固有表現のカテゴリ分類への DNN の応用方法の検討

Examination of DNN Application to Category Classification of Named Entities

朝倉 遼
Ryo Asakura

岡山大学 太田研究室
Ohta Laboratory, Okayama University

概要　商品や施設などのレビューには書き手の感情の極性があり, それを特定するのがレビュー分析のタスクの一つである. さらに, 極性は, レビュー文書中の固有表現に対して付与されるのが望ましいため, 固有表現を認識し, それにカテゴリを付与するというタスクが必要になる. カテゴリ付与の際には SVM, CRF, 多層パーセプトロンによる分類がなされてきたが, 本研究では CNN, LSTM などを効果的に用いることでカテゴリ分類精度の向上を試みる.

1　はじめに

文書の解析において, 固有表現認識, 極性判定は重要なタスクであり, 様々な手法を用いて行われる. 従来は, 固有表現認識では特に, 文字列へのタグ付けの手法である Conditional Random Fields (CRF) や, 線形分類器である Support Vector Machine (SVM) によるクラス分類によって行われてきた. 近年, ニューラルネットワークによる手法も登場している. 例えば, n-gram に基づく多層パーセプトロンによる固有表現認識[1] が知られている. 本研究では, Collobert らの手法[1] をもとに, Deep Neural Network (DNN) による固有表現認識を行う.

本研究では, SemEval-2015 Task 12 [2] において提供されたデータを用いて実験を行う. SemEval-2015 task 12 において提供されるデータには, レストランドメイン, コンピュータドメインといったドメインが用意されており, それぞれレビュー文書と Gold Annotation と呼ばれるレビュー文書内の固有表現, そのカテゴリ, 極性が明示されたデータから構成されている. レビュー文書の一例は “The food was delicious.” のような短い文章である. これに対し, カテゴリと極性が {category=“FOOD#QUALITY”, target=“food”, polarity= “positive”} のように付与される. target は文中に現れる固有表現を, category はその固有表現に対して割り当てられるカテゴリを表す. カテゴリは固有表現とその周辺の文脈によって決定されている. また, カテゴリは表 1 のとおり 13 種類が前もって用意されている. polarity は固有表現に対する極性であり, positive, negative, neutral の 3 種類がある.

本研究では, DNN を用いて, レストランドメインのレビュー文書中の各単語へカテゴリ付与する実験を行う.

表 1: レストランドメインにおけるカテゴリ一覧

RESTAURANT#GENERAL	DRINKS#QUALITY
RESTAURANT#PRICES	DRINKS#PRICES
RESTAURANT#MISCELLANEOUS	DRINKS#STYLE_OPTIONS
FOOD#GENERAL	AMBIENCE#GENERAL
FOOD#QUALITY	LOCATION#GENERAL
FOOD#PRICES	SERVICE#GENERAL
FOOD#STYLE_OPTIONS	

2　提案手法

本研究では, レビュー文書中の単語に対するカテゴリ付与を DNN を用いて行う. 具体的には, Convolutional Neural Network (CNN), Long Short-term Memory (LSTM) を用いる. レビュー文書は SemEval-2015 task 12 のレストランドメインのものを用い, Gold Annotation をカテゴリ付与における正解データとした.

2.1　データの前処理

文書内の固有表現とその任意の情報は, 基本的にはある程度近い場所に現れると考えられる. よって, レビュー文書から n-gram を作成し, n 個の単語を同時に学習することで単語に対して適切にラベルを付与することができると予想した. 本研究では, n は 3 とした. また, SemEval-2015 task 12 の 13 のカテゴリに「その他」カテゴリを追加して 14 カテゴリとし, それらをすべての単語に割り当て, これを学習時の DNN への入力とする.

2.2　ネットワーク構造

実験は図 1~3 の 3 つのネットワークについて行った. いずれも, 入力層 (Input Layer), 単語埋め込み層 (Embedding), 最後の 2 つの全結合層 (Fully Connected Layer) , ソフトマックス活性化関数 (Softmax) は共通している.

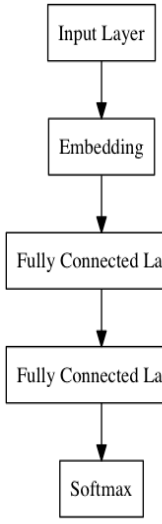

図 1: ベースラインのネットワーク

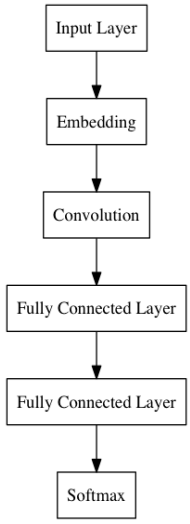

図 2: 畳み込みを用いたネットワーク (CNN)

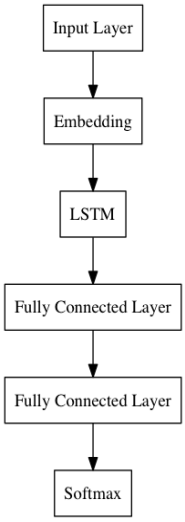

図 3: LSTM を用いたネットワーク

Embedding は n 個の単語を入力に, n 個の単語ベクトルを出力する層である. 本研究では単語 n-gram を入力とした.

図 1 のネットワークは単語埋め込み層の出力を単純に全結合層へ渡している.

