係り受け情報に基づくゼロ代名詞推測手法の検討

幸田　翼†　　木村　昌臣†

芝浦工業大学工学部情報工学科†

1. はじめに

近年，Web 上には製品の評価サイトやレビューを行うブログ等が増加し，企業にとって有益な情報が数多く見られるようになった．そうした背景を基にテキストマイニング技術を用いた評判情報を抽出する研究が多くなされている[1]．評判抽出の手法として，「価格→高い」のように，対象となる語とそれを評価する語との対応関係を求める事が多いが，評価語の対象となる語が省略されている場合には，評判情報が書かれているにも拘らず正しく評判抽出が出来ない．本研究では評価語の対象を単語間の係り受け情報に基づいて推測し，より多くの評判情報を抽出する手法を提案する．

2. 既存研究

対象となる語が省略されている場合の評判情報抽出問題について，小林らはこの問題と照応解析における照応詞と先行詞の同定やゼロ代名詞の推測問題との類似性に着目し，省略された対象の推測に照応解析手法を適用する方法を提案した[2]．しかし，照応解析を行う際に辞書情報として使用する単語間の係り受けによる評価語と評価対象の共起関係を抽出する為には大量の学習データを用意する必要があるが，それらの共起関係を完全に網羅することは困難である．その為，共起関係を完全に学習する必要のない辞書情報を作成することが望ましい．そこで，本研究では単語間の係り受け関係の情報をネットワーク構造で保持し，より柔軟な方法で省略された評価対象の推測を行う手法を検討する．

3. 提案手法

主語となる名詞とその名詞に対し係り受け関係がみられる評価語（動詞・形容詞・形容動詞）をそれぞれノードとし，係り受け関係がある場合にノード間をエッジで繋いだネットワークを作成する（以下，係り受け関係ネットワークと呼ぶ）．日本語は直前に使用した単語を省略する傾向にある事を考慮すると，主語を伴わない評価語より前に出現する名詞がゼロ代名詞の先行詞（評価の対象）である可能性が高いと考えられる．そこで，評価語と評価対象の候補である複数の名詞の内，係り受け関係ネットワーク上で直接接続している名詞（共起関係となる名詞）があればそれを評価対象として推測し，ない場合は他のノードを経由して接続している名詞ノードを探すことで評価の対象を推測する．これにより評価語との共起関係が学習されていなかった評価の対象を抽出できる為，評判情報抽出の再現率が向上することが見込まれる．

3-1. 係り受け関係ネットワーク

評価サイトからレビューテキストを読み込み，CaboCha[3]を用いて係り受け解析を実行する．文節中に格助詞「が」や係助詞「は、も」を伴う名詞が出現し，その文節が評価語の含まれる文節に係っている場合に「対象，評価語」をペアとして抽出する．図1はペアを基にして作成したネットワークの一部を Pajek[4]を用いて可視化したものである．

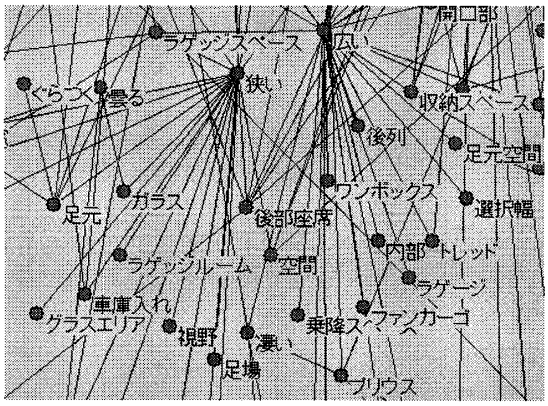

図1 係り受け関係ネットワークの例（一部）

ネットワークは名詞とそれ以外の単語との二部グラフとして形成される．

3-2. ゼロ代名詞推測の方法

評価対象の省略がみられた場合，係り受け関係ネットワーク上でその評価語を始点とし，文章中で評価語の前にある名詞全てを評価対象の候補として終点に定める．始点から終点までの

A Study of Japanese Zero pronoun resolution based on dependency information.

