2U-03

品詞と係り受けを考慮したDual Embeddings CNNによる属性抽出

前田裕一朗† 遠藤聡志‡ 山田孝治‡ 當間愛晃‡ 赤嶺有平‡
琉球大学大学院理工学研究科† 琉球大学工学部工学科‡

1. はじめに

ショッピングサイトのレビューは商品の評判を知る上で有益だが、全てに目を通すのは困難である。そこで、商品の特徴を示した属性を抽出することで有益な情報を得やすくする研究が行われている[1]。従来の属性抽出はルールベースで行われていた[2]が、近年では機械学習により高い精度での抽出が行えるようになった[3]。Xu らは機械学習手法である Dual embeddings CNN を用いてドメイン毎の単語の意味変化に対応した属性抽出を行った。しかしこのモデルは、ルールベースで重視されている品詞と係り受けが考慮されていない。そこで本研究では、Dual Embeddings CNN に係り受けを含んだ品詞情報を与えることで品詞パターンを考慮した抽出を行うモデルを提案する。

2. 関連研究

駒田らは商品評価ツイートから品詞や係り受け情報を利用しルールベースによる属性抽出を行った[2]。係り先の文節に特定の品詞が含まれていた場合に属性候補にするといったルールを設け、単語が属性になるパターンとならないパターンの判定を行った。

Xu ら は機械学習を用いてドメインによる言葉の意味変化に対応した属性抽出を行うために Dual embeddings CNN を提案した[3]。ドメインによる言葉の意味変化とは、例えば「速さ」という言葉が一般的には単位時間あたりの移動距離を意味するが、コンピュータの話題では単位時間あたりの命令数を意味するといったことである。また属性抽出は、文章から単語の意味を学習し、そこで獲得した単語埋め込みと呼ばれるベクトル表現を特徴として学習器に与えることで行う。Xu らは意味の変化を捉えるために、意味の異なる 2 つの単語埋め込みを連結したものを dual embeddings と定義し、これを学習器である CNN に与える Dual embeddings CNN の提案を行った。実験では、単語埋め込みを独立に扱ったものと dual embeddings をそれぞれ学習器に入力として与え、抽出結果の比較を行った。実験の結果 dual embeddings が最も高い精度を示し、dual embeddings の有効性を明らかにした。しかし Dual Embeddings CNN には単語の品詞や係り受けなどの情報が含まれていない。本研究では dual embeddings にこれらの情報を加える提案を行う。

3. 技術要素

3.1 Dual Embeddings CNN (DE-CNN)

DE-CNN では一般的な単語埋め込みである general embeddings とドメイン固有の単語埋め込みである domain embeddings を CNN の入力に用いる。単語埋め込みは、自然言語を処理する際に用いる単語のベクトル化手法で、skip-gram[4]という単語の周辺に出現する語を推定する学習器を用い、その中間層の重みを取り出すことで獲得を行う。単語の意味が近いほど単語埋め込みの類似度も高いという特徴がある。先行研究では、general embeddings を一般的なコーパスである Wikipedia から学習し、domain embeddings を商品のレビューから学習することで獲得し、これらを連結した dual embeddings で 1 単語を表現した。

CNN は周囲の情報を畳み込んで特徴抽出を行うモデルで、文脈を考慮した特徴抽出や並列化による高速計算が行える。図 1 は DE-CNN の入力から出力までの流れである。入力層では単語を dual embeddings に変換している。太枠はフィルタの役割があり、フィルタで抽出した値が次の畳み込み層に渡される。複数の畳み込み層を挟むことで、ベクトルからラベルの推定において特徴となる値を抽出している。出力層では入力した単語が属性の先頭の単語か、2 番目以降の単語か、属性でないかを各単語に対応した位置に 3 次元の one-hot 表現で示している。

