1J3-1

ツイート中の地理情報に対する時間的極性の自動推定

Automatic Estimation of Temporal Awareness of Location Reference Expression in Tweets

珊瑚 彩主紀 *1 Mizuki Sango　松田 耕史 *2 Koji Matsuda　岡崎 直観 *3 Naoaki Okazaki　乾 健太郎 *4 Kentaro Inui

*1東北大学工学部 School of Engineering, Tohoku University

*2東北大学大学院情報科学研究科 Graduate School of Information Sciences, Tohoku University

We propose a novel model for automatically estimating Temporal Awareness of Location Reference Expressions in short text messages such as tweet. We modeled this task as target-dependent text classification problem, based on a Convolutional Neural Network (CNN) model. In our experiments, we investigated several target-dependent features for temporal awareness classification with crowd-sourced data. With target-dependent features, our model achieved promising accuracy compared to the baseline model.

1. はじめに

近年，モバイル端末を通じて，屋外や外出先でその場の情報をリアルタイムに SNS に投稿するユーザーが増えている．特定の場所について述べているツイートを抽出することで，観光産業への応用や，災害時の対応など，さまざまなアプリケーションが広がる．

しかしながら，地名や場所名を含んでいるツイートを単純に抽出するだけでは，その場の状況を適切に要約するには不十分であり，リアルタイムな情報を取り出すためには，著者がその場所に対してどのように認識しているかについて，抽象化を行った分析が不可欠である．

我々は，リアルタイムな情報や，予測情報を効率的に収集し，分析に活かすための分析軸の一つとして，**場所参照表現に対する著者の主観的な時間極性** ([Li 14] においては，*Temporal Awareness* と呼んでいる．本稿では *TA* と略記する) をとりあげる (図 1)．ツイートの著者が，現在その場所で見ている情報を述べているのか，そうでないのか，(さらには，その場所に実際に行ったのか，それともこれから行く予定なのか) を把握することができれば，その場所に対するより信頼性の高い情報を集めることが可能になる．我々のモチベーションを説明するための例として，以下のようなツイートを取り上げる．

(1) 昨日は 仙台 に行って来たんだけど，今日は 名古屋 にいます．明日は 横浜 だ〜

上記の例においては，仙台，名古屋，横浜という二つの場所への言及が行われているが，それぞれの場所への著者の時間認識は異なる．仙台は「過去にいた」場所，名古屋は「現在いる」場所，横浜は「これから行く予定のある」場所である．観光情報マイニングや，マーケティングへの応用を考えると，これらの言及を区別することは非常に重要である．これらを区別して扱うことによって，例えば，ある場所へ行く予定のあるユーザーのみに広告を配信したり，ある場所に実際にいたユーザーの生の声を抽出することが可能になる．この問題は，ツイート中で地名語がどのように言及されているかに依存する問題であり，通常の Bag-of-Words 素性に基づいた文書分類に基づくアプローチでは性能に限界がある．我々は，単語の埋め込みをベースにした畳込みニューラルネットワーク (CNN) を用いた分類モデルを提案する．さらに，CNN に，地名語に依存する特徴量や，形態論的な特徴量を導入することによって，性能の向上がはかれることを示す．

連絡先: 松田耕史，東北大学大学院情報科学研究科，matsuda@ecei.tohoku.ac.jp

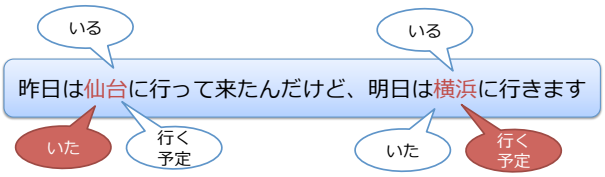

図 1: ツイート中の地理情報に対する時間極性の自動付与

2. データ作成

以下に示す 10 種の地名語 (ターゲット語) を含むツイートを収集し，それぞれ 1,200 件のツイートをランダムに抽出した．

秋葉原，清水寺，スカイツリー，渋谷駅，仙台，市役所，交差点，病院，改札，動物園

各ツイートに対して，7 人の作業者が表 1 のラベル集合を元にアノテーションを行った *1．アノテーションは，Yahoo! クラウドソーシングのプラットフォームを用いた．地名語の選定にあたっては，エンティティの特定性が高い地名語を 5 個，エンティティの特定性が低い地名語を 5 個選び，地名，施設名，観光地名等をバランス良く含むように考慮した．作成したデータの概要を表 2 に示す．ラベルの分布の傾向としては，`Non-Temporal` がやや多いものの，それ以外のラベルは それぞれ概ね同じ割合で含まれている．データを観察したところ，以下のような観察が得られた．

