構文片の改良と評判分析への適用

瀧川 和樹　　山本 和英
長岡技術科学大学 電気系
{takigawa, yamamoto}@jnlp.org

1　はじめに

現在、言語処理の研究にはいくつかの処理単位が用いられている。使用される処理単位は対象にするタスクによって違うが、そのほとんどが単語集合やn-gramである。しかし、単語集合ではその単語が対象内でどのように使われているかを正確に知ることはできない。またn-gramでは言語的に意味の繋がらないものが大量に生成され不要なデータが多くなる。このような問題を解決できる処理単位として構文片がある。構文片とは意味のある要素を処理単位としたものである。現在は、構文解析の結果から係り受け情報を取得し、そこから修飾節と被修飾節の対を抽出したものを構文片として使用している。構文片は構文解析を行った結果から生成されるため、単語集合やn-gramなどに比べて要素自体に意味を持たせることができる。しかし構文片は文節の対であるため、抽出される要素の数が非常に多くなる。またこのような方法で生成される文節対のなかには、「こと-が ⇒ ある」のように意味を持たない文節対も抽出されてしまう。そこで本稿では、現在の構文片の実装が持つ問題を解決するための手法を提案する。そして改良した構文片を用いて評判分析に適用させ、その有効性を調査する。

2　関連研究

藤村ら [3] は、評判分析の処理単位として文節n-gramを使用した。文節n-gramは構文片と似た要素が抽出される。しかし、この処理単位では単純に隣接している文節の連続しか取得することができず、意味のない要素が取得される。構文片は係り受け関係の情報を用いているため、意味のある文節対を処理単位として用いることが可能である。

また、青木ら [1] は構文片を用いて評判分析を行った。しかし、他の処理単位に比べ特に再現率が下回る結果となっていた。これは、構文片という文字列の長い要素を処理単位とすることで対応できる評価表現が減少し、分類できなかった文が多く存在することが主な原因である。本稿では、この問題にも対応できる手法を提案している。

3　構文片

構文片とは意味のある要素を処理単位とすることを目的とした、修飾節と被修飾節の対からなる要素である。つまり係り受け関係の情報さえあれば抽出できるため誰でも容易に扱うことができる。また、他の処理単位と同じように統計情報がとりやすいなどの特徴をもつ処理単位である。

構文片は修飾節と被修飾節それぞれの性質により5種類に分類される。以下にその種類を示す。

・格フレーム:名詞 (-格助詞)⇒ 述語
e.g. 未来-が ⇒ 明るい
・副詞修飾:副詞 ⇒ 述語
e.g. とても ⇒ めんどくさい
・名詞修飾:名詞-の ⇒ 名詞
e.g. 彼-の ⇒ かばん
・動詞修飾:動詞 ⇒ 名詞
e.g. 走る ⇒ 車
・形容詞修飾:形容詞 ⇒ 名詞
e.g. おいしい ⇒ ごはん

しかし、現状の構文片には以下の問題がある。

1. 要素数が他の処理単位に比べ多くなるため、過疎性の問題が発生する
2. 要素の文字列が他の処理単位より長く、辞書として扱う場合取得できる要素が少なくなる
3. 構文解析の結果をそのまま使用すると、意味を持たない文節対も抽出されてしまう

本稿ではこれらの問題を解決するために2つの手法を提案する。

4 提案手法

4.1 同類表現の統一

構文片には、別の素性として扱われているが意味の似ている表現 (同類表現) が存在する。具体例を例 1 に示す。

例 1

ケーキ-が ⇒ おいしい / おいしい ⇒ ケーキ

これら 2 つの表現は、構造は違っていてもそれぞれが示す事象の意味は類似していると言える。そこで同類表現の統一を行う。具体的には、同じ内容語を持つ格フレームと形容詞修飾の構文片群、もしくは格フレームと動詞修飾の構文片群を同類表現として統一する。ここでの統一とは、例えば統計をとるときにこの定義にあてはまる構文片群はすべて同じものとし、出現頻度を合計するという意味である。

同類表現を統一して扱うことで、以下の効果が期待できると考える。

1. 統計をとる場合、過疎性の問題を軽減できる
2. 構文片を辞書のように扱う場合、表層表現が違っていても同類表現であればそれを手がかりに検索が可能となる

4.2 形式的内容語の対処

現在の構文片は修飾節と被修飾節の対であり、構文解析結果から係り受け関係を持つ文節の対を取得することで抽出している。しかしこの方法で抽出すると、意味を持たない文節対を取得することがある。例えば、「とても満足することができる」という入力を構文解析すると (1) とても ⇒ 満足する (2) 満足する ⇒ こと (3) ことが ⇒ できるの 3 つの文節対が抽出される。このとき (2)(3) の文節対は、それ単体では意味を持たない。これは「こと」という単語が形式上内容語に分類されているために発生する。しかし「こと」という単語はそれ単体では意味を持たず、実質的には機能的表現に分類されるべきである。これらの文節対を構文片とすることは「意味のある要素を処理単位とする」という本来の目的とは外れてしまう。