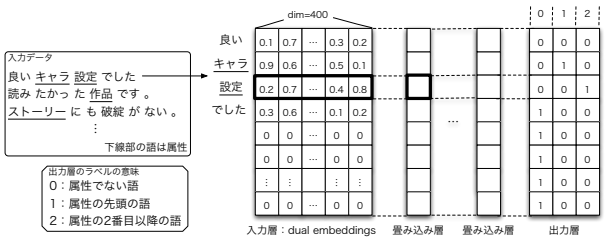

図 1：DE-CNN の流れ

3.2 Graph Attention Networks (GAT)

GAT はグラフ構造を畳み込むモデルである。グラフは特徴ベクトルと隣接行列で表現される。Veličković ら[5]は論文のカテゴリ推定を行うために、図 2-1 のように特徴ベクトルを各論文の単語の出現頻度(bag-of-words)、隣接行列を論文の参照関係としてグラフを生成した。グラフ構造の畳み込みは、図 2-2 のように、まず自身のノードと隣接するノードの特徴ベクトルに共通の重み W を掛け、ノード間の attention である α を求める。そこから自身のノードと隣接するノードの特徴ベクトルに各 α の値と重み W をかけ平均をとることで、隣接するノード情報を畳み込んだベクトルが生成できる。

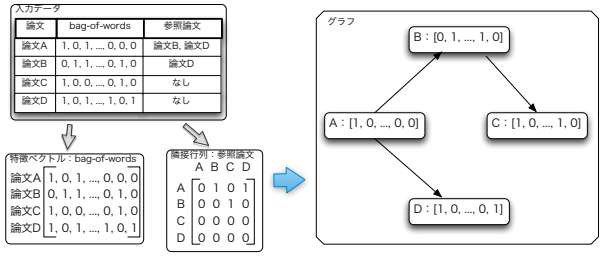

図 2-1：GAT におけるグラフ生成の流れ

Aspect term extraction using Dual Embeddings CNN based on part of speech and dependency
Yuichiro Maeda, Satoshi Endo, Koji Yamada, Naruaki Toma, and Yuhei Akamine
University of The Ryukyus

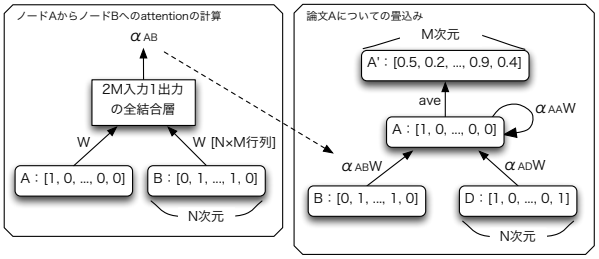

図 2-2：ノード A の畳み込み処理

4. 提案手法

品詞パターンの獲得を行うために、従来の dual embeddings に品詞や係り受けを含んだ特徴ベクトルを連結する。特徴ベクトルは、品詞の種類を表す品詞 ID と周辺に出現する品詞の頻度情報を持った品詞埋め込み、GAT でこの 2 つに係り先の情報を持たせた GAT(品詞 ID)と GAT(品詞埋め込みを)の 4 種類となっている。

4.1 品詞 ID

品詞の種類をベクトル長とし、品詞 ID に対応する配列番号に 1、それ以外に 0 を入力した one-hot で表現する。

4.2 品詞埋め込み

skip-gram を用いて中心単語から周辺品詞を推定するよう学習することでその単語の周辺に出現する品詞の頻度を獲得する。

4.3 GAT(品詞 ID)と GAT(品詞埋め込み)

品詞 ID と品詞埋め込みを GAT で畳み込むことで獲得する。特徴ベクトルを品詞 ID と品詞埋め込み、隣接行列を単語の係り先としてグラフを生成する。図 3 は品詞 ID を特徴ベクトルとした場合の入力データと畳み込み処理の一例である。

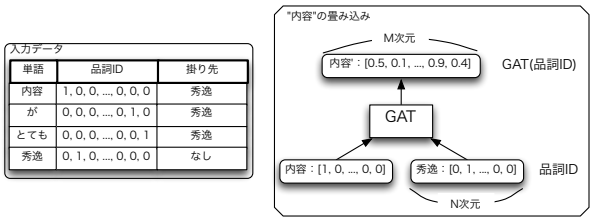

図 3：入力データと品詞 ID の畳み込みの例

5. 実験

提案した 4 つの特徴ベクトルが属性抽出に与える影響を、dual embeddings にこれらの特徴ベクトルを連結したモデルを用いて比較実験した。また、ベースラインとして DE-CNN を用いた。

