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

１．背景

現在、lang-8(http://lang-8.com/)や Livemocha(http://www.livemocha.com/)といった言語学習者用 SNS で日本語学習者が多くいるように、日本語を学習したいという需要は多い。

また、教育者が日本語学習者の書いた文章を添削する際に、自動である程度訂正しておいてくれるとかなり助かるという話を日本語教師から聞いた。特に「ご飯を食べる」を「ご飯が食べる」と書いてしまうといった格助詞の誤りや、コロケーションと呼ばれる語と語の組み合わせの問題、例えば「猫が鳴く」を「猫が吠える」と書いてしまうといったものは、間違いが起こりやすい（図 1）上に、文章が表す意味を考慮しなくても自然言語処理の技術のみで訂正できる余地がある。英文での統計的自動校正プロジェクトは前例があるが、日本語での学習者向けアプリケーション前例はいまのところない。Microsoft 社の Word などのワープロソフトでも簡単な文法のチェックなどは行ってくれるが、おかしな言い回しや、コロケーションの誤りはチェックしてくれない事が多い。また、格助詞は日本語学習者にとって誤りを起こしやすいものの一つだが、格助詞誤りも現在のところ検出されない。

近年、インターネット上での情報爆発により、大規模な日本語の言語データを利用できる状況になっている。

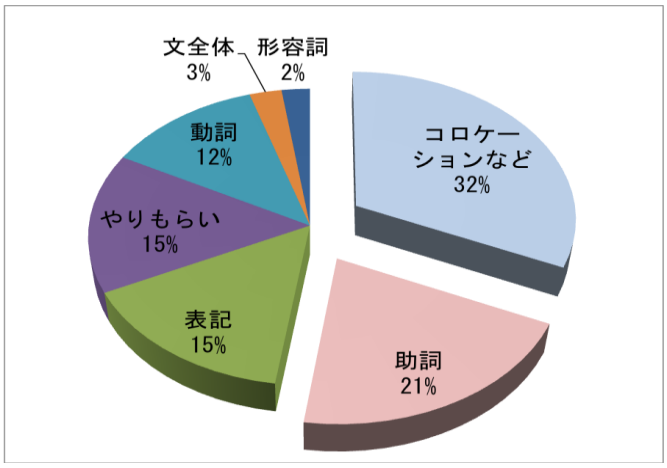

図１　学習者の誤り箇所

２．目的

コンピュータで、与えられた文章中のおかしな部分を検出し、正しそうな候補を提示するアプリケーションを提供できれば、日本語学習者と教育者の両者に大きなメリットがあるとし、本プロジェクトを実施した。

統計的自然言語処理の手法により日本語学習者の入力支援アプリケーションを作成し、多くの方が利用できるように Web アプリケーションという形で無料公開し、日本語学習者と教育者に活用してもらい、日本語使用人口の増加に貢献することが目的である。同時にソースコードはオープンソースソフトウェアとして公開し、日本語入力支援ソフトウェアのさらなる発展を推し進めたい。

３．開発の内容

・格助詞誤りの検出・訂正候補提示

格助詞の誤りを検出し、訂正候補を提示するため、単語 n-gram モデルと呼ばれる手法を用いた。格助詞とは用言に付き、文の中での意味関係を表す品詞であり、現代語では、「が」「の」「を」「に」「へ」「と」「より」「から」「で」などがある。

格助詞の誤りを検出する場合は、あらかじめ大量のテキストを学習データとして、単語 n-gram の条件付き確率を求めて語の並びと確率を辞書として保持しておく。

次に誤り検出をしたいテキスト中での格助詞前後の単語 n-gram について、辞書を参照して行き、著しく確率の低いものがあれば誤りとして検出できるという仕組みである。さらに、検出された単語列に対し、格助詞を他の格助詞に入れ替えた場合の確率も調べ、その中で確率の高い組み合わせから順に提示すれば、学習者が独力で訂正する手助けになる。

・コロケーション誤りの検出・訂正候補提示

コロケーションの誤りを検出する場合には、どの名詞とどの動詞が使われやすいかという情報を得るため、係り受け解析を行い、共起関係を推定の一部に取り込む事が有効である。共起関係を調べる手段として、相互情報量を用いた。

