

8L-01 形態素列間に定義した類似度による中国人の 日本語文誤りパターン獲得システム

耿 セイ 小島丈幸 乾伸雄 小谷善行
(東京農工大学 工学部 電子情報工学科)

1. はじめに

外国人が日本語を勉強するとき、すでに理解している母国語の文法と対照しながら勉強することが多い。外国人それぞれの母国語の文法が違うので、母国語に応じて日本語文章の誤りも違う。このため、日本語の文章を読むだけで、中国人が書いたのかそれともアメリカ人が書いたのかがわかる。このことは誤りがパターンになっていることを示していて、その誤りパターンを機械によって獲得することができること考えられる

本研究は、中国人が書いた日本語誤り文を入力することによって、中国人の日本語誤りパターンを獲得することを試みる。

2. システムの設計について

2.1 システムの概要

本システムが文の形態素解析した結果に対して、誤り文と同じパターンを持つ例をデータベースより検索するための類似度を提案する。ユーザーが中国人の誤り文を入力し、定義された文間の類似度によってデータベースから検索された誤りパターンの候補文をユーザーに示す。これらの作業は、ダイアログボックスを使ってユーザーと対話する形で進めていく(図1に示す)。

本システムは入力した中国人の誤り文を誤りパターンに帰納し、誤り訂正に用いることを目的とする。入力文がシステム中の誤りパターンに含まれているかの判別がきわめて重要である。システムは入力文と同じ誤りパターンの候補の例文を

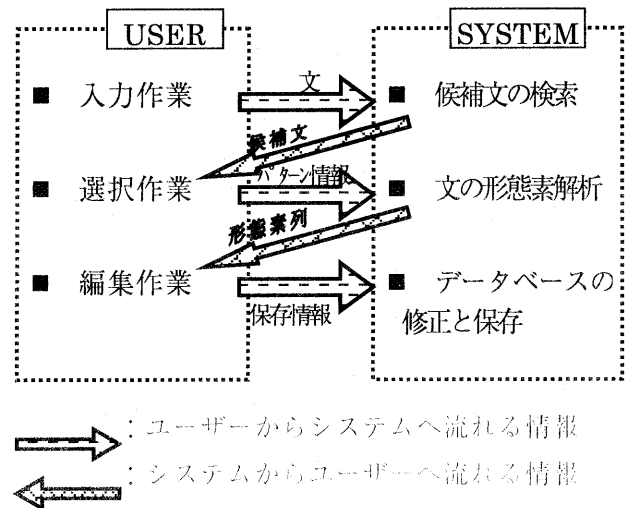


図1 システムとユーザーの情報交換

ユーザーに示す。その作業は、入力文とデータベースにある誤りパターンとの類似度の計算によって行う。この論文ではその類似度の計算方法を中心に説明していく。

2.2 文間の類似度の計算

システムはユーザーに次の二つの情報を要求する。

誤り文(中国人が書いた誤り文)

キー(誤りのポイントだと思われる単語)

システムはこの二つの情報によって、入力文とデータベースにある誤りパターンとの類似度を計算する。

まず、キーによる類似度の計算を行う。キーとなる単語とデータベースに入っている誤りパターンとの対応関係を保存するファイルを用意する。入力したキーがそのファイルに入っていないければ、新規のキーとして、同じファイルに保存する。そのキーがファイルに入っていれば、それと対応する誤りパターンの類似度を最大にして、候補誤りパターンとして出力する。

キーによる類似度の計算が失敗した（候補文の中に入力文と同じパターンである文が存在しない）場合には、形態素間の類似度による類似度の計算を行う。

- (1) 茶筌（形態素解析ソフト）を用いて、入力文に形態素解析を行い、品詞列に変換する。
- (2) DP マッチングを用いて、入力文の品詞列と、知識としてデータベースに保存してある各誤りパターンの品詞列との間の類似コストを計算する。

i) 品詞の挿入と脱落についての処理

日本語では「ご」、「お」などの接頭詞や「た」、「よ」などの語尾が存在する。しかしこういう成分が存在することによって、文の全体の構成に影響は少ない。よって、二つの品詞列間に少し挿入と脱落があっても問題にならない場合が多い。しかしながら、接頭詞や語尾などがあっても品詞成分としては一つ占めないで、連続の挿入や脱落などの現象は許してはならない。そのため、品詞間のコストの設定を、次のように決める。

- 違う品詞間の類似コスト： 2
- 挿入、脱落品詞の類似コスト： 1

ii) 助詞についての特別扱い

i) で述べた方法で類似コストを計算すると、「私がやる」という文と「テーブルを拭く」という文で、「が」と「を」が表す文法的な意味が全く違うのに、同じ助詞であるため、二つ文間のコストは 0 となってしまう。これを避けるため、品詞が助詞である場合には、語が異なるだけで類似コストを 2 とする。

- (3) 品詞列間の類似コストが大きければ大きいほど文の間の類似度が低くなるように文の間の類似度を決める。

$$\text{類似度} = (1 - \text{類似コスト} / (\text{入力文の長さ} * 2)) * 100$$

- (4) 類似度の大きい順で前十通りの誤りパターンの例文を候補文として、ユーザーを選択させる。

3. 類似度の例

候補パターンの類似コストと類似度の計算結果の例を載せる：

入力文：寒かったの私

誤りパターンの例文	Cost	類似度
白いのシャツ	1	88
日本語勉強	2	75
優しいだ	2	75
工場に見学する	4	50
日本語ほど中国語が難しい	5	38
帰えられない	6	25
話すをする	7	13

入力文：休むする

誤りパターンの例文	Cost	類似度
話すをする	1	75
帰えられない	2	50
日本語勉強	4	0
優しいだ	4	0
工場に見学する	4	4
白いのシャツ	5	-25
日本語ほど中国語が難しい	7	-75

上の例で示すようにほとんどの誤りパターンは入力文と近い順で表示されている。しかし、例 1 のところで、「寒かったの私」と「日本語勉強」の類似度が高いことが少し不自然である。なぜなら、形態素解析は誤り文を正しい文として解釈しようとするので、助詞であるはずの「の」を「名詞」として判断してしまうからである。ゆえに、形態素解析が誤り文への対応はこれからの重要な課題である。その方法の一つとしてはあらかじめ助詞辞書を作っておくことを考えられる。

4. おわりに

本稿では DP マッチングを日本語文の品詞列の特徴を考えた上で加工し、品詞列間の類似度を計算する方法を提案した。現在、この方法を用い、中国人の日本語文誤りパターンを獲得するシステムを開発している。

参考文献

- [1] 石崎 俊著：自然言語処理 pp1-3, 昭晃堂, 1995.
- [2] web 「DP マッチングとは？」：
<http://sail.i.shikawa-nct.ac.jp/pattern/onsei/>