そこでこのような単語を含む文節は、直前の内容語に対する機能表現として扱う。上記の例にあてはめると、(1) とても ⇒ 満足する (2) 満足すること-が ⇒ できるという文節対に整形することで意味のない要素を省き、本来の目的に沿った構文片を抽出することができる。

本研究ではこのような単語を人手で収集し、「形式的内容語」と定義した。収集した形式的内容語を以下に示す。

形式的内容語

こと, ところ, とき, 内, 部, 前, 後, 割に, なる

5 評判分析への適用

改良した構文片の有効性を調査するため、評判分析に適用させる。評判分析の対象は文とし、1 文を肯定・否定・その他に分類する。分析手法は青木ら [1] を参考にした。以下に手法の詳細を示す。

5.1 種辞書の作成

人手で用意した肯定文・否定文を教師データとして構文解析する。そして解析結果から修飾節と被修飾節の対を取得する。このとき、記号や被修飾節にある助詞・助動詞は削除する。次に得られた各構文片に極性スコアを与える。極性スコアの計算には藤村ら [2] の手法を使用した。

$$score(p_i) = \frac{P(p_i) - N(p_i)}{P(p_i) + N(p_i)} \quad (1)$$

$$(-1 < score(p_i) < 1)$$

ここで p_i は構文片、$score(p_i)$ は p_i の極性スコア、$P(p_i)$ は肯定文内での p_i の出現確率、$N(p_i)$ は否定文内での p_i の出現確率を表す。得られた極性スコアと構文片を種辞書として用いる。

5.2 辞書の拡張

種辞書だけでは対応できる表現の数が少ないため、辞書の拡張を行う。まず種辞書を用いて大規模コーパスから肯定文と否定文を取得する。そして取得された肯定文・否定文を新たな教師データとして使用することでさらに辞書を拡張させる。

5.3 文分類

文の極性は文中に出現する評価表現で決定すると仮定する。つまり本手法の場合、文中に出現する構文片の極性によって肯定文か否定文かに決定される。そこで、作成した辞書をもとに文に極性スコアを与え、肯定・否定・その他に分類する。文に与える極性スコア

は、作成した辞書の極性スコアから総和をとることにする。

$$s\_score(S) = \sum_{p_i \subset S} score(p_i) \quad (2)$$

ここで S は分類対象となる文、$s\_score(S)$ は文 S に付与する極性スコア、p_i は文 S 中の構文片、$score(p_i)$ は構文片 p_i 極性スコアを表す。$s\_score(S) > 0$ のとき肯定文、$s\_score(S) < 0$ のとき否定文、そして $s\_score(S) = 0$ のときはその他とする。

5.4 評判分析に本手法が有効と考えた理由

評判分析を用いた理由として以下の 2 点があげられる。

(1) 辞書を用いる手法である
(2) 種辞書の作成時に、統計情報を用いている

辞書を用いて極性スコアを文に与えるという手法であるため、同類表現の統一により辞書の検索能力の向上が期待できる。また種辞書の作成時に出現確率を用いているため、統計情報が必要となる。この点も同類表現の統一により従来の構文片よりも過疎性を軽減できると考える。さらに辞書内には意味の持たない表現は少ないほうが良いといえる。この点は形式的内容語を対処することにより改善できる。

以上の点から、評判分析に本手法を適用した。

6 評価実験

教師データとして、人手により分類した肯定 1,966 文・否定 1,019 文の計 2,985 文を用意した。また、辞書拡張用の大規模コーパスには約 210,000 文を用意した。教師データ・大規模コーパスともに、Yahoo!API(2) を利用して取得した Yahoo!ショッピングレビューから作成した。そして教師データを 5 分割し、1 つをテストデータ、残りを学習データとして評価を行った。構文解析には構文解析器 Cabocha(1) を用いた。提案手法の有効性を調査するため、以下の手法を用いて評判分析を行った。

(1) 同類表現の統一のみ
(2) 形式的内容語の対処のみ
(3) (1) と (2) を組み合わせた手法

ベースラインとして、提案手法を使用しない従来の構文片でも同様の実験を行った。

7 実験結果および考察

7.1 実験結果

文分類の結果を表 1 に示す。この結果から、従来の構文片に比べすべての手法で適合率の向上が確認できた。特に同類表現を統一したことで再現率も同時に向上させることができた。

表 1: 結果比較

処理単位	再現率 (%)	適合率 (%)
(1) 従来の構文片	47.1	75.5
(2) 同類表現のみ	49.8	77.1
(3) 形式的内容語の対処	44.6	77.3
(2)+(3)	47.7	78.7

7.2 同類表現の統一

同類表現を統一に扱うことで、再現率・適合率ともに従来の構文片よりも高い結果を出すことができた。この手法の目的は、辞書作成時における過疎性問題の解消と辞書検索範囲の向上である。これらの点が有効に働いていたかを考察する。

