7Q-04

複数の音声認識エンジンによる音声記録の書き起こし文補正方式

森田 瑛登† 石井 翔大† 秋吉 政徳†
神奈川大学†

1 はじめに

音声記録の自動書き起こしソフトの増加や情報機器端末・Web 上において音声認識の導入など近年音声認識技術が急速に広まっており、手を動かすことなく、音声で機械への動作命令を行うことが可能になってきている。しかし、現在使用されている音声認識技術の多くは短文を主としたものばかりであり、認識対象の文章が長くなるにつれて認識率が低下している。さらに、日本語は数多くの同音異義語・短縮語が存在している上、文章の区切りが曖昧で認識が非常に困難な言語である。音声認識に対する事前トレーニングを行い、理想的な環境下での実験では 8 割の認識率を示す場合もあるが、現実的にそのような環境下で行う事は難しい。音声認識エンジンは大抵の場合に学習能力が備わっており、声や人物登録や、長期間使用し続けることで認識率は向上する。しかし、現実的に数年単位での学習をしなければならないため、非現実的である。

そこで、書き起こしデータ文に対し誤認識箇所を訂正する方式[1]が提案されているが、書き起こし文の補正を行う際の範囲指定は手作業となっており膨大なデータに対する音声認識となると手作業は非現実的である。また、品詞 N-gram による誤り検出法[2]もあるが、音声認識で発生する同音異義語を修正することが不可能である。

本稿では、音声記録データに対して、複数の音声認識エンジンによる書き起こし文に対しての正誤判定を行い、正しい書き起こし箇所のみを自動抽出する方式を提案する。

2 書き起こし文の補正

2.1 方式の構成

図 1 に示すように、まず複数の音声認識エンジンにより作成された書き起こしデータ同士のマッチング処理を行う。その際、複数の音声認識エンジンで同じ単語が発生している場合は正しい書き起こし箇所とする。また、形態素解析により区切られた書き起こしデータがコーパス内においてどの程度出現しているかに対して、単語 N-gram による単語列を作成し、その単語列でコーパス内の一定以上の出現頻度を持つものを抽出する。更に、検索エンジンにより単語列の完全一致型検索を行い、検索結果数が一定以上の文字列のみ抽出を行う。その結果、抽出された文字を一致データに加えることで、音声記録文書き起こしデータの補正を行う。

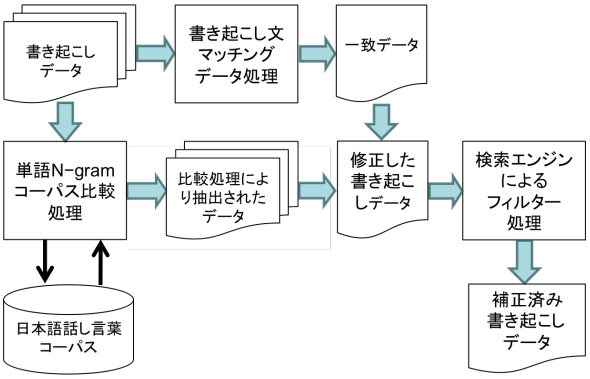

図 1: 書き起こし文補正方式

2.2 書き起こし文マッチング処理

複数の音声認識エンジンにより書き起こされた文を比較し、二つ以上の音声認識エンジンが同じ文章を書き起こしている箇所を正しく認識された箇所として、抽出を行う。結果として、複数の書き起こし文の一致箇所のみを抽出したデータ(以下、一致データと呼ぶ)が出力される。この段階では一致データの中身はデータが歯抜けの状態であり、接続語や語尾が抜けていることが非常に多い。この処理は既存のテキスト比較ツール“diff”[3]を利用する。

A Correction Method of Transcriptions of Speech Data de- rived by Multiple Speech Recognition Engines
†Akito Morita, Shodai Ishii, Masanori Akiyoshi
Kanagawa University

