寄稿

自然言語処理と語学教育

奥村 学*

* 東京工業大学 精密工学研究所

Manabu Okumura**

** Precision and Intelligence Laboratory, Tokyo Institute of Technology

1. はじめに

我々人間同様,「ことば」を操ることができる計算機を目指した研究とも言える自然言語処理研究は、半世紀に及ぶ歴史を重ね、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、専門家でなくとも利用できるようになってきている。

一方、近年のインターネットの普及、教育現場へのコンピュータの導入などにより、語学教育においても、音声、言語処理技術など、情報処理技術を導入する動きが盛んになってきており、CALL(Computer Assisted Language Learning)、e-learningなどというキーワードを目にする機会も増えてきている[1]。

そこで、本稿では、一般に利用可能になっており、語学教育で有用と思われる、自然言語処理関連ツール、言語資源を概観し、紹介したい。また、それらを利用した語学学習支援システムの具体的な事例も紹介したい。なお、現在一般に利用可能な自然言語処理関連ツールの詳細に関しては、[2]の特集記事を参照して頂きたい。

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、言語資源にはどのような種類があるのだろうか。自然言語処理と言えば「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言えば「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。

さらに、古くから自然言語処理研究を行う上で不可欠であった用例検索システム、KWIC(Key Word In Context)は、語学教育においても、収集した用例を提示するのに有用と考えられる。また、近年 Web 上でも利用可能になっている、自動翻訳システム、辞書も、いろいろな形で利用が可能と思われる。

3 節では、これらの自然言語処理関連ツール、言語資源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、次節でまず、形態素解析、構文解析ツール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。4 節では、自然言語処理関連ツール等を実際に利用した具体例を紹介する。

2. ツールのベースとなる自然言語処理技術

本節では、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[3,4]などを参照して頂きたい。

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ない、(多くの場合)形態素解析結果として、もっと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)もっとも妥当な構造のみを出力する。

3. 自然言語処理関連ツールあれこれ

本節では、自然言語解析ツールと、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。

3.1 自然言語解析ツール

・形態素解析ツール

－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/

－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

・構文解析ツール

－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

3.2 その他

本節では、形態素解析ツール、構文解析ツール以外の、さまざまな自然言語処理関連ツールを紹介する。

・用例検索システム、KWIC

自然言語処理研究では、実際の言語現象を収集して、分析し、それを自動的、あるいは、人手で規則化することで、それらの言語現象を扱うシステムを構築するという手法が伝統的に採られてきた。そのため、人間が、膨大な量の収集した用例(コーパス)を見通し良く閲覧できるためのツールとして、KWIC が古くから利用されている。KWIC は、注目する単語とその周囲の文脈の情報をソートして提示する。この KWIC は現在では、語学教育の分野でも広く利用されるようになってきているし、また、非常に多くのツールが存在するので、列挙することはしない。(たとえばhttp://prairie.lang.nagoya-u.ac.jp/など)。このようなツールを語学教育では、

- collocation(単語間の共起情報)を提示する、
- 単語間の使い分け情報を提示する、
- 語義間の差異の情報を提示する。

などに利用できるだろう。この用例検索は、単語をキーとして検索するだけなら、UNIX のコマンドである‘grep’を利用することで容易に実現できる。また、高速に実現するため、

－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。

が開発されており、利用可能である。

一方、最近の自然言語処理研究では、単に生のコーパスを分析するのではなく、コーパスにいろいろな情報をタグとして付与したもの(タグ付コーパス)(たとえば、単語に品詞を付与したコーパス、文の構造を付与したコーパス、など)を分析の対象としている。そのため、単に単語で用例が検索できるだけでなく、品詞の

並びを表現した正規表現を入力とするなど、より複雑な検索が可能であるような用例検索システムを作成する例が多い(たとえば、[5]、[6])。今後これらのツールの開発が増え、一般の利用に供されるようになれば、語学教育においても、これまで以上に有用なツールとなるであろう。

・自動翻訳システム

自然言語処理研究の歴史の当初から研究、開発が進められてきた自動翻訳(機械翻訳)システムはおそらくもっとも代表的な自然言語処理関連ツールではないだろうか。自動翻訳ソフトも、インターネットの普及に合わせて急速に価格が安くなり、身近となった。Web 上のサービスも数多く、サーチエンジンなどでも検索結果を自動翻訳してくれる機能がついているものもあり(たとえば AltaVista の Advanced Text Search)また、翻訳サイトも数多い(http://www.excite.co.jp/world/)。したがって、このリストアップも不要であろう。精度的に問題がある可能性はあるが、語学教育への利用の可能性はあるのではないだろうか。

