AAMT

The Asia-Pacific Association for Machine Translation

Journal

No.26
April 1999

アジア太平洋機械翻訳協会

目　　次

エッセイ

機械翻訳とそのマーケット

― シンガポールサミットにむけて ―

東京大学大学院教授　辻井　潤一

「機械翻訳の歴史は、過剰な期待とその反動としての失望の繰り返しである」とよくいわれる。実際、翻訳というのは、本職の翻訳家が行っても、なかなか満足できるものはできない。また、わかりにくいので、原文を読んでみたら原文もやはりわかりにくかった、というのもよく経験することである。翻訳は、機械がやろうと人間の専門家がやろうと難しい。２人の人間のコミュニケーションを助ける善意の第３者が、最後に誤解の責任をとらされて非難される、というのはよくあることである。

この第３者の役割、しかも、異なる言語を使用するものの仲介者を機械がしようとするのが機械翻訳システムだから、賞賛されることのない損な役割を担うシステムではある。

また、機械翻訳のように、人間ができることを機械にやらせる技術はその難しさが理解されず、なかなか感心してもらえない損な技術分野だといわれてきた。しかし実際には、翻訳というのは、このように人間もうまくできないものであるから、多少でもうまくできれば、膨大な科学計算を行う計算機システムが賞賛の目でみられるように、機械翻訳も賞賛の目で見られることは十分にあり得る。

機械翻訳が損な役割を担うシステムだと思っていたのは、実は、これまで翻訳の必要性がそれほど幅広くなく、使用者の満足のいく翻訳は人間であれば作り出せる、また、そのような人間（翻訳家）が十分にいる、という幻想を社会全体が持っていたからではないか？

たとえば、ヨーロッパにいると、彼らにとって中国語、日本語でかかれたテキストが英語やフランス語に翻訳されることは一種のマジックである、ということがよくわかる。あるいは、韓国語のテキストが日本語に翻訳されるのをみても、我々日本人には、一種のマジックのように見えるだろう。語族の近い日本語と韓国語の翻訳は、英語と日本語の翻訳にくらべて、技術的にははるかに「マジック」の度合いが低いのだが、そのように感じる。

結局は、使用者の両言語に関する知識がどの程度かで、システムがマジックに見えるかどうかが決まり、賞賛されるかどうかが決まる。使用者の能力がシステムの評価に決定的に影響するというのが、機械翻訳を道具としてみたときの大きな特徴であろう。

いままで、日本では、その経済効果という点から英語と日本語の翻訳システムに焦点が当てられてきた。また、システム使用者も英語を使う必要のある人、したがって相応の英語の能力がある人を前提としていた。このために、「自分が翻訳すればもっとうまく翻訳できる」という自信過剰な使用者から、技術のマジックとしての度合いを不当に低く評価されてきたのではないか？

この状況は、ここ数年で大きく変わりつつある。実際、ウェブブラウザー用の翻訳システムでは、英語能力をさほど持たないユーザがそれを使用するようになってきている。このような傾向は、インターネットが社会に浸透すれば浸透するほど強くなるだろう。同時に、インターネットの浸透は国際的なコミュニケーションにおける英語の役割を相対的に低下させているようでもある。インターネットのユーザが広がれば広がるほど国内の情報マーケットが大きくなると同時に、英語をコミュニケーションの手段としない人たちが国際的なコミュニケーションに参加するようになってくる。

その結果として、国際的な情報収集が日常的に行われ、たとえば、日本のインターネット使用者が韓国語や中国語のウェブサイトにアクセスする必要性、英語の知識が全くない使用者が、英語テキストを読

む必要性は、近未来どころか現在すでに起こりつつある。機械翻訳がマジックとみられる状況は整ってきているのである。

ヨーロッパにおける多言語主義の復活は、文化面での米国中心主義への反発という復古主義というより、むしろ、インターネットを通じたコミュニケーションの広がりへの適応という現代的、前向きな意味を持っている。あるいは、インターネットを通じた商取引（Electric Commerce）における個別言語の重要性を意識した、きわめてプラグマティクなスタンスだと考えるべきであろう。

アメリカやヨーロッパにおける機械翻訳、あるいは多言語情報処理に対する急激な関心の高まり、とくに、中国語、韓国語、日本語処理への関心の高まりは、このような具体的なニーズの高まりを反映している。機械翻訳が過剰な期待、あるいは夢を売り、具体的なニーズを掘り起こす時期はすでに終わった。むしろ、具体的なニーズがありそれに技術がどう応えるか、という時代になっている。　あるいは、機械翻訳技術は、いま人間の能力が応えることのできない社会の要請、その要請に応えるためのマジックとなりえる時代を迎えつつある。

機械翻訳の技術者や研究者が、企業家、使用者としての翻訳者と出会い、真剣に議論できる唯一の会議として、MTサミットがある。この会議は80年代、日本における機械翻訳の研究者が、この分野における技術の使用者の重要さ、使用者と開発者の交流の重要さをいち早く認識し、日本の主導権で始まった会議である。２年に一度、ヨーロッパ、北米、アジアの機械翻訳協会が交代に主催してきており、今回、その７回目を迎える。

今回のサミットは、アジア太平洋機械翻訳協会の主催で、インターネットによるコミュニケーションの世界的な広がりと深化を象徴するように、東洋と西洋のコミュニケーションの接点的な役割を担うシンガポールで行われる（本年９月14日―16日）。アジアや日本の経済危機という逆風の中ではあるが、ヨーロッパ、北米の企業家、開発者、研究者の関心は高い。機械翻訳という技術の特殊性は、使用者、企業家、開発者、研究者の交流を不可欠なものとしている。

日本からも多くの人が参加し、機械翻訳技術がグローバルコミュニケーションの時代のマジックとなることに貢献していただければ、と思っている。

会員投稿

MT翻訳者の生活と意見

森本　健一（ロサンゼルス在住、当会会員）

1．私とコンピュータ

AAMT Journalにはどちらかというと、生産者ないしは研究者からの視点で書かれた記事が多いようなので、翻訳者ユーザとしてのMTの使用体験をフィードバックさせていただきます。

関数機能付き電卓－シャープのポケコン－Tandyのラップトップ－（以下デスクトップ）NECの9801E－PowerMac6100－Dell Dimension Pentium166C/XPS。これは、北欧の航空会社で航空機運行管理者として社会への第一歩を踏み出してから、旅客営業、人事を経て停年前退職を決行し、在米在宅翻訳者として今日に至るまでに使用した機種を年代順に並べたものです。関数機能付き電卓を含めたのはこれも立派な電子計算機だからです。シャープのポケコンではプログラミング言語にBasicを使い航空運賃計算ソフトを完成、仕事に使っていました。会社を辞めたのはTandyのラップトップを使用していたときです。コンピュータが市販されるようになる前からコンピュータの可能性に強い関心を抱いていた私は、今思えば必要もないのに微積分の復習に没頭し、通勤電車ではもっぱら数学関係の本を読んでいました。そして、Tandyのラップトップに巡り合ったとき、これさえあれば不可能な作業はないと思い込んでしまったのでした。こうしたパソコンを使って航空運賃計算の能率向上を会社に提言して受け入れられなかった私は自前で情報サービスを始めることをもくろみ、停年を遥か前にして退職を決意しました。NEC 9801Eほか周辺機器を買い込み開業の準備を始めていた私はある日書店で将来実現することを夢見ていた情報誌の創刊号が書棚に平積みされているのを見て肝をつぶします。大企業の出版社が私の夢をブルトーザーで踏み潰したのです。気を取り直した私は自宅でできる翻訳業へと転身をはかり、現在に至っています。コンピュータさえあれば何でもできると早とちりして会社を辞めた私の不覚でした。

それはともかく、コンピュータはその普及前には厳然としていた理系と文系の垣根を取り除かないまでも、引き下げたのではないでしょうか。現に、MTの世界では文系と理系の共同作業が不可欠になっています。そこでは文系にも理系の発想が、理系にも文系の知識が求められているはずです。文系でもコンピュータ操作の技能がなければ少なくともアメリカではビジネス社会で受け入れられませんし、理系でも業務内容の知識がなければコンピュータは無用の長物です。これからも文理融合の傾向は強くなっていくのではないでしょうか。

2．機械翻訳との出会い

1985年、最初の市販機械翻訳ソフトBravisが世に出たときには、早速デモを覗きに出かけたものです。それでも、猜疑心を捨て切れない私は開発企業の用意原稿によるデモでなく、私の持参原稿で機械翻訳を見たいという意地悪な動機で、Bravisのデモ係員に英語の原稿を手渡したのでした。予想通りの無残な出来栄えでしたが、とにかく英語が日本語に変換される事実に感動したものでした。しかし、時代の先端を行き過ぎていたのでしょうか、Bravisはやがて姿を消しました。でも、その日以来MTが私の頭を離れることはありませんでした。

英日の翻訳であれば400字詰め原稿用紙に手書き、日英の翻訳であればタイプライターというのが通り相場でしたから、パソコンによるワープロ処理をほどこした私の印字原稿はクライアントに喜ばれました。ワープロ専用機は市販されて間もないころですから、英文処理のWangとかIBMのワープロは高価で個人の翻訳者には手が出ませんでした。

ロサンゼルスのパソコンショップで富士通のMTソフトAtlasのカタログを見たのはそれから11年経過してからです。ほかにどのようなMTソフトが市販されているのかも分かりませんでした。比較する対象の競合ソフトを調べることもままならず、最近のMTソフトの価格に比べると5倍くらいはする英日・日英翻訳の製品を別料金の専用専門用語辞書2分野とともに求めました。早速一括翻訳で受注英文原

稿の全文をこのソフトにかけてみました。翻訳の出来不出来はともかく、英文が日本語に変換されて数分で訳出されることに素朴な感動を覚えました。しばらくAtlas　V3.0にお世話になった後、V4.0にアップグレードし、bekkoameのホームページで知った東芝のAstransac英日版V3.0を購入して現在では両者に英日の競訳をさせながら愛用しています。英日翻訳の受注が圧倒的に多いため、日英についてはご披露するだけの体験の蓄積はありませんが、後日機会があれば感想を述べたいと思います。

両者の優劣ですが、Atlasが翻訳不能で原文のまま返してくるときでも、Astransacは律義に訳をつけてくることがよくある点を除いては、実力は伯仲しているというのが私の感想です。その後両ソフトはバージョンアップされたり、Atlasは専門用語辞書と併用、Astransacでは専門用語辞書が不在というような事情から公平な製品比較にはなっていないかもしれません。

