機械学習を用いた日本語ゼロ代名詞照応関係の同定

吉野圭一、竹内和広、松本裕治
奈良先端科学技術大学院大学 情報科学研究科
E-mail: {keich-yo, kazuh-ta, matsu}@is.aist-nara.ac.jp

1　はじめに

文章中に表れた照応詞が何を表しているかを正しく解釈するためには、広範な言語的情報や背景知識を用いる必要があり、それが高度な自然言語処理システムを構築する上での大きな課題となっている。

これまで照応解析においてはセンタリング理論 [1] を用いる手法や、解析規則を人手で作成する手法 [2, 3] などが提案されてきた。これらの手法は現在までに大きな成果をあげているが、取り扱う問題が限定されていたり、解析規則の作成に多大なコストが必要であるという問題があった。

本研究では日本語テキスト中に出現するゼロ代名詞に対して、照応関係の解析規則を機械学習によって獲得することを目的とする。機械学習を用いる手法は、学習コーパスの作成にコストがかかるという問題があるが、学習コーパスが用意できれば人手を用いるよりも効率的に解析規則を獲得することが期待できる。そこで、日本語の新聞記事に対して人手でゼロ代名詞や照応関係の情報を付与して学習コーパスを作成し、それをもとに機械学習を行い、その手法の有効性を検討する。新聞記事に対しては、既に形態素の情報等が付与された大規模なコーパスが整備されており、最小限の情報の付与で機械学習に用いることが可能である。

2　本研究のアプローチ

本節ではゼロ代名詞照応解析における本研究のアプローチおよび設定したタスクについて述べる。

一般に機械学習で良好な精度を得るためには多大なデータが必要となるが、まず小規模なコーパスで手法の有効性を検討し、その結果をコーパス作成にフィードバックしていくことを考えた。つまり、学習された解析器による解析を行い、そこから得られた結果を元に人手による修正を行っていくことで、より規模の大きなコーパス作成への支援を行うことが出来る。規模の大きなコーパスが作成されればよりいっそうの精度向上が期待される。

2.1　タスクの設定

ゼロ代名詞の照応は、先行詞が同一文章内に存在するもの (文章内照応) とそうでないもの (文章外照応) の大きく二つに分けることが出来る。このうち文章外照応は、対話文においては頻繁に出現するが、本研究で対象としている新聞記事においては出現頻度が少なく、また解析のためには高度な推論や背景知識が要求されるため、現時点では考慮しない。

文章内照応はさらに先行詞が同一文内に出現するもの (文内照応) とそうでないもの (文脈照応) の二つに分類することが出来る。

文内照応が起こるのは、複文を単文に分割した際に初めて出現するゼロ代名詞に対してのみであり、それ以外のゼロ代名詞は文内照応とはならない。したがって、文内照応はそれ以外の照応と区別して考え、異なる手法を適用することが可能である。文内照応に限定すれば、中岩ら [3] では高い精度での解析が可能であることが報告されている。

本研究ではこれらの点を踏まえて、文脈照応における適切な先行詞の同定を取り扱う。

2.2　Support Vector Machine

本研究ではSupport Vector Machine(SVM)[4] を用いて学習を行う。SVM は学習サンプルと分類境界の間隔を最大化するような戦略に基づいて二値分類を行う学習アルゴリズムである。SVM は学習データの次元数 (素性集合) に依存しない極めて高い汎化能力を持ち合わせており、さらにKernel関数を導入することによって非線形のモデル空間を仮定したり、複数の素性の組合せを考慮して学習を行うことができる。このような点においてSVM は他の手法に対して優位性があるため、本研究では学習アルゴリズムにSVM を採用した。

