予測入力の拡張

- 効率的な日本語入力方法 -

1．背景

本プロジェクトは日本語予測入力システムである PRIME (PRedictive Input Method Editor) の作成および発展である。なお、以下 PRIME という名称は「本プロジェクトで開発を行った日本語予測入力システム」という意味を持つ。本節では、日本語予測入力の背景述べる。

日本語予測入力とは、日本語入力手法のひとつであり、POBox などが代表例である。日本語予測入力の特徴は、利用者が入力した最初の数文字から目的の単語を予測することである。予測された単語の候補は、利用者の入力が進むにつれて、順次更新されていく。利用者は、入力したい単語が候補に現れた時点で、目的の単語を選択すればよい。図１は、入力「こん」から「こんにちは」を予測し、入力単語として選択した例である。

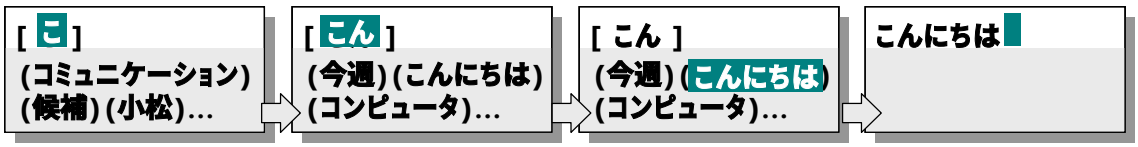

図 1：予測入力の例

日本語予測入力を用いれば、少ないキー入力での文章作成が可能である。そのため日本語予測入力は、携帯電話や PDA などのキー入力のコストが高いデバイスを中心に、広く普及している。

2．目的

本プロジェクトの目的は大きく分けて「新しい予測方法の作成」および「実用性の高いソフトウェアの作成」の 2 点である。

「新しい予測方法の作成」とは、例えば、利用者の状況に応じた単語候補の予測や、手書き入力時におけるひらがなと漢字の混じった入力からの単語候補の予測などである。これらの新しい予測方法は、日本語入力をさらに効率的にかつ利用者の思考に近い形とする。

「実用性の高いソフトウェアの作成」とは、プロトタイプの作成に留まらずに実際に日常利用可能ソフトウェアを作成することを目標とする。具体的な内容には、辞書およびマニュアル等の周辺資料の充実。簡便な導入方法の確立、長期運用を見据えたソースコードの再設計などが含まれる。

3．開発の内容

主な開発内容は「連文節予測」「手書き予測」「周辺環境の整備」「単語登録の改善」である。以下、順に説明を行う。

・連文節予測

まず、連文節予測について説明する。例えば、「未踏ソフトウェア」という単語は登録されておらず、「未踏」と「ソフトウェア」だけが単語登録されている条件を考える。連文節に対応した予測とは、この条件下で「みとうそ」という入力に対して「未踏ソフトウェア」という単語の予測を可能にすることである。また逆に、常に「未踏」と「ソフトウェア」を続けて入力しているのであれば、「みと」という入力からも「未踏ソフトウェア」の予測を行うことである（図 2）。

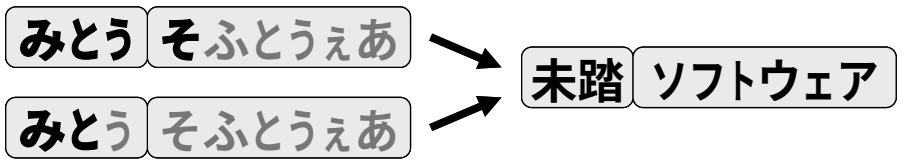

図 2：連文節予測の例

本プロジェクトでは、通常の連文節変換に加えて、上記の連文節予測も新たに作成した。連文節にまたがった入力候補の予測では、候補数が膨大な数になるため候補の絞り込みが課題となるが、入力履歴を元にした単語間の共起関係により、候補数の発散を抑止している。

・手書き予測

手書き予測とは、ペンデバイスなどによる手書き入力のための候補予測のことである。

手書き入力は、キーボード入力とは入力の特徴が異なるため、手書き入力のための予測手法を作成する必要がある。例えば手書き入力の場合、利用者は漢字を直接入力することが可能であるし、また「しょう油」のようにひらがなと漢字を交ぜて入力することもある。本プロジェクトではこれらの特徴を踏まえ、手書き入力に対応した単語予測を実現した。

