1

自然言語処理関連ツールあれこれ
－使えるフリーソフト－

奥村　学
東京工業大学 精密工学研究所
oku@pi.titech.ac.jp

自然言語処理関連ツールの全体像

自然言語処理研究もある程度の歴史を重ね，近年，商品としてのソフトウェア以外に，質の高い自然言語処理ツールがフリーソフトとして公開され，一般に利用できるようになってきている．本稿では，特集の最初の記事として，一般に利用可能になっている，日本語を対象とした自然言語処理関連ツールを概観し，可能な限り網羅的に紹介したい．

本特集では，本稿の後いくつかの記事で個々のツールが取り上げられ紹介されているが，本稿では，取り上げられていないツール群も合わせて紹介し，個々の記事で紹介されるツールとの関係も示したい．また，ツールを含め，自然言語処理関連の資源の共有化を促進する目的で設立された言語資源共有機構（GSK）の取り組みなどについても言及したい．なお，英語を対象とした自然言語処理関連ツールに関しては，第4編の関根氏の記事を参照していただきたい．

さて，一般に利用可能になっている自然言語処理関連ツールにはどのような種類があるのだろうか．

自然言語処理といえば「ことば」を処理することであるから，自然言語処理ツールといえば「ことば」で書かれたテキストを処理するツールがまず思いつく．自然言語処理の要素技術である形態素解析，構文解析を，テキストに対して高精度で頑健に行ってくれるツールは，自然言語処理研究者にとって大変有用であるし，全文検索システムで形態素解析ツールがよく利用されるのが一例のように，他のツールの中に組み込まれて利用されるようになってきている．一般のユーザが直接利用するツールとしては，テキストを書く際に利用するエディタや，ネットニュースを読むのに利用するニュースリーダなどがある．さらに，近年の情報洪水の中で，ユーザがテキストデータをうまく利用できることを支援するツールとして，全文検索システム，要約システムなどが挙げられる．一方，「ことば」で話された対話を処理するための，自然言語処理と関連するツールとしては，音声認識，音声合成ツールなども考えられる．

「事例紹介」の章では，これらの自然言語処理関連ツールを，種類等の観点から分類し，個別に紹介するが，それに先立ち，次章でまず，形態素解析，構文解析ツール，全文検索，要約システムなどの自然言語処理ツールは何をするツールなのか，その原理はどうなっているのかを簡単に説明する．

ツールのベースとなる自然言語処理技術

本章では，自然言語処理ツールを実現する上で基礎となる，いくつかの自然言語処理技術を簡単に説明する．なお，自然言語処理や情報検索技術などに関しては，最近良い入門書が出版されているので，詳細は，たとえば文献1）～3）などを参照していただきたい．

自然言語処理は，言語表現を計算機で解析する自然言語解析と，言語表現を計算機が作り出す自然言語生成に大きく分けられる．そして，自然言語解析の処理は，形態素解析，構文解析，意味解析，文脈解析の4つのステップに大きく分けられるとされる．

自然言語解析の最初のステップである形態素解析では，1）日本語のように，単語間に区切りのない言語では，テキストを単語（厳密には，言語学的に，意味を持つ最小単位と定義されている形態素）の並びに分割し（segmentation），2）単語が語形変化（たとえば，動詞の活用変化）している場合には，原形へ戻し（stemming），3）単語の品詞を決定する（part-of-speech

tagging).

形態素解析では，単語辞書と，単語（品詞）間の連接可能性を規定する連接規則を用いて，入力された文字列から，可能な単語の並びを生成し，連接のしやすさを表すコストなどの数値や優先規則を用いて，単語の並びの優先順位づけを行い，（多くの場合）形態素解析結果として，最も良い単語の並びを出力する．

構文解析では，形態素解析で得られた単語（品詞）の並びに対し，単語間の構文的関係を決定し，文の構造を得る．文の構造としては，日本語の場合，文節間の係り受け関係を有向辺で表す係り受け構造として得ることが多い．構文解析は，文法を用いて可能な文の構造を生成し，文の構造に関する優先規則やコーパスから計算された確率等を用いて，文の構造を順位づけ，（場合により）最も妥当な構造のみを出力する．