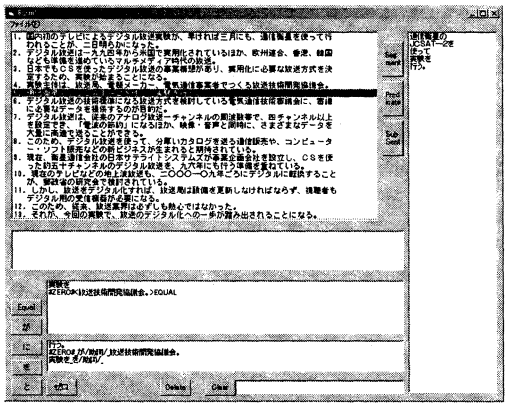

図 1: データ作成支援ツール

3 実験手法

本節では、本研究で提案する手法の有効性を調べるために行った実験について述べる。

3.1 実験コーパス

本研究では実験データとして、京都大学テキストコーパス [7] から無作為に 40 記事を選び、これに人手で以下の情報を付与した。

- 節への分割
- 節中の述語
- 述語の補語となる文節
- ゼロ代名詞の位置および述語との格関係
- (ゼロ) 代名詞の先行詞

コーパスの作成にはデータ作成支援ツールを用いた。このツールは、計算機の扱いに不慣れな人であっても効率的にデータ作成を行えるように設計されたものである。図 1 にこのツールの実行画面を示す。

3.2 素性

学習に用いる素性の選択は極めて重要な問題である。ここでは、学習に用いた素性について述べる。

距離 一般的にゼロ代名詞に対する先行詞は、ゼロ代名詞の近くに出現する。したがって、先行詞を選ぶ際に距離は重要な素性となる。距離に関しては文節、文の 2 種類の距離を学習の素性として用いた。

形式段落 形式段落は話題の境界と一致するといわれている。そこで、先行詞の候補がゼロ代名詞と同じ形式段落内にあるか否かを素性とした。

ゼロ代名詞の格 解析対象のゼロ代名詞の格によって、先行詞の選び方が変わってくると考えられる。そこで、省略されている格が何であるかを素性とした。

先行詞候補の格 センタリング理論では先行詞の候補となる語の格要素が、文の遷移関係を決める重要な要素となる。このことから、先行詞候補が文中で何格で出現しているかは、先行詞を同定するのに重要な情報であると考えられる。そこで、先行詞の候補が何格で出現しているかを素性とした。

固有名詞 先行詞候補に固有名詞が含まれるか、また含まれている固有名詞が解析している記事中で出現頻度の高い固有名詞かを素性とした。

品詞 本実験では、先行詞の候補を探索範囲中のすべての文節としている。そのため、名詞を含まない (先行詞になり難い) 文節も候補となっている。そこで、先行詞と品詞の関係を考慮するため、先行詞候補の品詞を素性とした。先行詞候補の品詞は主辞の品詞とそれ以外の品詞とを分けて素性としている。

共起頻度 ある単語が文中に出現した際に同一文中に出現する傾向の強い単語は、ゼロ代名詞の先行詞になりやすいのではないかと考えられる。そこで、2 つの単語の同一文内での出現頻度を調べて、それを素性とした。

4 年分の新聞記事に対し茶筌 [5] を用いて形態素解析を行い、その解析結果をもとに二つの単語の同一文内での共起頻度を調べた。二つの単語の組合わせは、名詞-名詞、名詞-動詞の 2 種類にわけて、それぞれについての共起頻度を調べた。

格フレームの制約 ゼロ代名詞の先行詞は、ゼロ代名詞を格要素にとる用言と意味的に整合的でなければならない。そこで、日本語語彙大系 [6] の格フレーム辞書および名詞意味分類を用いて、格フレームの制約を満たしているかどうかを利用した。格フレームの制約を調べる際には、述語の主辞、先行詞の候補の主辞、ゼロ代名詞以外の格要素の主辞を用いる。述語の主辞から格フレーム辞書を引き、格要素の主辞から意味分類を調べ、それらが格フレームの制約を満たしているかを素性とした。

語の重要度　記事中で重要度の高い語は出現する回数が多く、また照応もされやすいと考えられる。ここでは、TF-IDF 値を重要度の尺度として用い、先行詞候補の主辞の TF-IDF 値および先行詞候補中に出現する語の中で最も高い TF-IDF 値を素性として用いた。

