罹患者への定型的応答を利用した罹患ツイートの自動獲得とRNN罹患判定器学習への適用

浅川　玲音　　秋葉　友良
豊橋技術科学大学 情報・知能工学専攻

r163301@edu.tut.ac.jp　akiba@cs.tut.ac.jp

1　はじめに

本研究では Twitter を用いた疾病サーベイランスのための病気罹患判定に取り組む．Twitter の保有する大量の情報は，リアルタイム性に優れている点などから，インフルエンザ流行の早期発見のアプローチのひとつとして Twitter ベースの手法が考えられてきた．Twitter ベースシステムの予測はインフルエンザ流行記録と相関があることは荒牧ら [1] によって報告されている．Twitter ベースのアプローチでは，対象疾患への罹患を表すツイートを検出し、集計することで流行を予測する．従って罹患判定の技術は Twitter を用いた疾患サーベイランスシステムの中心技術である．罹患判定の先行研究の多くは教師ありの機械学習を利用しており，いずれもある程度の学習コーパスを人手でアノテーションして作成する必要がある [2][3]．しかしながら人手によるラベリングは非常に高コストであることなどから，学習コーパスを十分に用意することは非常に困難である．医療言語処理のシェアドタスク“NTCIR-13 MedWeb task[4]”が開催され，ツイートの罹患判定のためのアノテーションガイドラインと少量のラベル付きデータが提供されたが，ツイートの多様な罹患表現を学習するのに十分な量であるとは言えない．

本研究では，罹患者への定型的応答を利用して自動的に学習コーパスを獲得し，そのコーパスを用いて RNN ベースの罹患判定器の学習データをデータ拡張するアプローチを提案する．この手法では，まず自動獲得した大量の自動獲得コーパスを用いて罹患判定器の各パラメータを学習しておく．それから前段階で学習したパラメータを初期値として，人手でラベル付けされた少量のコーパスを用いて学習を進める．このように学習を二段階に分けることで，互いの性質を補完し合う形で二種類のコーパスを効果的に組み合わせることができる．本稿の構成は次のとおりになっている．2 節でツイートの罹患判定というタスクについてより詳しく説明し，提案する自動的な学習コーパスの獲得方法とそれを用いたデータ拡張法についてまとめる．提案手法を評価する実験については 3 節でまとめ，4 節で結論を述べた．

2　提案手法

2.1　問題設定

本研究では，NTCIR MedWeb タスクのような複数疾患への罹患判定タスクの基盤となる，疾患を区別しない 2 クラス分類に取り組む．すなわち，対象のツイートについて，その投稿者または周囲の人間が今現在何らかの疾患/症状に罹患していることを意味しているか否かを推定するタスクである．何らかの疾患への罹患が有れば罹患 Positive クラス，無ければ罹患 Negative クラスであると判定する．表 1 にそれぞれのクラスの例を示す．

例から分かるように罹患判定は難しいタスクであり，病名だけでなくその周囲の表現も罹患判定の重要な手がかりとなる．従って，教師ありの手法で罹患判定器を学習するには多様な表現をカバーできるだけの大量な学習データが必要となる．提案する自動獲得手法では，様々な疾患への罹患を有するツイートのサンプルを自動的に大量に獲得でき，幅広い疾患へのマルチラベル分類タスクへの応用が期待できる．

表 1: 罹患 Positive と罹患 Negative の具体例

No.	Tweet	Class
1	咳つらいくるしいたすけてえ	Positive
2	頭痛が痛い昼寝しよ	Positive
3	頭痛が痛い（笑）よく言っちゃう www	Negative

2.2 見做し罹患ツイートの自動獲得手法

本研究では日常的な会話に着目し,「お大事に」という罹患者への定型的応答を利用して，罹患 positive と推測されるツイート（以降これを見做し罹患ツイートと称する）を自動的に取得する手法を提案する [5] [6]. 見做し罹患ツイートの自動獲得手法は以下の 3 ステップで構成されている.

> 見做し罹患ツイート
>
> 罹患者への定型的応答（「お大事に」を含むツイート）にリプライされているツイート

1. 「お大事に」をキーワードとしてツイートを検索
2. 該当のツイートのリプライ先のツイート ID を調べる
3. ツイート ID からツイート本文を取得

2.3 自動獲得学習データと人力作成学習データ

我々の提案法では，見做し罹患ツイートを正例，一般ツイートを負例と仮定してラベルを付けたデータセットをツイート罹患判定器の学習データ拡張に利用する（以降はこれを自動獲得学習データと称する）. ここで一般ツイートとは，一定期間無作為に収集したツイートを指す.

