

AAMT

Asia-Pacific
Association for
Machine
Translation

Journal

No. 36

目 次

巻頭言：	「新しい機械翻訳システムを目指して」..... 長尾 真	1
研究報告：	「日本語から多種の言語への機械翻訳」..... 池田 尚志	3
解説：	「小さなテキストからの自動用語抽出」..... 中川 裕志	8
総会・関連行事：	「第 14 回通常総会および関連行事の報告」.....	12
役員リスト：	AAMT 役員名	13
活動報告：	協会活動報告 (2004 年 1 月～8 月)	14
調査報告：	「IJC/NLP-04」ならびに関連会議に参加して	16
Report：	「インドにおける MT 活用法」..... Sivaji Bandyopadhyay	25
会員投稿：	「機械翻訳に対するユーザとしての要望」..... 吉川 潔	32
MT サービス一覧：	インターネット機械翻訳サービス一覧	34
新製品紹介：	(1)「CROSSROAD for Enterprise V2」..... 日本電気(株).....	38
	(2)「The 翻訳オフィス™ V7.0」..... 東芝ソリューション(株).....	40
	(3)「専門分野 翻訳これ一本 ビジネス・科学技術編 V2」..... シャープ(株).....	42
	(4)「コリヤ英和！一発翻訳 Ver.6.0 マルチリンガル」..... ログヴィスタ(株).....	44
	(5)「MAC-Transer V2」..... (株) クロスランゲージ	46
	(6)「ATLAS V11」..... 富士通(株).....	48
MT の歴史：	「日本の機械翻訳の歴史」..... AAMT インターネット WG	51
MT サミット：	MT Summit X 開催案内	53
事務局だより：	役員交代のお知らせ.....11 AAMT取材記事のご紹介.....31 事務局長交代のお知らせ.....47	
	編集後記	52

CONTENTS

Foreword:	“Towards a New MT Paradigm”	M. Nagao	1
Report on Research:	“Machine Translation from Japanese to Various Languages”	N. Ikeda	3
Introductory Article:	“Term Extraction from Small Text”	H. Nakagawa	8
General Meeting:	Report on the 14th AAMT General Meeting and related events		12
Board Members:	AAMT Board Members		13
AAMT Activities:	AAMT Activities (January 2004~August 2004)		14
Report:	“Report on ‘IJC/NLP-04’ and associated workshops/events”	S. Yokoyama, T. Ehara	16
ク	“Use of Machine Translation in India”	S. Bandyopadhyay	25
AAMT Member:	“Request from User for Machine Translation”	K. Kikkawa	32
MT Web Links:	MT Software of Internet Service	AAMT Internet Working Group	34
New Products:	(1) “CROSSROAD for Enterprise V2”	NEC Corporation	38
	(2) “The Honyaku Office™ V7.0”	Toshiba Solutions Corporation	40
	(3) “Senmon-Bunya Honyaku-Koreippon for Business & Sci-Tech Ver.2”	Sharp Corp.	42
	(4) “KoryaEiwa! Ippatsu-Honyaku V6.0 Multilingual”	LogoVista Corporation	44
	(5) “MAC-Transer V2”	Cross Language Inc.	46
	(6) “ATLAS V11”	Fujitsu Ltd.	48
History of MT:	“History of MT in Japan”	AAMT Internet Working Group	51
MT Summit X:	Announcement of MT Summit X		53
AAMT News:	New Directors	11 MT-related article	31 New Secretary-General
	Message from the Chair of the AAMT Journal Editorial Committee	H. Isahara	52

新しい機械翻訳システムを目指して

初代 AAMT 会長 長尾 真

機械翻訳は、1960年前後から研究が始まり、あと数年で半世紀になる。ずい分長い期間にわたって研究開発がなされてきたが、その結果はどう評価すべきであろうか。

ヨーロッパ共同体（EU）では、英仏、仏伊、英独などの言語相互間の機械翻訳を行っており、翻訳の満足度がかなり高く、後修正も簡単にすますことのできるシステムとなっている。米国では英西について良いシステムを持っている。日本では韓国語との間の翻訳が実用としてかなりの程度のもとなっているが、最も力を入れている日英、英日ではまだまだ一般の人達に満足感を与えるところまで行っていない。

日本語と英語では文の構造が全くちがうし、単語の意味・用法についても大きな差があって、文構造の変換と単語の置き換えだけでは不十分である。したがって理想的なことを言えば、英語（あるいは日本語）の文を意味表現にまでおとし、それを日本語（あるいは英語）の表現として半ば独立に生成するという過程が必要になるだろう。これは人間の行っている優れた翻訳そのものであるから機械的に実現することは容易ではない。

このような理想の方法は別として、機械が行えることで、現在の機械翻訳よりも良い方法は何かをよく考えることが必要である。意味を担う基本的単位は単語であるといつてよい。しかし1つの単語単位では意味を確定できない場合がいろいろとある。その代表例は慣用句である。そこで慣用句はこれを1つの単位とみなして扱い、訳語（あるいは訳表現（句））を与える。多義の単語の意味の決定は、文中の他の単語との関係で決めることになるが、慣用句のように固定的表現でないところに問題がある。

良質な翻訳表現を作るためには、単語ごとの翻訳の構文的組合せよりは、慣用句表現をはじめとしてもっと長い表現を単位とする方がよい。いわゆるチャンクと呼ばれている単語列や句単位の翻訳を行うのがよいだろう。しかし、単語だけでも少なくとも数万語あるから、名詞+前置詞+名詞という組み合わせの句を考えるだけでも数百万では納まらない数となる。これらの中で句単位で訳さないと良い表現にならないものの典型例を用例辞書に登録して、他の類似表現は用例表現の翻訳例を参考にして翻訳することになる。

文全体の翻訳をこのような方法ですべて行うことはできないから、句単位の用例翻訳を他の言語の対応する文構造にしたがって並べかえるという方法をとる。すなわち部分的には用例主導翻訳を行いながら、構文的な変換とを組み合わせることを重層的に行うシステムである。どの部分を用例主導翻訳にし、どこを構文翻訳

で処理するかは固定的でなく柔軟な判断に基づく必要があり、このような柔軟な処理システムを作ることが大切な課題となるだろう。

それにしてもどれだけの数の用例句の翻訳対を用意しなければならないだろうか。これを人手で全て整備することは難しい。したがって、翻訳対のテキスト、いわゆるパラレルテキストの巨大なデータベースを用意し、そこから自動的に最小単位の用例句対を取り出すという作業を重層的に行う必要がある。これを自動的に行うプログラムを開発しなければならない。まずはある専門分野の文章の翻訳をすることにし、その分野の翻訳対テキストデータベースを作ることから始めるのがよいだろう。

最近では照応関係、省略語句の推定、文脈関係の把握等においても計算言語学的にある程度の研究成果が得られているので、これをさらに精密に研究し、上記の翻訳システムに組み込むことをすれば、これまでの翻訳方式では越えられなかった壁を打ち破ることができるのではないだろうか。

1980年代の翻訳方式が一定の成果をあげながらも大きな壁につきあたっている今日、これを乗り越える新しい翻訳方式を考えねばならない時に来ているが、上記のような方法はその1つの道だろうと考えている。これまでの方式とは違ったより良い方式をもっといろいろと提案していただき、それらの可能性についてよく検討し、有力な2、3の方式について相互協力によって一段と優れた機械翻訳システムを作る努力をする時が来ているのではないだろうか。

特にこれからは東アジア諸国の言語を積極的にとりあげて機械翻訳システムを作る努力をする必要があるだろう。

日本語から多種の言語への機械翻訳

岐阜大学 池田 尚志

1. はじめに

我々の研究室では、日本語からいろいろな言語への翻訳を行う機械翻訳エンジン *jaw* の開発を行っており、そのエンジンを用いていくつかのアジアの言語（中国語、ベトナム語、ミャンマー語、シンハラ語）への機械翻訳システムの開発を試みている。また、日本語テキストから手話テキストへの機械翻訳についての試みも始めた。アジアの言語への翻訳はアジアからの留学生と協同しての研究である。未だ辞書の規模も小さくプロトタイプシステムの段階である。

2. 機械翻訳エンジン *jaw* と翻訳システム

翻訳エンジン *jaw* の概要を図 1 に示した[8]。 *jaw* は、現在研究の主流となっている用例翻訳方式や統計的翻訳方式ではなく、表現パターン翻訳規則との照合による変換と、変換によって作り出された目的言語の内部表現（我々は目的言語の表現構造と呼んでいる）からの目的言語の生成（我々は表現構造からの線状化と呼んでいる）という、いわば古典的なトランスファー方式の枠組みである。変換規則には図 2 に示す 3 つのタイプを設けた。図中の二重の枠で囲った語がパターン記述のキーとなる語であり、矢印で結ばれている部分がパターンを制約する条件となる語句である。これらの規則は RDB（関係データベース）に格納されている。パターン照合において翻訳に必要な範囲での意味解析を行うことになる。複数の照合結果がある場合には、ある指標で最適な照合結果が選択される。（base type の規則で広く照合する規則を優先、字面で照合する規則を優先、など）

パターンには目的言語への翻訳に必要な表現部品が対応させられており、我々のシステムではこの変換規則は、一つ一つが目的言語の表現構造を作り出すプログラムに書き換えられ、コンパイルされてライブラリプログラム(dll)として格納されている。目的言語の表現構造は C++ のオブジェクトを利用して実装されており、変換規則のプログラムを再帰的に実行すると図 3 のようなオブジェクトのネットワークが作り出される。これがすなわち表現構造であり、必要な表現部品はこのオブジェクトのメンバーとして表されている。（図 3 は、「付き合ってみると彼は面白い男だった」という日本語に対するシンハラ語の表現構造である。）

表現構造から文を作り出す線状化は、我々のシステムでは C++ のクラスメソッドとして実装されており、そこに目的言語の構造を反映した生成規則が組み込まれている。目的言語毎に適切なクラスを設計し生成規則を書く。

なお日本語では、用言等に後接する機能語（テンス・アスペクト・モダリティ等）や体言等に後接する機能語（取立て）が重要な役割を受け持つが、これらに関連する翻訳処理については、先のパターン翻訳規則とは別に表形式の翻訳規則を設けている。その処理結果も目的言語の表現構造の中に書き込まれる。図 4 にミャンマー語の場合の例を示す。

現在、*jaw* を用いて中国語、ベトナム語、ミャンマー語、シンハラ語への機械翻訳システムの構築を試みている。中国語を除けば、これらの言語に関する機械処理の観点（特に日本語との関わりで）からの研究・文献はほとんどない。我々は、これら

の言語の文法構造を整理し、日本語の構造との対照を分析しながら、第 1 段階として辞書の用例文等から集めた 200 文ほどのサンプル文の翻訳を jaw の上で実現することを行ってきた。現在その段階をほぼクリアしたというレベルである。言語的内容に関しては、中国語に関して取立て詞[12]、時

制[11]、否定表現[3]、場所表現[7]等の翻訳規則について、シンハラ語、ミャンマー語に関して用言後接機能語の翻訳規則について[1][5]、ベトナム語に関して名詞句の翻訳規則について[6]等々の研究を行ってきた。

