インフルエンザ流行検出のための事実性解析

北川 善彬 * [1]　小町 守 [1]　荒牧英治 [2,4]　岡崎直観 [3,4]　石川 博 [1]

[1] 首都大学東京　[2] 東北大学　[3] 京都大学　[4] 科学技術振興機構さきがけ

1　はじめに

現在，Twitter などの SNS の普及により，情報の発信が容易になってきた．それに伴って，これまでにない大量の情報をリアルタイムに SNS から抽出する技術が注目されている．インフルエンザの流行情報の検出のためには，実際にインフルエンザに感染している人がどの程度いるのかを判断する必要がある．しかし，機械的に「インフルエンザ」を含む発言を集めるだけでは感染している人がどの程度存在するかは分からない．このために，集めた発言を感染者に関する発言かそうでないかの分類を行うことにより流行情報の検出を行う．

このような分類のためには，文に記述されている事象が，実際に起こったことなのか，そうでないことなのかの事実性を判定する技術が必要となる．これは，事実性解析と呼ばれる．

事実性解析が必要な例は以下のような例である．

(1) 熱があったので、病院に行ったら インフルエンザ **だった**。

(2) インフルエンザ に罹った **かもしれない**。

(3) インフルエンザ に罹っていた **ら**、休まざるを得ないだろう。

これらの中で，インフルエンザであることを「だった」として断定したり，「かもしれない」と推量をしたり，「たら」と仮定をしたりしていることが分かる．これにより，(1) は，「インフルエンザに感染する」という事実を持っており，反対に (2)，(3) はこの事実を持たないと判断できる．このような表現はモダリティと呼ばれ，人間が情報の真偽を考える上で重要な手がかりになる．

本研究では，インフルエンザ流行検出のために，話し手の判断や感じ方を表す言語表現であるモダリティを利用した手法を検討する．実験の結果，インフルエンザ流行検出に対する事実性解析においてモダリティを利用することで 3.5 ポイントの精度向上がみられ，提案手法の有効性を示した．

*kitagawa-yoshiaki@ed.tmu.ac.jp

2　関連研究

風邪，インフルエンザなどの疾患情報の取得を目的としたこのような分類タスクは先行研究 [3, 4] 等がある．ここで問題になるのはその疾患に実際にかかったのか，単なる言及なのを分類することであり，文献 [3] では，SNS 上で言及される「風邪」という単語を含んだ表現の半数は，単に疾患について言及したのみであり，疾患を罹患している事実はないと報告している．すなわち，単に疾患名で検索して頻度を数えるだけでは 50%近いノイズが含まれることになる．そこで，彼らは風邪とその諸症状に対して，命題に対してその事実性を判断する命題識別の分類器とモダリティによって事実性を判断するモダリティ分類器とを 2 つに分け，両方の分類器が正例と判断した場合のみ正例と判断するという手法をとった．この結果，一部の症状においては精度の向上が見られたものの，全体としては，モダリティは精度に貢献しないとされた．

先行研究 [3] では，モダリティに関しての事例を集めたコーパスを作成することでモダリティ情報を利用しているが，本研究では，既存のリソースからモダリティに関しての素性を作成することで行っているため，コーパス作成の手間を省く事を可能にした．

日本語においては，モダリティを用いて事象の事実性を判断するために，文献 [5] が態度表明，時制，仮想，態度，真偽判断，価値判断，焦点などについて詳細に事象アノテーションを行っている．焦点を除いた 6 種の項目を拡張モダリティと呼び，情報抽出や含意認識といった自然言語処理のタスクへの応用に向けて研究が行われている．

また，このような研究は日本語だけでなく英語に関しても活発であり，文献 [2] がモダリティを用いて，事実性の度合いを判断する研究を行っている．

表 1: 発言（事例）とラベル付けの例

ラベル	発言（事例）
正例	やっぱり インフルエンザ だったか…こりゃ家族内で蔓延しそうだな…
負例	まぁ俺 インフルエンザ のワクチンとか打ったことなんですけどね

Web 応用のタスクでは，文献 [1] がモダリティに関しての素性を利用している．特にモダリティの一部である否定 (Negation) や疑い (Suspicion) については，専門のワークショップ [NeSp-NLP 2010] が開催されるなど盛んに研究されてきた．

