4X-06

日本語入力手法評価のためのフレーズ集合の開発*

柳橋 良亮† 郷 健太郎‡ 木下 雄一朗§

山梨大学工学部コンピュータ理工学科

1 はじめに

文字入力手法の評価実験では, 文字入力手法を用いてフレーズを入力するというタスクを被験者に与え, 1 分間あたりに入力できる文字数 (CPM) や誤入力の割合 (エラー率) などを計測して評価する. しかし, 評価実験の際に被験者に提示されるフレーズが, その文字入力手法に適さない場合には, 正しい性能評価が不可能になってしまう. また, それぞれの文字入力手法の評価実験において提示される課題文が異なっていると, 異なる文字入力手法の性能比較が困難になってしまう. これらの問題を解決するため, MacKenzie ら [1] は文字入力手法の性能評価に用いるフレーズ集合の開発を行った. また, Tim ら [2] は, N-gram を用いて大きなコーパスから自動でフレーズ集合を生成するアルゴリズムの構築を行った. Leiva ら [3] は, フレーズの覚えやすさを考慮してフレーズ集合を生成するアルゴリズムの構築を行った.

このように文字入力手法評価のためのフレーズ集合が近年開発されているが, いずれも主として英語を対象としたものであり, 日本語文字入力手法の評価にそのまま利用することができない. また, 標準となるような日本語フレーズ集合の開発は行われていないため, 日本語入力手法の評価実験において提示されるフレーズは研究者ごとに異なっている. つまり異なる日本語入力手法の直接的な性能比較が困難である.

そこで本研究では, 日本語文字入力手法を対象に正確な性能評価と異なる手法間での性能比較を可能にするフレーズ集合の開発を行う.

2 フレーズ生成の課題

文字入力手法の評価実験を行う際に, フレーズを提示せず被験者に思いついたことを自由に入力してもらう手法を用いると, 文字入力の途中でどんな文字を入力するかを被験者が考えこんでしまう可能性がある [1]. したがって, 文字入力手法そのものの正確な性能評価を行うためには, フレーズを提示する必要がある.

また, 文字入力手法の評価実験で使用するフレーズ集合は「その言語とフレーズ集合の文字出現頻度が近いこと」, 「フレーズが長すぎないこと」, 「フレーズが覚えやすいこと」を考慮することが重要である [1].

評価対象となる文字入力手法は, SNS での入力を想定しているものや, インターネット検索での入力を想定しているものなど, 各入力手法によって課題としたいフレーズが異なる. よって, それぞれの用途に合わせたフレーズ集合を開発する必要がある. そこで, 元コーパスと作成したフレーズ集合の文字の出現頻度を近づけることによって元コーパスの特徴に近いフレーズ集合を作成する. しかし, 文字の出現頻度を考慮したうえで複数のフレーズ集合を手作業で作成することは困難であるため, 自動でフレーズ集合を作成するアルゴリズムを構築する.

3 フレーズ生成の流れ

3.1 マルコフ連鎖テーブル

マルコフ連鎖を利用して文生成を行うためマルコフ連鎖テーブルを作成する. マルコフ連鎖テーブルとは N-1 単語に後続しうる 1 単語を探索するためのテーブルである. 「山梨に行ったことはあるが、東京に行ったことはない。」という文章を元に N=4 のマルコフ連鎖テーブルを作成した例を, 表 1 に示す. 例えば「た/こと/は」という 3 単語の単語列に後続する単語をマルコフ連鎖テーブルから探索すると, 表 1 の 4 行目の「ある」という単語と, 13 行目の「ない」という単語が後続することと, それぞれ後続する確率が 1/2 であることがわかる.

表 1: N=4 マルコフ連鎖テーブル

1	山梨	に	行っ	た
2	に	行っ	た	こと
3	行っ	た	こと	は
4	た	こと	は	ある
5	こと	は	ある	が
6	は	ある	が	、
7	ある	が	、	東京

8	が	、	東京	に
9	、	東京	に	行っ
10	東京	に	行っ	た
11	に	行っ	た	こと
12	行っ	た	こと	は
13	た	こと	は	ない
14	こと	は	ない	。

3.2 文生成アルゴリズム

N=4 のマルコフ連鎖テーブルを用いた文生成の流れを以下に示す.

1. マルコフ連鎖テーブルの任意の行から先頭の 3 単語を生成文の先頭としてコピーする.
2. 生成文を品詞解析し, 先頭の単語が { 名詞, 動詞, 形容詞 } でない場合は 1 の処理を再度行う.
3. 生成文の末尾 3 単語を探索キーとしてマルコフ連鎖テーブルの全ての行における先頭 3 単語とマッチングを行う.
4. 生成文の末尾 3 単語と, ある行の先頭 3 単語が一致したとき, その行の 4 単語目を後続する単語のリストに追加する.

*Developing Phrase Sets for Evaluating Japanese Text Entry Techniques

†Yoshiaki Yanagihashi - University of Yamanashi
‡Kentaro Go - University of Yamanashi
§Yuichiro Kinoshita - University of Yamanashi

5. 全ての行に対する探索が終了した時点で後続する単語のリストの中から 1 単語を無作為に選んで生成文の末尾に追加する.
6. 後続する単語のリストを空にする.

ステップ 3 から 6 の処理は, ステップ 5 の処理における後続する単語のリストが空になるまで繰り返す.

