IPA

2009 年度下期未踏 IT 人材発掘・育成事業　採択案件評価書

1. 担当PM

勝屋　久 PM（Venture BEAT Project 主宰）

2. 採択者氏名

チーフクリエータ：松本　一輝（フリーランス）
コクリエータ：なし

3. プロジェクト管理組織

株式会社京王 IT ソリューションズ

4. 委託金支払額

4,500,000 円

5. テーマ名

英文添削コーパスを活用した英文入力支援・校正ソフトウェアの開発

6. 関連Webサイト

なし

7. テーマ概要

インターネットの発達、グローバル化の進展に伴い、ビジネス・プライベートを問わず、個人が英文を作成し、外国人とコミュニケーションを行う局面は非常に多く見られるようになった。しかし英語ノンネイティブが英文を作成する上で助けとなる、英文入

力支援ソフトウェアは、スペル チェッカーの域を出ない未熟な製品しか市場に登場しておらず、この分野において未だ IT はエンドユーザの潜在ニーズに応えることができていない。

本提案では、語学学習 SNS「Lang-8」の利用者により生成された膨大な英文添削データを活用することで、従来の英文入力支援ツールとは一線を画する、実用的な入力支援・校正アプリケーションを開発する。

本提案で開発するアプリケーションは、Web ブラウザ上で動作するツール（プラグイン）として実装される。入力者が Web ブラウザ上で英文を入力すると、自動的かつ非同期に API サーバに問い合わせを行い、バックエンドデータベースに蓄積された添削データの解析結果を元に、文法上・表現上の誤りである可能性が高い部位を指摘表示すると共に、より適切な修正候補をわかりやすくリストアップするものである。本提案では、この一連のプロセスに必要な、Web ブラウザプラグイン、バックエンドデータベース並びに API サーバソフトウェアのすべてを開発する。

本アプリケーションは、特に東アジア圏の英語ノンネイティブが全世界に情報を発信する上で、言語の壁を克服するための有力なツールと成り得るものである。

8. 採択理由

既存の英文校正支援ツールはスペルチェッカー機能どまりで、文脈に踏み込んだ単語の並びなどの校正には限界がある。当開発プロジェクトはネイティブ（外国人）が実際に手作業で作成した既存保有添削データにもとずく、文構造を元とする的確な英文校正機能の提供ができる点と、クライアント側に DB をもつことはなく、オンライン上の膨大な言語資源にリアルタイムで問い合わせをすることで、高い出力精度を実現できる点が優位性であると感じた。

開発者の松本さんは、期間中は現職を休職をし、当プロジェクトに専念する意気込みもあり、より質の高いサービスをグローバルへ展開できる可能性も感じられた。

9. 開発目標

本提案で開発するアプリケーションでは、インターネット上で人手によって作成された豊富な数のコーパス（データ）を利用することで、この問題の解決を図る。オンラインで展開される各種の語学学習 Web サービスにおいて、記述言語が英語であり、かつインターネット上に公開されているものを対象としてデータの分析を行い、これによ

り「添削前の誤りを含んだ文」及び「添削後の正確な文」それぞれについての構文解析データを生成し、これを構文木データベースに蓄積する。

誰もが犯しがちな作文ミスは、似通った文脈において、すでに誰かの手によってオンライン上で引き起こされているはずであり、また、別の誰かの手によって添削がなされているはずである。つまり現在作成中の英文について、この構文木データベースを横断的に検索することで、類似した過去の誰かの過ちを発見することができ、その添削データを基に入力文を機械的に校正することが可能になるのである。このデータベースに対して高速にリアルタイム探索を行う API サーバソフトウェアの開発を行うことで、オンラインで動作可能な、全く新しい英文入力・校正支援アプリケーションの実現が可能となる。

従来より存在するワープロソフト等に組み込みの「英文入力校正ツール」は、単語のスペルミスの他には、例えば以下のような、極めて明快な文法上の単純ミスにおいてのみ、効力を発揮するものであった。

しかし、たとえば下記のような例においては、入力文中にスペルミスもしくは極めて明快な文法上の誤りが存在しているわけではなく、これらの文がネイティブに与える「違和感」を機械的に指摘・校正することは非常に困難であった。

I am learning Japanese for almost three years.
I am working in mobile communication company in Japan now.
Anyway, I’m happy because he came back.

本提案で開発するアプリケーションは、上記のような入力文に対しても、以下のように校正案を自動で提供することが可能である。

I am learning Japanese for almost three years.
→ I ~~am~~ **have been** learning Japanese for almost three years.

I am working in mobile communication company in Japan now.
→ I am working ~~in~~ **at a** mobile communication company in Japan now.

Anyway, I’m happy because he came back.
→ Anyway, I’m happy ~~because~~ he**’s** ~~came~~ back **now**.

