日本語学習者作文支援のための機械学習による日本語格助詞の正誤判定

大山浩美

奈良先端科学技術大学院大学　情報科学研究科

hiromi-o@is.naist.jp

1. はじめに

第二言語として日本語を学習する人たちは、年々増えており、特に海外でその人数は多くなっている（2,356,745 人 2003 年現在）。近年、年々普及するコンピュータネットワークを使った日本語学習者のための学習支援などの開発がさかんで、その一環として作文教育の支援や研究などが様々な形で行われはじめている。ウェブやネットワークを使った作文･添削システムを目指すためのものとして、ネットワーク型添削支援システム CoCoA[3]や中国語の作文や添削を行うシステム[2]などがすでに見られる。CoCoA は、教師が誤り箇所を指摘し訂正したりするのを支援するシステムである。中国語の作文添削システムでは、添削者がオンラインエディターを使って、文章を添削すると、XML タグにより欠落、余剰、組み合わせの不適を指摘し学習者に返却することができる。この中国語のシステムは日本語に応用することもできると思うが、上記いずれの場合もやはり教師の関与が必要である。今後、ウェブやネットワークシステムを利用することを考えれば、日本語学習者の大量の作文がシステムに流れ込んでくることが予想される。その際、コンピュータを使って日本語学習者の誤用を自動的に判定することができるようになれば、大量の作文を添削することにも対処でき、添削の信頼性、統一性を保つことにも役立つだろうと予測される。もし教師の手を煩わすにしても、その大量の作文を添削する際の負担を減らす助けにもなることが期待される。さらに、オンラインで学生の作文自動添削はもちろん、作文の評価も可能になってくるかもしれない。本研究においては、日本語の正用文と誤用文を区別するために、まず正用文を機械学習により認識することを試みた。

2.日本語学習者の誤用について

日本語学習者がおかす誤用には、表記の間違い、活用形の間違いや語彙選択の間違いなどいろいろな種類のものがある。また、日本語学習者の母語は多岐にわたるので、その誤用の理由も様々である。その多様な誤用の中で、今回はまず格助詞に注目した。その理由は、格助詞の誤りが日本語学習者の誤りの中で最も多く出現するものの中の一つであること、そして、比較的機械的に訂正することが容易ではないかと考えられるということの二点である。例えば、日本語学習者の実際の格助詞の誤用には以下のようなものがある。

- 私＜が＞国際経済＜　＞関して非常に興味を持っています。

→　＜が＞部分を＜は＞に変換、＜　＞部分に“に”を挿入

- 毎年＜に＞、日本と同じく、中国でも一番にぎやかな日は、やっぱり正月だった。

→　＜に＞部分を削除

- 今から、友達＜が＞いっぱい＜を＞つくるように頑張ります。

→　＜が＞を＜を＞に変換、＜を＞を削除

これらの格助詞の間違いには、変換が必要なもの、不必要なもの、挿入しなければならないものなどがある。今回は、格助詞「が」「を」「に」「で」「と」「より」「から」「へ」と係り助詞「は」の 9 つの検出に焦点をあてた。「は」は係り助詞であり、格助詞ではないが、「は」の誤用は日本語学習者の作文に多く見受けられるためここで対象に入れた。

3. 実験方法

今回の実験では、格助詞の正用をとるために毎日新聞の 2003 年度の半年分（888MB、2,394,638 語）を日本語係り受け解析器 CaboCha[1]で解析したものを使用した。その新聞のデータの中で「が」「を」「に」「で」「と」「より」「から」「へ」と係り助詞「は」が使われている文の前後 5 つの単語の表層情報と品詞情報を素性として用い、機械学習でパターンを認識し、正しい格助詞を選択できるかどうかを実験した。機械学習器は、Support Vector Machines[1]（以下 SVM）、TinySVM[2]を使用した。SVM は、2 クラス識別法のひとつであり、データで正例と負例を与えると、パターンを学習し正例と負例を判別するモデルを作る。今回、毎日新聞半年分から10万、20万、30万語のデータを取り出し、それらを元にして正例、負例を判別するモデルを学習させ、1 万、2 万、3 万のテストデータで正しい格助詞を選択することができるかどうかを実験してみた。その際、ひとつの格助詞（例えば、「が」）を正例とし、ほかの 8 つの格助詞（「が」以外）を負例とし、正例（ここでは「が」）をどれくらい正しく見つけられるかをみた。他の8つの各格助詞でも同様の実験を試みた。

4.結果

表1は、各格助詞に対してどれだけ正しく見つけられたかを F 値で示したものである。これをみると、「が」の値が最も高く、「を」、「に」、「と」、「より」、「で」と続いている。「へ」、「から」においては 60%を少し超えた値となっている。30 万語のテストにおいては、「が」が 100%に近い値となり、検出率が高くなっていることを示している。また、他の格助詞においても、学習データの語が多くなると 70%近くの検出率を出していることが見られる。しかし、「は」になると、正用判定が難しく、50%を少し超えたくらいの一番低い値となった。

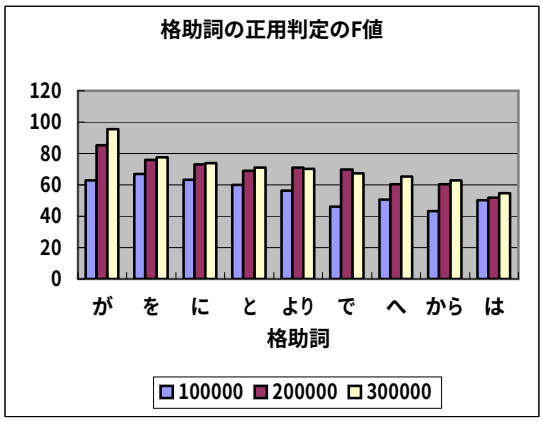

表 1

5.まとめと今後の課題

今回は、新聞における格助詞の使用をモデルとし、格助詞の正用を判定する実験を行った。平均約70%の値で格助詞の正用が検出できることがわかった。今回の実験では、格助詞の素性として単語の表層情報と品詞情報しか取らなかったが、今後、その格助詞を含む文節の係り受け情報を素性として加え、正用判定の値が伸びるかどうかみたいと考えている。それから、実際の誤用文に機械学習したモデルをあて、負例として判定できるかという実験も試みたい。また、他の機械学習手法でも判定値が上がるかどうか実験を試みたい。さらに、将来的な日本語作文指導のためには、誤り箇所を指摘、訂正したり、欠落、余剰、組み合わせの不適を知らせたりするのに加え、様々な誤用文に対し、誤用の種類に応じた誤り検出、訂正、誤りの理由の提示なども視野に入れていかなければならないと考えている。

参考文献

[1] Vladimir N. Vapnik. *The Statistical Learning Theory.* Spinger,1998.

[2]劉松、砂岡和子、浦野義頼、“誤用データ統計機能を備える中国語作文･添削支援システム”、Computer & Education,Vol.20,pp.74-79. 2006.

[3]脇田里子、緒方広明、矢野米雄、“作文教育のためのネットワーク型添削支援システム CoCoA の実践と評価”、Vol.15, No.4, pp270-275, 1999.

[1] http://chasen.org/~taku/Software/CaboCha/
[2] http://chasen.org/~taku/Software/TinySVM