2ZD-02

IBM Watsonによる授業質問Tweetの自動分類

柏原　直志†　　藤澤　公也†
東京工科大学メディア学部†

1. はじめに

大学における大人数授業では講義者が受講者から質問を効率的に受け付けることは難しい．これまでにも Twitter でリアルタイムに質問を受け付けるなどの試みが行われている[1]．Twitter を用いることで匿名性が高まり，質問をする敷居が低くなる一方で，授業とは関係のない，あるいは質問ではない内容の Tweet も多くなってしまう[2]．

本研究では授業における質問を Twitter でリアルタイムに受け付けた際に，講義者に伝えるべきかどうかを，Watson NLC による自動的な分類・判定を試み，その精度について検証を行った．

2. Watson NLC による Tweet の分類

Watson NLC とは IBM が開発した AI である Watson の自然言語分類器である．ラベル付けされたテキストを NLC に学習させることで，新たに入力されたテキストと各ラベルとの関連度を算出することができる[3]．

当初，授業に対して行われた Tweet のラベルは質問，感想，無意味，テスト投稿，時事ネタ，授業内容メモ，他授業関連，要望の 8 個に分かれていたが，実験過程においてテスト投稿，時事ネタ，授業内容メモ，他授業関連，要望は Tweet 数が少なく，学習データとするには適さないと判断し除外した．また，複数のラベルに分類されると判断された Tweet も除外した．

ラベル付けの基準は，質問ラベルが授業内容に関連し，質問の Tweet，感想ラベルが授業内容に関連し，質問ではない Tweet，無意味ラベルが授業内容に関連しない Tweet となっている．

3. 分類精度の検証実験

3.1. 実験方法

ラベル付けしたデータの一部を検証データとして残し，残りを学習データとし NLC に学習させる．NLC の学習結果に検証データを分類させ事前に人の手で分類したものと比べ，正確に分類できているかを検証した．

Automatic classification of questions tweets by IBM Watson
†Kashiwabara Naoshi, Fujisawa Kimiya, School of media science, Tokyo University of Technology

3.2. 全体の 8 割学習した場合

質問・感想・無意味に分類された 472 個の Tweet の内，それぞれの分類の Tweet の 8 割を学習データとして NLC に学習させた．残りの 2 割を検証データとして検証した結果が表 1である．

左側にある事前の分類の各ラベル横の数字は検証データの数，表の数値は事前の分類に対して NLC によって行われた分類の割合を示している．表は全て同様の読み方をする．

表 1 全体の 8 割を学習した際の結果

		NLC による分類結果		
		質問	感想	無意味
事前の分類	質問(14)	79%	14%	7%
	感想(30)	7%	77%	17%
	無意味(50)	0%	8%	92%

質問 Tweet の 21%を質問として判断できていないが，感想・無意味を質問と判断している割合はどちらも 10%未満となっている．

3.3. 学習時の比率を考慮した場合

3.2の実験では無意味の学習データが質問よりも多いいため Tweet の分類が無意味に偏ってしまう傾向にあるのではないかと考えた．そのため，感想・無意味は質問の学習データの数との比率を考慮して再度実験を行った．質問の学習データは 70 個の質問 Tweet の内 50 個とした．

(1)質問 Tweet とそれ以外をそれぞれ同数とした場合

質問・感想・無意味に分類された 472 個の Tweet の内，それぞれ 50 個の Tweet を学習データとして NLC に学習させた．残りの 322 個の Tweet を検証データとして検証した際の結果が表 2である．

表 2 質問・感想・無意味を同数ずつ学習した際の結果

		NLC による分類結果		
		質問	感想	無意味
事前の分類	質問(20)	85%	15%	0%
	感想(100)	9%	70%	21%
	無意味(202)	4%	21%	75%

