多数の低信頼性文字列からの文章の復元

瀬戸　就一†、川辺　弘之‡、下村　有子‡
金城大学短期大学部†　　金城大学社会福祉学部‡

1．はじめに

近年、聴覚障害のある大学生は年々増加している。聴覚障害学生の聴講を支援する手段は、パソコンを用いた要約筆記、速記式入力、音声認識ソフトによる音声入力など、様々な方法が用いられている。しかし、これらの方法は、学生ボランティアに高度な技術を要求する。例えば、速記入力の習得は、最低でも４ヶ月かかり、熟練のボランティアになるためには約２年の年月を要する。学生ボランティアを養成しても、学生は４年間で大学を離れてしまうために、熟練のボランティアを養成できない。そのため、ボランティアをする人々が極端に少ないという問題を抱えている。[1][2]

それを打開するために、いろいろな方法で支援する方法が考え出されている。例えば、教室内の画像や音声をインターネットで遠隔ボランティアへ送り、ボランティアはそれらをノートテイクして教室内の学生へ送り返し、講義内容を支援するシステム[3]などである。しかし、いずれの方法もインターネットを用いた大規模なシステムであり、熟練の音声入力者や熟練のノートテイカーが必要である。

本研究の目的は、多人数の初心者学生とテキストマイニング手法によって、熟練のノートテイカーの代替をさせ、リアルタイムに講義ノートを作ることである。

2．文章復元問題

学生ボランティアが入力するのは、表１の講師の講義内容（音声で説明した部分）である。表２は複数の短期大学学生ボランティア６人によって聞き取り入力された例である。

Text reproduction from many low reliable strings
†Shuichi Seto・Kinjo College
‡Hiroyuki Kawabe and Yuko Shimomura・Kinjo University

入力された単語には、いくつかのタイプミスが確認できる。各人の入力データには、次の特徴がある。

１）時間に順序がある断片データである。
２）データに任意の隙間を含んでいる。
３）打ち間違いや文字の欠落がある。

入力データ全体の集合は、以下の特徴が発生する。

１）複数の入力データに重複部分が生じる。
２）隙間が、個々の入力データごとに異なる。

このような講義内容入力データからデータ整列とテキスト抽出を行い、文章を復元する。

表１．講師の講義内容（音声）例

This is my sister Lisa. She lives in Canada.

表２．初心者６人による入力例

A:	This is my sister Risa. She lives in Canada.
B:	This is my sister. She is from Kanada.
C:	This is my sister. She lives in Canada.
D:	This is my sister Risa.
E:	This is my sister Lisa. She lives in Canada.
F:	This is my sister. She is from Canada.

3．データ整列処理

データ整列プログラムは、最初に、データを読込み、単語ごとに区切る。次に、類似した単語をそれぞれ位置合わせする。最後に、適切な

文章を作り出す。通常、２つの単語間の類似度を計算するにはLevenshtein距離[4]を使用するが、本論のケースでは多くの単語ペアが発生し計算に時間がかかる。そのため、文字コード自身を類似度の基準として用いた。単語の適切な位置合せを検索する手順を、以下に示す。

(1) 単語といくつかの空白を連結した後で、連結された単語の先頭から４文字を切り取る。

(2) アスキーコードを用い、単語をコード化する。

(3) 各列でコード化した数値の分散を計算し、列に対する分散を評価関数として使用する。もし、列に単語数がゼロ、あるいは１つならば、分散の値はゼロとなる。このケースを避けるために、ゼロや１つの場合は分散の値に大きなペナルティを与える。

(4) 分散の合計値が大きいとき、単語を後方へ移動し、空白コードを埋める。

(5) 単語のコードがすべて空白だけの列を除くために、分散値を再計算する。

(6) 分散の合計値がほとんどゼロに近づいたとき、単語の配置の最適化を終了する。

最後に多数決原理に基づいて単語を決定する。

４．結果

表 3 で示されているのは、本論のデータ整列処理で得られた結果である。ここでは講義内容の文章と、聞き取った単語を３人、４人、６人の初心者で入力した場合の抽出結果と比較して示されている。表中の４つのアスタリスク（＊）は、未確定な言葉を意味している。初心者の入力のため、文章中にはいくつかのタイプミスがある。抽出結果より、３人や４人より６人の方が良い結果が示されている。このことから、入力者のサンプル数が増えると良い結果が得られることがわかった。

５．まとめ

本論では、講義内容抽出システムについて報告した。このシステムは、複数人から入力した誤りの多い文章からでも元の文章を復元することが特長である。しかしながら、以下の点において改善が必要である。まず、テキストの再配置に対して、適切な評価関数を検討する必要がある。次に、テキストの配置に対して別の最適化手法を試す必要がある。また、サンプル数の依存についても、多くの入力者と長い文章があれば、精度が向上するのかどうかの検討が必要である。さらに、アルゴリズムの改善やハッシュコードの改良もしなければならない。

表 3. 数人の入力で得られた復元結果

学生	入力テキストから復元した内容
A-C:	This is my sister ****. She **** **** Kanada Canada.
A-D:	This is **** **** sister. She **** **** Kanada Canada.
A-F:	This is my sister ****. She lives in Canada.

謝辞　この研究は平成 19 年度金城大学特別研究費の支援を受けて行われている。謝意を表する。

＜参考文献＞

[1] 小林庸浩、(2004)「パソコン要約筆記の遠隔支援に関する現状報告」筑波技術短期大学テクノレポートVol. 11(1) pp. 15-20

[2] 日本聴覚障害学生高等教育支援ネットワークPEPNet-Japan、(2006)「ノートテイカー指導者養成講座」

[3] 立入哉 、井上かおり、宮武由佳、(2003)「音声認識を利用した聴覚障害学生学習保障システムについて」電子情報通信学会技術研究報告、NLC、言語理解とコミュニケーション Vol. 103, No. 115(20030606) pp. 43-48

[4] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10: pp. 707-710.