以上のような機能を実現させるため、本システムでは、インターネット上で人手によって作成された豊富な数のコーパス（データ）を利用した。これらのデータにより構築される構文木データベースは、極めて大規模化することが予想されるため、本システムはクライアントサーバシステムの形態をとり、構文木データベースはサーバ側でのみ保持することとした。クライアントソフトウェアは必要に応じてサーバにクエリを発行することで、修正済みの英文情報を取得することができる。

本システムが持つ構文木データベースは、運用期間の長さに比例して各種データを蓄積していくため、最終的にはデータベース容量が数百 GB（ギガバイト）～数 TB（テラバイト）へと拡大することが予想される。またその容量増加は連続であるため、サービスの停止を伴った大規模メンテナンスによる拡張で、都度対応することは運用上難しい。さらに本システムでは入力された英文の解析等にかなりの演算能力を要するが、瞬間利用ユーザ数が時間帯により大きく変動することが予想される実運用において、必要とされる最大演算能力を常時確保することは、インフラコストの面から許容できない。

この問題に対処するため、本システムの設計ではデータベースを数多くの小規模サブデータベースに分割し、それぞれが PC サーバ上で運用可能なサイズに留めることとした。また英文解析等を行うワーカープログラムも同様に分散処理化し、ピーク負荷に合わせた動的なワーカ数調整等、柔軟な運用が可能となる設計を採用した。

また、クライアントプログラムとして、Firefox 上で動作する拡張プラグイン（Greasemonkey スクリプト）を作成した。本プラグインは任意の Web ページ を閲覧中に、透過的に動作するものであり、ページ内に設置されている入力フォーム（TextArea 等）を検出すると、本システムの入力画面を起動するトリガーアイコンを表示する。このトリガーをクリックすることで、ユーザは Web ページの入力フォームに直接文章を入力する代わりに、本システム専用編集フォームを通して英文の入力・編集作業ができるようになり、編集後の結果は元のフォームへと適宜反映される。

さらに、本システムにおいては、出力精度の向上を目指すため、オンラインコーパスと併せて、英文校正者が手動で入力した、本システム専用の校正コーパスも利用することとした。これを実現するため、英文校正データの効率的な入力を可能とする専用の GUI 入力インターフェース画面を構築し、上述の Firefox プラグインから呼び出し可能とした。

10. 進捗概要

中間報告会でもコンセプトとプロトタイプを作成しており、ほぼ計画どおりプロジェクトは進捗した。期間後半に、コーパス（データ）により構築される構文木データベースの構築において、パフォーマンスを最適化するために、データベースを数多くの小規模サブデータベースに分割し、それぞれが PC サーバ上で運用可能なサイズに留めることや英文解析等を行うワーカープログラムも同様に分散処理化し、ピーク負荷に合わせた動的なワーカ数調整等、柔軟な運用が可能となる設計など開発面においても課題を抱えたが、高いレベルでのプログラミング能力と情熱で乗り越え、成果物を期間内に開発をした。

11. 成果

Microsoft Word や JustSystem ATOK に代表される製品が有する校正機能と比較して、大幅に踏み込んだ提案を行う文書校正ソフトウェアの開発に成功した。また、Web サービスとして実装しているため導入が比較的容易である点、ブラウザ上で透過的に使用できるため、様々な Web サービスと併せて幅広く利用出来る点も利点である。さらに、英文校正に用いるデータベースをクライアントから分離し、オンラインで提供する設計を採用したことによって、本システムの本運用に伴いデータベースが順次拡充されることから、英文校正精度の大幅な精度向上が期待される。

12. プロジェクト評価

従来のワープロソフト等に組み込みの「英文入力校正ツール」の単語のスペルミスの領域を超えて、ネット上の人手によって作成された豊富な数のコーパス（データ）を利用し、文章自体にネイティブに与える「違和感」を機械的に指摘・校正することを支援するソフトウェアの開発をした。オンラインで展開される各種の語学学習 Web サービスにおいて、記述言語が英語であり、かつインターネット上に公開されているものを対象としてデータの分析を行い、これにより「添削前の誤りを含んだ文」及び「添削後の正確な文」それぞれについての構文解析データを生成し、これを構文木データベースに蓄積する。

誰もが犯しがちな作文ミスは、似通った文脈において、すでに誰かの手によってオンライン上で引き起こされているはずであり、また、別の誰かの手によって添削がなされているはずである。つまり現在作成中の英文について、この構文木データベースを横断的に検索することで、類似した過去の誰かの過ちを発見することができ、その添

削データを基に入力文を機械的に校正することが可能となる。このデータベースに対して高速にリアルタイム探索を行う API サーバソフトウェアの開発を行うことで、オンラインで動作可能な、全く新しい英文入力・校正支援アプリケーションの実現が可能となる。

いままでにないユーザーインタフェースと利用可能なレベルまで開発できたことの開発実現力・情熱・独創性は評価できる。特に、コーパス（データ）により構築される構文木データベースの構築において、パフォーマンスを最適化するために、データベースを数多くの小規模サブデータベースに分割し、それぞれが PC サーバ上で運用可能なサイズに留めることや英文解析等を行うワーカープログラムも同様に分散処理化し、ピーク負荷に合わせた動的なワーカ数調整等、柔軟な運用が可能となる設計など開発面においても創意工夫をした努力を認めたい。また、最終報告会では多くのお客様にも賞賛され、確実に多くの人のニーズを満たすソフトウェアであることが確信できた。

13. 今後の課題

本格的な一般公開を実施する上で、サービス提供に必要十分なサーバ処理能力を確保するため、米 Amazon 社が提供するクラウドサービス「EC2」等を活用する予定である。これにあたり、瞬間負荷に応じて動的にワーカ数を増減させ、サービスランニングコストを最小に留めるための制御機構を導入する必要があるため、近日中に当該機構の開発を行う予定である。上記課題を解決した後一般公開を行い、一般ユーザの試用感触を得た後、適宜改善等を施し、順次サービスの商用化等に踏み切る予定である。今後も必要な時点で個別にアドバイスを行ってゆきたい。