20 個の質問の内 15%を感想だと判断してしまっているが，感想と無意味を質問だと判断しているのはどちらも 10%未満にとどまっている．

(2)質問 Tweet とそれ以外のすべてを同数とした場合

質問・感想・無意味に分類された 472 個の Tweet の内，質問を 50 個，感想・無意味をあわせて 50 個の合計 100 個を学習データとして NLC に学習させた．残りの 372 個の Tweet を検証データとして検証した際の結果が表 3である．なお，元のデータと同じ比率になるように感想・無意味の学習データは感想を 19 個，無意味を 31 個とした．

表 3 質問と同数の感想・無意味を学習した際の結果

		NLC による分類結果		
		質問	感想	無意味
事前の分類	質問(20)	95%	5%	0%
	感想(131)	37%	24%	39%
	無意味(221)	6%	6%	87%

20 個の質問 Tweet の内 5%を感想と判断しているが，(1)の実験と比べ質問 Tweet を質問と判断する精度は上がっている．しかし，感想 Tweet の 37%が質問と質問と判断されており，(1)の実験と比べて精度は大幅に下がった．

3.4. 質問 Tweet とその他 Tweet に分けた場合

ラベルを講義者に伝えるべき Tweet とそうでない Tweet の 2 種類に減らし実験を行った．

質問・感想・無意味に分類された 472 個の Tweet の内，50 個の質問は前の実験と同様に質問，感想・無意味を合わせた 100 個をその他とし，合計 150 個を学習データとして NLC に学習させた．残りの 322 個の Tweet を検証データとして検証した際の結果が表 4である．

表 4 質問・その他を学習した際の検証結果

		NLC による分類結果	
		質問	その他
事前の分類	質問(20)	80%	20%
	その他(302)	2%	98%

質問とすべき Tweet の 20%がその他と判断されてしまっており，3.3の(1)(2)と比べて精度が低下している．しかし，その他にすべき Tweet が質問と判断されてしまっているのは 2%にとどまっており，3.3の(1)(2)とくらべて精度が高まっている．

4. 実験結果と考察

4 つの実験の結果の中で最も質問 Tweet を質問と判断した確率が高かったのは3.3の(2)の実験であった．しかし，感想・無意味 Tweet を質問と判断した確率が最も高かったのも同様の実験であった．これらの要因として考えられることは学習データとして用意した質問 Tweet と感想・無意味 Tweet の数の違いである．この実験では質問 Tweet の学習データが感想・無意味 Tweet よりも多くなっているため最も多いデータに影響されて質問に判断された Tweet が増加したのではないかと考えられる．

実際に講義者に質問を伝えるツールとして授業で使うことを考えたとき，どの実験の学習方法を使うべきか．3.3の(2)の学習方法では正しく質問と分類された Tweet は最も多いが，誤って質問と分類された Tweet も最も多い．3.3の(1)の学習方法では質問と誤って分類された Tweet は少ないものの，正しく質問と分類された Tweet も少ない．

いずれの学習方法が授業に適切かは，Tweet の数などによって変わってくる．授業中の投稿 Tweet が多く多少の取りこぼしを無視できる場合は3.3の(1)の学習方法を使い，授業中の投稿 Tweet が少なくなるべく多くの Tweet を拾いたい場合は3.3の(2)の学習方法を使うなど，状況に応じた使い分けをすることが必要となる．

5. おわりに

今回の実験では学習データとなるラベル付けされた Tweet のラベルが正しく信頼性があるかの検証を行っていない．そのため，学習データを見直すことで実験の質の向上が見込める．また，1 年度分の同じ授業で投稿された Tweet だけをデータとして扱ったため，異なる年度や授業の Tweet を学習させたときの結果やそれらを混ぜたときの結果がどうなるかなどの実験を試みていきたい．

参考文献

[1] 藤澤公也．"授業支援への twitter の活用: スライドに tweet を表示する試み" システム/制御/情報 55.10 (2011): 446-451.

[2] 藤澤公也，天野直紀．"大規模授業における Twitter によるアクティブラーニング." JeLA 会誌 12 (2012): 90-98.

[3] IBM - Natural Language Classifier 自然言語分類 | Watson Developer Cloud - Japan, https://www.ibm.com/watson/jp-ja/developercloud/nl-classifier.html