多言語プラットフォーム UNA

小島 裕一　井田 裕子　本間 咲子　望主 雅子
伊藤 篤　佐藤 奈穂子*　亀田 雅之
（株）リコーソフトウェア研究開発本部　*（株）リコー研究開発本部

１．はじめに

近年、我々の日常では電子メールや企業内文書、インターネットコンテンツなどの電子化文書が増大し、これら大量の文書の中から有用な情報を抽出することが困難になってきている。

検索技術は、このような状況を解決するための技術の 1 つである。データベース等に登録されている文書を検索する場合は、検索するためのキーワードとして入力された検索語と一致する単語が文書に含まれているか否かによって判断する。

検索対象となる文書は、データベース等に登録する際に文書中の文章を単語に分割し、異表記の単語をすべて正規化表記に変換して、文書と関連付けを行い、検索対象のインデックス情報として登録する。一方、検索語が入力された際には、同様に、検索文字列を単語分割し、正規化表記に変換する。変換された表記を用いて、データベース等に登録されている文書に関連付けされたインデックス情報を検索することにより、文書を検索することができる[1]。

このように検索技術では、自然文から検索対象のインデックス情報や検索語を取り出すための言語解析処理が必要となる。

言語処理エンジン UNA（以下 UNA）は、単語分割と異表記正規化、ステミング機能を有する言語処理エンジンで、リコーの文書管理システム等の検索機能に利用されている。

本稿では、UNA の概要について紹介する。

２．開発経緯

これまでに我々は、軽量・高速な日本語解析ツール『簡易日本語解析系 QJP』[2][3]や、これを応用しキーワード抽出や重要文抽出機能を実現した『日本語文書読解支援系 QJR』[4][5]を開発してきた。しかし QJP は、検索他、様々なアプリケーションで利用する上で次の問題があった。

(1) 機能面での問題
　複合語分割や異表記正規化、ステミング機能を有さず、解析対象は日本語文書のみであった。
(2) 解析精度の問題
　日本語文書の単語分割に解析用辞書を使用しないため、解析用辞書を用意することにより解析精度の向上が見込まれた。

そこで、これらの問題を解決するために UNA の開発を行った。

本稿では、(1)の問題から UNA に実装されている機能を中心に説明する。

３．概要

3.1.　特長

以下は、UNA の主な特長である。

(1) 言語の指定により、複数言語に対して動作
(2) 単語分割機能
(3) 欧州 6 言語（英、仏、独、日、伊、蘭）のステミング機能
(4) カタカナや漢字の正規化機能
(5) マルチプラットフォーム上で動作
(6) ユニコード（Unicode）対応

各詳細については後述する。

3.2.　仕様

・対応 OS
Windows 95/98/NT4.0/2000/XP
（日本語、欧州 6 言語、中国語（簡体字/繁体字）の各国語版に対応）
Sun Solaris 8（SPARC 版）
Red Hat Linux 9, AS
・文字コード
Unicode（UTF-8）
・実行時必要メモリ
日本語：7.5Mbyte
欧州 6 言語：2.0Mbyte
中国語：2.2Mbyte
・解析速度（Pentium M 900MHz）
日本語：100MB/2min
欧州 6 言語：100MB/6min
中国語：100MB/3min

3.3.　処理の流れ

UNA は、図 1 に示す通り、外部アプリケーションからの言語指定（4.1 に後述）に応じて、解析文書を単語分割し、分割された単語に対して異表記正規化もしくはステミングを行い、単語の見出し（出現形）と異表記正規化もしくはステミング表記、品詞情報をアプリケーションに返す。