次に，本特集ではNamazu（第5編）が例として取り上げられている全文検索システムであるが，情報検索システム，サーチエンジンなどとも呼ばれる．インターネット上のサービスとしても多くのサイトで提供されており，利用されている方も多いだろう．

このシステムは，大量のテキスト集合から，ユーザの検索要求に対して適合する（と思われる）テキストを検索結果として返す．そのため，テキスト集合をあらかじめ索引づけし，検索要求が入力されると，索引を元に，検索要求と各テキストの照合を行い，適合するテキストを適合度の順に出力する．この索引づけの過程で，索引語集合をテキストから生成する際，形態素解析ツールを用いた，単語への分割処理を行うシステムもある．一方，照合を文字列ベースで行い，検索要求中の文字列を含むテキストを検索するシステムもある．Namazuは前者であり，単語を索引として用いるタイプのシステムということができる．

要約システムは，テキストを入力すると，そのテキストの重要と思われる部分を残した，より短いテキストを出力する．市販のワープロソフトの中にも，要約機能を付けたものがいくつか見受けられるようになってきた．現在の自動要約手法の多くは，テキスト中の文（あるいは，形式段落）を1つの単位とし，それらに何らかの情報を元に重要度を付与し，その重要度で順序づけ，重要な文（形式段落）を選択し，それらを寄せ集めることで，要約を作成する．すなわち，要約は重要文抽出により行われてきている．

事例紹介

本章では，形態素解析ツール，構文解析ツール，全文検索システム，その他に分類し，自然言語処理関連ツールを個別に紹介する．

形態素解析ツール☆1

(a) 日本語形態素解析システムChasen「茶筌」

最新版: Ver.2.02（UNIX版，Windows版がある）

URL: http://cl.aist-nara.ac.jp/lab/nlt/chasen/

詳細は，本特集第2編参照．

(b) 日本語形態素解析システムJUMAN

最新版: Ver.3.61（UNIX版，Windows版がある）

URL: http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html

(c) 日本語形態素解析パッケージALTJAWS

公開されている版: ALTJAWS2.0

URL: http://www.kecl.ntt.co.jp/icl/mtg/resources/altjaws.html

動作環境: OSは，SunOS 4.1.x，Solaris 2.x（SPARC），Linux 2.0.x（ホームページに詳細な記述あり）．

特記事項:

1) 各単語に対し，品詞だけでなく，NTT日本語語彙大系中の意味分類も付与する．

2) 文節区切りを決定して出力する．

3) 大学・研究機関での研究目的に限る．利用に関する詳細は問合せ先まで．

(d) 日本語形態素解析システムBreakfast

公開されている版: Breakfast4.0.4f

URL: http://www.fujitsu.co.jp/hypertext/free/breakfast/

動作環境: Windows95, WindowsNT 3.51, WindowsNT 4.0（ホームページに詳細な記述あり）．

(e) 日本語形態素解析システム「すもも」

公開されている版: Ver.1.4

☆1 奈良先端科学技術大学院大学の松本裕治教授とNTTサイバースペース研究所の永田昌明氏が，10月に開催されたACL2000のチュートリアル（Morphology for Asian Languages）に合わせて，ツールとして利用可能な日本語形態素解析システムに関するサーベイを行っておられるようである．サーベイの内容は，チュートリアルで報告されるとともに，今後ホームページなどで公開されるとのこと．

URL: http://www.t.onlab.ntt.co.jp/sumomo/index.html
動作環境: 正式にはSun SS20 (SunOS 4.1.4) でのみサポート．この他に多くのマシン，多くのUNIX上での動作が確認されている（ホームページに詳細な記述あり）．

構文解析ツール

(a) 日本語構文解析システムKNP

最新版: Ver.2.0b6
URL: http://pine.kuee.kyoto-u.ac.jp/nl-resource/knp.html
詳細は，本特集第3編参照．