具体的な内容としては、例えば「未」という入力から「未踏事業」の予測や、「しょう油」から「醤油」の予測などを可能とした。この手書き予測は本プロジェクトの応募時には計画に盛り込まれていない。プロジェクト実行中に新規に実装した内容である。

・周辺環境の整備

周辺環境の整備とは、本プロジェクトの成果物をより実用的に利用するための、辞書や簡易な導入方法の実現などのことである。

具体的な内容は「辞書の拡充」、「簡易な導入方法の実現」、「既存の IME との親和性の向上」、「辞書のデータ構造の再設計」などである。

「辞書の拡充」により、語彙数は 14 万語から 24 万語へと増加した。また辞書データは毎月更新されており、その中には時事単語も登録されている。また、利用者ごとに独自に辞書を作成する機能も用意した。

「簡易な導入方法の実現」として、GNU の Autotools を活用したインストール方法を提供している。そのため、導入者は特別な設定をする必要なく本成果物の導入が可能である。また各 OS 用のパッケージも用意した。用意したパッケージは Debian, RPM (RedHat, Turbo など), Gentoo, FreeBSD, MacOS である。これらのパッケージは本プロジェクトのフリーソフトウェア活動の成果物でもある。

「既存の IME との親和性の向上」の内容は、各 IME 辞書からのコンバータの作成と、他の IME の変換プロトコルとの語幹機能の実現である。これらを活用することにより、利用者がこれまでに利用していた IME のデータをそのまま活用可能するとともに、本成果物への移行をより容易にする。

「辞書のデータ構造の再設計」により、従来よりもより多くの学習データを素早く扱えるようになった。また、漢字からひらがなへの逆変換なども可能にした。

・単語登録の改善

これまでは利用者による手作業であった単語の登録を、自動的に、または半自動的に行う方法を提案した。

具体的には、利用者がこれまでに参照した文書のデータベースや、インターネットなどを活用した方法をもちいて、登録単語の品詞判定や単語情報そのものを自動的取得する方法を提案した。

4. 従来の技術（または機能）との相違

本プロジェクトの特徴のひとつは、成果物の PRIME がフリーソフトウェアであることである。現在開発が続けられている日本語予測入力システムのフリーソフトウェアは、この PRIME のみである。また技術的な特徴として「連文節予測」および「手書き予測」が挙げられる。これらの詳細は、第 3 節において述べている。

5. 期待される効果

・短期的な効果

現在開発が続けられている日本語予測入力システムのフリーソフトウェアは、本プロジェクトで開発されている PRIME だけであり、本プロジェクトは、フリーソフトウェアによる日本語入力環境の選択肢を広げ、入力効率を大きく改善するものである。

また、PRIME の開発に伴い作成された各種ライブラリや辞書は、汎用性を持つように設計されている。そのため、ライブラリや辞書のみを開発等に利用して、PRIME 自体は使用しないという使い方も可能である。

・長期的な効果

最近では、ネットワーク対応固定電話からのメール作成、ビデオレコーダでの検索・データ入力、オンラインゲームでのチャットなど、日本語入力に対応した家電・機器類は多い。しかし、これらの家電・機器類にはキーボードを備えているものが少なく、文字の入力コストは非常に高い。そのため現時点では、家電・機器類の日本語入力環境は効率的であるとは言い難い。

本提案の予測入力は、入力コストの高いデバイスとの相性がよい。そのため、本事業を通じて蓄積されたノウハウが、これらの家電・機器類の日本語入力環境を向上させることは間違いない。また入力コストを下げるという点には、アクセシビリティの向上という側面もあり、身体的ハンディキャップを和らげるための活用も期待できる。

6．普及（または活用）の見通し

PRIME はすでにフリーソフトウェアとして公開されており、自由に活用することができる。ダウンロード数・ユーザ数などの具体的な数字は不明であるが、第３節で述べた通り、各ＯＳ向けのパッケージが多数作成されていることから、需要の高さを伺える。

今後、Windows や携帯端末向けの実装を実現することにより、さらなる普及が期待できる。

7．開発者名

小松弘幸
東京工業大学 情報理工学研究科 数理・計算科学専攻 博士課程 2 年
komatsu@taiyaki.org

参考 URL:
http://taiyaki.org/prime/
http://sourceforge.jp/projects/prime/