アノテーションとアノテーション作業者の信頼性推定

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
‡ 情報通信研究機構 ユニバーサルコミュニケーション研究所
†{mitsudak,take}@cl.cs.titech.ac.jp, ‡ryu.iida@nict.go.jp

1 はじめに

近年，自然言語処理の様々な分野で，正解となる情報（タグ）をコーパスに人手でアノテーションし，その結果を利用することで，対象としている問題の分析や，自動解析のモデルを構築するという試みがなされている．特に，機械学習を利用したモデルの構築では，アノテーションされた結果に基づいて学習が行われるため，コーパス中のタグがアノテーションの仕様に基づいて正しくアノテーションされていることが重要となる．しかし，人手でアノテーションされた結果にはアノテーション仕様に対する誤解や作業漏れ，仕様に記述されない例外的な事例に対する作業者の誤った判断に基づくアノテーションなど，さまざまな理由によりノイズが混入することになる．このため，完全に正しくアノテーションされた場合と比較してノイズが含まれたアノテーション結果を学習の対象とした場合，必ずしも期待されるような学習結果が得られるとは限らない．この問題を解決するために，アノテーションされたタグの信頼性を推定することで，その推定の結果に基づいた事例の取捨選択や，既に構築されたタグ付きコーパス中に含まれるアノテーションの誤りを効率的に検出し修正を行う手法が必要となるが，これらについては形態素解析や文書分類などに関する誤りの検出や修正 [1,3] については研究が進められているものの，他の研究分野では Kappa 係数 [2] のようなアノテーション作業者の一致率に基づいたアノテーションの品質評価が行われているだけである．一致率に基づくアノテーションの品質評価は，文献 [5] でも指摘されているように，個々のアノテーション作業者の特性や，個別の事例の特徴や，その事例に対するアノテーション作業者のアノテーション行為を捨象した評価であるため，この一致率の数値だけではコーパスに人手でアノテーションされた結果の傾向を見積ることが難しいことがわかる．

このような背景から，我々のこれまでの一連の研究 [9,10,12] では，述語と項の関係をアノテーションする問題を例題に，アノテーション作業者の振舞い，特に，アノテーション時の視線の動きに着目したデータの収集とその分析を行い，さらに，アノテーション漏れやアノテーション誤りの自動検出に関する問題に取り組んできた．ただし，これまで行ったや誤り自動検出では，1 回のアノテーション作業で文章中の複数の述語に対してアノテーションした結果を利用したため，各述語に対する視線情報の特定や，その視線に基づいた素性の設計が困難になるという問題が生じた．

そこで，本研究では，アノテーション作業者の視線を計測しながら，1 回の作業で 1 つの述語の 1 つの格要素のみを選択するアノテーションの結果を収集したデータ [11] を対象に，アノテーション作業者が行った各事例へのアノテーションの信頼度を，(1) 与えられた問題の潜在的な難しさ，(2) アノテーション時の視線の動きや作業時間，(3) アノテーション結果からわかる情報の 3 つに基づいて推定する手法を提案し，人手でアノテーション誤りを修正する際の指標となるアノテーションの信頼度の推定手法を提案する．さらに，各事例へのアノテーションの信頼度に基づき，その作業を行ったアノテーション作業者の信頼性を推定する手法も提案し，その手法で推定する信頼度とアノテーション対象となる問題集合の正解率の関係を調べることで，作業者の信頼度推定手法の有効性を調査する．本稿では，まず 2 節でアノテーション課題と作業者の振舞いに関するデータ収集について述べ，次に 3 節でアノテーションの誤りを修正する手順を説明し，そこで利用する事例に対する信頼性推定モデルを提案する． 4 節で誤り検出の評価実験を行い信頼性推定に用いた特徴の種類の有効性について示し，さらに 5 節で作成した検出モデルを利用した作業者の信頼性推定の結果について報告する．最後に 6 節でまとめと今後の課題を述べる．