3．機械翻訳の実際

こうしたMTソフトを使用して最終製品を仕上げる訳ですが、まず私の作業の手順をご紹介しましょう。私はMS Wordをワープロとして使っていますが、作業はセンテンスごとに区切って行います。まずソース原稿とターゲット原稿（最初は空白）を画面に開き、Wordのウィンドウ機能を利用してソース原稿とターゲット原稿を上下に配します。ソース原稿のセンテンスをコピーして一方のMTソフトにかけ、訳文をターゲット原稿にペーストし、同じ作業をもう一方のMTソフトについて行います。こうしてできた訳文を上下に並べます。前編集はしません。出力された両者の訳でより優れた方を採用、後編集というより、訳文の表現、訳語で適切なものを採用、文章の構成はほとんど自力で行います。訳文をそのまま採用できることもまれにはあり、そのときには快哉を叫び指を鳴らします。そうまでしてMTにこだわるメリットは何でしょうか。これは専門用語辞書を使用しているAtlasについて特に言えることですが、辞書引きをかなり省けるということがひとつ。もうひとつは、原点に返った訳文に教えられるということです。考えすぎて訳文が的を外すことがこれで防げます。それと、MTに誤訳はあっても訳し漏れはまずありませんので、「抜け」がなくなるという効果が期待できます。このためには、完成した訳文をフレーズごとに区切ってMTの訳と突き合わせるという作業が必要になります。こうした一連の作業は本来MTが得意とする翻訳の高速化に逆行するのではないかという疑問がでてくるでしょう。特に時間との競争を強いられる翻訳者の場合、MTの省時間メリットがないように思えるかもしれませんが、慣れればそれほど時間を取る作業ではありません。確かに、全体としてMTを使用しために納期を短縮できるというほど時間の節約にはなっていませんが、上記のように捨て難いメリットがあります。MTソフトのバージョンアップも頻繁に行われています。その都度上位バージョンを買い続けるほど翻訳者の懐は豊かではありません。確かに近年MTソフトの価格は誰でも気楽に買えるレベルまで安くなりましたが、常に最新バージョンに買い替える余裕はありません。メーカーの皆さんへのお願いですが、少なくともワンバージョンで2年はアップグレードしないようご配慮ください。上記MTソフトのひとつでバージョンアップした版を購入、試したところ前に比べほとんど変化が見られなかった経験があります。バグの解消、機能の進化、他社との競合等、メーカーとしてバージョンアップしたい事情はあるでしょうが、アップグレードできずに、あせりを感じながら旧バージョンのMTソフトを使い続けるユーザの心情にも思いを馳せてくださるようお願いします。

4．機械翻訳の泣き所

MTで翻訳者が苦労するのは当然のことながら翻訳原稿が電子ファイルに限定されることです。最近ではソース原稿も6、7割程度まで電子ファイルで貰えるようになりましたが、そうでない場合はFAXか宅急便によるハードコピーです。いずれの場合でも、印字品質がよければスキャナーとOCRでハードコピー原稿の電子ファイルへの変換が可能ですが、そういう場合でもハードコピーが虫眼鏡でも読めなかったり、読めてもスキャナーにかからなかったり、結局キーボードからの手作業入力になってしまうことが多く、折角のスキャナー・OCRもその本領を発揮できないことが少なくありません。残念なことにFAXによる原稿受信が圧倒的に多く、当然FAXの印字品質は一般的に言ってよくありません。勢い手作業の入力作業に頼らざるを得なくなるのです。

それでも入力にかかった時間を料金に上乗せすることはできません。キーボードのブラインドタッチには多少自信があるのですが、それも判読に問題のない原稿の場合であって、推理を最大限に働かせて一字一句読み進んでいくの場合には、折角のキーボード入力技能が活かせません。スキャナーとOCRが使用できる場合でもスペルチェックは欠かせず、結構時間のかかる非生産的な作業となります。こういう翻訳工程以外の作業が入るため翻訳作業の合計時間は手作業とさほど変わらない結果になるのです。ソース原稿が100%電子ファイルになれば、作業のスピードアップは確実に実現できます。

ついでながら、MTを利用した翻訳例をご覧いただきます。

例1

In a circuit, the amount of current 'I' depends on its resistance 'R' and the voltage 'E'.*

Atlas（MT）

回路では、現在の＊I＊の量はその抵抗＊R＊と電圧＊E＊に依存する。

Astransac（MT）

回路では、現在の＊I＊の量がその抵抗＊R＊および電圧＊E＊に依存する。

最終訳例1

回路に流れる電流Iの量は、回路内の抵抗Rと回路に加えられた電圧Eによって決まる。＊

例2

Depending on the type of materials, this cutting machine can handle a diameter of 10 inches.*

Atlas（MT）

材料のタイプに頼っていて、この打抜き機は10インチの直径を扱うことができる。

Astransac（MT）

用品のタイプによると、この切断機械は、10インチの直径を扱うことができる。

最終訳例2

素材の種類にもよるが、この切削機械では直径10インチのものを加工できる。＊

5．電子辞書

以上私の翻訳作業の手順をご紹介しましたが、ここで欠かせないのが電子辞書です。手狭になってきた2Gb（執筆時14Gbにアップグレード）のHDD（ハードディスクドライブ）のスペースを節約するため、電子辞書でCD-ROMのままで使える辞書はできるだけHDDにインストールしないで使うようにしています。そのため、5枚のCD-ROMを収容できるCD-ROMチェンジャーを本体に内蔵、セットした電子辞書を必要になった時点でデスクトップに開いて使用します。使用している電子辞書はMicrosoft／小学館のBookshelf、Merriam-Webster Dictionary、Microsoft Press Computer Dictionary、Mosby's Medical encyclopedia（図解人体各部用語集）がCD-ROMのまま、HDDには富士通の専門用語辞書（医学、薬学）追加した英和・和英（電辞海）辞書を使用しています。Bookshelfは英和、和英は言うに及ばず国語辞書、類義語辞書、時事用語集などが入っていて重宝しています。Merriam-Websterは英英辞書、Microsoft Press Computer Dictionary、Mosby's Medical encyclopediaも英語版です。MTソフトとともにこうした電子辞書をタスクバーに並べ、その都度必要な辞書のボタンをクリックしてデスクトップに表示し辞書引きを行うのですが、これがどれだけ能率向上に寄与しているか計り知れないものがあります。ハードコピーの小学館Random House大英和辞書もページをめくることがめっきり少なくなりました。コンピュータ関連用語を調べるコンピュータ用語集も毎年最新のものを1点は買って、絶え間なく生成されるコンピュータ用語の検索に万全を期すよう心がけています。英日の用語集が出るのを待っていては間に合わないので、アメリカで入手できる、それもできるだけ新しいものを求めるようにしています。ちなみに、上記のComputer Dictionaryは同名の書籍に付録として入っていたものですが、今では付録が本体のお株を奪っています。ハードカーバーの本誌はほとんど使用しません。このようにマルチタスキングを最大限に利用するためコンピュータ本体のEDO RAMは128Mbと本機としては最大限度まで増

設しました。ついでながらコンピュータ本体付属のその他の周辺機器をご紹介しますと、通常のCD-ROMドライブ、1 GbのJazドライブ、フロッピーディスクドライブがあります。CD-ROMドライブはソフトウエアのインストール用としてのほか、作業中にCD（音楽）を再生BGM演奏をさせるのに使用、Jazドライブはもっぱら作成した文書ファイルのバックアップ用に使用しています。以上、MTの本題から外れたようですが、電子辞書は私の翻訳作業に欠かせない時間短縮用ツールなのであえてご紹介させていただきました。

6．MT翻訳者は少数派

既述のMTのメリットにもかかわらず、MTを使用する翻訳者は少数派のようです。限られた私の交友範囲でもMTを利用している翻訳者は皆無です。ベンダーによればMTソフトの購入者に翻訳者も若干名いるということですが、少数派であることには間違いないようです。少なくとも出版翻訳家ないし文芸翻訳家にはMT使用者は皆無に近いものと思われます。

私のようにMTの限界を承知の上で、そのメリットを利用する職業翻訳者が少ないのは、ひとつにはMTへの期待と現実の落差に幻滅、翻訳者としてのプライドから機械翻訳とか自動翻訳とかを利用することに抵抗を覚える、翻訳時間の短縮に寄与しない、MTソフトがパソコンにバンドルされて一般人が使うようになりプロとしての自尊心を傷つける、などいろいろ理由は考えられますが、現行MTに不可避の制約を認識した上そのメリットを最大限に利用する発想に転換すれば翻訳者のMT使用人口は飛躍的に増えるのではないかと予想されます。このためには、メーカー側から翻訳者を対象とする積極的な働きかけが必要かと思われます。この拙文がその一助になれば幸いです。

Webブラウザー専用のMTソフトが多く市販されるようになりましたが、使い勝手はどうなのでしょうか。使用体験者の感想を聞きたいと思います。少なくとも、こうしたソフトは翻訳者を対象としてはいないようです。翻訳者にとってのメリットが考えられないからです。

7．多言語の機械翻訳

MTと言えば、日本人関係者の間では通常、英日、日英を指すものと相場が決まっていますが、これからは中国語、韓国語、英語以外の欧州系言語を対象とするMTも重要度を増してくるものと思われます。特に英語と欧州系言語との近似性は日本語と英語のそれとはくらべものになりませんから、MTソフトの実用度はかなり高いものと考えられます。今年から単一通貨ユーロが現実のものとなった現在、将来にかけて欧州との関わり合いはますます深まり、このようなMTソフトに対する需要は著しく増大するものと思われます。

日本で発売されているGlobalink社の日本語対欧州系言語のMTソフトも英語経由と聞きました。日本語対欧州系言語MTソフトはまだ少ないようですが、これは日英以外の言語ペアの研究の遅れ、日本側のコーパス（用例集）の蓄積の少なさなどから考えられる結果とは思いますが、将来にかけての進化を期待したいものです。私の限られた予備調査では、Windows95Jでも上記の親ソフトGlobalinkのPower Translator（英語対独仏伊西各語）は作動するようです。近日中に実験してみたい希望を持っています。個人的な語学の背景になりますが、少年時代を中国で過ごし、大学の教養学部でフランス語、ラテン語、ドイツ語、中国語の授業を受講ないし聴講し、北欧系の企業に奉職し、海外出張先の言語を泥縄で学習した経歴から多言語に対する興味には人一倍強いものがあります。

8．機械翻訳者の夢

MTが今日の隆盛を見るに至った背景としてハードウエアの飛躍的進化があります。CPUの高速化、HDD容量やRAM容量の爆発的増大、CD-ROM等のメディアの出現などBravisデビュー時とは比較にならないほどハードウエア環境の充実があげられます。これからもハードウエア環境は同じペースあるいはそれ以上のペースで高度化することが予想されます。コーパスの蓄積および体系化には公的機関による本格的介入と速やかな着手と実行、メーカー間の協力が必要でしょうが、MTソフトのアルゴリズムの革新、AIの実用化とMTへの取り込みが進めば冗句とはならないMTによる翻訳精度の向上が現実のものとなるのではないでしょうか。

MTの主題からは外れますが、日頃私ども翻訳を生業とする者の名称の不統一について触れておきたいと思います。通常、出版書籍の翻訳を行えば翻訳家、技術文書の翻訳を行えば翻訳者、翻訳技術の審査機関から翻訳レベルの認定を受けて翻訳を行えば翻訳士となるようですが、用法上の便宜はともかく、翻訳者の私には何となく階級差別の匂いがします。将来は一律に翻訳者という呼び名に統一されることを望みます。その上で、出版翻訳者、産業（または技術）翻訳者、XX機関資格認定翻訳者のように表現すればよいのではないでしょうか。

