XML 文書からの意味情報の自動推定

— 意味情報の XML 化による日本語助詞の校正システム事例から

中挾知延子

Learning Semantic Information based on XML documents

- Correcting Japanese Postpositional Words with Semantic Representation in XML

Chieko NAKABASAMI

Abstract

This work is a report on the use of a language application using semantic representation in an XML format. The aim of the application is to correct the misuse of Japanese postpositional words, called *joshi*, as they have been used in sentences written by non-Japanese students. First, the sentences written by the students are processed morphologically so that pairs of *joshi* nouns and pairs of joshi verbs are extracted. Next, document data that include the same extracted nouns are picked up from an EDR corpus where these nouns are stored with other pairs of *joshi* that are used as training examples in the correction process. The data are automatically transformed into semantic representations in an XML form. "Association rules" are applied in the correction process to assist in learning how to use the appropriate *joshi* after target nouns. Semantic representation is used to provide suitable explanations of how *joshi* should be used in a sentence.

1. はじめに

筆者の所属する大学の学部においては外国人学生が多く、彼らにとって日本語の学習は必須となっている。筆者を含めて日本語と英語の専門家と共同で、日本語文章校正システムを開発している。外国人学生が日本語を学ぶにあたって苦労する分野の一つに作文があげられるが、その中でも習得が困難なものとして助詞の用法がある。本稿では、助詞の校正に焦点をあてて開発を進めている日本語文章校正システムの開発事例を報告する。

本システムでは、より一般的に、多くの知識源からの情報の集積ができるようにするため、知識源としての自然言語の意味情報を XML 形式 [1]で貯えている。システムにおける校正の方法は、EDR の日本語コーパス [2]を利用して多くの用例を集め、学生が入力した文と照合して、文中の助詞をより尤らしい語に置き換えるものである。この際、尤らしい助詞は、学生の入力文にある名詞と同じ語を含むコーパスの用例から連想ルール（Association Rule[3]）を使って決定される。XML 化される意味情報はコーパスから取り出されたものであり、この XML データからパーザ [4]を適用して柔軟に入力文中の名詞に付与された意味タグが抽出される。名詞の意味タグから、名詞に続く助詞の役割が同定されるので、意味を伴った名詞と助詞の使用例を作成し保持することができる。

本システムは将来、インターネットでの利用を想定しており、学生が自由な時間に作文の校正に利用できる環境を目指している。指導教員が既存の用例データとは別に、独自の作文例をシステムに組み入れる場合、データを XML 化して加えることで、システムに即座に組み込まれ、それ

以降の学習に役立てることができる。また、日本語文章に対応する英文の用例をXML化したデータで持たせることで、英語の文例も提示して、校正結果についての説明をすることができる。このようにして XML という共通の形式でデータを貯えていくことにより、異なる知識源を統合したデータベースが構築できる。

以降の章では、2 章ではシステムの概要について述べる。3 章では意味情報の推定に関する現状を報告し、4 章で結びとする。

2. システムの概要

助詞の校正に必要な情報として、入力文とコーパスにおける名詞と助詞の共起対と、そのときの助詞がどのような意味役割を担っているのかについての情報があげられる。本システムではこのような情報をクラスのインスタンス集合として保持するため、JAVA を用いて開発している。ここでのクラスは名称を「文節」とし、以下の属性を有する。例えば、(1)「11月はコースが始まりました」という文章を入力文とする。助詞の誤用集[5]を参照すると、(1)は誤りがちな用法となっており、校正例として「11月にはコースが始まりました」があげられている。(1)の文からは、インスタンス 1 が作られる。意味情報はコーパスにある用例から抽出されるものとするので、ここでは与えられていない。また、文情報は入力文ならば-1 としてコーパスの用例と区別する。

文節クラス	インスタンス 1
文番号	文情報:-1
自立語(名詞、動詞など)	自立語:11月
助詞	助詞:は
意味情報	意味情報:なし

次にコーパスからインスタンスの自立語属性の語を含む用例を抽出する。用例として(2)のような意味構造データが得られたとする。これらの意味構造データから(2)'のようなXML形式のデータに自動的に変換する。

(2) 夏休みが終わり、11月に推薦入学の試験が終わるまで忘れていた。
[[main 15:忘れ:10eed0][attribute already end][time-to [[main 12:終わ:3cea4c][object [[main 10:試験:2dd49a][modifier 8:推薦入学:0f8647]]][time 6:11月:3bd04a][sequence [[main 3:終わ:3cea4c][object 1:夏休み:101b00]]]]]]

