論文 / 著書情報
Article / Book Information

論題(和文)	構文解析器を利用した作文支援システムの開発ー形容詞表現に関して
Title(English)	Development of a System for Composition in Japanese Utilising the Dependency Structure Analyser--Focusing on Adjectives
著者(和文)	戸次徳久, 仁科喜久子
Authors(English)	KIKUKO NISHINA
出典(和文)	第３回「日本語教育とコンピュータ」国際会議, Vol. , No. , pp. 67-70
Citation(English)	Castel/J 2002 Proceedings-The Third International Conference on Computer Assisted Systems For Teaching & Learning Japanese Computer Technology and Japanese Language Education, Vol. , No. , pp. 67-70
発行日 / Pub. date	2002, 7

構文解析器を利用した作文支援システムの開発　ー形容詞表現に関して

Development of a System for Composition in Japanese by Utilizing the Dependency Structure Analyser --Focusing on Adjectives

戸次 徳久, 仁科 喜久子 (東京工業大学)
Norihisa Totsugi, Kikuko Nishina (Tokyo Institute of Technology)

概要： We propose a system to help the Japanese learners to correct their composition. We utilise NLP applications such as the morphological analyser and the dependency structure analyser to make our system flexible and useful. Adjectives which modify nouns are focused in this paper. A simple syntactic dictionary and a thesaurus are made to construct the system. The system processes the user's input, detects unnatural collocations and shows substitutive expressions.

キーワード： 作文, コロケーション, 自然言語処理, 形容詞, 類語, 統計

1. はじめに

母語でない言語で作文をすると、文章構成法、統語、表記、表現などに誤りが生じることがある。上級の学習者になると文章構成、統語、表記に関しては、ほぼ間違えることはなくなる。しかし、表現に関する間違いは依然として残る。例えば、「強烈な印象」と書くべきところを「？猛烈な印象」とする過ちは、辞書を引いてもわからないことが多く、学習者自身では直せない過ちである。言語習得の際にコロケーションを学ぶことの重要性は、Granger(1998)等の一連の第 2 言語習得研究で指摘されている。

また、楊・赤堀(1996)らの研究に代表されるように、従来の作文支援システムは、固定の場面の固定の表現に関して作文練習をするものであった。近年、語学教育分野において、掛川他(2000)など、自然言語処理の方面からの取り組みがあり、自由な入力に柔軟に対応できるように、システム開発に新たな面が生まれてきている。

2. 研究目的

この研究は、日本語学習者用の作文支援システムを開発することを目標としている。ここでは、特に形容詞表現に関して述べる。本システムでは、自然言語処理技術を用いることにより、自由な入力に対して助言を与え得るようにする。

