擬似罹患ラベル付きデータを用いたツイートのマルチラベル疾病分類

奥田 智大　　秋葉 友良
豊橋技術科学大学 情報・知能工学科
{okuda.tomohiro.oj, akiba.tomoyoshi.tk}@tut.jp

1 はじめに

近年ソーシャルメディアの発展に伴い個人の情報発信が増加している．それにより個人が病気に関する情報を気軽に発信するようになった．なかでも Twitter のツイートを用いたインフルエンザの流行予測が，リアルタイム性に優れいている点などから注目を集めている．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によって報告されている.[1]

このような状況に関係して医療情報処理シェアドタスク NTCIR-13 MedWeb task[2] が実施された．MedWeb は，ツイートに対し 8 疾病 (インフルエンザ, 下痢, 花粉症, 咳・たん, 頭痛, 熱, 鼻水・鼻づまり, 風邪) に罹患しているか否かをマルチラベル分類するタスクである．以下に例を挙げる．

(1) なんか今日は熱出てるし頭も痛い

(2) きつい上司、頭痛の種

これらをマルチラベル分類した結果を表 1 に示す。

表 1: 疾病分類結果

	(1)	(2)
インフルエンザ	n	n
下痢	n	n
花粉症	n	n
咳・たん	n	n
頭痛	p	n
熱	p	n
鼻水・鼻づまり	n	n
風邪	n	n

(1) は頭痛と熱罹患していることが読み取れ, 表 1 にも該当する箇所にポジティブラベルがついていることが分かる．(2) は文章内に頭痛と出てくるもののこれは慣用句としての表現であり, 実際には頭痛に罹患しているわけではないためすべての罹患にネガティブラベルがついている．

ただし NTCIR-13 で配布されたデータは学習用とテスト用共に疑似ツイートであり, 現実のツイートではない．そのため絵文字が一切使われておらず, また文章も現実のツイートと比べて表現が砕けていないという違いがある．また配布されたデータ量も合計で 1280 件と十分な量とは言えない．

そこで本研究では現実のツイートに対し疾病分類することを目標とする．そのためには分類器の学習データも現実のツイートであることが望ましい．しかし現実のツイートにはラベルがついておらずそのままでは学習に使用できない．人手でラベルを付けるのも非常にコストがかかる．よって機械的な方法でつけることができる疑似罹患ラベルの考案を行う．

2 データセット

本研究では NTCIR データと見做し罹患ツイートを使用する．本章では各データセットの内容について記述する．

2.1 NTCIR データセット

NTCIR データセットは”NTCIR-13 MedWeb task”で配布された疾病分類のデータセットである．ツイートの文章データに対しインフルエンザ, 下痢, 花粉症, 咳・たん, 頭痛, 熱, 鼻水・鼻づまり, 風邪の 8 疾患への罹患がポジティブかネガティヴなのかラベルがついている．NTCIR データセットの一例を図 1 に示す．

図 1 の 1 は本人が風邪にかかっていると明言はされていないが「だるくなる」と症状に関する内容が記されていることから現時点で風邪をひいいていると推測

1. 風邪を引くと全身がだるくなる。
2. あかん。咳込みすぎて頭まで痛くなってきた
3. この契約が上手くいかないのは頭痛の種でしかない。

図 1: NTCIR データセット

1. インフルになっちゃったあ(T_T)
辛すぎるッッッ😭
2. うわぁ〜風邪ひいたっぽい😭
今日中に治さないと😠
はぁ😞💨連休なのに…🌀引きこもろう🏠

図 2: 見做し罹患ツイート

できるため風邪にポジティブのラベルがついている. 2 は咳が出て頭が痛いというところから明らかだが, 咳・たんと頭痛の 2 疾病で罹患ポジティブのラベルがついている. 3 は頭痛の種という表現を使っているが実際に病気として頭痛があるわけではないため全ての疾病で罹患ネガティヴのラベルがついている.

このデータセットは学習用とテスト用にそれぞれ 640 件ずつ用意されており, 本研究では学習用を使用した.

2.2 見做し罹患ツイート

見做し罹患ツイートとは, 浅川ら [6] が考案した「お大事に」という文字列を含んだリプライをもらったツイートのことである.

実際のツイートから獲得した見做し罹患ツイートの例を図 2 に示す.

この場合図 2.1 はインフルエンザに, 図 2.2 は風邪に罹患していることが読み取れる. また絵文字や顔文字が使用されていることも分かる. 見做し罹患ツイートは 2018 年 12 月に収集された 180,107 件を使用する。

見做し罹患ツイートは，ある疾病に罹患しているツイートを高い確率で自動的に収集したものである。しかし，どの疾病に罹患しているかは分からず，8 疾病のマルチラベルが与えられていないため，そのままマルチラベル分類の学習データとしては利用できない.

