厚生労働科学研究委託費（障害者対策総合研究事業）
委託業務成果報告（総括及び業務項目）

目的文節数文章の抽出プログラム作成に関する研究

業務主任者又は担当責任者　小町守　首都大学東京システムデザイン学部准教授

研究要旨
長文節文章の聴取検査は、中枢聴覚機能を反映させることが期待される。長文節文章の聴取検査を作成するに当たり、長文節の文章及びその音声を多数収集する必要がある。今回我々は、日本語係り受け受け解析器 CaboCha を用い音声コーパスより目的の文節数の文章を抽出するプログラムを作成した。本プログラムにより、指定した文節数の文章及び音声をコーパスより抽出することが可能となった。

A．研究目的

老人では末梢性の聴覚障害のほかに、音信号の分析・抽出・認知・記憶・理解といった中枢における聴覚処理の障害が大きく影響していることが明らかとなっている。しかしながら、従来の聴力検査では中枢における聴覚処理は測定が不可能であった。一般に文章の聴取は文節数が多くなるにつれ困難になる。これは文節数が増加するにつてれ、単音節・単語レベルの聴取・理解に加え、それらを認知・記憶し総合的に理解する必要が出てくるためである。長文節文の聴取の検査は聴覚の中枢処理能力を反映することが期待される。長文節文章の聞き取り検査を作成するに当たり、長文節の文及びその音声を多数収集する必要がある。このような背景から今回、文章と音声のデータベースである、音声コーパスより各種文節数の文を抽出するためのプログラム開発を行った。

B．研究方法

本研究で用いた日本音響学会新聞記事読み上げコーパス（JNAS コーパス）には単語境界が付与されているが、文節境界が付与されていないため、
日本語文章から長文節文章を抽出するに当たって、以下のアルゴリズムを採用した。
日本語テキストに対して1文ずつ形態素解析（単語分割）および係り受け解析（文節境界推定）を行う。
文節数をカウントし、指定した文節数の文章を抽出する。
抽出した文章のIDにより、文章データおよび音声データを対応づける。
プログラムはプログラミング言語 Python および Bash を用いて実装した。形態素解析には MeCab 0.996（UniDic 2.1.2）、チャンキングには CaboCha 0.68 を用いた。CaboCha は MeCab の出力を入力とする前提で動作するため、JNAS コーパスに付与されている単語境界は使用せず、MeCab で付与した。
（倫理面への配慮）プログラムの作成であり被験者などの資料は扱わない。

C．研究結果

前節で作成したアルゴリズムを（JNAS）に対して適用した。JNASコーパスには16,178文が含まれ、5-7文節の文を抜き出し、5,322文が得られた。
実際に得られた文の例を以下に示す。

（5文節の文）
新 外国人 選手 の プレー が ファン の 注目 を 集め そう だ
（6文節の文）
米国 の 復興 援助 に 加え 朝鮮 戦争 特需 の 神風 も 吹い た
（7文節の文）
場合 によって は 衆院 本会議 で の 同 法案 の 趣旨 説明 を ボイコット する 構え も 見せ ている

D．考察

5-7文節の文を抽出することにより、コーパスの大きさが約1/3となったが、5,000文あれば聴取テスト作成には十分と考えられる。また、本研究で作成したプログラムは任意の文節数の文を抽出できるが、予備実験の結果5-7文節の文でも従来の聴取テストより難易度が高いと推測されるため、8文節以上のものは取り除いた。
一般に、形態素解析や係り受け解析は自動付与されるため、解析誤りが存在することが多いが、今回用いた日本語テキストが比較的整った文章である新聞記事であったため、形態素解析・係り受け解析誤りによる長文節文の抽出への影響は見られなかった。
一方、前節で例文を示したように、今回用いたコーパスが新聞記事の読み上げコーパスであるため、文節数が5-7文節でも比較的単語数の多い文章が含まれる。また、政治やスポーツの記事など、新聞に比較的多く見られるジャンルのテキストが多い。聴取テスト作成の際にはジャンルの影響を考慮する必要があると考えられるが、聴取テストに適したジャンルの音声コーパスはどういったものか、今後検討していく必要がある。

E．結論

音声コーパスより指定した文節数の文章・音声を抽出するプログラムを作製した。

G．研究発表

1． 論文発表

林部祐太，小町守，松本裕治．述語と項の位置関係ごとの候補比較による日本語述語項構造解析．自然言語処理，Vol.21，No.1,3-26，2014.

2． 学会発表

Budi Irmawati, Mamoru Komachi, Yuji Matsumoto. Towards Construction of an Error-Corrected Corpus of Indonesian Second Language Learners. 6th International Conference on Corpus Linguistics. 5.22-24, 2014. Spain

Yinchen Zhao, Mamoru Komachi and Hiroshi Ishikawa. Extracting a Chinese Learner Corpus from the Web: Grammatical Error Correction for Learning Chinese as a Foreign Language with Statistical Machine Translation. The 22nd Conference on Computers in Education. 11.30-12.4, 2014. Nara, Japan.

Kenichi Ohwada, Ryosuke Miyazaki and Mamoru Komachi. Predicate-Argument Structure-based Preordering for Japanese-English Statistical Machine Translation. The 1st Workshop on Asian Translation. 10.4, 2014. Tokyo, Japan.

H．知的財産権の出願・登録状況

なし