(b) MSLR Parser

公開されている版: Ver.1.03
URL: http://tanaka-www.cs.titech.ac.jp/pub/mslr/index-j.html
動作環境: 動作するOSはUNIXのみ
特記事項:
1) 形態素解析と構文解析を同時に行う．
2) パーザの辞書引きモジュールに，後述する高速文字列検索システムSUFARYが用いられている．

(c) LiLFeS Parsing System

URL: http://www-tsujii.is.s.u-tokyo.ac.jp
動作環境: 本体はLinuxとSolaris, GUIはJAVA2が動く環境
特記事項:
1) 原稿執筆時は，部分公開のみ．
2) 詳細は，文献4) を参照．

全文検索システム

全文検索システムに関しては，http://www.kusastro.kyoto-u.ac.jp/~baba/wais/other-system.htmlに，京都大学の馬場氏が作成された日本語全文検索エンジンソフトウェアのリストがある．このページは，「現在流通している，自サイトで簡単に構築できる全文検索ソフトウェアの数々を，独自に調査した豊富な導入事例とともに，網羅的に解説・紹介して」おり，「ソフトウェアを提供しているWebページへのポインタを，フリーソフトと商用製品とにおおまかに分けた上で，簡単なコメントを加えて紹介して」いる（「」内はホームページからの引用）．非常に有用なページであり，ここにその抜粋となるようなリストを記載するよりも，読者に直接このリストをご覧いただく方が有益と考えた．なお，いうまでもなく，Namazu (第5編参照) も紹介されている．

その他

ここでは，形態素解析ツール，構文解析ツール，全文検索システム以外の，さまざまな自然言語処理関連ツールを紹介する．また，ツールの利用事例も紹介する．

(a) タグ付けツール

形態素解析ツールの出力結果を視覚的に表示し，形態素解析結果のタグが付与されたコーパスを作成するのを支援するツールとして，美茶と美寿満がある．美茶 (ViCha)，美寿満 (ViJUMAN) はそれぞれ，日本語形態素解析システム「茶筌」，JUMANの解析結果を視覚化するシステムである（http://cl.aist-nara.ac.jp/lab/nlt/vi4ma.html)．

(b) 高速文字列検索システムSUFARY

最新版: Ver.2.11
URL: http://cactus.aist-nara.ac.jp/lab/nlt/ss/
suffix arrayというデータ構造を用いて高速な文字列検索を行うためのライブラリをパッケージ化したもので，大規模なデータを対象とした全文検索や辞書検索に利用できる．前述した馬場氏の全文検索システムのリストでも紹介されているが，辞書検索ツールとして他の自然言語処理ツール（たとえば，MSLR Parser) でも利用されているので，あえて紹介した．

(c) 語彙的連鎖生成プログラムLexical Chainers

URL: http://galaga.jaist.ac.jp:8000/pub/tools/index.html
Lexical Chainersは，テキスト中で互いに語彙的結束性を持つ（意味的に関連する）語の連続である語彙的連鎖 (lexical chain) を計算するプログラムである．Lexical Chainersでは，語彙的連鎖の計算方法として以下の3種類を選択できる．

1. 同じ語の繰り返し（同語反復）で連鎖を構成する．
2. シソーラス中の同一の分類に属す語で連鎖を構成する．
3. 共起関係（共起頻度）に基づき意味的に関連すると考えられる語で連鎖を構成する．

語彙的連鎖は，意味的に関連する語の連続であるため，次の特徴を持つ．

a. テキスト中の局所的な文脈情報を提供する．
b. テキスト中に存在する話題の範囲と種類を知る手

がかりになる．

この特徴を利用して語彙的連鎖は，テキストの話題境界を同定するテキストセグメンテーションなど，さまざまな研究に応用されている．

(d) 重要文抽出プログラムPOSUM

URL: http://galaga.jaist.ac.jp:8000/pub/tools/index.html

Posumは，基本的な重要文抽出の手法を用いて，要約を作成する簡易要約器である．文の重要度を計算するための情報として以下の2種類を利用できる．

1. テキスト中の単語の重要度．
2. 単語間の意味的なつながり（語彙的連鎖）．

2.の語彙的連鎖計算には，(c) のLexical Chainersの一部が用いられており，シソーラスに基づく語彙的連鎖と，同じ語の繰り返しに基づく語彙的連鎖の2種類が利用できる．

