アラビヤ語形態素解析エンジンの開発と学習者向け辞書システムへの応用

１．背景

現在、国際社会におけるアラビヤ語の重要性は日々高まりつつあり、実際に日本でも学習者の増加が見られる。しかし、アラビヤ語の学習において「辞書が引けない」ということが問題となっており、アラビヤ語の辞書を自在に引けるようになるには５年かかる、とすら言われている。どのような言語においても、「文中に出現する語の形」と「辞書の見出し語」は異なり、辞書を引く際には「辞書の見出し語」を把握している必要がある。例えば、英語の文章を読んでいる際に、developed という語を調べたいときは、ed をはずして develop という語で辞書を引く。アラビヤ語の辞書引きが困難である理由は、活用が激しく、語同士の結合や文字の欠落が頻繁に起こり、「文中に出現する語の形」と「辞書の見出し語」が大きく異なることである。

２．目的

当ソフトウェアは、自然言語処理の基礎技術である形態素解析を用いることで、アラビヤ語学習者の辞書を引く作業を補助することを目的としている。形態素解析によって、通常は学習者が行う「文中に出現する語」から「辞書の見出し語」を導く作業をコンピューターが代行できるため、学習者は「文中に出現する語の形」そのままで辞書を引くことが可能となる(図１)。またアラビヤ語学習を支援することで、最終的にはアラブ・イスラーム地域との相互理解と相互交流を促進することも、このソフトウェアの目的である。

図１.
辞書引きにおける、
形態素解析の役割

3．開発の内容

開発したソフトウェアについて、システムアーキテクチャ、実際の使用例、オリジナルのアラビヤ語形態素解析エンジンの３項目に分けて記述する。

3.1 システムアーキテクチャ

本システムは Java と Postgres を用いたウェブアプリケーションとして実装されている。サーブレットコンテナとしては Tomcat を用いている。図２に、本システムのアーキテクチャと利用形態をまとめた。プログラムはサーバーに設置されており、ユーザーはブラウザもしくは Flash のインターフェースを通じて、本ソフトウェアを利用する。

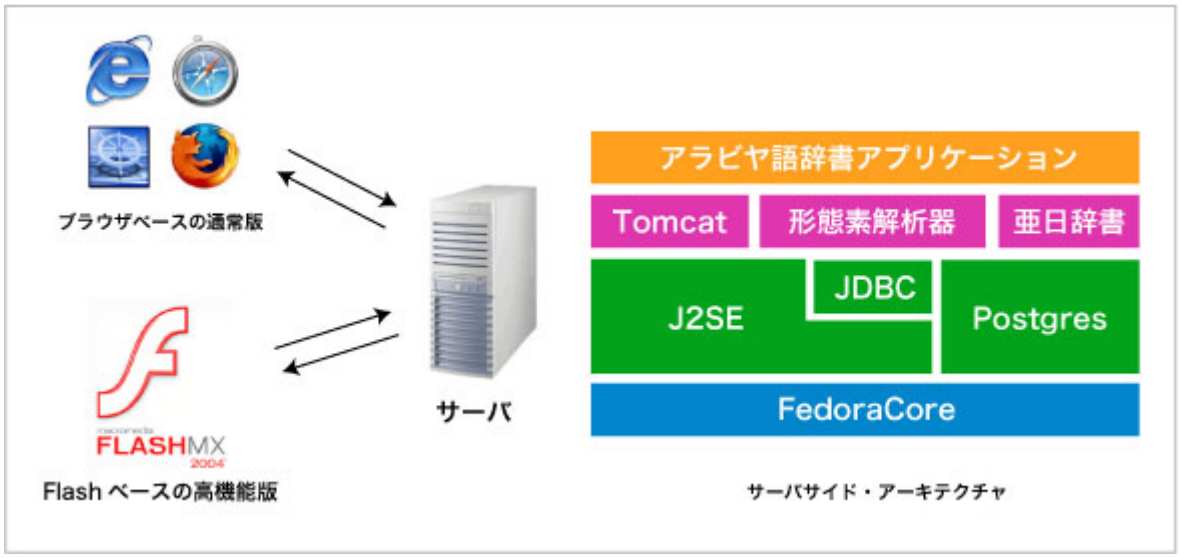

図２．システムアーキテクチャと利用形態

3.2 実際の使用例

ここでは例として أتكلم باليابانية وبالعربية (私は日本語とアラビヤ語を話します；アタキャッラム・ビルヤーバーニーヤ・ワビルアラビーヤ)という文章を、Flash ベースの高機能版を用いて読んでみる。スクリーンショットは次ページの図３となる。

