1S-3

チャンキングとカテゴリを利用した携帯電話向け予測入力手法の提案

蛯澤 綾乃　　松原 雅文　　Goutam Chakraborty　　馬淵 浩司
岩手県立大学ソフトウェア情報学部

1. はじめに

近年，携帯電話の普及により，携帯電話上で日本語入力をする機会が増加した．しかしながら，このような端末は携帯性を重視しているため，通常の PC で使用されるような大きなキーボードや多数のキーを備えることができない．よって，必然的にキー数が少なくなり，それを補うため打鍵数が多くなり，迅速な入力が難しくなってしまう．

そこで本研究では，携帯端末上において文字入力に要する打鍵数を減少させるために，チャンキングとカテゴリに着目し，予測変換を組み合わせる手法を提案する．チャンキングを利用することで，少ない打鍵数で複数の語を予測入力することができ，打鍵数減少につながる．また，カテゴリを使用することで話題に沿った予測が可能となり，かつ返信元メールだけでは予測できない部分をカバーできる．

本稿では，提案手法の概要を示し，行った実験の結果から提案手法の有効性を示す．

2. 既存手法

動的略語展開手法 Nanashiki とはメールの返信を対象としている予測機能で，返信元のメールに含まれる単語が予測変換候補の上位に出力される[1]．返信元メールに着目することにより，メールのやりとりなどのコミュニケーションにおける入力において，効率的な予測が可能となる．

しかし，一単語ごとの予測しか出来なかったり，動的略語展開を行うための文脈情報が引用メッセージ中に存在していない場合，ユーザが入力したい単語の提示率が低くなってしまうなどの問題点がある．

3. 提案手法

3.1. 概要

上記で述べた問題を解決するために，チャンキングとカテゴリを利用した手法を提案する．チャンキングを利用することで，少ない打鍵数で複数の語を一度に予測入力することができ，打鍵数減少につながる．また，同じカテゴリのメールに含まれるチャンクを利用することでより話題に沿った効率のよい入力が可能となり，かつ返信元メールだけでは予測できない部分をカバーできる．

Prediction Input Method for Mobile Phone Using Chunk and Category
Ayano Ebisawa, Masafumi Matsuhara, Goutam Chakraborty and Hiroshi Mabuchi
Faculty of Software and Information Science, Iwate Prefectural University

処理の流れを図 1 に示す．具体的には，受信したメールを形態素解析により単語へと分割し，名詞が連続した場合にチャンキングを行い，チャンクを作成する．次に，受信メールを対象にクラスタリングを行い，返信元のメールが含まれるクラスタを最適カテゴリとし，そこに含まれるチャンクを予測候補とする．そして，それぞれのチャンクに対し重み付けを行い，優先順位を決定する．

返信の際は文字を入力し，入力文字に当てはまる予測変換候補を優先順に表示する．

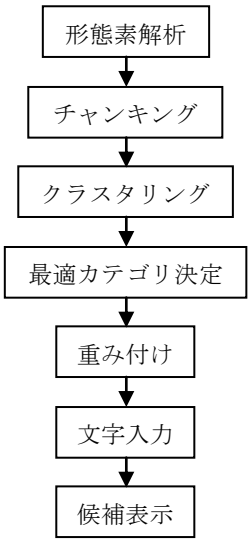

図 1　処理の流れ

3.2. チャンキング

チャンキングとは，複合語を同定する際に用いられる処理である[2]．本手法では，形態素解析を行い，名詞が連続した場合にチャンクを生成する処理を行う．

3.3. カテゴリ

カテゴリの優先候補設定とは，プライベートやビジネスなど TPO に合わせて設定したカテゴリの単語が，優先的に候補として表示される手法である[3]．しかしながら，適切なカテゴリが何かわからない場合があったり，優先候補は手動設定しなければならないなど，ユーザに負担がかかってしまう．この問題解決のため，提案手法においてはカテゴリをクラスタリングで決定する．受信メールを対象にクラスタリングを行い，返信元のメールが含まれるクラスタを最適カテゴリとする．

3.4. 優先順位（重み付け）

予測変換候補のチャンクを並び替えるため，重み付けを行い，優先順位を決定する．

優先順位は以下の通りである．
1. 返信元メールに含まれるチャンク
2. 最適カテゴリ内のメールに含まれるチャンク

この優先順位において，それぞれ別に重み付けを行う．その重み付けは以下の通りである．
1. チャンクの出現頻度
2. 出現頻度の平均－標準偏差
3. 時系列

チャンクそのものが多く出現している方がよく使われると考えられるので，最初の重み付けにチャンクの出現頻度を利用する．チャンクの出現頻度が同じ値であった場合，次の重み付けを利用する．