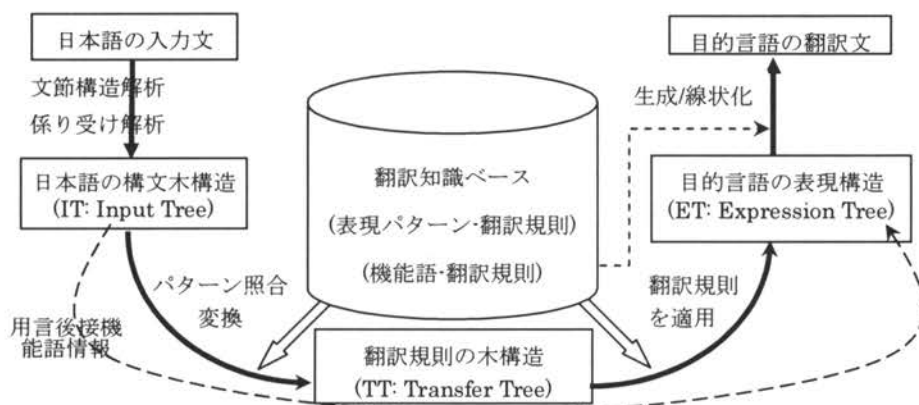


図 1 機械翻訳エンジン jaw

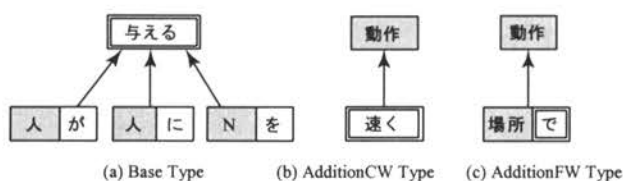


図 2 パターン翻訳規則の 3 つのタイプ

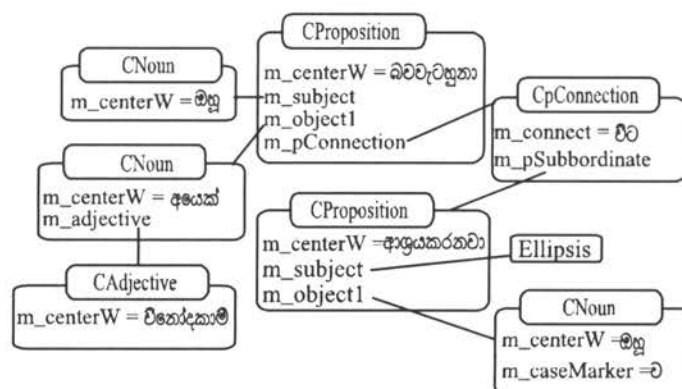


図3. 目的言語の表現構造 (jaw/Sinhalese)

		日本語 (K=かもしれない)	ミャンマー語 (H= ဖြစ်)	
肯定	未	V+る+K	V+လျှင်+V	①
	来	V+たい+K	V+ချင်+လျှင်+V+ချင်	②
	非	V+た+K	V+ခဲ့+လျှင်+V+ခဲ့	③
			V+ခဲ့+တာ+H+လျှင်+H	
	未	V+たかった+K	V+ချင်ခဲ့+လျှင်+V+ချင်ခဲ့	④
	来	V+ている+K	V+နေတာ+H+လျှင်+H	
	来	V+ていた+K	V+နေခဲ့တာ+H+လျှင်+H	
否定	未	V+ない+K	V+ချင်မှ+V	⑤
	来	V+たくない+K	မ+V+ချင်တာ+H+လျှင်+H	⑥
	非	V+なかつた +K	V+ခဲ့+ချင်မှ+V+ခဲ့	⑦
		V+たくなかつた	မ+V+ချင်ခဲ့တာ+H+လျှင်+H	⑧
	未	V+ていない+K	မ+V+နေတာ+H+လျှင်+H	
	来	V+ていなかった	မ+V+နေခဲ့တာ+H+လျှင်+H	

彼 が 行っ たかもしれません。

図 4. 用言後接機能語の翻訳規則 (jaw/Myanmar)

3. 手話への機械翻訳について

手話も 1 個の独立した言語である。とすると、同じように機械翻訳の構図が成り立つはずである。我々は同じ jaw の枠組みで、日本語テキストから手話テキストへの機械翻訳の試みも始めた (jaw/SL)[2][4]。この場合まず問題になるのは、手話をテキストに書き取る表記法が確立していないこ

とである。つまり文字が未だ無い。いくつかそのような試みはあるが、動作を書き取る方に力点があるようで、機械翻訳など言語処理には向いていない。我々は表記法に関する提案からはじめて、手話の教科書等にある日本語一手話の対訳例文を jaw で実現することを行っている(図 5)。

1. 「本は、私が買う」	{<topic> 本} 私 買う。
2. 「あなたの手話の先生は誰ですか」	{<topic> あなた 手話 先生}{<whq> 誰 PT(x)}?
3. 「雨なら車、晴れなら自転車です」	{<cond> 雨} 自動車, {<cond> 晴れ} 自転車。
4. 「あなたは佐藤さんですか」	{<q> あなた 佐藤 あなた} ?
5. 「読みなさい」	読む PT(2; 命令) !
6. 「彼に読ませる」	読む 彼(x) 与える(→ x)。
7. 「本を読みながら食事をする」	{<gaze(x)> 本(x) 食べる}。

図 5. 手話への機械翻訳 (jaw/SL)

4. おわりに

我々はこのほかに、いわゆる翻訳という範疇ではないが、日本語テキストを点訳(点字への翻訳)するシステム IBUKI-TEN を開発しており、こちらは実用レベルのシステムとしてホームページ上で公開し多くの方々に利用して頂いている[10]。jaw

の方は未だ実用レベルからはずいぶん遠い。次の課題の一つとして大域パターンの翻訳規則を取り入れることを考えている。辞書を少しずつでも拡充していくこともどうしてもやらなければならない。母語話者との協同を模索しつつ取り組んでいくつもりである。

文献

- [1] Ma Ngin Khaw Cing, 福本真哉, 池田尚志, 日本語-ミャンマー語機械翻訳システム jaw / Myanmar における述語部構造の翻訳について, FIT, pp139-142, 2004
- [2] 松本忠博, 田中伸明, 吉田鑑地, 谷口真代, 池田尚志, 手話の表記法とテキストレベルの日本語-手話機械翻訳システムの試みについて, 電子情報通信学会, 思考と言語研究会, TL2004-12, pp7-12, 2004
- [3] ト朝暉, 池田尚志, 日中機械翻訳における否定文の翻訳, 自然言語処理, 11-3, pp97-122, 2004
- [4] T. Matsumoto, N. Tanaka, A. Yoshida, Y. Imai, T. Ikeda, The First Step Toward a Machine Translation System from Japanese Text to

- Sign Language, IJCLP-04, Companion Volume, pp61-66, 2004
- [5] S.Thelijjagoda,Y.Imai,N.Elikewala,T.Ikeda, Japanese-Shinhalese MT System (jaw/Shinha-lese) and a Solution for the Translation of Japanese Function Words after Predicate into Shinhalese, IJCNLP-04, Companion Volume, pp73-78, 2004
- [6] C.Nguyen,Y.Imai,T.Ikeda, Translation of Adnominal Embedding Clause in Japanese -Vietnamese Machine Translation, IJCNLP-04, Companion Volume, pp41-44, 2004
- [7] 宇野修一,ト朝暉,今井啓允,池田尚志, 日中機械翻訳システム jaw/Chinese における連体修飾表現・場所表現等の処理, 言語処理学会第10回年次大会論文集, pp640-643, 2004
- [8] 今井啓允,池田尚志, オブジェクト指向言語のパラダイムを利用した機械翻訳エンジン jaw, pp125-128, 言語処理学会第10回年次大会論文集, pp125-128, 2004
- [9] 謝軍,今井啓允,池田尚志, 日中機械翻訳システム jaw/Chinese における変換・生成の方式, 自然言語処理, 11-1, pp43-80, 2004
- [10] 高松大地,岸井謙一,伊佐地和哉,松本忠博,池田尚志, 自動点訳システム IBUKI-TEN と新点字規則への対応, 言語処理学会第10回年次大会論文集, pp401-404, 2004
- [11] 謝軍,ト朝暉,池田尚志, 日中機械翻訳におけるテンス・アスペクトの処理, 自然言語処理, 10-4, pp177-200, 2003
- [12] ト朝暉,謝軍,池田尚志, 日中機械翻訳におけるとりたて表現の翻訳について――「も」「さえ」「でも」――, 自然言語処理, 9-5, pp111-130, 2002

小さなテキストからの自動用語抽出

東京大学 中川 裕志

1. 背景

日々、新たに起こる多数の事件やイベント、また学問や技術の急速な発展に伴い、新たな概念を表現する用語は増大の一途を辿っている。機械翻訳においては、このような新しい用語の対訳辞書が必要になるが、そのためにはまず新しい用語を取り出す必要がある。では、どこから取り出すか？大規模なコーパスから取り出すのが普通だが、コーパスサイズが大きくなると人間が取り出すわけにはいかなくなる。小さなコーパスであっても人手がかかる。そこで重要な用語を計算機で認識して抽出する自動用語抽出への要請が生まれる。

自動用語抽出は、以下の流れになる。

コーパス → 用語候補の抽出と枚举 → 用語候補の重み付け → 用語選択

用語候補の抽出と枚举では、コーパスを形態素解析し、適当な品詞列を指定して、それに一致する形態素列を用語候補として取り出す。日本語の場合は名詞（あるいは未知語）の連鎖を抽出することが多い。名詞連鎖とはすなわち複合名詞である。

用語候補の重み付けでは、取り出した個々の用語候補に対して、用語の重要度を反映した何らかの定量的な重みを付ける。以下で主に議論するのはこの問題である。用語候補は重みの大きい順に並べる。ただし複合名詞の場合は単名詞の頻度よりは込み入った状況となる。

用語選択は、上記のように重みの順に並んだ用語候補から対訳辞書に登録したい用語を選択するが、最終的には人間の判断によらざるをえない。よって、ここではこの問題は扱わない。

大規模なコーパスからの重要な用語の自動抽出（重み付け）は当然、統計的手法によって行われる。この手法は辞書に記述すべき用語の選択に近い技術であるため辞書学でも重視された。文献[1]には従来の用語重み付けの方法がまとめられている。用語重み付けは情報検索におけるキーワード重み付けとも類似している。全文検索が当たり前になり、文書中に現れる全ての単語をキーワードにできるようにはなった。それらの単語を用いるベクトル空間モデルにおいては個々のキーワードに重みを与えて、これを検索時に文書ランキングのために用いる。情報検索において頻繁に用いられる $tf \times idf$ ¹ という重みは用語重み付けにおいても有力な方法である。

$tf \times idf$ のような出現頻度を利用する方法はコーパスサイズが大きい方ほど信憑性が高いと考えられるため、用語抽出の情報源となるコーパスはサイズが大きいほど良いと考えられた。また、 idf の観点からは網羅性の良いコーパスのほうが分野に固有の単語の重要度を上げることができて良いと考えられる。しかし、この常識的な考えには次の2点で疑問がある。

- (1) 大規模で網羅性の良いコーパスは収集に時間と手間がかかる。
- (2) コーパスサイズが大きいと、狭い分野で少ない出現頻度だが重要な用語を取り出しにくくなる可能性がある。

こういった背景に基づき、比較的小さなコーパスからの自動用語抽出も視野に入れて、用語重み付けの技術について眺めてみることにする。

*1 $tf \times idf$ は、ある文書において出現頻度 (tf) が大きく、かつ少数の文書にしか出現しないという性質 (idf) を持つ単語に高い重みを与える重要度である。

2. 分野性とコーパスサイズの関係

対象となるコーパスの性質で重要なのはサイズと分野である。これを次の表にまとめた。

表1 コーパスサイズと分野

	コーパスサイズ 大	コーパスサイズ 小
多分野を網羅	(A)	(B)
特定の分野、 特定のトピック	(C)	(D)

