AAMT

The Asia-Pacific Association for Machine Translation

Journal

No.21
January 1998

アジア太平洋機械翻訳協会

目　　次

事務局だより

AAMT 事務局のご案内

'98年２月２日から郵便番号が７ケタとなり、「105」から「105－0011」に変わりますのでお知らせいたします。

また、会員の皆様にとって AAMT 事務局がオアシスとなるよう応接コーナーを広げましたので、どうぞ近くにお越しの節は気軽にお立ち寄りください。

〒105－0011
東京都港区芝公園３－５－12　芝公園真田ビル
電話　03－5473－7135
FAX　03－5473－0569

最寄駅は、地下鉄の日比谷線「神谷町」または、
三田線「御成門」です。

右が東京タワー、
左が AAMT のある芝公園真田ビル

エッセイ

High–End な機械翻訳システムの研究・開発

副会長　辻井潤一（東京大学教授）

ネットワークの発展にともなって、機械翻訳が一般に広く流布するようになってきた。秋葉原の電器店にいくと、さまざまなゲームソフトとともに、翻訳ソフトが売られているし、翻訳ソフトの優劣を比較する記事が雑誌などに頻繁に掲載される。商売にはつながらないといわれていた10年前が、夢のようである。ただ、機械翻訳をとりまく状況は、必ずしも、ばら色ではない。他のソフトとバンドルされて販売されるために、一本あたりの利益が低いことや、PCのブームが一段落して、それにともなって新規のユーザが増加しないことが影響している。それ以上に、翻訳の質、ユーザ・インターフェースなどの点で、まだ、多くのユーザを本当に引き付けるだけの商品になっていないことも、問題であろう。

昨年の秋には、アメリカ・サンディエゴでMT SUMMITが開催され、私は残念ながら出席できなかったが、多くの参加者を集めたそうである。日本からの参加者に会議の様子を聞いたところでは、翻訳者を中心としたユーザの参加が非常に多かったそうである。また、アメリカでは機械翻訳の研究開発の黄金期を迎えている、というアメリカからの参加者の感想もあったとのこと。

日本での機械翻訳の商品化が、ネットワーク利用の一般ユーザをターゲットにし、それ一辺倒になっていったのに対して、翻訳家や情報分析専門家のためのシステムへの興味も、ヨーロッパやアメリカでは強いように見える。情報分析のための機械翻訳という意味では、ヨーロッパにおける多言語情報検索の研究・開発の動きもある。これまで、別個に研究・開発されてきた情報検索と機械翻訳を統合するシステムの研究である。

ソフトウェアの、一般社会への普及は、拡張期と反省期を繰り返しながら、大きな傾向としては、徐々に浸透していくものである。ネットワークの広がりと相関して、広がった日本の翻訳ソフトは、いま小さな反省サイクルに入りつつあるようである。ただ、表面的なブームではなく、本当に使われて広がるためには、翻訳家や情報分析の専門家が業務に使えるようなシステムがあらわれる必要があると思う。これまでの延長で、ネットワーク用の機械翻訳だけでなく、業務に使われる翻訳ソフトである。この種の、値段は高いが、本当の業務に使うことのできるシステムが開発され、その技術がLow–Endの低価格の一般ユーザ用のシステムに還元されるという、ポジティブなサイクルが確立されてはじめて、機械翻訳システムが商品として安定したものになる、と思われる。

ネットワーク用の翻訳ソフトは、機械翻訳が浮世離れしたお話しではなく、現実的な商品になり得ることを示した点では、大きな役割を果たした。ただ、ネットワーク用の翻訳ソフト一辺倒では、ブームは一過性のものになる。再び、翻訳を業務とするマーケットを目標にした、High–Endの機械翻訳システムを目指す時期がきているように思われる。

1999年のMT　SUMMITは、本協会の田中会長が中心となり、私がプログラム委員長を引き受けさせてもらって、本協会が中心となってシンガポールで開催される。そのサミットで、アジア・日本でも、「機械翻訳の黄金期だ」といえるようにするためには、ユーザと開発・研究者をむすぶ本協会の役割は、ますます重要になてきている。

第6回 MT Summit が、昨年の秋アメリカ・サンディエゴ市にて開催され、日本からも AAMT 調査団11名を含む28名が参加し討議、研究、交流を深めて来ましたので、その状況を報告します。

概況

2年毎にアジア太平洋地域（AAMT)、ヨーロッパ（EAMT)、アメリカ（AMTA）と順番に開催されるMTサミットは、今回はAMTAが担当となりカリフォルニア州最南端のサンディエゴが会議場となった。

期間は'97年10月29日から11月1日までの4日間。世界23カ国から300名を越える学者、エンジニア、翻訳家などが一堂に会し活発な討議、交流が行われた。

発表のうち日本からは、京都大学長尾真教授、東京工業大学田中穂積教授を含む5件6名、展示会には、日本からAAMT（UPF活動の成果)、東芝、NTT、富士通の4社が参加した。

また、今回からIAMTのために顕著な貢献をした会員に対し、賞（IAMT Award of Honour）が贈られることとなり、第1回受賞者としてAAMT長尾真前会長が選ばれた。

さらに並行して行われたIAMT理事会および総会において、IAMT会長に田中穂積AAMT会長が就任、次回のMT SummitⅦを2年後の1999年にAAMTが当番となってシンガポールで開催することを発表し、多くの参加者から期待と歓迎の声で了承された。

MT Summit の開催状況

	開催地			年月日	参加人員
第1回	AAMT	日本	箱根	1997年9月17日～19日	200名
第2回	EAMT	西ドイツ	ミュンヘン	1989年8月16日～18日	284名
第3回	AMTA	アメリカ	ワシントン	1991年7月1日～4日	310名
第4回	AAMT	日本	神戸	1993年7月19日～22日	230名
第5回	EAMT	ルクセンブルグ	ヘミサイクル	1995年7月11日～13日	350名
第6回	AMTA	アメリカ	サンディエゴ	1997年10月29日～11月1日	320名
第7回	AAMT	シンガポール		1999年	

会場のカタマランリゾートホテル

第1回IAMT賞を受賞された長尾真前会長（左）と田中穂積現会長

MT SUMMIT VI 報告

九州工業大学　情報工学部教授　野村　浩郷

MT SUMMIT VIは1997年10月30日より11月１日にかけて米国San Diegoにおいて「機械翻訳の過去、現在、未来」のテーマの下に開催された。参加者数は314名（日本からの参加者数は28名）であった。

テーマに沿った各種講演と機械翻訳システムの紹介、および機械翻訳システムなどの展示があった。セッションは、オープニングおよびキーノートの他に30の課題別のものがあった。機械翻訳システムの紹介は、15あった。機械翻訳システムなどの展示は26社であった。

アジア太平洋機械翻訳協会（AAMT）と日本電子工業振興協会（JEIDA）はそれぞれ参加団を派遣した。本稿は、紙面の都合により、両参加団のメンバーが協力して作成した報告書から抜粋した内容の概要である。詳細な報告書はAAMT Journalの URLで閲覧できるようにする予定である。また、平成９年度末にJEIDAが作成する報告書にも掲載される予定である。

アジア太平洋機械翻訳協会からの参加団のメンバーは、田中穂積（東工大、団長）、雨宮弘和（東芝）、石原佳典（NEC情報システムズ）、伊藤悦雄（東芝）、押金章吾（富士通）、白木沢佳子（科技振）、福沢桃子（イデア）、美馬秀樹（ATR）、森本秀樹（富士通）、吉川昌隆（ブラザー）、中島泰雄（AAMT事務局）である。

日本電子工業振興協会からの参加団のメンバーは、野村浩郷（九工大、団長）、亀井真一郎（NEC）、北村美穂子（沖）、佐藤光弘（松下）、高山泰博（三菱）、田中裕一（ジャストシステム）、松井くにお（富士通）、宮川孝次郎（JEIDA、事務局）である。

機械翻訳は、国際化のさらなる進展およびインターネットの急速なる普及などにより、新しい局面を迎えている。多言語処理を含む機械翻訳は、避けて通ることのできない技術である。これに、昨今のインターネット高度活用技術である知的情報アクセスの技術を融合していく必要があろう。機械翻訳への取り組み開始50周年に当たり、日米欧における次世代機械翻訳へむけての重点的な取り組みへの胎動が感じられる。

Opening Session

M. Vasconcellos & W.S. Bennett

今年は、1947年にロックフェラー財団の自然科学部門長だったWarren WeaverがNorbert Wienerに宛てた手紙の中で機械翻訳の可能性を示した時から数えてちょうど50周年に当たる。それを記念して、今回の第６回MTサミットは「機械翻訳の過去、現在、未来」をテーマとした。また、機械翻訳の普及発展における特に顕著な功績に対して贈られるIAMT「Award of Honour」を創設した。その第１回目の授与者に京都大学の長尾真教授を選んだ。

Keynote : U.S Government Suport and Use of Machine Translation : Current Status

T. Pedtke（National Air Inteligence Center, USAF）

米国政府機関は機械翻訳技術の発展を長年に渡り支援してきた。米国政府機関における機械翻訳技術の重要性の認識が最近とみに高まっている。50年前に機械翻訳が提唱されて以来、概念検討期、初期開発期を経て1966年のALPACレポートによって暗黒期に陥るが、1979頃からルネッサンス期を迎え、1993年から現在に至る黄金期を迎えている。

現在、機械翻訳の重要性が再認識されている要因には、技術的なものと社会的なものとがある。技術的な要因としては特にインターネットの発展が挙げられる。これに伴って流通する情報の量が爆発的に増大しているため、機械による言語処理の重要性が再認識された。社会的な要因としては特に冷戦終結が挙げられる。これに伴い、多数の民族紛争が勃発するなど、それまで扱っていなかった言語によって書かれた情報源からの情報収集の必要性が生じ、機械翻訳への新たな期待が高まっている。

日本の場合、機械翻訳の必要性は経済的側面からのみ考えられるので、日本語英語間の翻訳以外には投資が行なわれにくい。これに対して米国では、上記のように経済的にはあまり必要性の高くない言語であっても、軍事的に重要な言語であれば機械翻訳の対象になり、政府機関からの投資が行なわれる。さらに研究開発人員の面においても、米国には世界各国から人が集まるため、他言語から英語へ翻訳するシステムを構築する人材の確保が容易である。これらの点において、米国の方が日本よりも多言語機械翻訳の研究開発が推進しやすい環境にあると言える。

Machine Translation Through Language Understanding
M. Nagao（Kyoto University）

現在商用システムとして主流になっている機械翻訳システムの問題は、文脈や語と語の関係を無視した変換ベースであることが大きな原因となっている。それを解決するため、語と語の共起関係を利用して翻訳する用例ベースの翻訳システムが注目されているが、それでも文の深い意味や話者の感情理解を必要とする翻訳には対処できない。この問題を解決する翻訳処理の１つが「言語理解による翻訳」である。

ここで言う言語理解とは、文章から意味ネットワークを獲得することを意味する。したがって言語理解による翻訳方法とは、まずパラグラフや文章全体を翻訳単位として知識辞書やヒューリスティックルールなどの多くの原言語の知識を用いて、構文解析、意味解析を行い、その文章の意味ネットワークを構築し、次に、その意味ネットワークから対象言語の文を生成する、という翻訳方法を意味する。

現在、その要素技術の１つである、意味ネットワークから日本語文を生成する実験を行っている。

今後、文化や個人の環境に深く関わる言語知識が大量に蓄積されれば、言語理解による機械翻訳が実現するであろう。

A Real-Time MT System for Translating Broadcast Captions
E. Nyberg, T. Mitamura（CMU）

テレビ放送の字幕（VBIでデータが送られるもの）を翻訳するシステムには、リアルタイムに処理できること、断片的な文や未知語、ミススペルにも対応できるロバストなシステムであること、事前／事後のエディットができないため、できるだけ精度の良い訳文を生成できること、などが求められる。そこで、事例ベースの翻訳エンジンと、知識ベースの翻訳エンジンとを並列に動作させ、それぞれの翻訳結果に与えられた点数から、一番良いものを訳文として選択する、というシステム構成を提案する。現在は、英語→ドイツ語の字幕翻訳を対象としたシステムを構築中である。

MT from the Production Perspective
M.Flanagan（CompuServe）

CompuServeはオンラインの文書翻訳・後編集サービスを５年間行っているが、今後期待されているサービスは、フォーラムなどのメッセージ翻訳やチャット翻訳、Email翻訳である。特に、即時性、マルチリンガル翻訳が望まれるチャット翻訳の需要は大きく、新しい翻訳ターゲットして注目されている。

チャットなどのインタラクティブテキストは、トピックが頻繁に変わる、スペルや文法ミスが多発する、文法に則らない記述や省略、略語表現などが多い、などの理由で変換ベースの機械翻訳システムでの翻訳が難しい。しかし、その一方で、一文が短い、そのフォーラム特有の決まった表現があるという利点もある。これらの特徴を生かした新しい翻訳手法が望まれる。

一方、ユーザインターフェースも難しい。チャット参加者に混乱を生じさせないよう、複数の言語の翻訳処理結果を会話の流れに沿ってユーザに提示する手法が望まれる。

F.Bruckert（Logos）

Logosは、情報産業分野などの国際企業を顧客に持つ翻訳サービス会社で、ネット経由での機械翻訳サービス、機械翻訳システム導入とその翻訳工程管理サービス、Translation Memory（TM）によるドキュメント管理までを含めた総合的サービスの３形態のサービスを提供している。翻訳は種々の文書形式（HTML,FrameMaker etc.）に準拠しており、膨大な

量の専門用語辞書を所有しており、多種多様なユーザのニーズに対応できるような細かなサービスを行っている。今後のMTの展望として、機械翻訳導入による文書作成までも含めた翻訳工程の効率化がある。

MT R&D in Europe
R.Havenith（EC）

ECはMTおよび自然言語処理関係のプロジェクトを長年に渡り推進してきた。現在のプロジェクトは以下のとおりである。

LRE	（Linguistic Research and Engineering program、1991年に開始）
ALEP	（Advanced Language Engineering Platform、1992年より開始）
LS-GRAM	（Large-scale Grammar、1994－1996）
SECC	（Statistical English Check and Correction）
TSNLP	（テストコーパスのデータベース、1997－1998）
LE	（Language Engineering Programme、1994年に開始）
Otello	（ネットワークの利用を推進するプロジェクト、1996－1997）
LinguaNet	（多言語通信システムの構築、1995－1998）、
Aventinus	（対麻薬のための国際警察プロジェクト、1995－1998年）

MT from an Everyday User's Point of View
A. Bech（Lingtech AS）

デンマークのLingtech社は、MTシステムを利用して、特許文書の英語、フランス語、ドイツ語からデンマーク語への翻訳サービスを行っている会社である。MTには、PaTransというシステムを利用している。MT利用により、50%程度のコスト削減に成功した。翻訳の質を向上するためには、辞書、特に専門用語の拡張が重要となってくるが、そこで問題となるのが複合語の登録である。現行のMTシステムに対する要求として、翻訳の質の向上ももちろんであるが、辞書登録などの周辺作業のためのツールの充実、Post-Editorの充実、翻訳精度向上のためのガイドツール、といったシステムとしての使い易さを向上させることが重要である。

R&D for Commercial MT
Exchange Interface for Translation Tools
Gr. Thurmair（GMS）

多言語処理におけるテキストと言語資源を交換するための互換性インタフェースの紹介。言語資源には、翻訳メモリ（翻訳対）、用語がある。テキストの交換については、SGML/HTML/XMLが優勢であるが、RTFも無視できない。最近提案された互換フォーマットの多くはSGMLによって規定されている。MARTIFという語彙の互換標準フォーマットがISOに提案（ISO FDIS12620）されているが、純粋に単語だけに特化した互換フォーマットだけでは多機能の語彙データベースには役に立たない。より実用的な語彙レベルのフォーマットとしてOLIF（Open LexiconInterchange Format）が提案されている。

また、EUのEuramisのフォーマットがすべての言語資源を統一的なフォーマットで記述している。

Laurie Gerber（SYSTRAN）

商用の機械翻訳システムでは、コスト省力化のための生産性がもっとも優先される。そのため、既存のシステムとコンポーネント、既存の辞書とコーパス、既存のツールを使いこなすことが重要。

The Current State of Machine Translation
H.L.Somers（UMIST）

機械翻訳研究はここ10年間で規模が非常に大きくなった。これは機械翻訳が研究、開発の段階を経て実用の段階に達していることを良く示している。機械翻訳システムの力点は、現在ではWEBのラフな翻訳など少し実用性にシフトしている傾向がある。例えばWEB翻訳では固有名詞処理、未登録語処理などの技術が重要であり、そのような技術が盛んに研究されている。機械翻訳は、すでに開発されている自然言語処理技術の統合システムとして、また新

