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

Developing the Arabic Morphological Analyzer and the Dictionary for Learners

岩井 貴史　慶應義塾大学環境情報学部
Takafumi IWAI　Faculty of Environmental Information, Keio University

植村 さおり　慶應義塾大学大学院政策・メディア研究科
Saori UEMURA　Graduate School of Media and Governance, Keio University

三浦 稔隆　慶應義塾大学総合政策学部
Toshitaka MIURA　Faculty of Policy Management, Keio University

奥田 敦　慶應義塾大学総合政策学部教授兼大学院政策・メディア研究科委員
Atsushi OKUDA　Professor, Faculty of Policy Management
Graduate School of Media and Governance, Keio University

{t03124ti, saori, s03890tm, assalam}@sfc.keio.ac.jp

概要

今日、国際社会におけるアラビヤ語の重要性は増加の一途をたどり、実際に学習者も広がりを見せているが、その学習において「辞書が引けない」ということが問題となっている。それは、文中に出現する語形と、辞書の見出し語の語形が大きく異なるためである。そこで我々は、形態素解析の機能を内蔵したアラビヤ語の辞書システムを開発している。形態素解析により見出し語を導くことで、学習者は文中に出現するそのままの語形で辞書を引く事が可能となる。本稿では、主にアラビヤ語の文法モデルと、Java を用いて実装されたアラビヤ語の形態素解析エンジンについて説明する。

1. はじめに

2. アラビヤ語学習の重要性と難しさ

2.1. アラビヤ語の現代的意義

アラビヤ語は世界に約２億人の母語としての話者を持つ言語であると同時に、約 15 億人とも言われているイスラーム教徒の信仰の中心にある言語である。ある文化圏を理解するための最も基本的な方法は、その地域に住む人々の言語を勉強することであり、その意味においてアラビヤ語の重要性は日に日に高まっている。実際にアラビヤ語の学習者も増加の傾向にあり、2003 年には NHK のテレビ外国語講座において、初めてアラビヤ語講座[1]が開設された。

2.2. 辞書の引けない言語

このようなアラビヤ語の一般化が進む中で、辞書が引けないということが学習の障壁となっている。辞書とはいうなれば「見出し語」をキーとした単語のデータベースである。辞書を引き、単語の意味を調べるためには「見出し語」を把握している必要がある。多くの言語において通常、文中に出現する語は、活用などにより見出し語とは異なる語形をしており、辞書を引く際には引く人が文中の語より「見出し語」を推察して言葉を探す必要がある。たとえば、英単語動詞の過去形"developed"を調べる際には"-ed"をはずして"develop"という語で辞書を引く。

しかし、アラビヤ語においては、この文中の語から「見出し語」の推察が困難を極めるほど、語形の変化が激しい。たとえば、「彼らは書く」という意味の動詞"＿＿＿"(ヤクトゥブーナ)は、「彼は書いた」という意味の"＿＿"(キャタバ)という語形で辞書に載っているが、このような人称と時勢を変えるという動詞の最も基本的な活用においてでさえ、文字数が２倍に増え、語頭の文字が変化する。

また、「手の中に」を意味する"＿＿"(ビヤディン)という語は、名詞の"＿"(ヤドゥ/手)と前置詞の"＿"(ビ/中に)が結合した語であるため、これらを誤ってひとつの語と判断して辞書を探しても、意味を調べることはできない。

このようにアラビヤ語では、複雑な活用や文字の結合・欠落が頻繁に起こり、見出し語の判定を困難にしている。そのため、アラビヤ語の辞書を自在に引けるようになるためには 5 年かかるとすら言われている。

3. 形態素解析への着目

3.1. 形態素解析とは

形態素解析とは自然言語処理の基礎技術であり、文章を、変化しない最も小さい単位である形態素に分割し、品詞を解析することである。日本語の形態素解析エンジンとしては ChaSen[2] や MeCab[3]が有名である。ChaSen を用いて「お待ちしております」という文章を解析すると図 1 のようになる[4]。

図 1. お待ちしておりますを形態素解析した例

ここで注目するべきは、形態素に分割され、活用を戻した原形は、辞書の見出し語そのものであるという点である。つまり、複雑な活用がされている文章であっても、形態素解析をかけることで、辞書の見出し語を得ることができ、それをキーとして単語のデータベースを検索し意味を調べることができる。図 2 に、辞書引きにおける形態素解析の効果についてまとめた。

図 2. 辞書引きにおける形態素解析の効果

3.2. アラビヤ語における形態素解析

文章に対して形態素解析を行うと、辞書の見出し語が得られるのはアラビヤ語においても同様であり、以上の根拠に基づき、我々は形態素解析機能を内蔵したアラビヤ語の辞書システムを開発している。たとえば図 3 のようにクルアーン[5]の一節を形態素解析すると、文中に現れる語から見出し語を判定するので、辞書を引くことが可能となる。

