3I4-4

ニューラルネットワーク言語モデルを用いたセンチメント辞書の改良

Improvement of Sentiment Dictionary Using Neural Network Language Model

柳本 豪一*1
Hidekazu Yanagimoto

*1大阪府立大学
Osaka Prefecture University

In this paper I improve a dictionary for sentiment analysis, which is constructed with semi-supervised learning, using neural network language model. The automatically constructed dictionary includes a lot of words which does not include sentiment polarities essentially. To reduce such words from the dictionary I use a neural network language model. I found the proposed method could constructed a dictionary which was similar to a manually constructed dictionary.

1. はじめに

インターネット上にレビューなどの評価が含まれている文書が膨大に存在している。そのため、その評価極性を推定するセンチメント解析に関する研究が盛んに行なわれている。一般的には、予め単語に評価極性を付加した辞書 (センチメント辞書) を作成し、その辞書を用いて文書の評価極性を推定している。我々のグループでは半教師学習を用いたセンチメント辞書の自動構築に関する研究を行っている。しかし、作成された辞書には本来評価極性がないと思われる単語にまで評価極性が割り当てられるという問題があった。

本論文では、ニューラルネットワーク言語モデルと組み合わせることで、上記の問題である評価極性がない単語を辞書から削除する方法を提案する。ニューラルネットワーク言語モデルには Continuous Bag-of-Words モデルを用い、得られた単語のベクトル表現から計算された類似度を用いて辞書から削除する単語を決定する。これにより、半教師学習における閾値調整に比べて、不要な語の削除が行なえることが確認できた。

2. 半教師学習とニューラルネットワーク言語モデルを用いたセンチメント辞書作成

半教師学習によりセンチメント解析のための辞書作成の負担を軽減する。そして、過剰に単語に評価極性が付加される問題をニューラルネットワーク言語モデルから得られる単語の類似度をもとに削減する。以上により、人手で作成された辞書に近いセンチメント解析用の辞書の自動作成を目指す。

2.1 半教師学習による辞書作成

文章の評価極性を決定するためにその文章に含まれる単語に着目し、その単語の評価極性が一般的に用いられている。このため、単語と評価極性を紐付けたセンチメント辞書をあらかじめ用意しておく必要がある。しかし、辞書作成は人手を要する負荷の高い作業である。このため、あらかじめ人手で決定した少数の評価極性を割り当てた単語を用いて、他の単語の評価極性を推定する半教師学習が効果的である。本手法では、評価極性が判明している単語との共起度合いを χ^2 値で評価し、その値を用いて評価極性が未知な単語の推定を行なう。

連絡先: 柳本 豪一，大阪府立大学，堺市中区学園町 1-1，072-254-9279，072-254-9279，hidekazu@cs.osakafu-u.ac.jp

χ^2 値は評価極性が分かっている単語群 w^s ごとに以下のように計算する。

$$\chi^2_s(w) = \frac{1}{|D^s|} \sum_{w' \in D^s} \frac{(\mathrm{freq}(w', w) - \frac{\mathrm{freq}(w')\mathrm{freq}(w)}{N})^2}{\frac{\mathrm{freq}(w')\mathrm{freq}(w)}{N}} \quad (1)$$

ここで、$\mathrm{freq}(w)$ は単語 w の出現頻度、$\mathrm{freq}(w_1, w_2)$ は単語 w_1、w_2 の共起頻度、N は全単語の出現頻度、D^s は評価極性 s を持つ単語の集合を表す。

χ^2 値を評価極性が未知の単語に対して Positive と Negative の両方について求め、その差が閾値を超えたときに単語の評価極性を決定する。この閾値を大きくすることで、片方の評価極性に偏りが大きい単語だけ抽出することができ、精度の高い辞書を作成することができる。しかし、辞書に登録される単語が減ることにより評価極性を推定できる文章が減り多くの文章が Neutral と判断され、評価極性の推定精度が低くなる可能性がある。したがって、適切な閾値を決定する必要がある。

2.2 ニューラルネットワーク言語モデルの辞書の改良

半教師学習を用いた辞書作成では、上記に示した適切な閾値を決定するという問題があった。また、評価極性が分かっている単語との共起頻度に基づいて推定を行なうため、コーパスにおける単語の出現頻度の偏りにより辞書の精度が左右されてしまう。その一例として、見かけ上の共起の偏りにより本来極性を持たない単語に対しても極性を割り当ててしまうという問題があった。例えば、閾値として 300 として Positive な 5 単語と Negative な 5 単語をあらかじめ設定し、半教師学習により単語の評価極性を推定すると 268 個の単語に評価極性が付加された。しかし、人手で作成した辞書では極性が付加されなかった単語が 189 単語も含まれている (表 1 参照)。このような評価極性がないと思われる単語を削除するため、本手法ではニューラルネットワーク言語モデルを用いる。

ニューラルネットワーク言語モデルでは、テキストコーパスを用いて単語をベクトルの形で表現することができる。そして、線形演算により単語間の関係を把握できることが知られている [Mikolov 13]。株式ニュースを用いた予備実験より、評価極性を有する単語 (極性は無関係) が近くに配置されることが分かった。このため、辞書に登録された単語間の類似度を計算し、ノードを単語とする重みつきグラフを作成し、閾値より小さい重みを削除することで孤立するノードに対応する単語を削除することで、上記の問題を解決を目指す。