規ドメインに適用されて新規技術を新たに研究しそれを試す土台として、各要素技術の完成度を高める上で最適の応用例であると言える。現在はまだ、話し言葉の翻訳に関する報告は少ない。話し言葉の翻訳は中心的な話題の一つになるであろう。さらに機械翻訳を有効活用するためには、機械翻訳システムをユーザを含めたトータルシステムとして広くとらえる観点が重要である。ユーザの行なうプリエディット、ポストエディットといった機械翻訳システム活用技術に関するさらなる研究がますます必要になる。

Sharable Formats and Their Supporting Environments for Exchanging User Dictionaries among Different MT Systems as a Part of AAMT Activities
S.Kamei（NEC），E.Itoh（Toshiba），
M.Fujii（NOVA），T.Hirai（Sharp），
Y.Saito（Fujitsu LAB），
M.Takahashi（Kyushu Matsushita），
T.Hiyama（NIS），K.Muraki（NEC）

アジア太平洋機械翻訳協会（AAMT）の技術動向調査委員会で行なっている機械翻訳ユーザ辞書の共通フォーマット設定活動（UPF）に関しての報告である。

機械翻訳システムを有効活用するには、新しい語彙やユーザごとに使用する語彙を新たにユーザ辞書として追加する必要があるが、翻訳辞書の構築には原言語目的言語双方に対する言語知識が必要であり、また実際の辞書作成には多大な時間と労力がかかるため、個々のユーザにとって大規模なユーザ辞書構築には限界がある。そこでこの問題を解決するため、機械翻訳ユーザ辞書の共通フォーマットを設定し、異なる機械翻訳システムのユーザ辞書の間でも辞書情報を交換できる電子環境（WEB ページ）を構築中である。

ユーザ辞書の共通フォーマットとしては、基本標準と拡張標準の二種類を設定する。基本標準とは、特に頻繁に登録される品詞に関して簡便にユーザ辞書が作成できることを目的としたフォーマットである。各機械翻訳ソフトは、このフォーマットに対応することが推奨される。拡張標準とは、ユーザ辞書間でデータ交換を行なう際に情報をもらさず変換できるようにするため、すべての辞書情報が記述できるように設計されたフォーマットである。共通フォーマットは、SGML 様のタグ形式による記述を採用している。また機械だけでなく、人間にとっての可読性を重視し、特別なツール無しにユーザが辞書内容を読むことができるようにするため、辞書のマスターはテキストファイルとしている。ただしユーザの辞書作成を支援するための辞書作成エディタは用意する。

現在は、基本仕様および拡張仕様をほぼFIX し、機械翻訳メーカ 6 社が協力して各メーカのシステムの独自形式のユーザ辞書と共通フォーマット形式のユーザ辞書との間の双方向変換を行なうコンバータを作成して翻訳実験を行ない、フォーマットの実用性を評価中である。評価終了後、1998年 3 月に全体の仕様を FIX し、一般公開の予定である。

JEIDA's Bilingual Corpus and Other Corpora for NLP Research in JAPAN
H.Isahara, JEIDA（通信総研）

電子協テキスト処理技術専門委員会で作成中の日英対訳コーパス（テキストデータ）を中心に、RWC コーパス、CRL 単文データも含めて、現在開発中の言語データについての発表である。テキスト処理技術専門委員会では、自然言語処理の研究のための一般公開を前提とした言語データとして、日英対訳コーパスを作成している。

この他、日本語の言語データとしては、RWC の作成している大規模言語データである RWC コーパスや、RWC のデータから係り受け関係を抽出した「CRL 単文データ」があり、研究用に利用することができる。

Multi-lingual Spoken Dialog Translation System Using Transfer-Driven Machine Translation
H.Mima, O.Furuse, and H.Iida
（ATR Interpreting Telecommunications Research Labs.）

ATR 音声翻訳通信研究所で話し言葉を多言語に翻訳することを目的に研究・開発された多言語対話翻訳システム Chat　Translation の発表である。Chat

Translation システムでは現在、日本語から英語、韓国語、ドイツ語、さらに英語、韓国語から日本語への5方向の翻訳が実現されている。ホテルの予約、道案内、トラブルシューティング等の旅行に関する会話全般の翻訳を対象としており、受け付けや窓口カウンターで使われる表現のほとんどはカバーしている。また、インターネット等を通じて多言語による会話シュミレーションを行なう機能も備える。

User–Friendly MT
David Farwell & Stephen Helmreich（NMSU）

ユーザフレンドリィな機械翻訳についての概念と、プラグマティックなアプローチによりそれを実現する方法。翻訳の過程は、翻訳者が特定の表現に表された原言語の著者の意図を確認することから始まる。これを解釈（Interpretation）と呼ぶ。その後で翻訳（Translation）が行われる。翻訳における最初の過程は、対象言語の発話者とは異なる言語学的慣習、異なる文化的慣習、及び異なる世界知識、すなわち著者の意図に基づいて、対象言語として表現することから始まる。次の過程は、著者の世界知識及び対象に関する知識を、翻訳者の世界知識及び対象に関する知識に置き換えることである。このような過程において、翻訳者は翻訳を生成しながらたくさんの選択を行わなければならない。

MT R&D in Asia
H. Tanaka（Tokyo Institute of Technology）

アジアにおけるMT研究は、CICCを中心とした多言語翻訳プロジェクト（MMT）がその代表であった。MMTでは、日本語、中国語、タイ語、インドネシア語、マレーシア語を対象とし、中間言語方式による多言語MTの開発を目指していた。プロジェクトは1995年に終了し、いくつかの成果を納めたが、実用的なMTシステムの開発には至らなかった。現在アジアでは、知識ベースから知識獲得へ、多言語から二言語へ、全文の翻訳から部分の翻訳へ、といったパラダイムシフトが起こっている。

From METAL to T1：Systems and Components for Machine Translation Applications
U.Schwall（GMS）

専門家利用型のMTは、スタンドアロン型のPCに変わりつつある。GMSのT1は、PC標準のユーザインターフェースをできるだけ採用し、かつ、徹底した翻訳ワークフローをユーザに与えることによって翻訳専門家の要求に答えた操作性の優れたPC上で動作する機械翻訳システムである。翻訳環境設定、ユーザ辞書作成などの通常のフローに加え、プロフェッショナルバージョンでは、TMの構築ツール、参照ツール、語彙拡張エディタ、語彙履歴機能などが提供されており、実際の翻訳現場（例：Siemens社）でも実績を持つ。

Java and its role in natural language Processing and Machine Translation
T.Read（Universitu of Granada）

Internet上でHTMLやCGIの使用により例えば下記のような自然言語ツールを提供しているサイトがある：
ARTFL：（仏英、英仏辞書）
http : //humanities.uchicago.edu/forms_unrest/FR–ENG.html
Logos：（多言語辞書）
http : //www.logos.it/forumget.html
travlang's translating dictionaries（38言語辞書）
http : //www.travlang.com/languages。

しかし、これらはデータタイプの制限やサーバの負荷などの問題があり十分に活用できるものではなかった。そういった状況下でサンマイクロシステムズがJAVAを提唱した。JAVAはプラットフォームによらないという触れ込みだったため、非常に高い関心を呼び、多くの自然言語研究者がJAVAを用いた自然言語システムやツールを開発しようと研究を重ねている。JAVAで書かれた自然言語処理ツールに関する情報源は以下である：
オンライン辞書参照システム
http : //www.ugr.es/~tread/molex.html
漢字検索ツール
http : //www.cs.arizona.edu/people/rudick/DrawApplet/index.html。

The MT Market Perspective（Europe）
C.Brace（LIM）

MTのヨーロッパの市場規模は$120billionに達し、年20%の成長率である。必要なのはワールドワイドの言語資源の共用。Unicodeの標準化などである。

The MT Market Perspective（World and America）
L.C.Miller（MCS）

1997年は1995年より20%上昇した。また、新しいマーケットとしてブラジルが注目されるようになった。今後注目すべき地域は中国とアラビア。今後の機械翻訳の発展に欠かせない項目は、機械翻訳をユーザの目から見ることであり、技術的にはUNICODE対応、Localization、標準化などである。

The MT Market Perspective（Japan）
Y.Shimamoto

日本ではインターネットの急速な普及が特徴であり、PCの家庭への普及率は1995年の2倍、プロバイダ数は7倍になっている。インターネットは2020年には10億ドルの市場になるだろう。日本では、機械翻訳の供給は、OEM、プレインストール、他のソフトとの組み合わせ、パッケージ小売り、オンラインサービスなどがある。ただし、小売価格が高いのでアメリカでは余り見かけないようである。100以上の英日システムがあるが、日英はその4分の1程度である。日本での機械翻訳の主な使われ方はマニュアルの翻訳やWEBの翻訳である。そのため、インターネット用の翻訳ソフトは大量に販売されている。

【システムプレゼンテーション】

CompuServe（MT of Real Time Chat : Mary Flanagan）

CompuServeのチャットを対象としてリアルタイムのオンライン機械翻訳を目的とするシステムであり、2人以上のチャット参加者の入力テキストを処理する。チャットのテキストには句表現、略語、省略が多いため、独特の困難があるが、今回開発したプロトタイプのデモでは、それほど問題あるような省略文等は入力されなかった。英語とフランス語の双方向翻訳。

Globalink

Globalinkのイントラネット上で、Webおよび電子メールの翻訳を行うもので、サーバ上にMTシステムが置かれている。

ALIS Technologies

1981年に始まった、インターネット、イントラネットを中心とした各種文書処理技術を提供する会社である。現在約20種類の言語を扱っており、90種類の言語の辞書を持つ。プロキシー方式のWeb翻訳システムを、社内イントラネット用の翻訳サーバーとして販売している。翻訳エンジンは自社で開発しているのではなく、ヨーロッパ言語はSYSTRAN、日本語はTSUNAMIを使っている。現在、LosAngels Times、GTE（ケベックの電話会社）、Xeroxがパイロットユーザとして利用している。

LOGOS Corporation

カスタマイズ機能（ユーザ環境指定）はかなり詳細化されている。辞書の分野は200種類。ターミノロジの収集にも力を入れている。その一方で、ユーザによる辞書構築や管理も簡単に行えるような環境を提供している。ユーザのフロントエンドはJAVAで実装している。

TRADOS

Translation Memory（TM）主導の翻訳支援システムである。TMのデータとなる翻訳テンプレートは種々のエディタで作成することが可能である。

【機械翻訳システムの展示】

展示していた機関・会社は以下の通りである。

SYSTRAN　NTT　NeocorTech LLC
LOGOS TRANSLATION SYSTEM　ESTEAM Inc.
ATR　CITAC Computer, Inc.
Fujitsu Software Corp.　Toshiba　AAMT
GMS　ALIS Technologies Inc.　IBT, Inc.

MCB Systems　PAHO MT Systems
USC Information Science Institute
EPI*USE Systems（Pty）Ltd.　Kielikone Ltd.
Universal Grammar
クローズドキャプションの翻訳

以下、いくつかのシステムについての展示内容の概要を順不動で述べる。

1）SYSTRAN Software Inc.
SYSTRAN PROfessional
英語をベースにフランス語、イタリア語、ドイツ語、スペイン語、ポルトガル語、日本語の各言語を対象とした双方向の翻訳システムと、ロシア語、中国語から英語への単方向の総14方向に対する翻訳システムが用意されている。機械翻訳システムのデベロッパーとしては歴史が古く、言語ペアのバリエーションが多いのが特徴である。14の言語ペアに対して総計250万語の辞書、ユーザ定義辞書、21種類の分野専門用語辞書を備えている。また、MS ワードを始めとする一般的なワードプロセッサ、WWW ページ等のレイアウトを保存したままの翻訳も可能である。Web サイトは http : //www.systransoft.com/

2）NTT Communication Science Labs.
ALT-J/E、日本語語彙体系
NTT コミュニケーション科学研究所で研究・開発されている日英翻訳システム。翻訳メカニズムとしては従来のトランスファ方式を基にしたものであり、製品として売られているものではないが、辞書、文法、語彙体系（シソーラス）共に研究・整備されており、"オープニング"、"超ベリーバッド" 等の日本語文中の未知のカタカナ語の翻訳も可能。また、現在、岩波書店より発売中の日本語語彙体系辞書が同時に展示されており、ブラウザ上でのデータ閲覧も可能であった。これは、3000種の意味属性を用いて日本語単語30万語の意味を定義したものであり、また、用言6000語の文型を結合価パターンにまとめたものである。

3）NeocorTech LLC
Tsunami MT for Windows（EJ），
Typhoon MT for Windows（JE）
亀島産業製の EJ-BANK と JE-BANK を翻訳エンジンとして英語版 Windows を始めとする IBM 互換パソコン上で実行可能な形にして提供している。Tsunami　MT（EJ）、TyphoonMT（JE）も日本のライセンス契約で使っている。約11万語のコア辞書と別売で生物学、物理学に関する各11万語の専門用語辞書が提供されている。Web サイトは、http : //www.neocor.com/

4）LOGOS TRANSLATION SYSTEM
Translation Express
英語からドイツ語、フランス語、スペイン語、イタリア語、日本語、及びドイツ語から英語、フランス語、イタリア語への機械翻訳システムを基に、インターネットを通じて翻訳サービスを行なうのを特徴とする。さらには個別の会社等のイントラネット向けの製品（Translation Control Center）も用意されている。さらには辞書作成ツールを用いて独自の単語対やフレーズ対を追加・作成することが可能となっており、翻訳処理自体を各ユーザでカスタマイズすることができるようである。また、Microsoft Word RTF を始めとする著名なワープロのファイルや HTML 文書等もオリジナルのフォーマットを保存したまま翻訳することが可能となっている。
Web サイトは、http : //www.logos-ca.com/

5）ESTEAM Inc.
ESTeam BTR
機械翻訳と翻訳メモリー（過去にリンクされた翻訳対を呼び出すことで翻訳処理を行なう技術）を用いた多言語翻訳システム。現在、英語、フランス語、ドイツ語、スペイン語、イタリア語を対象に、英語を中間言語としたピボット方式を採用することで、全ての言語に関して双方向の翻訳が可能となっている。

6）ATR 音声翻訳通信研究所
多言語話し言葉翻訳システム Chat Translation
話し言葉を多言語に翻訳することを目的に研究・開発された Chat　Translation システム。現在、日本語から英語、韓国語、ドイツ語、さらに英語、韓国語から日本語への５方向の翻訳システムが開発されている。ホテルの予約、道案内、トラブルシューテ

ィング等の旅行に関する会話全般を対象としており、一般に市販されている旅行会話集などにある表現のほとんどはカバーしている。また、インターネット等を通じて他言語との会話シュミレーションを行なう機能も備える。Webサイトは、
http : //www.itl.atr.co.jp/chattrans/index.html

7）AAMT（UPF活動の紹介）
AAMTのUPFによるユーザ辞書の変換。

8）Fujitsu Software Corporation
Altas JE/EJ, TransLingo
富士通のAtlasJE/EJで、日本語版Windowsが必要。メニューも日本語表示。一方、WWWブラウザーを翻訳するTransLingoは、英語Windows上のWWWブラウザーで日英翻訳をする。

9）KIELIKONE Ltd
Transmart
フィンランドで唯一の機械翻訳開発会社。
MS Wordで作られたフィンランド語のマニュアルをレイアウトを保ったまま、英語に翻訳する。

新刊図書の紹介

AAMTの会員である大阪大学　成田　一教授が、「パソコン翻訳の世界」という本を発行し、AAMTにも寄贈がありましたのでご紹介します。

書　名：　「パソコン翻訳の世界」
著　者：　成田　一（大阪大学　言語文化部　教授）
出　版：　講談社現代新書
価　格：　640円
内　容：　「この本のタイトルを『パソコン翻訳の世界』としたのは、あまりテーマを絞らないで、機械翻訳に関わる問題をあらゆる角度から扱いたいと思ったからです。パソコンでの翻訳の仕組みはどうなっているのか。人間の翻訳とどう違うのか。どの程度の文章が翻訳できるのか。機械翻訳を利用するには語学力が関係するのか。翻訳能力をどう評価すればいいのか。どういう使い方があるのか。機械翻訳にはどういう歴史があるのか。日本における開発状況はどうなっているか。そもそも翻訳とはどういうものなのか。言語によって翻訳が易しい難しいということはあるのか。翻訳を育む文化というものがあるのか。こうした誰もが知りたいと思うことについてお話ししたいと思います。」

（著者のまえがきより）

イベント報告

テクニカルコミュニケーター協会

「テクニカルコミュニケーションシンポジウム'97」の報告

1997年9月4・5日　　東京・工学院大学

「かたち、色、音、動き……電子空間でのTC」というテーマで、去る9月4・5の両日、テクニカルコミュニケーター協会の「テクニカルコミュニケーションシンポジウム'97」が開催された。