この表の分類に基づいて用語重み付けの方法を概説する。

(A)：多分野網羅、コーパスサイズ大の場合は、従来の大規模コーパスを対象にする統計的自然言語処理の方法が有力である。第1回 NTCIR の用語抽出タスクでは、 $tf \times idf$ による重み付け方法が良い成果をあげている。1. 節で述べたように複合名詞を対象にする、あるいはより広くコロケーションを対象にする場合は工夫が必要である。この目的を狙う代表的な方法として Xtract[3]、C-value[4]などがある。[3]はさらにコロケーションの対訳の研究にも発展した[5]。

(B)：多分野網羅、コーパスサイズ小の場合は、個別分野のコーパスが非常に小さく、意味のある結果は得られない。

(C)：特定分野でもコーパスサイズが大きければ、(A)と同様の手法が有力であるため、最近ではゲノムなど他分野へ応用する試みがなされている。

(D)の問題は、分野依存性の高い少量のテキストからの抽出であり、これまであまり試みられてこなかった。そこで以下で少量テキストからの用語抽出についてのポイントをまとめる。

3. 少量テキストからの用語抽出のポイント

少量のテキストとはいえ、メガバイトに近い大きさがあり、対象の分野が絞られているなら表1の(C)と似ているので語彙の出現頻度に重点を置く方法が効果的であろう。しかし、極端に小さなテキスト、

たとえば単一の Web ページを対象とするとなると多くの用語候補は出現頻度が1であったり、多くとも2程度で、出現頻度に全幅の信頼をおけない。前述の C-value[4]の場合は、コロケーションの入れ子構造に注目して出現頻度以外の情報を勘案した。具体的にはコロケーション Wall Street がより1単語長いコロケーション Wall Street Journal などに含まれる場合の C-value は次式となる。

C-value(Wall Street)

$$= \text{freq}(\text{Wall Street}) - n(\text{Wall Street}) / t(\text{Wall Street})$$

ただし、 $\text{freq}(A)$ はAの出現頻度、 $n(A)$ はAより1単語長いコロケーションの出現頻度、 $t(A)$ はAより1単語長いコロケーションの種類数である。よって、Wall Street が常に Wall Street Journal という形で現れるなら、 $n(\text{Wall Street}) = \text{freq}(\text{Wall Street})$ かつ $t(\text{Wall Street}) = 1$ であるため、 $n(\text{Wall Street}) / t(\text{Wall Street}) = \text{freq}(\text{Wall Street})$ となって Wall Street の C-value はゼロ、一方、Wall Street より1単語長いコロケーションがすべて異なるなら、 $n(\text{Wall Street}) / t(\text{Wall Street})$ は1になり C-value は $\text{freq}(\text{Wall Street})$ になり大きい。これはあるコロケーションとより長いコロケーションとの関係を使っている。すなわち出現頻度だけではなく語彙の構造を反映した尺度になっている。なお、ここで語彙とは用語の集合を意味する。したがって、小さなテキストでも、コロケーションあるいは複合語の成分となる単語(あるいは複合語)の用語としての重要さを測れる重み付けになっている。このようにテキストが小さく出現頻度だけに頼れない場合は、同じ統計量といっても語彙の構造を利用する方法が有力である。我々の実験でも C-value は小さなテキストで優秀な性能を示すことが確認された。

4. 言選 Web での実験

小さいテキストからの用語抽出における重み付け

で語彙の構造を利用する考え方をより徹底した方法が言選 Web[6]で用いられている重み付け方法である。たとえば、ある文書から以下のような用語候補が抽出されたとしよう。

形態素、形態素、英語形態素、日本語形態素、形態素レベル、形態素解析、形態素認識

「形態素」という単語は多くの複合名詞を構成しているため、この文書の扱う分野では重要な概念を表す名詞であると考えられる。そこで、「形態素」の重みを「形態素」がより長い複合語の形成に寄与したかで測る。具体的には、形態素の左側に接続する名詞の数：L(形態素)と右側に接続する名詞数：R(形態素)を用いる。上の例では、L(形態素)=2、R(形態素)=3 となる。これをまとめて、LR(形態素) $= (L(形態素) \times R(形態素))^{1/2} = (2 \times 3)^{1/2}$ とする。N個の単名詞が連接する複合語のLR重みは、長さNに依存しないほうが望ましい。よって、LR重みの相乗平均とする。例えば、「形態素解析」のLRは

$LR(形態素解析) = (LR(形態素) \times LR(解析))^{1/2}$
ところでこの重みは「形態素」という単語が他の複合語の一部ではなく、単独で出現した頻度は利用していない。そこで、これも乗じてFLRという重みを用いた。例えば、「形態素」の場合、上の例で「形態素」の単独出現回数が2なので

$FLR(形態素) = freq(形態素) \times LR(形態素)$
 $= 2 \times (2 \times 3)^{1/2}$ となる。

このサービスは約1年間、[6]で一般に提供しているが、月間1000アクセス程度、およびソフトのダウンロードが月間50本程度とかなり使われている。アルゴリズムの詳細は[7]を参照されたい。

一例として、この文書の以上の部分に言選 Web を適用して重みの大きい順に用語をいくつか挙げる。

用語、形態素、Wall Street、用語候補、コーパス、コーパスサイズ、出現頻度、
用語抽出、用語重み付け、分野、重み付け、自動用語抽出、方法、名詞、抽出、重み、複合名詞、サイズ、LR、用語選択、形態素解析、コロケーション、テキスト

これらはこの文書の特徴付けるものと思われる。ちなみに重みのもっとも小さい用語は、

前述、形、考え方、性質、要請、視野

であり、この文書の性質を特定するようなものではないことが分かる。この方法もC-value法も形態素解析システム以外にはなんらの言語資源(辞書、ソーラスなど)を必要しない点も特記できる。

5. まとめ

従来の大規模コーパスを対象にした統計的手法による自動用語抽出とは少し違う視点として、小さなテキストを対象にした自動用語抽出について考察し、有力な方法を紹介した。

参考文献

- [1] Kageura, K. and Umio, B. 1996. "Methods of automatic term recognition: a review". *Terminology* 3(2), 259-289.
- [2] Kageura, K. et al, 1999. "TMREC Task: Overview and Evaluation". In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, 411-440.
- [3] Smadja F, 1993, "Retrieving collocations from text:Xtract", *Computational Linguistics*, 19(1), 143-177
- [4] K.T. Frantzi, and Sophia Ananiadou, 1996, "Extracting nested collocations", 16th COLING, 41-46

- [5] Frank Smadja and Kathleen R. McKeown and Vasileios Hatzivassiloglou, 1996, "Translating Collocation for Bilingual Lexicons: A Statistical Approach", Computational Linguistics, 22(1), 1-38
- [6] 言選 Web
<http://gensen.dl.itc.u-tokyo.ac.jp/>
- [7] Hiroshi Nakagawa, Tatsunori Mori: 2003. "Automatic Term Recognition based on Statistics of Compound Nouns and their Components", Terminology, 9(2), 201-219

事務局だより

役員交代のお知らせ

以下のように AAMT 役員が交代になりました。(P13 をご参照ください。)

理事 (旧)	古川 一夫 氏	株式会社 日立製作所
	↓	
理事 (新)	篠本 学 氏	〃 〃
監事 (旧)	戸田 宗孝 氏	社団法人 電子情報技術産業協会
	↓	
監事 (新)	小笠原 一晃 氏	〃 〃

第 14 回 通常総会および関連行事の報告

当協会の第 14 回通常総会が、6 月 18 日（金）午後 1 時から神田駿河台、総評会館で開催され、2003 年度事業および決算、2004 年度事業および予算案等について、滞りなく承認が得られた。

総会后、各委員会からの報告会と東京工業大学・田中穂積氏（元 AAMT 会長）をはじめ 3 名の講師による講演会が盛況のうちに終わった。

【報告会】井佐原均委員長（技術動向調査委員会）、坂本義行委員長（市場動向調査委員会）、および横山晶一座長（ネットワーク翻訳研究会）より 2003 年度の活動について報告があり、続いて富士秀 WG リーダ（インターネット WG）により AAMT の HP 関連の紹介がなされた。

【講演会】講演 I. 「機械翻訳システムのブレイクスルーを求めて」 講師 田中 穂積 氏（東工大）
MT の歴史、現在の状況、問題点などにふれながら、ブレイクスルーの必要性、今後の展望などについて述べられた。

講演 II. 「音声言語翻訳技術の応用」 講師 奥村 明俊 氏（NEC）
音声言語翻訳技術によって実現した旅行会話自動通訳システムの事例紹介があり、課題や今後の発展性について述べられた。

講演 III. 「特許翻訳における機械翻訳の活用」 講師 川上 溢喜 氏（Japio）
Japio で運用を開始し、機械翻訳を活用した特許要約文の翻訳支援システムの紹介と特許分野特有の機械翻訳システムに求められる機能について提案がなされた。

【懇親会】講演会終了後、会場を JEITA に移して行われたが、多数の参加者があり有意義な交流の場となった。

行 事	内 容	時 間
総 会	第 14 回 通常総会 1. 開会の辞 2. 会長挨拶 東京大学・辻井潤一 氏 来賓挨拶 経済産業省 商務情報政策局 情報通信機器課 課長補佐・平井淳生 氏 3. 出席会員の確認 4. 議事録署名人の選出 5. 議案 第 1 号議案 2003 年度事業報告(案) 第 2 号議案 2003 年度決算報告(案) 第 3 号議案 2004 年度事業計画(案) 第 4 号議案 2004 年度収支予算(案) 第 5 号議案 役員改選 (案) * その他、会員提案事項、 6. 報告事項 7. 閉会の辞	13:00~13:45
報 告 会	各委員会からの 2003 年度活動報告： 技術動向調査委員会、市場動向調査委員会、 ネットワーク翻訳研究会、インターネット WG	14:00~15:00
講 演 会	I. 「機械翻訳システムのブレイクスルーを求めて」 東京工業大学 田中 穂積 氏 II. 「音声言語翻訳技術の応用」 日本電気株式会社 奥村 明俊 氏 III. 「特許翻訳における機械翻訳の活用」 財団法人日本特許情報機構 川上 溢喜 氏	15:15~17:15
懇 親 会	懇親会：JEITA305 会議室（立食パーティ形式による懇談）	17:30~19:30