本研究では，これら過去のモダリティに関するリソースを活用し，事実性判定の精度を向上することを目指す．

3 モダリティを利用した事実性解析

3.1 データとタスク設定

本研究では「インフルエンザ」を含む 10443 件の発言に対してアノテーションされたデータを使用する．アノテーションの基準については，文献 [3] の風邪疾患についてのアノテーションに準拠している．このデータでは発言をした本人，または，その周りの人物がインフルエンザに感染していると判断される発言に関しては正例，感染していないと判断される発言に対しては負例とラベル付けがされている．具体例を表 1 に示す．今回のデータでは正例数が 1319，負例数が 9124 となっている．また，発言それぞれに時間情報があり，発言のあった年月日が記録されている．[1]

3.2 ベースライン

本研究のような風邪等の疾患情報を検出するために発言の分類を行うタスクは先行研究 [3, 4] があり，分類のためには，注目している単語の周辺の単語が良い素性となることが知られている．ここでは，形態素解析により，分かち書きを行い，形態素を 1 つの単位としたウィンドウを作成した．インフルエンザが分かち書きされてしまうケースなどは，インフルエンザを含むウィンドウが出来るように形態素の連結を行い新たなウィンドウを作成した．「インフルエンザ」を含むウィンドウを中心として，左側の 3 つの形態素と右側の 3 つの形態素を Bag of Words（BoW）の素性とし，これにモダリティに関しての素性以外を加えたものをベースラインの分類器を作成した．素性の詳細については 4 節で説明する．

表 2: つつじによる意味 ID 素性の例

発言例	つつじの意味 ID
インフルエンザ **ですか**…	で→ r32 です→ D41 か→ Q31 し→ n13
びっくり **し** ま **し** た。	

3.3 つつじによる素性

1 つ目の手法としてつつじ [2] の利用を試みた．日本語の文を構成する要素には，主に内容的な意味を表す要素（内容語）以外に，助詞や助動詞といった，主に文の構成に関わる要素がある．ここでは，後者を総称して,「機能語」と呼び,「に対して」や「なければならない」のように，複数の語から構成され，かつ，全体として機能語のように働く表現である「複合辞」と合わせてこれらを機能表現と呼ぶ [6]．つつじは 16801 の機能表現の表層形を階層的に分類しており，同じような意味を持つ機能表現には同じ意味 ID が振られている．

本タスクは Twitter のデータを使用しており，発言の中には文が複数ある場合も多い．これにより，注目しているインフルエンザ感染に関連する機能表現と関係のない機能表現も多く存在すると考えられる．そこで,「インフルエンザ」の右の 15 文字中につつじの機能表現の表層形が含まれる場合にその意味 ID を素性として利用することにした．形態素解析を行うと機能表現の表層形を分かち書きしてしまうため，文字列での処理を行った．つつじによる意味 ID 素性の具体例を表 2 に示す．

3.4 Zunda による素性

つぎに，2 つ目のモダリティの利用法として，Zunda[3] の解析結果を利用する手法を提案する．Zunda は文中のイベント（動詞や形容詞，事態性名詞など）に対して，その真偽判断（イベントが起こったかどうか），仮想性（仮定の話かどうか）などを解析することのできる日本語拡張モダリティ解析器である．本手法におい

[1] このデータは実際のアプリケーションに利用されており,「インフル君」 http://mednlp.jp/influ/ の中で使用されているものである．

[2] つつじ 日本語機能表現辞書 http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

[3] Zunda 拡張モダリティ解析器 https://code.google.com/p/zunda/

表 3: Zunda による素性の例

発言例	Zunda による素性
左隣の患者さんが インフルエンザ **発覚**	発覚=成立

ては，Zunda の出力する真偽判断のタグを利用した．真偽判断についてのラベルとしては,「成立」,「不成立」,「不成立から成立」,「成立から不成立」,「高確率」,「低確率」,「低確率から高確率」,「高確率から低確率」,「0」のラベルが存在する．