2 単一述語項関係アノテーションのデータ収集

先行研究 [11] で行ったデータ収集では，図 1 に示すアノテーションの作業画面において，青枠で示された対象述語に対し，ガ格となる項を灰色の背景で示された項の候補集合から選択する作業を行う．ただし，図 1 からわかるように，正解となる項は必ずしも 1 つとは限らず，アノテーション作業者は正解となる項の集合のうち，いずれかを選択すれば正解となる．作業にはマウスを用い，正解となる項をクリックすることで作業が完了する．作業対象は BCCWJ [8] のコアデータの書籍レジスタ（PB）から抽出した 221 事例を用いる．このうち，述語と同一文内に項が出現する事例が 122 事例，述語とは異なる文に項が出現する事例が 99 事例である．作業時にはマウスの動作に加え，視線計測装置 Tobii T60 を利用して視線情報を記録した．データ収集のために 20 名のアノテーション作業者を雇用し，各作業者が同一の文章集合に対して作業を行った．このうち，4 名が述語項関係のアノテーションの経験者である．データ収集の詳細は文献 [11] を

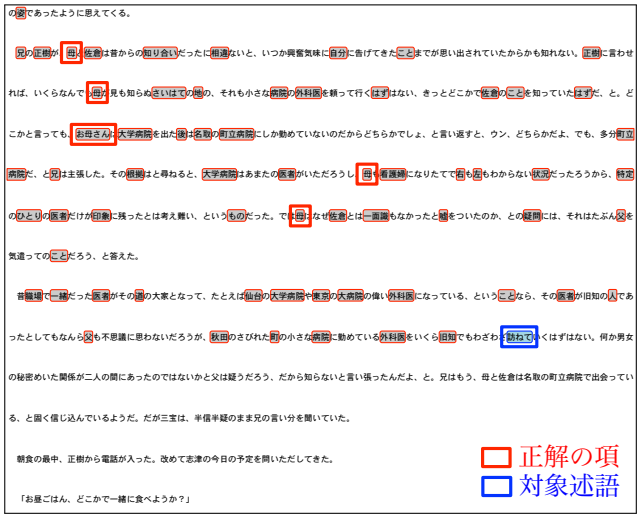

図 1: 単一述語項構造アノテーションの作業画面（1 事例）

参照されたい．

収集した 20 名のデータのうち，視線の平均計測エラー率が 0.15 を越える 4 名の作業者と，視線の画面鉛直方向の計測誤差が大きい 1 名の作業者を除いた 15 名の作業者 a〜t のデータを本研究で利用する．この 15 名の作業者のアノテーション作業結果の誤り率（アノテーション結果が不正解である割合）を表 1 に示す．表 1 から，ア

表 1: 作業者ごとのアノテーション誤り率

作業者	誤り率	作業者	誤り率
n*	0.09 (20/221)	g	0.31 (69/221)
c*	0.15 (33/221)	h	0.32 (70/221)
t*	0.17 (38/221)	d	0.36 (79/221)
k*	0.18 (39/221)	a	0.37 (81/221)
b	0.19 (43/221)	q	0.37 (82/221)
e	0.22 (48/221)	l	0.39 (87/221)
p	0.24 (53/221)	m	0.41 (91/221)
i	0.26 (58/221)	計	0.27 (891/3,315)

*は述語項関係アノテーションの経験者であることを表す．

ノテーション経験者 4 名の誤り率が 0.09〜0.18 である一方で，未経験者は 0.19〜0.41 と高い誤り率を示しており，アノテーションの作業全体の誤り率は経験があるか否かに関係していることがわかる．また，任意の 2 名の作業者が共通してアノテーションを誤った事例集合の重複率（Jaccard 係数）の平均は 0.69 であり，作業者ごとに異なる事例でアノテーションを誤っていることがわかる．この結果から，与えられた問題から得られる言語的な情報に基づいて問題の難易度を推定するだけでは，個々のアノテーション作業者の各事例に対する信頼度を推定することが困難であることがわかる．

3 アノテーション誤り修正作業とアノテーションの信頼度推定モデル

アノテーションされた結果には少なからず仕様と異なったアノテーションのような誤りが含まれるため，これを効率的に修正することが必要となる．そこで，本研究で提案する信頼度推定モデルが出力する各事例の信頼度にしたがって事例集合をソートし，その順序でアノテーションの誤りを修正した場合に，どの程度修正作業の効率化が可能かを調査する*1．

表 2: アノテーション誤り検出に使用する素性

カテゴリ	素性名	説明
pre	pred_pos	述語の品詞
	posi_cooc_flag	新聞記事から推定した〈名詞, 格助詞, 述語〉の共起スコア [4] が正である項候補が述語と同一段落内に出現
mid	anno_duration	作業開始から項を選択するまでの時間とその合計時間
	gazed_cands_ratio	文章中で注視が起きた項候補の割合
post	arg_case	アノテーションされた項の直後に出現する助詞
	arg_case_freq	アノテーションされた項と同一テキストの項候補の直後にある助詞の出現頻度
	arg_distractors	アノテーションされた項の意味カテゴリと同一のカテゴリとなる項候補の個数