グラフィックデザイナー杉浦康平氏の基調講演のあと、8っの分科会、42人の発表、6っの特別セミナーおよび展示会と盛りだくさんのプログラムで運営され、昨年を上回る1050人が参加した。

AAMTも協力団体の一つとしてこのシンポジウムに協賛した。

SanDiego 参加印象記

アメリカのバイタリティ

九州工業大学情報工学部　野村浩郷

機械翻訳は、国際化社会および情報化社会において避けて通ることのできない技術である。いわゆる自然言語処理技術を基盤として50年に渡る研究開発が実をむすび、現在の機械翻訳システムが実現・活用されるようになった。しかし、技術は未だ初歩の段階にあり、多言語処理に拘わる諸技術のさらなる発展と高度情報技術との融合を今後も重点的に推進しなければならない。

今回の MT　SUMMIT に参加して強く感じたことを敢えて二つに絞って述べると、次のようになる。

まず第一は、米国のたゆまない挑戦である。米国における機械翻訳への取り組みは常に政府の直接的な支援と政府との密接なる関連において推進されてきたといっても過言ではない。50年前の将来技術としての機械翻訳の提案もさることながら、いわゆるスプートニクショックに端を発する研究開発の歴史はいくつかの困難な局面に遭遇しながらも着実に発展し、今また世界をリードしようとしている。その背景となるキーワードは、インターネットの構築・活用と種々の紛争に対処するための情報である。インターネットは21世紀の最も重要な社会基盤であり、社会構造と生活様式を変革する技術である。その GII（Global Information Infrastructure）としての性格上、多言語処理・機械翻訳は必須である。ここで注目すべきことは、そのような要請に答えようとして日々挑戦する財政的支援者と実務的研究者・技術者の存在である。常に問題点を探求しかつ明確にし、その解決に挑むバイタリティは、機械翻訳のみならずすべての分野における優位化・差別化の原動力であるかのように見える。

次に第二は、高度情報技術としての知的情報アクセスとの融合である。これに関しての言及はほとんどなかったように思える。言葉としてはインターネットや Java などがいくつかでていたが、問題点を掘り下げるには至っていなかったように思える。現在は、ブラウザ上での翻訳や言語データの開発・共有などが主たる論点となっている。現時点での機械翻訳の活用やさらなる機械翻訳技術の発展にこれらが重要であることは言うまでもない。しかし、そもそも自然言語処理技術が知的情報アクセスという観点から追究されるべきものであるといってもよいものであるから、時代を先取りするという観点からは、情報アクセスにおける機械翻訳が中心的な論点となっていてもよかったであろう。機械翻訳の未来は言語バリアを克服する知的情報アクセスにある。機械翻訳の新しい局面はインターネットへの知的情報アクセスの要請に関して顕在化するものである。

実り多い調査団

富士通株式会社　押金章悟

青い空、青い海、海に浮かぶたくさんの小船――のどかな街というサンディエゴの第一印象の通り、MT サミットの会場となったホテルも海を望む絶好の場所にあり、リラックスして講演を聞くことができました。

セッションの内容としては、機械翻訳の歴史や現在の研究動向についてのものが多かったようです。ユーザの視点からの発表も多く、これは機械翻訳が現実の業務の中で実用的に利用されているためと考えてよいでしょう。全体的には、ネットワークの普及による情報のオンライン化で翻訳の需要は増大するとともに、膨大な量かつ短期間の納期の翻訳業務に対しては機械翻訳が必要不可欠のものであると関係者に認識されていることが確信できました。

講演や展示などで見受けられましたが、テレビの字幕翻訳に的を絞ったツールなどは日本でも今後、利用の場やニーズが増えて来るのではないかと思います。衛星放送などの普及により海外のコンテンツが日本に流入してくれば、機械翻訳が必要にならざるを得ないようになるでしょう。

サミット開催中にハロウィンがあり、思い思いに

着飾った仮装を見てアメリカの風俗に触れるという貴重な機会もありました。実り多い調査団に参加できて、同行の参加者の皆さんに深く感謝いたします。

まだまだ伸びる確信

株式会社東芝　伊藤悦雄

欧米の機械翻訳技術や製品に生に接する滅多にない機会であったので色々と学ぶことが多く有意義な調査団でした。特に、かつての翻訳結果を利用する「翻訳メモリ」や独自のインターフェースを持たずに「MS-WORDとの連携」を重視する姿勢に特長を感じました。

MTサミットを通じ、機械翻訳はまだまだ伸びていく分野であり、いずれ、サンディエゴの様に明るい日差しを浴びる日が来ることを確信しました。

San Diegoの印象

科学技術振興事業団　白木澤佳子

First impression

San Diegoの空港からバスで会場＆宿泊先のカタマラン・リゾート・ホテルに向かいました。バスの窓から見えるのは青い空、青い海、ゆったりとした街並み。ホテルのロビーには熱帯の植物が鬱蒼と茂り、時間がゆっくりと流れているよう。

Weather

毎日、気持ちのよい晴天続き。昼間はTシャツ1枚で大丈夫ですが、朝晩は結構冷えますので上着が必要です。例によって会議場は冷房がしっかりと効いていて、上着が手放せません。それにしても、どうしてアメリカ人はいつでも、どこでも、Tシャツ、短パンで寒くないのでしょうか？

Box lunch

予め3日分のボックスランチ（3日分で18ドル）を頼んでいたので、どこで何を食べようか、と悩むことはありませんでしたが、中身はと言うと、丸いパンのサンドイッチ1個（自分でマヨネーズやマスタードを付けて食べる）、リンゴ1個、ものすごく甘いケーキ、ポテトチップス小1袋、缶ジュースと、いかにもアメリカ的。ちょっとがっかりしたのも事実です。

Catamaran Resort Hotel

カタマラン・リゾート・ホテルは高層の建物1棟と、それを取り囲むように立っている2階建てのコテージ風の長屋からなっています。私はコテージ風の方に泊まりましたが、朝は庭を歩いている黒鳥の鳴き声で目が覚めました。庭にはインコもいて、止まり木に”Don't feed! They bite!”という札が下げられていました。のんびりとしたいいホテルでした。

Halloween

10月31日がハロウィンだったので、家のドアや窓に飾り付けをしたり、カボチャを飾っている家を見かけました。夜は若者達が天使、ゾンビ、動物など思い思いに仮装して大騒ぎ。ハロウィンというと子供たちのお祭りというイメージが強かったのですが、一番楽しんでいるのは大人達かもしれません。会議の参加者にも、カボチャ柄のネクタイをしている人がいました。

次回は仕事ではなく、のんびりと滞在を楽しむために訪ねたいと思いました。

ラッキーな経験

㈱イデア・インスティテュート　福澤桃子

自分がこれからかかわろうとしている分野への最初の一歩がこのような国際会議の出席から始まったことは、大変ラッキーだと思う。ヨーロッパや、アジア、アメリカの多くの人々が機械翻訳に携わり、興味をもっていること知った。また、機械翻訳とはどのようなもので、どのような研究が行われているのか、私たちのような翻訳会社や翻訳者が実際に使用できるものがあるのか等、日本にいて本や資料を探そうとしてもなかなか見つけられないが、このように一堂にそろった場所で実際のソフトを見せてもらいながら話を聞けるという機会は、このMTサミットを除いてほかにあり得ない、ということも実感することができた。

サン・ディエゴは、気候も良く、新しいものと古いものが交ざり合ったすてきな街である。サミット以外でも、毎晩楽しい催しが行われ、知り合って間もない方々との交流を深める絶好の機会が提供された。おかげで、自分の世界が少し広がったように思う。貴重な経験であった。

UPF説明員の喜びと失敗と

㈱NEC情報システムズ　石原佳典

今回のMTサミットには、AAMTが取り組んでいるUPF活動の展示説明員として参加しました。

展示はUPF活動の説明を行い、実際にノートパソコンを操作してUPFによるユーザ辞書の変換ができることを見せる予定でした。しかし、実際にはUPFの活動を説明することが主になり、ノートパソコンを使ってのデモまで行うことはあまりありませんでした。活動内容を説明すると、一様に好意的な反応が返ってきました。また、UPF活動を日英・英日のシステム以外にも広げる予定はないのか、または、広げてほしいという質問や要望を多く受けました。UPF活動の紹介という性格上、かなり地味なブースでしたが、最終的に当ブースを訪れた人は70名ほどと地味な展示の割りには人が集まり、よく聴いてくれましたのでホット一安心しました。

このように全般的に順調に終わった展示ですが、失敗談というか笑い話（私が笑われた話）は、日本電子工業振興協会の団長として参加しておられた九工大の野村先生に対して、UPFの説明を英語でしてしまったということでした。野村先生はお名前を知っていただけでご本人にお会いしたことがなかったのと、無言のまま私の前に立たれ、その何ともいえない雰囲気（迫力）に圧倒され、エンドレステープレコーダーのように自動英語説明モードになっていた私は思わず、英語で説明を始めてしまったのです。まわりの様子から、どうやらこの方は日本人だなと気が付いたのですがあとのまつり。今更日本語にも変えられず、そのまま冷や汗たらたらで貫きました。

夕食の席でさっそく調査団の笑いの種になったのは言うまでもありません。

ビジネスへの期待

㈱ATR音声翻訳通信研究所　美馬秀樹

サンディエゴは11月と言っても昼間は泳ぐことが可能なほど温暖で、朝夕はジョギングには最良の気候となります。サミットの会場となったホテルが海岸に面していることもあり、カリフォルニアのイメージとしては最高のシチュエイションで、早速ジョギングシューズを手に入れ、映画のワン・シーンを演じている気分で、毎朝、海辺で汗を流してから会議に臨むという、（日本ではまずやらないような）健全な生活を送ってしまいました。

発表では、全体として、技術的内容の詳細を述べると言うより、それぞれの翻訳技術の特徴、主張点を明確にすると言った傾向が強く、専門の技術者のみならず、（政治的な意味あいも含め）様々な分野の人間を対象とした主張を行なう傾向が見られました。

また、システムプレゼンテーション等では、多言語を対象とした、インターネットホームページ、ワードプロセッサ等の既存のフォーマット付き文書の翻訳に大きな興味が持たれていることが伺えました。

近年では、機械翻訳黄金時代とも言われているようで、展示会場においては、既に商品化されているもの、ネットワーク経由による機械翻訳サービス、今後実用化が期待される音声翻訳システム等、様々な機械翻訳システムが展示され、非常に盛況でした。日本のみではなく、アメリカにおいても、（ビジネスチャンスとして）商用機械翻訳システムに大きな関心が寄せられていることが伺えました。

また、クローズドキャプション（字幕）の翻訳システムのように、ハードウェア化された機器も既に発売されており、今後の機械翻訳ビジネスの可能性が大いに期待されるところです。

技術早分かり

機械翻訳と統計手法

NHK 放送技術研究所　浦谷　則好

この「技術早分かり」シリーズができたのは日本機会翻訳協会であったころの第3号（１９９１年１１月）からです。最初の４年間で機械翻訳の基本となる要素技術の解説は完了しています。そして，AAMTジャーナルNo.14からは人間の翻訳と機械翻訳を比較する観点から新たな技術解説が始まったのですが，品詞，文法，辞書，曖昧さの問題，対話の処理，翻訳の方法と解説してきました。読者にとってはまだまだ知りたいことは多いのかも知れませんが，私が書くことのできるテーマはすでに尽きている気がしています。そこで，今回は視点を変えて，最近機械翻訳に限らず，コンピュータによる言語処理で盛んに使われている統計手法について概説してみたいと思います。数学に全く興味がない方も，まずは読んでみて統計手法のおもしろさを漠然とでも理解していただければ幸いです。

言語処理で使われる統計手法について概説すると言っても，私は統計の専門家ではありませんので細部で不正確なことを書いてしまうかも知れません。ここではあくまで統計の素人にちょっとでも統計手法のことを理解してもらおうという気持ちで書きますから，そのつもりで読んでください。

１．平均値ってなんだろう？

何かの数値の与えられたデータをもらうとすぐに平均値を出したがる人がいます。そして，平均値を出してデータ全体を理解したような気になっています。もちろん，もうちょっと用心深い人は分散まで求めるのでしょうが，しかし平均値ってそんなにありがたいものなんでしょうか。別に私は偏差値教育を批判するためにこの話を持ちだしたわけではないのです。純粋に科学的な意味で疑義を挟みたいのです。

平均値はある集団（あるいはそこからの標本）が与えられたとき，「典型的な値」あるいは「位置を示す尺度」として用いられるものです。ですから，もし集団から少数の標本が抜け落ちたからといって大きく変化するようでは困るわけです。つまりはちょっとした測定の違いに対してロバスト（頑強）でなければなりません。正規分布（ガウス分布ともいわれるあの中央が丸い山になっている良く見る分布形です）の場合には確かにあらゆる意味で平均値は最良の尺度であると言われています。しかし，日本の家計の平均収入を聞いて，「俺はそんなにもらってないぞ」と思う人は多いはずです。それは収入の分布が正規分布から大きくずれているからです。少数のお金持ちが平均を押し上げている結果です。このように左右対称とはならない分布の場合には平均が良い尺度でないことは直感的にも分かりますが，コーシー分布（自由度１のｔ分布）のようにちょっと見には正規分布と似た形の場合でも平均はほとんど意味を持たないことがあるのです。ある集合からランダムに２０個選んで，平均値を求めるということを何回も実行することを想定します。そのとき，平均値のロバスト度合いをその分散で見てみると，母集団（標本を選ぶ元となる集合）が正規分布の場合を１．０としたとき，コーシー分布では１００００を超えてしまいます。それでは，もっと良い尺度はないものでしょうか。メディアン（中央値）はコーシー分布でも分散は３程度でずっと良い尺度ということが言えます。ただし，正規分布の場合には１．５となって平均値より劣ります。トリム平均（標本の上位α％と下位α％を除いて平均を取ること）でαを１０（つまり上下２個ずつ除去）としたときには，コーシー分布で７．３，正規分布で１．０６程度になって極めて良い尺度ということが分かります。これからは，正規分布から外れた分布を相手にするときは代表値としてトリム平均かメディアンを用いることにしましょう。

２．ｎグラム

「翻訳X_1になっX_2らやめX_3まで毎X_4が勉強X_5のに，X_6械は『X_7職』しX_8ら毎回X_9字通りX_{10}すだけX_{11}全然勉X_{12}しないX_{13}。」

さて，X_1～X_{13}までそれぞれ1文字が入るのですが，どんな文字か分かりますか。正解は，X_1＝家，X_2＝た，X_3＝る，X_4＝日，X_5＝な，X_6＝機，X_7＝就，X_8＝た，X_9＝文，X_{10}＝訳，X_{11}＝で，X_{12}＝強，X_{13}＝ね，です。この文は前号の「技術早分かり」の先頭の文章です。あなたはどれくらい分かりましたか。全問正解というわけにはいかなくても，かなりあなたは良い点数が取れたはずです。その理由は，「頭が良いから」に違いありません。さて，それでは何を手掛かりにX_1～X_{13}までを推定したのでしょうか。一言で説明すれば，この文章の文脈とあなたが保持している常識を使って推定したのではないでしょうか。

それでは，「頭の悪い」コンピュータに同じことをやらせるにはどうしたらよいでしょうか。エドガ・アラン・ポーの小説「黄金虫」には暗号を解くのに英語文字の出現頻度の違いを用いる方法が出てきます。英語文ではスペースを除けば，一番多く出現する文字は‘e’です。こうした傾向を使って暗号を解読するわけです。このように個々の文字の出現の傾向を確率を使って表すことをユニグラム(unigram)と呼んでいます。ある記号が‘t’だと分かったとき，その次に来る記号を解読するには，ユニグラムだけではなく‘t’に続く文字の傾向を使うことができます。‘t’の次に来る可能性の一番高い文字は‘h’です。他には‘a’，‘e’，‘i’，‘r’，‘o’，‘u’などが候補として有力です。ある文字が出現したという前提のもとで次の文字の出現傾向を確率（条件付き確率）で表すことをバイグラム(bigram)と呼んでいます。一般的には（n－1）個の前提となる文字の出現のもとでn番目の出現傾向を確率で表すことをnグラム(n-gram)と呼ぶわけです。式で書くと，$P(x_n|x_1x_2\cdots x_{n-1})$という具合になります。