これらのラベルが各イベントに対してついているが，つつじを使用した場合と同様にインフルエンザに関連するイベントがどこに存在するかを考えなければならない．我々は Zunda が動詞，事態性名詞を「イベント」として解析していることから,「インフルエンザ」の右に続く動詞，事態性名詞で一番近いものをインフルエンザに関連するイベントとみなし，そのイベントとイベントに付けられたラベルの組み合わせを素性として利用した．具体例を表 3 に示す．

4 インフルエンザ感染か否かの 2 値分類の実験・評価

本タスクは，発言をした人物，あるいはその周りの人物がインフルエンザにかかっているか否かを分類する 2 値分類タスクであり，L2 正則化ロジスティック回帰を用いて分類を行った．評価は 5 分割交差検定による適合率，再現率， F 値で行った．ツールとしては，Classias (ver.1.1) [4] を使用した．

ウィンドウを決めるための形態素解析器としては MeCab (ver.0.996) [5] を利用し，MeCab の辞書は IPA-Dic (ver.2.7.0) を用いた．

ベースラインに使用した素性を 4 に示す．つつじによる素性と Zunda による素性に関しては 3 節の説明によるものとする．

インフルエンザ感染に関しての 2 値分類を行った結果を表 5 に示す．

つつじに関しての素性と，Zunda に関しての素性を両方用いた All の場合において最高精度となった．この結果は，モダリティを抜いた baseline より，3.5 ポイントの F 値の向上が見られるので，モダリティに関する素性が有用であることを支持する．

表 4: ベースラインに使用した素性とその説明

window6BoW:「インフルエンザ」を含むウィンドウを中心とする左側 3 つの形態素，右側 3 つの形態素の Bag of Words の素性
URL: 発言における URL の有無の素性
Atmark: 会話等によるリプライの有無の素性
N-gram: インフルエンザ」の前後の文字 N-gram の素性．前後の文字 1-gram から 4-gram の素性
Season: 12 月から 2 月にかけての発言なのかそうでないかの素性

表 5: インフルエンザ感染に関しての 2 値分類の結果

素性の組み合わせ	適合率	再現率	F 値
window6BoW	0.740	0.305	0.432
window6BoW+URL	0.699	0.313	0.432
window6BoW+Atmark	0.740	0.305	0.432
window6BoW+N-gram	0.707	0.345	0.464
window6BoW+Season	0.724	0.333	0.456
window6BoW+tsutsuji	0.764	0.321	0.452
window6BoW+Zunda	0.699	0.313	0.432
baseline	0.697	0.392	0.502
baseline+tsutsuji	0.702	0.420	0.526
baseline+Zunda	0.679	0.412	0.513
All	0.689	0.440	**0.537**

5 考察

本論文では，モダリティに関しての素性の貢献を見ることができたが，実際にどのような事例に対して貢献が見られたか，また，うまくいかなくなった事例はどのようなものかについて調査する．表 6 に実際の発言例を示す．

5.1 つつじによる素性

つつじにおける素性について，分類器の判断に大きく影響を与える素性を調べた．その結果を表 7 に示す．

表 6 の発言例 1 において，表 7 における,「らしい」のような推量のモダリティ表現により，正しい出力を得るようになった．

逆に，発言例 2 においては，ひらがな 1 文字のものが多くマッチしてしまい「と」や「え」などはつつじの意味 ID「Q11」,「I23」に該当し分類器はこれらに負の重みをつけており，分類に失敗することとなった．「I23」の意味 ID をもつひらがな「え」は表 7 に示している「だろう」に該当するものであり，ひらがな一文字の場合は前後の文脈を見ないとうまく素性として機能しないことが分かる．

[4]Classias http://www.chokkan.org/software/classias/index.html.ja
[5]Mecab 日本語形態素解析器
http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

表 6: 分類に成功した例と失敗した例

発言例 1	@*** 強力な インフルエンザ **らしく** てですね、まだまだ完治しておりませぬ…ひい
発言例 2	まさかの インフルエンザ ... 全身鳥肌 **と** 震 **え** 半端ない寒気が... タミフルが効いて楽になったと思い熱...
発言例 3	10 年ぶりに インフルエンザ というものに **かかり** ました wwww
発言例 4	ASPARAGUS 渡邊忍が インフルエンザ に **感染** してしまい、本日の柏 DOMe での...