本研究で提案するアノテーションの信頼度推定モデルでは，述語項関係アノテーションにおける作業者の各アノテーションに対し，正しくアノテーションできた場合には 1 位，それ以外の場合には 2 位といった異なる順位のラベルを割り当て，それを Ranking SVM [7] で学習することでランカーを得る．評価時には，アノテーション作業者の作業結果集合を入力し，ランカーが出力する順序に基づいて事例を並べ，その順序で修正作業を行う．

ランカーの学習には表 2 に示す 3 種類のカテゴリの素性を利用する．一つはアノテーション対象となる文章自身から得られる対象述語の品詞等の素性（pre）であり，このカテゴリの素性のみを使ったランカーの学習がベースラインとなる．もう一つは，アノテーション作業にかけた時間や，作業者の視線情報など，作業中に収集した情報から得られる素性（mid）であり，最後の一つがアノテーションされた項に関係する情報から得られる素性（post）である．このうち，アノテーションされた項の意味カテゴリについては事前調査の結果，有生物とそれ以外で異なった問題の難易度となることがわかったため，その特徴を捉えるために日本語語彙大系 [6] の名詞意味カテゴリを利用して項候補を有生物とそれ以外に分類して，その分類結果に基づく素性を導入した（表 2 の arg_distractors）．

4 評価実験

アノテーション結果の信頼度推定モデルの出力する信頼度が低い順に事例を並べ，その順序でアノテーションの誤りを修正する状況で，3 節で導入したどの素性タイプを利用することが効率的な修正作業になるのかを調査する．評価時には，図 2 に示す設定のように，対象データとなる 221 記事を 10 分割し，さらに 15 名を独立に評価する．学習時には 14 名の 9/10 のデータを学習事例として学習し，評価時には残りの 1 名の残りの 1/10 を評価対象とすることで学習と評価のデータの依存関係を排除する．この設定により，既知の作業者・データに基づいて学習を行ったモデルの性能を，未知の作業者に対する未知のデータへのアノテーション結果に対して評価することになる．

*1 ただし，提示された誤りはそのタイミングで必ず正しく修正されることとする．

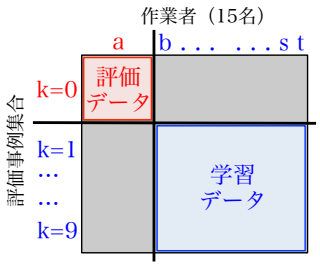

図 2: 学習データと評価データの分割

表 3: 誤り検出実験の評価結果

モデル	平均精度
ランダム	0.365
pre	0.421
mid	0.552
post	0.542
pre+mid	0.562
pre+post	0.549
mid+post	0.594
pre+mid+post（全素性）	**0.602**

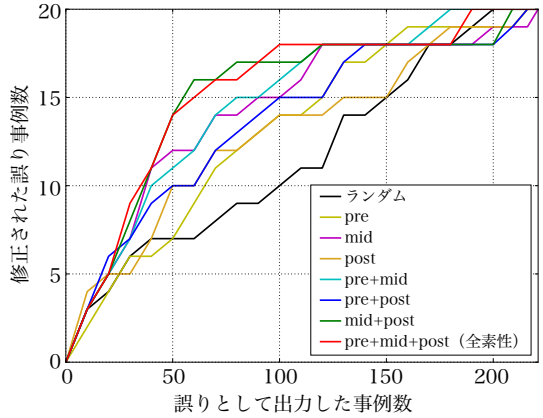

図 3: 出力事例数と誤り事例数の関係（作業者 n）

4.1 実験結果

修正作業の効率性を評価する評価尺度として，各 1/10 の評価事例集合における誤り検出問題においてアノテーション誤りの事例を適合事例とみなした平均精度のマクロ平均を採用した．この評価尺度に基づく評価結果を表 3 に示す．pre 素性だけを利用したベースラインに加え，ランダムに事例を選択した場合の結果も掲載する．表 3 に示した結果から，作業中の情報である mid 素性や，作業結果に関する情報から得られた post 素性を導入することで，誤り検出の平均精度が向上していることがわかる．