Posumが要約を作成するアルゴリズムは以下のようになる．

1. テキストを形態素解析する．
2. テキスト中の各語に重み付けを行う．
   語の重要度としては，tf（語のテキスト中での出現頻度）に基づくものと語彙的連鎖の重要度に基づくものの2種類から選択できる．
3. テキスト中の各文の重要度をその文に含まれている語の重要度の合計として計算する．
4. 重要度の高い文を要約率分だけ選択し，元テキストでの出現順に並べて出力する．

(e) 日本語読解支援システム

形態素解析ツールや構文解析ツールの，直接的な利用事例といえるのが，日本語学習者のための日本語読解支援システムである．日本語学習者が日本語の文章を読む場合，文章中の単語の区切りが分からないこと，単語の読み方が分からないため辞書を引くのに時間がかかること，文の構造が分からないことなどが大きな問題となる．そこで，形態素解析ツール，構文解析ツールを利用し，文章を形態素解析，構文解析した結果を得，画面上に，文章中の単語の区切りや文の構造を表示したり，文章中の単語に対して，あらかじめ用意した辞書中のその単語の項目（読みや，意味が記述されている）へのリンクを付与して，即座に辞書引きができるようにするなどの機能を有する読解支援システムがいくつか開発され，実際に学習支援システムとして利用に供されている（たとえば，「DL」(http://www.jaist.ac.jp/~tera/) や，「あすなろ」(http://hinoki.ryu.titech.ac.jp/)）．

(f) 情報処理振興事業協会（IPA）独創的情報技術育成事業に係る開発システム

IPAの独創的情報技術育成事業（http://www.ipa.go.jp/STC/dokusou.html）は，フリーソフトの開発，公開を目指すプロジェクトを支援する事業と考えられるが，近年この事業により，いくつかの自然言語処理関連ツールが開発されている（あるいは，開発中である）．

(f-1) 日本語ディクテーション基本ソフトウェア

URL: http://www.lang.astem.or.jp/dictation-tk/

動作環境: UNIXベースのOS（Solaris, Irix, PC Linux等）

連続した音声で入力された日本語を，統計的音声認識手法により，高速かつ正確に文章に変換するための基本ソフトウェアである．

- 大語彙連続音声認識プログラムJulius
- 音韻モデル
- 言語モデル・単語辞書
- 形態素解析・読み付与ツール（ChaSen/ChaWan）

などが含まれており，すでに公開されている．

(f-2) 知的ニュースリーダHISHO

URL: http://www.tis.co.jp/research/ipa-hisho/

2000年3月まで行われた「話題関連性に基づく情報検索手法を実装したニュースリーダの開発」というテーマのプロジェクトでは，インターネット上の大量のニュース記事の中から利用者が興味を持った記事に関連する話題を抽出し，利用者が必要とする情報を効率的に入手できる高性能の知的ニュースリーダを開発している．従来のニュースリーダの機能に加え，記事中のキーワードを抽出し，その情報を元に，ニュース記事集合中の話題の転換点を検出することで，大量のニュース記事を話題ごとに分類し，「リファレンスツリー」として表示する，各話題の内容をキーワードにより簡潔に表示する，また，ある話題に類似した話題を検索するなどの機能を持ち大量の記事の中からユーザが必要なものを効率的に入手するための支援を実現している．詳細は，文献5）を参照していただきたい．

(f-3) 汎用連想計算エンジンGETA

「汎用連想計算エンジンの開発と大規模文書分析への応用」というテーマで，昨年度開始されたプロジェクトでは，テキスト間あるいは単語間の類似度計算（連想計算）を高速に行う汎用連想計算エンジンを開発中であり，公開予定である．

動的クラスタリング，類似文書を検索する連想検索などは，計算コストが大きいことから，大規模な文書集合へ適用することが難しいとされてきた．GETAは，計算コストの大きい連想計算を，データ圧縮と高速データアクセスにより，高速実行することから，大規模な文書集合を対象とした連想検索を可能にする．また，連想計算に用いる類似度は，容易に差し替えが可能なように設計されており，汎用性が高い．