＊株式会社アルク翻訳事典98佐藤洋一「プロの翻訳に必要な原文解釈力と表現力」p160／161

（森本健一、Eメール、jemtrans@earthlink.net）

事務局だより

委員の交替

次のとおり委員の交替がありましたので、お知らせいたします。

［運営委員会］

退任　石崎　　俊　　慶應義塾大学
就任　坂本　義行　　東京家政学院筑波女子大学

［編集委員会］

新任　熊野　　明　　株式会社　東芝

技術早分かり

コーパスからの動詞格フレームの獲得

奈良先端科学技術大学院大学 情報科学研究科　宇津呂武仁

1．はじめに

動詞の格フレーム（または下位範疇化フレーム）とは、動詞がどのような格をとり、また格要素としてどのような名詞をとるかを記述したもので、構文解析をはじめとして、自然言語処理の様々な局面で重要な役割を担っている情報です。例えば、市販の機械翻訳システムなどでは、格フレームに相当する情報を内蔵して、構文解析などの段階で、何らかの形で使用していると考えられます。格フレームの情報を一般利用者に対して公開しているものとしては、情報処理振興事業協会技術センター（IPA）によって開発されたIPAL動詞辞書［IPA87］があります。IPAL動詞辞書は、日本語の和語動詞861語について、格フレームを含む様々な情報を計算機上で記述したものです。例えば、「買う」という動詞の辞書記述においては、四つの語義が設定されており、各語義について以下のような格フレームが記述されています。

1	（語義）代金を払って何かを自分のものにする N 1 [hum/org]ガ N 2 [con/abs]ヲ買う
2	（語義）相手のために自分が代金を払って 品物を与える N 1 [hum/org]ガ N 2 [hum/ani]ニ N 3 [con/abs]ヲ買う
3	（語義）自分の行為や関連した事柄で、 他人に悪感情を起こさせる N 1 [hum/org]ガ（N 2 [abs]デ） （N 3 [hum/org/abs]ニ／カラ） N 4 [men]ヲ買う
4	（語義）値打を認める N 1 [hum/org]ガ N 2 [cha/men/act]ヲ買う

ここで、hum（人間）、org（組織・機関）、con（具体名詞）、abs（抽象名詞）、ani（動物）、men（精神）、cha（性質）、act（動作・作用）などは、名詞の意味素性を表しており、IPAL動詞辞書では18個程度の意味素性を用いて、それぞれの格がとりうる名詞の意味的な範囲を限定しています。

一方、機械翻訳における相手言語への訳し分けの観点から、大規模かつ詳細な動詞格フレームを記述した辞書として、NTTで開発された「日本語語彙体系」［池原97］があります。日本語の用言については、6,000語の用言の格フレーム14,000パターンを記述しており、格要素の名詞の意味的制約の記述においては、3,000種類の意味属性を用いています。また、一つの動詞の複数の語義を分類する際には、翻訳の際に英語訳が同じになるかどうかという観点で、語義が分類されています。

これらの格フレーム辞書は、いずれも、人間の辞書記述者がたくさんの例文を収集し、それらの例文などを分析した後、それぞれの動詞の用法を格フレームという形で書き下すことによって構築されたものです。一方、近年、大規模なコーパスが利用可能になるのに伴って、形態素・構文情報の付与されていないコーパスや、逆に形態素・構文情報の付与されているコーパスから、動詞の格フレームを自動もしくは半自動で獲得する手法の研究が行われています。残念ながら、現時点においては、辞書の規模の面において、人手で構築された格フレーム辞書を上回るものが自動的または半自動的に構築されたという事例はありませんが、一般に、自動もしくは半自動で格フレーム辞書を獲得するというアプローチには、以下のような利点があります。[1]

- 膨大な人手コストを軽減できる。
- 計算機により客観的に判断できる部分が多くなるので、構築された辞書中で互いに矛盾する部分を減らすことができる。
- 分野ごとに異なる格フレーム辞書が必要な場合

1コーパスを利用した自然言語処理手法については、近年盛んに研究が行われており、品詞付け・形態素解析、構文解析、語義曖昧性解消などの処理については、規則に基づく手法と比較しても、同程度かそれ以上の性能が得られているものも少なくありません（たとえば、[Charniak97、Ide98、言語処理学会99]）。

に、それぞれの分野のコーパスが用意できれば、低コストで分野依存の格フレーム辞書が構築できる。

- 格フレーム獲得の際の格フレームの評価基準を柔軟に調節することにより、獲得される格フレーム辞書が柔軟に調節できる可能性がある。

ここでは、コーパスから格フレームを獲得する手法に関する研究を以下のような観点で分類し、それぞれの研究について解説していきます。まず、元となるコーパスがあらかじめ解析済であるか否かによって、大きく以下の二つの分類することができます。

(1)未解析コーパスからの動詞格パターンの抽出
(2)構文解析済コーパスからの動詞格フレームの獲得

さらに、(2)については、格フレームの記述内容のうち、どの部分の獲得に焦点を当てるかによって、以下の研究に細かく分類することができます。

(2a) 格要素の意味的汎化レベルの学習
(2b) 格の間の依存関係の学習
(2c) 動詞の語義の分類
(2d) 下位範疇化優先度の学習

以下では、これらの分類にしたがって、解説をしていきます。

2. 未解析コーパスからの動詞格パターンの抽出

このタイプの研究［Brent93、Manning93、Briscoe97］では、対象としている言語は、そのほとんどが英語で、未解析の英語コーパスに対して、パターンマッチングあるいは簡単な解析を施すことにより、各動詞がとりうる格パターンを抽出します。ここで、獲得の対象となる格パターンと呼んでいるものは、前節のIPAL動詞辞書の例のように、動詞がとりうる格および格要素の名詞の意味的制約のうち、動詞がとりうる格のみを記述したものです。例えば、英語の場合、主語＋目的語、主語＋that節、主語＋不定詞句、主語＋目的語＋that節、主語＋目的語＋不定詞句、主語＋二重目的語、などの文型パターンがここでの格パターンに相当します。これらの研究においては、以下の手順にしたがって、動詞の格パターンが抽出されます。

2.1 未解析コーパスに対する言語解析

まず、未解析コーパスに対して、何らかの言語解析を行います。この際には、言語解析結果の曖昧性は考慮せずに比較的浅い処理のみを行い、もっとも信頼性が高く、解析誤りも少ないと考えられる結果のみを選定し、後の処理に利用します。たとえば、［Manning93］では、品詞付けの後、有限状態パーサーによる構文解析結果で曖昧性のない部分のみを同定し、また、[Briscoe97］では、同様に品詞付けの後、確率的LRパーサーの最尤解のみを利用します。一方、[Brent93］では、品詞付け・構文解析などは行わなくても確実に抽出できる格パターンのみを対象とし、代名詞・冠詞・助動詞・接続詞などのクローズ・クラス語の辞書とこれを利用したパターンのみを用いています。

2.2 パターンマッチングによる格パターンの検出

次に、あらかじめ抽出の対象とする格パターンについて、その形を記述しておき、言語解析結果との間でパターンマッチングを行うことにより、格パターンを検出します。[Brent93］では、品詞付け・構文解析などは行わなくても確実に抽出できる格パターンとして、主語＋目的語、主語＋that節、主語＋不定詞句、主語＋目的語＋that節、主語＋目的語＋不定詞句、主語＋二重目的語の6種類を対象としています。一方、[Manning93］では、19種類のテンプレートおよびそのうち前置詞を語彙化した格パターンを、[Briscoe97］では、160種類の格パターンを対象としています。

2.3 統計的フィルタリング

最後の処理として、[Brent93］は、抽出された格パターンがどの程度信頼できるかを、統計的検定法を用いて判定する方法を提案しました。また、［Manning93、Briscoe97］についても、この方法を採用しています。

2.4 評価

事例として挙げた各研究［Brent93、Manning93、Briscoe97］は、対象とする格パターンの数と種類が異なるので、単純な比較はできませんが、対象と

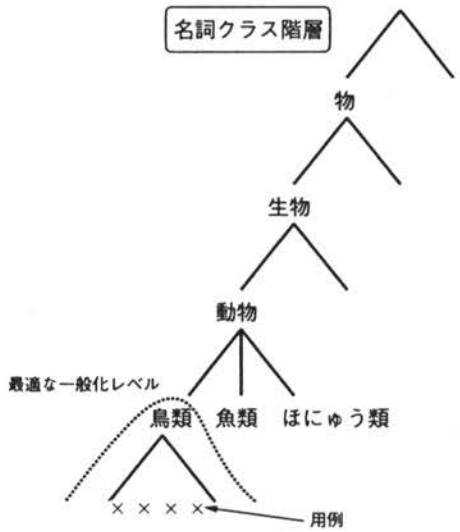

図1：動詞「飛ぶ」の主語となる名詞クラスの最適な一般化レベル

する格パターンを6種類に限定し、再現率は低くてもよいから信頼性の高い結果を目標とした［Brent 93］においては、元となったコーパスに対して、格パターンのタイプのレベルで60%の再現率・96%の適合率を達成しています。一方、［Briscoe97］では、対象とした160種類の格パターンに対して、43%の再現率・77%の適合率を達成しています。

3．格要素の意味的汎化レベルの学習

格フレームの記述においては、動詞がそれぞれの格の格要素としてどのような名詞をとりうるのかを指定する必要があり、通常、何らかの意味素性またはシソーラスなどの大規模な意味体系中の意味属性を用いて、格要素の名詞の意味的な範囲を指定します。一方、コーパスから格フレームを獲得する際には、格要素の名詞の事例を入力として、シソーラスなどの大規模な意味体系中で、格要素の名詞の意味的制約の最適な一般化レベルを決定する必要があります。これは、例えば、図1に示すように、動詞「飛ぶ」の主語の用例が多数与えられた時に、図中の名詞クラス階層において最適な一般化レベルを決定する問題に相当します。

この問題に対して、［Resnik92］は、［Church90］の相互情報量を用いた単語間共起の抽出手法を拡張して、動詞と格要素の名詞クラス間の共起を測定するAssociation Scoreを提案し、この尺度を用いて、動詞と目的語の名詞クラスの間の共起性を測定しました。ある動詞 v が格 r の格要素として名詞クラス c をとる場合の共起性を表すAssociation Score $A(c \mid v, r)$ は、v の格 r の格要素に c が出現する条件付き確率と格 r のもとでの v と c の相互情報量を全コーパスから求め、これらの積として以下のように定義されます。

$$A(c \mid v, r) \equiv P(c \mid v, r) \log \frac{P(c \mid v, r)}{P(c)}$$

ここで、第一項の条件付き確率が v と c の共起の一般性を表し、第二項の相互情報量が v と c の共起の強さを表します。表1に、無作為に抽出した15個の英語動詞について、Association Scoreの値が最も高かった目的語の名詞クラスを示します。ただし、名詞クラスとしては、WordNet［Miller95］の名詞クラスを用いており、それぞれ、〝単語、類義語の集合〟によって記述されています。

また、［Li98］は、木構造状のシソーラスのカット上の確率モデルを用いて格要素の名詞の意味的制約を表す確率モデルを定義し、記述長最小（MDL）原理［Rissanen89］に基づいて、動詞と格要素の名詞の事例からこの確率モデルを学習する手法を提案しました。例として、WordNet［Miller95］を木構造状に変換したシソーラス上で、動詞“*eat*”の目的語の意味的制約を表す確率モデルを学習した結果を、図2に示します。図中の数値は、それぞれの意味的制約が動詞“*eat*”の目的語となる確率値を表しています。［Li98］はまた、学習された意味的制約を英語の前置詞句付加の問題に応用し、その評価を行っています。その結果、［Resnik92］のAssociation Scoreによく似たSelectional Association［Resnik93］と比較して、［Li98］の確率モデルの方が高い精度を達成したと報告しています。

