

日本語学習用対話例データベースのための学習項目に基づいた インデクシングの自動化について

馬目知徳 加藤伸隆 神田久幸 伊丹誠 伊藤紘二

東京理科大学 基礎工学研究科 電子応用工学専攻

{manome,peechan,kanda,itami,itoh}@itlb.te.noda.sut.ac.jp

1 はじめに

第2言語としての日本語の教育現場の特徴の一つに、学習者の言語的、文化的背景が一様でないという点がある[1]。学習者のレベルも、ばらばらなことが多く、このため教師にとっては、様々な学習者にあわせて授業を行なわなければならないという負担がある。例えば、授業で教師が学習者に提示する対話例(スキットと呼ばれる)は、その対話例を理解するために必要な文法知識が、学習者にとって未習なものばかりの場合には問題があるため、適切なものを選ぶ必要があるが、実際には、様々な学習者に合わせるために、教師自身がスキット書きに追われてしまうこともある。

ここで、教科書、授業のカリキュラムにおいて教授される事項を学習項目と呼ぶ。電子化された対話例に対して、学習項目による検索が可能であれば、教師の負担を軽減することが期待できる。

検索の前提として、対話例に含まれる一つ一つの文と、その文の理解に必要な学習項目との対応づけ(リンク)が行なわれている必要があるが、この作業は自動化できることが望ましい。この作業に形態素解析ツールを利用することを考えると、形態素解析のみでは、例えば、係り受け関係はわからないといった限界はもちろんあり、ある文にたいして、その文と対応する学習項目を正確に絞りこむことはできない。今回、文と学習項目とをリンクする作業を、対応する学習項目の候補の一覧の提示までを自動化し、最終的に編集者の判断により決定する、つまりリンク作業を支援するツールを提案する。図1にその概略を示す。

2 文と学習項目の対応

日本語教師向けの文法参考書[2]から文と学習項目の対応の例を引用する。

これは本です。
これは本ですか。
はい、それは本です。
いいえ、それは本ではありません。

たとえば、日本語の教科書のはじまりによく見られる上記の例文の中には、次の学習項目が含まれている。

1. 文型 「～は N(名詞) + だ」
2. コソアド
3. 「か」 助詞。文末について疑問文を作る
4. 「はい、いいえ」 応答詞

ここで、「これは本ですか。」に対しては、1,2,3の学習項目が対応することに見られるように、文と学習項目とは1対多の対応関係を持つ。

3 学習項目と形態素情報

形態素解析ツールとして、奈良先端大の茶釜[4]を利用している。茶釜の解析結果の単位となる形態素の構造は、その形態素の形態品詞、品詞細分類、活用型、活用形、見出し語がそれぞれ H_1, H_2, K_1, K_2, M であるとき、

$$(H_1 H_2 K_1 K_2 M)$$

という5項組で表わされる。図1において、まず、「これは本です。」という文が形態素解析され、その結果として5項組の系列が得られる。

(指示詞 名詞形態指示詞 * * これ)
(助詞 副助詞 * * は)
(名詞 普通名詞 * * 本)
(判定詞 * * デス列基本形 だ)
(特殊 句点 * * 。)

