The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

イベント知識抽出のための畳み込みニューラルネットワーク
Convolutional Neural Network for Event Knowledge Extraction

石川 開[*1] Kai Ishikawa
高村 大也[*2] Hiroya Takamura
奥村 学[*2] Manabu Okumura

[*1] NEC データサイエンス研究所
Data Science Research Laboratories, NEC Corporation

[*2] 東京工業大学
Tokyo Institute of Technology

1. はじめに

今日、高度な知識処理を備えた対話システムは、スマートフォンやスマートスピーカーなどの日常的なデバイスを通じて我々の生活に浸透し、身近なツールとなりつつある。また、社会や産業の高度化に伴って、ますます増大しつつある知識労働者の負担を、知識処理システムで支援するという試みも、医療、法務などの様々な領域で活発化している。このように、AI による社会変革への関心の高まりと共に、実社会課題への適用が進み始めた知識処理システムではあるが、その実現には、システムの適用領域や目的に応じた膨大な知識が必要となる。この知識の開発・メンテナンスコストが、これまで知識処理システムを実用化する上での大きな障壁となってきた。知識処理システムの社会適用を本格的化する上で、知識開発を自動化するブレークスルーが必要となる。それにより、例えば、知識処理システム自体が、運用の過程でシステムに不足する知識を自己診断し、自律的に文書群や専門家などの情報源にアクセス、システムの知識の網羅性、信頼性、鮮度を維持する、といったアプローチが現実的なものとなる。

本研究では、このような知識開発の自動化の導入として、シェアードタスク NIST TAC KBP2016 Event Nugget Detection [Mitamura 2016] を対象に、ベースラインとなる知識抽出システムを構築し、改善方法を提案する。同タスクでは、課題となるイベントに対し、ラベル付きの事例が与えられるが、事例に含まれない語彙や表現も正しく抽出しなければならない。このため、教師データの限られた語彙や表現から、汎用性の高いモデルを構築する必要がある。そこで我々は、ベースラインに文分類のための畳み込みニューラルネットワーク [Kim 2014] を選択する。分散表現 [Mikolov 2013] の適用は、大規模離散的な語彙空間を低次元の連続的な意味空間に写像するため、事例の表層的な違いを吸収する効果が期待できる。また、畳み込みニューラルワーク(CNN)については、学習データに含まれない単語列の意味を、分散表現の畳み込みで構成、推定できる様な効果を期待している。並列処理向きで GPU による高速化の効果が大きいという利点もある。一方、CNN の固定長入力、近接する入力要素の階層的な畳み込みといった特性が、知識抽出タスクでは次のような問題を生じる。

(1) 短いフレーズでイベントが確定する事例では、入力中の周辺語がノイズとなって精度が低下する。一方、入力を短くすると、文脈を考慮する必要のある事例が正しく検出できない。

(2) 近接する入力の畳み込みを階層的に繰り返す CNN では、構成要素の離れた複単語表現(MWE)の学習が難しい。

以降、上記の問題に対する改善方法を提案、評価実験でベースラインと比較、実証する。また、NIST TAC KBP2016 の参加システムの公式評価の比較で、提案手法が KBP2016 の全ての参加ステムの性能を上回る事を示す。

2. イベント知識抽出システムの概要

本研究では、NIST TAC KBP2016 の Event Nugget Detection に対する評価システムを構築し、提案手法の評価実験を行う。Event Nugget Detection では、システムは定義された全 18 種類のイベントタイプのいずれかに該当するトークン列をテキスト中から検出する。この際、トークン列のテキスト中の位置や事実性の分類を含めたシステム全体での精度が評価される。

本研究の知識抽出システムの処理フローを図1に示す。本システムでは、入力文書に簡易なトークナイザー [Kim 2014] を適用してトークン列を生成する。

次に、イベントトークン検出では、各文書のトークン列から、全 18 種類のイベントタイプのいずれかに該当するトークンを検出する。さらに、検出されたイベントトークンに対し、イベントタイプ分類と、事実性分類が適用される。イベントタイプ分類は、検出されたイベントトークンのタイプを一意に特定し、事実性分類は、イベントトークンの事実性の状態が、事実(actual)、一般(general)、その他(others)のいずれに該当するかを特定する。.