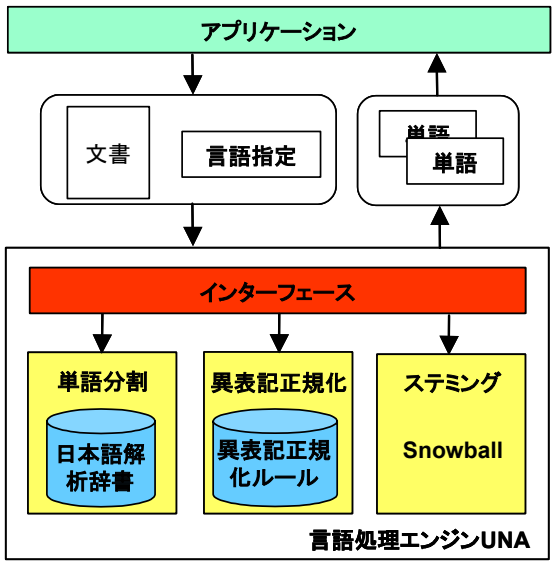

図１　UNA における処理の流れ

４．機能の詳細

本章では、3.1 に記述した特長(1)～ (4)の詳細について説明する。特長(5)(6)は 3.2 に前述の通りである。

4.1.　言語指定

UNA は、外部アプリケーションから指定された言語に応じて、入力文書を単語分割し、分割された単語に対して異表記正規化、ステミングを行う。以下は、UNA で指定可能な言語である。

日本語
中国語(簡体字)
中国語(繁体字)
英語
ドイツ語
フランス語
イタリア語
スペイン語
オランダ語

言語指定が日本語および中国語の場合、アルファベット文字列は英語と見なして処理する。

4.2.　単語分割

単語分割では、言語指定に応じて入力された文字列を単語単位に分割する。

検索システムにおける単語単位検索では、検索時に一致を調べる単位は単語であるため、検索に適切なように単語を分割することが望まれる。

各言語指定時の単語分割機能について説明する。

■日本語指定時

言語指定が日本語の場合について説明する。

単語分割の動作

・漢字/ひらがな/カタカナ

解析用辞書を参照して日本語の単語単位に分割する。解析用辞書に該当する単語が見つからない場合は、文字種に応じた未登録語処理を行い、１単語と見なし分割する。

・その他

アルファベットと数字（後述）を除いたその他の文字は記号とみなし、１文字１単語として分割する。

・改行にまたがる単語の検出

日本語文書では行末で単語が分割される「行分れ」という現象が生じる。「ソ(改行)フトウェア」のように改行にまたがる単語を「ソフトウェア」として検出することにより、改行による解析誤りの発生を防ぐ。この動作は ON/OFF の設定が可能であり、デフォルトの設定では改行をまたがる単語は検出しない。

機能

・出力情報

言語指定が日本語の場合、以下の情報を出力する。

- 見出し（出現形）
- 品詞情報（解析用辞書により単語毎に設定された品詞）
- 複合語の語構成情報

・短単位長単位の解析の切替（複合語の分割）

「ネットワーク化」などの複合語を分割するか否か、機能の ON/OFF の設定が可能である。設定が ON の場合は、上記の通り、複合語の語構成情報を出力する。検索時には、「ネットワーク」によって「ネットワーク化」を含む文書もヒットさせるためである。

解析用辞書

形態素解析辞書：収録語数約 10 万語
品詞体系：品詞大分類 22 品詞

精度

形態素解析精度 約 98.5%

解析アルゴリズム

単語出現頻度に基づく単語コストとヒューリスティックによる品詞間接続コスト設定 [6]とコスト最小法による解探索手法[7]をベースに改良した方法を用いた。

■中国語指定時

言語指定が中国語の場合について説明する。

単語分割の動作

・漢字

漢字１文字を１単語とみなして分割する。言語解析精度の向上を考えた場合は中国語解析のための辞書が必要であるが、検索精度の観点からは便宜的に１文字を１単語と見なした検索を行なっても現状の単語単位の検索と遜色ない検索精度を得ることができる。辞書開発に必要なコストを考慮し、現状では上記仕様とした。