・提供方法

Web アプリケーションとして公開するメリットとして、Web ブラウザ上で実行でき、インストールなどの複雑な処理をしなくても簡単に利用できる、という点がある。また、大規模な辞書データをサーバー側のメモリに読み込めば、クライアント側では読み込む必要がないので、メモリや記憶装置を占有する事なく、またモバイル機器などのそれほど高機能でない端末でも利用できる可能性を持っている。

図 2 にシステム構成の概要を示す。

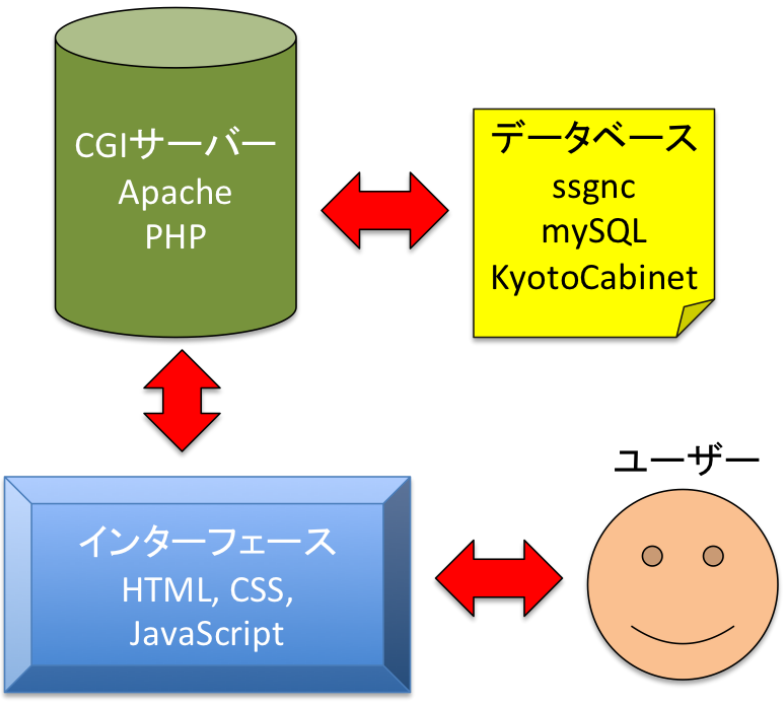

図 2　システムの構成

・インターフェース
図 3 に完成したインターフェースを示す。
(a) インタラクティブモード
　このインターフェースでは入力すると同時に、誤り箇所に誤りの種類に対応した色の下線が引かれる。
(b) ディテールモード
　ディテールモードはインタラクティブモードと異なり、文入力エリアと誤り表示エリアが別々に表示されている。
　アプリケーション内の中心あたりにある青い枠のテキストエリアにテキストを入力することができ、入力内容は即座に下の誤り表示部に反映される。誤り表示エリアでは、文は単語ごとに区切られて表示されており、誤り箇所を下線で示すというインタラクティブモードと同じ表現を用いることで、混乱することなく二つのモードを使いこなすことができる。
　訂正候補は誤り検出された語の真下に表示され、クリックすることで、誤り提示部、入力エリアの両方を訂正することができる。入力文と訂正候補の位置を合わせるために table を使用している。
　インタラクティブモードとディテールモードには相互にリンクが張ってあり、入力内容を保持したまま、二つのモードを行き来することができる。