図 3. クルアーンの一節を形態素解析した例

ところで、通常形態素解析は文法規則と辞書に基づき解析を行うが、本ソフトウェアの開発を通じてアラビヤ語の形態素解析においては論理的には辞書を用いることなく、すべての解析の候補を導くことが可能であるという点に気づいた。アラビヤ語はセム系の言語であり、7 割から 8 割という単語が、語根と呼ばれる通常 3 文字の並びから、派生によって生じる。この派生のパターンは有限であり、また語と語の結合・欠落の可能性も有限である。よって、派生のパターンをプログラムが内部的に保持し、品詞ごとの結合・欠落のモデルを作り処理することで、論理的にその文字の並びになりうる全ての語の組み合わせを推測することが可能となる。

本形態素解析エンジンでも、上記の方法ではじめに最大限の候補を導いた後に、辞書データを用いて、存在しない語をはじき絞込みを行うという処理の流れを採用している。

4. 関連研究

4.1. アラビヤ語辞書

Hans Wehr による Arabic English Dictionary of Modern Written Arabic[6]は、前述の語根と呼ばれる文字の並びから単語を探す方式のため、初学者には不向きだが、豊富な収録語により中上級者や研究者のスタンダードとなっているアラビヤ語・英語辞書である。

また初学者向けの亜日辞書として本田孝一によるパスポート初級アラビア語辞典[7] が挙げられる。語彙は Hans Wehr の辞書に比べると少ないが、意味が日本語で載っている点や、語根ではなく文字の並び順で語を探せる点が特徴である。

アラビヤ語・アラビヤ語辞典の古典として、Ibn Manzur による Lisan Al-Arab[8]がある。「アラブの言葉」というタイトルを冠するこの辞書は、現存する最大のアラビヤ語辞書であり、アラビヤ語の拠り所となっている。

4.2. アラビヤ語辞書ソフトウェア

コンピューター上で動作するアラビヤ語日本語辞書として、汎用辞書ソフトウェア PDIC[9]のための、アラビア語－日本語電子辞書データ[10]がある。このデータを用いることで、27000 語の語彙に日本語でアクセスすることができる。また、online アラビア語辞書[11]というサイトからは同データを web 経由で引くことが可能である。

4.3. アラビヤ語形態素解析エンジン

言語処理の国際企業、Basis Technology がアラビヤ語の形態素解析エンジン、ARLA[12]を開発・商品化している他、XEROX の欧州研究所においてもアラビヤ語形態素解析エンジンの研究開発が行われており、デモバージョン[13]を web 上で体験することができる。

4.4. 形態素解析機能付きの辞書

日本語の学習者向けの、形態素解析機能を内蔵した辞書システムとしてリーディング チュウ太[14]が公開されている。これは茶筅を用いて日

本語の文章を形態素解析し、英語・ドイツ語の辞書と関連付けて意味を表示するもので、分かち書きが必要な日本語の辞書引きを強く支援するツールと言える。

4.5. 関連研究の評価

言葉を探すという観点からは、紙媒体の辞書よりも、電子媒体の辞書のほうが優れている。しかし、4.2 にて紹介したソフトウェアは、入力された文字列をほぼそのまま前方一致で検索するため、見出し語の判定のできない初学者にとっての不便は相変わらず存在する。

また、アラビヤ語の形態素解析エンジンは、XEROX のものが web 上で利用できるので、それにより解析した結果を使って辞書を引けば同じことであるが、解析の精度にやや難点があり、また JavaApplet という実装形態なので実行速度に問題があり、長文を読んでいこうとすると、多くの時間がかかる。

よって、これらが融合した形でソフトウェアを仕上げる必要性が生じる。リーティング チュウ太の成功は、形態素解析と辞書 DB の結合が学習のツールとして有効であることを示唆しており、特にアラビヤ語においては学習初期の負担を軽減させるツールとして期待が持てる。

5. システムアーキテクチャとクラス構成

5.1. システム構成と利用形態

本辞書システムは、最終的にブラウザ経由で使うウェブアプリケーションとしての実装を行っている。また、たとえば現地研修に行った時などネットワークのない場所でも使える、ローカルバージョンの作成も合わせて検討している。現行の形態素解析エンジンは J2SE1.4 を用いて実装されているので、他環境への移植が比較的容易である。データベースのエンジンとしては Postgres8 を用いているので、ローカル版の作成の際には DB 部分の機能をどのように実現するかが課題となる。

図 4 に現状のウェブアプリケーション版のシステムアーキテクチャを示す。Fedora Core で運営されているサーバーの上に、Java と Postgres が走り、その上の層で tomcat と本エンジンが稼動し、最終的な辞書アプリケーションとしての機能を提供している。