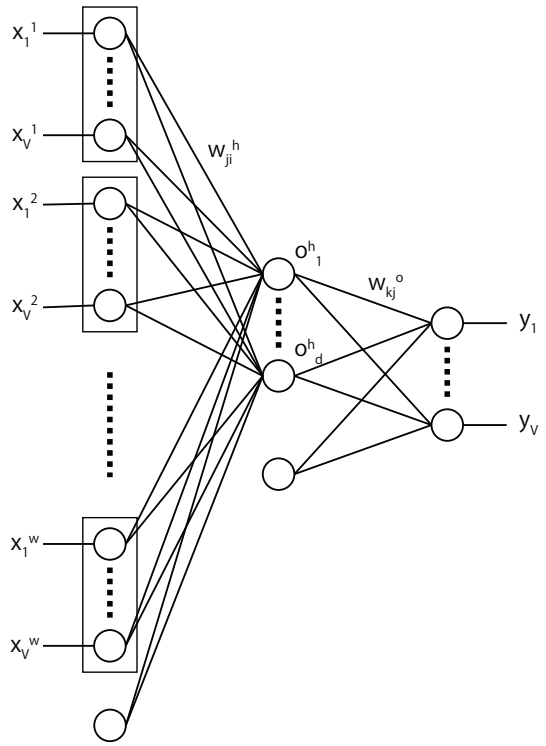

図 1: Continuous Bag-of-Words モデルの構成

ニューラルネットワーク言語モデルを構築するニューラルネットワークとしては Continuous Bag-of-Words モデルを用いる。実際の構成を図 1 に示す。このモデルは注目する単語の前後 n ワードを入力として、注目単語を予測するニューラルネットワークである。各入力単語と注目単語は 1-of-n coding により表現され、ニューラルネットワークの入力と出力とする。入力単語と隠れ層間の重みは入力単語間で共有している。このニューラルネットワークは誤差逆伝搬法を用いて学習される。本実験では、隠れ層のニューロンの発火が疎となるような補正を加えた学習を行なう [Yanagimoto 13]。学習終了後の入力単語と隠れ層の重みが単語をベクトルに変換するテーブルとなっているため、これを用いて単語をベクトルとして表現する。このベクトルの内積を単語間の類似度として用いる。

3. 実験

株式ニュースを用いたセンチメント辞書作成の実験を行なう。あらかじめ専門家により極性を持つと思われる単語の選別を行なってもらい、これを正解データとして用いる。

3.1 実験環境

テキストコーパスとして投資家に配信されている 2010 年度の株式ニュース T&C ニュースを用いた。これには 432,221 文の評価極性が不明な文章と 1,184 文の専門家による評価極性を付加した文章が含まれている。本実験では評価極性が不明な 432,221 文を用いて半教師学習による辞書作成とニューラルネットワーク減誤モデルの構築を行なった。

人手でセンチメント辞書を作成した。この辞書には 484 単語に Positive、428 単語に Negative の評価極性が付加されている。これを正解のセンチメント辞書とし、提案手法で作成された辞書がどれだけ近い辞書となるか評価する。

表 1: 登録単語の精度

		Positive	Negative	Neutral
Positive	SSL	32	7	123
	SSL+NNLM	26	6	58
	SSL(699)	21	4	68
Negative	SSL	8	32	66
	SSL+NNLM	4	18	17
	SSL(699)	3	13	20

3.2 結果と考察

まず、半教師学習のみで作成したセンチメント辞書について説明する。本実験では、χ^2 値の差の閾値として 300 を用いる。この結果は表 1 の SSL で表されている。

次に、上記の半教師学習で得られた辞書にニューラルネットワーク言語モデルを組み合わせ、単語の削除を行なう。このとき、単語間の類似度は 0.4 より大きいものとした。得られた単語数は 129 単語である。結果は表 1 の SSL+NNML で表す。

最後に半教師学習で閾値を変更した結果を検討する。辞書に登録される単語数が 129 語になるように閾値を修正した。閾値は 699 となり、結果を表 1 の SSL(699) で表す。

この結果より、センチメント辞書に登録される単語数が同じ場合に、半教師学習の閾値を修正する手法より、ニューラルネットワーク言語モデルを組み合わせた方が正しく推定された評価極性の単語を残しつつ、評価極性がない単語を効率的に削除できていることが分かる。閾値の調整手法では共起頻度のみに着目しているが、ニューラルネットワーク言語モデルでは共起頻度だけではなく、文章中での単語の置き換えられ易さなどを考慮した意味的な近さを評価できるため、センチメント辞書の改良を実現できたと考えられる。

ただし、正しく評価極性を推定された単語も削除されているため、ニューラルネットワーク言語モデルの設計および学習方法の検討を行う必要がある。予備実験より、極性を持つ単語が近くに配置される傾向があることは分かっているが、必ずしも正確である訳ではない。出現頻度や文章での利用形態による影響についてさらに詳細な議論をする必要がある。

4. おわりに

ニューラルネットワーク言語モデルを用いたセンチメント辞書の改良手法について提案した。評価実験より、半教師学習の閾値を調整する手法に比べ、人手で作成した辞書に近い辞書を構築できることが分かった。

作成された辞書はまだ人手で作成したものとは大きく異なっている。これを解決するために、半教師学習による辞書作成手法の改良、不要な単語を辞書から削除するためのニューラルネットワーク言語モデルの改良などを行う必要がある。

参考文献

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G, and Dean, J.: Efficient Estimation of Word Representations in Vector Space, Proceedings of Workshop at ICLR(2013)

[Yanagimoto 13] Yanagimoto, H.: Sparse Neural Network Language Model, Proceedings of IS2014(2014)