・ひらがな/カタカナ

記号とみなし、１文字１単語として分割する。

・その他

アルファベットと数字（後述）を除いたその他の文字は記号とみなし、１文字１単語として分割する。

機能
・出力情報
　言語指定が中国語の場合、以下の情報を出力する。
　- 見出し（出現形）
　- 品詞情報
　　アルファベットや数字は文字種毎に定義した品詞情報、その他の場合は未登録語品詞情報を出力する。
■欧州 6 言語指定時
　言語指定が欧州 6 言語の場合について説明する。
単語分割の動作
　基本的には空白をデリミタとみなし分割する。その他は以下の仕様に従う。
・漢字/ひらがな/カタカナ
　記号とみなし、1 文字 1 単語として分割する。
・その他
　アルファベットと数字（後述）を除いたその他の文字は記号とみなし、1 文字 1 単語で分割する。
機能
・出力情報
　言語指定が欧州 6 言語の場合、以下の情報を出力する。
　- 見出し（出現形）
　- 品詞情報
　　アルファベットや数字は文字種毎に定義した品詞情報、その他の場合は未登録語品詞情報を出力する。
■言語指定共通
　以降の文字は、言語指定によらず同一の処理を行う。
単語分割
・アルファベット/数字
　空白あるいは文字種の変わり目ごとに分割する。
・ハイフンで連結された文字列
　欧州 6 言語においては「abc-(改行)def」のような表記が存在するが、これとは別に、「abc-def」のような形で複合語が存在し得るため、改行をまたがった複合語との判断の問題が生じる。UNA では、改行をはさんだハイフン連結語を以下のように扱う。
　- 基本的には、ハイフンで連結された文字列を 1 単語とみなし分割する。
　- 複合語分割が指定されている場合はハイフン前後で分割する。
　なお、この動作は ON/OFF の設定が可能である。

4.3. 異表記正規化

　日本語や中国語では、同じ意味を有する単語を表記する場合でも複数の表記が存在し、特に外来語においては、複数の表記方法が可能な場合が多い。
　例) バイオリン、ヴァイオリン
　検索システムでは、このような表記ゆれに対応し、異表記の正規化と展開処理を行う[1]。文書検索の際には、検索語を正規化処理により統一した表記に変換する。さらに正規化処理のみでは、別々に扱われるべき語まで統一されてしまったり、統一されるべき語が統一されなかったりといった現象が生じる可能性があるため、展開処理を行い、正規化による統一の不足に起因する検索漏れを回避しつつ、過剰な統一によるノイズの発生を抑える。一方、文書登録の際には、インデックス情報の生成において正規化処理のみ行う。
　異表記の正規化や展開では、辞書ではなく、文字列単位の変換ルールに基づき変換する。また、異表記正規化機能は ON/OFF の指定を可能である。これは、入力文字列をそのままの表記で検索したいケースに対応するためである。
■日本語指定時
　言語指定が日本語の場合について説明する。
・ひらがな/カタカナ
　ひらがなとカタカナに対しては、以下の正規化を行う。
　- 半角から全角への変換
　　例) 「ｶﾀｶﾅ」→「カタカナ」
　- 旧字から新字への変換
　　例)「ゐ」→「い」
　- 合字
　　か行、さ行などのひらがな、カタカナ文字と濁点や半濁点が分離した表記を 1 文字の表記に置き換える。
　　例)「か」「゛」→「が」
・漢字
　日本語の漢字には、異体字という同じ意味を有する複数の表記が存在する。これらの表記を統一するために漢字 1 文字単位に正規化を行う。
　例) 「齋」→「斉」
・カタカナ表記正規化
　日本語においては、特にカタカナ表記の揺れが大きく検索漏れの原因となる[8]。カタカナ単語の表記揺れを吸収するため、表記の正規化を行う。
　例) 「コンピューター」→「コンピュタ」
・カタカナ表記の展開
　例えば、「シリコーン」「シリコン」はそれぞれ「シリコウン」→「シリコン」と正規化されるが、検索において両方をヒットさせるため、「シリコーン」を展開すると「シリコウン」「シリコン」の 2 つの展開結果を返す。この機能は検索時に入力された単語のための機能である。
■中国語指定時
　言語指定が中国語の場合について説明する。
・漢字
　漢字 1 文字単位に繁体字から簡体字への変換を行う。中国語には中国本土およびシンガポールで使用される「簡体字」と台湾および香港で使用される「繁体字」という二つの文字体系がある。繁体字では同じ意味に対して複数の文字が存在し[9]、検索漏れが生じる原因となるため、同じ意味を持つ複数の文字を 1 つの文字に統一する文字として簡体字に変換する。
■言語指定共通
　以降の文字は言語指定によらず同一の処理を行う。

・ラテン/キリル/アラビア文字
- 大文字から小文字への変換
例)「A」→「a」
- 全角から半角への変換
例)「ａ（全角）」→「a（半角）」
- 音標符号との合字
音標符号と合字を行う。
例) 0061;LATIN SMALL LETTER A
0300;COMBINING GRAVE ACCENT
↓
00E0;LATIN SMALL LETTER A WITH GRAVE
- 音標付き文字から音標なし文字への変換
欧州 6 言語を除く言語が指定された場合、音標付き文字を音標なし文字に変換する。
例) 00E0;LATIN SMALL LETTER A WITH GRAVE
↓
0061;LATIN SMALL LETTER A

