

## SLP・NL 合同セッション「ここまでできるぞ音声/言語処理技術」 — 言語編 —

松本 裕治 (奈良先端大) 武田 浩一 (日本 IBM) 永田 昌明 (NTT)  
宇津呂 武仁 (奈良先端大) 田代敏久 (マイクロソフト) 山下達雄 (奈良先端大)  
林 良彦 (NTT) 渡辺日出雄 (日本 IBM) 竹澤 寿幸 (ATR)

近年、電子化テキストの急激な増加、および、インターネットによる一般利用者の電子媒体への日常的なアクセスに伴って、言語処理研究と言語に関する実用技術の間のギャップが徐々に狭まってきており、実用的な自然言語処理研究という言葉が真に現実的な意味を持ち出してきた。本報告では、そのような実用的言語処理技術の事例のいくつかを「ここまでできるぞ言語処理技術」というタイトルで紹介する。

## SIG-SLP/SIG-NL Joint Session “Recent Advances in Speech and Language Processing Technologies” — Language Processing Technologies —

Yuji Matsumoto (NAIST) Kouichi Takeda (IBM) Masaaki Nagata (NTT)  
Takehito Utsuro (NAIST) Toshihisa Tashiro (Microsoft) Tatu Yamashita (NAIST)  
Yoshihiko Hayashi (NTT) Hideo Watanebe (IBM) Toshiyuki Takezawa (ATR)

In recent years, the amount of online text data rapidly increased, and ordinary users have come to access online texts through internet in their daily lives. This change around the situation of language data and language processing technologies gradually made the gap between research on natural language processing and practical technologies for processing natural language narrower, and has given the term “research on practical natural language processing” much more reality than before. This article reports case studies of those practical natural language processing technologies with the title “Recent Advances in Language Processing Technologies”.

### 1 言語処理技術の現状

実用的な自然言語処理研究という言葉が真に現実的な意味を持ち出したのはほんの最近のことである。極端な言い方をすれば、自然言語処理研究の成果が現実的な応用に活かされることは、残念ながらこれまでほとんどなかったといってよい。もちろん、自然言語に関連した実用技術はいくつも存在する。かな漢字変換や文書検索システムは既に実社会に深く浸透しているし、多くの機械翻訳プログラムが一般利用者向けに安価に販売されている。しかし、これらの実用技術と自然言語処理研究が互いに

フィードバックされつつ両者の発展に貢献するという段階にはいたっていなかった。近年の電子化テキストの急激な増加、および、インターネットによる一般利用者の電子媒体への日常的なアクセスという状況が、言語処理研究と言語に関する実用技術の隙間を否応なく埋め始めている。

本報告では、そのような事例のいくつかを「ここまでできるぞ言語処理技術」というタイトルで紹介する。以下で示すように、かな漢字変換、形態素解析、文字認識誤り訂正、多言語検索、機械翻訳など、バランスの取れた応用研究の事例を集めることができた。もちろん、ここで紹介する事例が網羅的

であることはなく、他にも多くの優れた研究事例がある。ここでの試みを手始めにして、今後の研究会と同様の機会を持って行きたいと考えている。

## 2 日本語入力システム MS-IME98

田代敏久 (toshta@microsoft.com)  
マイクロソフト株式会社

### 2.1 概要

コンピュータ上での快適な日本語入力を可能にするためには、変換精度が高く、かつ操作性のよい日本語入力システム(仮名漢字変換システム)が必要である。ここでは、本年3月に発売されたMS-IME98<sup>1</sup>の特徴を、主として言語処理技術の点から説明する。

### 2.2 最小コスト法の自然な拡張としてのN-BEST手法

従来、かな漢字変換システムの解析処理法としては、文節分かちを決定するための形態素解析処理として最小コスト法を利用し、同音異義語の最終決定に格フレーム処理や係り受け処理などの意味解析手法を利用することが一般的であった。しかし、この従来方法では、形態素解析処理と意味解析処理の統合方法が不明瞭になりやすく、より高い変換精度を実現するための辞書・文法のチューニングを機械的に行うことや、より操作しやすい編集方法を提供することが困難であった。そこで、MS-IME98では、最小コスト法の自然な拡張であるN-BEST探索手法[Nagata94]を仮名漢字変換システムに応用することにより、システムチックな辞書・文法の改良作業を可能にし、高い精度と、既存の仮名漢字変換システムにない新たな編集手法を実現した。