図 2 のネットワークは単語埋め込み層の出力に対して 1 次元の Convolution を行う. ある単語のカテゴリに対し, 周辺の単語, 品詞の並び方は一定のパターンをなしていることが多いと考えられる. 畳み込みは, フィルターと呼ばれる重みが単語の出現パターンを抽出することができるため, 本研究で扱う問題に有効に働くと予想した.

図 3 のネットワークは単語埋め込み層の出力を LSTM に入力する. LSTM は 再帰型ニューラルネットワークの一つであり, 与えられた単語群を見て, それらの次に来る, 単語などの任意の事象を出力するユニットである. 本研究では, 与えられた n-gram の中央の単語のカテゴリは, その n-gram の次に出現する単語と関連が深いという仮定のもとに LSTM を用いた.

いずれのネットワークも, 最後に 2 つの全結合層を持ち, その出力を Softmax と呼ばれる活性化関数に通すことで最終的な確率分布を出力する. 全結合層は, 実験により, 2 層の場合が最も良い結果を示したため 2 層とした. 確率分布は 14 本とし, それぞれの確率は, 入力である n-gram の中央の単語がある カテゴリに属する確率となる. 全体を通すと, ネットワークは 単語 n-gram を入力として受け取り, その中心に位置する単語のカテゴリを確率分布の形で出力するという構造となる.

3 性能評価

SemEval-2015 Task 12 のレストランのレビューデータを用いて図 1 ~ 3 のそれぞれのモデルを学習し, それらの予測精度の評価を行った. 学習データは 833, テストデータは 401 のレビュー文書からなる.

最初に学習の収束の速さを評価した. 結果は図 4 のとおりとなった. 学習は 8 回までとし, 学習 1 回ごとにそれぞれのモデルによるカテゴリ付与の正解率を測った. 畳み込みを用いたネットワークは収束が速いことが読み取れる. また, LSTM を用いたネットワークは比較的収束が遅く, 単純に Embedding の出力を連結するだけであるベースラインのネットワークより学習時精度が悪いことが分かる.

次に, テストデータを用い, 学習済みモデルによるカテゴリ付与の精度を測った. 結果は 表 2 のとおりとなった. ベースラインは最も正解率が低く, 学習の収束が遅かった LSTM を用いたネットワークは最も精度が高くなっている. 一方, 学習の収束が非常に速かった畳み込みを用いたネットワークは学習時精度ほど高精度ではないという結果となった. これは学習データが少ないことによって過学習が起きていることが原因であると考えられ, さらなる検証が必要である. また, レビュー文書中に登場する単語のほとんどが「その他」カテゴリに分類されるため, 全単語を「その他」に割り当てた場合の結果も表 2 に記している. 以上より, テストデータにおけるカテゴリ付与において LSTM, CNN ともに有効であることが分かる.

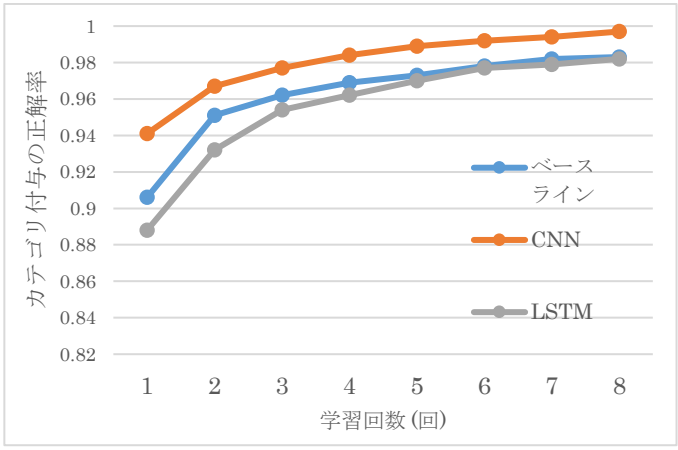

図 4: 学習の収束精度

表 2: テストデータにおけるカテゴリ付与の正解率

	ベースライン	CNN	LSTM	全単語に「その他」カテゴリを割り当てた場合
正解率	0.935	0.949	0.959	0.854

4 まとめ

本研究では, 文書中の単語のカテゴリ分類に DNN を用い, その有効性を確かめた. しかし, 文書の特徴量は単純な単語の並び以外にも様々なものが考えられる. 例として, 単語の品詞はその単語に割り当てられるカテゴリと大いに関連していることが予想できる. 品詞はすべての単語に割り当てることができるため, 単語列と同様のデータ構造の特徴量を作成することが可能である. さらに, 品詞列は単語列と同様に, パターンをなして出現すると考えられるため, 今回行った実験と同様に CNN, LSTM などの適用を検討している. また, より高精度なモデルを作成するためにはより多くの学習データが必要であるのは明らかであるため, それについても検討する. また, 現在は, カテゴリ付与の学習中に, 同時に単語ベクトルを学習する仕組みになっているが, word2vec などを用い単語ベクトルを事前に学習して決定することによる精度の向上も試みる.

参考文献

[1] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. “Natural Language Processing (Almost) from Scratch”. *Journal of Machine Learning Research*, 12:2493-2537, 2011.

[2] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos. “SemEval-2015 Task 12: Aspect Based Sentiment Analysis”, http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval082.pdf

[3] “Neural Networks for Named Entity Recognition”, http://nlp.stanford.edu/~socherr/pa4_ner