4．格の間の依存関係の学習

格フレームの記述においては、それぞれのフレームごとに、動詞がとりうる格の組合わせを指定しておく必要がありますが、コーパスからの格フレーム獲得においても、動詞と格要素の名詞の共起の事例から、動詞がどのような格の組合せをとりうるのかを決定しなければなりません。また、一般に、動詞

表1：［Resnik92］によって抽出された典型的な目的語のクラス

$A(v,c)$	動詞 v	目的語のクラス c	目的語の例
0.16	call	⟨someone, [person]⟩	(pname, man, ...)
0.30	catch	⟨looking_at, [look]⟩	(eye)
2.39	climb	⟨stair, [step]⟩	(step,stair)
1.15	close	⟨movable_barrier, [...]⟩	(door)
3.64	cook	⟨meal, [repast]⟩	(meal,supper,dinner)
0.27	draw	⟨cord, [cord]⟩	(line,thread,yarn)
1.76	eat	⟨nutrient, [food]⟩	(egg,cereal,meal,mussel,celery,chicken,...)
0.45	forget	⟨conception, [concept]⟩	(possibility,name,word,rule,order)
0.81	ignore	⟨question, [problem]⟩	(problem,question)
2.29	mix	⟨intoxicant, [alcohol]⟩	(martini,liquor)
0.26	move	⟨article_of_commerce, [...]⟩	(comb,arm,driver,switch,bit,furniture,...)
0.39	need	⟨helping, [aid]⟩	(assistance,hellp,support,service,...)
0.66	stop	⟨vehicle, [vehicle]⟩	(engine,wheel,car,tractor,machine,bus)
0.34	take	⟨spatial_property, [...]⟩	(position,attitude,place,shape,form,...)
0.45	work	⟨change_of_place, [...]⟩	(way,shift)

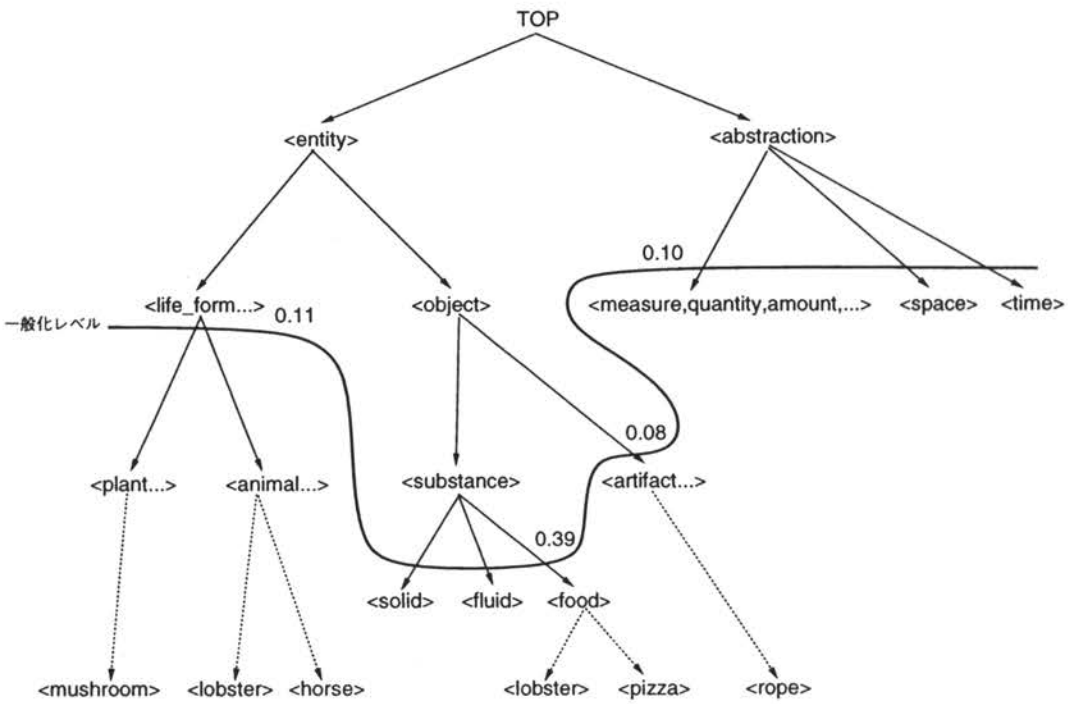

図2：［Li98］により学習された動詞“eat”の目的語の名詞クラスの最適な一般化レベル

表2：［Li96］により抽出された格の間の依存関係

動詞	依存関係にある格	
add	object	to
blame	object	for
buy	object	for
climb	object	from
compare	object	with
convert	object	to
defend	object	against
explain	object	to
file	object	against
focus	object	on
acquire	from	for
apply	for	to
boost	from	to
climb	from	to
fall	from	to
grow	from	to
improve	from	to
raise	from	to
sell	to	for
think	of	as

の格には、必ず共起しなければならない必須格とよばれる格と、必ずしも共起しなくてもよい任意格という格の区別があり、より厳密には、この区別も決定しなければなりません。また、二つ以上の格の間に、正または負の共起の依存関係（すなわち共起的または排他的な依存関係）がある場合には、この依存関係も学習する必要があります。

これらの問題に対して、[Li96]は、格が二つの場合に限定して格の間の依存関係を学習する手法を提案しています。[Li96]は、記述長最小（MDL）原理［Rissanen89］に基づいて、二つの確率変数の間の依存関係を判定する手法［Suzuki93］を用いることにより、二つの格が依存関係にあるか否かを判定しています。この方法で、英語の動詞の格のうち依存関係があると判定されたものの例を表2に示します。また、[Li96]は、これらの依存関係を、英語の前置詞句付加の問題に適用した結果、格の依存関係を考慮しない［Li98］の方法よりも高い精度を達成したと報告しています。ただし、[Li98]で依存関係が判定できたのは、格のレベルの間のみで、格要素の意味的制約の間では依存関係を検出することはできませんでした。

また、[春野95]は、複数の格およびそれらの格の格要素の意味的制約の記述を許した一般的な格フレームの獲得において、記述長最小（MDL）原理［Rissanen89］と同様の考え方に基づく手法を提案しています。この方法により、日本語動詞の格フレーム獲得を行った結果、複数の格に対して意味的制約を持つ格フレームが獲得されています。この方法は、格の依存関係を明示的に判定しているわけではありませんが、格フレームのコンパクトさの観点から、複数の格を持つ格フレームがより適切であると判定されていると考えることができます。

5．動詞の語義の分類

一般に、格フレームを記述する際には、動詞の語義の違いを考慮して、語義ごとに格フレームを記述する必要があるため、コーパスからの格フレーム獲得においても、自動的もしくは半自動的に動詞の語義の区別を行い、語義ごとに格フレームを獲得する必要があります。

この問題に対して、たとえば、[春野95]の方法では、格フレームのコンパクトさに基づいて、一つの動詞に対して複数の格フレームを獲得することができます。また、[宇津呂93、Utsuro96]では、単言語の情報だけでなく、二言語が対訳になったコーパスを利用することにより、単言語の動詞の多義性をよりきめ細かく類別する手法を提案しています。一般に、単言語の情報だけを利用する場合には、格要素の名詞がその言語のシソーラス中でどのような分布をしているかを調べることにより、動詞の複数の語義を類別することしかできません。しかし、二言語の情報を利用することができれば、格要素の名詞のシソーラス中での分布の微妙な違いを、他言語への翻訳結果の違いによって検出することができるという利点があります。

例えば、[Utsuro96]では、日英の対訳文を［Matsumoto93］の方法により構造的に照合した結果において、英語動詞の英語シソーラス中の意味クラスと、格要素の名詞の日本語シソーラス中の意味クラスの間で相互情報量を計算し、用例全体の（平均）相互情報量が最大になるように、日本語動詞の

表3：［Utsuro96］による他動詞／自動詞の用法の類別の例

日本語動詞	英語クラス (c_E)/日本語フレーム (f_J) (シソーラス中のレベル，例)	用例数	相互情報量
張る	*expensive*(Leaf)/が:値 (Leaf)	3	0.299
	Special Sensation(Leaf−3, *freeze*)/が:15130-11-10(Leaf, 氷)	3	0.237
	Acts(Leaf−2, *persist,stick to*)/を:13040(Level5, 強情)	7	0.459
ひく	Decrease(Leaf−1, *subside*)/が:151(Level3, 洪水)	2	0.109
	Results of Reasoning(Leaf−2, *catch,have*)/を:15860-11(Level6, 風邪)	26	0.421
ひらく	Intellect(Level1, *open*)/が:14(生産物)(Level2, 戸)	12	0.339
	hold(Leaf)/を:13510-1(Level6, 会合)	3	0.114
かなう	Completion(Leaf−1, *realize*)/が:1304(Level4, 願い)	8	0.460
	Quantity(Leaf−3, *equal*)/に:12000-3-10(Leaf, 彼)	8	0.504

語義分類を決定する方法を提案しています。例として、英語シソーラスとしてRogetのシソーラス［Roget11］を、日本語シソーラスとして分類語彙表［国研93］を用いた場合に、［Utsuro96］の方法により、日本語の多義動詞の他動詞／自動詞用法を類別した結果を表3に示します。表3から分かるように、英語側の動詞の意味クラスと日本語側の格および格要素の名詞の意味クラスの組の一つ一つが、多義動詞のそれぞれの語義に対応するという結果が得られています。また、実験の規模は小さいですが、［Utsuro96］の方法により動詞の用例の語義分類を行った結果と、人手で語義分類を行った結果を比較すると、複数の語義が誤って混在する割合は約6.7%で、また、語義数の比較においては、人手による分類の語義数と比べて、自動分類では約2.5倍の語義数になるいう結果が得られています。

6．下位範疇化優先度の学習

これまで、3、4、5節では、主として、格フレームの記述内容の様々な部分に焦点を当てて、それらの部分に関する情報をコーパスから獲得する手法について解説してきました。一方、自然言語処理の過程において格フレームを利用する場合、最も重要な利用法の一つとして、構文解析における統語的曖昧性解消での利用があります。したがって、コーパスから格フレームを獲得する際にも、構文解析における統語的曖昧性解消を主目的とし、この性能が最適化されるように格フレームを獲得するというアプローチを考えることができます。

例えば、これまで紹介した研究のうち、格要素の名詞の汎化レベルの学習を行う手法（3節）では、［Resnik93］のSelectional Associationや［Li98］の確率モデルを英語の前置詞句付加の問題に適用し、統語的曖昧性解消の性能を評価していました。しかし、これらの研究は、

- 統語的曖昧性解消の性能を最適化することを主目的として格フレームを獲得していない。
- 一つの格の格要素の意味的汎化レベルの学習にとどまっており、複数の格を持つ一般の格フレームを対象としていない