3.3 評価方法

実験の評価はすべて交差検定によって行う。学習データを記事単位で 10 分割し、1 つを評価データ、残りを学習データとして交差検定を行い解析精度を求めた。精度は正解候補を上位 3 位まで挙げ、その中に正解が含まれているか否かを判定することで求めた。

4 実験結果および評価

ここでは実験結果および結果の考察を行う。

4.1 実験結果

前節で述べた素性を用いて、実験を行った。

最初に、データを 10 記事ずつ増やしていき、それによる解析精度の変化を調べた。学習データ量に対する解析精度の変化をグラフにしたものを図 2 に示す。

次に、各々の素性に対してその有効性を検証するために、その素性を使わない学習データを用いた実験を行った。実験の結果を表 1 に示す。

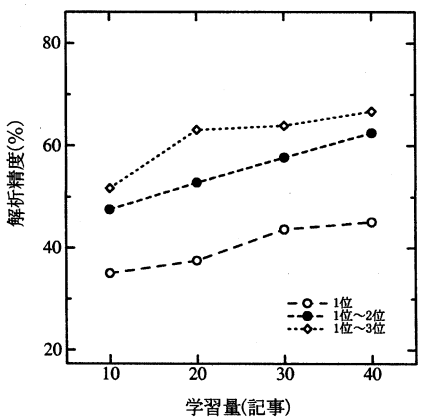

図 2: 学習量と解析精度

表 1: 使用素性と解析精度

除去した素性	正解数 (解析精度 (%)) 1 位	1~2 位	1~3 位
-	54(45.0)	75(62.5)	80(66.7)
距離	−32(10.0)	−47(23.3)	43(30.8)
段落	+1(45.8)	−1(61.7)	0(66.7)
ゼロ代名詞の格	+4(48.3)	−1(61.4)	+1(67.5)
先行詞の格	−9(37.5)	−9(55.0)	+1(67.5)
先行詞の品詞	+3(47.5)	−6(57.5)	−1(65.8)
固有名詞	+3(47.5)	−6(57.5)	−4(63.3)
共起頻度	+6(50.0)	−2(60.8)	0(66.7)
語の重要度	+2(46.7)	−5(58.3)	−4(63.3)
格フレームの制約	0(45.0)	−3(60.0)	+1(67.5)

ゼロ代名詞 120 個を含む 40 記事による実験

数値は全素性を用いた場合との正解数の差分を表す

4.2 評価および素性の検討

ここでは、実験の結果について検討を行う。

4.2.1 学習データ数と解析精度

図 2 からわかるように、学習データの増加に伴い、全体的に解析精度も向上している。この傾向はまだ続くと考えられ、今後学習データが増加すれば、より一層の精度向上が期待される。

4.2.2 学習素性の解析精度への影響

ここでは、表 1 に示した実験結果から、各々の学習素性がどの程度先行詞の同定に寄与しているかを検討する。

距離　表 1 より、距離に関する素性を除いた場合に最も大きく解析精度が減少することがわかる。文外照応の起こるゼロ代名詞の多くは、先行詞が直前の文に出現しており、離れた文に先行詞が出現することは少ない。そのため、先行詞までの距離を考慮しなければ適切な先行詞を選択するのが難しいことがわかる。

形式段落　段落に関する素性を取り除いて解析を行った場合には、あまり精度の変化が見られない。このような結果となったのは、新聞記事の形式段落が比較的小さくまとめられており、形式段落の切れ目と話題の境界が必ずしも一致しないためであると考えられる。本実験では形式段落は先行詞の同定に必ずしも寄与しているとはいえないが、小説文などの形式段落の大きい文章では有効な素性となる可能性がある。

ゼロ代名詞の格　ゼロ代名詞が何格として出現するかによって、先行詞の同定に異なる特徴があると考えられる。しかし、実験結果からはそのような傾向を見ることができなかった。これは、ゼロ代名詞のほとんどがガ格として出現しており、その他の格のゼロ代名詞が出現することが少なく、また実験データの数も少ないためと考えられる。

先行詞の格　先行詞の格に関する素性を取り除くと上位2位までの先行詞候補で精度が大きく減少している。このことから、先行詞を正確に同定するためには、先行詞の候補が何格で出現しているかということが有効な特徴であるといえる。これはセンタリング理論とも一貫性を持っている。

