アノテーションされた結果の人手分析
–ポストアノテーションと事例クラスタリング–

飯田 龍
東京工業大学 大学院情報理工学研究科
ryu-i@cl.cs.titech.ac.jp

近年，自然言語処理の問題の多くは，その問題の出力として得たい情報（タグ）を人手でコーパスにアノテーションし，そのコーパスを訓練・評価用データとして機械学習などの統計的な手法で利用する方法に基づき研究が進められている．この研究のパラダイムは，形態素・構文解析のような基盤処理から，情報抽出や機械翻訳などの応用処理まで広く普及している．特に，CoNLL や TREC，NTCIR などの評価側のワークショップ・国際会議では構文解析や意味役割付与，質問応答や含意関係認識など，さまざまな自然言語処理の問題に対する訓練・評価用データが人手で作成され，研究者間でアノテーション済みのデータが共有されることで一定の成果をあげてきた．

このアノテーションと機械学習に基づく研究の多くは，典型的に，既存研究で扱うことができていない点を改善する，もしくは既存研究で扱っている観点に関するモデル化を洗練することを試みるという方針で研究が進められている．また，提案された手法の妥当性は，その手法が対象とする問題を含む訓練・評価用データを利用した定量的な評価実験の結果を，例えば，再現率・精度・F 値などの評価尺度を用いて評価することで見積られている．ただし，それらの研究では，ある評価尺度に基づいた定量的な評価しか行われていない場合が多いため，仮説としてあげた観点に関して問題が解決できたことによって精度が向上したと断言できないという問題がある．このような定量的な評価結果のみを採用する背景には，一度人手で評価用データを分析してしまうと，そのデータを使った評価結果がオープンテストを用いた評価結果とみなすことができないという点に関連しているのかもしれないが，実データを見ずに研究者の直観だけで問題を解くことが繰り返されると，解くべき問題と解析手法のずれが大きくなる可能性がある．

これを解決するには，実際に問題を解いた結果を人手で分析れば良いが，誤り事例を分析するための方法論については，どのような種類の事例をどの程度見ればよいかという点について客観的な判断基準が無いというのが現状である．このため，研究者ごとに，個別の誤り分析の方法を採用し，異なる方針で誤り事例の分析が進められている．例えば，再現率・精度・F 値で評価される問題については，各事例に関して解析の信頼度を導入し，信頼度が高く false positive な事例を分析したり，信頼度が低いが false negative となっている事例を分析することで手法と実際の問題のずれを調査するというヒューリスティックなどが採用されているが，この分析方法の妥当性[1]や，分野にどの程度浸透しているかについての議論は無い．

この状況を変革するために，研究者間で共有すべき誤り分析の方法論を策定することが急務であるが，分析の方法は必ずしも論文に記述されるとは限らないため，それを収集することからはじめる必要がある．具体的には，各研究者がどのような誤り分析の方法を採用し，それに基づいて実際にどのような分析を行ったかの履歴を残す作業を考える．問題を解く前に正解を作成するアノテーションをここではプレアノテーションと呼ぶことにするのに対し，問題を解いた後に分析する対象に対して分類項目などを人手でアノテーションする作業をポストアノテーションと呼ぶ．例えば，既存研究 [1] では，ゼロ照応関係を暫定的に 6 種類に分類しているが，その分類にしたがって誤り事例に対してポストアノテーションを行うことが考えられる．このポストアノテーションの作業結果は，分析者によって分析する粒度や内容が異なることが考えられるが，異なる誤り分析の結果であっても，それらを公開されているデータにアノテーションし，研究者間で共有することで，誤り分析の方法そのものが共有されたり，解析手法横断的にどのような種類の誤りが多いのかと

[1] この分析方法では解析手法を一つ固定してその出力と正解とのずれを調査することになるが，一方で比較対象とするベースラインを想定している場合はそのベースラインと解析手法の間の差分を調査すべきかもしれない．

いう調査を明示的に行うことができると考えられる．

ポストアノテーションの作業を行うことで，正解データの内容を調べてしまうことになり，オープンテストとして利用できるデータが無くなるという問題が起こるが，近年では CoNLL などで同一タスクに対して複数回異なるデータが作成される状況にあるため，誤り分析を明示的に行い，その結果解けなかった種類の問題を中心的に事例を収集し，プレアノテーションすることで，今後解くべき問題に注力した事例収集を行うことが可能になると考えられる．このような中・長期的なデータ収集ということについても分野全体でどのように行っていくかを検討する必要があると考えられる．

上述のポストアノテーションの考え方は，与えられた問題を解いた結果に対する人間の分析の方針を議論したものだが，同様の考え方は機械学習を用いて学習する前に問題の傾向を調査するための訓練用データの人手分析の際にも適用できる．例えば，機械学習に基づく手法では言語理論や人間の内省に基づき素性集合が設計されるが，事例を人手分析するための方法論は素性設計の妥当性を調査するのにも利用可能である．機械学習に基づく手法ではある事例に対して抽出された素性集合に基づいて分類先のラベルを選択するため，同じラベルに分類されるべき事例から抽出される素性集合は類似していることが望ましいが，素性の設計によっては必ずしもそうなるとは限らない．そこで，訓練事例集合をあらかじめ設計された素性にしたがって素性空間上でクラスタリングし，そのクラスタリングの結果を用いて素性設計の良さを評価することを考える．ここで，このクラスタリングを事例クラスタリングと呼ぶ．このクラスタリングの結果，得られたクラスタと各クラスタに含まれるアノテーション済みのラベルの純度（同じラベルが含まれる割合）を調査することで，素性設計の妥当性を見積ることができる．例えば，あるクラスタに複数のラベルが密に混在する場合には，それらを弁別するような特徴（素性）を新規に導入しない限り，そのクラスタに含まれる事例を適切に分類することができないことがわかる．そのような場合は，クラスタ内の事例の実データを比較することで，新規に導入すべき素性を顕在化できると考えられる．

この事例クラスタリングを採用することで，問題を解くことなく，訓練事例内で閉じた分析が可能となる．上述の説明では素性抽出の結果得られた素性ベクトルをそのまま利用してクラスタリングすることを書いたが，訓練事例集合を用いて何らかの機械学習アルゴリズム（線形 SVM や最大エントロピーモデルなど）を用いてあらかじめ素性の重みベクトルも求め，そのベクトル積をとった結果に基づいて類似度を求めることで，素性の重要度も加味してクラスタリングを行うことも可能となる．

このように，ポストアノテーションと事例クラスタリングを駆使することで，どのような種類の問題が解けないのかを顕在化することができると考えられる．本発表では，この 2 種類の方法の実現可能性などについて議論したい．

参考文献

[1] 飯田龍, 笹野遼平. 日本語ゼロ照応関係に対する特徴分類とそのアノテーション．テキストアノテーションワークショップ・コンテスト, 2012.