これに対して，人の手で作成されラベルを付与されたデータセットを人力作成学習データと称することとする．本研究では，前述した NTCIR-13 MedWeb タスクで配布された 1920 件のデータセットを本タスクに合わせて修正し，人力作成学習データとして利用する．前述したように MedWeb タスクは 8 疾患への罹患の有無をラベル付けするマルチラベル分類タスクである（インフルエンザ，下痢／腹痛，花粉症，咳／喉の痛み，頭痛，熱，鼻水／鼻づまり，風邪）. したがって，疾患に拘らない 2 クラス罹患判定タスクに転用するために，8 疾患のうち 1 つ以上の疾患への罹患が有るツイートを罹患 positive，どの疾患にも罹患していないツイートを罹患 negative としてラベルを修正した.

自動獲得学習データは僅かなコストで大量に用意することができるという利点があるが、クラスラベルの信頼性が高くないという欠点がある。一方で、人力作成学習データはクラスラベルの信頼性が高い、ツイート内容がタスク学習用に考慮して作られているといった利点があるが、病気への罹患を表す多様な表現を学習させるだけのデータを揃えるのは難しいという欠点がある。

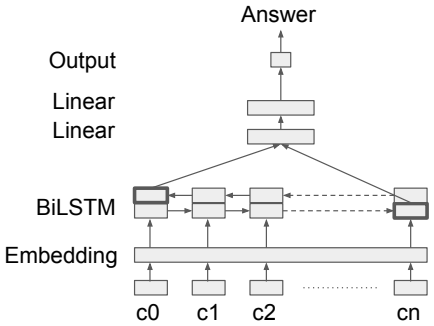

図 1: RNN based model

そこで，本研究では異なる性格を持つこれら 2 種類のデータセットを組み合わせて使用するアプローチを提案する.

2.4 RNN ベース罹患判定器

罹患判定器は，ツイート本文をモデル化する為の LSTM-BRNN 層と，罹患クラスを判定するための全結合層で構成される NN を用いた（図 1）.

まず，ツイートを文字の one-hot ベクトル $\mathbf{x}$ の集合で表現し，Embedding 層の入力とした．Embedding 層では埋込行列 $\mathbf{W_e}$ によって one-hot ベクトルを埋め込みベクトル $\mathbf{e(x)}$ に変換する.

$$\mathbf{e}(\mathbf{x_i}) = \mathbf{W_e} \cdot \mathbf{x_i} \tag{1}$$

その後，埋め込みベクトルは双方向 LSTM 層の入力となり，ツイートの文字系列のモデルが形成される.

$$\overrightarrow{\mathbf{h_{lstm}}} = \overrightarrow{LSTM}(\mathbf{e(x)});\ \overleftarrow{\mathbf{h_{lstm}}} = \overleftarrow{LSTM}(\mathbf{e(x)});\tag{2}$$

続いて，順方向と逆方向の LSTM 層のそれぞれの末尾の出力ベクトルを取り出して結合したベクトルを，2 層の全結合層に入れた．各全結合層の活性化関数には ReLU を適用した.

$$\mathbf{h_{blstm}} = concat[\overrightarrow{h_{lstm}}(n), \overleftarrow{h_{lstm}}(1)] \tag{3}$$

$$\mathbf{h_{l1}} = ReLU(\mathbf{W_{l1}} \cdot \mathbf{h_{blstm}} + \mathbf{b_{l1}}) \tag{4}$$

$$\mathbf{h_{l2}} = ReLU(\mathbf{W_{l2}} \cdot \mathbf{h_{l1}} + \mathbf{b_{l2}}) \tag{5}$$

最後に出力層によって次元数をクラス数と同じ 2 次元にまで落とし，Sigmoid 関数と argmax 関数を通して予測クラスを得た.

$$Answer = Sigmoid(\mathbf{W_o} \cdot \mathbf{h_{l2}} + \mathbf{b_o}) \tag{6}$$

2.5 二段階学習法

自動獲得学習データと人力作成学習データの間には，前述した利点欠点等のギャップが存在する．それらのギャップを埋めるため，Liu らの研究 [7] を参考に，二種類の学習データを効果的に利用するアプローチを提案する．二段階学習法では以下に示すように分類器の学習を二度行う．