それではその確率はどのようにして求めたらよいのでしょうか。その答えは単純です。データを集めて，確率を計算すればよいのです。もちろん，データは今から推定しようとしているものにできるかぎり近いものを選ばなければなりません。上の例題のX_1～X_{13}を推定するのに数学の参考書のデータを使のはあまり有用ではないでしょう。それから，nグラムのnはできるかぎり大きいことが望ましいわけですが，表が大きくなる（表の大きさは文字の種類のn乗に比例するので日本語の場合ＪＩＳにある文字だけに限定しても，nが2で5千万，nが3なら3千兆を超える）ことと，表がスパース（値の入らない場所が多くなること）になって確率の推定が不正確になるため，nは2や3に取られることが多いようです。nが2ならX_2を求めるのに‘っ’が前にあるという事実しか使えません。（逆向きもゆるせば‘ら’が後であるという事実も使えますが）人間に比べてなんて貧弱な知識しか使えないものでしょうか。でも，この技術を使えば推定の範囲をずいぶん狭められるのがお分かりになるでしょう。

nグラムは音声認識の音素の推定などに良く使われています。機械翻訳でも形態素解析の品詞推定やスペルチェックなどに使われることがあります。ひとつ補足しておかなければならないことがあります。nグラムの確率はデータから推定すると言いましたが，データが相当に大きなものでなければ観測値そのものを推定値とするのは危険です。データには1回も出現しなかった文字のつながりが現実には起こりうるかもしれないからです。もし，そういうことに遭遇すると大抵のプログラムは破綻をきたします。そこで，観測値に修正を加えて推定値としています。一番単純には観測値では0となってしまったもの全てに小さな値（例えば0.0001）を与えるという方法が取られます。通常は

$$\begin{aligned}P(x_n|x_1x_2\cdots x_{n-1}) = {}& \lambda_n P(x_n|x_1x_2\cdots x_{n-1})\\ &+ \lambda_{n-1} P(x_n|x_2x_3\cdots x_{n-1}) + \cdots\\ &+ \lambda_2 P(x_n|x_{n-1}) + \lambda_1 P(x_n) + \lambda_0\end{aligned}$$

のように1～nグラムまでの観測値にλ_nからλ_1の重みを掛けて修正する方法が取られます。λ_0は全ての推定値を底上げする定数です。このλ_n～λ_0を決めるのにいろいろな方法が提案されていますが，ここでは触れません。データが十分にあるときはλ_nは1に近い値を取り，他は0に近い値になりますが，データが十分でないときはλ_nを1に近付けること（したがって他のλを0に近付けること）ができません。

3．tスコア

「N_1がV_1しているN_2をV_2した」という文があるとします。このときV_1の（意味上の）主語はN_1でしょうかN_2でしょうか。V_2の主語はN_1なのでしょうか。具体的なもの，例えばN_1＝鳥，V_1＝飛ぶ，

N_2＝空，V_2＝見る　などが入れば人なら簡単に何が何に係っているのか，その意味は何なのかを理解することができます。N_2＝虫，V_2＝食べる　なら係り受けやV_1の意味上の主語が変ってしまうことがお分かりでしょう。つまり，上の文は品詞だけでは係り受け関係を決められないわけです。それでは，「頭の悪い」コンピュータに人まねをさせるのにどんな方法が考えられるでしょうか。

このとき共起関係（ある単語と別の単語が一緒に出現すること）を使うことが考えられます。すなわち，「飛ぶ」の主語となりうるものを辞書に全て列挙しておく，「見る」の主語となりうるもの，目的語となりうるもの，・・・という具合です。しかし，これは事実上不可能です。そこで，言語データが集めてあればそこから統計的にこうした関係を導き出せないかということを考えます。

ｔスコアというのは，２つの集団において，平均値に差があるかどうかを検定するのに用いる指標で，

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

で計算されるものです。$\bar{x}_1$，n_1，σ_1はそれぞれ，x_1の平均値，標本数，標準偏差で，x_2側も同様です。x_1とx_2が独立で正規分布しているとき，このｔもまた（規準）正規分布することが分かっていますので，ｔの絶対値が１．９６より大きければ５％の有意水準（つまり間違いの判定をする確率が５％以下のこと）で，「もとの集団の平均値は同じでない」と判定することができるわけです。

さて，それではこのｔスコアを使って，単語と単語の共起関係をどのように計算したらよいのでしょうか。まず，２つの単語が同時に出現すると見なせる範囲を決めます。例えば，「同一の文」とか，「同一の段落」とか，「２つの単語の距離が１０単語以内」だとか，応用に応じて決めるわけです。そして，同時に起こる回数f_{xy}と個々の単語の出現回数f_x, f_yを求めます。全単語の出現回数をＮとすれば，同時に起こる確率（共起確率）p_{xy}，個々の生起確率p_x，p_yは，

$$p_{xy} = f_{xy}/N,\quad p_x = f_x/N,\quad p_y = f_y/N$$

で求められます。今，x_1として，２つの語が共起する確率を取り，x_2としてそれらが独立に共起するという前提のもとで，（たまたま）２つの語が共起する確率をとれば，ｔスコアは「２つの語が独立に共起している」という前提からのずれの大きさを表していることになります。言い換えれば，ｔが大きいほど２つの語は同時に出てくる傾向が強いわけですから共起関係にある語と言うことができます。この場合のｔスコアの式をf_{xy}，f_x，f_y，Ｎで書き直すと，

$$t = \frac{\frac{f_{xy}}{N} - \frac{f_x}{N}\frac{f_y}{N}}{\sqrt{\frac{f_{xy}}{N^2}}}$$

となります。前述の式の分母のσ_2を含む項が消えた形になっています。これは，ｐx，ｐyが小さいときにσ_2^2/Nがf_{xy}/N^2よりずっと小さくなって無視しても計算値にほとんど誤差がでないからです。

ここで，「共起関係の強さはｔの大きさに比例する」とは言えないことに注意が必要です。ｔスコアは単に前述したように「同じ」という前提からずれの大きさを示しているだけで，決して直接的に共起関係の強さを表しているわけではないのです。後述する相互情報量とは反対に，ｔスコアは頻度の多い語の方が大きくなる傾向があるのです。

４．相互情報量

情報量はその情報が持っている情報の大きさを示すもので，$-\log p$ で計算されます。ｐは事象が起こる確率で，logの底としては普通２が用いられます。コインの表裏を当てるような賭けではそれぞれが1/2の確率ですから情報量は１（ビット）となります。「1000回に１回しか起きないような事象が起きた」という情報なら2^{10}=1024ですから，約１０ビットとなり，情報量の観点からはコインの表裏より１０倍価値ある情報と言えます。

ところで，ある事象ｘの生起を別の事象ｙから推測することを考えます。このとき事象ｙが起きたことを知ったとき，事象ｘの情報量の変化（差分）は次の式で計算することができます。

$$\begin{aligned} I(x;y) &= I(x) - I(x \mid y) \\ &= -\log p(x) + \log p(x \mid y) \\ &= \log \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

これを相互情報量と呼んでいます。この式は見てすぐわかるようにxとyに対して対称となっています。

この相互情報量も単語と単語の共起関係を測ることに使うことができます。p(x,y)を2つの単語が同時に共起する確率とし，p(x)，p(y)をそれぞれが共起する確率とすれば，もし，xとyに共起関係がないならば，p(x,y)=p(x)p(y)となり相互情報量は0となります。つまり，yの生起を知ることはxの生起を知る何の手掛かりにもならないということになります。逆に，相互情報量が大きければ，yの生起の情報はxの生起の良い予測を与えることになるわけです。xとyの頻度を用いて上式を書き直すと，

$$I(x;y) = \log \frac{\frac{f_{xy}}{N}}{\frac{f_x}{N}\frac{f_y}{N}}$$

となります。相互情報量は情報量の差分ですから，頻度の少ないものが大きくなる傾向があります。例えば，対象とした集合の中でxもyも1回しか出てこなくて，しかも共起と見なす範囲にその両方が出現した場足には，I(x;y)はlogNとなって非常に大きな値になってしまします。したがって，相互情報量はf_x，f_yがある一定値以上ないと意味をなさないことに注意が必要です。

5．カイ2乗分布

サイコロは昔からよくギャンブルに利用されますが，法律論をさておけば，まともなサイコロを使用すれば問題はないはずです。しかし，サイコロが「まとも」であるかどうかはどうやって検定すればよいのでしょうか。1つのサイコロを60回転がせばどの目も10回でることが期待できます。しかし，実際にはある目は9回しかでないかも知れないし，別の目は12回出るかも知れません。それでは，次のような結果となったとき，このサイコロはまともと言えるでしょうか。

目	1	2	3	4	5	6
観測度数	15	7	5	11	6	16

こうしたとき，観測度数と期待度数とがどの程度一致しているかを測る指標があります。この測度はカイ2乗といい次の式で定義されます。

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

ここでoおよびeは観測度数と期待度数を表しています。kは種類数（ここでは6）です。上の表をもとにχ^2を計算してみると11.2になります。このχ^2の値はある形の分布をすることが分かっています。これがカイ2乗分布と呼ばれるもので，自由度（上のkから1を引いたもの）によって大きく形が変化するような分布形です。分布形が分かれば，そこから何％のものがそのχ^2以上（あるいは以下）の値を持つのか知ることができます。上の例の場合ですと自由度が5の場合に相当しますので，χ^2の値が11.07で95％になります。したがって，上のような場合にはまともなサイコロでないと疑った方が良いようです。（専門的に言えば有意差5％で棄却されると言います）

それでは，この分布は言語処理ではどんな場面で利用できるのでしょうか。ある特定の単語の存在から，その文章がどんなカテゴリーに属しているのかを推定できる場合があります。例えば「選手」という言葉が出てくれば，スポーツに関係する文書だということが推定できます。ところで，こうした単語を候補の中から自動的に選び出す方法はないでしょうか。分類したいカテゴリーが予め決まっていれば，候補となる単語の出現頻度を求め，下のような表（分割表と言われます）を作ることができます。

		カテゴリー A	B	C	D	計
単語	α	18	29	70	115	232
	β	17	28	30	41	116
	γ	11	10	11	20	52
計		46	67	111	176	400

もし，カテゴリーと単語は関係ないとすれば，2つの分類水準は独立であるということになります。もし，χ^2の値が大きくなれば独立であるという仮説は棄却（すなわちカテゴリーと単語には関係がある）されます。ところで，サイコロの場合と違って上のような表の場合，どうやって期待度数を求めたらよいのでしょうか。それには，周辺分布を利用します。もし，独立ならば表のマス中の値は周辺分布から計

算できるはずだからです。例えば（A，α）のところの期待度数は

４００×２３２／４００×４６／４００

で計算できます。自由度は（列数－1）×（行数－1）で，上の場合には６となります。この表のχ^2の値を求めると，２０．７となります。自由度６の５％棄却点は１２．６ですから，この結果は有意（独立でない）と判定できます。しかし，これでは「単語とカテゴリーは独立でない」ことが分かっただけです。実際には，カテゴリーにあまり依存しない単語もあれば，カテゴリー依存度の高いものもあるわけです。カテゴリー依存度の高い単語を探すには，単に$(o_i - e_i)^2 / e_i$の大きいものを選べは良いことは容易に想像できることでしょう。

６．相関

相関はいくつかの変数の相互の関係を調べるために用いられます。しかし，相関係数では２つの変数が直線的に関連しているときにのみ有効な測度となりうることを忘れてはいけません。たとえば，xとyとに$x^2+y^2=1$のような関係があるとき，この２つの変量の関連性が強いことは明白ですが，得られるサンプルデータの相関をとってみれば０に近くなるはずです。相関係数は，

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

で定義されます。ここで，s_x,s_yは標本標準偏差を示しています。

２つのカテゴリーに属する文書から５０文ランダムに文を抜き出し，その長さの相関係数を測ったら0.2が得られたとします。さて，この２つのカテゴリーには相関があると言えるでしょうか。もちろん，0.2は相関があるとしてもかなり小さな値ではありますが，そもそも0.2が意味のある数字かどうかという問題です。言い換えますと，相関係数rの信頼性をどう考えたらよいのかという問題です。

実はxとyがどちらも正規分布をするという仮定の下でｒの理論的な分布形を求めることができます。この分布形はサンプル数とρ（rの真値）に依存し，かつ非対称な分布となりますから簡単ではありません。しかし，次のような変換をするとwはほぼ正規分布するということが分かっています。

$$w = \frac{1}{2}\log_e \frac{1+r}{1-r}$$

つまり，このwを使って，正規分布による検定を行なえば良いわけです。

wの平均値μ_wと標準偏差σ_wは，

$$\mu_w = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \qquad \sigma_w = \frac{1}{\sqrt{n-3}}$$

で得られます。ですから，上の問題を$\rho=0$の検定問題（つまり相関が０）として扱えばよいわけです。$\rho=0$，n=50から$\sigma_w=0.146$，$\mu_w=0$が得られますから，９５％の信頼区間は

$$-0.29 < \rho < 0.29$$

となることが分かります。r=0.2はw=0.20に対応しますから上の範囲に入ります。つまり，仮説は採択され，r=0.2は有意でないということになります。

ところで，標本の大きさが２８でr=0.6の場合のρの９５％信頼区間はどの程度になるでしょうか。計算してみると

$$0.29 < \rho < 0.79$$

となります。r=0.6が中心値にならないこと，区間の幅が広いことに注意してください。つまり，標本の大きさがかなり大きくないかぎり相関係数はあまり信頼できる測度でないことが分かります。

７．むすび

言語データベースの整備が進むにつれて，それを機械翻訳に利用するという動きも活発化しています。前回の「技術早分かり」では，対訳例文を機械翻訳に利用する方法を紹介しましたが，対訳データベースでなくても，現実の言語データを観察すれば，内省では気がつかない様々なものが見えてきます。統計手法は大量のデータから意味のある情報を自動的に抽出するための強力な道具となります。ここでは，６つの手法を紹介しましたが，主成分分析，クラスタリング，エントロピーなどまだまだ知っていた方がよいと思われる手法が存在しています。以下に，統計の勉強に役立つと思われる本２冊を紹介して今回の終りとします。

参考文献

Ｐ．Ｇ．ホーエル著：初等統計学，培風館
奥野忠一ほか：多変量解析法，日科技連

研究機関紹介

KAIST
Korea Advanced Institute of Science and Technology

Key-Sun Choi
kschoi@cs.kaist.ac.kr

1．はじめに

本記事では、KAIST（Korea Advanced Institute of Science and Technology）における自然言語処理研究の活動について述べる。機械翻訳に限定したものではなく過去５年間のあらゆる研究開発に触れている。２章では、KAISTが現在行なっている基本的な状況に触れる。次に、基盤技術とその応用について言及する。最後に研究者の紹介を行なう。

2．基礎—インフラストラクチャと標準

韓国語の利用者は世界で８番目に多い。利用者の大部分は南北韓国の人々であり、次いで中国北西部、日本、ロシアに生活する利用者が多い。韓国語の方言は多様である。基本的な部分における問題はほとんどないが、技術的なコミュニケーションで幾つかの問題がある。

同じ韓国語でも他地域の言語で書かれた文書に対しては翻訳が必要な場合があるため、それぞれの地方や地域の現代語利用者のための韓国語のインフラを構築する必要がある。これを目的としたプロジェクト［２］が、文化省（Ministry of Culture）、科学技術省（Ministry of Science & Technology）の支援で始まっている。プロジェクトは1992年より計画策定が始まり、1994年に10年計画が承認された。同年後半より始まったプロジェクトは大きく分けて「インフラストラクチャ構築（infrastructure building）」、「標準化（standardization）」、「応用（application）」の３つのカテゴリからなる。最初の２つのカテゴリではKAISTの指導のもと、いくつかのサブプロジェクトが国家規模の研究機関の15チームと２つのソフトウェアハウスにより実施されている。インフラストラクチャ構築プロジェクトは、情報ベース（information base）の構築と、その開発／管理環境とから成る。情報ベースとは、生コーパス、タグ付けコーパス、ツリーバンク、辞書、専門用語集、のことである。開発／管理は、タギングワークベンチ、構文木タギングワークベンチ、意味タグワークベンチ、バイリンガル対訳コーパス、辞書／コーパス管理システムなどである。コーパスとしては、生コーパスが100Mトークン、形態素／構文タグ付けコーパスが50Mトークン、ツリーバンクが200Kなどがある。我々はこのタスクに既に３年間を費やしているが、我々の10年計画は３つのフェーズから成る[13]。最初のフェーズは「単語」に焦点を置いたものである。このフェーズはほとんど完了した［http://kibs.kaist.ac.kr,written in Korean］。1997年の後半から始まる第二のフェーズは「文」レベルに重きを置くものであり、最終フェーズは「文脈」レベルをキャッチフレーズにしている。