・数字/記号
全角/半角による表記ゆれを吸収するために、数字や記号を全角から半角に変換する。
例)「１（全角）」→「1（半角）」

・空白文字
Unicode の 3000 や 0020 などの文字を対象に全角から半角に変換する。

4.4. ステミング

欧州言語では動詞の活用や複数形など文法的な表記ゆれが生じる。ステミングは、このような表記表記を正規化する機能である。

例）teacher, teachers, teaching,taught→teach

従来、UNA では自社開発のステマー[10]を使用していたが英語ステマーのみであったため、UNA を多言語対応化するにあたっては英語以外の欧州言語のステマーを開発する必要があった。しかし、ステマー開発には各言語に関する知識が必要とされ、作業コストを要することから BSD ライセンスで提供される Snowball ステマー[11]を採用した。

なお、ステミング機能は、異表記正規化機能が ON の場合、ON/OFF の指定が可能である。これは、入力文字列をそのままの表記で検索したいケースに対応するためである。なお、ステミングは大文字小文字などの正規化がされていることが前提であるため異表記正規化機能が OFF になった場合はステミング機能も OFF になる。

5．現状の問題点

我々は、欧州 6 言語、中国語の検索性能を評価するために CLEF2003、NTCIR-4 の検索コンテストに参加した。この結果、ドイツ語の解析処理に問題があることが分った[12][13]。現状の UNA は、複合語分割を行っていないが、検索精度向上のためには複合語分割を行うことが必要であると考えている。

また、中国語文書の検索では文字単位の正規化でも十分な性能を得ることができたが、中国語文字列を単語単位に分割して正規化することでさらに検索精度が向上する可能性があると考えている。

6．今後の展開

我々は UNA を多言語解析のためのプラットフォームと位置付けている。現在、UNA は多言語の文書を処理可能であるが、日本語以外には非常に限られた機能しか提供していない。今後は、UNA で解析対象とする言語に対して日本語と同レベルの機能を提供するために、以下の機能の開発をステマーと同様に外部技術の導入も含めて検討している。

・中国語の単語分割
・日本語以外の品詞付与
・欧州 6 言語の複合語分割

参考文献

[1] Ikeda, Mano, Itoh, Takegawa, Hiraoka, Horibe, Ogawa, “TRMeister: a DBMS with high-performance full-text search functions”, ICDE2005, 2005.4 発表予定

[2] 亀田, “軽量・高速な日本語解析ツール『簡易日本語解析系ＱＪＰ』“, 言語処理学会 第１回年次大会 発表論文集, pp.349-352, 1995

[3] Kameda, “A Portable & Quick Japanese Parser：QJP”, Coling '96, 1996.8

[4] 亀田, ”擬似キーワード相関法による重要キーワードと重要文の抽出”, 言語処理学会 第２回年次大会 発表論文集, pp.97-100, 1996

[5] 亀田, ”段落間及び文間関連度を利用した段落シフト法に基づく重要文抽出”, 情報処理学会 自然言語処理研究会 121-179, 情報学基礎 47-9 (共催), 1997.9

[6] 佐藤,小松, ”コスト最小法を用いた形態素解析におけるコスト設定の一方法”, 情報処理学会第 47 回全国大会講演論文集, 3-153, 1993

[7] 小松, ”コスト最小法に基づく逐次確定型・形態素解析”, 情報処理学会第 47 回全国大会講演論文集, 3-151, 1993

[8] 増山, 関根, “大規模コーパスからのカタカナ語の表記の揺れリストの自動構築”, 言語処理学会第 10 回年次大会, A1-5, 2004

[9] Halpern, J., Kerman J., “The Pitfalls and Complexities of Chinese to Chinese Conversion”, Proc. of the Fourteenth International Unicode Conference in Cambridge, MA, 1999

[10] 本間, ”全文検索における英語接辞処理の評価”, 情報処理学会研究報告 2000-NL-138, No.138-008, 2000

[11] Snowball, http://snowball.tartarus.org/ 2005/2 現在

[12] Kojima, Itoh, “Ricoh in the NTCIR-4 CLIR Tasks”, NTCIR-4, 2004.6

[13] Kojima, Itoh, “Ricoh at CLEF 2003”, CLEF2003, 2003.8