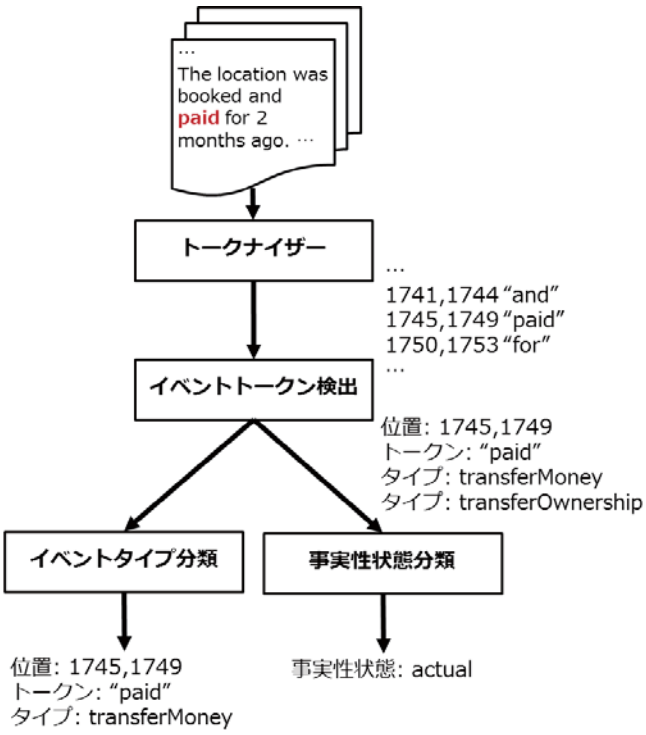

図１： 本研究のイベント知識抽出システムの処理フロー

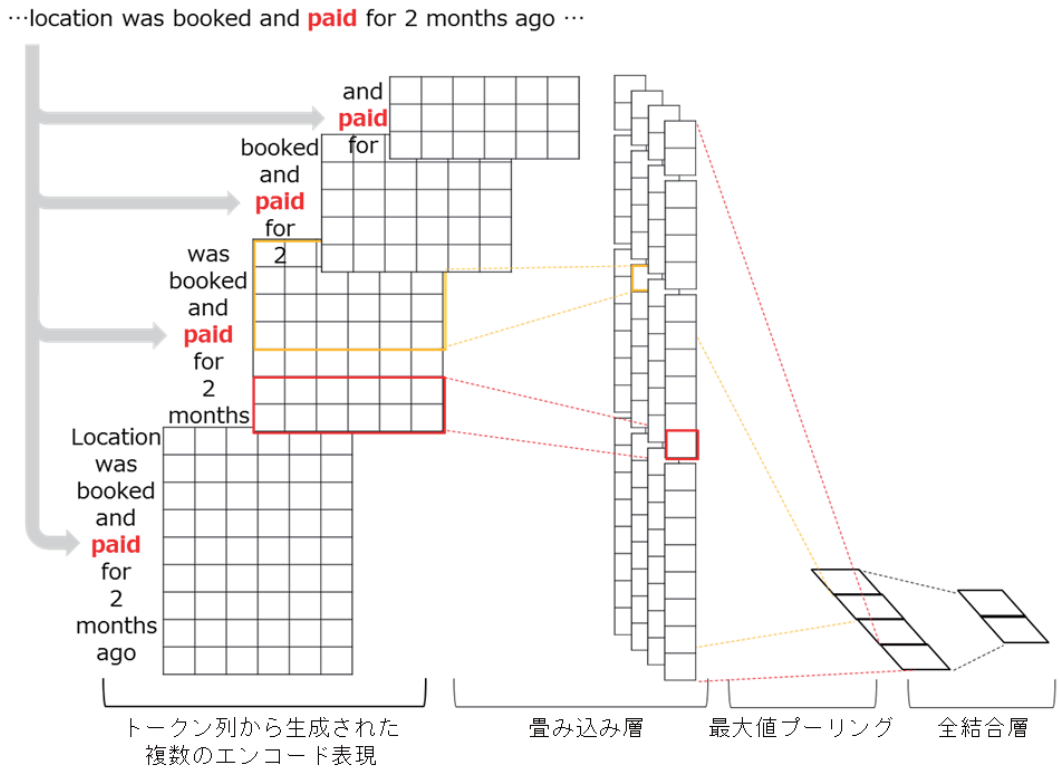

図2： MW 法におけるトークン列からの入力表現の生成例とニューラルネットワークの構成

3. イベントトークン検出モデル

3.1 ベースライン： Single Window (SW) 法

前述のように、イベント検知では、各文書から生成されたトークン列中の各トークンに対し、全 18 種類の各イベントタイプに該当するかどうかの二値分類を行う。本研究では、この二値分類のベースラインとして、Collobert 等のモデル [Collobert 2011] の改良版である文分類のための畳み込みニューラルネットワーク [Kim 2014] を用いる。

この分類手法では、各トークンのエンコーディングに、の word2vec [Mikolov 2013] によって得られる 300 次元の単語ベクトル表現を用いる。そして、分類対象のコンテキストの長さを制限するために、固定長の窓を導入する。入力文がこの窓よりも長い場合は窓の長さにトリミングし、この窓の長さよりも短い場合は、パディング（padding）により入力を窓の長さに一致させる。この文分類の手法を元に、窓の中心にあるトークンを検出の対象トークン、窓内の周辺トークンを、対象トークンの文脈と読み替えれば、イベントトークンの検出に適用することが出来る。以降、この手法をベースライン手法とし、Single Window (SW) 法と呼ぶ。

SW 法において、ある対象トークンのニューラルネットワークへの入力は、対象トークンと、前後に長さ l のトークン列を加えた、式(1)に示すような長さ 2l+1 のトークン列に対する単語ベクトル表現である。

$$x_{-l:l} = x_{-l} \oplus \ldots \oplus x_0 \oplus \ldots \oplus x_l \qquad (1)$$