インフラストラクチャ構築プロジェクトは高度な研究の達成を目的としたものではなく、既知の技術やリソースを完成させること、すなわち様々な研究開発のために利用できる基本モジュールや大規模なリソースを構築することを主眼に置いた。そのためには標準化が不可欠になった。その際、我々は標準化というものが研究と異なることを学んだ。仕様は１つ１つ議論する必要があり、時間を要する作業である。さらに正式の標準化は多くの努力、忍耐、実用性の検証を要する。我々は、標準というものは正式決定の前に実際に利用されるべきであると考え、プロトタイプ段階ですぐに利用できるように配布してきた。

各モジュール並びにリソースの記述に関する標準化は達成できた。現在は、それらが国内の標準（少なくともデファクトスタンダード）となるよう進めている。リソースの標準化は形態素－構文タグ、構文構造のタグセット、TEI（Text Encoding Initiative）やSDML（以下で述べる）のようなコーパス・辞書記述言語から成る。言語処理モジュールの標準で

は、あらゆるインタフェースにおいてSDMLだけでデータがやり取りされるように提案する。

データ・情報の構築は研究とは別のものである。今回のプロジェクトを通じて我々は研究者やリソースをいかに管理するかを学んでいるが、更に重要な事は、各研究者のアイデアをいかに管理するか、そして創造性を制限せずにいかに自由を与えるか、にある。言語工学の複雑さは知識工学やソフトウェア工学以上である。ソフトウェア工学の「滝（water-fall）」モデルでは、ステップ毎の逐次的な洗練化並びに同期したフィードバックが必要とされるが、これに対し専門家の知識を計算機に移植するには表現や学習が解決されなければならばい。言語工学のタスクは「滝」モデルを超えその本質は分散・非同期である。言語工学の知識構築のタスクは知識工学者に限ったものではない。たとえば、辞書の中身を編集する辞書編纂者は人間知識のソースであるだけでなく知識工学者でもある。言語工学における知識構築の本質は、知識工学者と人間知識資源の間のコミュニケーションが問題なのではなく人間の知識資源と巨大な情報ベースの間に問題があることにある。その次の問題は、各研究者が開発したそのようなリソースやシステムをいかに統合するかである。

この考えは「分散情報アーキテクチャ（distributed information architecture）」と呼ばれるコンセプトに通じている。その1つは、辞書や情報の記述用言語SDML（Standard Dictionary Markup Language）の特殊な版で実現されている［15］。また、TDMS（Text and Dictionary Management System）と呼ぶ情報管理システムも実装している［14］。

人間に限らずあらゆるエージェントはコミュニケートし合っている。エージェントには次の条件を満たす情報伝達者を要する。すなわち、どのようなエージェントにでもデコード可能で、いかなる内容を表現でき、自分自身の表現スキーマの中に存在する他のエージェントへエンコード機構を提供できる情報伝達者である。エージェント標準化組織FIPA（Foundation for Intelligent and Physical Agent）に対し、SDMLの更新版としてSKDL（Specificationof Knowledge Description Language）を提案した。［Tokyo 96；Turin97，http：//drogo.cselt.stet.it］

この概念に基づき、KAISTではABMT（Agent-BasedMachine Translation）の研究を進めている。同様のメカニズムは”shake-and-bake machine translat-ion”や”multi-engine machine translation”として提案されていた。違いはプロセスがエージェントによって非同期に制御される点にある。すなわち、各モジュールに対する多重エンジンは最適な解を求めて互いに協力／競合しあいながら計算を進める。その時のコミュニケーションはデコーディング手法を含む標準化言語SKDLにより処理されるが、各コンポーネントは随時その時点の処理に割り込むことができる。

以上、全体構成を述べてきた。次に文献リストを参考に各研究を簡単に紹介する。

3．部品と応用

名詞句認識には正確な形態素解析とタグ付け結果が必要である［11］。情報検索を目的とした名詞句認識に関してはいくつかの発表がある［1］，［12］，［19］。初期の研究では、形態素解析を用いて1語に絞り込み、この結果を構文解析と格構造にかける。しかしながら、この手法は語彙の不足によりすぐに行き詰まり、研究の焦点は統計的手法による自動シソーラス構築に移った。関連語の「クラスタリング」を見つけるのに有用とされる「コロケーションマップ」の概念が必要となったが、無限の文書空間並びに単語のペア間の無限の対応を前提にしているコロケーションマップは計算量があまりに大きいので現実的な解決法ではない。サンプリングサイズが自動的に発見できればこの問題は減少する可能性はあり［8］、そのための基礎的な統計が［6］で示された。

重みだけでなくシソーラスの構造を動的に制御するというアプローチもある［9］。また、情報検索システムを現実の問題に適用すると、［16］で述べられるように未登録語や外来語が重大な問題となる。バイリンガル辞書［10］の構築に効率的な手法を見つけるのに、英語と韓国語間のバイリンガルコーパスワークベンチを研究してきた。この成果は機械翻訳に適用可能であり、現在は英—韓［3］［7］、韓—英［10］、中—韓［17］の3種類の機械翻訳を研究開

発している。

構文解析では統計的手法とルールベースシステムとを統合するように試みている。その1つとしてHMM（Hidden Markov Model）をRTN（Recursive Transition Network）に適用する研究がある[4],[5]。また、依存文法（dependency grammar）をインサイドーアウトサイドアルゴリズムのような統計手法に統合することも試みている[18]。

4．メンバー

研究所のメンバーは教授2名、研究奨学生4名、常勤の博士過程の学生15名、修士学生10名、非常勤の研究生4名、スタッフ4名から成る。4名の研究奨学生のバックグラウンドは様々であり、異なる国の博士号を持っている。オースチンの Texas 大学からの心理言語学者、Paris 第7大学からの計算機語彙学者、Tuebingen 大学の意味論の研究者、Harbin 工科大学の自然言語処理研究者である。中国からの修士学生も一人いる。4名のスタッフの業務は文書管理、図書出版、中国語の語彙編集、秘書業務である。KAIST はキャンパス内にベンチャーパークを持っており、研究成果は現在、ソフトウェア会社へ供与／協力されている。

References

1．Park, Hyoukro, GeneSung Chung, Key-Sun Choi," Linguistic Information Based Index Term Weighting", International Symposium on Natural Language Understanding and AI(NLU & AI), Kyushu, 1992, 7.

2．Choi, Key-Sun, Young S. Han and Oh W. Kwon," KAIST Tree Bank Project for Korean : Present andFuture Development", Proceedings of International Workshop on Sharable Natural Language Resources (SNLR`94), Nara, 1994, 8.

3．Choi, Key-Sun, Seungmi Lee, Hiongun Kim,Deok-bong Kim, Gil-Chang Kim, "An English-to-Korean Machine Translation," PRICAI '94, Beijing, 1994, 8.

4．Choi, Key-Sun and Young S. Han, "A Reestimation Algorithm for Probabilistic Recursive Transition", COLING '94, Kyoto, 1994, 8.

5．Han, Young S. and Key-Sun Choi, "A Chart Re-estimation Algorithm for Probabilistic Recursive Transition Network",Computational Linguistics, Vol.22, No.3, 1996.

6．Park, Hyuk R., Young S. Han, Joong H. Shin and Key-Sun Choi, "An Upper Bound Estimate for the Entropy of Korean Texts", Literary and Linguistic Computing, Vol.11, No.3, 1996.

7．Mitkov, Ruslan, Kang-Hyuk Lee, Hiongun Kim, Key - Sun Choi , " English - to - Korean Machine Translation and the Problem of Anaphora Resolution", Literary and Linguistic Computing,Vol.12, No.1, 1997.

8．Kim, Moonjoo, Young S. Han, and Key-Sun Choi," Collocation Map for Overcoming Data Sparseness",The 7 th Conference of the European Chapter of the Association for Computational Linguistics(EACL-95), Dublin, 1995. 3.

9．Park, Young C., Young S. Han and Key-Sun Choi," Automatic Thesaurus Construction using Bayesian Networks",Conference on Information and Knowledge Management(CIKM '95), Baltimore, Maryland, 1995, 11.

10. Shin, Jung H. and Young S. Han and Key-Sun Choi,"Bilingual Knowledge Acquisition from Korean-English Parallel Corpus Using Alignment Method (Korean-English Alignment at Word and Phrase Level)," COLING '96, Copenhagen, Denmark, 1996, 8.

11. Jung, Sung-Young, Young C. Park and Key-Sun Choi,"Markov random field based English Part-of-Tagging system",COLING '96,Copenhagen, Denmark, 1996, 8.

12. Han, Young S., Hyuk R. Park, Key-Sun Choi and Kang H. Lee,"A Probabilistic approach to compound

noun indexing in Korean texts",COLING '96, Copenhagen, Denmark, 1996, 8.

13. Kim, Seongyong and Key-Sun Choi,"Korean Language Engineering : Current Status ofthe Information Platform", COLING '96, Copenhagen, Denmark, 1996, 8.

14. Choi, Byung-Jin, Jae Sung Lee, Woon Jae Leeand Key-Sun Choi, "A Logical Structure for theConstruction of Machine Readable Dictionaries", The 11 th Pacific Asia Conference on Language, Information and Computation, Seoul, Korea, 1996, 12

15. Choi, Byung-Jin, Jae Sung Lee, Woon-Jae Lee,Key -Sun Choi, "The Processing of Dictionary Definitions and Maintenance in Text and Dictionary Management System(TDMS)",International Conference on Computer Processing of Oriental Languages(ICCPOL '97), Hong Kong, 1997, 4

16. Lee, Jae Sung and Key-Sun Choi, "Two Approaches to the Roman to Hangul Transcription Based on Statistical Translation Method", ICCPOL '97, Hong Kong, 1997, 4.

17. Li, Jun-Jie and Key-Sun Choi, "Corpus-Based Chinese-Korean Abstracting Translation System", IJCAI '97, Nagoya, Japan, 1997. 8.

18. Lee, Seungmi, Key-Sun Choi, "Reestimation and Best First Parsing Algorithms for Probabilistic Dependency Grammars", The Fifth Workshop on Very Large Corpora(WVLC-5), Hong Kong, 1997, 8.

19. Nam, Jee-Sun, Key-Sun Choi, "A Local Grammar -based Approach to Recognizing of Proper Names in Korean Texts", WVLC-5, Hong Kong, 1997, 8.

イベント報告

日本翻訳連盟

「国際コミュニケーション・フォーラム」の報告

10月22・23日　東京・東条会館

恒例のJTF翻訳祭は、今年は「国際コミュニケーション・フォーラム」と名称を変え、内容もブラッシュアップされて去る10月22・23の両日東条会館で開催された。

デジタル色が一段と濃厚になってきた<情報ビッグバン>への途上にあって、2001年に向かう国際コミュニケーションのあり方など、盛りだくさんのテーマで熱心な講演、討論が繰りひろげられた。

初日の記念講演では、米国マイクロソフト社のD.Brooks氏が、バーチャルリアリティ、サイバー・コミュニケーションなどをキーコンセプトにコンピュータ・ネットワークなどディジタル・コミュニケーション環境が創る「21世紀の異文化交流」について、また、基調講演では、ディジタル時代への教育を展開する富士通ラーニングメディア・山口俊治氏が来るべき情報民主主義について語るなど、近未来における期待と課題が語りかけられた。

２日目の基調講演では、国際ジャーナリストの日高義樹氏が異文化交流の難しさについて語り、「カルチャー・リテラシー」つまり比較とジャッジメント、そして自らのアイデンティティが必要であることを説かれた。そのあと、翻訳事業者、翻訳者、さらには翻訳発注企業から現状と課題について、また、放送、映像などのメディアやディジタル・コミュニケーションにおけるコンテンツ・ローカライゼーションなどに関連して、現場からでなければ得られない多くの提言がなされた。

AAMTは、この「国際コミュニケーション・フォーラム」を後援するとともに、並行して行われた10社のエキジビションにUPF活動の成果を展示し、参観者のうちの約50名に詳細な説明を行った。

Semantics Implementation based on Extended Idiom for English to Korean Machine Translation

Yu Seop Kim, Yung Taek Kim

Department of Computer Engineering
Seoul National University
Seoul, 151-742, South Korea
E-mail: yuseop@plaza.snu.ac.kr

Abstract

In English-Korean machine translation, the meaning difference between the source sentence and the object sentence usually comes from the result of literal translation. The meaning of the source sentence would be distorted when translated literally and delivered not properly to object language speakers, especially between the two languages, English and Korean.

For the resolution of this problem, this paper proposes Extended Idiom which usually has meaning fixed. The Extended Idiom includes the conventional idioms, collocations, verb patterns and etc., in the source language as well as in the object language. Sometimes the Extended Idiom can have various meaning depending on the context of the sentence. And the meaning of idiom is determined by using the context constraints.

The algorithms for acquisition and recognition of idioms are presented in this paper and the related problems are shown with the solutions. The meaning accuracy and the efficiency of the time and memory space for the translations are shown to be improved by extended idiom through the experiments.

1. Introduction

In machine translation, it is difficult to get the precise meaning of object sentence for the source sentence[1]. The reason is that the meaning of a word or a phrase in a source sentence does not match exactly the meaning of a word or a phrase in an object sentence[11]. This leads to a mismatching of the meanings between the two languages.

Fig.1 shows a frequent translational equivalent between a source sentence and the corresponding object sentence. Area B is the portion of a sentence that the meaning in the source language and the object language is matched properly. But area A and area C are shown as the poor translated portion of the sentence. For example, when the source sentence *"I have a lot of money"* is translated literally, the phrase *"a lot of"* would cause to generate a poor portion of translation.

The source sentence may include some idioms which can be understood only by source language native speakers, so the meaning of the literal translation of the sentence would not be proper for the object

language speakers. In this paper, we use the Extended Idiom to recover the meaning deficiency for the translation. During the translation, an idiom plays a role as a keyword in each sentence, which helps the parsing to reduce time complexity and memory space and also helps semantics analysis to get the precise meaning for the object language speakers[3][8].

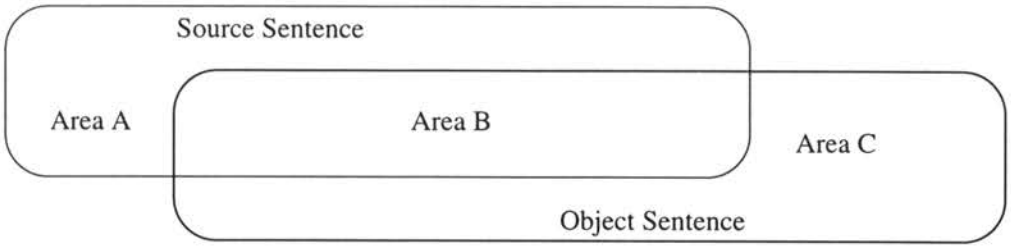


Fig.1 Translational Equivalent

The Extended Idiom in section 2 consists of conventional idiom, collocation, verb pattern, idiomatic prepositional attachment, sentence-level idiom and transformation. A conventional idiom has fixed meaning with relatively high probability in a sentence. And it plays very important role in machine translation for semantics implementation as well as for syntactic analysis[3]. A collocation is another form of an idiom which is a transformation of the example sentences. The verb pattern is also dealt as another form of an idiom. Each pattern of the verb provides own meaning playing a role of an idiom. And some nouns and adjectives have also a preposition for the idiomatic prepositional attachment. The scope of idiom would range from a word to a complete sentence.

In section 3, the process of idiom acquisition consists of two phases. That is, the first is candidate phrase discovery phase and the second is the verification phase. And the problems caused during the acquisition and recognition of idioms are shown and the solution for these are proposed. In section 4, the results of experiments are given with graphs, which show the improvement of meaning correctness and memory space and processing time.

The structure of current our system with four phases is shown in Fig. 2 and the phases are explained briefly below.

Lexical Analysis

In this phase, the source sentence is analyzed lexically. Using the lexical dictionary, most of the part of speeches of the words for the sentence are decided.

Syntactic Analysis

After idiom recognition, the parse tree for the input sentence is constructed by the set of grammar rules.

Semantic Analysis

With semantics implementation, the transfer phase is performed using the transfer dictionary.

Generation

The object sentence is generated using generation dictionary.

