NLP2019
Poster
P7-24

対義語対の差分ベクトルを使用した評価極性辞書の拡張

川島寛乃† 松田 寛‡ 毛利 研‡

† 慶應義塾大学環境情報学部 ‡ 株式会社リクルート Megagon Labs, Tokyo, Japan

本研究では極性既知の単語とそれに類似する単語ベクトルをもつ拡張候補リストから、反対極性の単語を除外するための方法として極性既知の単語と対義関係にありかつ反対極性が既知の単語とのベクトルの差分に着目した方法を提案する。

背景

特定のドメインにおける評価極性辞書の拡張において
高品質な辞書を**効率的に**構築する方法の検討が必要

	コスト	品質
単語に対する評価極性の付与を人手で行う場合	×	○
既存の辞書に含まれる極性既知の単語を元に自動で追加する単語を収集し, 辞書の拡張を行う場合	○	△

課題点: 評価極性を1単語ごとに人手で付与する方法に比べて追加する単語を効率的に獲得することができるが, **対義語などの逆極性の適さない単語**も含まれてしまう可能性がある

関連研究

1. 評価極性辞書の自動拡張
- ブートストラッピング法を使いた評価極性辞書の自動拡張 [鳥倉, 2004]
- 金融データに特化した評価極性辞書の拡張 [五島ら, 2015]

2. 分散表現を用いた単語の極性の判定
- 分散表現から極性を分類する分類器を訓練 [佐藤ら, 2016]
- 極性を考慮した分散表現を学習 [中村ら, 2018]

本研究では分散表現を用いた評価極性辞書の拡張として対義語も分散表現の類似度が高くなりやすい点に対処した方法を提案

提案手法

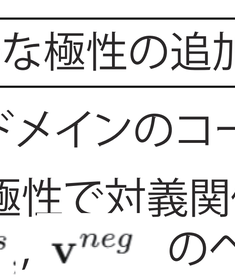
対義語対の差分ベクトルを用いた評価極性辞書の拡張

positiveな極性の追加候補単語を獲得する方法

1. 対象ドメインのコーパスを用いて単語の分散表現を学習
2. 反対極性で対義関係にある2単語の単語ベクトル $\mathbf{v}^{pos}$, $\mathbf{v}^{neg}$ のペアを**対義語対**とする
3. 対義語対の差分を**差分ベクトル** $\mathbf{d}^{+}$とする
$$\mathbf{d}^{+} = \mathbf{v}^{pos} - \mathbf{v}^{neg}$$
4. $\mathbf{v}^{pos}$ に差分ベクトル $\mathbf{d}^{+}$を加え $\mathbf{v}^{\hat{pos}}$ とする
$$\mathbf{v}^{\hat{pos}} = \mathbf{v}^{pos} + \mathbf{d}^{+}$$
5. $\mathbf{v}^{\hat{pos}}$に対して**類似度が高い単語ベクトルを持つ上位n件の単語**を拡張した辞書への追加候補語として獲得

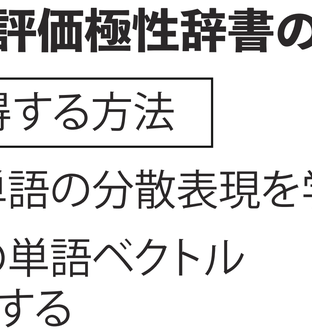
ベースライン(Baseline)

類似度の高い単語の中には類似する文脈で出現する逆極性の単語も含まれる



提案手法(Ours)

差分ベクトルを加えることでよりpositiveな極性を強めたベクトルを算出



逆極性の単語を除外した追加候補語を獲得

- Positiveの極性既知単語
- Negativeの極性既知単語
- ○ 極性不明の単語
- ⇐ 差分ベクトル
- 追加候補語とする範囲

※ Negativeの評価極性辞書への追加単語を得る場合は同様の方法で, 差分ベクトルに $\mathbf{d}^{-} = -\mathbf{d}^{+}$ を使用

実験

分散表現の学習コーパス
- じゃらんnetの口コミデータの一部(**約1600万文**)を使用
- 語彙に含める単語の最低出現回数を3回に制限

対義語対
下記の手順で得られた**94個**の単語ペアを使用
1. 東北大学の日本語評価極性辞書より極性単語を取得
2. 1の極性単語のうちJUMAN辞書における意味項目内に反義情報が存在し対義関係が明らかな単語のペアを作成
3. 2の単語ペアのうち極性が異なるペアを対義語対とする