といった点で、格フレームを用いた統語的曖昧性解消という観点からは十分ではありませんでした。

これに対して、［Utsuro98］では、統語的曖昧性解消の性能を最適化することを主目的として、複数の格を持つ一般の格フレームを対象として、動詞が格要素をとる（「下位範疇化する」ともいいます）優先度（以下では、下位範疇化優先度と呼びます）を学習する手法を提案しています。［Utsuro98］の方法は、基本的には、［Li98］の方法と同様に、動詞が格要素の名詞を下位範疇化する確率を表す確率モデルを学習するというものです。この確率モデルの学習の際には、i）それぞれの格要素の名詞の意味的制約はどの程度の汎化レベルで記述すればよいか、ii）格の間にどのような依存関係があるか、の二つのことを決定する必要があります。これらの二つの情報はお互いに複雑に関係し合うので、［Utsuro98］では、これらの複雑に関係し合う情報を統一的に取り扱えるモデルとして、最大エントロピーモデル［Della

Pietra97、Berger96] を採用しています。また、モデル選択の際のモデル評価基準としては、1）記述長最小（MDL）原理 [Rissanen89] に基づく基準、2）統語的曖昧性解消の性能を直接評価する基準として、別途用意したデータセットを用いた下位範疇化テストに基づく基準、の二つの基準を併用しています。具体的には、まず、データスパース性の問題を考慮し、また、モデルの探索空間を制限するために、1）のMDL原理を用いてモデル探索を行い、次に、2）の下位範疇化テストによってモデル探索過程で得られた各モデルを評価し、最適なモデルを選択します。小規模な実験の結果では、格の依存関係を許さないモデル、最も汎化レベルの高いモデル、記述長最小モデルなどのどの対照モデルと比較しても、この手順で選択されたモデルの方が高い性能を達成しています。

7．おわりに

コーパスから格フレームを獲得する手法について、利用するコーパスの形態の違いや、格フレームの記述のうち獲得対象とする部分の違いを考慮して、それぞれの研究事例の紹介を行ってきました。最後に、これまでで触れることができなかった事項についていくつか触れたいと思います。

まず、（特に構文解析済コーパスから格フレームを獲得する際に）獲得された格フレームをどのように評価すればよいかについてですが、これまでの研究で行われてきた評価方法を整理すると、大きく次の二つに分けることができます。

a）人間用の辞書や人手で記述された格フレーム辞書などと比較して、どの程度の格フレームが獲得されたかを評価する。
b）獲得された格フレームを統語解析等に利用し、統語的曖昧性解消の性能で評価する。

このうち、a）の方は、いわば、人間が見て格フレームがどの程度理解しやすく直観と一致しているかという基準に相当するのに対し、逆にb）の方は、計算機が統語的曖昧性解消の処理を行う際に、そこで利用する格フレームがどの程度妥当であるかという基準に相当します。したがって、これらの二つの方法による評価結果が矛盾する可能性は十分にあり、例えば、統語的曖昧性解消においては、動詞の複数の語義をどのように分類するかということ（5節）はあまり問題にはなりません。また、この二つの評価方法のうち、b）の方は、統語的曖昧性解消の性能という客観的な尺度で評価できるのに対して、a）の方は、評価の際に用いる既存の辞書の影響を受けやすく、あまり客観的な評価とはいえません。このように、現在のところ、人間が見て格フレームがどの程度理解しやすく直観と一致しているかという基準で、格フレームの客観的評価を行うことは容易ではありません。

このような現状をふまえて、[中塚98] では、格フレームの完全な自動獲得を目指すのではなく、格フレーム獲得における計算機と人間の役割分担を行い、計算機による判断が困難な部分や計算機の判断が誤った部分について、人間が柔軟に判定・修正を行えるような支援環境を作成しています。この支援環境は、図3のようなインタフェースを持ち、人間は、格フレームの階層的クラスタリング・格要素の名詞の意味的制約のシソーラス・コーパス中の用例などを柔軟に閲覧しながら、各人の基準にしたがって、格フレームの取捨・選択を行うことができます。

また、これまで紹介してきた技術は、いずれもコーパスから格フレームを獲得するために開発された要素技術ですが、これらの技術は、人手で構築された既存の格フレーム辞書を客観的に検証するという目的にも適用できると考えられます。現在のところ、そのような研究事例はまだありませんが、今後はそのような方向の研究も行われると思われます。また、機械翻訳での利用を目的として、ここで紹介した技術と密接な関係にある手法によって翻訳知識を獲得するという研究もいくつかあります（例えば、[Almuallim94、田中95、北村96]）。なお、格フレーム獲得と翻訳知識獲得の大きな違いとしては、翻訳知識の場合には、相手言語への訳し分けの観点から翻訳知識が細かく分類されるという点が挙げられるでしょう。

参考文献

[Almuallim94]
Almuallim, H., Akiba, Y., Yamazaki, T. and Kaneda, S. : Induction of Japanese-English Translation Rules from Ambiguous Examples and a Large Semantic Hierarchy,

図３：コーパスからの格フレーム半自動獲得のための支援環境［中塚98］

人工知能学会誌, Vol. 9, No. 5, pp. 730－740(1994)。

［Berger96］

Berger, A. L., Della Pietra, S. A. and Della Pietra, V. J. : A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics,* Vol. 22, No. 1, pp. 39－71(1996)。

［Brent93］

Brent, M. R. : From Grammar to Lexicon : Unsupervised Learning of Lexical Syntax, *Computational Linguistics,* Vol. 19, No. 2, pp. 243－262(1993)。

［Briscoe97］

Briscoe, T. and Carroll, J. : Automatic Extraction of Subcategorization from Corpora, *Proceedings of the 5 th ANLP,* pp. 356－363(1997)。

［Charniak97］

Charniak, E. : Statistical Techniques for Natural Language Processing, *AI Magazine,* Vol. 18, No. 4, pp. 33－43(1997)。

［Church90］

Church, K. W. and Hanks, P. : Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics,* Vol. 16, No. 1, pp. 22－29(1990)。

［Della Pietra97］

Della Pietra, S., Della Pietra, V. and Lafferty, J. : Inducing Features of Random Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 19, No. 4, pp. 380－393(1997)。

［言語処理学会99］

言語処理学会第５回年次大会併設ワークショップ「構文解析－現状の分析と今後の展望－」論文集(1999)。

［春野95］

春野雅彦：最小汎化とオッカムの原理を用いた動詞格フレーム学習、情報処理学会研究報告、Vol.95, No. 69 (95－NL－108), pp. 29－36(1995)。

［Ide98］

Ide, N. and Veronis, J. : Introduction to the Special Issue on Word Sense Disambiguation : The State of the *Art,*

Computational Linguistics, Vol. 24, No. 1, pp. 1－40 (1998)。

[池原97]
池原悟、宮崎正弘、白井諭、横尾昭男、中岩浩巳、小倉健太郎、大山芳史、林良彦（編）：日本語語彙大系、岩波書店（1997)。

[IPA87]
情報処理振興事業協会技術センター：計算機用日本語基本動詞辞書 IPAL(Basic Verbs)説明書(1987)。

[北村96]
北村美穂子、松本裕治：対訳コーパスを利用した翻訳規則の自動獲得、情報処理学会論文誌、Vol. 37, No. 6, pp. 1030－1040(1996)。

[国研93]
国立国語研究所：分類語彙表、秀英出版（1964、1993)。

[Li96]
Li, H. and Abe, N.: Learning Dependencies between Case Frame Slots, *Proceedings of the 16th COLING,* pp. 10－15(1996)。

[Li98]
Li, H. and Abe, N.: Generalizing Case Frames Using a Thesaurus and the MDL Principle, *Computational Linguistics,* Vol. 24, No. 2, pp. 217－244(1998)。

[Manning93]
Manning, C. D.: Automatic Acquisition of a Large Subcategorization Dictionary from Corpora, *Proceedings of the 31st Annual Meeting of ACL,* pp. 235－242(1993)。

[Matsumoto93]
Matsumoto, Y., Ishimoto, H. and Utsuro, T.: Structural Matching of Parallel Texts, *Proceedings of the 31st Annual Meeting of ACL,* pp. 23－30(1993)。

[Miller95]
Miller, G. A.: WordNet: A Lexical Database for English, *Communications of the ACM,* Vol. 38, No. 11, pp. 39－41(1995)。

[中塚98]
中塚幸毅、宇津呂武仁、松本裕治：コーパスからの格フレーム半自動獲得のための支援環境の構築、言語処理学会第4回年次大会論文集、pp. 442－445、言語処理学会（1998)。

[Resnik92]
Resnik, P.: WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery, *Proceedings of the AAAI-92 Workshop on Statistically-Based Natural Language Programming Techniques,* pp. 48－56(1992)。

[Resnik93]
Resnik, P.: Semantic Classes and Syntactic Ambiguity, *Proceedings of the Human Language Technology Workshop,* pp. 278－283(1993)。

[Rissanen89]
Rissanen, J.: *Stochastic Complexity in Statistical Inquiry,* Vol. 15 of *Series in Computer Science,* World Scientific Publishing Company(1989)。

[Roget11]
Roget, S. R.: *Roget's Thesaurus,* Crowell Co.(1911)。

[Suzuki93]
Suzuki, J.: A Construction of Bayesian Networks from Databases Based on an MDL Principle, *Proceedings of Uncertainty in AI,* pp. 266－273(1993)。

[田中95]
田中秀輝：動詞訳語選択のための「格フレーム木」の統計的な学習、自然言語処理、Vol. 2、No. 3, pp. 49－72(1995)。

[宇津呂93]
宇津呂武仁、松本裕治、長尾眞：二言語対訳コーパ

スからの動詞の格フレーム獲得、情報処理学会論文誌、Vol. 34, No. 5, pp. 913－924(1993)。

[Utsuro96]
Utsuro, T.: Sense Classification of Verbal Polysemy based-on Bilingual Class/Class Association, *Proceedings of the 16th COLING*, pp. 968－973 (1996)。

[Utsuro98]
Utsuro, T., Miyata, T. and Matsumoto, Y.: General-to-Specific Model Selection for Subcategorization Preference, *Proceedings of the 17th COLING and the 36th Annual Meeting of ACL*, pp. 1314－1320(1998)。

事務局だより

Eメールアドレスの変更

AAMT事務局のEメールアドレスが、4月1日からいずれもor → neに変更になりましたので、訂正くださいますようお願いいたします。

(旧)		(新)
JWD 05467@nifty.or.jp	→	JDW 05467@nifty.ne.jp
aamt 0001@infotokyo.or.jp	→	aamt 0001@infotokyo.ne.jp
aamt 0002@infotokyo.or.jp	→	aamt 0002@infotokyo.ne.jp
aamt 0003@infotokyo.or.jp	→	aamt 0003@infotokyo.ne.jp

Java言語による機械翻訳システムの実現

沖電気工業株式会社　研究開発本部
関西総合研究所　自然言語プロジェクト　永田　淳次

1　はじめに

機械翻訳システムを設計するにあたり常に重要視しているのは翻訳品質である。現在、より品質の高い翻訳結果を得るために、新たな機械翻訳システムの研究開発を行っており、Java言語で記述することを検討した。

