2010 年度 未踏 IT 人材発掘・育成事業 採択案件評価書

1. 担当PM

後藤 真孝 PM（産業技術総合研究所 情報技術研究部門
メディアインタラクション研究グループ長）

2. 採択者氏名

チーフクリエータ： 笠原 誠司（奈良先端科学技術大学院大学）
コクリエータ： なし

3. 委託金支払額

1,792,000 円

4. テーマ名

大規模データを用いた統計的日本語校正アプリケーション

5. 関連Webサイト

http://cl.naist.jp/chantokun/

6. テーマ概要

日本語学習者の書いた文章の校正を大量の言語データを利用して修正するアプリケーションの開発を行う。

現在、lang-8 や Livemocha などといった、言語学習 SNS に日本語を学習したい外国人が多くの作文を寄せている。また、日本語学校に通い熱心に学習している外国人もいる。しかし本人の意思にもかかわらず、日本人に指示してもらうことができる量と時間は限られている。

また近年、日本語版 Wikipedia や青空文庫など大量の日本語言語データが手に入

りやすくなっている。それに加え、情報技術の発達により、インタラクティブなアプリケーションを Web ブラウザのみで利用できる仕組みが発達してきている。

これらの状況をふまえ表題の通り、言語データを統計的に処理し、日本語文章の校正を行うアプリケーションを作成する。完成品は Web アプリケーションとして公開する事で広く世界で使用してもらい、日本の文化、社会に興味を持ってもらいたいと考えている。

7. 採択理由

多量の日本語テキストに基づいて、日本語学習者が書く日本語文章中の誤りを検出・指摘し、訂正候補を提示するシステムの提案である。正しい文章だけを正例の学習データとして利用するのではなく、「学習者がどの単語をどの単語と間違えやすいか」という負例のデータを日本語学習者が集うWebサイトから収集して活用することで、精度を高めることができる特長を持つ。「200 億文を使いこなす」、「無料で本当に使える日本語校正アプリケーションを目指す」と明言しているところが素晴らしい。

笠原君は、自身の英語学習時に苦労して大きく上達した経験から、語学学習の支援に高いモチベーションをもっているところが素晴らしく、学内で運用されている twitter アプリ関連の開発経験からも、実用的な校正アプリケーションを公開し、広く使われるところまで根気強く頑張ることが期待できる。日本語学習者用の Web サイトから人手による添削データを入手するだけでなく、まだ添削されていない文章を、笠原君の実現した日本語校正アプリケーションによって自動的に添削し、その結果を元の Web サイトにアップロードして還元するところまで、是非挑戦して欲しい。笠原君の頑張りに大いに期待したい。

8. 開発目標

本プロジェクトの目標は、日本語学習者と教育者を支援するため、言語データを統計的に処理し、日本語文章の校正を行う日本語学習者の入力支援アプリケーションを開発することである。具体的には、以下の項目等に取り組む。

① 学習用テキストの下処理
② 格助詞誤りエンジンの実装・調整
③ コロケーション誤り訂正エンジンの実装・調整
④ Web アプリケーションの作成

9. 進捗概要

未踏プロジェクト開始段階では、言語データを統計的に処理することで格助詞の誤りを検出する手法を部分的に実現し、バッチ処理のような単純なインタフェースのプロトタイプを作成していたに過ぎなかったが、プロジェクト開始後、コロケーション訂正エンジンの作成、統計処理をする上でのゼロ頻度問題を解決するスムージング処理の実装、より完成度の高いインタフェースへ向けた Web アプリケーションの実装、誤りを含む文も柔軟に解析できる形態素解析器の作成等に取り組んだ。4 月にプロジェクトレビューをした際には、訂正エンジンとインタフェースとの結合をしたWebアプリケーションのプロトタイプシステムの実装がある程度進んでおり、そのデモンストレーションを交えた有意義な議論ができた。その後、Web アプリケーションとしてのクオリティを高める作業を着々と進めると共に、訂正エンジンの精度向上と高速化に取り組んで、着実にプロジェクトを進めた。成果報告会前には、Web アプリケーションを公開してソーシャルメディアとの連携も開始し始め、成果報告会ではデモンストレーションを交えて魅力的に成果を発表した。

10. プロジェクト評価

実際に日本語学習者が、ウェブブラウザ上で日本語文章を入力すると、その誤りを自動検出して訂正候補を提示し、選択等により訂正が可能なインタフェースをウェブアプリケーション「Chantokun」(http://cl.naist.jp/chantokun/)として実現した。日本語学習者が誤り易く、しかも既存のソフトウェアでは訂正できない「格助詞」(「が」「の」「を」「に」「へ」「と」等)と「コロケーション」(名詞と動詞との対応等)の誤りを、ウェブ上から集められた大規模なテキストデータを用いて、統計的自然言語処理に基づいた手法で自動検出した。格助詞では、単語トライグラムを用いて誤り検出をしただけでなく、訂正候補の提示もおこなった。当初実現したプロトタイプのウェブアプリケーションは操作性が低かったが、利便性を追求して改良を重ね、最終的には、テキストの各文字を入力する最中に、リアルタイムにその場で誤り箇所が提示されるだけでなく、テキストの入力に専念するインタラクティブモードと、訂正候補も一緒に見やすく提示されるディテールモードの二つを用意して、非常に使いやすい機能を実現した。これをすでに一般公開し、エンドユーザが誰でも利用できるようにした点を極めて高く評価する。当初の計画を越えて、和英辞書などのウィジェットの配置や、ソーシャルメディアとの連携、未知語の検出等の機能も実現した点も優れている。

11. 今後の課題

Web アプリケーションの公開まではしたものの、対外的なアピールはまだまだこれからであり、広く使ってもらえる状態まで、引き続き改良とアピールを進めていくことを

期待したい。また、語学学習者が作文添削を他の人から受けることが出来る Lang-8 等の既存 Web サービスとの連携にも、今後是非取り組んで欲しい。