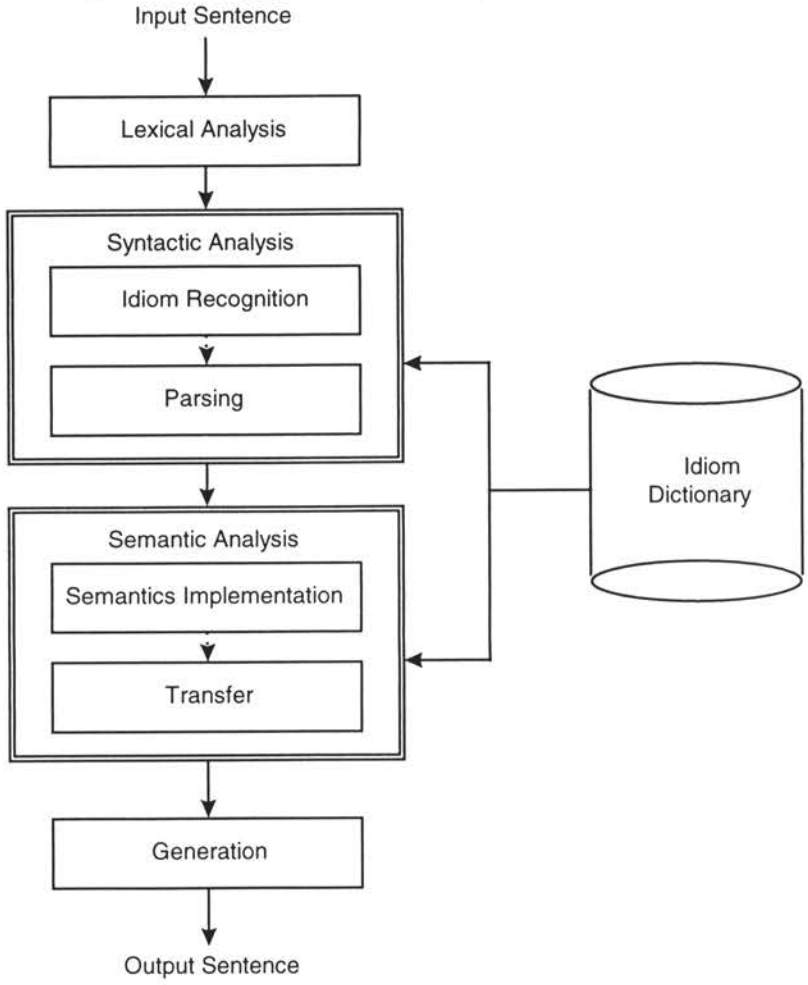


Fig. 2 Structure of current system

2. Extended Idiom

Extended idiom is explained as follows.

2.1 Conventional Idiom

The conventional idiom includes source idiom which is admitted by source language speakers and object idiom which is recognized as an idiom by object language speakers.

Source Idiom

The word stream of source idiom has different meaning from the sum of meaning of each words in idiom stream. The definition of source idiom is as follows.

$M(id_s) = M(W_1W_2...W_n) \neq \Sigma M(W_i)$
(id_s : *source idiom*, W_i : *ith word*, $M(\alpha)$: *meaning of* α)

Followings are some source idioms listed by its part-of-speech.

1) Noun Idiom
"A hot line" --> "직통 전화 연결" (a direct, secret telephone link between two important people)
"the pros and cons" --> "찬반 양론" (the arguments for and against a matter)
i.e. "the casting vote", "the cold war", and etc
2) Verb Idiom
"read between the lines" --> "눈치 채다" (understand or sense more than the actual words say)
"make up one's mind" --> " 결정하다" (take a decision)
i.e. "keep one's word", "bring up" and etc
3) Adjective Idiom
"quite a few" --> "상당히 많은" (much numerous)
"afraid of _" --> "_를 두려워하는" (fearful)
i.e. "good at _", "as sharp as a deedle" and etc
4) Adverb Idiom
"so far" --> "지금까지" (till now)
"again and again" --> "몇번이고" (repeatedly)
i.e., "all over again", "for God's sake" and etc
5) Other Idiom
"as well as" --> "뿐만 아니라" (and ... also)

Object Idiom

Object idiom can be formed as an idiom when there exists more suitable target expression in object language even though a corresponding meaning in literal translation can be found. The definition of object idiom is

$M(id_o) = M(W_1W_2...W_n) \cong \Sigma M(W_i),\ E_o(M(id_o)) \neq E_o(\Sigma M(W_i))$
(id_o : *object idiom*, $E_o(\alpha)$: *expression of object idiom*)

Some of object idioms are listed below by its part-of-speech.

1) Noun Idiom
"book reading" --> "독서(reading)"
"school book" --> "교과서(textbook)"
i.e. "younger brother", "institute of technology" and etc
2) Verb Idiom
"be at war" --> "전쟁 중이다(situation)"
"make haste" --> "서두르다(be in a hurry)"
i.e. "agree on one point", "bear in mind" and etc
3) Adjective Idiom
"my own" --> "나 자신의(my)"
"any other" --> "그밖에 다른(anything else)"
i.e. "little or no", "many more" and etc
4) Adverb Idiom
"even more" --> "더더욱(moreover)
"a long time ago" --> "오래 전에(long ago)"
i.e. "all alone", "many more" and etc

2.2 Collocation

A collocation comes from the example sentences for each word[5][6]. The meaning of word can be fixed variously in accordance with co-occurrence word.

$$M(W_\alpha) = M_\beta \text{ if } C(W_\alpha, W_\beta)$$
$$M_\chi \text{ if } C(W_\alpha, W_\chi)$$
$$\ldots\ldots$$
$$M_\delta \text{ otherwise}$$

($C(W_1, W_2)$: W_1 and W2 are co-occurred, M_i : meaning)

Following is the list of some collocations classified by its part-of-speech.

Verb Transitive Collocation
"pick"
(The thief) picks a lock. --> "자물쇠를 풀다(unlock)"
(The thief) picks a key. --> "열쇠를 집다(take)"
Verb Intransitive Collocation
"run"
boys run (in the garden) --> "소년들이 달리다(rush)"
computer runs (fast) --> "컴퓨터가 수행되다(execute)"
Adjective Collocation

"fair"

(I saw a) fair lady --> "아름다운(beautiful) 숙녀"

(I saw a) fair judge --> "공평한(impartial) 재판관"

Adverb Collocation

"hard"

(I) studied hard --> "열심히(eagerly) 공부하다"

(I) pulled (the door) hard --> "강하게(strongly) 당기다"

Preposition Collocation

"with"

(I) went (to the airport) with (my) mother --> "나는 나의 어머니*와 함께*(together) 공항에 갔다"

(I) observed (the cell) with (my) microscope --> "나는 나의 현미경*을 사용하여*(tool) 세포를 관찰하였다"

2.3 Verb Pattern

The verb patterns are defined in the dictionaries for some verbs[2][5] and each pattern has its own frozen meaning.

$$M(V_\alpha(P_\varphi, P_\chi, \ldots, P_\omega)) = \begin{cases} M_\varphi & \text{if } P(V_\alpha) = \varphi \\ M_\chi & \text{if } P(V_\alpha) = \chi \\ \quad\ldots. & \\ M_\omega & \text{if } P(V_\alpha) = \omega \end{cases}$$

(V_i : a verb word, P_i : pattern I, $P(V)$: the pattern number of verb V)

For example,

"make"

P4 --> "하려고 하다" (will)

"I ***made to leave*** the tent" --> "나는 천막을 *나오려고 했다*"

P5 --> "되다" (become)

"I'll ***make a dentist***" --> "나는 *치과의사가 될 것이다*"

P6 --> "만들다" (create)

"God ***made man***" --> "하느님이 *인류를 창조하셨다*"

P14 --> "_에게 _을 만들어 주다" (give)

"I ***made him a new suit***" --> "나는 *그에게 새 양복을 맞춰주었다*"

The verb patterns in this paper are listed below.

P1 : V	P2a : V + Adverb	P2b : V + (for) Adverb
P3 : V + Preposition + O	P4 : V + Infinitive	P5 : V + C

P6 : V + O	P7 : V + O + Adverb	P8 : V + Infinitive
P9 : V + Gerund	P10 : V + what + Infinitive	P11 : V + that Clause
P12 : V + what Clause	P13 : V + O + Preposition + O	P14 : V + I.O + D.O
P15 : V + O + that Clause	P16 : V + O + what + Inf	P17 : V + O + what Clause
P18 : V + O + Adjective	P19 : V + O + Noun	P20 : V + O + Infinitive
P21 : V + O + (to be) C	P22 : V + O + Bare Infinitive	P23 : V + O + Present Participle
P24 : V + O + Past Participle	P25 : V.Aux + Bare Infinitive	P26 : V.Aux + Infinitive
P27 : V.Aux + Present Participle	P28a : V.Aux + Past Participle	P28b : V.Aux + Past Participle

2.4 Idiomatic Prepositional Attachment

A prepositional attachment of high frequency is defined as an idiom as follows,

$$M(Pr_\alpha) = M_\beta \text{ if } C(Pr_\alpha, V_\beta)$$
$$M_\chi \text{ if } C(Pr_\alpha, N_\chi)$$
$$M_\delta \text{ if } C(Pr_\alpha, A_\delta)$$
$$\ldots\ldots$$

(Pr_i : a preposition word, N_i : a noun word, A_i : a adjective word)

A main word and a preposition are combined to be an idiom. The main word can be a verb word, a noun word, and an adjective word. The prepositional phrase underlined below can be translated into two or more different phrases in Korean because of the semantic ambiguity of the preposition, but using idiomatic prepositional attachment, correct one can be selected.

The verb attachment is composed of a verb word and a preposition word.
"***go to*** school" (n) --> 1) "학교*에 가다*" (direction)
2) "학교*까지 가다*" (goal)

The noun prepositional attachment is composed of a noun word and a preposition. For example,
"the ***access to*** the gate" (n) --> 1) "문*으로의 접근*" (direction)
2) "문*까지의 접근*" (point to)

The adjective prepositional attachment is composed of an adjective and a preposition. For example,
"***sad for*** his failure" (n) --> 1) "그의 실패*에 대해 슬퍼하다*" (reason)
2) "그의 실패*를 위해 슬퍼하다*" (sake)

In the above, 1) is the correct translation for each attachment.

2.5 Sentence-level idiom

A full sentence can be recognized as an idiom. And some sentences used in dialog also can be classified as the sentence-level idiom. A single word, a phrase, or a sentence can be realized as an idiom sentence.

A single word roles as an object sentence.
"Congratulations" --> "축하합니다"
"Fine" --> "좋습니다"

A phrase can be translated into an object sentence.
"No problem" --> "문제없습니다"
"Of course not" --> "물론 아닙니다"

A complete source sentence can be translated into a phrases of object language.
"It depends" --> "경우에 따라서는"
"I bet you" --> "틀림없이"

And a complete source sentence can be translated into an object sentence.

Fixed form Idiom

"All around is fair" --> "날씨가 좋다"
"Fools rush in where angels fear to tread" --> "하룻강아지 범 무서운 줄 모른다"
"Hold on a minute" --> "잠깐만 기다리세요"
"So you say?" --> "그것이 정말입니까?"

Variable form Idiom

"The same was true of _" --> _ 도 마찬가지다
"There live _" --> "_ 가 살다"
"Would you mind if *sentence?*" --> "*sentence* 해도 괜찮습니까?"
"It is time *to-infinitive*" --> "*to-infinitive* 할 시간입니다"

2.6 Transformation

For the translation, the structural information of a sentence such as tense, number, voice and etc. is reformed in view of object language. Some phrases in source sentence can be ignored and some indicators in source sentence are to be changed, and even part-of-speech can be changed for proper meaning of the object language.

Tense

The tense difference between source language and object language should be resolved. For example,
"He ***is*** ready." --> "준비가 되었다" (was)
"It ***is sold out***." --> "매진되었다" (was)

Number

The number in person should be changed for object language. For example,
"***My*** school" --> "***우리(our)*** 학교"
"***My*** parents" --> "***우리(our)*** 부모님"

Voice

The voice has to be changed in some cases. For example,

"I must ***be dressed up***." (pass. voice) --> "나는 ***정장해야*** 한다" (act. voice)

"I ***am prepared for*** anything." (pass. voice) --> "나는 무언가를 ***준비한다***" (act. voice)

Order of modification

The order of description has to be changed. For example,

"***billions of*** stars" --> "***수십억의*** 별" (stars of billions)

"***a large amount of*** products" --> "***많은 양의*** 생산품" (products of a large amount)

Part-of-speech

In the result of translation, POS can be changed. For example,

"***Nothing***!" --> "***전혀***! (Noun --> Adverb)

"***Congratulations*** on your birthday." --> "당신의 생일을 ***축하합니다***." (Noun->Verb)

Mood

A sentence mood is changed by elision of sentential element. For example,

"***You just*** cheer up." (indicative mood) --> "기운내시오" (imperative mood)

"***You just*** wait." (indicative mood) --> "기다리시오" (imperative mood)

Meaning weakening

Some insignificant words in source sentence increase the ambiguity of syntax and semantics. Such words can be ignored during the translation for object language speakers. But the ignored words may be recovered when it is necessary. For example,

"I read ***even*** a word in today's newspaper."

i.e. even, just, also, well, however, ...

The insignificant phrases can be ignored without any affection on the meaning. For example,

"What ***do you think*** the book is?"

Transformation for contrary meaning

A meaning in source language may have contrary meaning in object language. For example,

"Is this ***seat taken***?" --> "***빈자리***입니까?" (Is this seat empty?)

Simplification

The simplification of source sentence simplifies the sentence structure and sentence meaning. For example,

"***not so much as saying good bye***" --> "***saying good bye***"

3. Algorithm

The algorithms for idiom acquisition and idiom recognition are explained below.

3.1 Algorithm for idiom acquisition

The idioms are acquired from the corpus and the process of idiom acquisition consists of two phases, which include idiom discovery phase and idiom verification phase. In the first phase, the phrases which seemed to be an idiom are found in the corpus and in the second phase, the phrases which are used actually with a consistent meaning are verified in other corpus. In the idiom discovery phase, sentences in the corpus are translated one by one. If the result expression is not natural for the object language speakers, it is registered as a candidate phrase of idiom. This procedure is illustrated in Fig. 3.

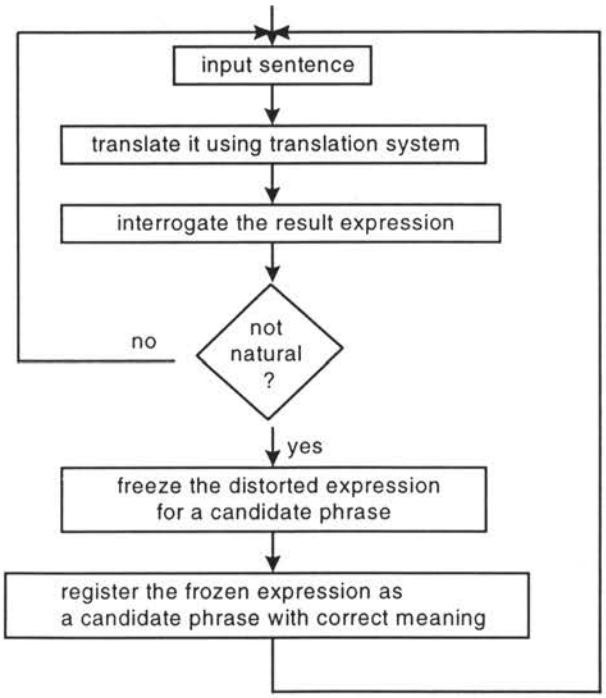


Fig. 3 Idiom discovery

In the second phase, the candidate phrases from above algorithm should be verified whether they are used with a same meaning in different corpus. After the sentences including the candidate phrase are extracted from the corpus, they are translated one by one. If the candidate phrase has a single meaning through corpus, it can be registered in the idiom dictionary as an idiom. If not, it is registered in the idiom dictionary with sentential contextual constraints. The procedure of idiom verification is shown in Fig. 4

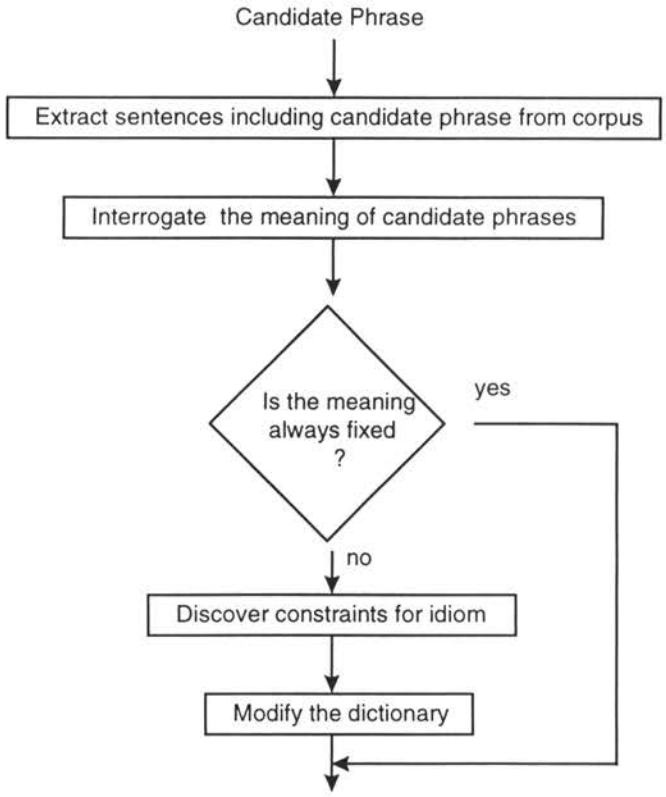


Fig. 4 Idiom Verification

The frequent constraints discovered during verification procedure are listed as below

- Collocation for objectives

The phrase is frozen with a single meaning in case of collocation for objectives. The definition of this is as follows

$$Id(W_1W_2...W_n) = yes \ if \ O(W_1W_2...W_n) \in CO(W_1W_2...W_n)$$
$$no \ otherwise$$

($Id(\alpha)$: Is α a idiom?, $O(\alpha)$: objectives of α, $CO(\alpha)$: a list of objective collocation)

For example, a verbal phrase

"live on _"

"My uncle lives on the hill." --> "언덕 위에서 살다" (literal translation allowed)

"The horse ***lives on*** hay." --> "건초를 먹고 살다" (idiom translation only)

i.e. "fall on _", "be in the mood for _" and etc

In the above, "live on" is frozen only when the objective belongs to the set, {hey, meat, corn, ...}. The phrases "drop in", "act on" are the same case.