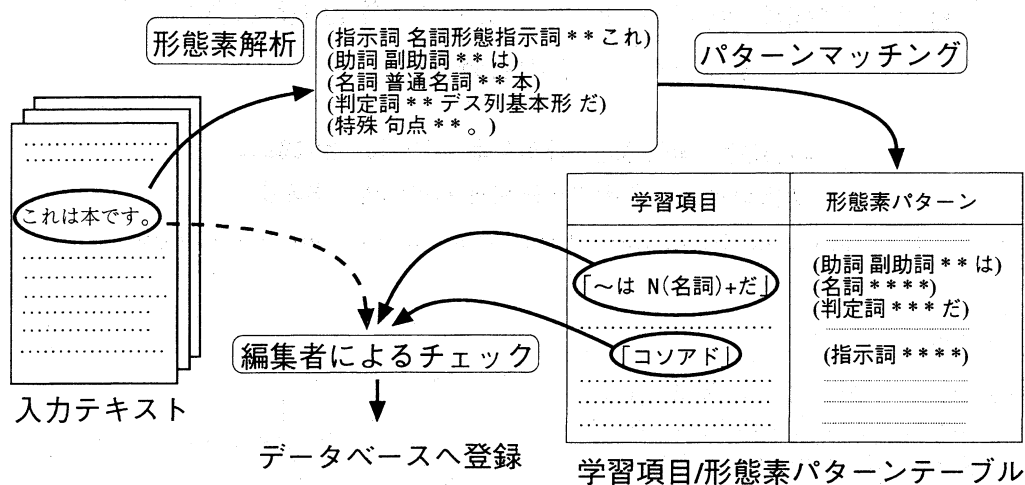


図 1: リンク作業の支援の概略

この出力結果の * は例えば名詞に活用がないように、「無し」という意味である。

この5項組を、そのまま学習項目と対応する形態素パターンの記述にも用いる。パターンの記述には更に以下の規則も利用することができるものとする。

- 5項組の要素としては don't care 項として * が使用できる。
- 5項組 α の 0 回以上の繰り返しが使用でき、 α^* と表記する。(don't care の * とは別)
- 5項組 $\alpha, \beta, \dots, \gamma$ にたいして、論理和が使用できる。{ $\alpha \beta \dots \gamma$ } のように表記する。
- 空の5項組が利用でき、*nil* と表記する。5項組 α が省略可能であることを表現するには、{ $\alpha \text{ nil}$ } と表記すればよい。

パターンマッチングに用いる形態素パターンの例をいくつか挙げる。

例: 「～は N(名詞)+だ」
(助詞 副助詞 ** は)
(名詞 ****)
(判定詞 *** だ)

「あれは看板だ。」 「これは時計です。」等にマッチする。「だ」の活用は自由で、任意の名詞

を一つ持つ文である。よって、形容詞や連体修飾を含むような文はマッチしないことになる。

例: 「(これ, それ, あれ) は～」
{(指示詞 *** これ)
(指示詞 *** それ)
(指示詞 *** あれ)}
(助詞 副助詞 * * は)

「あれは学校ですか」、「いいえ、あれは病院ですよ。」等

例: 「(この、その、あの)+N(名詞)」
{(指示詞 *** この)
(指示詞 *** その)
(指示詞 *** あの)}
(名詞 * * * *)

「昨日あの店に入ってみたよ。」 「ああ、あの店ね、どうでしたか。」等

4 検索

文と学習項目のリンク作業を支援するツールにより、候補の一覧が編集者に提示される。編集者はそれをチェックし、必要であれば修正を行ない、その結果が図2の検索用のテーブルに登録される。文カラムには、その文の文字列データ、学

文	学習項目(配列)	形態素(配列)	ファイル名	位置
これは本ですか。	「～はN(名詞)+だ」 「コソアド」 「～か」(疑問)	(指示詞 名詞形態指示詞 ** これ) (助詞 副助詞 ** は) (名詞 普通名詞 ** 本) (判定詞 ** デス列基本形 だ) (助詞 終助詞 ** か) (特殊 句点 ** 。)

図 2: 検索用テーブル

学習項目カラムには、配列としてその文にリンクされた学習項目が、形態素カラムには、配列としてこの文の形態素解析の結果が、そして、ファイル名カラムと位置カラムには、その文がどのファイルのどこにあるのかが登録される。

文を学習項目から検索する際には、この検索用テーブルの学習項目カラムが用いられる。また、形態素カラムには、文の形態素解析の結果が登録されるが、これは、直接、形態素パターンを用いて文が検索された時にキャッシュとして用いられる。

5 問題点及び今後の課題

本稿では、形態素解析の後の処理、例えば、自立語の意味あるいは意味カテゴリーを調べたり、構文解析により係り受けを調べたりといったことはしていない。また、形態素パターンは、句点で終了する一つの文に対してマッチングを行っていて、文の連続、あるいは談話に対して解析は行っていない。これらによる制約としては、例えば次のような問題がある。

テ形

テ形で複数の動作を並べる時、テ形の前後の動詞によって、テ形の持つ意味が順次動作(「朝起きて歯をみがく」)、並行動作(「手をかざして遠くを見る」)、手段・方法(「ブラシを使って掃除する」)、原因・理由(「途中で追いぬかれて、負けた」)のように変化するが、これらを区別する学習項目の分類を要求された場合、違いを判別することは、形態素パターンのルールだけでは当然できない。

のに

「店に行って、お金がないのに気がついた。」
「店に行って、お金がないのに買ってしまった。」、この両者の区別には、「のに」が逆接かどうか、前後の意味がわかっている必要がある。

文を超えた解析

例えば、指示詞の「それ」が、前方の語句を指しているのか、それともこれから表われる後方の語句を指しているのか、それとも漠然と状況を指しているのか、この区別は、もちろん一文に限った今回の方法では判別できない。

本稿の手法は、形態素解析のみを使うものであり、単純な方法である。今後、この方法の可能性をさらに検討するため、様々な文型、文法事項について適用可能かどうかを詳しく調べていきたい。

参考文献

- [1] 石田敏子: 日本語教授法, 大修館書店,(1988).
- [2] 吉川武時: 日本語文法入門, NAFL 選書 6, アルク,(1989).
- [3] 生田目 弥寿: 日本語教師のための現代日本語表現文典, 凡人社,(1996).
- [4] 松本裕治 北内啓 山下達雄 今一修 今村友明: 日本語形態素解析システム『茶筌』 version 1.0 使用説明書, NAIST Technical Report, NAIST-IS-TR97007,(1997).