* 第 5 議案：新役員選出後、臨時理事会が開催され、会長、副会長が再選された。

【AAMT 役員名 / AAMT Board Members】

2004 年 6 月現在

- 会 長
President 辻井 潤一 (東京大学大学院 教授)
Jun-ichi Tsujii (Professor, University of Tokyo)
- 副会長
Vice President 小谷 泰造 (株式会社インターグループ 代表取締役社長)
Taizo Kotani (President, Inter Group Corporation)
- 理 事
Director 長尾 真 (独立行政法人情報通信研究機構(NICT) 理事長)
Makoto Nagao (President, National Institute of Information and Communications Technology)
- 理 事
Director 田中 穂積 (東京工業大学大学院 教授)
Hozumi Tanaka (Professor, Tokyo Institute of Technology)
- 理 事
Director 石崎 俊 (慶應義塾大学 教授)
Shun Ishizaki (Professor, Keio University)
- 理 事
Director 横山 晶一 (山形大学 教授)
Shoichi Yokoyama (Professor, Yamagata University)
- 理 事
Director 飯田 仁 (東京工科大学 教授)
Hitoshi Iida (Professor, Tokyo University of Technology)
- 理 事
Director 井佐原 均 (独立行政法人情報通信研究機構(NICT) 自然言語グループリーダー)
Hitoshi Isahara (Group Leader, Computational Linguistics Group, NICT)
- 理 事
Director ^{チャイ キソン}崔 杞鮮 (韓国 KAIST 教授)
Key-Sun Choi (Professor, Korea Advanced Institute of Science and Technology)
- 理 事
Director ソンラートラムワーニッチ・ウィラット (タイ自然言語ラボラトリー(TCL) ディレクター)
Virach Sornlertlamvanich (Director, Thai Computational Linguistics Laboratory)
- 理 事
Director 河村 進介 (東芝ソリューション株式会社 取締役社長)
Shinsuke Kawamura (President & CEO, Toshiba Solutions Corporation)
- 理 事
Director 前山 淳次 (富士通株式会社 取締役専務/経営執行役専務)
Junji Maeyama (Member of the Board/Corporate Executive V.P., Fujitsu Ltd.)
- 理 事
Director 鷺塚 諫 (シャープ株式会社 顧問)
Isamu Washizuka (Adviser, Sharp Corporation)
- 理 事
Director 笠原 裕 (日本電気株式会社 NEC ソリューション開発研究本部長)
Yutaka Kasahara (Executive General Manager, Solution Development Laboratories, NEC Corporation)
- 理 事
Director 篠本 学 (株式会社日立製作所 執行役常務)
Manabu Shinomoto (Corporate Officer, Hitachi, Ltd.)
- 理 事
Director 杉本 晴重 (沖電気工業株式会社 常務取締役)
Harushige Sugimoto (Senior Vice President, Oki Electric Industry Co., Ltd.)
- 理 事
Director 吹譯 正憲 (社団法人電子情報技術産業協会(JEITA) 専務理事)
Masanori Fukiwake (President, Japan Electronics and Information Technology Industries Association)
- 監 事
Auditor 小笠原 一晃 (社団法人電子情報技術産業協会(JEITA) 常務理事)
Kazuaki Ogasawara (Executive Vice President, JEITA)
- 監 事
Auditor 勝田 美保子 (株式会社十印 代表取締役会長)
Mihoko Katsuta (CEO, Toin Corporation)

協 会 活 動 報 告

(2004 年 1 月～2004 年 8 月)

第 14 回通常総会	6 月 18 日	①03 年度事業報告 ③04 年度事業計画 ⑤役員改選	②03 年度決算報告 ④04 年度収支予算 ⑥その他
成果報告会	6 月 18 日	①技術動向調査委員会 ③ネットワーク翻訳研究会 (AAMT/Japio 研究会) ④インターネットワーキンググループ	②市場動向調査委員会
講演会	6 月 18 日	①「機械翻訳システムのブレイクスルーを求めて」 東京工業大学 田中穂積 氏 ②「音声言語翻訳技術の応用」 NEC 奥村明俊 氏 ③「特許翻訳における機械翻訳の活用」 Japio 川上溢喜 氏	
懇親会	6 月 18 日	講演会終了後、JEITA で開催	
理事会	3 月 18 日	①04 年度事業計画案 ③決算理事会および総会日程 ⑤その他	②04 年度収支予算案 ④新事務局長紹介
	5 月 17 日	①03 年度事業報告案 ③総会について	②03 年度決算報告案 ④その他
会計監査	5 月 7 日	①03 年度の会計監査	②その他
運営委員会	1 月 23 日	①運営諸問題討議	②アジアサミット討議 ③その他
	3 月 9 日	①予算理事会準備 ③新事務局長紹介	②運営諸問題討議 ④その他
	4 月 13 日	①運営諸問題討議	②アジアサミット討議 ③その他
	5 月 12 日	①決算理事会準備	②6/18 総会準備 ③その他
	6 月 8 日	①総会、報告会、講演会の企画、進行について ②その他	
サミット準備委員会	4 月 16 日	①開催日程/場所について ③その他	②組織体制について
	6 月 8 日	①開催日程について (2005 年 9 月 12 日～16 日に決定) ②会場ホテルについて ③組織体制について ④その他	
	7 月 29 日	①会場候補について ③組織体制について ⑤協賛・協力団体の依頼について	②登録・受付業務について ④CFP 印刷・配布について ⑥その他

技術動向調査委員会	1月6日	①評価関係討議	②今後の計画討議	③その他
	2月20日	①評価関係討議	②今後の計画討議	③その他
	6月4日	①報告発表について	②本年度の計画討議	
		③新委員の紹介	④その他	
	7月30日	①今後の計画討議	②MT サミットについて	
③JTF 翻訳祭について		④その他		
市場動向調査委員会	1月23日	①個人向けアンケート詳細討議		②その他
	2月20日	①ベンダー向けアンケート実施計画討議		②その他
	3月26日	①ベンダー向けアンケート実施計画討議		
		②今後の計画討議	③新事務局長紹介	④その他
	4月23日	①ベンダー向けアンケート実施計画討議		
		②今後の計画討議	③日経新聞取材について	④その他
	5月31日	①報告会に向けた最終討議（ベンダー調査）		
		②	” ” ”	（ユーザーアンケート）
		③その他		
7月16日	①6/18 報告会発表について			
	②9月以降の活動について	③その他		
ネットワーク翻訳研究会 — （AAMT/Japio 研究会）				
	1月29日	①報告書について	②関連する研究のご紹介	③その他
	3月15日	①報告書のまとめについて		
		②対訳文書からの専門用語抽出について		③その他
	5月14日	①報告書(3/31 発行)についての報告		
		②中国「海南島」出張報告		
		③本年度の研究計画について		
		④出席者からの提案について（討議）	⑤その他	
	6月17日	①新事務局長、新メンバー紹介について		
		②Japio 研究用データについて		
③本年度の研究計画について				
④AAMT 報告会での発表について		⑤その他		
インターネットWG	6月4日	①本年度の活動について	②6/18 報告発表について	
		③その他		
編集委員会	6月4日	①前号反省	②次号（36号）企画	③その他

第1回自然言語処理国際連合会議(IJCNLP-04)ならびに関連会議に参加して

山形大学 横山 晶一、諏訪東京理科大学 江原 暉将

まえがき

The First International Joint Conference on Natural Language Processing (IJCNLP-04) は、Asian Federation of Natural Language Processing (AFNLP) の主催、Chinese Information Processing Society of China (CIPS) が local host となり、Association for Computational Linguistics (ACL) などの後援のもとで、3月22日～24日 Resort Golden Palm, Sanya, Hainan Island, China (中国海南島三亚) で開催された。

関連の会議として、同じ場所で、2件の Pre-conference Tutorials と Pre-conference Workshop (Beyond shallow analysis) (いずれも3月21日開催)、3件の Post-Conference Workshops ((i) Fourth Workshop on Asian Language Resources (ALR-04) (3月25日), (ii) Workshop on Named Entity Recognition for Natural Language Processing Applications (3月26日), (iii) International Workshop on Machine Translation and Multilingual Information Retrieval (MTMIR2004) (3月26日)) と1件の Satellite Event (Asian Symposium on Natural Language Processing to Overcome Language Barriers) (3月25～26日) が開かれている。

ここでは、IJCNLP-04を中心に、ALR-04, MTMIR2004, Satellite Eventの一部について参加報告を行う。

1. IJCNLP-04 報告

上記の場所で開催され、Proceedings ならびに会議終了時の発表によると、20カ国から226名の参加者(うち中国国内53名、他19カ国173名)、211件の論文投稿があり、うち101件が採録(口頭発表

66件、ポスター発表35件、そのうち Proceedings には95件が掲載)された。以下では、開会講演、口頭発表、ポスター発表、パネル討論など、いくつかについて簡単に解説する。

1.1 開会講演 (3月22日 9:15～9:55)

題目: Taxonomy of Natural Language Processing by Prof. Makoto Nagao

長尾真・前京大総長による基調講演であった。まず自然言語処理研究のトピックである、テキスト処理、辞書構築、形態素解析、構文・意味解析、談話解析、文生成、学習過程、応用(機械翻訳、情報検索、テキスト検索、マンマシン対話システム)について各細目をあげて簡単に解説し、これまでの経緯を振り返った後、現在も残る問題点や困難さを指摘した。たとえば、談話解析では、照応や省略現象、文書の種類の同定、文をこえた文脈、会話やシーンの文脈、知識などが問題になり、さらに、言い換え、比喩の同定、同じ意味を種々の言い方で述べたときにそれをどう区別するか、普通の場合と正式な場合の状況をどう判断するかなどの問題があり、ほとんどの問題については未解決の部分が残されていることを指摘した。そして、特に機械翻訳について、将来高度な機械翻訳を実現するための指針を示し、理想的な機械翻訳環境の構築に対する見通しを示した。

1.2 招待講演

(1) “Probabilistic Parsing with a Wide Variety of Features” by Prof. Mark Johnson (Brown Univ.) (3月22日 10:00～10:50)

開会講演の後で行われた招待講演である。当初プログラムにのせられていた題名とは多少異なっている。上記の題名の通り、いろいろな特徴量を用いて、確率的な構文解析を行う場合の一般的な解説と事例を示したものである。統計的、確率的なパラメータを導入するときに、その特徴量とモデルに対する整合性が異なる（たとえば最大尤度は生成モデルに対しては明確であるが、条件付き、あるいは弁別的モデルに対しては制御しにくくなる）ことを示し、粗い近似から細かい近似へ持っていくためにはどのようなことが必要であるかを解説した。実際のデータに対していろいろな特徴量をテストし、それが構文解析木の抽出にどのような影響を与えるかについても、エラーの分析とともに示した。この構文解析は英語についてのものであったが、将来的には中国語にも適用したいということであった。

(2) “The Impact of Information Technology on Communication and Linguistics” by Prof. Ching-Chun Hsieh (Chair Professor of Hsuan Chnang University, Taiwan) (3月23日 9:00～9:50)

題名の通り、大局的な観点から情報技術がコミュニケーションや言語学に与えた影響を論じたものである。情報技術の発達、最初オーディオ世界（音声、メディアとしては音声波形）から始まり、やがて書かれたテキストによる視覚世界（文字、資料）、そして再びオーディオ世界（ただし電子世界。マルチメディア、エネルギー安定のメディア）へと移行した。情報は内容と形式に分けられる。前者は知るべきことであり、後者は表現形式である。この二重性は美学における内容と形式、記号論における意味の考え方と記号、言語学における内容的な意味と要素、情報科学における知識と実際などに見られる。情報は「知りたいこと」（サンسكريット語における「所知」）、デジタルノード、デジタル装置、応用表現システムによって4つのカテゴリに分けられる。デジタル化は、これらの要素を加味して展開することができるという趣旨の講演であった。

(3) “Language Technologies for E-Memory Applications” by Prof. Hans Uszkoreit (DFKI, Saarland Univ.) (3月24日 9:00～9:50)

言語工学がネット社会へ与える影響という観点での講演であった。ウェブのリンクをたどって情報を得て、オントロジー、論理などを抽出し、人間が日常記憶しているようなメカニズムを模倣するには、言語工学のどのような要素が必要であるかを解説した。いくつかのシステムの結果や、かなり古い時代からの知的文脈の構築の先駆者たちの業績について述べた。

1.3 口頭発表

以下には、日本人以外の発表の中で、いくつかを取り上げて概説する。示したページは、Proceedingsのページである。

- “High Speed Unknown Word Prediction Using Support Vector Machine For Chinese Text-to-Speech Systems” by Juhong Ha, Yu Zheng and Gary Geunbae Lee (POSTECH) (pp. 1-6)