### 2.3 コーパスを利用した辞書・文法チューニング

パーソナルコンピュータのユーザ層が広まるに伴い、日本語入力システムの精度に対する要求はますます高まる一方である。こうしたユーザの声をいち早く製品に反映するためには、辞書や文法の改

<sup>1</sup><http://www.microsoft.com/japan/office/ime/>

良作業を、できる限り合理的に行うことが必要である。MS-IME98では、言語解析処理部にN-BEST手法を採用することにより、辞書や文法の情報をコーパスに基づいて改良するできる仕組みを実現し、より短期間で質のよい辞書・文法チューニングを開発することが可能になった。

## 3 日本語形態素解析システム 茶筌

山下 達雄 (tatuoy@is.aist-nara.ac.jp)  
奈良先端科学技術大学院大学 情報科学研究科

茶筌は、奈良先端科学技術大学院大学松本研究室で開発されたコスト最小法による日本語形態素解析システムである。JUMAN2.0[松本94]が原形となっている。

広く一般に利用されることを目的としてフリーの形態素解析システムとして公開されている。機械翻訳や対話処理などの自然言語処理システムの前処理に利用されているだけでなく、freeWAISなどの日本語全文検索ソフトのインデキシングなどにも利用されている。UNIX版だけでなく、Windows 95版も提供されており、国文・教育関連の文系の研究者にも利用されている。また、付属の約20万エントリの形態素辞書はフリーのものとしては最大規模で、他の形態素解析システムや、形態素解析以外の研究にも多く利用されている。

茶筌の特徴として、品詞体系、形態素辞書の項目、接続規則がユーザにより自由にカスタマイズできるという点があげられる。現在、公式にサポートされている品詞体系には益岡田窪文法とRWCPで採用されているもの(学校文法に準拠)がある。後者は現在辞書の整備を行なっている最中で、近日中に公開予定である。

また、文法だけでなく、コスト最小法で利用する、形態素コスト・接続コストを辞書や接続規則定義ファイルで設定することができる。これにより、コーパスから統計的に学習した確率値をコストに変換して用いれば、茶筌を確率モデルに基づく形態素解析システムとして利用できる。

