3N-07

利用者の入力単語予測のための単語共起頻度データベース選択方式

小松 淳　　鷹野 孝典

神奈川工科大学 情報学部 情報工学科

1. はじめに

モバイル端末の入力方式エディタ（IME）では，利用者の入力単語を予測して提示する機能が普及している．一方，多くの検索エンジンでは，問い合わせの単語列を入力する際に，問い合わせ単語及びその問い合わせ単語に関連のある単語群の候補を提示する，キーワード提示機能が見受けられる．本研究では，両機能のことを「入力単語予測機能」と呼んで参照する．入力単語予測機能では，利用者個人の入力履歴や複数の利用者による大規模な入力履歴を統計的に解析したもの（ホットキーワード[3]）がしばしば用いられる．モバイル環境においては，利用者の現在地点に応じた絞込みも可能である．

しかしながら，文書ソフト等で入力単語予測機能を使用しながら文章を入力している際に，最初は個人の入力履歴で予測しているだけで十分であったが，次第に，検索エンジンを利用して関連単語を調べ始めたならば，「ホットキーワード」を利用したり，「入力している文章の内容」を分析して予測した方が，利用者の暗黙的な要求に応じた入力単語の予測が可能であると考えられる． 利用者の入力状況に応じて，入力単語を予測する有用な手法の研究[1-3]が盛んに行われている一方，このように，「利用者個人の入力履歴」，「ホットキーワード」，あるいは別の入力単語履歴のうち適切なものを選択することで，利用者の入力意図に適った予測結果を提示することには十分対応できてない．

本研究では，利用者の入力単語予測を目的として，システムにより単語予測結果として提示された候補単語集合のうち，利用者が実際に選択した単語情報に基づいて，適切な単語共起頻度データベースを選択する方式について提案し，実際に構築したプロトタイプを用いた実験により，提案方式の実現可能性を検証する．

2. 提案方式

提案方式のシステム構成を図 1 に示す．提案方式では，利用者の最初の入力文字に応じて提示する「予測単語」と，利用者の入力単語の想起支援を目的として，入力した単語に関連のある「関連単語」の提示も行う（図 2）．本方式では，利用者が Web ブラウザ上等で単語入力を行う際に，特定の分野・目的・状況ごとに分類して単語の共起頻度データベースを構築しておき，利用者の入力意図に応じた入力単語予測に利用する． 利用者の実際の単語選択結果をフィードバック情報として得ることで，各単語共起頻度データベースに対する優先順位を動的に設定し，利用者の入力意図に適した入力単語予測が可能となる．

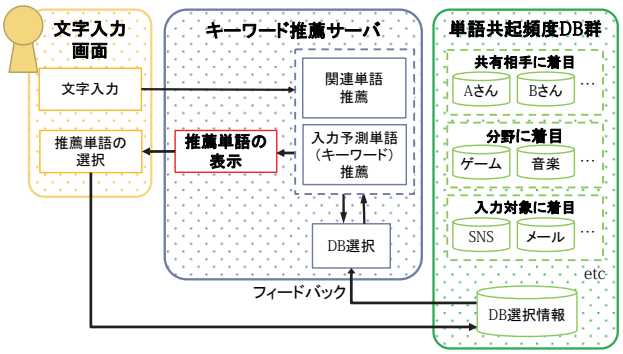

図 1 システム構成図

図 2 単語が推薦されている様子

単語共起頻度データベースの選択は，以下の手順で実行される．初期状態では，予測単語は文字コードでソートして提示されるものとする．また，アプリケーションやサービスごとに独立に作成された単語共起頻度データベースを $r_1, r_2, \ldots, r_n$ とする．

Step-1: 各単語共起頻度データベース $r_1, r_2, \ldots, r_n$ の単語集合 T を単語のスコア s 順にソートして提示する．

$$T = T_1 \cap T_2 \cdots \cap T_n, \quad T_i = \{t \in r_i\}$$

A method for selecting co-occurrence word databases for word prediction

Atsushi Komatsu, Kosuke Takano, Department of Computer and Information Sciences, Faculty of Information Technology, Kanagawa Institute of Technology

ここで，スコアは単語 t の出現頻度 $freq(t)$である．ただし，同じ単語が異なるデータベース中に存在する場合は，値の大きい方，加算，平均等を用いてスコア s を算出する．

Step-2: 利用者が選択した単語 t が格納されているデータベース r_x を特定し，データベース r_x の選択回数 c_x に 1 を加算する．ただし，t が複数の n 個のデータベースに格納されている場合は，全てのデータベースに対して 1 または平均値である $1/n$ を加算する．

Step-3: データベース r_x の選択回数 c_x に応じて，データベース r_x 中の単語スコア s を増加させる．

$$s = \alpha \times freq(t) \times log_2(c_x + \beta) \quad (\alpha \text{ と} \beta \text{は定数})$$