次の重み付けでは，チャンクを構成する単語の出現頻度の平均から標準偏差を引いたものを利用する．例えばチャンクを構成する単語のうち 1 つの単語の頻度が 100 で，他の単語の頻度が 1 だった場合，他の単語の頻度が低いのにも関わらず 1 つの単語の頻度が高いだけで平均は高くなってしまい，正しい重み付けが出来ない可能性がある．よって，平均だけではなく，平均－標準偏差を利用することで適切な値にすることが出来ると考えられる．平均－標準偏差が同じだった場合，次の重み付けを利用する．

古いメールよりも新しいメールのほうが話題に沿った予測が可能と考えられるので，最後の重み付けに時系列を利用する．1 通のメールの中でも，文頭より文末のほうが新しい話題と考え，これが優先される．時系列には，メールのヘッダから取得した時間情報と，本文先頭から数えて何文字目にチャンクが存在するかという 2 つの重み付けを利用する．時系列を利用することで，必ず全てのチャンクを並び替えることが可能となる．

4. 実験

返信対象チャンクが予測変換候補の上位に出力されることで打鍵数減少につながると考え，予測変換順位に関する検討を行った．

4.1. 実験方法

著者が保有している受信メール 200 件を対象とする．200 件からランダムに選択した 10 件のメールに対する返信メールを利用する．クラスタリングツール bayon[1]を使用する．クラスタリングの際，名詞とその出現頻度を入力データとし，全てのデータの所属度が 0.5 以上になるまでクラスタリングを行う．文字入力は最初の 1 文字のみとする．そして，入力したいチャンクが予測候補の何番目に出力されるかを調査した．

提案手法との比較対象として，カテゴリを使用せず返信元メールのみを利用した手法，返信元メールを優先せずカテゴリのみを利用した手法，両方使用せず全ての候補を提示した手法を用いる．

4.2. 実験結果および考察

実験結果を表 1 に示す．返信元のみの場合，第 1 候補に表示された数は多かったが，提案手法と比べて候補として提示できないものが 22 個存在した．また，カテゴリのみは提示できる候補が提案手法と同じであったが，第 5 候補以下が 12 個と精度が低くなってしまった．さらに，両方使用しない場合は提示できた候補が 92 個と最も多かったが，第 5 候補以下が 54 個と多く，最も低いもので第 232 候補が存在するなど，精度が低かった．

提案手法は，第 1 候補に表示されたチャンクが 47 個で 66.2%と高い割合であり，全てのチャンクが第 4 候補以内に出力され，精度も高い．よって返信元メールの他にカテゴリを利用した本手法は有効であるといえる．

表 1　実験結果

	提案手法	返信元	カテゴリ	なし
第 1 候補	47	38	31	19
第 2 候補	16	8	18	2
第 3 候補	6	3	5	12
第 4 候補	2	0	5	5
第 5 候補以下	0	0	12	54
合計	71	49	71	92

5. おわりに

本稿では，チャンキングとカテゴリに着目し，予測変換を組み合わせることにより，携帯端末上において文字入力に要する打鍵数を減少させることを可能とする手法を提案した．

優先順位に合わせて重み付けを行い，予測変換候補を並び替えることで，入力したいチャンクを予測変換候補の上位に出力させることが出来た．少ない打鍵数でチャンクを予測入力することで，打鍵数の減少が期待できる．

今後は，返信元メールに含まれるチャンクと最適カテゴリ内のメールに含まれるチャンクにおいて，それぞれ別に行っていた重み付けを一本化するための検討や，クラスタリングの所属度の最適な値に関する検討などを行う予定である．

参考文献

[1] 小松弘幸，高林哲，増井俊之：“動的略語展開を利用した文脈をとらえた予測入力”，情報処理学会論文誌，44(11)，pp.2538-2546，November 2003.
[2] 土屋雅稔，注連隆夫，高木俊宏，内元清貴，松吉俊，宇津呂武仁，佐藤理史，中川聖一：“機械学習を用いた日本語機能表現のチャンキング”，自然言語処理，14(1)，pp.111-138，January 2007.
[3] SHARP ケータイ Shoin6：“http://www.sharp.co.jp/products/sh905i/text/08_shoin.html”，2007.

[1] bayon：http://alpha.mixi.co.jp/blog/?p=1049