中国語のテキスト音声合成に利用することを目的とした未知語予測手法である。未知語の予測には高精度な機械学習として定評のある SVM (Support Vector Machine) を用いている。ただし、SVM は処理速度が遅いため、入力文のすべての部分に対して未知語を予測するのではなく、前処理を用いて、未知語の可能性のある部分のみを事前に抽出し処理している。その結果、高速で高精度な未知語予測が実現できた。実験の結果、F 値は人名に対して 89.49%、地名に対して 91.14%であった。また、前処理を行うことで、処理速度を 25 倍に高速化することができた。

- “Chinese Chunk Identification Using SVMs plus Sigmoid” by Yongmei Tan, Tianshun Yao, Qing Chen and Jingbo Zhu (Northeastern University) (pp. 7-13)

SVMを用いて中国語のチャンク認定を行うものである。利用している素性は、チャンク認定を行う単語および前後2単語の範囲の語形、品詞、チャンク認定結果(前の単語のみ)である。チャンク種別は7種類であり、各種類ごとにチャンクの先頭、内部の2種の区別がある。従って、出力としては、チャンク外も含めて15種類となる。SVMは2値分類器であるので、各出力種別ごとにSVMを作成し、それらの出力から最終的な認定結果を得ている。ここで、各SVMの出力を正規化して比較するために、Sigmoid関数を用いてSVM出力を[0, 1]区間に写像している。実験の結果、F値は77%ないし100%であった。

- “Automatic Learning of Parallel Dependency Treelet Pairs” by Yuan Ding and Martha Palmer (Univ. Pennsylvania) (pp.30-37)

木構造の変換を用いた統計的な機械翻訳システムを提案したもので、著者がすでに発表したパラレルコーパスを用いて依存木の語の配置を行うアルゴリズムを発展させて、木構造のペアを文の長さに対する多項式時間で自動学習し、英語と中国語との木構造の対応をとるシステムを作成している。これは、まず対応する木のペアをtreelet pairと称し、それに対して各々の木を語のかたまりとみなし、語と語の翻訳確率を計算するための統計的翻訳モデルをすべての木のペアに適用して訓練する。次に各treelet pairに対して、根以外の節に対する語間のマッピング確率を計算し、最もよく対応する語のペアをヒューリスティックな関数で選択する。この語のマッピングが構文的な条件を満たしていれば、木の部分的な対応をとってもとのtreelet pairと置き換え、この操作を繰り返す。以上のような手順で、約78万のtreelet pair(各ペアが5.26個の語を含む)を得て、約半数が正しいという結果を得ることができた。

- “Automatic Genre Detection of Web Documents” by Chul Su Lim, Kong Joo Lee, and Gil Chang Kim (KAIST) (pp.69-76)

韓国のグループの発表。ここで述べるジャンルとは、ニュース、小説、詩といった、ウェブのドキュメントの区別で、内容であるトピック(スポーツ、経済など)とは異なる。この研究では、1224のドキュメントを使って、まずウェブのジャンルを非語彙的なもの(個人のホームページ、公共機関のホームページ、商用のホームページなど8つ)と語彙的なもの(ジャーナル、研究、FAQなど8つ)に人間が分類してコーパスを作成した。そしてこれらから特徴抽出する。特徴は、URLやHTMLタグからの特徴、テキスト(文ごとの語数など)や品詞から得られるトークン情報、語彙的な特徴(たとえば論文では語彙が豊富)、構造的な情報(名詞句、動詞句などの句から得られる情報や、日付、時間などの情報)といったものを用いた。これでドキュメントを分類したところ、約75%の精度を得た。

- “Statistical Substring Reduction in Linear Time” by Xueqiang Lue, Le Zhang, and Junfeng Hu (Peking Univ.) (pp.77-82)

北京大学のグループによる発表。n-gram string (n個の文字連鎖)を統計的に求めるときに、同じ頻度を持つ部分文字列(たとえば、「中華人民共和国」に対する「華人民共和国」)を取り除くことは重要である。これまでのアルゴリズムでは、このようなことを行うためには、文字列の長さnに対して、2乗のオーダーの計算時間がかかっていたが、文字列集合の中にある文字列を昇順にソートすることによって、この計算時間のオーダーをnの範囲で押さえるアルゴリズムを開発した。これによって、高速に、無駄な部分文字列を取り除くことができるようになった。

- “BBS Based Hot Topic Retrieval Using Back-Propagation Neural Network” by Lan You, Yongping Du, Jiayin Ge, Xuanjing Huang, Lide Wu (Fudan Univ.) (pp. 99-106)

復旦大学（上海）のグループによる発表。BBS (Bulletin Board System)は、インターネット上でテーマを決めて議論を行うシステムであるが、その中で今何が最も話題になっているかを自動抽出しようという試みである。そのような話題の特徴は、多くの質の高いメッセージがあること、多くの著者が論じていること、話題への反応が速いこと、その話題の権威によって書かれていること、多くの読者の関心が高いことなどがあげられる。このシステムは、著者の解析から権威者を見つけるとともに、BBSからスレッドを抽出して解析し、そのデータをバックプロパゲーションニューラルネットワークを用いてクラスタリングする。これらの解析結果をもとに、何が話題になっているかを同定するものである。これを用いてF値を計算した結果、7000弱のBBSに対して84%程度になり、他の方法よりよい結果を得ることができた。

- “A Novel Pattern Learning Method for Open Domain Question Answering” by Yongping Du, Xuanjing Huang, Xin Li, and Lide Wu (Fudan Univ.) (pp. 286-293)

復旦大学（上海）のグループによる発表。分野を限定しない質問応答システムに対して、質問のパターンを分類し、それに対する答をウェブから自動学習しようとするものである。質問を、焦点（何を聞いているのか）、固有名詞（質問中に現れる固有名詞）、動詞（質問の主動詞）、名詞句（質問中に現れる名詞句）を用いて解析し、それに対する答の型を場所、人名、数、日付、組織、名詞句の6つに分ける。これらの質問と答のパターンを学習し、それぞれにスコアをつける。これを用いて、Text Retrieval Conference (TREC)の500の質問に適用してウェブ

のサーチエンジン(Google)から答を学習させた。評価は、MRRという指標を用いた。この指標で最もよい結果で精度が0.309になった。

- “Chinese Named Entity Recognition Based on Multilevel Linguistic Features” by Honglei Guo, Jianmin Jiang, Gang Hu, and Tong Zhang (IBM) (pp. 294-301)

IBMグループによる発表。中国語の解析で問題になっているのは、固有名詞をどのように同定するかという問題である。すべて漢字で書かれ、語間に切れ目のない中国語では、どの部分が固有名詞であるかを認識するのが重要な問題である。ここでは、robust risk minimization (RRM)という手法を用いて文字、単語ベースで中国語の固有名詞を認識するシステムについて述べる。ここでは、文字に対する条件付き確率を用いたRRMという尺度を用いて大量のデータで訓練した確率を計算する。それを用いて実際の固有名詞を認識したところ、人名については85.48%、地名については85.18%、組織については72.92%など、全体で79%程度のF値を得た。

- “A Study of Semi-Discrete Matrix Decomposition for LSI in Automated Text Categorization” by Qiang Wang, XiaoLong Wang and Yi Guan (Harbin Institute of Technology) (pp. 302-309)

中国語の文書をカテゴリー分類する手法に関する発表である。文書を索引語のリストで表現するベクトル空間モデル(VSM)において、Latent Semantic Indexing(LSI)が用いられているが、本論文では、文書・索引語行列をSingular Value Decomposition(SVD)ではなくSemi-Discrete Matrix Decomposition(SDD)で分解する手法を用いている。実際の分類は、k-NN(k最近傍)法によって行われる。実験の結果、LSIを用いる方法が単純なVSMを用いる方法よりも8.21%精度が向上している。

- “Combining Labeled and Unlabeled Data for Learning Cross-document Structural Relationships” by Zhu Zhang and Dragomir Radev (Univ. Michigan) (pp. 342-349)

修辞構造理論(Rhetorical Structure Theory, RST)は、ドキュメントの談話構造を木のような形で表すことのできる理論である。この理論に基づいて、ドキュメント間の構造的な関係を表すドキュメント間構造理論(Cross-document Structure Theory, CST)が著者の一人によって提案された。これは、話題の関係するドキュメントの間の文どうしの対応関係を示そうというもので、すでに複数ドキュメントの要約などに応用されている。この論文では、あらかじめラベル付けされたドキュメントと、ラベル付けされていないドキュメントとを、教師付きの学習を用いて関連付けする実験について述べた。パラメータとしては、文の持つ6つの品詞の一致度を見る浅い構文レベルの特徴量、トップレベルの句(名詞句や動詞句)の対応や配置を見る深い構文レベルの特徴量などがある。これによって、約0.87程度の精度が得られている。

- “Window-based Method for Information Retrieval” by Qianli Jin, Jun Zhao, and Bo Xu (Chinese Academy of Science) (pp. 382-389)

情報検索の評価尺度としてよく用いられるものにtf-idfがある。しかしながら、見かけ上この値が高くても、実は本来検索すべき情報ではないドキュメントが含まれてしまうことがある。これを防ぐために、文の中に窓を設定して、その窓の範囲内に関連する検索語が含まれていれば、より求めるドキュメントに近いものが得られるのではないかというのがこの論文のアイデアである。窓を固定したり、動的に窓を動かしたり、核となる語を中心とした窓を設けたりすることによって、従来のtf-idf法よりもよい結果を得ることができた。

- “The Role of Semantic Information in Learning Question Classifiers” by Xin Li, Dan Roth, and Kevin Small (Univ. Illinois at Urbana-Champaign) (pp. 451-458)

質問文をクラス分けするのは、通常のテキストのクラス分けとは異なる。質問文は一般的には短いので、単語ベースでの情報が取りにくい。この論文では、質問に対する答の型をまず6つの基本的なクラス(省略、実体、記述、人間、場所、数値)に分けるとともに、それらをさらに細かい50のクラスに細分化する。これらを機械学習させて質問文のクラス分けに用いる。上記の上海のグループとやや類似したアプローチのやり方である。結果は、クラス分けの精度として、85~90%程度が得られている。

- “Annotation of Gene Products in the Literature with Gene Ontology Terms using Syntactic Dependencies” by Jung-jae Kim and Jong C. Park (KAIST) (pp. 528-534)

韓国のグループによる発表。医学文献から自動的に遺伝子にかかわる項目を抽出するシステムについて述べている。その方法は、形態素解析と構文解析を用いてオントロジー的な項目を抽出し、遺伝子に関する情報と、タンパク質の名称とが対応することを用いて遺伝子にかかわる項目を引き出そうとするものである。実際には211の医学文献のアブストラクトをコーパスに入れてその半分を訓練用を使用し、のこりの半分でどの程度情報が抽出できるかを評価した。この手法では、従来の手法に比べて著しくよい89.7%が正しくコード付けされた。

1.4 パネル討論

- Panel on Emerging Asian Language Processing Efforts

アジアの少数言語に関する言語処理の状況に関するパネルである。インドネシア・マレー語について Hamman Riza から、モンゴル語について Ochir から、

チベット語について Di Jiang から、ウイグル語について Yusup Abaidula から、それぞれ報告があった。座長は Benjamin Tsou である。インドネシア・マレー語を除いて、日本のワープロ黎明期を思わせる状況であり、文字コード、フォント、入力システムについての報告がなされた。言語処理に関して技術的・経済的な支援を要請する意見が出ていた。標準化動向について質問したが、明確な方向性はないようであった。

・ Panel on Multilingual NLP for Public Information Services (2008 Digital Olympics)

