6T-04

ツイートコーパスを用いた意見抽出システムの構築

野崎雄太† 櫻井義尚†
明治大学†

1. はじめに

近年企業が Twitter をマーケティングに利用する「ソーシャルリスニング」が広く行われている．しかし，膨大なツイート集合から人手で意見を抽出するのは困難であり，また教師あり学習手法を用いた意見抽出システムも Twitter の特性上，アノテーションにおいて教師データが不均衡になり，抽出精度が落ちるという課題がある．先行研究ではこの課題に対し，アンダーサンプリングやオーバーサンプリングなど教師データレベルで対策を行っているが，本研究ではツイート集合に段階的にフィルタリングをかけることによって，不均衡化が緩和されたツイートコーパスを構築する，アノテーションレベルでの対策手法を提案する.(図 1)

2. 関連研究

不均衡データに対するアプローチの先行研究として澤崎ら[1]の研究が挙げられる．澤崎ら[1]は不均衡データである livedoor ニュースコーパスを利用して類似単語の入れ替えや文節の並び替えを行う事によって文書をオーバーサンプリングする手法を提案し，SVM，Random Forest を使って学習を行うことによって有効性を検証した．本研究では教師データ作成時のアノテーション手法を提案する．

Twitter などの Web 上からの意見抽出に関する研究では立石ら[2]は評価表現辞書を用いて意見を抽出した．本研究では辞書に含まれている意見表現が明示されているツイートだけでなく，辞書に含まれていないが，文脈上等で意見と判別されるツイートもコーパスに含ませる構築手法を提案する．

3. 意見抽出システム

本研究で提案するのは，提案手法で構築したコーパスで学習した分類モデルを利用した意見抽出システムである．システムの構築フローは

① ツイート収集
② コーパスの構築
③ 機械学習

である．本章ではシステム構築のそれぞれのステップについて述べる．

3.1 ツイート収集

Twitter から収集する意見の対象となる事物名でキーワード検索を行い，ツイート集合を収集する．

3.2 コーパスの構築

前項で収集したツイート集合を「意見」の定義に基づいてアノテーション作業を行い，教師データのコーパスを構築する．

アノテーション作業には膨大な人手と時間的コストがかかるため，より信頼性の高い教師データの構築手法が求められる．しかし，収集したツイート集合から教師データの作成のためにランダムサンプリングを行ってアノテーションを行うと，Twitter などの SNS の特性上，「意見」ではないツイートが非常に多く抽出され，不均衡なデータになり，学習が難しくなる．

また，意見表現をまとめた辞書を構築し，それを収集したツイート集合にフィルタリングを行い，抽出されたツイート集合にアノテーション作業を行う立石ら[2]の研究が存在するが，辞書に収録されている表現を含むツイートのみが教師データとして学習され，辞書に収録されていない表現ではあるが「意見」として抽出されるべきツイートが教師データに含まれず，抽出漏れが起きる可能性がある．また，反対に辞書でフィルタリングして抽出されたツイート全てが必ずしもアノテーションで「意見」と判定されるわけではない．そのため本研究では収集したツイート集合を以下の 4 段階の機械的なプレフィルタリングを用いたサンプリングによって教師データを作成することによって不均衡化を緩和するアノテーション手法を提案する．

第 1 段階フィルタ　評価表現辞書[3]を，収集したツイート集合にプレフィルタリングする．フィルタで抽出されたツイート集合と抽出されなかったツイート集合それぞれ5:5になるようにランダムサンプリングを行った上でアノテーションを行う．

第 2 段階フィルタ　評価表現辞書[3]を，収集したツイート集合にプレフィルタリングする．フィルタで抽出されたツイート集合と抽出されなかったツイート集合それぞれ6:4になるようにランダムサンプリングを行った上でアノテーションを行う．

Construction of Opinion Mining System by Twitter
† YUTA NOZAKI, YOSHITAKA SAKURAI, Meiji University