表 7: 重みの絶対値の大きい意味 ID 素性とその表層形の例（左が正，右が負の素性）

ID	重み	表層形の例	ID	重み	表層形の例
D11	0.55	ないと駄目	z25	0.88	ではどう
l31	0.54	あまりに	k51	0.69	には
r12	0.47	事にしてる	h11	0.68	について
s23	0.41	おかげで	R11	0.55	みたい
G21	0.37	ことにする	I23	0.46	だろう
i11	0.36	といった	A31	0.44	そう
I12	0.34	らしい	I11	0.43	かも

表 8: Zunda による重みの絶対値の大きい素性

正の素性	重み	負の素性	重み
罹患=成立	0.80	注射=成立	0.62
かかり=成立	0.65	対策=成立	0.50
診断=成立	0.52	かかり=0	0.48
寝=成立	0.47	なる=成立	0.45
診断=成立	0.52	する=成立	0.45
発覚=成立	0.47	死亡=成立	0.42
回復=成立	0.44	行っ=成立	0.39
ダウン=成立	0.40	注意=成立	0.38
うつっ=成立	0.39	感染=不成立	0.37
潜伏=成立	0.37	なっ=不成立	0.34

5.2 Zunda による素性

次に，Zunda による素性について，分類器の判断に大きく影響を与える素性を調べた．つつじの場合と同様に，重みの大きな素性を大きい順に並べた結果を表 8 に示す．

表 8 を見るとつつじの素性に比べて直感的に理解できるものが多い．インフルエンザの発言では，注意を呼びかける発言，予防接種の内容の発言，ニュースに関する発言等が多く，負の重みによくそれが現れている．正の重みに関しては直接疾患に関係のある名詞や動詞が多くなっている．

発言例 3 においては，「かかり＝成立」の素性により，判別できるようになった．Zunda においてはこのようにクリティカルに素性がうまく働いている例が多く見られた．

発言例 4 においては，実際に感染しているのは発言をしている本人でもなく周りの人でもないため，ここでは，負例を正解とするのが正しいが，「感染=成立」という素性のために正例になってしまっている．このことから，状態を所有する主体を捉えることが重要であることが分かる．

6 おわりに

本研究では，インフルエンザの流行検出のため，モダリティの素性を組み込む手法を提案し，これが，Web 応用のタスクの精度の向上に貢献することを示した．

今後の課題としては，風邪，頭痛等の他の感染症に適用できるかを検証すること，インフルエンザに関係のあるモダリティ表現を係り受けを捉えて判断すること等が挙げられる．本タスクにおいては，「インフルエンザに感染する」ということに対しての事実性を捉えることが目標であり，それに関連する情報のある場所を判断することが難しい問題であった．また，事実性の問題だけでなく，表 6 の発言例 4 で挙げたような症状を所有する主体を正しく捉えることも重要になる．

参考文献

[1] Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. Major life event extraction from Twitter based on congratulations/condolences speech acts. In *EMNLP*, pp. 1997–2007, 2014.

[2] Roser Saurí and James Pustejovsky. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, Vol. 38, No. 2, pp. 261–299, 2012.

[3] 荒牧英治, 森田瑞樹, 篠原 (山田) 恵美子, 岡瑞起. ウェブからの疾病情報の大規模かつ即時的な抽出手法. 言語処理学会第 17 回年次大会発表論文集, pp. 838–841, 2011.

[4] 荒牧英治, 増川佐知子, 森田瑞樹. 文章分類と疾患モデルの融合によるソーシャルメディアからの感染症把握. 自然言語処理, Vol. 19, No. 5, pp. 419–435, 2012.

[5] 松吉俊, 江口萌, 佐尾ちとせ, 村上浩司, 乾健太郎, 松本裕治. テキスト情報分析のための判断情報アノテーション. 電子情報通信学会論文誌 D, Vol. 93, No. 6, pp. 705–713, 2010.

[6] 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂. 自然言語処理, Vol. 14, No. 5, pp. 123–146, 2007.