- 文頭や文末，ターゲット語の近くなどの，文中の特定の範囲に，大きな手がかりになり得る表現があることが多い．
- 対象語それ自体に対するラベルの依存性はそれほど強くない．たとえば，ある文脈の中に含まれる対象語それ自

*1 http://crowdsourcing.yahoo.co.jp/request/detail/3588468313

表 1: *TA* のアノテーションに用いたラベル集合

ラベル	説明	例 (ターゲットを太字で示す)
`Present`	現在，対象語で表される場所か，その近くにいる	間近で見る**スカイツリー**はきれい.
`Past`	現在，対象語で表される場所にいないが，過去にいた	週末の思い出は，曇ってる**スカイツリー**と燃え盛るテーブル.
`Future`	現在，対象語で表される場所にいないが，これから行くつもりであるようだ	今から**スカイツリー**へ．下で行くから 1 時間くらい.
`Non-Temporal`	対象語で表される場所に言及しているだけで，行く予定があるわけでも，いたわけでもない	**スカイツリー**って何時から開くんだろ ??
`Non-Mention`	対象語で表される場所に言及していない	**スカイツリー**ラインの冷風，台風並みに強い

表 2: クラウドソーシングで作成したデータの概要 (ラベルは 5 人一致したものを採用した)

合計ツイート数		12318
一致度	7 人一致	2212(18%)
	6 人一致	5452(44%)
	5 人一致	3797(31%)
ラベル	`Present`	2413(20%)
	`Past`	2342(19%)
	`Future`	2134(17%)
	`Non-Temporal`	4416(36%)
	`Non-Mention`	962(8%)

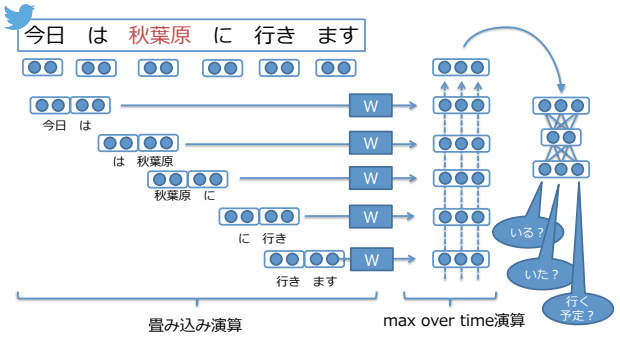

図 2: CNN に基づく文分類モデル．それぞれの単語は 2 次元のベクトルで表現されており，2 単語長のフィルタを 3 つ適用する ($\mathbf{W} \in \mathbb{R}^{4\times3}$) である場合を模式的に表す

体が「秋葉原」であるか,「仙台」であるかによってラベルが異なる例は殆どなく，多くの事例は対象語の文脈からラベルが決定可能であった.

- 複数のラベルが付与されうる事例が存在する．例えば，以下の文における仙台は，`Past` と `Future` のどちらを付与すべきか現状のガイドラインからは定かではない.

(2) 清水寺$_{\texttt{Past?Future?}}$，またじっくり行ってみたい

3. 提案手法

本稿において，我々はこの問題を，ターゲットに依存した文書分類問題 [Jiang 11] として解く．単語のつながりと，単語ベクトルの持つソフトな特徴量を同時に考慮するために，CNN(畳み込みニューラルネットワーク) を用いる (図 2)．まず，我々がベースとして用いた，Kim らの CNN に基づく文分類モデル [Kim 14] を概説する．Kim らのモデルにおいては，文を単語ベクトルの連結で表現し，そこに固定サイズの畳み込み（フィルタ）を適用することにより，単語ベクトルから素性抽出を行う．$\mathbf{x} \in \mathbb{R}^k$ を k 次元の単語ベクトルとする．ある文 は単語ベクトルの連結を用いて，以下のように表される:

$$\mathbf{x}_{1:n} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus ... \oplus \mathbf{x}_n \tag{1}$$