5.1 実験データ

通販サイトの小説一作品と Laptop PC 一商品のレビューから、評価を述べた文章それぞれ 1000 文を選択した。小説のレビューはドメインの依存性が低く、Laptop PC のレビューはドメインの依存性が高いデータの例となっている。

5.2 実験結果と考察

800 文を学習データ、200 文をテストデータに用いた時の結果を適合率と再現率、および F 値で評価した。小説レビューの結果を表 1 に、Laptop レビューの結果を表 2 に示す。

モデル	適合率	再現率	F 値
DE-CNN	0.516	0.648	0.574
DE-CNN+品詞 ID	0.547	0.616	0.579
DE-CNN+GAT(品詞 ID)	**0.576**	0.634	0.603
DE-CNN+品詞埋込	0.540	**0.682**	0.602
DE-CNN+GAT(品詞埋め込み)	0.551	0.681	**0.609**

表 1：小説レビューにおける属性抽出の結果

モデル	適合率	再現率	F 値
DE-CNN	0.543	0.732	0.623
DE-CNN+品詞 ID	0.555	0.693	0.617
DE-CNN+GAT(品詞 ID)	**0.601**	0.751	0.667
DE-CNN+品詞埋込	0.574	**0.765**	**0.655**
DE-CNN+GAT(品詞埋め込み)	0.557	0.763	0.643

表 2：Laptop PC レビューにおける属性抽出の結果

表 1 と表 2 より、提案した 4 モデルは DE-CNN と比べて適合率が向上し、再現率は DE-CNN+品詞 ID と小説レビューの DE-CNN+GAT(品詞 ID)以外のモデルで向上した。また、適合率がもっとも高かったモデルは DE-CNN+GAT(品詞 ID)、再現率は DE-CNN+品詞埋め込みとなっている。

DE-CNN+GAT(品詞 ID)の適合率が最も高かったのは掛り先の品詞という特徴が属性のパターンの獲得に最も適しており、DE-CNN+品詞埋め込みの再現率が最も高かったのは周辺品詞が属性でないパターンを獲得するのに適しているからだと考えられる。また、DE-CNN+品詞 ID で再現率が低下しているが、これは、単語に付随する品詞単体では属性か否かを判定することができないからと考えられる。2 つのレビューについて、いずれも DE-CNN+品詞 ID 以外の提案モデルで F 値が向上していることから、周辺情報を加味した品詞情報がドメインの依存性に関わらず有効に働くことが確認できた。

6. おわりに

本稿では、機械学習の属性抽出法である DE-CNN に対し、ルールベースで重視されている品詞と係り受けの情報を追加して抽出実験を行なった。実験の結果、適合率は DE-CNN に GAT(品詞 ID)を加えたモデルが最も高く、再現率は品詞埋め込みを加えたモデルが最も高い値となり、品詞情報の重要性を示した。これらを合わせたモデルは適合率と再現率の両方の向上が期待されるが実験は行っておらず、また与えた品詞情報からどのようなパターンが獲得できたかについても未検証であり、これらは今後の課題となる。

参考文献

[1]山西良典, 古田周史, 福本淳一, 西原陽子, “出現頻度と構文特徴を用いたレビュー構造の俯瞰のための評価視点の抽出”, 知能と情報(日本知能ファジィ学会誌) Vol.27, No.1, pp.501-511(2015)
[2]駒田康孝, 山名早人, “商品評価ツイートからの属性語自動抽出手法の提案”, DEIM Forum 2014, B5-6, (2014)
[3]Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu, “Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction”, In ACL (2018)
[4]Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, “Enriching Word Vectors with Subword Information” arXiv:1607.04606(2016)
[5]Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio`, P., Bengio, Y.: Graph attention networks. In: ICLR (2018)