過疎性問題の解消

種辞書作成時に過疎性が解消しているとすれば、従来の構文片を用いて作成するよりも頑健な辞書が作成できると考える。そこで、従来の構文片で作成した辞書と本手法で作成した辞書を入れ替えて文分類を行った。文分類の手法自体にはどちらも従来の構文片を用いた。その結果、再現率・適合率ともに同じ値となった。このことから、同類表現を統一しても過疎性を改善できると証明できなかった。この原因として、構文片の極性スコアはもともと+1 や-1 を付与したものが多く、同類表現を統一してもそのスコアがかわることが少なかったと考える。

辞書検索範囲の拡大

辞書を拡張する際には、種辞書を用いて大規模データを分類することで新たな教師データを獲得する手法を用いた。このとき、同類表現を扱うことで従来の構文片よりも約 14,000 文 (約 5.7%増加) 多く新しい教師データを獲得することができた。この結果から、同類表現を統一することで従来の構文片よりも多くの文に極性スコアを付与でき、拡張辞書の規模が増加したと言える。大規模コーパスから取得した教師データの数を表 2 に示す。

表 2: 大規模コーパスから取得した教師データの数

処理単位	取得できた新しい教師データ (文)
従来の構文片	246,477
同類表現を使用	260,438
差分	13,961

7.3 形式的内容語への対処

今までの構文片の抽出法では意味のない文節対も同時に抽出してしまう問題を、形式的内容語への対処をすることで解決しようと試みた。結果を観察すると、従来の構文片による手法では、教師データの偏りにより偶然に正解となった文が存在した。例えば、従来の手法では「なると ⇒ 思う」という要素が肯定表現として辞書に登録されている。これは、教師データの肯定文内に「なると ⇒ 思う」という表現が多く存在したことが原因である。しかしこの要素は直前の表現によって極性が変化する表現である。よって、「なると ⇒ 思う」という表現により正解を出力できたとしても、それは表現の意味を正しく扱えているわけではなく、偶然に正解しただけである。一方、本手法では「邪魔に-なると ⇒ 思う」が否定表現、「プレゼントに-なると ⇒ 思う」が肯定表現としてそれぞれ辞書に登録されている。このため、本手法では従来の手法よりも正確に表現の極性を扱ったうえで文分類を行えていると言える。

また、結果を見ると再現率が減少してしまっているが、これは同類表現を統一する手法と組み合わせることで、従来の構文片と同程度の再現率を保てており、さらに適合率も向上している。

7.4 その他の処理単位との比較

本手法とその他の処理単位を比較した。比較対象として、単語集合、単語 2-gram, 単語 3-gram を用意した。各処理単位の結果を表 3 に示す。

表 3: その他の処理単位の結果

処理単位	再現率 (%)	適合率 (%)
単語集合	57.0	57.0
単語 2-gram	78.8	79.9
単語 3-gram	75.3	78.0
同類表現のみ	49.8	77.1

単語集合に対しては再現率・適合率ともに上回ったが、逆に単語 2-gram、単語 3-gram に対してはどちらも劣る結果となった。特に再現率の値は、本手法の中で最も良い手法 (同類表現の統一のみ) と比較しても 20 ポイント以上差がついている。このことから、本手法では従来の構文片よりは分類できる文の数は多くなったものの、単語 2-gram、単語 3-gram に比べると少ないといえる。つまり、本手法はまだ他の処理単位に比べ辞書と一致する表現を集めきれていないと言える。この問題は、教師データを増やして辞書の規模を拡大することが最も単純な解決策である。しかし、教師データを増やすにはコストがかかるため、今回の同類表現をまとめた処理をさらに拡張することが望ましい。

8 おわりに

本研究では構文片の改良のため、同類表現を統一して扱う手法と、今まで抽出された意味を持たない文節対を適切な形に整形する手法を提案した。そして改良を行った構文片の有効性を検証するため、評判分析に適用させた。その結果、従来の構文片よりも適合率・再現率ともに向上し、本手法の有効性を検証することができた。しかし一方で、単語 2-gram や 3-gram よりも大きく再現率が劣る結果となった。今後は他の処理単位に特に劣っている再現率を向上させることが課題である。

使用した言語資源及びツール

(1) 構文解析器 CaboCha ,Ver.0.53, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/~taku/software/cabocha/

(2) Yahoo!API ,http://developer.yahoo.co.jp/

参考文献

[1] 青木優, 山本和英. 構文片を用いた分野の同定を必要としない意見・評判情報抽出. 電子情報通信学会 技術研究報告, 言語理解とコミュニケーション研究会, 「主観表現処理の最前線」シンポジウム, NLC2007-88, pp. 7–12, 2008.

[2] 藤村滋, 豊田正史, 喜連川優. Web からの評判および評価表現抽出に関する一考察. 情報処理学会研究報告, 2004-DBS-134(II)-63, Vol. 72, pp. 461–468, 2004.

[3] 藤村滋, 豊田正史, 喜連川優. 文の構造を考慮した評判抽出手法. 電子情報通信学会第 16 回データ高額ワークショップ, pp. 57–60, 2006.