ここで，$\oplus$ はベクトルの連結を表す演算子である．$\mathbf{x}_{i:i+j}$ を，文中の i 番目の単語から $i+j$ 番目までの単語の連結とする．ここで，フィルタ行列 $\mathbf{W} \in \mathbb{R}^{hk\times L}$ を導入する．ここで，h はフィルタのサイズ (n-gram モデルの n に相当する) を表しており，今回は 2 単語とした．また，L は適用するフィルタの数を表す．l 番目のフィルタ ($\mathbf{W}$ の l 行目のベクトル) を i 番目の単語に適用した結果は以下のように計算される:

$$c_{i,l} = f(\mathbf{W}_{\cdot l} \cdot \mathbf{x}_{i:i+h-1} + b) \tag{2}$$

ここで，$b \in \mathbb{R}$ はバイアス項である．f は非線形関数であり，本稿ではシグモイド関数を用いた．このフィルタを，文中のすべての位置 $(1, 2, ...n-h+1)$ に対して適用し，以下のような素性マップ $\mathbf{c} \in \mathbb{R}^{n-h+1}$ を得る：

$$\mathbf{c}_l = [c_{1,l}, c_{2,l}, ..., c_{n-h+1,l}] \tag{3}$$

その後，素性マップベクトルの最大値を取り出す (max-over-time pooling).

$$\hat{c}_l = max(\mathbf{c}_l) \tag{4}$$

これによって，あるフィルタベクトル $\mathbf{W}_{\cdot l}$ で計算された素性の最大値 (最も重要な素性) を取り出す．これを各フィルタベクトルに対して行い，全結合層に対して入力する．最終的なラベルの分布は，行列 $\mathbf{A}$, ベクトル $\mathbf{b}$ をパラメータとして持つソフトマックス層を通して以下のように計算される:

$$p(\hat{y}|x) = softmax(\mathbf{Ac} + \mathbf{b}) \tag{5}$$

目的関数として，交差エントロピー損失関数を用い，誤差逆伝播法で学習を行う.

3.1 バイナリ次元の追加に基づくターゲット情報の導入

さらに，対象語の文脈に依存した *TA* を推測するために，対象に関する情報をモデルに導入することを考える．我々は，ベクトルに次元を追加することで，単語ごとにターゲットか否かを判断できるようにすることを考える．図 3 に示すように，単

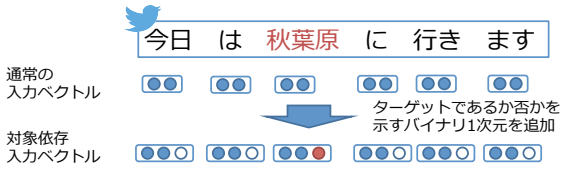

図 3: バイナリ次元の導入に基づくターゲット情報の導入

語毎に構成される入力ベクトルに，対象語であれば 1，そうでなければ 0 になるような次元を追加した．これによって，たとえば，対象語の近隣に出現する特定の語に対して働くような特徴抽出器 (フィルタ行列 **W**) が学習されるようになることが期待される．

同様に，2. 節の観察に基いて，強い手がかりになり得る以下の位置に対して，同様の手続きでバイナリの次元を追加した．

1. 文頭，文末の単語に対して発火する素性 (文頭文末素性)
2. 対象語と同じ文のすべての語に対して発火する素性 (対象同文素性)

3.2 形態論的な素性の導入

形態素解析の結果得られる活用形の情報なども，離散的な情報として取り入れることが望ましいと考えた．しかし，すべての語，すべての形態論的な特徴を 1-of-K 表現の形で入力することは，データスパースネスの問題から難しい．今回は，1-of-K 形式とベクトル埋め込みの折衷案として，以下の 2 つの付加情報を元のベクトルに追加した．

1. *TA* の識別のためには，用言の活用の情報が不可欠であると考えたため，形態素解析の結果得られる活用形の形を 27 次元の 1-of-K 表現の形で入力ベクトルに連結した．
2. 予備実験において，Bag-of-words モデルにおいて大きな重みが学習された単語を確認したところ，いずれも *TA* の識別に重要な手がかりになることが分かった．そのため，訓練データから学習したベースラインモデル (Bag-of-words を素性とした基づく多クラスロジスティク回帰) の重みの絶対値が大きいものから 100 単語を 1-of-K 形式の形で追加した．

4. 実験

我々の提案するモデルの性能を検証するために，クラウドソーシングを用いて，ツイートと地理情報に対する *TA* のアノテーションを収集し，実験を行った．

今回の自動付与実験においては，アノテーションを行ったすべてのツイートのうち，5 人によって付与されているラベルが一致しているもののみを学習/評価に用いた (全体の 93%のツイート)．さらに，ラベルが `Non-Mention` であるものを除いた．これは，地名語の抽出とグラウンディング (特定の場所を指しているか否かの判断を含む) は，本稿で述べる判別モデルの前処理として切り出すことが可能であるためである．作成したデータを，各キーワードに対して 700 件の訓練データ，100 件の開発データ，100 件のテストデータに分割し，実験に用いた．