次に，各素性の効果を具体的に示すために，作業者のうちアノテーションの正解率が最も高かった作業者 n の作業結果をランカーが出力する順位まで修正した場合に，何件の誤り事例を修正できたことになるのかの関係をグラフにまとめたものを図 3 に示す．このグラフから，例えば，50 件修正した時点ではランダムベースラインと比較して，提案する 3 種類の素性を導入したモデル（図 3 の pre+mid+post）が約 2 倍の誤りを検出できていることがわかる．この結果から，作業中の視線などのアノテーションの行為や作業時間の情報，作業結果から得られる情報が誤り検出において有効に働くことがわかる．

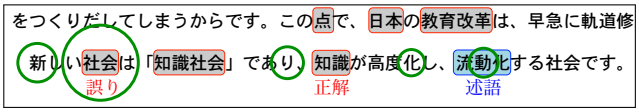

図 4: 推定信頼度が高いがアノテーション誤りの事例

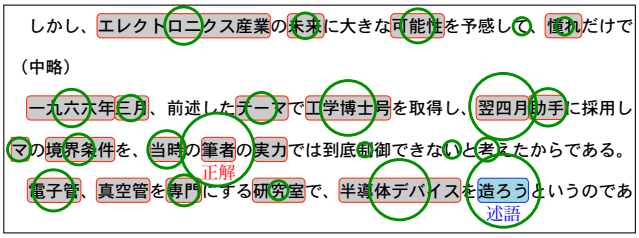

図 5: 推定信頼度が低いが正しいアノテーションの事例

4.2 誤り分析

次に，(1) モデルが出力した信頼度が高いがアノテーション誤りであった事例，(2) 信頼度が低いが正しいアノテーションであった事例のそれぞれの具体例を分析し，提案する信頼度推定モデルがどのような場合に信頼度を適切に推定できないかを調査した．まず，(1) の推定信頼度が高いアノテーション誤りの事例では，例えば，図 4[*2] で述語「流動化（する）」に対し，「（新しい）社会」ではなく「知識」をアノテーションする必要があるが，このように述語と項が近傍に出現する，もしくは同一文内に出現するような場合では，正しいアノテーションをした場合と誤ったアノテーションをした場合で視線や作業時間などの作業者の振舞いにほとんど差が無いため，本研究で導入した素性セットではその特徴が捉えられなかったと考えれられる．

また，(2) の推定信頼度が低いが正しいアノテーションの事例として，例えば，図 5 のように，多くの項候補に視線の停留が起こり，作業時間も長い場合はアノテーションを誤る場合が多いが，この例ではその傾向に反して正しくアノテーションされている．提案する信頼度推定モデルでは，時間をかけて作業している場合は一般に信頼度が低く見積られるが，例外的に時間がかかっても正しくアノテーションされる場合も存在するため，今後はその例外的な特徴を pre 素性，post 素性に導入することにより，信頼度推定を精緻化する必要がある．

5 作業者の信頼性推定

3 節で問題とした各事例のアノテーションの信頼度推定に加え，アノテーション作業者の作業の信頼度を推定することも，高品質なコーパス構築のために重要となる．本研究では，この作業者の信頼度推定を作業者の全事例に対するアノテーションの信頼度を組み合わせることで実現し，その結果と全事例に対する正解率の関係を調べることで，手法の有効性を調査する．

3 節で提案した手法が推定する信頼度の傾向を調査した結果，アノテーションの結果の正解率が高い作業者はアノテーションを行う際に典型的に同じような振舞いをするため，どの事例に対しても高い信頼度が保たれるが，正解率の低い作業者は作業を行う際に迷いが生じ，それ

[*2] 緑の円が作業者の視線を表し，円の大きさは視線がその位置に留まった時間の長さを表わす．

表 4: 推定された作業者の信頼度と正解率との相関

	相関係数	Kendall
ベースライン（全事例）	0.152	0.085
提案手法（文内最大値）	**0.624**	**0.467**

が視線の動きや作業時間に反映されるため，事例に対する信頼度が低く推定される傾向があることがわかった．この傾向にしたがい，作業者の各事例に対するアノテーションの信頼度の平均を信頼度スコアとすることで，その作業者の信頼性推定が実現できると考えられる．

ただし，この方法では，作業者が例外的な振舞いをしたために，個別の事例に対するアノテーションの推定信頼度が真に求まるべき信頼度の値と大きく異なる事例も含まれてしまうため，そのような例外を除いて作業者の信頼度を推定することが望ましい．そこで，前述の傾向が最もよく現れると考えられる文内の項がアノテーションされた事例のみを対象に，10 分割した評価事例集合のそれぞれから信頼度が最も高いアノテーションのスコアを抽出し，その平均をとった値を信頼度スコアとして利用する方法を提案する．