表 2: 擬似疾病ラベル用単語

疾病	単語
インフルエンザ	インフル
下痢	下痢, 腹, おなか
花粉症	花粉
咳・たん	咳, たん, 痰
頭痛	頭, あたま
熱	熱
鼻水・鼻づまり	鼻水, 鼻づまり, 花粉
風邪	風邪

2.3 擬似疾病ラベル

見做し罹患ツイートを学習に使用するために，擬似的な疾病ラベルを付与することを考える. 本研究では, 次の 2 つの手法を試みた.

手法 T 比較的単純な手法として，ツイートに 8 疾病に関する単語が含まれている場合にその疾病に対してポジティブラベルを付与することにした. 疾病毎に利用した単語を表 2 に示す. これらは, NTCIR の判定基準をもとに決定した.

手法 C NTCIR のラベル付きデータでマルチラベル分類器を学習しておき, 見做し罹患ツイートをマルチラベル分類する. この分類結果に従って, 疑似ラベル付けを行った.

これらの手法で付けたラベルのことを擬似疾病ラベルと呼ぶ.

3 実験

3.1 マルチラベル疾病分類器

本研究では，畳み込みニューラルネットワークを用いたマルチラベル疾病分類器を実装した. 分類器の構成を図 3 に示す. このモデルは kim らが提案したモデル [3] を参考にし, マルチラベル用に拡張したものである. 本研究の対象データであるツイートデータは砕けた表現やネットスラング, 方言の処理などの問題により形態素解析が難しいことが知られている.[4] そこで本研究では文章を文字単位で分割を行うものとする. 日本語において, 文字ごとに分割し畳み込みを行った場合でも高い性能を発揮できることは佐藤ら [5] によって報告されている.

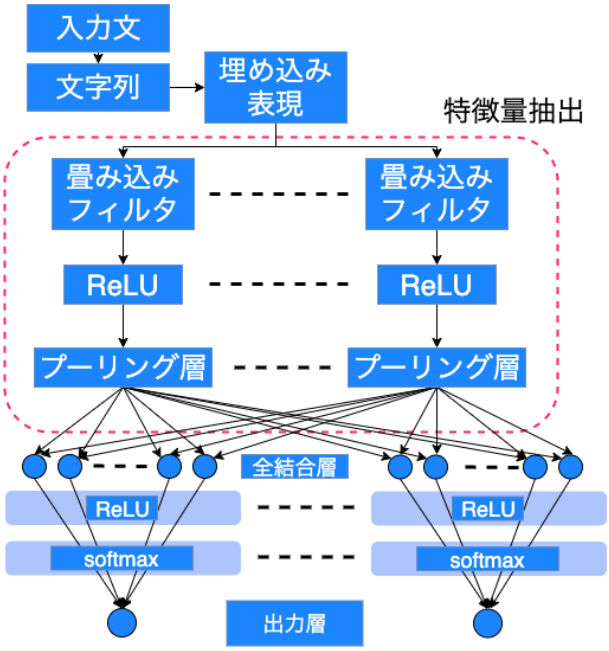

図 3: 本実験システムの構成

1. 2週間位咳が止まらんのやけどそろそろ死ぬ？
2. "インフルじゃなかった〜✌✌
でも頭の痛さがレベル違う😂笑
心配してお母さんが駆けつけてきた、笑
3. インフルの予防注射打ちに行くがてらにお散歩のんびりしてこよー

図 4: 罹患ツイートテストセット

まず入力文章を文字単位に分割し 140 次元の埋め込み表現に変換する．次にカーネルサイズ 3 のフィルター 128 個ので畳み込みを行い各フィルターごとに 138 次元の特徴量を得る．これをマックスプーリングすることで 128 次元に圧縮する．その後全結合層に与え各疾病ごとに罹患分類を行う．

3.2 テストセット

分類器の性能を確かめるためのテストセットとして疾病名を文章内に含む現実のツイート 639 件集めた．この一例を図 4 に示す．

これらのツイートに対し一人のアノテーターが NTCIR データの基準を用いて 8 疾患ごとに罹患ポジティブかネガティヴかラベル付けした．図 4 の 1 は文章内に明確に咳と記されているため咳・たんに罹患していることが分かる．一方，2 は「インフル」と文章内に出てくるが「インフルじゃなかった」という箇所からインフルエンザは罹患ネガティヴであると分かる．しかし「頭の痛さがレベル違う」という点から頭痛の罹患はポジティブとなる．3 はインフルエンザの予防接種に行くというだけの内容であるため 8 疾患全てで罹患ネガティヴとなる．

3.3 比較手法

分類器の学習に用いるラベル付きデータの違いによる分類性能を比較する．学習に用いるマルチラベル付きデータは以下の 4 種類である．

NTCIR 疑似ツイートに対する人手ラベル付けデータ.640 件.