このパネルは、2008 年の北京オリンピックに向けて、自然言語処理技術を用いた情報サービスを提供しようという試みである。このシステムを MISS (Multilingual Intelligent Service System) と称し、スポーツ、観光、予約、ショッピングといった情報を音声、テキストで 3~6 言語で PC、mobile の形で提供し、人間と機械との自動インタフェースを目指そうという意欲的な試みである。技術的には、音声認識、テキストの読み上げ、機械翻訳、自然言語理解、対話操作、コンテンツなどが必要になる。2002 年から 2004 年までに、パイロット的な研究を行い (第 1 フェーズ)、2005 年から 2006 年にかけて拡大テストを行って、標準仕様を決め (第 2 フェーズ)、2007 年に統合と試行を行おう (第 3 フェーズ) というものである。中国のいくつかの機関が参加している。Yixin Zhong から発表された。

これに対して、Key-Sun Choi (Daejon), Robert Dale (Sydney), Keh-Yih Su (Hsinchu), Kumiko Tanaka-Ishii (Tokyo), Hans Uszkoreit (Saarbruecken) の各パネラーからそれぞれコメントがあった。多くのパネラーがこのプロジェクトの意欲を買いながらも、あまりテーマを広げ過ぎて実現できるかを危惧する声が多かった。また、もし実現できなかった場合のマイナスの影響を心配する声もあった。総じて、意欲は買いながらも、もっとテーマを絞り込んだ方がよいという意見であった。

また、4 年ごとのオリンピックに向けてこのような試みが継続されることを望む意見も出された。

1.5 ポスター発表

- ・ “Using a Smoothing Maximum Entropy Model for Chinese Nominal Entity Tagging” by Jinying Chen, Nianwen Xue and Martha Palmer (University of Pennsylvania) (pp.123-128)

中国語の固有名詞認定を ME 法 (Maximum Entropy) を用いて行うものである。出力クラスは 6 種類である。ME 法を適用するにあたって Gaussian prior を用いた Smoothing を適用している。11 万字の学習データで学習し、13.5 万字の試験データで試験した結果、F 値にして 55~79% であった。これは、5 回未満のデータを用いない単純なカットオフ手法に比べて 10~20% 精度が高い。また意味に関する素性を利用しないようにすると、2~3%、クラスによっては 15% 以上の性能低下となった。さらに、学習データを約 2 倍にして実験したところ 1~15% の精度向上が得られた。誤りデータを分析した結果、前処理に利用しているタガーの性能不足および広範囲な文脈を利用していないことが誤りの主な原因であることが分かった。

- ・ “Building a parallel bilingual syntactically annotated corpus” by Jan Cufín, Martin Čmejrek, Jiří Havelka and Vladislav Kuboň (Charles University) (pp.141-146)

チェコのグループによる発表である。bilingual syntactically annotated corpus である PCEDT (Prague Czech-English Dependency Treebank) についての設計思想と内容を述べた。既存の主な Czech-English bilingual corpus として Reader's Digest corpus があるが、これは自由な翻訳によるものであり、機械学習に利用しにくいという問題がある。そこで PCEDT では、既存の Penn Treebank をチェコ語に翻訳し、それに構文構造を付与するという方法をとった。構成文法を利用した

Penn Treebank の構文構造をチェコ語に適用するのは困難である。チェコ語は語順が比較的自由で分離した構成素などもあり、依存文法に基づく構文構造が有効である。従って、PCEDT では、そのような構文構造として Prague Dependency Treebank で用いられてきたものを利用する。その構文構造には、analytical なものと tectogrammatical なものの 2 種類がある。Penn Treebank の構文構造はツールを用いて analytical dependency structure と tectogrammatical dependency structure に変換される。チェコ語への翻訳結果もいくつかのツールを用いて上記の構文構造が付与される。

- “A Graph Grammar Approach to Map between Dependency Trees and Topological Models” by Bernd Bohnet (University of Stuttgart) (pp. 548-553)

依存文法に基づく自然言語生成においては、語の出現順序を依存木から生成しなければならないが、語順が自由な言語においては、このことが困難である。そこで、ドイツ語の生成においては、依存木と表層文の中間に topological model を設定する方法がとられている。しかし、従来の topological model は過度に複雑であり、言語依存であった。さらに依存木から topological model へ自動的に書き換える手法は得られていなかった。本論文では、適切な topological model を設定することで、上記の書き換えを Graph grammar によって行えるようにした。また、コーパスから書き換え規則を自動学習する手法についても述べている。

- “An English-Hindi Statistical Machine Translation System” by Raghavendra Udapa U and Tanveer A Faruque (IBM India Research Lab) (pp. 626-631)

英語からヒンディ語への統計的機械翻訳に関する論文である。IBM モデル 1、2、3 を用いている。

fertility parameter をモデル 2 から 3 に効率良く transfer することを特徴としている。15 万文の英語・ヒンディ語対訳コーパスを用いて学習し、1032 文で試験したところ BLEU Score が 0.1391、NIST Score が 4.6296 であった。

2. Fourth Workshop on Asian Language Resources (ALR-04) 報告

アジアの言語は、文字や文法を取ってみても非常に多様である。このワークショップでは、それらについて情報交換し、アジアにおける自然言語処理の発展に寄与するという目的がある。全部で 13 件の発表（1 件のキャンセルを除く）が行われ、そのうち午後 3 時までに行われた 10 件の発表を聞いた。このうち日本からの発表を除く 8 件についてごく簡単に解説する。

“Developing Asian language corpora: standards and practice” by Zhonghua Xiao, Tony McEnery, Paul Baker, and Andrew Hardie (Lancaster University) は、有名な英国のランカスター大学のコーパスに、ベンガル語やヒンズー語などの 14 の南アジアの言語コーパス（音声を含む）を構築しようとするものであった。

“The Formosan Language Archive: Language Processing and Linguistic Analysis” by Elizabeth Zeitoun and Ching-hua The Formosan Language Archive: Language Processing and Linguistic Analysis” by Elizabeth Zeitoun and Ching-hua Yu (Institute of Linguistics, Academia Sinica, Taipei) は、台湾列島の諸言語（24 以上の言語がある）に対して 2006 年までにコーパスを作成し、語のカテゴリなどを付与しようという試みである。

“Handling Dislocated and Discontinuous Constituents in Chinese Semantic Role Labeling” by Nianwen Xue (Jinying Chen による代理発表) は、ペンシルベニア大学の中国語のトリバンクに意味を付与しようとする試みであった。

“An Introduction to Chinese Corpora Resources of 863 Program for Chinese Language Processing and Human-Machine Interaction” by Qian YueLiang et al. (Institute of Computing Technology of Chinese Academy of Sciences)は、中国の種々の機関の協力で作成されている 863 プログラムという中国語関連（テキスト、オーディオ（音声合成、音声認識）、イメージ（文字画像））のコーパスの紹介であった。

“Lexical descriptions for Vietnamese language processing” by Thauh Bon Nguyen et al. (IFI, Hanoi)は、文字レベルから、品詞、意味カテゴリなどを付与したベトナム語のデータベースの話であった。

“Acquiring Semantic Information in the TCL’s Computational Lexicon” by Thatsanee Charoenporn et al. (Communication Research Laboratory)は、タイ語の電子辞書に、上下関係、全体部分関係などの意味情報を付与した試みであった。

残りの2つの short paper, “Implementation of an OLAC-based Linguistic Metadata System over a Set of Heterogeneous Language Archives” by Bu-Yng Chang et al. (Academia Sinica, Taipei)と、

“MINECoP: An Integrated Visualization Tool for Corpus Mining” by Asance Kawtrakul et al. (Kasetsart University, Bangkok)は、前者が中国語（繁体字）、後者がタイ語のコーパスに関するものであった。

このように、アジアの種々の言語のコーパス化、データベース化が進んできている。

3. International Workshop on Machine Translation and Multilingual Information Retrieval (MTMIR2004) 報告

6 件の発表予定であったが、1 件キャンセルされ、5 件の発表であった。そのうち日本以外からの発表は次の 3 件である。

- “Approach to Automatic Translation Template Acquisition Based on Unannotated Bilingual Grammar Induction” by Rile HU et al. (Chinese Academy of Science)

EBMT の課題の一つとして、いかに翻訳用例 (translation template) を集めるかがある。対訳コーパスを両言語の parser を用いて解析し、部分木を対応させて翻訳用例とする手法では、高精度な parser が必要である。本論文では、対訳文に現れる語を spatially clustering と temporally clustering という手法でクラスタリングし、句のグループ (phrasal structure group) として文を表現しておき、原言語と目標言語の句の対として翻訳用例を求める。対応付けには Dynamic Programming を用いている。これによって unannotated bilingual corpus から翻訳用例を収集することができる。

- “An Improved Chinese Sentence Similarity Measure” by Rongbo Wang et al. (Hong Kong Polytechnic University)

EBMT の重要な要素技術の一つである文間の類似度計算に関して、中国語文を対象に新しい指標を提案している。従来は類似度として単に共通する語の割合を用いていたが、ここで提案されている方法は、語だけでなく語列について考慮するとともに各語にウエイト付けして類似度を計算するものである。

- “CI-based NLU Methodology: Theory and Application” by YiXin ZHONG

自然言語理解を、構文解析、意味解析、語用解析の 3 段階でとらえ、すべての段階をベクトル空間モデルで表す方法を提案している。

4. Asian Symposium on Natural Language Processing to Overcome Language Barriers 報告

Satellite Event として2日にわたって行われた。4件の招待講演、1件のパネル、8件の口頭発表、7件のポスター発表が行われた。そのうちいくつかについて以下に報告する。以下ページ番号は Proceedings of Asian Symposium on Natural Language Processing to Overcome Language Barriers のものである。

- ・ “Trends in Natural Language Processing to Overcome Language Barriers” by Eduard Hovy (University of Southern California) (Invited Speech)

「自然言語処理は table look up である」という大胆な方法論を展開した。手法には各種あるものの、利用する情報源(corpus など)の量を増やしてゆくと手法による性能の差は縮まってくる。つまり、情報源の量を増やすことが重要であるという主張である。統計的機械翻訳などを例に引いてユーモアを交えた魅力的な講演であった。

- ・ “Projects corresponding to multilingual environment in the “Asia Broadband Program” by Shuichi Honda (Japanese Ministry of Public Management, Home Affairs, Posts and Telecommunications) (Invited Speech, pp. 105-127)

Asia Broadband Program の紹介と CRL や ATR での研究やアジアとの collaboration に関する発表である。また、先に行われた Asia Broadband Symposium の内容についても講演した。

- ・ Example-Based Machine Translation Based on Translation Corresponding Tree Representation” by WONG Fai, TANG Chi Wai, DONG Ming Chui, MAO Yu Hang, LI Yi Ping (Tsinghua

University et al.) (pp. 31-38)

EBMT の翻訳用例を bilingual corpus から収集する方法に関する論文である。通常、翻訳用例は bilingual corpus の原言語側と目標言語側の双方を構文解析し、構文木を対応させることで収集するが、ポルトガル語と中国語のように文法構造が大きく異なる言語間では、構文木の対応が見つからない場合が多く、翻訳用例を多く集めるのが困難である。そこで、本論文では原言語側のみ構文解析し、原言語側の構文木の各ノードに目標言語側の単語(列)を対応させるという手法を用いている。これによって木同士の対応が見つからないという問題が回避できる。さらに、原言語の構文解析を Chomsky normal form の文法で行うことで、翻訳用例に語順の入れ替え情報を指定することが自然にできる。

- ・ “Putting NLP Tools into Action: Incorporating NLP Tools in Web-based Language Learning Environment” by Md Maruf Hasan, Kazuhiro Takeuchi, Virach Sornlertlamvanich, Hitoshi Isahara (CRL) (pp. 85-90)

形態素解析や電子辞書など NLP tools を組み合わせて外国語学習のための学習環境を Web 上に実現した。読解支援、作文支援、教材作成支援(教員用)などを行うものである。

あとがき

今回の IJCNLP は、アジアで初めての NLP に関する international joint conference であった。参加者も多く充実した議論が展開された。日本、韓国からの発表はいつもどおり多かったが、中国からの発表も多くかつ優れたものであった。次回の IJCNLP は、2005 年 10 月ごろに韓国済州島で開かれる予定である。

謝辞

本文は AAMT/Japio 研究会の報告書を基に記述したものである。転載を許可された(財)日本特許情報機構(Japio)に感謝いたします。

Use of Machine Translation in India

Sivaji Bandyopadhyay, Jadavpur University

Introduction

India has 18 officially recognized languages which are written in 10 different scripts: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu, and Urdu. Hindi is the official language of India. English is very widely used in the media, commerce, science and technology and education. Many of the states have their own regional language, which is either Hindi or one of the other constitutional languages. Only about 5% of the population speaks English.