まず、ユーザーは画面右下の INPUT タブより、意味を調べたい文章を入力し、検索ボタンを押す。通常の辞書アプリケーションは、一度にひとつの単語しか調べることができないが、当アプリケーションでは複数の語を一度にまとめて調べることができる。ユーザーの入力語はサーバーサイドで形態素解析され、辞書 DB に送られ、言葉の意味と分かち書きの結果とともに、ユーザーに返される。

ユーザーはキーボードの左右キーで、文章を移動させ、入力した語すべての意味を確認することができる。入力した語の下に表示されているのは、その語の最も一般的な意味だが、キーボードの上下キーを用いることで、別の意味を探すことが可能である。

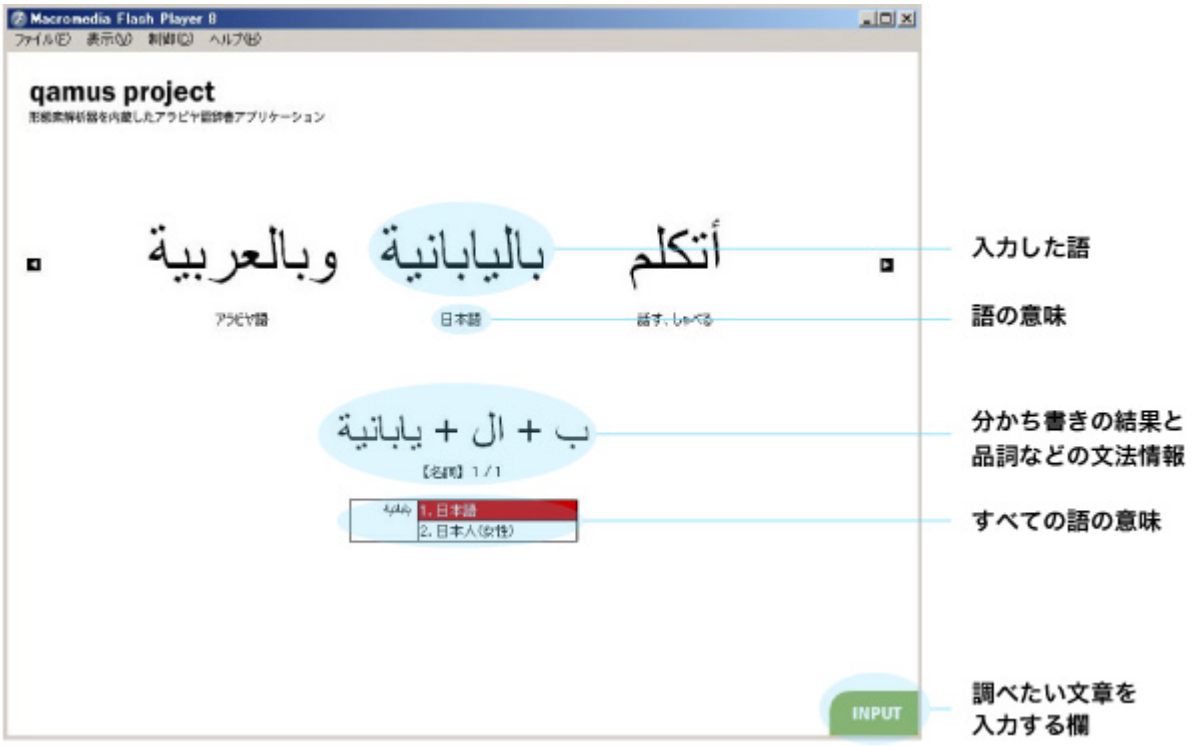

図３．実際の使用例とスクリーンショット

3.3 アラビヤ語形態素解析エンジンの詳細

本アプリケーションの実現に当たっては、アラビヤ語の形態素解析エンジンをスクラッチで実装した。アラビヤ語は英語と同じように、語と語の間にスペースがあり、その単位では分かち書きの必要はないが、スペースで区切られた単位内で形態素が連結するため、分かち書きが必要となる。また、アラビヤ語の品詞は大きく動詞・名詞・文字の３つに分類できる。