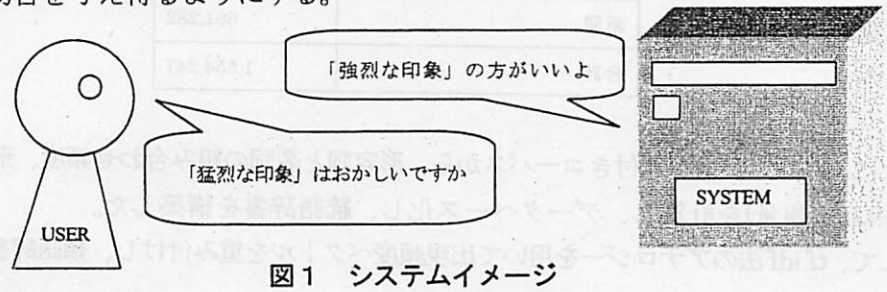

図1　システムイメージ

3．　システム

3．　1．　利用形態

学習者が自学自習できるようにするため、インターネットを介してシステムを利用できる形態を取ることにした。

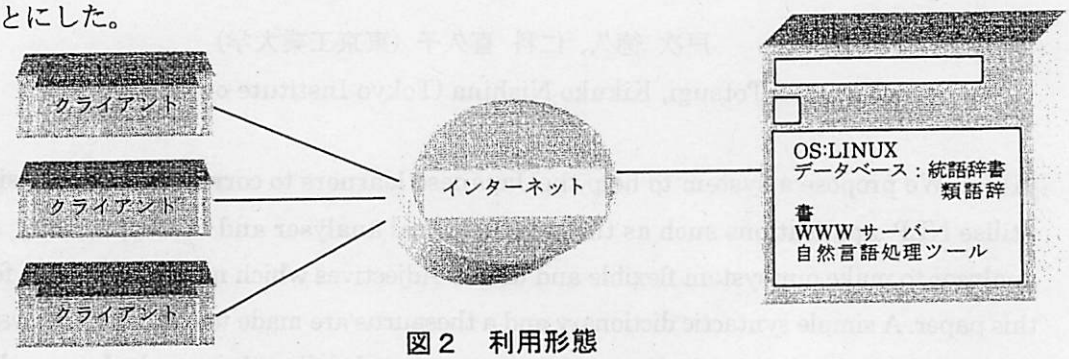

図２　利用形態

3．　2．　システムの概要

システムは、学習者の入力を受け付け、その中で誤った表現を発見し、代替候補を出力する。

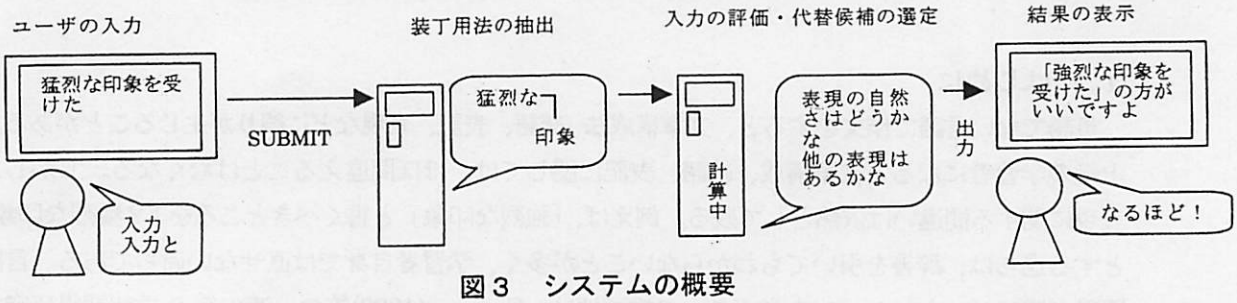

図３　システムの概要

3．　3．　システムの構築

まず、表１に示すテキストから、形態素解析器 JUMAN、構文解析器 KNP を使用して構文木付きコーパスを作成した。テキストは、毎日新聞 CD-ROM95 年度版、CASTEL/J2000 など約 150 万文から成る。

表１　コーパス

ジャンル	文の数(文)
小説	701,524
エッセイ・日記・評論	191,441
新聞	661,282
合計	1,554,247

次に、作成した構文木付きコーパスから、形容詞と名詞の組み合わせ頻度、形容詞と名詞の関連度(相互情報量)を計算し、データベース化し、統語辞書を構築した。
そして、tf.idf法のアナロジーを用いて出現頻度ベクトルを重み付けし、類語辞書を作成した。手順

を以下に示す。

1. 構文木付きコーパスから形容詞と名詞の関係について学習し、名詞の異なり数 NN、形容詞に対する名詞の頻度 nf(名詞, 形容詞)、形容詞の異なり数 NA を求める。
2. 形容詞の特徴ベクトルを以下のように求める。

$$\vec{x}=\frac{\vec{c}}{|\vec{c}|},\quad \vec{c}'=(nf(n_1,a_x)\cdot iaf(n_1)\quad nf(n_2,a_x)\cdot iaf(n_2)\quad \cdots\quad nf(n_{NN-1},a_x)\cdot iaf(n_{NN-1})\quad nf(n_{NN},a_x)\cdot iaf(n_{NN}))$$