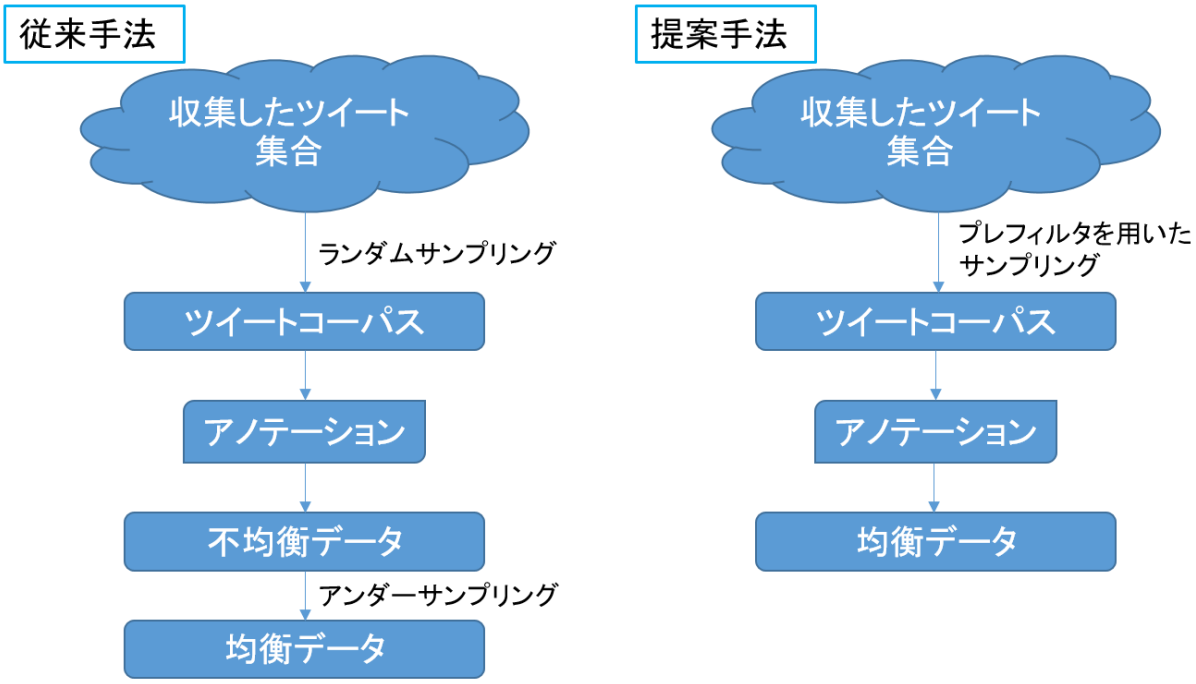

図 1　従来手法と提案手法の不均衡データに対する対策手法の比較

第 3 段階フィルタ　評価表現辞書[3]に MeCab[4]で形態素解析を行い,「意見」と判別される文章に頻出する「形容詞」,「副詞助詞類」,「助動詞」,「名詞副詞接続」,「名詞形容動詞語幹」に該当する品詞が含まれている表現を抽出した新しい辞書を作成する．また，収集したツイート集合に対して表記ゆれの解消，分かち書きを行った上で作成した新しい辞書でプレフィルタリングを行う，抽出されたツイート集合と抽出されなかったツイート集合それぞれ 8:2 になるようにランダムサンプリングを行った上でアノテーションを行う．

第 4 段階フィルタ　評価表現辞書[3]に MeCab[4]で形態素解析を行い，第 3 段階フィルタよりも厳格に「意見」と判別される文章に頻出する「形容詞」,「副詞助詞類」,「助動詞」,「名詞副詞接続」に該当する品詞が含まれている表現を抽出した新しい辞書を作成する．また，収集したツイート集合に対して表記ゆれの解消，分かち書きを行った上で作成した新しい辞書でフィルタリングし，その上で「意見」とされるツイートは文字数が多くなると仮定して，70 文字以上の文字数があるツイートのみを抽出した．その後，抽出されたツイート集合と抽出されなかったツイート集合それぞれ 8:2 になるようにランダムサンプリングを行った上でアノテーションを行う．

4 段階のフィルタに分けること，またフィルタで抽出されなかったツイート集合も一定の割合で含めることによって不不均衡化を緩和し，特定の表現や文字数への教師データの偏りを防ぐことができる．

3.3　機械学習

前項での提案手法を用いて作成したツイートコーパスを用いて機械学習モデルの構築を行う．

学習モデルは文書分類タスクで用いられる SVM, Random Forest，Logistic regression を利用し，それぞれのモデルで意見抽出システムを構築する．

4.　おわりに

本研究では不均衡データになりやすい意見抽出タスクの教師データ作成において，新しいコーパス構築手法を提案し，この手法を用いて作成した教師データを学習させ，意見抽出システムを構築する手法を示した．

今後の課題として，提案手法を用いて実際にアノテーションを行うことによってコーパスを構築し，データの不均衡化の緩和効果を検証する．その上で，構築したコーパスを用いた機械学習モデルの構築と精度評価を行い，従来からの不均衡データに対する手法と比較することで，コーパス構築まで含めた機械学習モデル構築手法の比較評価を行い，その有効性を検証する．

参考文献

[1] 澤崎夏希,遠藤聡志,當間愛晃,山田孝治,赤嶺有平(2017),「量的不均衡データに対する学習精度改善のための文書かさ増し手法」『WI2 研究会第 11 回研究会オンライン・プロシーディングス』
[2] 立石健二,石黒義英,福島俊一(2001),「インターネットからの評判情報検索」『情報処理学会研究報告』NL144-11,　pp.75-82
[3] 「評価値表現辞書 (評価表現辞書) :Syncha」 http://www.syncha.org/evaluative_expressions.html(2020 年 1 月 9 日参照)
[4] 「MeCab」https://taku910.github.io/mecab/(2020 年 1 月 9 日参照)