グラフィカルユーザインターフェース(図1)の開発や可変長接続規則の実装などの機能拡張が行われており、今後も継続されていく予定である。

茶筌に関する詳しい情報はマニュアル[松本97]を、最新の動向に関しては茶筌ホームページ<sup>2</sup>を参

<sup>2</sup><http://cl.aist-nara.ac.jp/lab/nlt/chasen.html>

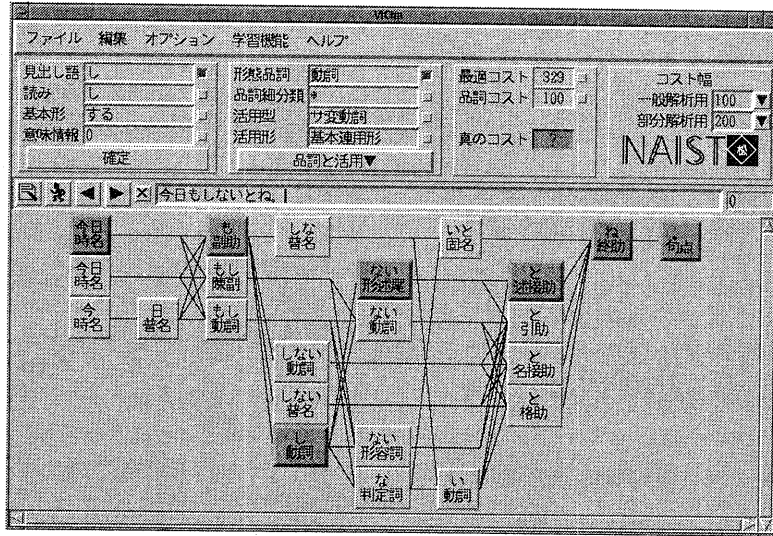


図 1: 茶釜の解析結果を表示する GUI

照されたい。

#### 4 文字認識用 日本語スペルチェッカ jspell

永田 昌明 (nagata@nttnly.isl.ntt.co.jp)  
NTT 情報通信研究所

我々は高精度の文字認識用日本語スペルチェッカーを開発した [Nagata96, Nagata98]. 認識誤りの半分から三分の二を自動訂正するので、従来は 70~90% だった光学的文字読み取り装置 (OCR) による手書き文字認識精度を、90~95% にまで改善できる (図 2). これにより、ファックスや郵便による申込書、アンケートなどの電子化作業を大幅に効率化できる。

文字認識は個々の文字を認識単位とするので、漢字の「口」と片仮名の「ロ」のように形状が似た文字を区別することは難しい。そこで、文字列が言語情報を表す場合、前後の文脈の情報を用いて正しい文字を判定する方法が従来より研究されている。

入力文字列が単語であり、かつ、住所や商品名のように語彙が制限されている場合は、文字候補列と単語辞書の単純な照合により認識誤りの検出・訂正が可能であり、これは既に実用化されている。一方、入力文字列が文章の場合は、単語の文法的接続可能性を検査する方法が提案されているが、単純

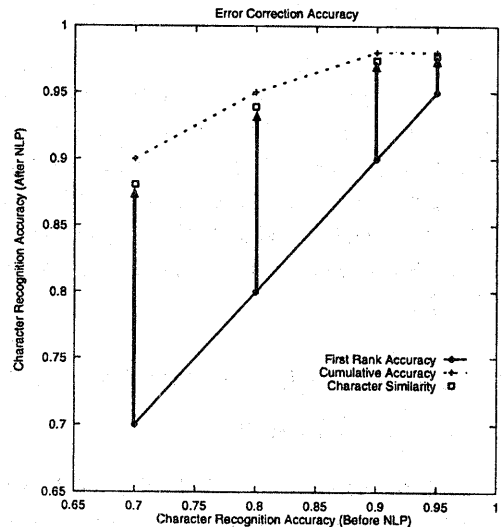


図 2: 誤り訂正の精度

な文法検査だけでは制約が弱すぎることや、誤読文字列と未知語の判別が難しいことから、実用化にされていなかった。

本技術は文章を構成する品詞・単語・文字の並びの出現確率を大量の日本語データから事前に学習し、誤りの前後の文脈から最適な単語候補を選び出す。その手順は以下の通り (図 3)。



なってきた。

TITAN では、検索結果である各ページにおける語の分布情報を基に on-the-fly でクラスタリングを行わせることができる。また、Xerox PARC により提案された Scatter and Gather と同様のインタラクティブな再クラスタリング機能もサポートしている。さらに、Query 中の語の結合の仕方 (and/or) に応じた動的な語の重み付けなども試みており、ユーザの直感に合うようなクラスタリング方法を模索している。このようなクラスタリング機能を用いることにより、ユーザはインタラクティブに望ましい情報へ接近できる可能性がある。ただし、有効性の評価については、評価法を含めて検討していく必要がある。

### 5.2.2 意味の一貫性に基づく Query Terms の翻訳

Query の翻訳の問題は、Query 中に含まれる語のリストをもっとも良く翻訳する問題として近似することができる。すなわち、リストの各語に対して、辞書に記載されている訳語の中から一つを選んで置き換えたものを“訳語セット”と呼ぶと、可能な訳語セットのなかから最適なものを選択することが問題である。我々は、“複数の訳語セットのうち、構成する単語が意味的に最も一貫している (coherent) ものを最も尤もらしい訳語セットとして選択する”という基準により最適な訳語セットを得る方法を検討している [Kikui98]。

ここで、意味の一貫性の尺度としては、コーパスから獲得した語の共起行列を SVD 手法により縮退させた多次元空間上で、訳語セットを構成する語がどの程度“狭い範囲”に集中しているかという尺度を用いる。すなわち、目的言語において類似した文脈に出現しやすいとコーパス統計から推定される語集合を意味的に一貫した最適な訳語セットとして選択する。新聞記事を対象とした実験によれば、この手法により 82% の翻訳精度が得られた。