†Tsubasa Kouta, Masaomi Kimura

Shibaura Institute of Technology

距離情報をダイクストラ法で計算し，距離が最短である単語をゼロ代名詞の先行詞であると推測する．但し，距離が長くなると経路上に含まれる評価語や名詞に，始点となる評価語と関係が無い語が含まれてしまう為，結果として評価語と評価対象として推測する名詞との関係性が小さくなる恐れがある．その為，最大で5ステップまでの距離の範囲で探索を行う．

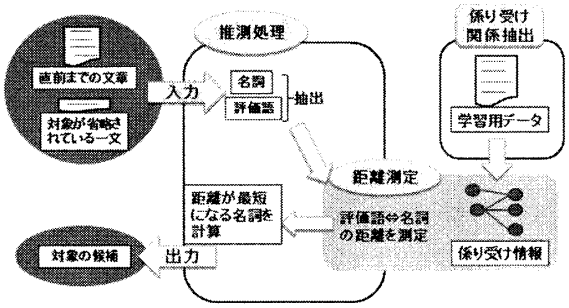

図2 ゼロ代名詞の推測手順

4. 実験

係り受け関係ネットワーク作成の為の学習用データとして価格.com[5]内のトヨタ車現行モデル48車種に対するレビュー文を取得したところ，2368個のノードからなるネットワークが作成された．この係り受け関係ネットワークを利用し，ゼロ代名詞の推測を行う．推測が必要である文章は，価格.com内の車に関するレビュー記事から40件(ホンダ15件，日産15件，トヨタ10件)取得し，推測すべき名詞が何であるかをあらかじめ主観で定めた．

システムに与える情報は評価対象が省略されている一文とその直前の文章とする．但し直前の文章とは，評価語の含まれる文から2文前までとし，それ以前に段落が存在する場合はその段落までと定義する．このように定義した理由は，直前の文を多く取りすぎると評価の対象となる語の候補が多くなりすぎてしまい，推測の精度が減少する恐れがあることと，ゼロ代名詞の先行詞はゼロ代名詞の位置から遠くない距離にある事が多い為である．

実験の評価として，推測を行った文章においてあらかじめ決定した評価語に対する正しい評価対象と，実験で得られた推測結果を照らし合わせて再現率，適合率，F値を求めた．

5. 結果・考察

学習データが共起関係のみである場合と係り受け関係ネットワークを利用した場合とで結果がどのように変わるかを比較し，評価を行った．表1に実験結果を示す．

表1 実験結果

	共起関係	係り受け関係ネットワーク
再現率	0.379	0.867
適合率	0.944	0.706
F値	0.541	0.778

表1より，学習データにおいて共起関係がある場合のみの推測結果よりも，係り受け関係ネットワークを利用した推測結果の方が良い再現率を得られたことが分かる．これは，係り受け関係ネットワーク上で評価語に対し接続している名詞ノードと，その名詞ノードが接続しているその他の評価語ノードから接続している名詞ノードとの使用例が近い為であると考えられる．係り受け関係ネットワークを利用した場合，適合率は共起関係のみの場合より低くなったが，F値は大きくなった為，適合率の低さを上回る高い再現率を得ることが出来たと考えられる．この結果から，本手法を用いれば，対象と評価語の対応が必ずしも網羅されていない学習データでも評価対象を推定することが可能であることが示された．

6. まとめと今後の課題

本稿では学習データに係り受けによる共起関係が網羅されていない場合でも評価語に対する評価対象の推測が可能である手法を検討した．その結果，再現率は大量の学習データを用いた既存手法と同等の結果が得られた．しかし，ネットワークを通して複数の単語を推測結果とすることで評価語と関係の無い名詞が含まれてしまうことがあった為，適合率に若干の低下が見られた．この問題の解決策として，係り受け関係ネットワークを学習する際に，単語ノードに出現頻度による重みや，使われやすい分野の情報を付与することで評価語と関係性が小さい名詞を候補から除外するなどの方法が考えられる．

参考文献

[1] 奥村学，他：意見分析エンジン-計算語学と社会学の接点-，コロナ社，(2007)．
[2] 小林のぞみ，他：照応解析手法を利用した属性－評価値対および意見性情報の抽出，言語処理学会年次大会発表論文集 Vol.11th, pp. 436-439(2005)．
[3] CaboCha/南瓜：http://chasen.org/~taku/software/Cabocha/
[4] Pajek：http://vlado.fmf.uni-lj.si/
[5] 価格.com：http://kakaku.com/