なお，選択回数 c_x の重みは，利用者が直近に選択した m 種類のデータベース中の単語スコアに適用する．

Step-4: Step-3 の単語スコアに基づき単語をランキングして提示する．以下，Step-2～4 を繰り返す．

3. 実験

Web ブラウザで動作するプロトタイプを用いて実験を行う．単語共起頻度 DB として表 1 に示すものを構築した．入力文章として，表 2 に示す 3 つの文章を用いた．下線単語は，評価のための着目単語で，添え字番号は着目単語 ID である．

表 1 単語共起頻度データベース

	説明
DB_1	走れメロス
DB_2	蜘蛛の糸
DB_3	安倍首相(前回)の所信表明演説
DB_4	麻生元首相の所信表明演説
DB_5	福田元首相の所信表明演説
DB_6	鳩山元首相の所信表明演説
DB_7	菅元首相の所信表明演説
DB_8	T 先生とのメール

表 2 入力文章

(1)	実験(1)のプログラム(2)を作るには、php(3)が良さそう。実験システム(4)の構成(5)についてスライド(6)にまとめておく。ゼミ(7)の時に、鷹野(8)先生にも相談(9)してみよう。卒論(10)のクロスチェックもしないといけないから、研究室(11)に顔を出す。停電(12)が終わった後に無線(13)機器(14)の電源を入れたら、ネットワーク(15)の接続状況が良くないようだ。
(2)	若者(1)のモラル(2)の低下(3)が問題(4)になっていると感じていたので、教育(5)の強化(6)プログラム(7)を推進(8)の意向に共感いたしました。他にも、女性(9)や高齢者(10)、ニート(11)やフリーター(12)の積極的(13)な雇用(14)を推進(15)するという意向についても、今後(16)の日本(17)の経済(18)成長のために良い取り組み(19)だと感じました。
(3)	地球温暖(1)化の防止(2)の取組(3)みとして太陽光(4)発電の導入(5)、建物(6)の緑化(7)、バイオマス(8)の利用など様々あるが、まずは身近(9)なことから始め、日本(10)を美しい(11)国にすることが大切(12)だ。

また，Step-3 の単語スコア s の算出において，$\alpha = 1, \beta = 2, m = 2$としている．

(1)提案手法と(2)単語の出現頻度のみでのランキング（DB 選択なし）により入力単語を予測し，比較した結果を図 3 に示す．図 3 において，着目単語は ID 順に入力されている．この結果より，提案方式では，利用者が選択した単語情報に基づいて，単語共起頻度データベースを選択し，そのデータベース中の単語スコアの重み付けを行う事により，単語の出現頻度のみをスコアとした場合の方式と比較して，入力単語予測の精度が向上していることが確認できる．

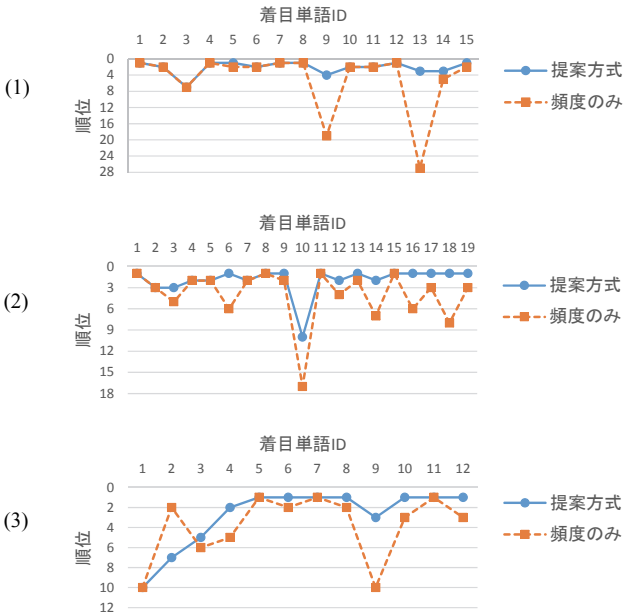

図 3 入力単語の予測結果の比較

4. まとめと今後の課題

提案方式を用いて，利用者の入力意図に応じた入力単語の予測が実現可能であることを確認した．今後は，大規模なテストデータを用いた既存手法との比較実験結果に基づいて単語共起頻度データベース選択アルゴリズムを改良するとともに，提案方式の有用性を検証していく予定である．

参考文献

[1] 中村将人，木村昌臣，”意外な検索キーワードを推薦する手法の提案”， 第 72 回情報処理学会 全国大会講演論文集(1)，pp.895-896，2010.
[2] 渡辺奈夕子，岡本昌之，菊池匡晃，飯田貴之，佐々木健太，堀内健介，山崎智弘，大村寿美，服部正典，”閲覧 Web ページからの第 1 検索キーワード抽出に基づく検索支援”，情報処理学会論文誌，Vol.53(7)，pp.1783-1796，2012.
[3] 竹内郁雄，”未踏の第 20 期スーパークリエータたち”，情報処理学会誌 情報処理，pp1324-1331，2014.