- Part-of-speech of the successor

The phrase is frozen for certain a successor. End-of-sentence mark and preposition, or conjunction word are likely to be the successor. The definition is below.

$$Id(W_1W_2...W_n) = yes \ if \ when \ W_1W_2...W_nS..., \ POS(S) \in \{EOS, CONJ, PREP\}$$
$$no \ otherwise$$
$$(POS(\alpha) : POS \ of \ \alpha)$$

For example, a prepositional phrase

"in public"
"He is much interested in public health." --> "공중 보건" (literal translation allowed)
"He used to criticize them ***in public***." --> "***공개적으로***" (idiom translation only)
i.e. "for short", "in a second", "for sure" and etc

In the phrase "in public health", the phrase "in public" could not be frozen because it is followed by the noun (not preposition, conjunction and EO-sentence) word "health". "turn around", "at a discount" are the same case.

- Context feature of following word

The verbal phrase is frozen for certain feature of the following word. For example,

"give the ax"
"I gave the ax to cut the *tree*." --> "도끼를 주다" (literal translation allowed)
"I ***gave the ax*** to Mr. Brown yesterday." --> "***해고하다***" (idiom translation only)
i.e. "bring home the bacon", "carry the baby" and etc

In the first sentence, the verbal phrase is not frozen because it is followed by the keyword "tree". An algorithm for the resolution of homonym word sense [17] is used for this case. The phrases "money for jam", "get the picture" are in this case.

The idioms are collected from the corpus using idiom collection algorithm to construct the idiom dictionary which is implemented for our system. The size of idioms is shown in Fig.5. The conventional idiom takes the biggest portion in the dictionary, but the portions of the other idioms will be getting bigger as the corpus tests go on.

Extended Idiom	Number
Conventional Idiom	85000
Collocation	6000
Verb Pattern	30
Idiom Prep. Attach.	600
Sentence-level Idiom	500
Transformation	350
Total	92480

Fig.5 Number of Idioms by types

3.2 Algorithm for idiom recognition

The Extended Idiom in a source sentence is recognized before the sentence translation. The algorithm for idiom recognition is as following.

For input sentence $W = \{W_1, W_2, ..., W_n\}$

```
I <- 0;
while (I <= n)
{
  Id <- GetIdiomListfromDict(Wi);
  if (j <- FindMatchString (Id, Wi,n))
    k <- ProcessMatchString (Idj, Wi,n);
  I <- I + k;
}
```

The description functions used in above algorithm is below

GetIdiomListfromDict (W) : Returns all the idioms registered in word W entry in the dictionary with argument W

FindMatchString (L, WL) : Returns the position of idiom phrase in the sentence with argument L (idiom list) and WL (word list in sentence)

ProcessMatchString (L, WL) : Returns the last position of idiom phrase in the sentence with L (idiom phrase) and WL (whole sentence)

When applying this algorithm for the recognition, several levels of difficult problem are observed as below.

- **Easy recognition**

Frequent idioms in source sentence could be recognized easily. For example,

"I have ***a lot of*** books."
"I ***kept*** *his lesson* ***in my mind***."

In this case, idioms are recognized without any semantic problem.

- **Difficult recognition**

Idiom recognition becomes difficult in cases.

(1) Modifier problem
A modifier inside the idiom makes the recognition more difficult. For example, for the phrase format

$id = w_1 w_2 w_3 ... M_i w_i .. w_n$ (id = idiom string, w = word in idiom, M = modifier, and M_i modifies w_i)

A modifier M_i causes the problems for the recognition. In the idiom "take care of _", a modifier "good" for word "care" brings in some difficulties for recognition.

"I *took care of* my baby." (easy recognition)
"I *took* ***good*** *care of* my baby." (difficult recognition)

For such problems, extra idiom entry for "take good care of _" are added into the idiom dictionary.

(2) Scope problem
In the idiom format,

$id = w_1 w_2 w_3 ... Mt$, where Mt is a Meta-word

Mt has a scope ambiguity. For example,

"I ***kept company with Mr. Kim in my laboratory.***" (Mt includes prepositional phrase)
"I ***kept company with Mr. Kim*** in the United States." (Mt does not include prepositional phrase)

For the resolution, we need an extra processing for the prepositional phrase attachment determination. And the prepositional phrases would be tested for prepositional attachment after the idiom recognition.

4. Experiment

Our system performed translation of 1000 sentences from Korean high school English textbook and the length of sentences ranges to 20 words for this experiment. We used SUN Workstation SPARC-20 with

Solaris 2.4 O.S. to measure meaning correctness and efficiency for time and space for this experiment. This system with idiom and without idiom were experimented to evaluate translation correctness and efficiency of time and space. In Fig. 6, E represents excellent translation, A means excellent translation with minor deficiency and U means unacceptable translation including parsing failures. In this experiment, the meaning correctness is improved from 44.8% to 87.9% for excellent level and 93.3% (from 68.9%) is achieved by idiom implementation for understandable translation.

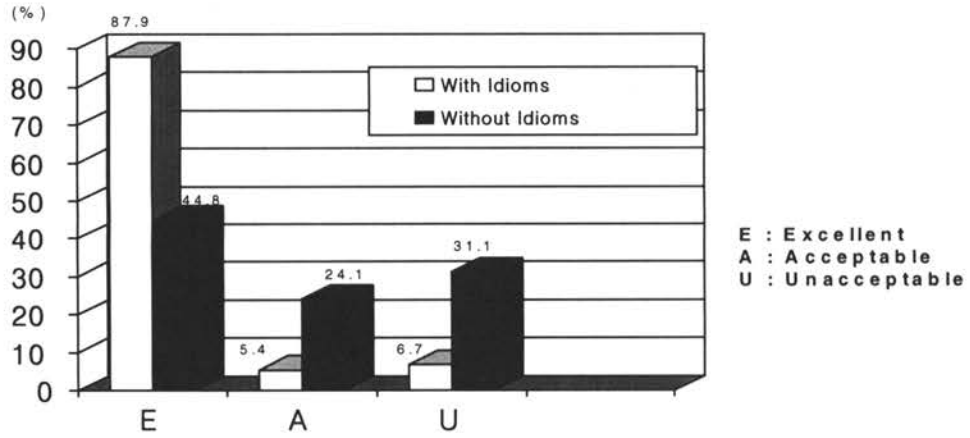


Fig. 6 Meaning Correctness

In Fig.7, the average time consumption for a sentence is reduced from 3.7 second to 2.87 second and also the space consumption is reduced from 6.47 Mbyte to 4.08 Mbyte by idiom approach.

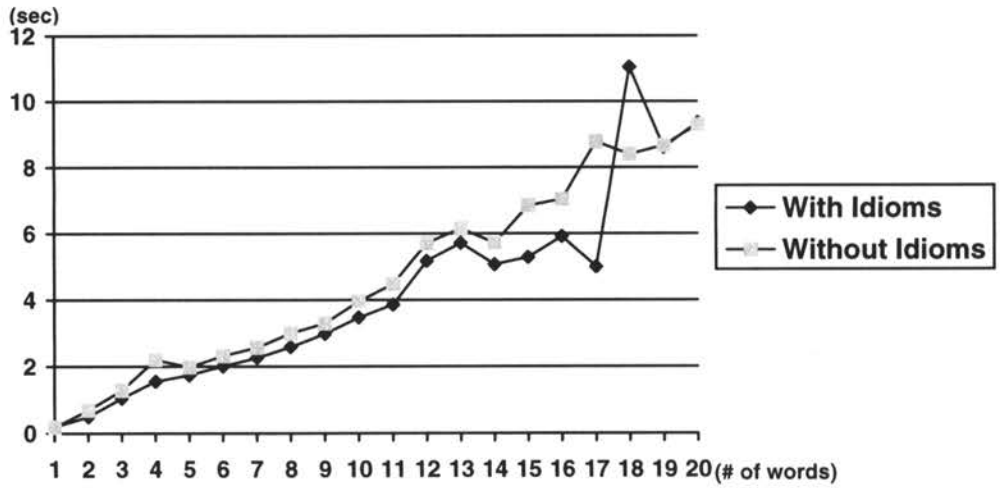


Fig. 7-a Time consumption Improvement

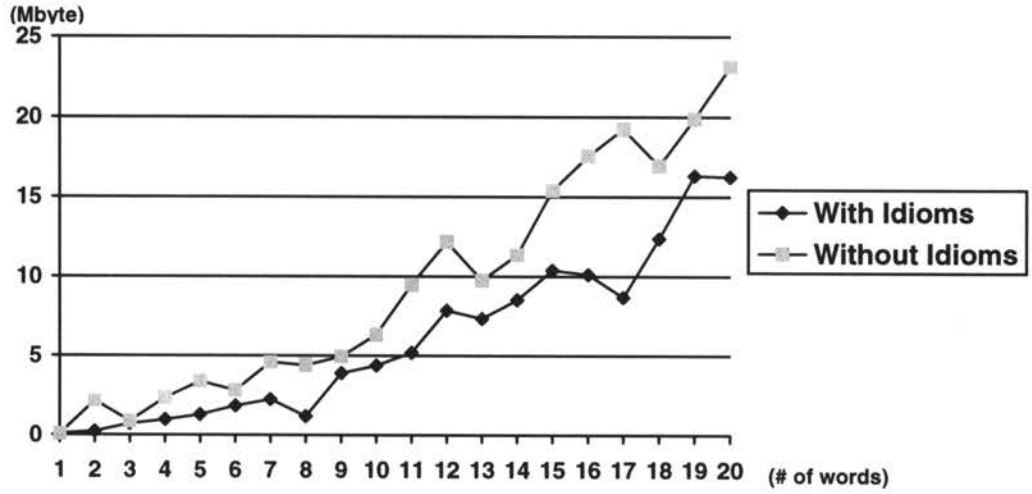


Fig. 7-b Memory Space Improvement

5. Conclusion

The Extended Idiom concept is proved to improve the meaning of the translated result and memory space and speed for the translation as shown in the previous graphs. The Extended Idiom usually has a fixed single meaning but sometimes the meaning is determined depending on the context among various possible meanings.

The system will collect more idioms and the system will be improved more as the corpus tests going on by this approach for semantics as well as for syntactic problems. And the domain-based idiom should be collected for the domain-dependent translation.

For longer sentences, more kinds of idiom are to be searched for the characteristics of longer sentences and more idioms have to be collected for improvement of the wide range of sentence translations.

reference

[1] Santos O., "Lexical Gaps and Idioms in Machine Translation," *Proc of COLING-90, pp330-335, 1990*

[2] Nakaiwa H, A. Yokoo and S. Ihehara, "A system of Verbal Semantic Attributes Focused on the Syntactic correspondence between Japanese and English," *Proc of COLING-94, pp672-678, 1994*

[3] Yoon S. and Y. Kim, "Idiomatical and Collocational Approach to English-Korean Machine Translation," *Proc of ICCPOL-92, pp56-60, 1992*

[4] Longman Dictionary of English Idioms, 1979

[5] The New World Comprehensive English-Korean Dictionary, Si-sa-yong-o-sa Inc. ,1987

[6] Essence English-Korean Dictionary, Minjungseorim, 1989

[7] Bos J., B. Gamback, and C. Lieske, "Compositional Semantics in Verbmobil," *Proc. of COLING-96*, pp131-136, 1996

[8] Lee H. and Y. Kim, "An Idiom-based Approach to Machine Translation," *Proc of fifth international conference on Theoretical and Mathodological Issues in machine translation, 1993*

[9] Alshawi H., R. Crouch, "Monotonic Semantic Interpretation," *30th annual meeting of the Association for Computational Linguistics*, pp32-39, 1992

[10] Fischer I., M. Keil, "Parsing decomposable idioms," *Proc of COLING-96*, pp388-393, 1996

[11] Kim N. and Y. Kim, "Determining Target Expression Using Parameterized Collocations from Corpus in Korean-English Machine Translation," *Proc. of PRICAI-94, pp732-736, 1994*

[12] Seidl J. and W. McMordie, Fifth edition English Idioms, Oxfort University Press., 1978

[13] Dorna M., M. C. Emele, "Semantic-based Transfer," *Proc. of COLING-96, pp316-321, 1996*

[14] Beaven J. L., "Shake-and-Bake Machine Translation," *Proc. of COLING-92, pp603-609, 1992*

[15] Frantzi K. T., S. Anaaiadou, "Extracting Nested Collocations," *Proc. of COLING-96, pp41-46, 1996*

[16] The Merriam Webster Thesaurus, Merriam Webster, 1989

[17] Seo M., "A Study on Homonym Sense Selection in English-Korean Machine Translation," *SNU master thesis, 1997*

事務局から

未納会費納入のお願い

AAMTは、会員の皆様からの会費のみを運営資金として活動を行っております。'97年度も残り少なくなりましたが、会費を未納の方は早急に納入くださいますようよろしくお願いいたします。

振込先：① 三和銀行　掘留支店　普通預金口座　3756971
東京三菱銀行　飯倉支店　普通預金口座　0121459
名義人　AAMT事務局（三和：エーエーエムティージムキョク・東京三菱：エイエイエムテイジムキョク）
② 郵便振替　00150－1－663538
名義人　アジア太平洋機械翻訳協会

新製品紹介

英日/日英機械翻訳システム ASTRANSAC® for Windows V3.0

株式会社　東芝

1．概要

東芝は、Microsoft® Windows®上で動作するパーソナル翻訳ソフトウェア「ASTRANSAC®（エイエストランザック）for Windows」をバージョンアップし、V3.0と致しました。

ASTRANSAC® for Windows は、パソコン上の電子文書を初心者でも簡単に翻訳できるソフトウェアで、対訳エディタを用いて文を一文ずつ確認して翻訳したり、ファイル指定で一気に翻訳したりすることができます。また、新製品の V3.0では、クリップボードに入った文を自動的に翻訳したり、Microsoft® WORD や一太郎、インターネットの情報をレイアウトを残したまま翻訳できる機能など、他のアプリケーションとの連携が強化されました。さらに、語句の使い方や例文を登録することにより、ユーザ固有の言い回しを翻訳できるなど、カスタマイズ機能もさらに強化されました。

2．新機能

ASTRANSAC® for Windows は、語彙遷移文法による原文解析ならびに意味トランスファ方式による言語間変換を用いることにより高品質な翻訳結果が得られ、豊富な翻訳環境設定、使いやすい対話画面などを備えた Windows®95／WindowsNT®上で動作する32bit アプリケーションです。新バージョンにおいては、上記の機能に下記の機能を拡張いたしました。

（1）クリップボード自動翻訳

Windows®のアプリケーションのコピー機能を使用して、クリップボードに翻訳したい文書に入れると、自動的に翻訳する機能を備えました。この機能を使用すると、例えば、OCR 入力データや WORD などワープロ中の文書、Microsoft® Excel など表計算ソフトの項目、プログラム中のコメント、メールなど、様々なアプリケーションとのシームレスな連携が可能となります。

翻訳結果は、翻訳結果補助画面や、クリップボードに格納することもでき、他のアプリケーションでペーストするだけで翻訳結果を挿入できます。

（2）RTF 文書・HTML 文書の翻訳

今までのテキスト文書や WORD や一太郎のドラッグ範囲の翻訳に加え、RTF 形式や HTML 形式の文書の翻訳ができるようになります。RTF 形式を使えば、WORD や一太郎の文書を、HTML 形式を使えばインターネットの文書をまるごと翻訳し、翻訳結果に原文のレイアウトを反映することができます。