In a large multi-lingual society like India, there is a great demand for translation of documents from one language to another. Most of the state governments work in the respective regional languages whereas the Union Government's official documents and reports are in bilingual form (Hindi/English). In order to have a proper communication there is a need to translate these documents and reports in the respective regional languages. The newspapers in regional languages are required to translate news in English received from International News Agencies. With the limitations of human translators most of this information (reports and documents) is missing and not percolating down. A machine assisted translation system or a translator's workstation would increase the efficiency of the human translators.

As is clear from above, the market is largest for translation from English into Indian languages, primarily Hindi. Hence, it is no surprise that a majority of the Indian Machine Translation (MT) systems are for English-Hindi translation.

Machine Translation in India is relatively young. The earliest efforts date from the late 80s and early 90s. The prominent among these are the projects at Indian Institute of Technology (IIT) Kanpur (<http://www.iitk.ac.in/>), University of Hyderabad (<http://www.uohyd.ernet.in/>), National Center for Software Technology (NCST) Mumbai (now, CDAC, Mumbai) (<http://www.ncst.ernet.in/>) and Center for Development of Advanced Computing (CDAC) Pune (<http://www.cdacindia.com/>). The Technology Development in Indian Languages (TDIL) (<http://tdil.mit.gov.in>), an initiative of the Department of IT, Ministry of Communications and Information Technology, Government of India, since 1990-'91 has played an instrumental role by funding these projects. The following domains have been identified for the development of domain specific translation systems: government administrative procedures and formats, parliamentary questions and answers, pharmaceutical information, legal terminology and important judgments, and so on.

Since the mid and late 90's, a few more projects have been initiated—at Indian Institute of Technology (IIT) Bombay (<http://www.iitb.ac.in/>), International Institute of Information Technology (IIIT) Hyderabad (<http://www.iiit.net/>), Anna University – KB Chandrasekhar Research Center (AU-KBC) Chennai (<http://www.au-kbc.org/>) and Jadavpur University Kolkata (<http://www.jadavpur.edu/>).

There are also a couple of efforts from the private sector - from Super Infosoft Private Limited, and more recently, the IBM India Research Laboratory.

A survey of the machine translation systems that have been developed in India for translation from English to Indian languages and among Indian languages reveals that the MT softwares are used in field testing or are available as web translation service. These systems are also used for teaching machine translation to the students and researchers. Most of these systems are in the English-Hindi or Indian language-Indian language domain with only one system in the English-Kannada language domain. The translation domains are mostly government documents/reports and news stories. There are a number of other MT systems that are at their various phases of development and have been demonstrated at various forums. Many of these systems cover other Indian languages beside Hindi.

1 MT Systems in Field Testing or as Web Service

The following machine translation systems are either being used in field testing or are available as a web service. Most of these systems are in the English-Hindi language domain with one exception of an English-Kannada machine translation system.

1.1 ANGLABHARATI Technology

The *ANGLABHARATI* project was launched by Professor R. M. K. Sinha at the Indian Institute of Technology, Kanpur in 1991 for machine aided translation from English to Indian languages. The system's approach and lexicon is general-purpose, but has been applied mainly in the domain of public health campaign. This technology is being extended to another domain for translation of Financial and Supplementary Rules of Government of India and related correspondence. The project is primarily based at IIT-Kanpur (<http://www.cse.iitk.ac.in/>), in collaboration with CDAC, Noida (<http://www.cdacnoida.com>), and has been funded by TDIL.

The *Anglabharati* methodology was used to design a functional prototype for English to Hindi on Sun system. Feasibility on extending this for English to Telugu/Tamil was also demonstrated. The outcome of this project has been release of the first version of the software named *AnglaHindi* (an English to Hindi version based on *Anglabharti* approach) that accepts English text with almost no constraints on its form. *AnglaHindi* has also been web-enabled and its beta version is available for on-line translation at URL: <http://anglahindi.iitk.ac.in>. The link to the *AnglaHindi* web translation service is also maintained at the TDIL

the British Computer Society Natural Language Translation Specialist Group (<http://www.bcs.org.uk/siggroup/nalatran/>). The Center for Development of Advanced Computing (CDAC), Noida has developed a machine aided translation system that translates public health campaign documents from English to Hindi using the *Anglabharati* approach. At present, IIT Kanpur is executing non-exclusive technology transfer of *Anglabharati* technology and the *AnglaHindi* system with the various resource centers of the TDIL project for development of machine translation systems using the *Anglabharati* technology from English to the respective language(s) catered by the resource center. *Anubharati* is a recent project at IIT Kanpur, dealing with template-based machine translation from Hindi to English, using a variation of example-based machine translation. An early prototype has been developed and is being extended. The contact person is Prof. RMK Sinha (rmk@cse.iitk.ac.in).

1.2 MANTRA MT System

MANTRA (MAchiNe assisted TRAnslation tool) translates English text into Hindi in a specified domain of personal administration, specifically gazette notifications, office orders, office memorandums and circulars. *Mantra* uses *Lexicalized Tree Adjoining Grammar (LTAG)* to represent the English as well as the Hindi grammar. It uses *Tree Adjoining Grammar (TAG)* for parsing and generation. The input to the *Mantra* system can be through text documents, outputs of speech recognition programs or OCR packages. The *Mantra* has become part of the permanent research collection on Information Technology at Smithsonian Institution's National Museum of American History. Further details about the *Mantra* system can be obtained from the CDAC website <http://www.cdacindia.com/html/aai/mantra.asp>.

The project has been funded by TDIL. Later, it received funding from the Department of Official Languages, Government of India for specific domain of Government of India Appointment Letters. The system is currently being tested at several ministries in the Central Government. The *Mantra* Technology is being expanded to translate English texts into other Indian languages such as Gujarati, Bengali, and Telugu as well as on extending to the domain of parliament proceeding summaries. The domain for Hindi translation is being expanded from the domain of personnel administration to other domains like banking, transportation and agriculture. The contact person is Dr Hemant Darbari (darbari@cdac.ernet.in).

1.3 ANUSAARAKA MAT System

A machine aided translation system (*ANUSAARAKA*) among Indian languages has been built with funding from the TDIL project. *Anusaaraka* is a Language accessor rather than a machine translation system in true sense. *Anusaarakas* have been built from Telugu, Kannada, Bengali, Marathi, and Punjabi to Hindi. Alpha versions of all of these have been released so that their field testing can be carried out. The beta-versions are expected to be released soon.

It is domain free but the system has mainly been applied for translating children's stories. An e-mail server

been established for the *Anusaarakas*. It currently holds alpha-versions of *Anusaarakas* from Telugu, Kannada, Marathi, Bengali, and Punjabi into Hindi. The purpose of the alpha-version is to invite people to join and work for its improvement. To run the *Anusaaraka* on a given text, e-mail has to be sent to nandi@anu.uohyd.ernet.in with the name of the language in the subject line. For example, if 'telugu' is put in the subject line, it automatically runs the Telugu to Hindi *Anusaaraka*. The output produced is sent back to the sender. The computer keeps a copy of the output for a later study of results. The text should be in 7-bit ISCII coding. Similarly help by mail is available if mail is sent with subject 'Help'. The systems with complete source code and language data are available as "free" software and can be downloaded from the website of the Language Technologies Research Center, Indian Institute of Information Technology, Hyderabad (<http://ltrc.iiit.net/>).

The project originated at IIT Kanpur, and later shifted mainly to the Centre for Applied Linguistics and Translation Studies (CALTS), Department of Humanities and Social Sciences, University of Hyderabad. It was funded by TDIL. It is now being carried out at the Language Technologies Research Center, Indian Institute of Information Technology, Hyderabad with financial support from Satyam Computers Private Limited. The contact person is Prof. Rajeev Sangal (sangal@iiit.net).

1.4 SHIVA and SHAKTI MT Systems

Two machine translation systems from English to Hindi, *Shiva* and *Shakti* are being developed jointly by Carnegie Mellon University USA, Indian Institute of Science, Bangalore, India, and International Institute of Information Technology, Hyderabad. The two experimental systems have been released for experiments, trials, and user feedback. The Example-based Machine Translation system (*Shiva*) can be run from the site (<http://ebmt.serc.iisc.ernet.in/mt/login.html>) and the machine translation system (*Shakti*) can be run from its own site (<http://shakti.iiit.net>).

1.5 UNL-based English-Hindi MT System

The Indian Institute of Technology, Bombay has developed a machine translation system for translating from English to Hindi using the Universal Networking Language (UNL) as the interlingua. The UNL is an international project of the United Nations University, with an aim to create an Interlingua for all major human languages. IIT Bombay is the Indian participant in UNL. The MT system can be used at the site <http://www.cfilt.iitb.ac.in/machine-translation/eng-hindi-mt>. They are also working on MT systems from English to Marathi and Bengali using the UNL formalism. The contact person is Prof. Pushpak Bhattacharya (pb@cse.iitb.ac.in).

1.6 MATRA MT System

The Natural Language group of the Knowledge Based Computer Systems (KBCS) division at the National Centre for Software Technology (NCST), Mumbai (currently CDAC, Mumbai) is working on *MaTra*,

a human-aided transfer-based translation system for English to Hindi. The work is supported under the TDIL Project. The *MaTra* lexicon and approach is general-purpose, but the system has been applied mainly in the domains of news, annual reports and technical phrases, and has been funded by TDIL.

A Hybrid approach system *Vaakya* has been developed at NCST, Bombay for translating English news stories to Hindi. The system can handle single verb sentences. Prototype *Vaakya* system is now being enhanced and adapted for providing web translation service to the news agencies from English news stories to Hindi. More details about the *MaTra* system can be obtained from the NCST website <http://www.ncst.ernet.in/matra/>. They are using the translation system in a project on Cross Lingual Information Retrieval that enables a person to query the web for documents related to health issues in Hindi (<http://www.ncst.ernet.in/projects/clir/>). The contact person is Ms. Kavitha Mohanraj (kavitham@ncst.ernet.in).

1.7 English – Kannada MT System

The Computer and Information Sciences Department at the University of Hyderabad has worked on an English-Kannada MT system, using the Universal Clause Structure Grammar (UCSG) formalism. This is essentially a transfer-based approach, and has been applied to the domain of government circulars, and funded by the Karnataka government. Currently, efforts are on to improve the system and apply it for English-Telugu translation as well. Further details about the English-Kannada machine aided translation system can be obtained from their website <http://www.language technologies.ac.in/lerc/mat/mat.htm>. The contact person is Prof. K Narayana Murthy (knmc@uohyd.ernet.in).