機械翻訳システムの実現において、その実装に使用するプログラミング言語によって実現する機能が変化することはない。どのプログラミング言語を選択するかが重要ではないという考え方がある。沖電気が開発した機械翻訳システムPENSEEは、C言語で記述されている。C言語もJava言語もプログラミング言語の一種であり、その実現の面だけをとらえると大差はないが、その潜在能力の違いは大きなものがある。我々は、Java言語で機械翻訳システムの実現を試みておりその差の大きさを実感している。

本稿では、新たに研究開発している機械翻訳システムの設計思想と、Java言語を選択した理由について述べる。Java言語の採用は研究開発者にとってメリットがあるだけでなく、利用者にとっても大きな利益があることを理解していただければ幸いである。

2　PENSEE

沖グループは、手軽に、安く、使い易くをコンセプトとして94年5月に機械翻訳システムWindows版PENSEEをリリース［1］した。パーソナルユースを目的とした本システムはPC市場の拡大とともにユーザに受け入れられた。このシステム以外にもDTPシステムと融合したPENSEE-GV［2］、「Let's Enjoy Internet Surfing.」を設計コンセプトとし、Webページのテキストを翻訳するPENSEE for Internet［3］を商品化し、提供してきた。

PENSEEは、このように多種多様なシステム上で動作している。特定の機種（OSやコンピュータ）には依存しない形で設計されている。PENSEEは、利用者との間をとるMMI部、翻訳知識が格納されている文法・辞書部、そして翻訳知識を解釈し実行する翻訳エンジンの三要素で構成されている。

文法・辞書は、翻訳結果の品質を左右する重要な部分である。より良い品質を得るためには十分な文法・辞書が必要であるが、短時間で大量に用意することは難しい。そのため、効率良く文法・辞書を開発できること、またコンピュータの知識が無くても開発できることを目的として、文法・辞書の部分とそれを解釈し実行する部分とに分けた。文法・辞書の追加を簡単に行うことができ、品質向上の簡単化が図れる。これにより翻訳知識を豊富に持つ人が、効率良く、文法・辞書を開発しており、PENSEEの翻訳品質を少ない労力で高いレベルにすることに貢献している。この実行部分を翻訳エンジンと呼んでいる。

機種に依存する部分はMMI部に集約し、新たなシステムで動作させるときは、MMI部のみを開発した。それに対し、文法・辞書および翻訳エンジンは、どのシステムも同一であり、変更は加えられていない。どのPENSEEも同じ翻訳エンジンで動作している。同じ版であれば、同じ文法・辞書であり、同じ翻訳エンジンである。

翻訳エンジンは、小さなリソース、小さなメモリー、高い翻訳品質を目指して設計され、C言語で記述されている。翻訳エンジンはどのプラットフォームでも動作するよう、C言語の標準的な仕様でのみ記述するようにしている。最新の技術は、時として特定の機種に依存せざるをえないことがあるため、安定した技術を使うようにして、どの機種にも対応できるように細心の注意を払ってきた。

3　用例主導型機械翻訳エンジン

インターネットの拡大、PCの普及につれ、機械翻訳への期待も大きくなってきた。機械翻訳システムがおかれた環境も変化しつつある［4］。機械翻訳システムは限られた利用者だけが使用するのではな

く、より多くの人たちの利用が始まっている。完全な翻訳結果を得ることはできないが限られた利用範囲では十分な能力を発揮する道具であるという認識が定着し始めている。

これは、より現実的なシステムの実現の期待が高まっていることを意味する。トランスレーションメモリーのような翻訳例を格納し出力する機能を重視した翻訳をしない翻訳者支援システムの要求も高まり、文法・辞書もユーザでカスタマイズ［5］したいという希望もより強くなってきている。また、コンピュータの能力の向上、先端技術の開発により、それらの実現の可能性も高くなってきた。

“機械翻訳の現状と今後の動向”［4］でも解説されているように、機械翻訳システムの翻訳品質の向上には、大量の翻訳知識が必要である。機械翻訳システムの実現では、多量でかつ高品質な翻訳知識をシステム内でどう表現し実現するか、また、どのようにして最適な翻訳知識を選択し翻訳過程に適用するかが重要である。ルールベースの機械翻訳システムでは、翻訳知識であるルールの追加コストの増大、ルール間の衝突の制御に課題を持ち、将来翻訳品質向上が飽和してしまう可能性がある。PENSEE も翻訳知識をルールで表現するシステムであり、同様のことが予想された。そのブレークスルー技術として、知識獲得の研究開発［6］,［7］を行っている。本成果を利用すると翻訳知識すなわち文法や辞書を自動的に生成することができる。大量のコーパスから自動獲得した知識には確率量も含まれており、高品質の翻訳結果を出力する情報として利用可能である。また、テンプレート化された対訳例を翻訳知識として利用する事例ベース方式の研究開発も行っている。これは文法と辞書の垣根をなくしている。文法も対訳例として表現し、辞書も対訳例として表現している。文法も辞書も同じ形式で表現されているため、追加、削除、さらに衝突制御も簡単である。これは、より簡単に翻訳品質の向上を行うことができ、知識獲得した翻訳知識も簡単に組込むことができる。日英という対訳だけでなく日英独という対訳も処理することができるため、他の言語対の翻訳へ拡張も容易である。

これらの研究の成果である新たな技術は、既存の翻訳エンジンとはなじまない。PENSEE の翻訳エンジンは、木構造の処理を中核としている。ルールベース方式では、述部をルートにし、主部や目的部を枝にして処理するため、木構造を簡単にかつ高速に処理できることが重要視されてきた。これにパターン処理を初めとする新たな枠組を加えることは難しいため、新たに翻訳エンジンを検討することとした。これを用例主導型機械翻訳エンジンと呼び、C 言語ではなく Java 言語で記述し実装をしている［8］。

4 Java と機械翻訳システム

用例主導型機械翻訳エンジンの設計では、Java 言語を採用した。Java 言語を採用する一番大きな理由は、オブジェクトにある。我々は機械翻訳システムをプログラムとしてとらえるのではなく、オブジェクトとして考えている。手順の集まりとして提供するのではなく、オブジェクトとして提供したいと考えている。機械翻訳システムは、それ単独のシステムではない。機械翻訳システムの利用者は、情報の収集のために翻訳したり、ワールドワイドに流通するドキュメントを作成するために、また使用言語が異なる人達とのコミュニケーションのために翻訳を行っている。翻訳は目的ではなく手段である。機械翻訳システムの利用も手段であり、目的を果たすシステムとの連動は当然のことであろう。現在も、Web ページの翻訳（HTML 翻訳）や文書処理（例えばワード、一太郎）との親和性を差別化技術として提供するシステムもある。機械翻訳システム単独で利用されるより、他のシステムと連動したり、他のシステムから呼び出されたりするのが当たり前となる。そうなるとプログラムとしてではなく、オブジェクトとしての完成度が求められる。

オブジェクト指向言語はすでに数多く提案されている。オブジェクトを指向するのであれば、Java でなくてもいいかもしれない。我々は機械翻訳システムをオブジェクトに向かって設計しようとしているのではなく、オブジェクトとして捉えている。オブジェクトが大前提である。これをコンセプトとしていてかつ身近に使用できるプログラミング言語としては現在のところ Java 言語しかない。

Java 言語の採用は数多くのメリットをもたらす。システム全部もオブジェクトであり、システムの一部もオブジェクトである。オブジェクト単位で簡単に変更ができるため、システムのバージョンアップが容易である。Java 言語で記述したアプリケーシ

ョンは、プラットフォームに依存しないため、どのコンピュータでも動作可能である。さらにネットワーク上での分散処理にも簡単に対応できる。

また、システムの構成部品もオブジェクトであり、APIが定義されている。例えば形態素解析APIがある。これによって機械翻訳システムだけでなく自然言語処理を行うアプリケーションシステムから形態素解析機能を呼び出すことが可能となる。

さらに、翻訳エンジンもAPIとして定義される。これは、エンジンの入れ替えができるため、話題に応じて利用者の好みの機械翻訳システムを利用することができることを意味する。

5　おわりに

機械翻訳システムは、それ単独で利用するシステムというより、ネットサーフィンやワープロや、より大規模なシステムの一部として動作すべきであると考えている。機械翻訳処理技術のみならず多くの機能をオブジェクトとして定義し提供していくことが重要である。

過去、ソフトウエアは、完結したシステムとして提供されてきた。しかし、類似のソフトウエアを個別に開発したり、同一の機能に対してそれぞれMMIを開発する時代ではない。特徴を持ったエンジンが複数存在し、共存する時代になると考えている。

技術動向調査委員会が提案した電子化言語情報の共有化［9］と同様、自然言語処理用APIの定義とその標準化も重要なテーマとなるであろう。その時代にはJavaに代表されるフレームワークのしっかりしたオブジェクト指向言語の果たす役割は大きいと考えている。

参考文献

1．石川："Windows版PENSEEの商品化と市場への取り組み", AAMT Journal No. 15pp 8 －10(1996)

2．沖電気工業(株)："PENSEE-GV", AAMT Journal No. 3pp20－21(1993)

3．沖ソフトウエア（株）："PENSEE for Internet", AAMT Journal No. 13pp22－23(1996)

4．山端，大倉，亀井，松井："機械翻訳の現状と今後の動向", AAMT Journal No. 24pp 3 －11(1998)

5．仁井："ユーザ主導の機械翻訳", AAMT Journal No. 7 pp 2 － 9 (1994)

6．北村，松本：対訳コーパスを利用した翻訳規則の自動獲得，情報処理学会論文誌，37（6），pp. 1030－1040(1996)

7．Shimohata, Sugio, Nagata : "Retrieving Collocations by Co-occurrences and Word Order Constrains", 35th annual meeting of the Association for Computational Linguistics, pp476－481(1997)

8．村　田：Java Conference Winter Seminar97in Yokohama(1997)

9．技術動向調査委員会："機械翻訳用電子化言語情報の共有に向けて", AAMT Journal No. 18(1997)

国際会議参加報告

ICCPOL '99 報告

東京大学大学院 理学系研究科 情報科学専攻 辻井研究室 野畑 周

3月24日から26日の3日間、徳島で開催されたICCPOL'99（the 18th International Conference on Computer Processing of Oriental Languages）に参加したので報告する。

1．概要

ICCPOL'99では、徳島大学を会場として110件（口頭発表67件、ポスター発表37件、デモンストレーション6件）の発表が行われた。国別では、開催国である日本が最も多く53件、以下韓国から41件、中国（香港を含む）から7件、台湾から6件の発表があり、アジア地域以外からも3件の発表があった。

1日目の午前に辻井潤一教授（東大）によって"NLP Techniques for Information Extraction and Information Retrieval"と題して基調講演が行われ、ひき続いてJong-Hyeok Lee教授（Pohang University of Science and Technology,Korea）による韓日・日韓翻訳システムの紹介・デモンストレーションが行われた。

その後口頭発表は2会場で並行して行われ、1日目の午後には4セッションの発表があった。2日目は6セッションの口頭発表が行われた後、4会場に分かれてデモンストレーションとポスター発表が行われた。3日目は8セッションの口頭発表があり、3日間を通じて朝早くから夕方まで精力的に発表と議論が行われた。