## 6 パターンベース英日翻訳システム PalmTree

渡辺 日出雄, 武田 浩一

(watanabe@trl.ibm.co.jp)

日本アイ・ビー・エム株式会社 東京基礎研究所

本システムは、パターンベース翻訳手法 [Takeda96a, Takeda96b] を用いた英語から日本語への翻訳システムであり、弊社のインターネット向け翻訳システム「翻訳の王様」の翻訳エンジンとして使われている。以下では、このシステムの特徴について簡単に述べることにする。

### 6.1 パターンベース翻訳

パターンベース翻訳手法とは、ソース側の CFG 規則とそれと対応する CFG 規則のペアを翻訳パターンとし、このソース側の CFG 規則を用いて構文解析を行い、解析結果を構成するソース側 CFG 規則と翻訳パターン中で対を成すターゲット側 CFG 規則から同期導出によりターゲット側構造を構築するというものである。

翻訳パターンは一般に以下のような形式をしている。

$$SR_1 \dots SR_n \Leftarrow SL \ TL \Rightarrow TR_1 \dots TR_m$$

ここで、 $SR_i$  はソース側右辺項、 $SL$  はソース側左辺項、 $TR_j$  はターゲット側右辺項、 $TL$  はターゲット側左辺項を表わす。それぞれの項は、

単語:品詞:IDX:属性

からなる。ここで、単語 (終端記号) と品詞 (非終端記号) はどちらか一つがある場合と両方ある場合が許される。IDX は右辺と左辺及びソース側とターゲット側の対応関係を表わす。属性としては、マッチング条件などを指定することが出来る。以下は、翻訳パターンの例である。

- (p1) take:VERB:1 a look at NP:2  $\Rightarrow$  VP:1  
VP:1  $\Leftarrow$  NP:2 wo(dobj) miru(see):VERB:1  
(p2) NP:1 VP:2  $\Rightarrow$  S:2 S:2  $\Leftarrow$  NP:1 ha VP:2  
(p3) PRON:1  $\Rightarrow$  NP:1 NP:1  $\Leftarrow$  PRON:1

(p1) は英語の慣用表現 “take a look at,” の翻訳パターンであり、(p2) と (p3) はそれぞれ一般の文法的な翻訳パターンをあらわしている。

構文解析は通常の CFG パージングの手法を用いる。例えば、チャートパージングでは、

- (1) ある不活性アークの規則の左辺項  $T_A$  とマッチする項を最右辺項  $T_B$  として持つ規則から新たなアークを作成する。

- (2) ある不活性アークAの始点を終点とする活性アークBがあるとき、Aの規則の左辺項 $T_A$ とBの規則の最左右辺活性項 $T_B$ がマッチすれば、AとBから新たなアークを作成する。

というのが、基本動作である。マッチするかどうかは、通常 $T_B$ の品詞・単語・マッチング属性等が $T_A$ のものとマッチするかどうかで決定される。通常のCFG規則を構成する項は単語(終端記号)か品詞(非終端記号)のどちらかであるが、パターンベース翻訳では両者があっても良い。そこで、両者がある場合は、AND条件でマッチングをチェックするというようにしている。

このシステムの利点としては、(1)解析、トランスファー、生成をひとつにまとめているのでシステムとしての複雑さが低減している、(2)翻訳知識が翻訳パターンに集約されているので管理が容易である、(3)訳質の向上が翻訳パターンの追加により容易にかつ漸進的に行なうことができる、などを上げることができる。

## 6.2 パターンコンパイラ

上記の(3)に関しては、パターンコンパイラという、簡易なレベルのパターン記述を上記の様な内部レベルのパターン表現に変換するものを作成し、これにより翻訳パターンの追加を容易にしている。例えば、以下のような簡易パターン<sup>4</sup>は、

between #:1 and #:2 =PP=  
#:1 と #:2 の 間に

次のような内部パターン表現<sup>5</sup>に変換される。

between NP:1 and NP:2  $\Leftarrow$  PP  
PP  $\Rightarrow$  NP:1 と NP:2 の 間に

## 6.3 まとめ

パターンベース翻訳システムでは、従来の手法に比べ訳質向上が非常に容易である。基本の文法的翻訳パターン部分は専門の文法開発者が必要であるが、その他の翻訳パターンは、通常の学校で習う程度の英語の知識がある人であれば十分に作成可能である。