ここで、x_iは窓中の i 番目の単語に対する k 次元の単語ベクトル、⊕は結合（concatenation）オペレータである。

SW 法は、原理的には窓を大きく設定することで長い文脈を考慮できるが、実際には冒頭で述べた通り、次のような問題が存在する。

・問題点1： 窓中の各トークンが同じ重みでネットワークに入力されるため、窓を大きくすると、表現によってはイベント検知と直接関係の無い入力情報が増え、窓の拡大によって全体的な検知精度が低下する。

・問題点2： 隣接トークン対に比べて離れたトークン対の意味の構成性は CNN では学習が難しく、非隣接の出現パタンで、イベント検知に有用な複単語表現(MWE)などが上手く学習されない可能性が生じる。

以降、これらの問題に対する改善手法の提案を行う。

3.2 提案1： Multiple Window (MW) 法

第1の提案は、ベースラインの問題点1の解消を目的とした Multi Window (MW) 法の導入である。MW 法は、対象トークンを共通の中心とし、大きさが段階的に異なる窓を複数用意し、各窓に対するエンコードを生成、多チャネルでネットワークに入力する。具体的には、式(2)に示すような、長さ 2l+1, 2l-1, …, 1 の l + 1 個の窓に対するエンコードの集合として表される。

$$\{x_{-l:l}, x_{-l+1:l-1}, \ldots, x_{-1:1}, x_0\}. \qquad (2)$$

図2は、MW 法におけるトークン列からエンコード表現を生成し、ネットワークに入力する全体構成を示している。図に示すように、式(2)で得られた l+1 個のエンコードを並列でネットワークの畳み込み層（入力層）に入力する。畳み込み層の出力は、最大値プーリングで圧縮後、全結合層に入力される。ソフトマックスにより、最終的に、統合された予測結果（二値分類の予測結果と予測確率）が得られる。