（3）パターン翻訳

例文を登録すると、それを参照しながら翻訳する機能を提供します。これにより今までより自然な翻訳結果を出力することができるようになります。

（4）詳細辞書登録

言葉の使われ方を細かく登録することができ、翻訳精度を上げることができます。今までの簡単な辞書登録手段も提供しますので、英語の理解力に応じて２種類の辞書登録方式を使い分けることが可能となります。

（5）標準辞書の語数が23万語に増加＊

業界最多語数23万語の標準辞書が利用できるため、高度な翻訳を実現します。利用者は20万語までユーザ辞書を登録可能であり、別売りの専門用語辞書を組み合わせることにより専門性の高い文書の翻訳にも高い品質が得られます。

＊：1997年11月現在（当社調べ）英日のみ。日英は８万語

3．機能

翻訳：全文翻訳、一文翻訳、次文翻訳、範囲指定翻訳、クリップボード翻訳、テキストファイル一括

翻訳、RTFファイル一括翻訳、HTMLファイル一括翻訳、クリップボード自動翻訳、WORDドラッグ範囲の翻訳（＊1）、一太郎ドラッグ範囲の翻訳（＊2）

辞書：簡易辞書登録、詳細辞書登録、辞書参照、辞書削除、一括辞書登録、パターン辞書登録、ユーザ辞書登録内容一覧表示

翻訳環境：パターン辞書使用設定、原文解析パラメータ設定、訳文解析パラメータ設定、辞書環境設定、その他の項目設定

原文編集：未知語検索、並列語検索、翻訳不要句設定、並列句設定、挿入句設定、一括不要句設定（＊3）、句指定解除、品詞指定（＊3）、数指定（＊4）

訳文編集：訳語選択／学習、大文字・小文字変換（＊4）、簡易置換（＊4）

ファイル：新規作成、既存文書の読み込み、挿入読み込み、上書き保存、名前を付けて保存、文書印刷（原文、訳文、対訳形式）、文書送信

編集：移動、複写、貼付、削除、取消、全文消去、検索、置換、文挿入、文削除、文分割、文接続

その他：フォント変更、HELP

＊1：対応するWORDのバージョンは6.0/95/97

＊2：対応する一太郎のバージョンは6.0/6.3/8.0

＊3：英日のみ

＊4：日英のみ

4．動作環境

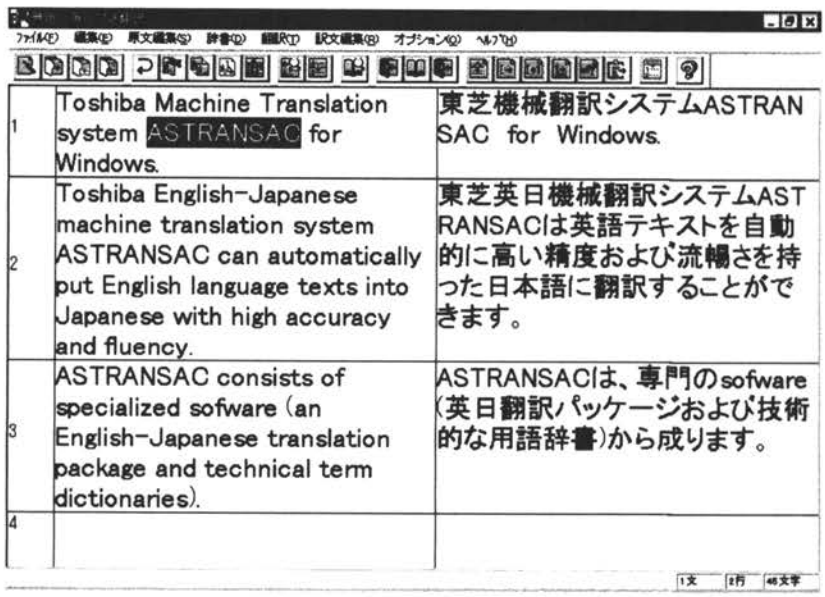
ハードウエア

対応機種	PC/AT互換機、PC98シリーズ
CPU	i486DX4，Pentium, Pentium Pro, Pentium-II
メモリ	16MB以上（推奨24MB以上）
ディスク	英日25MB以上、日英25MB以上、両方向50MB

ソフトウエア

OS	日本語Windows®95または日本語WindowsNT® SERVER/WORKSTATION3.51/4.0

（株）東芝
コンピュータ・ネットワークプロダクト事業部
〒105－8001 東京都港区芝浦1－1－1 東芝ビル
電話：03－3457－2725 FAX：03－5444－9234



1	Toshiba Machine Translation system ASTRANSAC for Windows.	東芝機械翻訳システムASTRANSAC for Windows.
2	Toshiba English-Japanese machine translation system ASTRANSAC can automatically put English language texts into Japanese with high accuracy and fluency.	東芝英日機械翻訳システムASTRANSACは英語テキストを自動的に高い精度および流暢さを持った日本語に翻訳することができます。
3	ASTRANSAC consists of specialized sofware (an English-Japanese translation package and technical term dictionaries).	ASTRANSACは、専門のsofware(英日翻訳パッケージおよび技術的な用語辞書)から成ります。
4		

ASTRANSACの対訳画面例

新製品紹介

インターネット連携日英翻訳ソフト　TransLinGO! V 1.0

富士通株式会社

1．はじめに

「TransLinGO！」は日本語がわからない外国人の方々が英語版 Windows95と Internet Explorer を利用して日本語 Web ページを英語で読むための日英翻訳ソフトです。このソフトを利用することで、英語を母国語とする方々がインターネット上の日本語の情報（企業情報、観光案内、日本文化紹介など）を英語で読んで、日本語の世界を気ままにネットサーフィンすることができます。

2．特徴

2．1　英語版 Windows95で動作

日本語環境は不要です。Internet Explorer にアドオンされている Multilanguage Support（Japanese）を利用することにより、英語版 Windows95で日本語表示が可能です。

2．2　利用形態で選択できる３つの翻訳機能

（1）ページ翻訳

Internet Explorer で表示される日本語 Web ページをレイアウトをそのままにしてページ全体を翻訳します。

（2）アウトライン翻訳

ページの概要をつかむために、ヘッダ（見出し）、リスト（項目）、リンク部分だけを翻訳します。

（3）クリップボード翻訳

翻訳させたい部分をクリップボードにコピーするだけで英語に翻訳します。翻訳結果は Internet Explorer の画面に表示されます。

2．3　高い翻訳品質

高度な意味処理方式を採用した「ATLAS」翻訳エンジンとインターネット用語を含め15万語の基本辞書で、高品質な翻訳を実現しています。

2．4　簡単な操作

Internet Explorer と連携するための複雑な設定は一切不要です。翻訳はリンクをたどるだけで自動で開始します。翻訳機能の切り替えもツールバーのボタン操作で簡単に行えます。

2．5　訳文表示の選択

設定により同一画面上で和文（原文）と英文（訳文）を交互に表示させることもできます。日本語が少し理解できる方や日本語を勉強している方は、日本語の情報を確認しながら読むことができます。

3．動作環境

CPU	i486以上
OS	Microsoft Windows95（英語版）
メモリ	16MB 以上（OS 含む）
ハードディスク	35MB 以上
必須ソフトウェア	Internet Explorer4.0（または3.02）※ Internet Explorer Add-ons : Multilanguage Support（Japanese）※

※CD-ROM 版に組み込まれています。

販売会社・価格

米国市場向け　Fujitsu Software Corporation

（http : //www.fsc.fujitsu.com/language／）

ダウンロード販売　$88（税別）

CD-ROM 通信販売　$98（税別）

日本市場向け　株式会社ジー・サーチ

（http : //shop.g-search.or.jp/manual/TLGO／）

ダウンロード販売　9、800円（税別）

CD-ROM 通信販売　10、800円（税別）

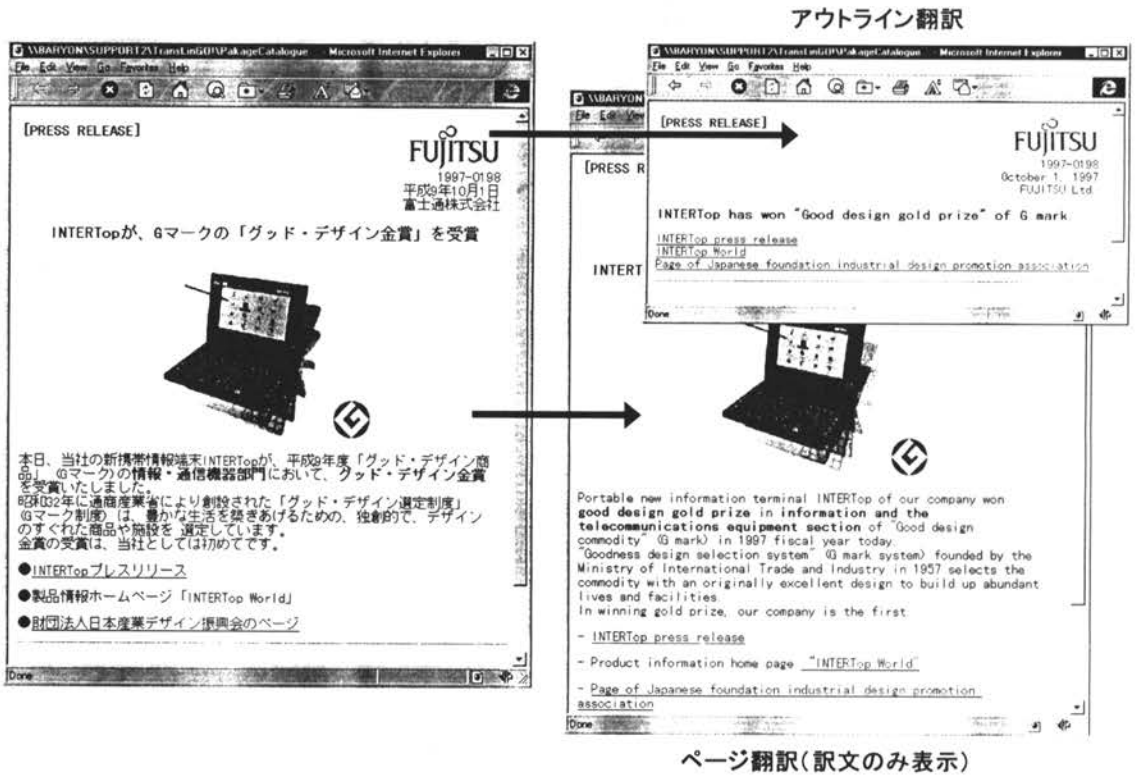
お問い合わせ先　富士通株式会社

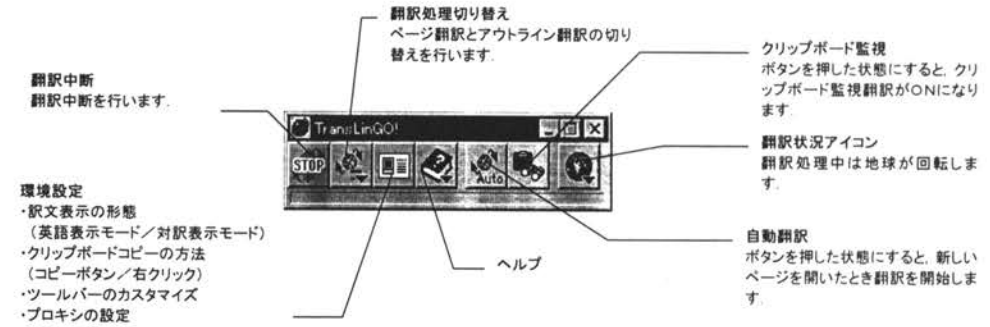
ソフトウェア販売推進部　市川

TEL：03－3730－3136

e-mail：dev_mail@softinfo.ssl.fujitsu.co.jp

富士通　TransLinGO! V 1.0





委員会だより

会員の皆様からの投稿歓迎！

編集委員会

AAMT Journalは、アジア太平洋機械翻訳協会の機関誌として、協会から会員の皆様への情報伝達のほかに、ひとりひとりの会員から他の会員への情報発信の場としての役目も大切なことだと思っております。世界がますます小さくなり、私たち機械翻訳に携わるものに新しい期待が寄せられるようになってまいりました。ぜひ、会員の皆様の声をお寄せくださるようお待ちしております。

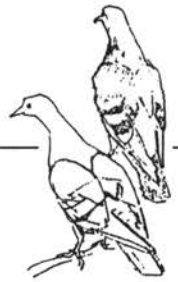
事務局だより

AAMT 前会長　京都大学教授
長尾　　真先生に三つの慶事

10月30日　第１回 IAMT 賞（IAMT Award of Honour）が長尾真前会長に授与されることとなり、アメリカサンディエゴ市での MT Summit Ⅵの会場で授賞式がとり行われました。

11月３日　恒例の秋の褒章受章者が政府から発令され、京都大学教授長尾真先生が知能情報学研究の分野で紫綬褒章を受章されました。

11月15日　京都大学の第23代総長に長尾真教授が選任され、12月16日に就任されました。



事務局だより

'96 '97年度入会のかたがた

－敬称略－

Minako O'Hagan	'96. ４月入会
小笹　和彦	'96. ９月入会
佐良木　昌	'96. ６月入会
鳴海　武史	'96. 10月入会
山岸　鉄男	'96. 10月入会
荒木　健治	'97. ７月入会
林　　典門	'97. ７月入会
藤並　稔弘	'97. ７月入会
恩地　晴美	'97. ７月入会
陸　　楽	'97. ８月入会
Yung Taek Kim	'97. ８月入会
Song Dong Kim	'97. ８月入会
Yuseop Kim	'97. ８月入会
田中　孝	'97. ８月入会
Key–Sun Choi	'97. 10月入会
福富　織枝	'97. 10月入会
草場　武司	'97. 10月入会
森本　健一	'97. 10月入会
Kaori Espanol	'97. 12月入会

事務局だより

役員・委員の交替

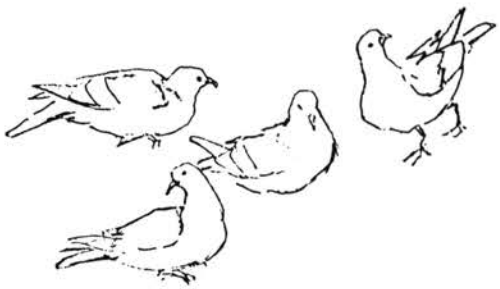
次のとおり、役員および委員の交替がありましたのでご連絡します。

理　事	沖電気工業株式会社	退任　榊　　靖夫	
		就任　牛尾真太郎	常務取締役
運営委員会	社団法人日本電子工業振興協会	退任　古沢　章	
		就任　樋口和雄	技術部兼業務情報化推進室　部長
編集委員会	日本電気株式会社	退任　田村真子	
		就任　山端　潔	C&C メディア研究所音声言語テクノロジーグループ主任
ネットワーク翻訳研究会	シャープ株式会社	退任　篠崎直子	

事務局だより

協　会　活　動　報　告

運営委員会	9月12日	①MT Summit Ⅵの参加方針　②MT Summit Ⅶの取組み　③事務局報告　④その他
	10月15日	①MT Summit Ⅵのプログラムと参加状況　②MT Summit Ⅶの方針　③AAMT の会員・会費状況　④その他
	11月21日	①MT Summit Ⅵ報告　②IAMT の新体制と田中穂積教授の会長就任　③MT Summit Ⅶの計画立案　④その他
市場動向調査委員会	9月30日	①ヒアリングの進め方　②他の委員会との交流の進め方　③事務局報告
	11月10日	①事務局報告・審議　②ヒアリング（スーパーアスキー小川編集長）
技術動向調査委員会	9月30日	①UPF 活動の進捗状況　②UPF の展示の進め方　③事務局報告
	11月11日	①UPF 活動の進捗状況　②UPF ホームページ開設　③MT Summit Ⅵ報告　④その他報告
編集委員会	10月6日	①AAMT Journal No.20の反省　②No.21以降の編集方針　③AAMT Journal の部数・配布先・費用等の検討　④MT Summit Ⅵの取扱　⑤その他報告
ネットワーク翻訳研究会	9月10日	①発足趣旨説明　②座長選出　③活動方針検討
	11月21日	①ヒアリング（長瀬産業堂野前進氏）　②MT Summit Ⅵ報告
インターネットWG	12月22日	①AAMT ホームページの運用　②'98年度の計画　③その他
シンガポールサミットWG	12月25日	①MT Summit Ⅶ準備の意見交流（CICC,JEIDA）



AAMT
ジャーナル

No. 21
(January 1998)

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 : KYN02317@niftyserve.or.jp （NIFTY）
ホームページ：http : //www.jeida.or.jp/aamt/
編集委員会　野村浩郷(委員長)　富士　秀　奥西稔幸　山端　潔
事 務 局　中島泰雄　神野千恵子
印 刷 所　伸光写植印刷株式会社