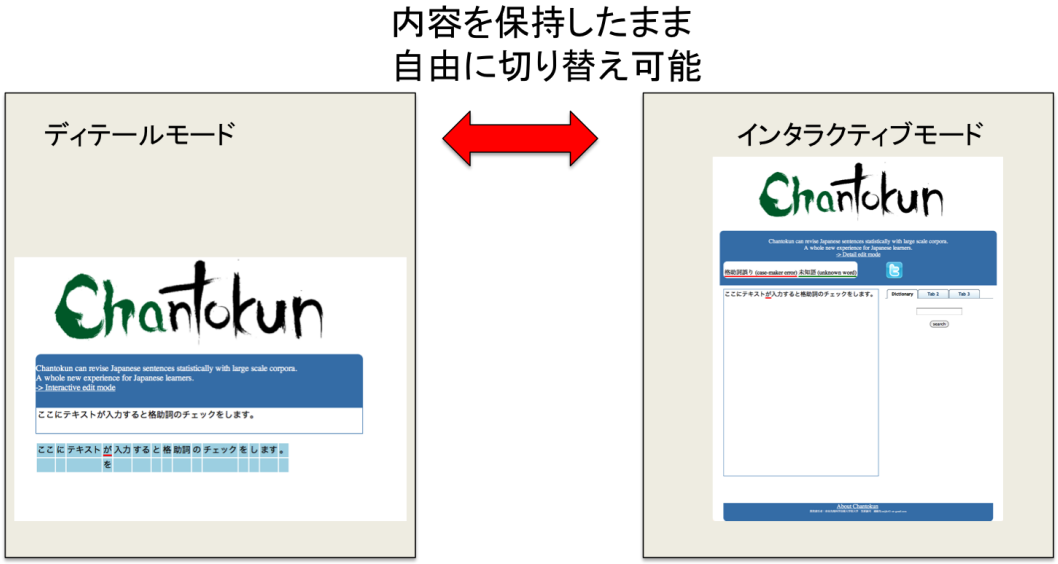

図 3　完成した二つのインターフェース

４．従来の技術との相違
　本プロジェクトの成果、Chantokunは日本語学習者が誤り易い格助詞などの誤りを訂正することができる。これは他のオンライン上の校正アプリケーションにはない機能である。大規模なテキストデータを使用しているため、高い精度で校正することができる。
　また日常的に使用するために、今までのオンライン上の校正ソフトウェアとくらべ、わかりやすいインターフェースを作成した。

Webアプリケーションであるためネットワークにつなげばすぐに使用することができる。また、リソースを消費することなく、大規模なデータの恩恵を得ることもできる。

表１　既存製品・サービスとの比較

	即応性	無料	柔軟性	誤りデータ	NoisyChannel	言語
ちゃんとくん	○	○	○	○	○	日
Word	○	×	×	×	×	日英
Just Write!	○	×	×	×	×	日
日本語校正ツール	○	○	×	×	×	日
なつめ	○	○	○	×	×	日
Lang-8	×	○	○			多
Lingo project	○	？	○	○	×	英

表 1 は既存製品・サービスとの比較表である。日本語校正機能を持つもので、一番広く使用されているものはマイクロソフトの Word であるが、Word では人手で作成したルールにマッチするものしか校正できない。一方、本プロジェクトで作成したアプリケーションでは、「ご飯が食べる」といった、単語の組み合わせからくる誤りも校正することができる。JustSystems の Just Right!や、日本語文章校正ツール(http://www.japaneseproofreader.com/)など、より校正に力を入れた製品もあるが、仕組みとしては Word と同じく人手でのルールによって校正している。

東京工業大学で研究されている「なつめ」は大規模なテキストデータを使用して日本語学習支援をするという点では同じであるが、なつめのインターフェースは辞書に近く、ワープロのようにインタラクティブに作文に集中することはできない。

2009 年度の未踏事業の成果に「Lingo」があるが、Lingo は対象言語が英語である、誤りデータしか用いていないのでカバーできない文が多い、2011 年 8 月現在は公開されていない、などの点で差がある。

学習者が作文添削を受ける際の選択肢として Lang-8 や Livemocha(http://www.livemocha.com/)などの添削 SNS があるが、これらは人手での添削なので、精度は高いが、すぐに添削結果が帰ってこないという特徴がある。

５．期待される効果

本システムにより、日本語に興味がある人が独学で学習をすすめることが出来るため、他国とのコミュニケーションを促進することが出来る。日本語教師も添削の負担が減り、高度な教育に集中することが出来る。

６．普及の見通し

海外には約300万人の日本語学習者がいるが、一方で教師の数は4万4千人しかおらず、かなり不足している。本アプリケーションは世界中から無料で使用できるため、需要が高く、大きく普及することが期待できる。

７．クリエータ名（所属）

笠原誠司（奈良先端科学技術大学院大学）

(参考). 関連Webサイト

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