表 3: 実験結果 (テストセットにおける正解率)

手法	インドメイン	クロスドメイン
Majority Baseline	0.390	0.390
多クラスロジスティック回帰	0.673	0.593
CNN 基本モデル [Kim 14]	0.693	0.630
+ 対象語素性 (3.1, 図 3)	0.698	0.647
+ 活用形素性 (3.2.1)	0.698	0.646
+ BL 重み素性 (3.2.2)	0.698	0.638
+ 文頭文末素性 (3.1.1)	0.697	0.638
+ 対象同文素性 (3.1.2)	0.700	0.647
CNN + 全追加素性	**0.712**	**0.649**

4.1 実験設定

提案モデルは Chainer を用いて実装した．CNN に入力する埋め込みベクトルには， word2vec を用いてツイートデータから訓練した 300 次元のベクトルを用いた．また，過学習を防ぐため，全結合層において，ドロップアウトに基づく正則化を適用した．バッチサイズは 500，フィルタ の数は 1000，サイズは 2，隠れ層のユニット数は 150 に固定している．最適化は Adam を用いて行った．

比較のためのベースラインとして，Bag-of-Words を素性として採用したロジスティク回帰モデルを用いた．定量的な評価指標として，テストデータに対する正解率を用いた．ただし，初期値の影響を排除するため，行列の初期値を変えながら 10 回実験を行い，その平均正解率で評価を行った．

エポック数は，開発データに対して正解率が最大となるように調節を行った．具体的には，開発データにおける正解率が 5 エポックの間最高値を更新しなかった場合に，最高正解率のモデルを用いてテストデータの解析を行った．実験は，以下の二つの設定で行った．

4.2 インドメイン実験

対象としたい地名語を含めたすべての訓練データを用いる設定．これは，判別を行いたい地名語に対する訓練データの入手が容易に行える場合を想定した実験である．

4.3 クロスドメイン実験

本実験においては，訓練において，テスト対象の地名語を含んだツイートを一切用いない．これは，判定を行いたい地名語に関する訓練データの入手が容易ではない場合を想定している．より実態に即しており，未知の地名語に対する *TA* の判別性能を見積もるのに適した設定である．

5. 結果と考察

実験結果を 表 3 に示す．CNN の導入により，ベースラインと比較して正解率が改善していることが分かる．また，CNN に対して，対象語の情報や，活用形の情報を追加素性として埋め込むことで，正解率がさらに向上していることが分かる．

CNN は，初期値に依存する局所解に収束するモデルであるため，初期値の影響を考慮した上で，それぞれの追加素性の効果を調査する必要がある．モデルの初期値を変えて，10 回の実験を行った際の平均正解率と標準偏差を 図 4 に示す．それぞれの追加素性が正解率の向上に寄与しており，インドメイン実験においてはそれらをすべて組み合わせることで，性能が大きく向上することが分かる．クロスドメイン実験においては，

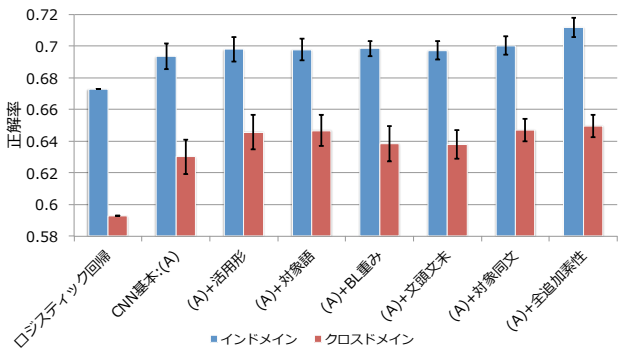

図 4: 各追加素性の効果

追加素性の組み合わせによる正解率の改善幅はあまり大きくなかった.

5.1 定性的なモデル分析

ターゲット情報の導入によって，正しく解析が行われるようになった例を示す.

(3) 昨日は 名古屋Past に行って来た．今日は 横浜Present だ.

ターゲット情報を導入していないモデルにおいては,「名古屋」にも「横浜」も `Present` ラベルが付与されたのに対して，ターゲット情報を導入したモデルにおいては，ターゲットの出現文脈に依存した *TA* を正しく認識することが可能になった.

