E-031

初等中等教育授業における教師発話の言語的特徴のモデル化のための学習データ選択方法の検討

南條 浩輝 *　　谷奥 大喜*

1　はじめに

現在，大学などにおいて，音声ドキュメント処理技術を用いた講義支援や学習支援，聴覚障害者への情報保障などの研究が試みられている [1][2][3]．一方，初等中等教育においては，そのような試みはほとんど行われていない．本稿では，初等中等教育授業を対象とし，音声ドキュメント処理の基礎技術である音声認識について述べる．具体的には，教師発話（子ども向けの発話）の言語的特徴をモデル化するための学習データを選択する方法について述べる．

2　初等中等教育における教師発話の特徴とモデル化の難しさ

初等中等教育での授業は児童・生徒（以下，子ども）が理解できるようにすすめられる．難しい語は基本的に使用されず，よびかけや確認が多い傾向がある．教師発話は話し言葉ではあるものの，大人向けの話し言葉（大学講義，学術講演，会議）とは言語的にも音響的にも大きく異なっており，大人向けテキストから学習した言語モデルを教師発話の音声認識に用いるのは難しい [4][5]．

3　web からの子ども向けテキストの抽出

これまでに我々は，子ども向け web サイトのテキスト（子ども向けニュースサイトコーパス（KNC）[5]）から，言語モデルを学習する効果を確認している．web 上には子ども向けであることが明示されていないサイトや，子どもが対象でないものの子ども向けとみなせる表現が含まれるサイトが存在する．本研究では，このようなテキスト資源を有効利用するため，web から収集したテキストから子ども向け表現テキストを抽出し，言語モデルの学習を試みる．実際に既に我々はこの課題に取り組んでおり [6]，一般の web サイトにも子ども向け表現のモデル化に有効なテキストが含まれていることを確認している．本稿ではこの抽出方法の改善方法について検討したので，その報告を行う．

3.1　従来の抽出法とその問題点

これまでに行っていた子ども向け表現の抽出手順は次のとおりであり，式（1）で表される [6]．

1. web から収集したテキストの一部 w を取り出す
2. w と，子ども向け表現モデル C および大人向け表現モデル A のそれぞれとの距離 $\mathrm{D}(A,w)$，$\mathrm{D}(C,w)$ を算出する
3. 距離の比をとり，比が定めたしきい値（Th_r）以下のときに w を子ども向け表現とする．

$$\frac{\mathrm{D}(C,w)}{\mathrm{D}(A,w)} \leq Th_r \rightarrow \text{slect } w \text{ as a kids-oriented text} \quad (1)$$

この抽出方法で抽出されるテキストを調査したところ，式（1）のしきい値 Th_r が小さいときに，子ども向け表現でないテキストが多く含まれることがわかった．実際に Th_r を 0.5 としたときに抽出されたテキストの先頭から 100 ページを調査したところ，大人向け web ページが 12%含まれていた．また，固有名詞などの単語のみのページが 74%含まれていた．これは，式（1）の左辺は小さいものの，$\mathrm{D}(C,w)$，$\mathrm{D}(A,w)$ ともに大きな値，すなわち，大人向け表現モデル A と子ども向け表現モデル C の両方とテキスト w が遠いときにテキスト w を抽出していることに主に起因する．

3.2　子ども向け表現の抽出法

本稿では，3.1 での問題を改善するために，抽出手順 3 を以下に置き換える．

3. $\mathrm{D}(C,w)$ の距離があるしきい値（Th_a）よりも小さく，かつ，距離の比が定めたしきい値（Th_r）以下のときに w を子ども向け表現とする．

4　評価実験

4.1　距離算出のためのモデルと距離尺度

本研究では，距離を測るためのモデルとして単語 3-gram 言語モデルを，距離尺度には補正パープレキシティを用いた．

子ども向け表現モデルは KNC と CSJ それぞれのコーパスから単語 3-gram モデルを学習し，それらを確率ベースで補間した．単語は Chasen-2.4.4+Unidic-1.3.12 を用いて決定し，地名および人名は 1 つのクラスとしてモデル化した．大人向け表現モデルは毎日新聞記事 1 年分（MAI）と CSJ から学習した．子ども向け表現言語モデルと同様に学習した．いずれも語彙サイズを約 20000 語とした．

4.2　抽出対象

子ども向け表現を抽出する対象のテキスト集合（web コーパス）として，子ども向けと大人向けの web サイト

*龍谷大学 理工学部

表 1: web コーパス

	ページ数	単語数	文数
子ども向け	66677	51951602	4946760
大人向け	51808	57573835	4688892

表 2: 各言語モデルによる補正パープレキシティ

LM の学習データ	APP
CSJ	598.4
KNC	1288.2
KC_{conv}	684.7
KC_{prop}	620.9
KNC + CSJ	571.7
KC_{conv} + CSJ	484.1
KC_{prop} + CSJ	464.0