2.3 単語 N-gram コーパス比較処理

単語 N-gram コーパス比較処理プログラムでは、形態素解析により書き起こし文を単語区切りにし、N-gram の単語列を含む文章がコーパス内に存在するかを調べ、コーパス比較処理により一致した文章は正しい書き起こし箇所として、一致データに加える。使用するコーパスは、国立国語研究所作成の「日本語話し言葉コーパス」を利用する。単語 N-gram でのコーパス比較処理により、一つの音声認識エンジンのみ正しく書き起こしている箇所の抽出を行う。

2.4 検索件数比較

検索エンジンによる検索結果抽出プログラムでは、マッチングプログラムとコーパス比較処理プログラムにより作成された一致データを検索エンジン“bing”により検索をかけ、検索ヒット件数による正誤判定を行う。検索を行う際に、一致データを形態素解析により単語ごとに分割を行う。そして単語 N-gram をもとに検索ワードを生成し、完全一致型検索による検索件数が k 件以下の箇所は誤っている箇所とし、データから削除を行う。

3 実験

3.1 実験概要

表 1 に、実験で用いた条件を示す。

表 1: 実験条件

音声認識エンジン	AmiVoice SP2 ドラゴンスピーチ 11 Google Web
読み上げ人数	6 人
読み上げ用テキスト文字数	約 2300 文字
自由対話による発話話時間	約 5 分
データベース内容件数	約650 万単語
検索エンジン	bing
検索方法	完全一致型検索

実験に使用するデータは読み上げ用テキストと自由対話形式の2種類である。読み上げ用テキストでは、ホームページ作成の学習サイトの一部が記載された約 2300 文字を読み上げた音声記録を使用した。自由対話による発話では、ホームページを作成する際の打ち合わせで、一人約 5 分程の音声記録である。本実験に使用する音声記録は全て音声認識エンジン “Ami Voice SP2”により記録されたものであり、記録を行う際、全員ヘッドセットを着用した。 一つの音声記録から 3 種類の書き起こし文が作成され、これらをもとに書き起こし文の補正を行う。

3.2 実験結果

結果を表 2 に示す。再現率は発話に対し、正しく書き起こされた文字数の割合を、精度は書き起こされた文章に対し、正しく書き起こされた文字数の割合を示している。

表 2: 実験結果

音声認識エンジン	テキスト読み上げ形式 再現率 (精度)	自由対話による発話 再現率 (精度)
AmiVoice	78%(85%)	60%(66%)
ドラゴンスピーチ	81%(89%)	67%(72%)
Web Speech	67%(81%)	37%(85%)
提案方式	79%(92%)	57%(80%)

自由対話による発話形式では、突発的な発話や脈絡のない単語が出現することが多く誤認識が増えてしまうため、テキスト読み上げ形式に対し、既存の音声認識エンジンと提案方式の結果が、ともに大きく下がってしまうことがわかる。しかしその中でも精度は約 80% という結果になった。精度だけを見ると既存の“Web Speech”が最も高い。しかし再現率が 37%と非常に低く、文章の認識という点においては結果は不十分である。また、テキスト読み上げ形式においては、再現率は既存の音声認識エンジンとあまり差はないが、精度が向上されたことがわかる。しかし、再現率においてはテキスト読み上げ形式、自由対話による発話形式が共に既存の音声認識エンジン以下の制度となった。

4 おわりに

複数の音声認識エンジンによる書き起こし文を元に、書き起こし文の自動補正を行う方式を提案した。評価実験の結果から、“再現率”に関するさらなる改良が必要であることがわかった。今後は、音声認識エンジンの追加における比較データ数の増加による結果への影響の調査、及び再現率の向上への手法の改良を試みる予定である。

参考文献

[1] 中島悠輔, 張志鵬, 仲信彦, “音声による文字入力の使いやすさ向上を目指した効率的な音声認識誤り 訂正技術”, NTTdocomo Technology Reports, Vol.17, No.2, pp.30-35 (2009)

[2] 石場正大, 竹山哲夫, 青木恒夫, 兵藤安昭, 池田尚 志, “品詞 N-gram 統計情報を用いた日本語文書における誤り検出法について”, 音声言語情報, pp.95-100 (1997)

[3]テキスト比較ツール, http://difff.jp