している．さらに，ling_or_gaze モデルのように言語と視線の特徴を分けて利用することで，比較したモデルの中で F 値が最大となることがわかった．この理由としては，利用している事例の数が少ないため，個別のモデルでそれぞれの特徴を学習し，特にアノテーション誤りだと思われる部分を選別して出力することで結果的に良い結果を得たのだと考えられる．このモデルごとの振舞いの違いを調べるため，以降で ling モデルと gaze モデルが捉えた特徴を人手分析した結果を報告する．

4.3 ling モデルと gaze モデルの人手分析

各モデルの特徴を調べる際，そのモデルの出力結果を調査することも重要であるが，モデルの学習した素性の重みを調べることで，そのモデルの特徴を端的に比較することができる．そこで，各作業者ごとに構築した ling モデルの素性の重みを調べた結果を表 7 にまとめる[*5]．この表より，ヲ格とニ格に関する学習結果の傾向はその値の極性（正負）が統一されているという意味で類似しているが，一方，ガ格は作業者ごとに極性が統一されていないことがわかる．これは，作業者によってガ格のアノテーション誤りの割合が異なったためだと考えられる．また，ニ格の値が正の重みを持っている点については，ニ格が他の格よりもアノテーションすべきか否かの判断が難しいため，特に誤りやすいことを表している．また，lemma 素性についてはどの見出しが上位にくるかは作業者ごとで異なるが，その素性の値は表 7 に示した結果よりも高い値を持つよう学習されていた．これは，作業者ごとにある特定の語が顕著に誤ることを学習したためだと考え

[*5] ただし，lemma 素性とその値は見出しごとに異なるため割愛する．

表 6: アノテーション誤り自動検出の実験結果

作業者	ling			gaze			ling+gaze			ling_or_gaze		
	P	R	F	P	R	F	P	R	F	P	R	F
A_0	0.15	0.30	0.18	0.16	0.16	0.12	0.29	0.26	0.27	0.21	0.52	**0.28**
A_1	0.38	0.39	0.38	0.37	0.30	0.27	0.32	0.33	0.32	0.33	0.69	**0.43**
A_2	0.23	0.30	0.25	0.11	0.19	0.13	0.25	0.26	0.25	0.23	0.53	**0.30**
A_3	0.25	0.34	0.27	0.24	0.21	0.17	0.30	0.34	0.31	0.28	0.66	**0.36**
A_4	0.33	0.35	0.33	0.25	0.21	0.18	0.33	0.33	0.32	0.29	0.62	**0.38**
平均	0.34	0.27	0.28	0.22	0.23	0.17	0.30	0.30	0.29	0.28	0.61	**0.35**

(i) は作業者の注視の順序を表す.

図 2: A_0 に特徴的な視線の動きの例

られる.

次に，gaze モデルの特徴を調べるために，ling モデルでは検出を誤ったが，gaze モデルでは検出できた 785 事例[*6]を人手で調査した．785 事例のうち，訓練事例中に同一の述語の見出しが出現した割合は約 6 割であり，残りの 4 割は学習データに同一の述語が出現していない．つまり，その 4 割については視線の素性が有効に機能して分類が成功したのだと考えられる.

これらの事例をより詳細に分析するために，注視の動きと解析対象となる述語やその項，近傍文脈を調べることで，どのような注視の系列が誤り検出に貢献しているのかを調査した．この結果，‘EA EB’，‘EB EA’，‘WA WB’，‘WB WA’ のように，アノテーション時間中にアノテーションした項とそれ以外の対立候補の間で注視が動く場合，つまり，典型的には作業者が付与する項を迷っている場合に，アノテーション誤りが生じていることがわかった．例えば，図 2 の例では，作業者は動詞「かしずく」の項を探すが，正解である「親類」と対立候補である「ものたち」の間で注視が遷移し，その後作業者は誤った項である「ものたち」にアノテーションをしてしまっている．このように特定の注視の系列は常にアノテーション誤りを検出する強い要因となるわけではないが，作業者の振舞いの典型的な動作を捉えることがあり，その結果，言語的な素性だけでは捉えることができない現象を検出するのに貢献している.

5 おわりに

本稿では，日本語の述語項構造アノテーションの問題を対象に，作業者がアノテーション作業を行った場合の誤りを自動検出するタスクに取り組んだ．自動検出のために，言語的な情報と作業者の視線に基づく情報を併用することで，言語情報のみを利用した場合と比較してより頑健に誤りを検出できることを示した.

先行研究 [6] や本研究で使用した注視の表現方法はヒューリスティックに決定したものであり，改善の余地があるため，今後は誤り検出のためにより適切な作業者の視線の表現方法を考える必要がある．また，言語情報と視線情報がそれぞれどのような状況で誤り検出に役立つのかの見極めも今後重要な課題となる．また，本研究では一つの文章全体に対してアノテーションを行った結果を利用して分析を行っているため，1 事例に対するアノテーション時間もヒューリスティックに決定したものに依存してしまっている．この問題を回避するため，アノテーション作業者に 1 事例ごとにアノテーション作業を行わせ，その際の振舞いを記録する収集実験を行っており，今後この新規に作成したデータを用いてより精密な分析を行うことでアノテーション作業者の振舞いの特徴をより明らかにできると考えられる.

[*6] 785 事例の作業者 A_0〜A_4 に関する内訳はそれぞれ 58，305，37，197，188 であり，作業者ごとに視線情報の有効性が異なることがわかる.

参考文献

[1] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, Vol. 34, No. 4, pp. 555–596, 2008.

[2] Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, Vol. 22, No. 2, pp. 249–254, 1996.

[3] Karën Fort, Claire François, Olivier Galibert, and Maha Ghribi. Analyzing the impact of prevalence on the evaluation of a manual annotation campaign. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 1474–1480, 2012.

[4] Dain Kaplan, Ryu Iida, Kikuko Nishina, and Takenobu Tokunaga. Slate – a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, Vol. 26, No. 2, pp. 89–101, 2012.

[5] 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 Eigth International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 1483–1486, 2010.

[6] Koh Mitsuda, Ryu Iida, and Takenobu Tokunaga. Detecting missing annotation disagreement using eye gaze information. In *Proceedings of the 11th Workshop on Asian Language Resources*, pp. 19–26, 2013.

[7] Rebecca Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 831–836, 2006.

[8] Takenobu Tokunaga, Ryu Iida, and Koh Mitsuda. Annotation for annotation – toward eliciting implicit linguistic knowledge through annotation –. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9)*, pp. 79–83, 2013.

[9] V. N. Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing Communications, and control. John Wiley & Sons, 1998.

[10] 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 回年次大会発表論文集, pp. 449–452, 2013.

[11] 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.