不規則な空白を含む時系列日本語断片データからの文章復元

瀬戸　就一[†]、杉森公一[‡]、川邊　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、聴覚障害学生を支援する広がりが顕著になってきている。日本聴覚障害学生高等教育支援ネットワーク（PEPNet-Japan）が実施した聴覚障害学生支援に関する全国調査[1, 2]の結果では、全国の大学の約 30～40%に聴覚障害学生が在籍すると報告されている。聴覚障害学生の講義受講に対する支援（情報保障）の手段には、手話、即時字幕、音声認識ソフトウェアによるテキスト提示、要約筆記（ノートテイク）など様々な方法が用いられている。特に、手話通訳による支援、パソコン要約筆記（講義などの情報保障の場合にはノートテイクとよばれることが多い）による支援は、聴覚障害学生が在籍している大学の 18.8%、13.6%とまだまだ多くはないが、今後増えていくと思われる。

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、手書きによるノートテイクの基本的な技術を習得するには少なくとも 4 ヶ月かかり、さらに熟練ボランティアとなるには約 2 年の年月を要する。また、せっかく学生ボランティアを養成しても、大学生であれば 4 年間、短大生は 2 年間で卒業を迎えてしまうため、熟練ボランティアを継続して確保することは非常に困難である[3, 4]。

そこで私たちは、多くの初心者学生ボランティアがノートテイクを行うことでこれらの問題を解決するサポートシステムの提案を行っている[5]。今までの研究では英文データで提案を行っていたが、今回は日本語データで文章を復元する新しい日本語用アルゴリズムを報告する。

2．入力データの特徴

学生ボランティアが入力するのは、講師の講義内容（音声で説明した部分）である。

入力されたデータは、いろいろなタイプミスが発生する。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。

また、入力データ全体の集合は以下の特徴がある。

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う。

3. 文字配列最適化と文章復元アルゴリズム

3.1 文字の配列および抽出の手続き

時間に順序がある文章の復元のために、文字整列および抽出を以下の 6 ステップで行う。

(a) 1 文字ずつに分解
(b) ローマ字による文字化
(c) 文字のコード化
(d) 評価関数で文字の再配列（空白の挿入）
(e) 不適当な隙間をコントロールする罰規則を備えた整列
(f) 文字抽出

3.2 キーボード配列に基づいたコード化

以前の研究[5]では、私たちは英文データで ASCII コードに基づいたコード化（ASCII コード化）を行い、文章の復元を試みた。しかし、ASCII コード化は、タイプミスを復元することはできなかった。そこで、QWERTY 配列のようなキーボード配列によってコード化する新しいコード化手法（キーボードコード化）を提案する。

3.3 分散の局所的な合計値による評価関数

2007 年度研究では、各列のコードの分散値はすべての列の合計値として計算した（全域合計評価関数）。分散の全合計値を取ることで、原理的には段落全体で良い配列になるよう探索はできるが、空白挿入数は不十分な結果であった。そこで、分散値の合計をすべての列で行うのではなく、空白挿入の位置から数列という局所的な範囲について分散の合計を取る新しい評価関数を提案する（局所評価関数）。

A Reproduction of Time Sequential Data from a Ser of Time Sequential Japanese Fragments with Random Gaps
[†]Shuichi Seto・Kinjo College
[‡]Kimikazu Sugimori, Hiroyuki Kawabe and Yuko Shimomura・Kinjo University

3.4 比較のためのデータセット

各方法の比較のために、A1～A6 のサンプル（表 1 参照）の組み合わせによる 14 個のテストセットを用いた。ソースコードは GCC version 3.4.4 でコンパイルし、実行した。

4. 結果と考察

4.1 文字のコード化

ASCII コード化とキーボードコード化を比較するために、評価関数は全域合計評価関数に固定した。結果は ASCII コード化とキーボードコード化の間には、注目するような差異が無いこと、あるいは、ミススペルはキーボード配列には依存していないことが判明した(表 2 参照)。

4.2 分散の局所合計による評価関数

局所評価関数は、文字の範囲を 20、15、10、5、4、3 文字間として適用した。単純に全体合計を局所的な合計と比較するために、私たちはキーボードコード化に固定した。新しい局所評価関数は前の全域合計評価関数より復元率が大きくなった。

4.3 入力者の人数による変化

以前の研究では、文字配列の性能は入力人数に依存しており、最良の結果を得るためには少なくとも 6 人の入力者が必要であった。局所評価関数の結果においては、5 人以上の入力者がいるとよい結果が得られることがわかった(表 2 参照)。

5. まとめ

本研究では、不規則な空白を含む日本語の時系列断片データから講義ノート（文章）の復元をするための新しいアルゴリズムを提案した。新しいアルゴリズムでは日本語のコード化、評価関数の改良を提案した。

キーボードコード化による改良で、文章復元の精度は向上しなかった。ミススペリングあるいはミスタイプについて詳細に検討する必要があると思われる。局所評価関数は文章復元の結果をよく改善し、より狭い範囲で段階的再配列が行われることで、適切な空白挿入が実行された。その計算時間は全域合計評価関数に比べると、約 2 倍の時間であった。良い性能を得るためには、空白挿入の範囲については 5 文字の範囲以内、入力者数については少なくとも 5 人以上が必要であることが明らかとなった。

今後はコード化法を改善し、ミススペリングやミスタイプの文字の復元を可能とすること、最適人数の復元率の向上を考慮し、システムの改良を行っていく予定である。

表 1. 日本語のサンプル文字入力

サンプル	誤入力･欠落データの入力例
正解	ワガハイハネコデアルナマエハマダナイ.
A1	ワイハネコテアルナエハマナイ.
A2	ワカハイハコテルナマハマタナイ.
A3	ワガイハネデアナマエマダイ.
A4	ガハハネコアルマエハダナイ.
A5	ワガイハコアルナエハマタナイ.
A6	ワカハイネデアルマエハマナイ.

表 2. 評価関数による文字復元率の比較

サンプル数（人）		6	5	4	3
全域	ASCII ｺｰﾄﾞ化	65%	55%	40%	40%
	キーボード コード化	60%	55%	65%	40%
局所	20 文字間	35%	55%	45%	40%
	15 文字間	35%	55%	40%	25%
	10 文字間	60%	70%	70%	25%
	5 文字間	80%	90%	65%	60%
	4 文字間	90%	100%	55%	40%
	3 文字間	95%	70%	65%	20%

全域：全域合計評価関数
局所：局所評価関数

謝辞　この研究は平成 20 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] 白澤麻弓(2005)一般大学における聴覚障害学生支援の現状と課題 ～全国調査の結果から～第 2 回「障害学生の高等教育国際会議」予稿集 pp. 9-10.
[2] 白澤麻弓(2005)聴覚障害学生に対するサポート体制についての全国調査 http://www.PEPNet-J.org.本文
[3] 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
[4] 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」
[5] Seto, S., Kawabe, H., Shimomura, Y. (2007) A Reproduction of Time Sequential Data from a Set of Time Sequential Fragments with Random Gaps, *Proceedings of Asia-Pacific Conference on Industrial Engineering and Management (APIEMS) 2007.*