1. 大量の自動獲得学習データを用いて基礎的なモデルを学習．
2. 1 で学習されたパラメータを初期値として，人力作成学習データを用いて罹患判定器を学習．

このアプローチでは，2 ステップ目にタスク専用に作られた人力作成学習データを用いることで，最終的なモデルがよりタスクに適したものになると考えられる．また 1 ステップ目に大量の自動獲得学習データを用いることで人力作成学習データの量不足による未知語問題を解決し，多様な罹患表現を補填する効果が期待できる．このように，ギャップもあるが共通の特徴を有する二種類の学習データを組み合わせて学習に用いることで，互いの欠点を補完し合うことが期待できる．

3 評価実験

前述の通り，二段階学習法は自動獲得学習データと人力作成学習データの二種類を効果的に組み合わせるためのアプローチである．これを評価するために，自動獲得学習データのみを学習したモデル $AUTOMATIC$ と，人力作成学習データのみを学習したモデル $ANNOTATED$，両データを合わせて学習したモデル MIX，二段階学習法で学習したモデル $2STEP$ の，3 つのモデルの分類精度（Accuracy 値，F 値）を比較する実験を行った．

3.1 データセット

$AUTOMATIC$ モデルの学習には 200,000 件の自動獲得学習データを，$ANNOTATED$ モデルの学習には 1,920 件の人力作成学習データを用いた．MIX モデルの学習には両データを組み合わせて用いた（201,920 件）．$2STEP$ の実験では $AUTOMATIC$ モデルのパラメータを初期値として $ANNOTATED$ モデルと同じ人力作成学習データ（1,920 件）を用いて学習を進めた．自動獲得学習データの収集期間は表 2 の通りである．

評価には NTCIR-13 MedWeb タスクで評価用に提供された 640 件の疑似ツイートからなるテストセット（以降は NTCIR テストデータと称する）と，それを参考に実際のツイートを使って作成した 639 件のテストセット（以降は実ツイートテストデータと称する）を利用した．実ツイートテストデータは，MedWeb タスクで定められた 8 疾患に関連するツイートを各 80 件ずつ手作業で収集，ガイドラインに従ってアノテーションして作成した．

表 2: 収集期間

	見做し罹患ツイート	一般ツイート
Tweet 数	180,107	329,610
keyword	お大事に	-filter:links lang:ja
収集期間	2017 年 5 月 1 日 ～2017 年 5 月 11 日 2017 年 8 月 19 日 ～2017 年 9 月 3 日	2017 年 5 月 16 日 ～2017 年 5 月 18 日

3.2 学習条件

RNN ベース罹患判定器は Pytorch を用いて実装した．埋め込み層と BLSTM 層の次元数は等しく，128 次元 と 256 次元 の 2 種類を用意した．最適化には確率的勾配降下法 (SGD) を用いた．損失関数として交差エントロピー誤差関数を用い，重みを課すことで学習データの不均衡に対応した．学習率の初期値は 0.005，ミニバッチサイズは 10，Dropout 率は 0.2 とした．学習は，学習データのサイズを考慮して $AUTOMATIC$ モデルと MIX モデルは 100epoch まで，$ANNOTATED$ モデルは 15,000epoch まで，$2STEP$ モデルは 1 ステップ目を 100epoch，2 ステップ目は 15,000epoch まで学習を進めるようにと予め設定した．

3.3 実験結果

表 3 に実験結果をまとめる．size の列は埋め込みベクトルのサイズを表している．分類精度の評価には，正解率を示す Accuracy と，各クラスの分類精度を示すクラスごとの F 値を用いた．$MAJORITY$ の行は，全てを罹患 positive なクラスと判定した場合の Accuracy 値を示してある．

3.4 考察と結論

全モデルの分類精度を比較すると Accuracy，F 値共に $2STEP$ が最も高いことが分かる．この結果から，自動獲得学習データや人力作成学習データを単体で又は単純に合併して学習するよりも，提案する二段階学習法を用いた方が精度良く分類できると言える．