・Web 上で検索できる辞書、辞書引きサイト

オンラインで辞書が引けるようになり、紙の時代に比べ、画期的に便利になったと、何年か前著者は思ったものであったが、最近では Web 上で辞書が引けるようになっている。もはや辞書さえ自前で持つ必要はない時代らしい。このようなサイトも枚挙にいとまがない(「翻訳のためのインターネットリソース」(http://www.kotoba.ne.jp/)や「辞書の辞書」(http://www.monjunet.ne.jp/PT/bin/dict.dll)は有用なリンク集の例)が、これらの利用も語学教育では有用であろう。

・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。

ー大語彙連続音声認識プログラム Julius

ー音韻モデル

ー言語モデル・単語辞書

ー形態素解析・読み付与ツール(ChaSen/ChaWan)

などが含まれている。

4. 事例研究紹介

本節では、上で取り上げた自然言語処理関連ツール等を実際に利用した具体例として、2 つの語学学習支援システムを紹介する。

4.1 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言えるのが、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わからないこと、単語の読み方がわからないため辞書を引くのに時間がかかること、文の構造がわ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等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(例えば「DL」(http://www.jaist.ac.jp/~tera/)や、「あすなろ」(http://hinoki.ryu.titech.ac.jp/)[7]。

次の図 1 は「あすなろ」の表示例である。図 1 では、画面 1 の 1 文を、形態素解析により、単語に分割した結果が画面 2 に表示されている。また、画面 2 の単語の下の「◇」をクリックすることで、画面 3 にその単語の意味が、母語による訳語により表示される(図 1 では英語)。さらに、画面 2 の`Show Tree'の表示をクリックすることで、その文の係り受け構造を見ることもできる。

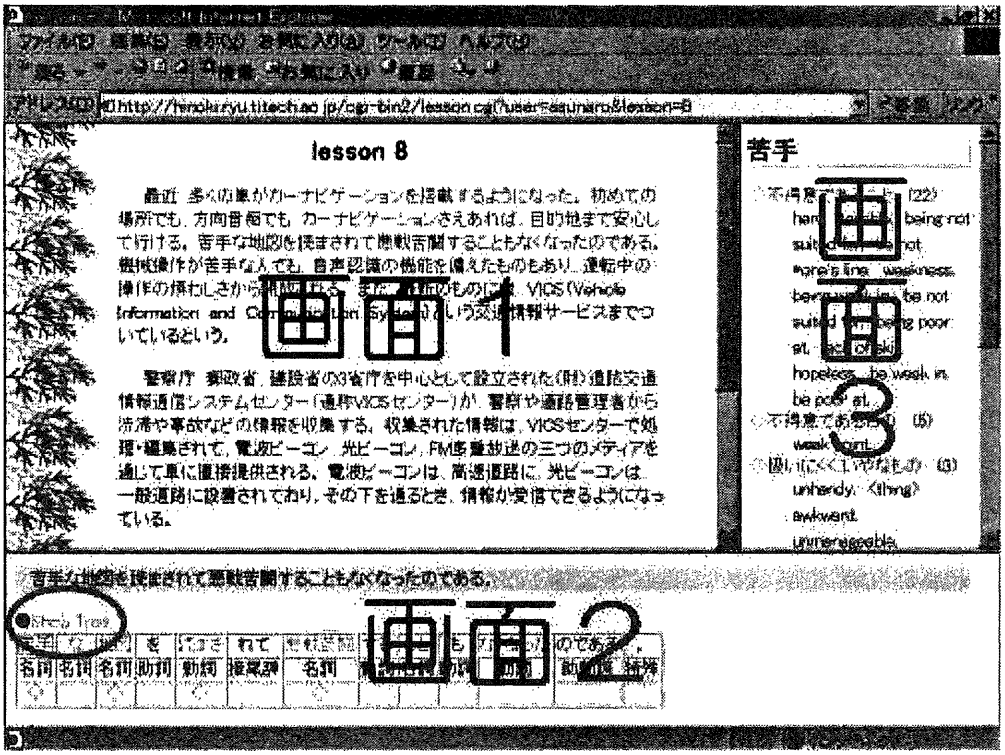

図 1: 「あすなろ」の画面表示

4.2 外国語「書取」学習支援システム

「書取」(ディクテーション)は、外国語の文章を聞き、それを正確に書き取ることである。外国語の学習法の1つに、このディクテーションにより外国語力を高める学習法がある。

従来ディクテーションによる外国語学習を行う場合、学習者はカセットテープ等を使用し、ディクテーションをした結果は教師が目で見て正解か不正解かを判断していると思われる。しかし、この一連の作業は、学習しづらく、また、教師にとってもつらい作業となっている。