但し、 $iaf(n_p)=\log\frac{NA}{af(n_p)}+1$

3. 以下の式で形容詞 x と y の類似度を得る。

$$Sim(x,y)=\vec{x}\cdot\vec{y}$$

4. 各形容詞毎に、類似度が高い順に他の形容詞を並べ、データベース化する。

しかしながら実際に作成した類語辞書を調べたところ、低頻度語が過小評価されていた。そこで、低頻度語での順位が高いものを評価するために、以下の式を用い、再評価した。この式を用いることにより、2 つの形容詞のどちらかから見た類語順位が高い場合に、高い評価を与えることができる。これにより、Sim'(x, y)=Sim'(y, x)となるので、副次的効果として、類似度の値を持つデータの量を半分にすることができた。

$$Sim'(x,y)=\frac{1}{\log r(x,y)+1}+\frac{1}{\log r(y,x)+1}$$

但し、$r(adjective1, adjective2)$は、形容詞 *adjective1* からみた形容詞 *adjective2* の類語辞書での順位

ユーザの入力を以下のように処理するようにした。

1. 入力文を形態素解析、構文解析し、係り受け関係を得る。

入力された文章に対して、形態素解析器(JUMAN)、構文解析器(KNP)を利用することにより、文章の係り受け構造を得る。

(例) 入力：猛烈な印象を受けた

形態素情報

猛烈な	印象	を	受けた	。
形容詞	名詞	助詞	動詞	特殊

係り受け関係

猛烈な——┐
印象を——┐
受けた。

図４　形態素情報と係り受け関係

2. 該当する語を含む係り受け関係(今回は形容詞の装丁用法に限る)に着目する。

1 項について用いられている形容詞の装丁用法の関係があった場合、その関係を抽出する。

(例)

猛烈な——┐　　抽出　　猛烈な——┐
印象を——┐　　印象
受けた。

図５　装丁用法の抽出

3. 当該語の類語を得、代替候補語とする。

類語辞書から該当した形容詞の上位30の類語を代替候補の形容詞とする。また、類語の幅を広げるため、そのうち上位5位の形容詞に関して、さらに類語辞典を引き、それぞれ上位10語の類語を代替候補に加え、さらに、入力された名詞と相互情報量の高い形容詞を順に10語加え、延べ100語の代替候補の形容詞を得る。

4. 代替候補語を、当該語との類似度や係り受け関係にある語との相性などを考慮に入れスコアリングする。

代替候補の形容詞から入力文を考慮に入れて相応しいものを選ぶためにスコアリングする。スコアリングには以下の式を用いる。

$$Score_x = \langle \text{表現の起こりやすさ} \rangle \times \langle \text{結びつき度} \rangle \times \langle \text{類似度} \rangle$$
$$= (1+\log_F freq(x, noun)) \cdot (1+\log_M mi(x, noun)) \cdot Sim'(x, adj)$$

但し、xは、候補形容詞、
Fは、頻度の組み合わせの最大値、
Mは、相互情報量の最大値、
$noun$は、当該名詞
$mi(adjective, noun)$は、形容詞$adjective$と名詞$noun$の相互情報量、
$freq(x, noun)$は、形容詞xと名詞$noun$の学習したコーパスでの頻度

スコアリングには3つの尺度、表現(形容詞+名詞)の起こりやすさ、表現(形容詞+名詞)の結びつきの強さ(汎用性のなさ)、表現(入力表現と候補表現)の類似度を用いている。

5. スコアに従い代替候補表現を出力する。

もとの入力のスコア、代替候補のスコアを評価し、結果を出力する。

4. おわりに

不自然な表現を訂正する作文支援システムを提案した。今後の課題として、システム評価を行うこと、論文を書くため、日記を書くため、手紙を書くため、など場面に特化して作文を支援できるようにすることおよび、データのスパースネスへの対処などが挙げられる。

参考文献

Granger, Prefabricated patterns in advanced EFL writing: collocations and formulae," in Cowie, A.P. (ed.), pp.145-160.

楊接期・赤堀侃司, 自然言語処理を用いた日本語作文学習支援システムの開発——受身について——, 日本教育工学会研究報告集, JET96-6, pp.61-68, 1996.

掛川淳一・神田久幸・藤岡英太郎・伊丹誠・伊藤紘二, 日本語学習支援システムにおける作文診断処理系の提案と試作, 電子情報通信学会論文誌, Vol.J83-D-I, No.6, pp.693-701, 2000.

黒橋禎夫, 日本語形態素解析システムJUMAN Version3.61 使用説明書, 京都大学大学院情報学研究科, 1998.

黒橋禎夫, けっこうやるなKNP, 情報処理学会誌, Vol.41, No.11, 2000.