論文発表以外にも特色ある催しが用意されていた。1日目の夕方にはバンケットが開催され、地元の阿波踊りグループによって阿波踊りが披露されたあと、簡単な指導のもとに参加者も一緒に阿波踊りを踊った。2日目には、徳島に本社を置くソフトウェア開発会社ジャストシステムの見学ツアーが催された。

2．論文発表

セッションとして立てられた分野は自然言語処理全般にわたるものであった。全18セッションのうちCharacter Recognition, Information Extraction, Morphological Analysisがそれぞれ2セッションを占めた。各セッション名は以下の通り：

- Character Recognition（I,II）
- Corpora and Tagging
- Error Recovery
- General Applications
- Image and Cross-Language Retrieval
- Information Extraction（I,II）
- Machine Translation
- Morphological Analysis（I,II）
- Natural Language Interface
- Natural Language Understanding
- Parsing
- Semantic and Disambiguation
- Text Categorization
- Retrieval Algorithms
- Voice Processing

デモンストレーション・ポスター発表は特に分野ごとに分けられていなかったが、同様に多岐のテーマにわたって研究が発表されていた。

以下、参加した口頭発表のセッションについて簡単にふれたい。

Information Extraction（I,II）

2セッションで7件の発表があった。定型的なデータを出力する本来の情報抽出に関する発表が1件あり、新製品情報の抽出結果を実際にWWW上で見ることができるようになっていた。他には情報抽出や機械翻訳システムのためのパターンや単語クラスタの自動獲得に関する発表が3件、特定の文字列の抽出に関する発表が2件あり、情報抽出システムのための知識データの自動獲得に関する関心が高ま

っていることを感じさせた。文章要約に関する研究では、クローズドキャプションのために情報量を変えずに文章を短縮するための手法が提案されていた。

Machine Translation

中韓、日韓、韓日、日中の4つの機械翻訳システムについて発表があった。中韓翻訳システムにおいてはパターンマッチングと統計的手法を組み合わせた機械翻訳システムの枠組が提案されたが、その他はシステム自体というよりも各システムの要素技術に関する発表が主であった。日韓翻訳システムについては、日本語の「て」の訳し分けに関する曖昧性解消の手法が提案・評価された。韓日翻訳システムについては、韓国語名詞の多義性解消をコーパスから教師なし学習によって行う実験が報告された。日中翻訳システムでは、トレーニングコーパスからパターンを獲得し、それらを用いて翻訳結果を後処理することによって翻訳の精度を向上する手法が提案・評価された。

Image and Cross-Language Retrieval

2件の発表があった。1件目は、天気図の画像から中国語の説明文を自動的に生成するという研究であった。得られた説明文を画像に対応づけて保存しておき、情報検索に利用しようとするものである。

2件目は、英日のCross-Language Information Retrievalにおけるqueryの翻訳手法についてであった。queryとして与えられた英語の複合語を、各単語の辞書引き結果とその結合確率によって日本語に翻訳し、情報検索を行うというものである。

Text Categorization

4件の発表があった。1件は情報検索の精度向上に関する研究であり、情報検索をした結果得られた記事についてクラスタリングを行ない、それらのクラスタをqueryに対してさらにランクづけすることで検索の精度を向上させる研究が発表された。テキスト分類については、シソーラスツールを用いてテキスト分類を行う研究が発表された。トレーニングコーパスから、各クラスタに対応する概念ベクトルをシソーラスツールによって作成し、テストコーパスの各テキストに対応する概念ベクトルをそれらと比較して分類を行うというものである。ウェブページの分類についての研究では、分類するための情報の質を良くする手法が提案・評価された。キーワードなどの素性とHTMLタグによって示される文書構造とに対してMemory Based Reasoningを適用し、ノイズを減少させてテキスト分類の精度を向上させるというものである。また、クラスタに分類されたテキストから、各クラスタの特徴を示す素性を自動的に取り出す研究も発表された。

Error Recovery

3件の発表があった。3件とも異なる対象についてエラー回復を行う手法を提案していた。まず自然言語インタフェース上で与えられた入力文を対象として、文の誤りを修正するシステムの速度を向上させる手法が発表された。また新聞記事の見出しを対象とし、格助詞の欠落を補う手法が発表された。OCRの文字認識結果を言語的知識を用いて修正する研究では、大規模なコーパスを基に誤って認識されている可能性が高い文字を認識し、それらを修正するための単語の候補を生成し、形態素情報や統計によって最適な修正候補を選択するという手法が発表された。

Retrieval Algorithms

4件の発表があったが、うち3件は情報検索に用いるためのデータを効率良く生成するための研究であった。トライ構造にキーワードを2値ベクトル化して格納していくことで情報検索に有効なデータベースを効率良く作成する手法や、リンクを導入したトライ構造によってデータベースをよりコンパクトに作成する手法、またn-gramを効率良く構成するための手法が提案・評価された。一方、情報検索に限らず、自然言語処理一般についてのツールキットの枠組が提案された。

Morphological Analysis (I,II)

2セッションで8件の発表があった。形態素解析

システム自体の精度向上に関する研究が2件あった。一方は形態素解析の品詞付けの誤りや切り分け誤りを後処理によって修正するという手法であり、他方はHMMに語彙情報を付加して拡張していく手法である。また、辞書のデータ構造を改良して形態素解析を高速化する研究も2件あった。単語の切り分け手法に絞った研究は2件あり、遺伝的アルゴリズムを用いて中国語の単語切り分けを行う手法や、PPM*アルゴリズムを用いて日本語の単語切り分けを行う手法が提案・評価された。日本語仮名漢字変換システムについては、連語に関するトレーニングコーパスを用いて精度を向上させる研究が発表された。また韓国語において、特定の分野に対して語形情報・統語的情報のルールを統合して形態素解析を行う枠組が提案された。

3．全体的な印象

アジア地域の言語に関する会議ということでひとつ気になった点は、発表において対象とする言語の例を出すときには注意しなければならない、という点である。聴衆にとっては、英語でも母国語でもない例文を見せられると、いくら口頭では英語で説明されてもその発表自体を理解しようとする気力が減少してしまう。発表する際には、できるだけ対象言語に依存しない例を用いて、多くの聴衆に理解してもらう工夫をする必要を感じた。

総じて運営は潤滑に行われ、運営委員会の努力がしのばれた。また、どちらかというと学生の発表者が多かったことが印象に残った。そのためか質疑応答もなごやかな雰囲気で行われることが多かった。また、休憩の間も盛んに議論をされている方々が見受けられ、できるだけこの場を有意義なものにしようという熱心さが感じとれた。

日本語・韓国語・中国語は、計算機を用いて自然言語処理を行う上で、欧米の言語とは異なる様々な問題を共通して抱えている。このような状況において、ICCPOLのような会議の重要性は今後も高まってくるものと予想される。

論文名・発表者名の正確な情報は、主催された徳島大学のホームページ

http : //iccpol99.is.tokushima-u.ac.jp/main/

に記載されているので、興味を持たれた方はご覧いただきたい。

事務局だより

AAMT事務局スタッフの交替

AAMT職員、木谷　恵さんが3月31日付で退職し、後任として高田佳代子（たかだかよこ）さんが3月23日から勤務しておりますので、お知らせいたします。

AAMTジャーナルの編集も担当いたしますので、ぜひ、ご指導ご支援くださいますようよろしくお願いいたします。

学会参加報告

言語処理学会第５回年次大会

富士通研究所　富士　秀

〈機械翻訳と学会活動〉

機械翻訳が製品化されている現在、機械翻訳に関する研究開発活動は全て公開されているわけではありません。しかしその一部は今回の言語処理学会のような学会活動を通じて発表されています。

なお、機械翻訳に関する学会発表は、大学、企業、研究機関によるものが多く、工学的な立場からのものがほとんどといって良いでしょう。

〈言語処理学会とは？〉

以前は、国内における機械翻訳や多言語処理に関する研究発表は、いくつかの主要学会の研究会などで行われてきました。しかし、1994年に言語処理専門の学会である言語処理学会が設立されて以来、多くの言語処理研究者が言語処理学会に研究発表の場を求める傾向が進みました。

言語処理学会の年次大会は今年が５回目となります。今回の会場は東京都調布市の電気通信大学で、日程は1999年３月15日がチュートリアル、16日～18日が本会議、19日が併設ワークショップでした。

〈発表内容〉

機械翻訳に関係すると思われる発表を織り交ぜながら、筆者が個人的に興味を持った技術ついてお話しします。

１．コーパス（大量文書群）からの対訳知識獲得

今回の大会では、特に、対訳でない日本語および英語コーパスから対訳知識を取り出す技術が目立っていたように思います。

数年前から、自然言語処理においてコーパスを利用した研究が増え、機械翻訳でもコーパス中の対訳知識を利用するような研究が紹介され始めました。当初は、英語コーパスが日本語コーパスの完全な翻訳であったり、逆に日本語コーパスが英語コーパスの完全な翻訳であるような、いわゆる対訳コーパスが処理の対象とされてきました。

しかし、現状では完全な形の対訳コーパスはあまり存在しないことから、対訳ではないがある程度内容的に対応のあるような日本語と英語のコーパスを対象とするような研究が増えてきました。また、コーパスからいきなり複雑な対訳情報を取り出す研究よりも、単語や複合語などのより基本的な対訳情報を対象とする研究が増えているようです。

発表番号Ａ３－４の「新聞記事における日英対訳コーパスの自動構築」では、日経の英日記事コーパスに対して記事中の数値情報などを手掛かりとして大まかな記事の対応付けを行い、対応の付いた記事対から対訳語句を抽出しています。また、自動的に対応付けた記事対から、対訳文を抽出する研究も行っています。

発表番号Ｐ２－３の「対訳コーパスからの訳語対抽出における辞書情報の利用について」では、学術論文抄録データベースを対象として、対訳専門用語を抽出する実験を行っています。研究では、「専門用語の対訳要素対列は、その周辺に、辞書中に存在するような対訳要素対を持つ場合が多い」という仮定のもとに分析を行い、分析データをもとに対訳語句の自動抽出を行っています。

発表番号Ｑ１－２の「言語横断検索システムQuest」は、機械翻訳に関する発表ではありませんが、言語横断検索用に対訳語句の判定を行っています。日英の学術論文データベースを処理の対象としています。複合語に対する訳語を確率モデルを用いて推定したり、翻字（カタカナとアルファベットの対応処理）を使って対訳語句の対応付けを行っています。

発表番号Ａ１－５の「対訳関係のないコーパスからの複合名詞対訳の獲得」では、対訳関係のない２言語のコーパスから抽出した単語列を、辞書や意味属性を用いて対応付けを行い、複合名詞対訳を収集する方法を示しています。

２．内容伝達型の情報と概要提示型の情報

文書分類や文書要約のセッションでは、内容伝達

型（informative）と概要提示型（indicative）の区別が話題になっていましたが、おそらく機械翻訳にも適用できる概念ではないかと思われます。

例えば文書要約で考えてみると、読者が要約文を読んでその文書が必要かどうか判断し、必要なら原文を読むというような使い方が概要提示型となります。これに対して、読者が原文を一切読まずに要約文のみから内容を判断するような使い方が内容伝達型となります。

機械翻訳でも、情報が読者にとって必要なものであるかを判断するだけなら概要提示型で十分であり、訳文の内容をしっかり把握しなければならない場合には、内容伝達型が必要になります。それぞれについて、別の形のインタフェースや評価方法が必要になるのではないでしょうか。