そのような面倒な作業を簡略化しスムーズに行えるように、我々は、ディクテーション方式英語学習支援システムを開発している[8]。本システムでは、英文を音声で流し、学習者が聞き取った英文をシステムに入力すると、システムが自動採点をする仕組みになっている。また、本システムは現在 WWW 上で利用可能になっている。

システムは、音声を出力するモジュール、ユーザが書きとって送信してきた答案と正解を照合するモジュール、照合結果からユーザの誤りに関するコメントを生成するモジュールから構成される。ユーザの書きとった結果と正解の照合および、自動的な誤り個所検出には、自然言語処理や音声認識の分野で日常的に利用されている DP マッチング技術を利用している。

我々が研究を発表した翌年には、同様に言語処理学会の年次大会で、東北大のグループが日本語版を開発したことを発表しており、また、我々の発表の際には、多くの英語教師から問い合わせがあったことから、この種の外国語の「書取」学習を支援するシステムは語学教育において有用であり、言語処理研究者と語学教育関係者が共同で、実際に使えるツールとして整備していくことが今後重要であると考える。

5. 今後の展望

本稿では、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、1999 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御尽力で設立された(詳

細は、http://tanaka-www.cs.titech.ac.jp/gsk/を参照して頂きたい)。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。

すでに、音声、言語処理技術を活用して、さまざまな語学教育教材、CALLシステムが開発されつつあるし、今後もその数は増えていくと思われる。語学教育関係者と、音声、言語処理研究者が共同でプロジェクトを組み、語学教育に有用なシステムを開発していくことも、音声、言語処理研究者にも関心が高まっていることから、今後ますます増えていくものと思われる。語学教育関係者が持っているニーズと、音声、言語処理研究者が持つ技術(シーズ)をお互いに率直に話し合い、一つの枠組で仕事をすることで、真に有用なシステムが実現できるのではないかと考える。

そういう意味で、現在、文部科学省科学研究費特定領域研究(A)「高等教育改革に資するマルチメディアの高度利用に関する研究」(平成11年--14年度)が活動中であり、そのような試みとなっているのは興味深い(http://www.nime.ac.jp/tokutei120/index.html)[1]。

最後に、本稿では列挙しなかった情報へのポインタを上げておく。人工知能学会誌に連載されている「私のブックマーク」(http://www.ai-gakkai.or.jp/isai/journal/mybookmark/)には自然言語処理関連のものもあり、有用な情報を含んでいる。藤氏による国内言語学関連研究機関WWWページリスト(http://www.sal.tohoku.ac.jp/~gothit/kanren.html)には膨大な情報があり、この分野のポータルサイトと言って良い。月刊言語に連載されている「インターネット言語学情報」にも、有用な情報が掲載されていることがある。また、我々の研究室にあるリンク集(http://lr-www.pi.titech.ac.jp/link_j.html)も良ければご参照頂きたい。

参考文献

[1] 壇辻正剛, IT化時代の語学環境としてのCALL, 情報処理, Vol.42, No.10,pp.1001-1005, 2001.

[2] 特集 使いやすくなった自然言語処理のフリーソフト -知っておきたいツールの中身,情報処理, Vol.41, No.11, 2000.

[3] 長尾真, 他, 自然言語処理, 岩波書店, 1996.

[4] 田中穂積監修, 自然言語処理 — 基礎と応用 —, 電子情報通信学会編, コロナ社, 1999.

[5] Oliver Christ, Corpus Exploration Tools for Lexicography, EURALEX'96 utorial Script, 1996.

[6] 工藤 拓, 松本裕治, RDBを利用したタグ付コーパス検索支援環境の構築, 情報処理学会自然言語処理研究会報告, 144-19, pp. 135-142, 2001.

[7] 仁科喜久子, 奥村学, 杉本茂樹, 八木豊, 傅亮, 阿辺川武, 戸次徳久, 外国人のための科学技術日本語読解支援システム「あすなろ」の開発, 教育工学関連学協会連合第6回全国大会講演論文集, 第1分冊, pp. 495-498, 2000.

[8] 神谷 泰弘, 望月 源, 奥村 学, 島津 明, ディクテーション方式英語学習支援システムの開発, 言語処理学会第6回年次大会発表論文集, P2-7, pp. 125-128, 2000.

著者紹介

奥村 学:1962年生. 1989年東京工業大学大学院博士課程修了. 同年, 東京工業大学工学部情報工学科助手. 1992年北陸先端科学技術大学院大学情報科学研究科助教授, 2000年東京工業大学精密工学研究所助教授, 現在に至る. 工学博士. 自然言語処理, 知的情報提示技術, 語学学習支援, テキストマイニングに関する研究に従事. E-mail: oku@pi.titech.ac.jp.