をそれぞれ起点として，そのリンクをたどって収集したテキストを使用した．子ども向けの web サイトの起点として，子ども向け検索サイトと企業や官公庁の子ども向け解説サイトを用いた．大人向け web サイトの起点として，龍谷大学の web サイトを用いた．表 1 に web コーパスの規模を示す．

4.3 抽出実験

4.3.1 実際の授業書き起こしに対する補正パープレキシティの比較

子ども向け表現テキストの抽出を行って音声認識用の 3-gram 言語モデルを学習した[1]．抽出時のしきい値は 2 分割交差検定で決定した．作成した言語モデルを 13 件の子ども向け授業 [5] の書き起こしに対する補正パープレキシティ（APP）で評価した．

CSJ および KNC と，抽出した子ども向け表現から種々の言語モデルを学習して比較した．以降，web コーパスから従来手法，提案法により抽出したテキストと KNC を混合したコーパスをそれぞれ KC_{conv}，KC_{prop} と表記する．結果を表 2 に示す．

抽出方法の改善により，APP を低くできていることがわかる．次に，KNC や KC と CSJ をともに用いて学習した言語モデルの評価を行った．テキストベースでコーパスを混合するとサイズの大きいコーパスでの単語 N-gram の出現カウントの影響が大きいため，それぞれのコーパスで学習した言語モデルを確率ベースで補間した [7]．

KNC と CSJ から学習した言語モデルでは APP は 571.7 であり，KC_{conv} と CSJ から学習した言語モデルでは APP は 484.1 であった．KC_{prop} と CSJ から学習した言語モデルにより，APP が 464.1 と最も小さくなった．科目ごとに分析したところ，抽出方法の改善により，社会や算数・国語（主要教科）の全てを含む 10 件の授業に対して APP の改善がみられた．

これらのことは，抽出方法の改善によって，より適切な表現が獲得されたことを示している．

[1] 4.1 節で述べた方法と同じ方法．ただし，全ての語をかな表記，語彙はコーパス中に 20 回より多く出現するものとした．

表 3: 100 文の特徴調査結果

テキストの特徴	従来抽出法	提案抽出法
子ども向け表現テキスト	43%	50%
大人向け表現テキスト	33%	27%
短い文，単語のみ	24%	23%

4.3.2 抽出されたテキストの比較

従来法で抽出したテキストと提案法で抽出したテキストそれぞれから 100web ページ取り出し，どのような特徴を持つテキストが抽出されているか調査を行った．

調査の結果を表 3 に示す．提案法により子ども向け表現テキストが含まれるページの割合が多くなっており，このことからも抽出方法が改善されていることがわかる．ただし，短い文や単語のみのテキストが抽出される問題は改善されていない．これらは文や単語の数をしきい値とすることで解決される可能性がある．また，大人向け表現がまだ多く含まれていることから，抽出方法には改善の余地が大きいことがわかった．

最後に，提案法で抽出できた子ども向け表現文の例を以下に示す．

- CD がなかったらどうなるでしょう
- 整理をしなくちゃいけないね
- ちょうないかいやボランティアなどが行うこともあるよ

5 おわりに

教師発話の言語的特徴のモデル化のために，子ども向け表現テキストの抽出法法を検討した．今後の課題としては，距離算出用の子ども向け/大人向け表現モデルの高精度化や他の抽出手法の検討などが挙げられる．また，抽出元のテキストをより大規模にして実験を行っていく予定である．

謝辞: 授業音声データを提供いただいた関係各位に感謝します．本研究は科研費の助成を受けた．

参考文献

[1] 藤原ら．講義音声認識のための LSA を利用した語彙推定手法の検討．第 3 回音声ドキュメント処理ワークショップ講演論文集, 2009.
[2] 中川ら．講義音声ドキュメントのコンテンツ化と視聴システム．信学論, Vol. J91-D, No. 2, pp. 238–249, 2008.
[3] 穂坂ら．授業音声字幕化のための学習データ分類に基づく話者依存音響モデル学習．第 4 回音声ドキュメント処理ワークショップ講演論文集, 2010.
[4] 久木ら．小学校授業の音声認識のための児童向けサイトを用いた言語モデルの構築．音講論（秋）, 1-10-17, 2011.
[5] 南條ら．初等中等教育における授業音声認識のための言語モデルの検討．信学技報, SP2011-54, pp. 13–18, 2011.
[6] 南條ら．初等中等教育の授業音声認識のための子供向け表現の抽出と言語モデル学習．音講論（秋）, 3-P-19, 2012.
[7] 長友ら．相補的バックオフを用いた言語モデル融合ツールの構築．情処論, Vol. 43, No. 9, pp. 2884–2893, 2002.