3．多言語検索

多言語検索（または言語横断検索）は、最近注目されており、今回も何件かの発表がありました。特に対訳キーワードの収集やキーワード変換に関しては、機械翻訳との共通部分が多いようです。

〈発表タイトル〉

機械翻訳および多言語検索に関連すると思われる発表を、筆者の独断でピックアップしました。

A1－1「英日機械翻訳システムにおける派生語推定とその訳語付与の改良」
A1－2「日本語記事の重要情報に基づく英文ヘッドラインの生成」
A1－3「機械翻訳における自動校正と日中翻訳への適用」
A1－4「多言語間翻訳情報共有システムの開発」
A1－5「対訳関係のないコーパスからの複合名詞対訳の獲得」
A1－6「スペル補完機能を組み込んだ英文作成支援」
A1－7「スクリプトを用いたハイブリッド翻訳処理」
A1－8「情態副詞の翻訳」
A7－1「原言語例文とその対訳に関する質問に基づく意味構造変換ルールの獲得」
A7－2 "Problems in Translating Modal Expressions"
A7－3「用例利用型翻訳に適した対話用例の自動作成について」
A7－4「目的言語の単語共起情報を利用した訳語選択と未知語の訳出」
A7－5「クリンゴン語翻訳システムを目指して～解析編～」
A7－6「中間概念としての認識構造の記述形式」
A7－7「日本語・ウイグル語機械翻訳における単語接続関係を用いたウイグル語文の生成方法」
A7－8「英字新聞記事見出し翻訳の自動前編集による改善」
A3－4「新聞記事における日英対応コーパスの自動構築」
A3－5「マルチリンガル・コーパスの作成」
P2－3「対訳コーパスからの訳語対抽出における辞書情報の利用について」
A5－2「情報検索の類似尺度を用いた検索要求文の単語分割」
A5－4「分類標数の相互参照に基づく多言語書誌データ検索システム」
Q1－2「言語横断検索システム Quest」
Q1－7「翻訳情報の提示によるクロスリンガル情報検索結果からの文書選択」

イベント

Language Expo '99 TOKYO 報告

1999年3月26日(金)～28日(日)新宿NSイベントホール

富士通研究所　大倉清司

1．概要

このExpoは、主に英語学習、教育、マルチメディア、留学、キャリアアップなどの情報を提供するもので、今年で4回目である。英会話スクールをはじめ、翻訳会社、パソコンを使った英語学習教材ソフトの会社など、多数出展していた。会場内は、「英語会話スクールゾーン」「キャリアアップゾーン」など、9つのゾーンに分かれている。Expoの来場者はさまざまで、大学生くらいから主婦など、英語を勉強しようという人が中心であるようだった。

今回注目したのは、「マルチメディア教育ゾーン」。マイクロソフトが4月に英語学習用ソフトを発売する他、IBMのViaVoiceを使った英語教育ソフト、英語の読み書きを支援するソフトなどを見学した。ここでは、個人的に興味を持ったものに限定して報告する。

2．辞書ソフト

2.1「English　Navigator」(オムロンソフトウェア株式会社) 定価14,800円

http : //www.omronsoft.co.jp/

英語の読み書きを強力にサポートするソフト。英語の読みについては、英語をマウスでクリックする(またはなぞる)と辞書引きしてくれるというもの。また、ブラウザなどで英語を読むときに、「大学入試レベル以上の単語をまとめて辞書引き」などという機能がある。

英語を書くための機能はなかなかのもの。F12を押すと、MS-IMEとEnglish Navigatorが融合してアプリケーション(例えばWord、Excel、メモ帳など)にはりつく。その状態で、例えば

yuuenchi

と入力して変換キーを押すと、普通にカナ漢字変換したように「遊園地」などと候補一覧が出て、そこで→(右矢印)キーを押すと、遊園地に相当する英語が出る。また、ittaと入力して変換すると、「言った」の場合には「said」、「行った」では「went」などとちゃんと過去形にしてくれる。あいまいな単語入力(understと入力)もできる。

3英語教育ソフト

3.1「Native　World」(沖北陸システム開発) 定価スターターキットで24,800円

http : //www.okisoft.co.jp/osg/ohsd/index.htm

昨秋、平岩賞「ソフトウェア・プロダクト・オブ・ザ・イヤー'98」に選ばれた英会話学習ソフトである。

大きく分けて「練習ステージ」「会話ステージ」に分かれている。「練習ステージ」ではビデオなどを見ながら単語やフレーズの練習、ヒアリング、発話練習ができる。自分が発話した文章は録音され、ネイティブの発音と聞き比べることができる。

「会話ステージ」では、画面上の外国人と会話をして進んでいくが、自分の発話によってストーリーの展開が変わってくる。また時と場合によっては同じことをいっても相手の反応が違うので、実践的な力がつく。

3.2「英語の魂①」(日本ユニシス・ソフトウェア(株)) 定価19,800円

http : //www.usk.co.jp/eitama/

IBM社ViaVoice搭載。以下、もらった資料から転載(一部省略)。

「英語の魂①」4つの特徴

(1)「英語の基礎」(中学3年～高校1年レベル)が体系的に整理されてます。

※各基本重要文型単位には全て音声・動画・静止画等の豊富な例題が添付されてます。

(2) 上記重要文型が参加型ドラマ、音声認識機能、ミュージカル等で感覚的に学習できます。約300画面の音声認識トレーニング、重要文型マスターで「英語の反射神経」が鍛えられます。聞く・話

す・読む・書く等のバランスの良い英語力が身につきます。

(3) 学校現場での副教材、社会人の為の「英語やり直し」用途に向いています。音声トレーニング・重要文型マスタは中学英語教科書の対応メニュー付です。

(4) 翻訳者・通訳者からも評価の高い見出し語104万語の高性能電子辞書を添付してます。自動意味表示・辞書編集等の機能がついており学習・仕事・インターネット等に最適です。

3.3「Microsoft ENCARTA インタラクティブ英会話」(マイクロソフト(株)) 定価12,800円

http : //www.asia.microsoft.com/japan/products/reference.htm

4月はじめ発売のマイクロソフトの英会話学習ソフト。このソフトは、今までの英会話学習経験者の悩みの声をもとに、マイクロソフトのソフトウェア技術を結集して生まれた英会話学習ソフトで、「学ぶ」より「慣れる」、「飽きずに学習を続ける」ための工夫をこらした、今までとは異なるタイプのソフトである。このソフトは、ECCの授業で正式採用とのこと。

実際にデモを見た。まず、プランナーという画面で、自分の学習プランをたて、それにそって学習していく。学習は、ビデオを見、聞きながら、会話を学ぶというもので、会話スピードも調節できる(遅いスピード、ナチュラルスピードで、別々に録音したという)。また、聞きとれなかったらその部分を文字で画面に表示したりできる。その日本語訳、文法解説、単語帳も表示できる。このようにして学習した成果を試せるのが、3Dバーチャルチャレンジ。これは、3Dグラフィックスの中を行き来して、そこにいる人たちと会話するというもの(人をクリックすると、その人が動いて本当に会話してるような感じになる)。会話次第で場面の展開が違ってくる。

3.4「ALC NetAcademy」(日立ソフトウェアエンジニアリング(株)) 1ライセンス・1年で35,000円(最低購入ライセンス:30ライセンス)

http : //www.hitachi-sk.co.jp/Products/ALC/

イントラネットを使用した、TOEICのための英語学習教材。デモを一通り見せてもらった。4部で構成されている。

(1) TOEIC診断テスト。TOEICの模擬試験を受ける。試験形態は本番と一緒だが、全部で20分で模擬試験が終るようになっている。

(2) Listeningの学習:聞く力をつける。面白いと思ったのが、「slow」ボタンを押すと、英語の読まれる速さが遅くなるのだが、もとの英語は1つで、コンピュータ処理により速度を早めたり遅くしたりする、というもの。なお、速さが変わっても声の高さは変わらないので面白い。

(3) Readingの学習:読む力をつける。ここで特徴的なのが、読む速さを自分で指定でき(200wpmなど。wpmは、words per minuteの略)、それにあわせて英語文が表示されていくというもの。「フレーズリーディング」、「キーワードリーディング」などができる。自分が読んだ実際の速さも表示され、内容については質問に答えて理解度がわかるようになっている。

(4) テスト:学習成果をはかる。TOEIC模擬試験。

協会活動報告

（1999年1月～3月）

理事会	3月30日	①1999年度事業計画案　②1999年度収支予算案 ③決算理事会および総会日程
運営委員会	1月26日	①来年度事業計画　②シンガポールサミット企画　③サミット合同会議
	3月3日	①来年度事業計画および予算　②総会、理事会の計画 ③シンガポールサミット
	3月24日	①1999年度事業計画案　②1999年度予算案　③シンガポールサミット ④MT白書
市場動向調査委員会	1月22日	①技術委との合同会議　②MT市場調査の進め方 ③MT白書への取り組み
	2月26日	①技術委との合同会議　②MT市場調査の回答報告 ③ユーザー調査の進め方
	3月19日	①技術委との合同会議　②MT市場調査のフォロー ③MT白書への協力
技術動向調査委員会	1月22日	①市場委との合同会議　②MT市場調査の進め方 ③MT白書への取り組み
	2月26日	①市場委との合同会議　②MT市場調査の回答報告 ③ユーザー調査の進め方
	3月19日	①市場委との合同会議　②MT市場調査のフォロー ③MT白書への協力
編集委員会	3月19日	①AAMT Journal No. 25の反省　②AAMT Journal No. 26の企画 ③ホームページ企画
ネットワーク翻訳研究会	1月18日	①JST対訳データ評価　②ホームページ翻訳評価　③MT白書協力
	2月23日	①対訳データの評価　②Webページ翻訳の評価　③来年度の方針
インターネットWG	2月12日	①メーリングリストの進め方 ②本年度の活動のまとめと来年度の取り組み
機械翻訳白書WG	1月18日	①技術委、市場委の活動報告　②MT白書の企画審議
	3月19日	①MT白書の基本方針　②予算および委託方法　③来年度の取り組み
MT Summit VII プログラム委員会		
	1月26日	①プログラム、実行の両委員会合同会議 ②企画、委員名、スケジュールの確認　③これからの分担協力
MT Summit VII 実行委員会		
	1月26日	①第1回目の挨拶、紹介　②シンガポールサミットの企画書、名簿の承認 ③これからの分担協力
	2月9日	①シンガポールサイドとの役割分担と連絡　②予算の考え方 ③推進体制づくり
	3月3日	①企画準備状況　②シンガポールサイドとの調整　③予算、後援、広報

AAMT
ジャーナル

No. 26
(April 1999)

発　行　アジア太平洋機械翻訳協会
所在地　〒105-0011　東京都港区芝公園3-5-12 芝公園真田ビル
TEL：03-5473-7135　FAX：03-5473-0569
E-mail : aamt 0001@infotokyo.ne.jp
ホームページ：http : //www.jeida.or.jp/aamt/
編集委員会　野村浩郷(委員長)　大倉清司　奥西稔幸　山端　潔
熊野　明
事務局　中島泰雄　高田佳代子
印刷所　伸光写植印刷株式会社