6. 関連研究

6.1 テキスト中の地理情報/時間情報の融合解析

鬼塚らは，観光情報の抽出を目的として，特定の地名語を含んだツイートの著者が，現地にいるかどうか，という二値分類を行った．具体的には，地名語を含んだツイートの前後のツイートに対してルールと機械学習に基づく判別を行い，地名語で表される場所にいるかどうかの二値分類問題として解いている [鬼塚 14]．また，Li らは，ツイートに含まれる地名を抽出する系列ラベリング問題 (POI Extraction) を解く際に，ツイート著者がその場にいるか，という情報をラベルとして導入することによって，CRF に基づく系列ラベリングに基いて，時間ラベル付きの場所名抽出を実現する方法を提案している [Li 14] .

6.2 CNN のテキスト分類への応用

近年，CNN をテキスト分類に用いる試みがなされている．特に，感情分類や比較的短いテキストに対して有効であるという報告がなされている [Johnson 14, Kim 14, Severyn 15]．しかしながら，我々の問題設定は，単にツイートを分類するだけではなく，ツイート中のターゲット（地名語）に対するツイート著者の主観的な認識を当てる，という，ターゲット依存の分類問題である点で異なる．ターゲット依存の分類問題は，感情分析の文脈で盛んに研究されている [Jiang 11, Dong 14, Vo 15] が，CNN を直接的に用いたモデルは我々が知る限り，まだ存在していない.

6.3 ツイートを対象とした事実性／時間関係認識

テキストに書かれているイベントが,「すでに起こった」ことであるのか「まだ起こっていない」ことであるのかを判定するタスクとして，事実性解析がある [Saurí 12]．我々が取り組んだ問題は，文中に明示されているとは限らない「著者が，ターゲット語で示される場所にいる」というイベントに対する事実性の解析ととらえることも可能である．事実性解析においては，モダリティや否定表現などの文法的な手がかりが重要であることが指摘されている．例えば，[叶内 15] は，ツイートから風邪やインフルエンザに罹患している人を抽出する公衆衛生サーベイランスというタスクにおいて，事実性解析が重要であることを指摘している．文中に明示されていないイベントに対する事実性の解析という点において，本稿における問題設定に近い.

7. おわりに

本論文では，ツイート中に出現する場所名への著者の主観的な時間極性 *TA* を自動推定するモデルとして，畳込みニューラルネットワークに基づく対象依存型の文分類モデルを提案した．提案手法においては，対象依存型の文分類を実現するために，対象の情報を単語ベクトルに追加した．提案手法の効果を示すために，クラウドソーシングを用いてツイートに対するアノテーションを収集し，その上で効果を実証した．我々が提案した手法は *TA* の解析だけではなく，対象依存型の感情分析等への応用が見込まれる.

謝辞

本研究は，東北大学工学部 情報知能システム総合学科「Step-QI スクール」の支援を受け，文部科学省受託研究「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」の一環として行われた．また，CNN によるテキスト分類器の実装を提供していただいた，東北大学工学部の五十嵐祐貴さんに感謝します.

参考文献

[Dong 14] Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., and Xu, K.: Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification., in *Proceedings of the ACL* (2014)

[Jiang 11] Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T.: Target-dependent twitter sentiment classification, in *Proceedings of the ACL-HLT* (2011)

[Johnson 14] Johnson, R. and Zhang, T.: Effective use of word order for text categorization with convolutional neural networks, *arXiv preprint arXiv:1412.1058* (2014)

[Kim 14] Kim, Y.: Convolutional neural networks for sentence classification, *arXiv preprint arXiv:1408.5882* (2014)

[Li 14] Li, C. and Sun, A.: Fine-grained location extraction from tweets with temporal awareness, in *Proceedings of the SIGIR* (2014)

[Saurí 12] Saurí, R. and Pustejovsky, J.: Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text, *Computational Linguistics*, Vol. 38, No. 2, pp. 261–299 (2012)

[Severyn 15] Severyn, A. and Moschitti, A.: Twitter sentiment analysis with deep convolutional neural networks, in *Proceedings of the SIGIR* (2015)

[Vo 15] Vo, D.-T. and Zhang, Y.: Target-dependent twitter sentiment classification with rich automatic features, in *Proceedings of the IJCAI* (2015)

[叶内 15] 叶内 晨, 北川 善彬, 荒牧 英治, 岡崎 直観, 小町 守：Web 情報からの罹患検出を対象とした事実性解析・主体解析の誤り分析, 自然言語処理, Vol. 22, No. 5, pp. 363–395 (2015)

[鬼塚 14] 鬼塚友里絵, 嶋田和孝：前後文脈を考慮した Tweet の現地性判断, 信学技報, 社団法人 電子情報通信学会 (2014)