RealTweet-T 見做し罹患ツイートに対して, 手法 T(疾病を表す単語の有無でラベル付けする方法) を用いて自動的に疑似疾病ラベルを付与した実際のツイート.180,107 件.

RealTweet-C 見做し罹患ツイートに対して, 手法 C(予め学習しておいた分類器でラベル付けする方法) を用いて自動的に疑似疾病ラベルを付与した実際のツイート.180,107 件.

Supervised 実際のツイートに対する人手ラベル付けデータ. テストデータ 639 件と同一. このデータを使用するときは,10-fold クロスバリデーションで評価を行い, 評価データに学習データが含まれないようにした.

これらの学習データを用いる, 以下の手法を比較した.

NTCIR NTCIR のみを使用して学習.

RealTweet-T RealTweet-T のみを使用して学習.

RealTweet-C RealTweet-C のみを使用して学習.

Supervised Supervised のみを使用して学習.

RealTweet-T+NTCIR RealTweet-T と NTCIR を混合して学習.

RealTweet-T → NTCIR RealTweet で学習したモデルを NTCIR でファインチューニングして学習. ファインチューニング時のエポック数は 10 とした.

RealTweet-T → Supervised RealTweet で学習したモデルを SupervisedTweet でファインチューニングして学習. ファインチューニング時のエポック数は 10 とした.

表 3: 実験結果

手法	F1-micro	F1-macro
NTCIR	0.695	0.637
RealTweet-T	**0.714**	**0.672**
RealTweet-C	0.653	0.587
Supervised	0.709	0.646
RealTweet-T+NTCIR	0.672	0.595
RealTweet-T → NTCIR	0.620	0.551
RealTweet-T → Supervised	**0.716**	**0.673**
RealTweet-ORACLE	*0.866*	*0.844*

RealTweet-ORACLE テストデータを全て使って分類器を作り, それを用いて見做し罹患ツイートを分類した結果を擬似疾病ラベルとして使用. このモデルは, 学習にテストデータを使っているため現実的ではないが, 見做し罹患ツイートのラベル付け精度を改善した場合の性能の上限を調べるために調査した.

3.4 実験結果

実験結果を表 3 に示す.

まず, 単独の学習データを用いた場合,RealTweet-T が最も良い性能を示した. 大量の実際のツイートに対して疑似疾病ラベルを付与することによって, 少量の人手ラベル付きデータを使う場合 (Supervised) よりも性能を改善することができた. ラベル付け手法は, 分類器を使う手法 (RealTweet-C) よりも, 疾病に関連する単語の有無だけを用いる単純な手法 (RealTweet-T) の方が効果があった. マルチラベル付けを行うツイートを見做し罹患ツイート (「お大事に」でリプライされているツイート) に制限することにより, 精度よく罹患に関係するツイートだけを対象にできたことで, 比較的単純な手法でも正しくラベル付けできたと考えられる. 分類器を用いる手法は, ベースとなる分類器 (NTCIR) の性能が低いため, あまりうまく機能しなかったと考えられる.

また,RealTweet-T → Supervised の結果より,RealTweet-T に対し教師あり学習データ (Supervised) でファインチューニングする手法により, さらに性能を改善することを確認した.

4 おわりに

本研究では見做し罹患ツイートに対し疑似罹患ラベルを作成することで疾病分類精度の向上を図った. 結果として, 疑似罹患ラベルの性能が高いものであることを確認できた. また, 人手でつけたデータでファインチューニングを行うことでより分類精度を上げられることが分かった.

今後はより高い精度で分類できる疑似罹患ラベルを考案し, モデルのハイパーパラメータやファインチューニングする際のエポック数を最適化したい.

謝辞 本研究は JSPS 科研費 19K11980 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. 15681576, 2011.

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

[3] Yoon Kim, New York University, Convolutional Neural Networks for Sentence Classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp.1746-1751, 2014.

[4] 大崎彩葉, 唐口翔平, 大迫拓矢, 佐々木俊哉, 北川善彬, 堺澤勇也, 小町守, 首都大学東京, Twitter 日本語形態素解析のためのコーパス構築, 言語処理学会 第 22 回年次大会 発表論文集, pp.1-6, 2016.

[5] 佐藤 拳斗, 折原 良平, 清 雄一, 田原 康之, 大須賀昭彦, 電気通信大学大学院情報システム学研究科, 文字レベル深層学習によるテキスト分類と転移学習, 人工知能基本問題研究会 102, pp.13-19, 2016.

[6] 浅川玲音, 秋葉友良, 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, 言語処理学会 第 25 回年次大会 発表論文集, pp.5-36, 2019.