<sup>4</sup>ここで、#:1や#:2は、任意要素とマッチすることを表す変数項である。

<sup>5</sup>誌面の関係で省略したが、実際にはこれ以外に各種の属性等が付加されている。

## 7 日英音声翻訳システム ATR MATRIX

竹澤 寿幸 (takezawa@itl.atr.co.jp)  
ATR 音声翻訳通信研究所

自然な話し言葉を対象とした日本語から英語への音声翻訳システム ATR MATRIX を構築した [竹澤 98, Reaves98]。普段我々が使うような自然な音声を認識し、英語へ翻訳し、合成音声で出力することができる。システム全体はワークステーション1台またはハイエンドパソコン1台で動作し、ほぼ実時間で処理を行うことができる。今回構築したシステムでは「ホテルの予約」を対象としている。

### 7.1 不特定話者音素環境依存音響モデルと可変長 $N$ -gram 言語モデルを用いた実時間音声認識

不特定話者の音素環境依存音響モデルを作成する統計的な手法を用いて、男性用の不特定話者音素 HMM モデルと女性用の不特定話者音素 HMM モデルを用意した [内藤 98]。自然な話し言葉に含まれる多様な言い回しをコンパクトにモデル化できる統計的な手法である可変長  $N$ -gram により言語モデルを用意した [内藤 98]。単語グラフに基づく効率的な単語仮説数削減手法を用いた実時間処理を実現している [山本 98]。

### 7.2 音声認識結果を扱うためのロバスタな言語翻訳

翻訳では、文の構造を判断するだけでなく、対訳用例を用いることにより、話し言葉に現れる種々の表現を取り扱うことができる [古瀬 97]。さらに、一部に誤りを含んだ音声認識結果を扱うためのロバスタな言語翻訳として部分翻訳機能を実現している。次の二つのヒューリスティクスを仮定し、確からしい部分を選択して翻訳することができる [脇田 97]。

- (1) 対訳用例に類似した言語表現は誤認識の度合も少なく、解析結果の信頼性も高い。
- (2) より長い語句の範囲で解析できた結果の方が誤認識の度合も少なく、解析結果の信頼性も高い。

### 7.3 個性豊かな音声合成

音声認識結果とともに音声認識過程で選ばれた話者モデルの情報を出力することができる。今回は、入力音声から話し手が男性か女性かを判断することができる。選ばれた話者モデルに関する情報に基づき、音声合成サブシステム [キャンベル 96] はそれに応じた声で音声合成を行う。話者モデルの数を増やせば、さらに個性豊かな音声合成を実現できる。

### 7.4 自然な会話を扱うための技術

自然な会話では、文ごとに区切らず、「ちょっと高いですね。もっと安い部屋は無いですか。」のように二つ以上の文をつないで発話することがある。その場合でも正しく1文ごとに翻訳する必要がある。そのような境界位置に無音区間(ポーズ)が挿入されることもあるが、そうでないこともある。境界位置の前2単語と後1単語の合計3単語の範囲の品詞・活用形・活用型を利用する手法により、発話を言語処理の単位に変換することができる [竹澤 97]。

さらに、自然な会話では、「部屋は空いています?」のように文末を上げることによって疑問を表すことがある。音の高さの変化(韻律)を検出して疑問文かどうかを判断することができるので、その情報を言語翻訳部に渡すことで“Rooms are available.”ではなく“Are rooms available?”のような翻訳を実現することができる。

### 7.5 今後の予定

現在、構築したシステムの性能評価を進めているところである。今後はさらに、英日方向の同様なシステムを構築し、最終的には日英、英日双方向での会話が可能なシステムを実現する予定である。また、多言語間の音声翻訳をめざし、C-STAR II という国際コンソーシアムを通じて世界各国の研究機関と研究協力を進めており、1999年度に多言語音声翻訳の国際共同実験を行う予定である。

## 参考文献

[キャンベル 96] ニック・キャンベル, アラン・ブラック: “CHATR: 自然音声波形接続型任意音声合成システム”, 信学技報, SP-96-7, pp. 45-52 (1996-05).