本エンジンは、語の分割コンポーネントと、３つの品詞を解析するコンポーネントの合計４つのコンポーネントから構成され、それぞれが連携することで解析を実現している。

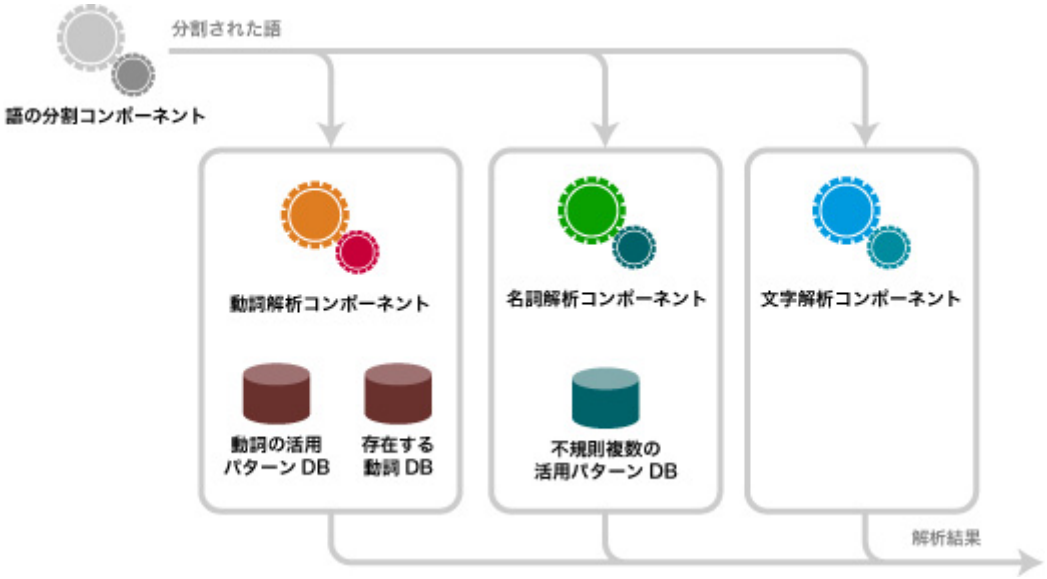

図４．形態素解析器のコンポーネントと処理の流れ

４．従来の技術（または機能）との相違

まず、既存のアラビヤ語辞書ソフトウェアと比較して、学習者が文中の語の形そのままで、意味を調べられるという、大きな優位性がある。数字としては、アルジャジーラの web サイトのニュースを、形態素解析を行わないで直接調べた場合、語が見つかる確率は 55.4%であったのに対し、形態素解析を行うことで 90.4%まで確率を高めることができた。

また、単体の形態素解析エンジンとしての特徴として、従来の解析手法は活用をすべて展開して辞書に載せていたが、本実装ではまず「活用パターンの辞書」を用い、予期されうる解を導き、その後、解の存在を辞書で確認するという手順を踏むことで、解析速度が速く、未知語への対応力が高いという点が挙げられる。

５．期待される効果

当ソフトウェアの実現により、アラビヤ語の初級者が中級・上級へと学習を続ける割合が向上することが期待される。アラビヤ語ははじめの上り坂が特にきつい言語であり、それを少しでも和らげることで、言葉を学ぶ辛さよりも楽しさを多く味わえるようになるはずである。さらには、まだ第二外国語としては一般的でないが、今後重要な意味を持つであろうこの言語が、少しでも多く一般に認知されれば、開発者としてこれほど嬉しい事はない。

６．普及（または活用）の見通し

本ソフトウェアは近日中に一般にリリースされ、誰でも自由に使うことが可能となる予定である。広報活動として、日本国内のアラビヤ語の教育機関や、SNS のアラビヤ語関係のコミュニティーを中心に告知を行う。またアラブ地域で活動をおこなっている NGO などの団体にも積極的に呼びかけを行い、実際の活動の中で役立てて頂ければと思う。

７．開発者名（所属）

岩井貴史（慶應義塾大学環境情報学部）
植村さおり（慶應義塾大学大学院政策・メディア研究科）

８．開発者ＵＲＬ

http://www.al-mintaz.org/