3.3 提案2： Pseudo Multiword Expression (PME) 法

第2の提案は、ベースラインの問題点2の解消を目的とした Pseudo Multiword Expression (PME) 法の導入である。PME 法は、対象トークンと、そこから一定距離内にある全てのトークンとの対を、擬似的に複数語表現（multiword expression）の構成要素とみなし、各トークン対に対するエンコードを生成、多チャネルでネットワークに入力する。対象トークンから距離 l 以内のトークンから得られるトークン対を用いて、対象トークンに対して式(3)のような 2l 個のエンコードが得られる。

$$\{x_{-l} \oplus x_0, \dots, x_{-1} \oplus x_0, \dots, x_0 \oplus x_l\}. \tag{3}$$

この 2l 個のエンコードを、MW 法で用いたのと同じネットワークに並列で入力することによって、最終的に、統合された予測結果（二値分類の予測結果と予測確率）が得られる。

3.4 イベント検知モデルの動的選択

イベント検知モデルは、窓サイズや用いるエンコード表現(SW 法、MW 法、PME 法)によって検知特性が異なる。システムの全体性能を向上するためには、対象トークンごとに相性の良いモデルが選択され、全体としてシステムの検知精度が向上することが望ましい。このような考えから、本研究では、CNN から出力される予測確率を確信度のように見なすことで、対象トークンごとに動的にモデルを選択する方法用いる。具体的には、式(4)に示すように、モデル(model)がトークン w をイベントタイプ(event type)として検知する際、予測確率 $P_{model}(event\ type|w)$ が最大となるモデルを選択し、その予測結果を、トークン w に対する複数モデルの統合的な予測結果として用いる。

$$\begin{aligned} model &= \underset{model}{\operatorname{argmax}}\{\max[P_{model}(event\ type|w), \\ &1 - P_{model}(event\ type|w)]\} \end{aligned} \tag{4}$$

4. 分類モデル

4.1 イベントタイプ分類

イベントタイプ分類では、前段の各イベントタイプ毎の検出結果の和集合に対し、複数のイベントタイプに判定されたトークンの再分類を行う事で曖昧性を解消し、タイプを一意に特定する。分類モデルには前述の SW 法、MW 法、PME 法を用いる。学習するデータセットは、イベントトークン検出においては、いずれのイベントタイプにも該当しないトークンを負例として学習するが、イベントタイプ分類においては、イベントとして検出されたトークンの分類が目的なので、学習対象から外す。得られた複数モデルの各イベントタイプに対する予測結果、予測確率を元に、式(5)を用いてイベントタイプの推定を行う。

$$\begin{aligned} &event\ type \\ &= \underset{event\ type}{\operatorname{argmax}} \left\{ \max_{model} P_{model}(event\ type|w) \right\} \end{aligned} \tag{5}$$

4.2 事実性状態(realis status)分類

事実性状態分類においても、イベントタイプ分類と同様、前段のイベントトークン検出の結果を元に、事実性状態を一意に特定する。分類モデルはここでも SW 法、MW 法、PME 法を用いる。ここでも、イベントとして検出されたトークンの分類が目的なので、訓練セットから、いずれのイベントタイプにも該当しないトークンを外して学習する。得られた複数モデルの各事実性状態に対する予測結果、予測確率を元に、式(6)を用いて事実性状態の推定を行う。

$$\begin{aligned} &realis \\ &= \underset{realis}{\operatorname{argmax}} \left\{ \max_{model} P_{model}(realis|w) \right\} \end{aligned} \tag{6}$$

5. 評価実験

5.1 データセットと学習条件

本評価実験では、NIST TAC KBP2016 の Event Nugget Detection (English)におけるシステムの公式の評価データセット・評価ツールを用いて評価セットを行った。また、訓練データには、KBP の訓練評価データをまとめた”TAC KBP Event Nugget Detection and Coreference - Comprehensive Training and Evaluation Data (LDC2017E02)”から KBP2016 の evaluation data を除いたデータと、”DEFT Rich ERE English Training Annotation (LDC2016E31, LDC2015E68, LDC2015E29)”を用いた。