[古瀬 97] 古瀬蔵, 美馬秀樹, 山本和英, Michael Paul, 飯田仁: “多言語話し言葉翻訳に関する変換主導翻訳システムの評価”, 言語処理学会第3回年次大会発表論文集, pp. 39-42 (1997-03).

[Hayashi97] Y. Hayashi, G. Kikui, and S. Susaki: “TITAN: A Cross-Linguistic Search Engine for the WWW”. *AAAI-97 Spring Symposium on Crosss-Language Text and Speech Retrieval*. (1997).

[Kikui97] G. Kikui: “Identifying the Coding System and Language of On-line Documents Using Statistical Language Models”. *Transactions of IPSJ*, Vol.38, No.12, pp.2440-2448. (1997).

[Kikui98] G. Kikui: Term-list Translation using Mono-lingual Word Co-occurrence Vectors. *Proc. of the 17th COLING and the 36th ACL*, 1998 (to appear).

[松本 94] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真: “日本語形態素解析システム JUMAN 使用説明書 version 2.0”, NAIST Technical Report, NAIST-IS-TR94025, July 1994.

[松本 97] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明: “日本語形態素解析システム [茶筌] version1.0 使用説明書”, NAIST Technical Report, NAIST-IS-TR97007, February 1997.

[Nagata94] M. Nagata: “A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A\* N-Best Search Algorithm”, *Proc. of the 15th COLING*, pp.201-207, 1994.

[Nagata96] M. Nagata: “Context-Based Spelling Correction for Japanese OCR”, *Proc. of the 16th COLING*, pp.806-811, 1996.

[Nagata98] M. Nagata: “Japanese OCR Error Correction using Character Shape Similarity and Statistical Language Model”, *Proc. of the 17th COLING and the 36th ACL*, 1998 (to appear).

[内藤 98] 内藤正樹, 政瀧浩和, Harald Singer, 塚田元, 匂坂芳典: “日英音声翻訳システム ATR-

- MATRIX における音声認識用音響・言語モデル”, 日本音響学会平成 10 年度春季研究発表会講演論文集, Vol. I, pp. 159-160, 1998.
- [Reaves98] B. Reaves, A. Nishino, and T. Takezawa: “ATR-MATRIX: Implementation of a Speech Translation System”, 日本音響学会平成 10 年度春季研究発表会講演論文集, Vol. I, pp. 53-54, 1998.
- [Takeda96a] K. Takeda: “Pattern-Based Context-Free Grammars for Machine Translation,” *Proc. of the 34th ACL*, pp. 144-151, 1996.
- [Takeda96b] T. Takeda: “Pattern-Based Machine Translation,” *Proc. of the 16th COLING*, Vol. 2, pp. 1155-1158, 1996.
- [竹澤 97] 竹澤寿幸, 森元遼: “発話単位の分割または接合による言語処理単位への変換”, 情報処理学会研究報告, 97-SLP-18-4, Vol. 97, No. 101, pp. 19-24, 1997.
- [竹澤 98] 竹澤寿幸, 森元遼, 匂坂芳典, Nick Campbell, 飯田仁: “日英音声翻訳システム ATR MATRIX”, 情報処理学会第 56 回全国大会, Vol. 2, pp. 279-280, 1998.
- [脇田 97] 脇田由実, 河井淳, 飯田仁: “意味的類似性を用いた音声認識正解部分の特定法と音声翻訳手法への応用”, 人工知能学会研究会資料, SIG-SLUD-9603-2, pp. 7-12, 1997.
- [山本 98] 山本博史, シンガーハラルド, リーブスベン, 匂坂芳典: “日英音声翻訳システム「ATR-MATRIX」における音声認識部分の構造と制御方法”, 日本音響学会平成 10 年度春季研究発表会講演論文集, Vol. I, pp. 161-162, 1998.