この二種類の方法で推定した作業者信頼度と，各作業者の全事例の作業結果の正解率との相関係数，Kendall の順位相関係数を調査した結果を表 4 に示す．この表より，我々の予想通り，事例を選別して利用することでより高い相関係数を得ることができていることがわかる．提案する作業者の信頼度推定を用いることで，相関係数で 0.624，Kendall の順位相関係数で 0.467 という高い相関係数の値を得ている．

また，提案手法が出力した作業者の信頼度スコアと作業者の正解率を作業者ごとにプロットしたものを図 6 に示す．この図より，作業者 b，k，t などの推定は誤っているものの，正解率の高い作業者 c と n の信頼度を適切に高く見積ることができていることがわかる．次に，信頼度の推定値が正解率と大きくかけ離れた作業者について，信頼度スコアの推定に使用された事例を，その事例をアノテーションしたときの視線の動きとともに調査した．この結果，これらの事例では，述語が含まれる行の最初から項を探しているために，典型的な視線の動きである述語から行頭に向かって項を探す場合よりも探索に時間がかかり，結果として信頼度が低下していることがわかった．今後はこのような通常とは異なる視線の動きに対して，個別に信頼性を推定するなどの細やかな対応をすることで，より正確な作業者の信頼性推定を行う必要がある．

6 おわりに

本稿では，述語項関係のアノテーションを例題に，高品質なコーパス構築のためのアノテーション事例の効率的な修正について述べ，それを実現するために事例単位のアノテーション事例の信頼度を推定する手法を提案し，推定された信頼度が低い順にアノテーションをやり直すことで効率的にアノテーションの誤りを修正できることについて述べた．さらに，提案したアノテーション事例に対する信頼度に基づき，アノテーション作業者の信頼度を推定する手法も提案し，作業者のアノテーションの正解率との相関を調査した結果，中程度の相関を得られることを示した．

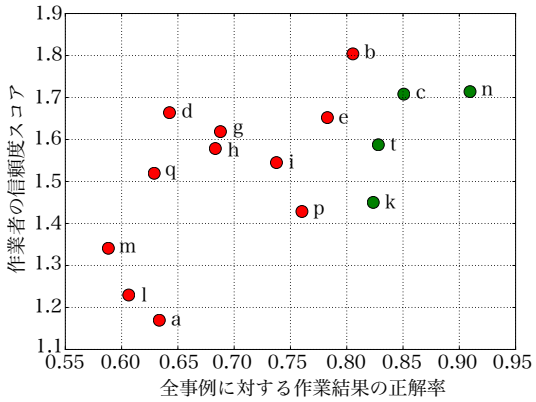

図 6: 正解率と推定した信頼性の相関

今後の課題としては，視線情報のより効果的な利用方法，アノテーション対象の事例やアノテーション結果から信頼度推定のために必要となる有益な言語情報を調査することに加え，述語項関係以外のアノテーションを対象とした場合に，同様のアプローチが有効であるかといったアノテーション課題横断的な調査を行いたいと考えている．

参考文献

[1] Steven Abney, Robert E. Schapire, and Yoram Singer. Boosting applied to tagging and PP attachment. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 38–45, 1999.

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

[3] Andrea Esuli and Fabrizio Sebastiani. Improving text classification accuracy by training label cleaning. *ACM Transactions on Information Systems*, Vol. 31, No. 4, 2013.

[4] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 6, No. 4, pp. 1–22, 2007.

[5] 飯田龍. 意味・談話処理課題の規格化とその緒問題. 人工知能学会誌 特集「ポスト経験主義の言語処理」, Vol. 27, No. 3, pp. 318–325, 2012.

[6] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系 CD-ROM 版, 1999.

[7] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 133–142, 2002.

[8] 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.

[9] 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.

[10] 光田航, 飯田龍, 徳永健伸. 視線と操作情報を利用した誤りアノテーションの検出. 言語処理学会第 20 回年次大会発表論文集, pp. 508–511, 2014.

[11] 光田航, 飯田龍, 徳永健伸. 単一述語項関係アノテーション課題における視線情報の収集と分析. 情報処理学会第 217 回自然言語処理研究会, pp. 1–8, 2014.

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