畳み込みニューラルネットワークのハイパーパラメーターはベースライン、提案手法の全てのモデルで共通のものを用いた。具体的には、Zhang によるパラメータ設定の分析[Zhang 2017]を参考に、フィルタサイズを[2, 3, 4, 5]、特徴マップの次元数を 100、ドロップアウト率を 0.5、ミニバッチサイズを 50 とした。学習には、Zeiler の Adadelta による更新ルールとミニバッチのシャッフルを加えた確率的最急降下法[Zeiler 2012]を適用、10fold の交差検定による F1 値に基づいてアーリーストッピングを実施した。

単語分散表現モデルには、Google News に対する公開モデルである GoogleNews-vectors-negative300 を用い[Mikolov 2013]、このモデルでカバーされない語についてはランダムな初期値を与え、逆伝播による word2vec のチューニングを実施した[Kim 2014]。

5.2 イベントトークン検出の評価

イベントトークン検出における、ベースラインと提案手法の各モデルの適合率、再現率、F1（マクロ平均）を表1に示す。以降の表では、手法とパラメータ l（式(1)、(2)、(3)で定義）を、次のように表記する。SW(1)は、SW 法で l=1 としたモデル、SW(1,3,5)は、SW 法でl=1, 3, 5 とした 3 つのモデルを動的モデル選択で統合したモデルを表す。PME(3,5)+MW(0,3,5)は、PME 法で l=3, 5、MW 法で l=0, 3, 5 とした計5つのモデルを動的モデル選択で統合したモデルを表す。

表1： ベースラインと提案手法のイベントトークン検出の性能

モデル(窓サイズ)	適合率	再現率	F1
SW(0)	49.5	50.0	49.8
SW(3)	43.8	33.6	38.1
SW(5)	16.1	13.3	14.6
SW(0,3,5)	49.9	38.1	43.2
MW(0)	49.5	50.0	49.8
MW(3)	53.4	45.5	49.1
MW(5)	51.2	47.1	49.1
PME(3)	52.0	48.7	50.3
PME(5)	50.4	51.1	50.8
PME(3,5)+MW(0,3,5)	**53.0**	**53.8**	**53.4**

表 1 から、SW 法では、周辺語を考慮しない SW(0)が最も F1 が高いことから、問題点1が実際に確認できる。PME 法では、最

も離れた周辺語を考慮する PME(5)が最も F1 が高く、SW(0)を上回ることから、PME 法が離れた周辺語の考慮で精度を向上している事が確認できる。MW 法は適合率で優れており、PM 法と MW 法のモデルの統合が、全モデル中最良の F1 を得た。

5.3 イベントタイプ分類の評価結果

続いて、イベントタイプ分類におけるベースラインと提案手法の各モデルの適合率、再現率、F1 を表2に示す。SW 法の入力には SW(0)で検出されたトークン、MW 法と PME 法の入力には、PME(3,5)+MW(0,3,5)で検出されたトークを用いた。ベースライン中、提案手法中で、モデル間の性能差が少ないが、これは、トークン検出で生じたイベントタイプの曖昧性が少ない事を意味する。わずかであるが、PME 法の優位性が確認できる。

表2: ベースラインと提案手法のイベントタイプ分類の性能

モデル(窓サイズ)	適合率	再現率	F1
SW(3)	**41.7**	**41.3**	**41.5**
SW(5)	41.6	41.2	41.4
SW(3,5)	41.7	41.3	41.5
MW(3)	45.2	45.8	45.5
MW(5)	45.4	46.0	45.7
PME(3)	45.5	46.1	45.8
PME(5)	**45.5**	**46.1**	**45.8**
PME(3,5)+MW(3,5)	45.5	46.1	45.8

5.4 事実性状態分類の評価結果

事実性状態分類におけるベースラインと提案手法の各モデルの適合率、再現率、F1 を表3に示す。SW 法の入力には SW(0)で検出されたトークン、MW 法と PME 法の入力には、PME(3,5)+MW(0,3,5)で検出されたトークを用いた。ベースライン中では SW(3,5)が、提案手法中では PME(5)+MW(3,5)が最も F1 が高く、いずれも単一のモデルよりも、動的モデル選択による統合モデルが優れた結果を得た。このことから、異なるモデルが同等の性能水準である場合に、動的モデル選択による統合が有効に働くと考えられる。イベントトークン検出(表1)において、SW(0,3,5)が SW(0)を下回り、動的モデル選択が上手く機能しなかったのは、SW(0)と SW(3)、SW(5)の間の性能差が大きく開いていたためと考えられる。