実験・評価方法

1. **追加候補語の獲得**: Baselineおよび提案手法を用いて追加候補語を獲得(1つのシードに対して類似度の高い上位100件を獲得)
2. **獲得した追加候補語の極性について**: positive/negativeが明らかな単語と極性未知の単語の数をそれぞれ算出
3. **人手のラベル付与実験**: 2の極性未知の単語に対して3人の被験者を用いてpositive/negative/neutralの3値を付与してもらうラベル付与実験を行い, その結果付与されたラベルごとの単語数をそれぞれ算出

実験結果

獲得した単語の結果

表1. 追加候補語のうち極性についてpositive/negativeが明らかな単語と極性未知の単語の数の比較

表1.1 Positiveの追加候補語の極性

極性	Baseline	Ours
Positive	893	895
Negative	488	242
極性未知	3,798	3,974
合計	5,179	5,111

表1.2 Negativeの追加候補語の極性

極性	Baseline	Ours
Positive	161	89
Negative	913	839
極性未知	2,587	2,683
合計	3,661	3,611

提案手法を用いた場合, 逆極性の極性既知語の数が減少

極性未知の追加語としてpositive 3,974個, negative 2,683個の単語を獲得

追加候補語の中の極性未知語に対するラベル付与実験の結果

表2. 3人の被験者に協力してもらい定性評価の評価ラベルを付与した結果.
(2人以上が同じ極性ラベルを選択した単語にのみ評価ラベルを付与)

表2.1 Positiveの追加候補語として獲得した単語について

	Baseline	Ours
評価単語数	710(100.0%)	626(100.0%)
逆極性ラベルの単語数	108(15.2%)	63(10.1%)

表2.2 Negativeの追加候補語として獲得した単語について

	Baseline	Ours
評価単語数	637(100.0%)	686(100.0%)
逆極性ラベルの単語数(割合)	55(8.3%)	44(6.1%)

Positiveの追加候補語・negativeの追加候補語ともに, 提案手法を用いた場合に極性の一致率が向上.

提案手法を用いた場合には極性既知の逆極性の候補語・逆極性のラベルが付与された候補語を除外できている
positiveのラベルが付与された単語
negativeのラベルが付与された単語

獲得した追加候補語の一例

表3. 対義語対「良い, 悪い」を用いた場合の追加候補語100件のうち類似度上位30件の一覧.
(極性既知語についてはpositive: ●, negative: ×と表現)

(a) Baseline.

極性	単語	類似度
●	良い (元の単語)	1.0
	よい	0.952
●	いい	0.849
●	イイ	0.728
●	素晴らしい	0.69
●	すばらしい	0.66
	よく	0.596
●	良好	0.59
●	気持ちいい	0.56
●	抜群	0.559
	グッド	0.558
●	最高	0.554
●	好い	0.553
●	嬉しい	0.552
×	悪い	0.548
	◎	0.538
	好	0.534
●	心地よい	0.53
	うれしい	0.52
	GOOD	0.516
	ヨカッタ	0.504
	good	0.504
●	バツグン	0.503
	上々	0.5
	バッチリ	0.498
×	イマイチ	0.496
	ありがたい	0.493
	今一つ	0.491
×	今ひとつ	0.489
●	気持ち良い	0.486
×	いまいち	0.486

(b) Difference vector.

極性	単語	類似度
●	良い (元の単語)	1.0
	よい	0.855
●	いい	0.749
●	素晴らしい	0.648
●	イイ	0.635
●	すばらしい	0.618
●	嬉しい	0.602
●	気持ちいい	0.58
●	最高	0.57
●	気持ち良い	0.564
●	素敵	0.558
	うれしい	0.558
	ありがたい	0.549
●	ステキ	0.547
●	心地よい	0.534
	グッド	0.526
	◎	0.512
	気持ちよい	0.508
	GOOD	0.507
	good	0.494
●	有難い	0.492
●	すてき	0.487
	よく	0.476
●	良好	0.469
●	抜群	0.468
●	満足	0.465
	好	0.463
	サイコー	0.462
	ヨカッタ	0.456
●	快適	0.444
●	ナイス	0.442

まとめ・今後の展望

- 差分ベクトルを用いることで逆極性の単語を除外した評価極性辞書への追加候補語の獲得する手法を提案
- 10個のシード語を用いて1000語の追加候補語を得たとき, 極性未知の単語のうち逆極性のラベルを付与された単語数の割合がpositiveで**108語から63語に減少。**
- **今後の展望**: 対義語対として使用する極性既知語の選出や類似語を獲得する際の閾値の探索