図 4. システムアーキテクチャ

5.2. 現状のクラス構成

現状で本プログラムは 7 つのパッケージ、1312 のクラスから構成されており、1200 程度のクラスは自動生成されたものである。パッケージ名と、それに含まれるクラスの役割を図 5 にまとめる。

図 5. パッケージ名一覧と提供する機能

このうち arabic パッケージに属するクラスは、プログラム全体を通じて使っているほか、ソースコードの自動生成の際にも使われているので、まずここで説明する。

5.3. arabic.ArChar クラス

Java 言語においては、ソースコードの中に直接アラビヤ語を書き込むことが可能であるが、それはソースコードの可読性や編集可能性の観点からすると望ましくない。よって我々は arabic というアラビヤ語の処理を円滑に行うためのパッケージを作成した。それに属する代表的なクラスは ArChar である。

本クラスはアラビヤ語の文字を表し、スタティック変数として Unicode に基づくアラビヤ語の各文字を図 6 のように保持している。

図 6. アラビヤ語の文字コード定義するの例

また逆にアラビヤ語の文字コードから、このクラスで定義されている文字名を String で返すスタティックメソッドが宣言されており、これが後述する入力したデータから Java のソースコードへの自動変換を可能にしている。その他、入力された文字コードが文字を示すのか発音記号を示すかを判定するメソッドなどが定義されている。

5.4. arabic パッケージのほかのクラス

arabic パッケージには、アラビヤ語の文字列を表現する ArString クラスや、アラビヤ語特有の正規表現の処理を、java.util.regex.Pattern と java.util.regex.Matcher クラスを委譲される形で保持し、円滑に進める ArRegex クラスが定義されている。

5.5. 残りのパッケージ

本ソフトウェアの開発に当たっては、実際のプログラミングに加え、それと同程度の作業量としてデータ入力の作業が発生した。またもちろんアラビヤ語の文法モデルを考案することも必要であった。以下より、アラビヤ語の文法モ

デル、データ入力、実際のプログラミング作業について順番に、残りのパッケージと関連付けながら説明する。

6. アラビヤ語の文法モデル

6.1. 三つの品詞と用語の整理

本エンジンでは、動詞・名詞・文字の３つをアラビヤ語の品詞とみなし解析を行っている。形態素の中で格を持つものを名詞とし、格を持たないものの中で活用するものを動詞、しないものを文字という規則に基づき分類している。この分類によると、形容詞は名詞であり、前置詞や接続詞は文字という扱いになる。一見大雑把過ぎる分類に見えるかもしれないが、文法上の特性は動詞・名詞・文字として括られたその他の品詞の間ですべて共通しており、実際にネイティブのアラビヤ語研究家の書いた教科書[15]においても、まずは上記の３つの品詞に分類されていた。

アラビヤ語は、英語のように語と語の間をスペースで区切って記述するので、形態素解析を行うときは、まずスペースにより文章を分割することが第一となる。しかし、スペースで区切られた語の塊(本プログラムでは token と呼称する)は、複数の形態素から成っており、それらを切り分ける必要がある。

token は 1 つの動詞・名詞・文字のいずれか(本プログラムでは stem と呼称する)と、複数の文字の結合によって構成される可能性があり、stem がどの品詞かによってその結合のモデルは異なる。本プログラムでは、品詞ごとの結合可能性をモデル化し、すべての可能な分割の案を導き出している。

図 7 に token と stem の関係を BNF(Backus Naur Form)を用いて記述するとともに、以下にさらに動詞・名詞・文字が stem だった場合の token について、詳細なモデルを述べる。

図 7.BNF による token と stem の関係

6.2. 動詞のモデル

図 8 に動詞を stem とする token のモデルについて示す。

図 8. BNF による動詞 token のモデル

6.3. 名詞のモデル

図 9 に名詞を stem とする token のモデルについて示す。

図 9. BNF による名詞 token のモデル

6.4. 文字のモデル

図 10 に文字を stem とする token のモデルについて示す。

図 10. BNF による文字 token のモデル

7. 解析に必要なデータ

7.1. 動詞に関するデータ

7.1.1. 存在する語根と活用形

7.1.2. ファアラを用いて記述したパターン

7.2. 不規則複数名詞のパターン

7.3. アラビヤ語の辞書データ

8. 解析の流れと実装

8.1. 解析の流れ

8.2. 単語の切り分け部の実装

8.3. 不規則複数名詞の単数予測の実装

8.4. 活用された動詞の解析部の実装

9. ネイティブによるテスト

9.1. テストの内容と方法

9.2. 結果と考察

10. 評価と今後の課題

本プログラムは 2006 年 2 月末を目指して作業が行われており、2005 年の 12 月初旬現在では、未完成の部分を残している。具体的には、存在する語根と活用形のデータを基に結果の絞込みを行う部分、解析の候補が得られた段階で辞書データベースに問い合わせを行い意味とともに表示する部分と、辞書のデータベースそのものである。まずこれらを完成させる必要がある。