(2)' <sentence><main>忘れ</main>
<attribute>end</attribute>
<time-to><main>終わ</main>

<object><main>試験</main>
<modifier>推薦入学</modifier>already</object></time-to>
<time>11月</time>
<sequence><main>終わ</main>
<object>夏休み</object></sequence></sentence>

これらの XML データからインスタンス 2 が作成される。これらのインスタンスは、XML 形式に変換した用例の意味情報から作成され、もしコーパスとは別にインスタンスを作成し、校正に反映したいときにも、同様の XML 形式のデータを加えることにより実現できる。

インスタンス 2
文情報:1
自立語:11月
助詞:より
意味情報:time

システムは校正結果として、インスタンス 1 の助詞を尤も適切なものに置き換える必要がある。もちろん、入力文の助詞が尤もらしいものであればその必要はないが、なぜそのままで良いのかについての説明は意味情報を基に提示する。ここで尤もらしい助詞を同定するために、Association Rule(連想ルール)を用いている。連想ルールは機械学習で用いられる手法の一つであり、各属性を複数個、条件部と結論部に配置することにより、正例データに潜む属性間のルールを導き出そうとするものである。連想ルールを助詞の同定に用いる場合、属性として自立語、助詞、意味情報ならびに自立語の名詞と共起する動詞を設定する。そしてこれらの属性間で成り立つルールを導き出し、自立語(ここでは「11月」)とその動詞が与えられたとき、共起しやすい助詞や意味情報を同定する。意味情報は校正結果を提示する際の学習者への説明事項として用いられる。

3. 連想ルールによる意味情報の推定

2 章で説明した方法でコーパス中の該当する用例から文節クラスのインスタンスを作成した後、その内容を連想ルール学習の入力とし、意味タグを推定する。入力データはあらかじめインスタンスの集合から内容を抽出して作成する。学習には WEKA[6]を用いている。WEKA は Java のクラスライブラリで構成された機械学習の実行環境である。WEKA に入力すべきデータとして(自立語、助詞、意味情報)の組から成るデータを作成した。このデータはコーパスから抽出されたインスタンス 2 で示した内容を値に持つ。現在システムは開発段階であり、途中結果ではあるが、2 章で示したインスタンス 1 に対して WEKA を用いて適切な助詞を推定させた。出力として得られた連想ルールのリストを図 1 に示す。

1. post=の 135 ==> word=11月 135 conf:(1)
2. sem=modifier 103 ==> word=11月 103 conf:(1)
3. sem=main post=に 93 ==> word=11月 93 conf:(1)
4. sem=main 236 ==> word=11月 236 conf:(1)
5. post=に 198 ==> word=11月 198 conf:(1)
6. sem=time 175 ==> word=11月 175 conf:(1)

図 1 得られた連想ルール

図 1 から、「11月」に続く助詞の候補として「に」があげられ、また、「11月」という名詞は“time”(時間)という意味情報を持つという結果が得られる。この結果を校正例に適用した場合、一例として“「11月」に続く助詞として、「に」にすると時間を表す”というような説明を提示できる。

4. 結び

本稿では、日本語文章における助詞の使用についての校正を行うシステムの開発事例を報告した。そのシステムの中で、知識源としての意味情報を XML 形式で貯えることを行った。システムは意味情報データから助詞の推定に必要な情報を抽出し、連想ルールの手法を用いて、目的の語に続く助詞の推定を行う。意味情報を XML 化することで可読性が増し、異なる知識源からの情報の集積が容易になると考えられる。本システムは、さらに機能を充実させ、インターネットを介した教育システムにする予定である。なお，本研究は『東洋大学平成13年度教材開発共同研究助成』を受けて行われている.

5. 参考文献

1) The World Wide Web Consortium (W3C), Extensible Markup Language (XML), http://www.w3.org/XML/, 1998
2) 日本電子化辞書研究所、日本語コーパス CD-ROM、1996
3) Agrawal, R., Imielinski, T., Swami, A., Mining Association Rules between Sets of Items in Large Databases, Proc. of the ACM SIGMOD Conference on Management of Data, 1993
4) Apache XML Project: Xerces, http://xml.apache.org/, 1999
5) 市川保子、日本語誤用例文小辞典、凡人社、1997
6) Weka Machine Learning Project, WEKA, http://www.cs.waikato.ac.nz/~ml/

中挾 知延子 東洋大学国際地域学部(〒374-0193 群馬県邑楽郡板倉町泉野 1-1-1)