固有名詞・語の重要度・先行詞の品詞　これら3つの素性に関していずれか一つの素性を取り除いた場合、1位の候補においては精度が向上しているのだが、全体としては精度は低下している。このことから、これらの情報は先行詞の候補の絞込みを行う上で有効に働いていると考えられる。

共起頻度　共起頻度に関する素性を取り除いた解析では、精度の減少はほとんど見られず、上位1位においては精度は向上している。このような結果となったのは、共起頻度の問題よりも、共起頻度のとり方に問題があったのではないかと考えられる。共起頻度のデータは新聞記事4年分をもとに作成したが、統語情報を用いず、単純な文内共起頻度を用いているため、本来あまり関係のない単語の共起頻度が高くなってしまうことが考えられる。

格フレームの制約　格フレームの制約に関する素性を取り除いた解析では、あまり精度の変化は見られない。格フレームの制約は先行詞を同定するのに重要な素性であると考えられる。しかし、実験においてその有効性が確かめられなかったのは、述語から格フレーム辞書を引くことが出来ないことがある(体言止など)ことと、意味的に正しい表現であっても格フレームの制約を満たさない例外が多いことによって、完全に格フレームの制約を満たす事例が少なかったためであると考えられる。

5　結論と展望

本研究ではSVMによる機械学習を用いて日本語ゼロ代名詞の照応関係の同定を行った。文脈情報や、言語的な情報を素性として実験を行った結果、新聞記事中に出現する文脈照応が必要なゼロ代名詞の先行詞を、先行詞候補を上位3位まで選択することによって67.5%の精度で同定することが出来た。この結果は先行研究に対して優れているとはいい難いが、今後素性の詳細な検討や、学習のためのコーパスの整備を進めることで、高い精度で照応解析を行なえる見通しが得られた。

また、異なる素性を学習に用いて各素性の効果を検討した結果、以下の知見が得られた。

- ゼロ代名詞と先行詞候補の位置関係は照応関係の同定において極めて重要な素性である
- 先行詞候補が文中で何格として出現しているかは、先行詞の同定において有効な素性といえる
- 先行詞候補の品詞や、文章中における重要度といった素性が先行詞候補を絞り込むのに有効に働いていると考えられる

今後は学習データの整備を進めると共に、より詳細な素性の検討を行っていく必要がある。

謝辞

本研究では、「京都大学テキストコーパス」,「毎日新聞CD-ROM版」,「日本語語彙大系」を利用させて頂きました。関係者各位に感謝致します。

参考文献

[1] Malilyn Walker, Masayo Iida, Sharon Cote "Japanese Discourse and the Process of Centering" Computational Linguistics, Vol.20, No.2 p193–p232 (Jun 1994)

[2] 村田真樹、長尾真 "用例や表層表現を用いた日本語文章中の指示詞・代名詞ゼロ代名詞の指示対象の推定" 自然言語処理, Vol.4, No.1 p87–p109 (Jan 1997)

[3] 中岩浩巳、池原悟 "語用論的・意味論的制約を用いた日本語ゼロ代名詞の文内照応解析" 自然言語処理, Vol.3, No.4 p49–p65 (Oct 1996)

[4] Vladimir N. Vapnik "Statistical Learning Theory" Wiley-Interscience, (1998)

[5] 松本裕治、北内啓、山下達雄、平野善隆、松田寛、浅原正幸 "日本語形態素解析システム『茶筌』version2.0 使用説明書 第二版" 奈良先端科学技術大学院大学, NAIST Technical Report NAIST-IS-TR99008, http://chasen.aist-nara.ac.jp/, (1999)

[6] 池原悟、宮崎正弘、白井諭、横尾昭男、中岩浩巳、小倉健太郎、大山芳史、林良彦 "日本語語彙大系 1 意味体系" 岩波書店 (1997)

[7] 黒橋禎夫、長尾眞 "京都大学テキストコーパス・プロジェクト" 言語処理学会 第3回年次大会, p115–118 (1997)