しかし、現在までに実装が行われた部分に関しては、特に単語の分割モジュールに関しては現地のテストで 9 割を超える正答率をあげ、あくまでモデルと語形に基づくアラビヤ語解析の有効性を示唆するものであった。

また、ソフトウェアは実際に人の手により使われてみることで、さらなる洗練を遂げるはずである。実際に学習者に使ってもらってのフィードバックの反映や、ユーザーインターフェースの設計なども今後の課題と言えよう。

語単位ではなく文を単位とする本格的な形態素解析エンジンへの改良や、オリジナルの高品

質な亜日辞書データの整備、さらには多言語版の作成、特に日本語を学ぶアラブ人のための形態素解析機能を有する日亜辞書システムなどにもモチベーションはあるが、何よりもまずは、遠くを目指しつつも常に確実な一歩を歩んでいければと思う。

謝辞

本プログラムの開発に当たっては、多くの方々にお世話になった。以下に記して御礼申し上げたい。まずなにより、大学に入るまでまさか興味などもっていなかったアラビヤ語の面白さと重要性を示し、学び始めるきっかけを作ってくださった慶應義塾大学総合政策学部教授兼大学院政策・メディア研究科委員の奥田敦先生と、慶應義塾大学総合政策学部マーヘル・カブラー訪問講師に感謝を申し上げる。

また、以前はただのアイディアでしかなかったこのソフトウェアが実際に開発される場は、2005 年度の未踏ソフトウェア創造事業未踏ユースへの採択により与えられたと言っても過言ではない。常に有形無形のアドバイスを下さる早稲田大学理工学部コンピュータ・ネットワーク工学科教授の筧捷彦 PM と、東京大学大学院情報理工学系研究科創造情報学専攻教授の竹内郁雄 PM、そしてプロジェクトの力強いバックアップを行ってくださる国際メディア研究財団の大野一生氏、情報処理推進機構未踏ソフトウェア創造事業事務局の後藤文博氏にも同じく感謝を申し上げる。

アラビヤ語を内容とする以上、アラビヤ語を母語とする方々の協力も欠かせなかった。普段の研修にとどまらず、本プログラムのテスト活動を受け入れてくださったシリア・アレッポ大学学術交流日本センターの皆様に感謝を申し上げる。特に副所長のアフマド・マンスール博士と、事務長のアブドラッザーク・バナナ氏に。また、実際にテストを行ってくださったアレッポ大学文学部大学院のファーギヤ先生、イーマーン先生、ラーウィヤ先生、リハーブ先生の 4 人の先生方に。自分の時間を無償で使い、いつもそばにいてくれる大親友のアレッポ大学医学部のムサンナ・アルアーボ君とアレッポ大学法学部のムハンマド・ハージ・ムハンマド君にも。

ここに名前を挙げることができたのは、本当に一部の人に過ぎない。日常を共に過ごす、友人や家族にも、もちろん最大限のありがとうの気持ちを表しつつ。

リファレンス

1. 日本放送協会ほか編, NHK テレビアラビア語入門, 日本放送出版協会, 2003
2. ChaSen's Wiki - FrontPage, http://chasen.naist.jp/hiki/ChaSen/
3. MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://chasen.org/~taku/software/mecab/
4. 形態素解析 - Wikipedia, http://ja.wikipedia.org/wiki/%E5%BD%A2%E6%85%8B%E7%B4%A0%E8%A7%A3%E6%9E%90
5. 徳増公明ほか編, 亜日対訳・注解 聖クルアーン, 日本ムスリム協会, 1982
6. Hans Wehr, Arabic English Dictionary of Modern Written Arabic, Spoken Language Services, 1993
7. 本田孝一, パスポート初級アラビア語辞典, 白水社, 1997
8. Muhammad Ibn Mukarram Ibn Manzur, Lisan Al-Arab, Dar Sadir, Bayrut, 1956
9. PDIC Home Page, http://homepage3.nifty.com/TaN/
10. アラビア語－日本語電子辞書データ, http://homepage1.nifty.com/A-JDIC/
11. online アラビア語辞書, http://www.arab.jp/
12. アラビア語形態素解析システム, http://www.basistech.co.jp/base%2Dlinguistics/arabic/
13. Arabic Morphological Analysis and Generation - Xerox XRCE, http://www.xrce.xerox.com/competencies/content-analysis/arabic/input/keyboard_input.html
14. Reading Tutor Homepage, http://language.tiu.ac.jp/
15. ラーウィヤ先生たちの教科書(todo)