表3: ベースラインと提案手法の事実性状態分類の性能

モデル(窓サイズ)	適合率	再現率	F1
SW(3)	37.0	37.5	37.2
SW(5)	36.0	36.6	36.3
SW(3,5)	**37.5**	**38.1**	**37.8**
MW(3)	39.9	40.4	40.1
MW(5)	40.5	40.8	40.7
PME(3)	40.4	40.9	40.6
PME(5)	40.7	41.2	41.0
PME(5)+MW(3,5)	**41.7**	**42.2**	**41.9**

5.5 タスク全体性能の評価結果

最後に、提案手法、ベースラインのタスク全体での性能水準を見るため、NIST TAC KBP2016 の Event Nugget Detection (English)の公式評価結の上位 10 チームの性能と比較した(表4)。その結果、MW 法と PME 法の動的モデル選択による提案手法 MW+PME が最も高い F1 値 35.85%となり、KBP2016 でトップの NYU1[Nguyen 2016]を上回ることが分かった。この事から、提案手法は、知識抽出システムとして現時点で世界トップレベルの性能水準にある事が示された。

表4: ベースライン、提案手法、KBP2016 参加システム上位 10 チームの Event Nugget Detection タスクの全体性能

参加システム/提案手法	適合率	再現率	F1
MW+PME (proposed)	**35.67**	**36.03**	**35.85**
NYU1	33.47	37.21	35.24
UTD1	34.05	33.12	33.58
wip3	38.38	29.1	33.1
SoochowNLP3	37.26	28.97	32.59
SW (baseline)	**31.13**	**30.85**	**30.99**
RPI-BLENDER2	31.92	28.59	30.16
LTI-CMU1	40.19	22.76	29.06
SYDNEY1	35.93	23.45	28.38
aipheshd-t161	29.94	23.81	26.53
Washington1	30.71	21.42	25.24
UMBC2	25.64	18.76	21.67

6. まとめ

本稿では、文書からの知識抽出の方式として、畳み込みニューラルネットワークを用いた知識抽出システムを開発し、NIST TAC KBP2016 Event Nugget Detection への適用を試みた。文書分類向けに開発されたベースライン SW 法における、固定長単一窓が不要な周辺語によるノイズを招く問題、構成要素が離れて出現する複単語表現(MWE)と畳み込みの相性問題への対処として、それぞれ MW 法、PME 法を提案した。KBP2016 の公式の評価セット・評価ツールを用いて評価を行った結果、提案手法(MW 法と PME 法の動的モデル選択による統合方式)は、NIST TAC KBP2016 の全参加チームを上回る F1 値 35.85%を得た。

今後、正例が極端に少ないイベントに対しても安定的に高品質なモデルを獲得することを課題に、さらなる改善を進めていく。

参考文献

[Kim 2014] Yoon Kim: Convolutional Neural Net-work for Sentence Classification, In Proceedings of EMNLP, 2014.

[Zhang 2017] Ye Zhang, Byron Wallace: A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification, IJCNLP, 2017.

[Mikolov 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean: Efficient estimation of word representations in vector space, In ICLR, 2013.

[Mitamura 2016] Teruko Mitamura, Zhengzhong Liu, and Eduard Hovy: Overview of TAC-KBP2016 Event Nugget Track, TAC, 2016.

[Nguyen 2016] Thien Huu Nguyen, Adam Meyers, and Ralph Grishman: New York University 2016 System for KBP Event Nugget: A Deep Learning Approach, TAC, 2016.

[Zeiler 2012] M. Zeiler: Adadelta: An adaptive learning rate method, CoRR, abs/1212.5701, 2012.

[Collobert 2011] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuglu, P. Kuksa: Natural Lan-guage Processing (Almost) from Scratch , Journal of Machine Learning Research12, 2011.