しかし, このアルゴリズムでは元コーパスによっては非常に長い文章を生成してしまう. 長い文章はフレーズとしては適さないため, 文生成アルゴリズムを応用し, ある程度の長さのフレーズを生成するアルゴリズムを開発した.

3.3 フレーズ生成アルゴリズム

ある程度の長さのフレーズを生成するため一定の単語数でフレーズの生成を打ち切ると, 日本語として成立しないフレーズが生成される可能性がある. そこで, 一定の単語数に達した際にフレーズの生成を打ち切るかどうかを判断するアルゴリズムを作成した. 以下にその流れを示す.

1. 生成文の品詞解析を行う.
2. 生成文の末尾の単語が { 名詞, 動詞, 形容詞, 助動詞} であるならば, 後続する単語の探索を継続する.
3. 文生成アルゴリズムと同様にマルコフ連鎖テーブルを探索し, 後続する単語のリストを更新する.
4. 後続する単語リストの中の { 動詞, 助詞, 助動詞 } となり得る単語から 1 単語を無作為に選んで生成文の末尾に追加する.
5. 後続する単語のリストを空にする.

以上の処理をステップ 4 の処理における後続する単語のリスト中に該当する単語が存在しなくなるまで繰り返す.

このフレーズ生成アルゴリズムを用いて複数のフレーズ集合を生成し, 文字の出現頻度が元コーパスと十分に近いものを選択して利用する.

4 フレーズ集合の試作と評価

本研究では, 吉川英治の著書である「宮本武蔵」を青空文庫 (http://www.aozora.gr.jp/index.html) から取得してコーパス (以下, 元コーパス) とした. 宮本武蔵は約 108 万文字からなる小説である. フレーズ生成における打ち切り単語数を 5 単語とし, N=4 のマルコフ連鎖テーブルを用いて約 4,000 文字からなる 500 フレーズを収録したフレーズ集合の開発を行った. 収録したフレーズの一部は以下のとおりである.

- 家として古い方
- 防風林に囲まれた
- 追われてきたので
- 情熱を込めた
- 他人の物みたいに

元コーパスと作成したフレーズ集合の文字出現頻度を降順に 5 つ取り出したものを表 2 に示す. 元コーパスと生成したフレーズ集合において出現頻度の上位 5 文字が一致していることがわかる.

表 2: 元コーパスとの文字出現頻度の比較

元コーパス「宮本武蔵」		生成したフレーズ集合	
カナ文字	文字の出現頻度 (%)	カナ文字	文字の出現頻度 (%)
イ	5.81	イ	5.71
ノ	4.56	ノ	5.23
ウ	4.36	ウ	4.29
シ	4.06	シ	4.21
タ	3.87	タ	4.21

4.1 文字入力評価実験

作成したフレーズ集合を用いて 2 本のジョイスティックを用いた文字入力手法「いとね」[4] の性能評価を行い, 従来研究で用いられたフレーズを用いた場合との比較を行った. 被験者は 1 人であり, 評価尺度は CPM とエラー率である. 従来研究では実験開始後の 5 セッションの平均で 34.97CPM, エラー率 3.98%であったが, 同様の条件で本研究のフレーズ集合を用いた場合は 39.33CPM, エラー率 4.81%であった. すなわち, CPM の差は少なく, エラー率は高くなった. 従来研究では, ことわざから選出した 105 フレーズ (1,094 文字) 中から 30 フレーズを被験者に提示しており, 表 3 に示すとおり従来研究で用いたフレーズにおける文字出現頻度が異なっていることが原因である可能性がある. 具体的には「カ」と「ハ」の出現頻度が高く現れている. つまり, [4] の実験の課題フレーズは十分に大きなコーパスの特徴を考慮できておらず, [4] の文字入力手法は提示したフレーズに特化している可能性がある.

表 3: 従来研究との文字出現頻度の比較

従来研究 [4]		生成したフレーズ集合	
カナ文字	文字の出現頻度 (%)	カナ文字	文字の出現頻度 (%)
カ	5.32	イ	5.71
ノ	4.95	ノ	5.23
ハ	4.95	ウ	4.29
イ	4.13	シ	4.21
シ	3.85	タ	4.21

5 おわりに

本稿では, 日本語文字入力手法の評価実験において提示されるフレーズが研究者ごとに異なっているために文字入力手法の性能比較が困難であるという問題を指摘した. その問題を解決するため, マルコフ連鎖テーブルを用いた文生成アルゴリズムを利用した日本語フレーズ集合の生成手法を提案した. 今後は, 提案手法を使って, 元コーパスをより大きなコーパスや SNS データなどに変更した新たな課題フレーズ集合を開発する. また, 4.1 節で示した実験の追試を被験者を増やして実施する.

参考文献

[1] MacKenzie, I, S. and Soukoreff, R, W. Phrase Sets For Evaluating Text Entry Techniques. In Proc. CHI EA '03, pp. 754-755, 2003.
[2] Paek, T. and Hsu, B, P. Sampling representative phrase sets for text entry experiments: a procedure and public resource. In Proc. CHI '11, pp. 2477-2480. 2011.
[3] Leiva, L, A. and Trilles, G, S. Representatively memorable: sampling the right phrase set to get the text entry experiment right. In Proc. CHI '14, pp, 1709-1712. 2014.
[4] Go, K. Konishi, H. and Matsuura, Y. Itone: A Japanese Text Input Method for a Dual Joystick Game Controller. In Proc. CHI '08, pp. 3141-3146, 2008.