なお，GETAは現在開発中であり公開はされていないが，GETAとインタフェースを用いて構築された連想検索システムが東大医科研ゲノムセンタでDB検索サービスとして提供されている（BACE（Bio Association CEntral），http://bace.ims.u-tokyo.ac.jp/）．

今後の展望

本稿では，日本語を対象とした自然言語処理関連ツールを紹介した．紹介してきた自然言語処理ツールを含め，各種コーパス，電子化辞書，シソーラスなどの言語資源の共有，流通を目指して，昨年5月言語資源共有機構（GSK）という組織が，東工大の田中教授のご尽力で設立された（詳細は，http://tanaka-www.cs.titech.ac.jp/gsk/を参照していただきたい）．今後この機構がうまく機能し，より多くの有用な自然言語処理ツールが開発されるとともに，開発されたツールがより多くの人に利用されるようになることを期待したい．

また，前章（f）で触れたIPAの独創的情報技術育成事業や，IREX, NTCIR（http://www.rd.nacsis.ac.jp/~ntcadm/workshop/cfp2-en.html）などの評価型ワークショップにおいて，フリーの自然言語処理関連ツールが今後も開発，公開されていくことを期待したい．

本稿を終えるにあたり，1つ残念なことは，自然言語処理関連ツールを可能なかぎり紹介するという目標はある程度達成できたかもしれないが，「それらのツールから，では，どのシステムをどういう基準で選べばよいのか」という指標を与えることが必ずしもできなかったことである．

いうまでもなく，この目標の達成は，非常に困難であり，この原稿を書く前から，著者の力量では無理なことは明らかであった．というのも，ツール，システムの比較は，さまざまな基準から可能であり，また，利用者がどういう目的，環境で利用するかということによっても，選択基準は変わってくるからである．

したがって，ここでは，ツール，システムを比較する（あるいは，選択する）際の，いくつかの基準を挙げるにとどめる．

システムの比較をする上で一番分かりやすい基準は，システムの精度だと思われる．しかし，複数のシステム間で精度を厳密に比較したデータを入手するのは現実には困難である．実際には，各システムについて記述する原論文にあたり，記述されているシステムの精度を比較することにより，ある程度の目安を得るというのが現実的な手法となる．

これ以外にも，ツールの選択基準となる要素はいくつも存在する．たとえば，ツールの動作環境，インストールのしやすさ，入出力データの形式，オプションの種類，ツールのインタフェースの使いやすさ，出力情報の種類，等々．

近年インターネット上では，機械翻訳ソフトなど，アプリケーションソフトウェアを比較，検討した情報が活発に流されるようになってきているようであり，今後本稿で紹介したような自然言語処理関連のフリーツールにおいても，同様の比較，検討が行われ，その情報が利用者に供されるようになることを期待したい．

謝辞　この記事を書くにあたり，いくつかのツールに関する情報を開発者の方に提供していただきました．提供していただいた方々に感謝いたします．また，ツール，システムを比較（選択）する基準に関して有益なご指摘をくださった，九州工業大学の乾助教授にも感謝いたします．

参考文献
1) 長尾　真他: 自然言語処理, 岩波書店 (1996).
2) 田中穂積監修: 自然言語処理ー基礎と応用ー, 電子情報通信学会編, コロナ社 (1999).
3) 徳永健伸: 情報検索と言語処理, 東京大学出版会 (1999).
4) Kanayama, H., Torisawa, K., Mitsuishi, Y. and Tsujii, J.: A Hybrid Japanese Parser with Hand-crafted Grammar and Statistics, in Proceedings of the 18th International Conference on Computational Linguistics, pp.411-417, Saarbruken, Germany (2000).
5) 井佐原均, 小作浩美, 内元清貴, 村田真樹, 藪田裕史, 三上宜宏, 西俣憲幸, 高橋　誠: 話題関連性に基づく情報検索手法を実装したニュースリーダの開発, 情報処理振興事業協会第19回技術発表会論文集 (2000).

（平成12年9月25日受付）