[1] 全てを罹患 negative クラスと判定する分類器となった

表 3: 2 ステップの学習法の分類精度

モデル	size	NTCIR テスト			実テスト		
		Accuracy	F 値		Accuracy	F 値	
			罹患 positive	罹患 negative		罹患 positive	罹患 negative
MAJORITY	-	0.695	-	-	0.559	-	-
AUTOMATIC	128	0.710	0.820	0.238	0.635	0.747	0.347
AUTOMATIC	256	0.697	0.815	0.164	0.517	0.737	0.290
ANNOTATED	128	0.842	0.889	0.726	0.632	0.687	0.554
ANNOTATED	256	0.856	0.900	0.742	0.578	0.670	0.413
MIX[1]	128	(0.305)	(0.000)	(0.467)	(0.441)	(0.000)	(0.612)
MIX	256	0.795	0.868	0.547	0.654	0.756	0.404
2STEP	128	0.856	0.899	0.753	0.664	0.742	0.517
2STEP	256	**0.878**	**0.913**	**0.795**	**0.700**	**0.773**	**0.558**

表 4: 二種類の学習データから獲得された語彙

	人力作成	自動獲得
ツイート数	1,920	200,000
文字タイプ数	959	6,691
総文字数	43,253	9,913,582
未知文字率（NTCIR テスト）	0.006	0.000
未知文字率（実テスト）	0.186	0.000

NTCIR テストデータの分類精度を見ると，$2STEP$ に続いて $ANNOTATED$ の分類精度が高く，$AUTOMATIC$ が最も悪い精度となっている．この結果は，人力作成学習データが MedWeb タスクの為に作成されている事と NTCIR テスト同様に擬似ツイートである事から，人力作成学習データを用いたモデルの方が NTSIR テストデータをより良く分類できるという我々の予想に沿った結果となったと言える．

全体を通して NTCIR テストの分類精度よりも実テストの分類精度の方が低いが，その下がり具合は $ANNOTATED$ が最も大きいことが分かる．このような結果となった一番の原因は自動獲得学習データから得られる語彙に比べて，人力作成学習データから得られる語彙の数が少なく実テストデータの語彙をカバーできていないことだと考えられる．表 4 に各テストデータの未知文字率を示す．人力作成学習データから得た語彙では NTCIR テストデータの未知文字率が 0.6%であるのに比べて実テストデータの未知文字率は 18.6%と非常に大きい．この結果から実際のツイートに含まれる文字が多様であると分かる．他方で，自動獲得学習データから得た語彙では両テストデータの未知文字率は 0.0%である．この結果から語彙の不足による分類精度への悪影響が確認でき，学習に大量の自動獲得学習データを用いて幅広い語彙を獲得することの利点を確かめられた．

4 おわりに

本稿では，ツイート罹患判定器の学習に十分な量のコーパスを人手で用意することが非常に困難であるという問題点の解決策として，見做し罹患ツイートの自動獲得手法を提案するとともに，人手でラベル付けされた少量の学習データの拡張にそれを利用するための効果的な方法を提案した．評価実験を行い，提案する二段階学習法による分類精度の向上を確認した．この結果から，自動獲得された大量の学習データからは実際のツイートにおける豊富な語彙と罹患表現を，人手で作成・アノテーションされた少量の学習データからは罹患判定タスクによりマッチした分類境界を，二段階学習法を用いることで二種類の学習データの利点を効果的に組み合わせて学習できることを確認した．

謝辞

本研究は JSPS 科研費 16K00153 の助成を受けた．

参考文献

[1] Eiji Aramaki, et al. Twitter catches the flu: Detecting influenza epidemics using twitter. *Proceedings of the Conference on EMNLP*, pp. 1568–1576, 2011.

[2] 松田紘伸ほか. Twitter を用いた病気の事実性解析及び知識ベース構築. 人工知能学会全国大会論文集, Vol. JSAI2016, pp. 2C5OS21b4–2C5OS21b4, 2016.

[3] Michael J. Paul, et al. Worldwide influenza surveillance through twitter. In *AAAI Workshop: WWW and Public Health Intelligence*, 2015.

[4] Eiji Aramaki, et al. Overview of the ntcir-13: Medweb task. *Proceeding of the NTCIR-13 Conference*, 2017.

[5] Reine Asakawa and Tomoyoshi Akiba. Akbl at the ntcir-13 medweb task. *Proceedings of the NTCIR-13 Conference*, 2017.

[6] 浅川玲音, 秋葉友良. 疾病サーベイランスのための罹患ツイートの自動獲得と事実性判定への利用. *NLP2018*, 2018.

[7] Ting Liu, et al. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. *Proceedings of the 55th Annual Meeting of ACL*, Vol. 1, pp. 102–111, 2017.