1.8 ANUVAADAK MT System

Super Infosoft Private Limited, Delhi have developed a general purpose English-Hindi machine translation tool *Anuvaadak 5.0* that supports post-editing. It has inbuilt *dictionaries for specific domains* e.g. official, formal, agriculture, linguistics, technical, and administrative. When Hindi meaning of the English word is not available in dictionary, facility of transliteration is provided. The software runs on any operating system in the Windows family. A demonstration version of the system can be downloaded from the site <http://www.mysmartschool.com/pls/portal/portal.MSSStatic.ProductAnuvaadak>. The contact person is Mrs. Anjali Rowchowdhury (anjalir@del16.vsnl.net.in).

2 Other MT Systems

The Computer Science and Engineering Department of Jadavpur University (<http://www.jadavpur.edu/>) is working on an machine translation system *ANUBAAD* for translating English news items to Bengali using hybrid approach. The current version of the system works at the sentence level. A separate activity is going for developing a machine translation system for translating English news headlines to Bengali. A demonstration version of the system has been developed. The English-Bengali MT system architecture has been used to develop a English – Hindi MT system that works at the simple sentence level. These

systems are used for teaching the students and researchers who work in the area of machine translation. The work was initiated as part of a Research Award granted to Prof. Sivaji Bandyopadhyay (sivaji_ju@vsnl.com) by the University Grants Commission (UGC) of the Government of India in 1999.

Work on a Tamil-Hindi *Anusaaraka* system and an English-Tamil human aided machine translation system is going on at the Anna University KB Chandrasekhar Research Centre (<http://www.au-kbc.org/>) at Chennai. The contact person is Prof. CN Krishnan (cnkrish@au-kbc.org). Utkal University, Vanivihar is working on a English-Oriya machine translation system OMTrans (<http://www.ils-utkal.org/omtrans.htm>). The contact person is Prof. Sanghamitra Mohanty (sangham1@rediffmail.com). The Department of Mathematics, Indian Institute of Technology, Delhi is working on an English-Hindi example based machine translation system and the contact person is Prof. Niladri Chatterjee (niladri@iitd.ernet.in). The IBM India Research Lab at New Delhi has recently initiated work on statistical MT between English and Indian languages, building on IBM's existing work on statistical MT.

3 Conclusion

MT is relatively new in India – about a decade old. In comparison with MT efforts in Europe and Japan, which are at least 3 decades old, it would seem that Indian MT has a long way to go. However, this can also be an advantage, because Indian researchers can learn from the experience of their global counterparts.

Development of Speech-Speech machine translation system from English to Indian languages and among Indian languages is the goal of the TDIL project and the various resource centers under the TDIL project are working towards that. Works on developing machine translation systems for other languages are also being initiated. There are governmental as well as voluntary efforts under way to develop common lexical resources, and to create forums for consolidating and coordinating NLP and MT efforts.

The time has come when translation activities in Indian languages should look beyond English as the source or the target language. The countries in the Asia-Pacific region are renewing their relations in the area of tourism and trade. Controlled language machine translation systems in these domains from Indian languages to the languages of the Asia-Pacific region like Japanese, Chinese and others and vice-versa will foster the development of people-to-people contact in this region.

Bibliography

1. R.M.K. Sinha and A. Jain. AnglaHindi: An English to Hindi Machine-Aided Translation System. Proceedings of MT SUMMIT IX, 2003.
<http://www.amtaweb.org/summit/MTSummit/papers.html>
2. D. Gupta and N. Chatterjee. Identification of Divergence for English to Hindi EBMT. Proceedings of MT SUMMIT IX, 2003.
<http://www.amtaweb.org/summit/MTSummit/papers.html>

3. Sivaji Bandyopadhyay. Teaching MT – An Indian Perspective. Proceedings of the 6th EAMT Workshop on Teaching Machine Translation, 2002.
<http://mull.ccl.umist.ac.uk/staff/harold/teachingMT/Bandyo.pdf>
4. Shachi Dave, Jignashu Parikh and Pushpak Bhattacharyya. Interlingua Based English Hindi Machine Translation and Language Divergence, Journal of Machine Translation, Volume 17, September, 2002.
5. Durgesh Rao. Machine Translation in India: A Brief Survey. SCALLA 2001 Conference, Bangalore.
<http://www.elda.fr/proj/scalla/SCALLA2001/SCALLA2001Rao.pdf>
6. Sivaji Bandyopadhyay. State and Role of Machine Translation in India. Machine Translation Review, Issue No. 11, December 2000.
<http://www.bcs.org.uk/siggroup/nalatran/mtreview/mtr-11/mtr-11-10.htm>.

事務局だより

AAMT 取材記事のご紹介

- (1) 日本経済新聞夕刊(取材記事) 2004年8月10日
「翻訳ソフト ビジネスや学習を強力にサポート」
- (2) eとらんす8月号(取材記事) パベルプレス発行
「AAMT 報告会・講演会(6/18/2004) 関連記事」 P1, P58～P61

機械翻訳に対するユーザとしての要望

翻訳者 吉川 潔

私は新会員ですが翻訳歴は古いです。20年前にアイビーエス社の桜井社長からパソコン通信で翻訳ネットワークの構築を考えている。第一号にと要請され新潟の田舎で翻訳を始めました。桜井社長は進取の気迫に溢れた人で、SOHO形態だけでなく、機械翻訳の実用化のためにユーザ側の先頭に立ちました。AAMTの設立総会で講演しているので、知る人も多いと思います。そういう事情で、私も機械翻訳に関係してきたので、ユーザとしての要望事項をまとめてみました。

(1) 自動翻訳ソフトの利点

●高速： 原稿がパソコン上に組み込んである場合、全てを瞬時に訳してくれます。私が10時間も要した英文和訳を10分で訳します。

●訳漏れが無い： 機械翻訳に漏れが無いです。原稿から一語でも翻訳漏れを生じるとは、翻訳者の信用低下になります。漏れが無いという心理効果は絶大です。

●数値の判読が正確： 人の眼には、20000 booksを2000冊の本と訳すように、数値を誤読する恐れがあります。この不安も解消されます。

●原稿が文法的に正確で短文の場合に、例えば、輸入家庭製品の取扱説明書は老若男女が迷わずに使用できるように作成しており、正確に高速で和訳してくれます。

●和文英訳の場合に、日本語が標準的であっても英訳文が思い浮かばない場合があります。その際に、ヒントを機械翻訳ソフトから得ることが毎日数回あります。

(2) 機械翻訳ソフトの欠点

様々に指摘されていますが、原文が適正な日本語（又は英語）で記してないと、適正に英訳（又は和訳）できないことです。長文になるにつれて不明瞭な訳になります。

翻訳者が受ける原稿は文法的に不適正な文章が多いので、機械翻訳時に障害になります。

(3) 欠点に対する対策として、ユーザからの一案

●原文の文法ミスに対して：

私が20年前にモルモット役となったソフトの場合、事前に原文を読み、文法ミスの有無を調べ訂正してから、機械翻訳していました。対策ソフトでは、原文にミスの疑いがある場合、強引に不鮮明な翻訳を強行せずに、エラーメッセージを表示できないでしょうか。

●原文が長文の場合：

前述の20年前のソフトは、事前に原文を読み、関係詞 when, which などがある場合、前後に「/」を挿入し、その部分を別個に機械翻訳していました。対策ソフトでは長文を自動的に短文化し、その枠内だけ翻訳し、前後の文節との連結はユーザの判断に任せてもかまいません。翻訳文が、不鮮明な状態で長々と続くより、箇条書きで巧みに表現したほうが理解しやすいかもしれません。

原文ミスは千差万別であり、統計的に論理的に分類不能と推定します。従って、ソフトでミスに完璧に対応することは無理でしょう。それを強引に訳しているから、不鮮明になると推測します。ミスをユーザに的確に示すことが、むしろ重要かと思います。

●原文の入力方法を一工夫すべきです。

英字新聞や書籍の記事を和訳する際に、キーボードから手で入力するのは厄介であり、誤入力の不安もあります。そこで、スキャナを用いてみました。しかし、英字の「i」と「l」と数字の「1」を正確に区別して訳してくれません。こういう問題をスキャナメーカーと共に解決してほしいです。

原文にミスが多いので、出力（機械翻訳）後にチェックが必要なことは覚悟しています。それだけに、入力だけは盲目的な作業で瞬時に確実にできないでしょうか。

機械翻訳には前述のように長所が多いですが、入力と出力の両方に細心の注意を払わなければならない現状は、普及に大きな障害になるでしょう。

（４）毎日使用している感想

機械翻訳は活用方法により部分的に非常に助かります。私は毎日用いています。ソフト価格は

1万円前後であり、多少の短所は我慢できる価格帯です。また、機械翻訳の研究開発から派生した翻訳支援ツールが、急激に進歩し、非常に役だっています。中世の欧州で錬金術の研究が行われ、化学の発展を促したのと同じだと思います。機械翻訳より古い30年以上も前から自動食器洗い機が開発されていました。期待されながら、未完成の状態は似たようなものと思います。最近、10万円前後で用途が限定されたモデルに人気が出てきたようです。こういうふうに、自動翻訳ソフトも、長所と短所の間に妥協点を見だし、ユーザ層を絞るべきかなと思います。メーカーにとって、販売戦略や採算性から課題があるでしょうが検討してください。

原文ミスや長文に起因する機械翻訳の悪例は、古い翻訳者ですので豊富にあります。紙数の都合もあるので、それらをアピールする方法の提案と共に、また機会があったら報告したいです。以上、皆様に役立てば幸いです。

（新潟県三条市在住）

インターネット機械翻訳サービス一覧

AAMT インターネットワークンググループ

- ★ 本ページでは、AAMT 独自調査による機械翻訳サービスを行なっている Web ページを記載しています。辞書引きサービスは原則として非掲載です。本ページに記載されていない機械翻訳サービス、あるいは記載されているサービスの URL 変更/停止などがありましたらご連絡頂けると幸いです。
- ★ 対象言語欄は、翻訳前あるいは翻訳後の言語として指定可能な言語を表し、翻訳方向は表していません。
- ★ 以下のサイトにアクセスしたことによる一切の損害を当協会は補償いたしません。
- ★ 2004 年 8 月 10 日現在/順不同

サイト (サービス) 名	翻訳対象/URL			対象言語
ajeeb	Web	Text		英アラビア
	http://english.ajeeb.com			
AltaVista Babel Fish Translation	Web	Text		英仏独西伊葡中日韓露
	http://world.altavista.com/			
AMIKAI		Text		英仏独西葡中日韓
	http://www.amikai.com			
@nifty	Web	Text		英日
	http://www.nifty.com/globalgate/			
ATLAS		Text		英日
	http://software.fujitsu.com/jp/atlas/sample.html			
BIGLOBE	Web		検索	英中日韓
	http://honyaku.biglobe.ne.jp/			
clickq Japan	Web			日韓
	http://www.clickq.com			
Cyber G-BOC	Web	Text		英日中
	http://www.g-boc.com/j/cyber/search/index.html			
英語ナビ	Web	Text		英日韓
	http://eigo-navi.com/			
ePal.com		Text		英仏独西伊葡中
	http://www.epals.com/translation/			
Excite	Web	Text	MS Office, 検索	英日中韓
	http://www.excite.co.jp/world/			
Foreignword.com		Text		25 言語
	http://www.foreignword.com/Tools/transnow.htm			
Free Translation.com	Web	Text		英仏独西蘭葡ノルウェイ
	http://www.FreeTranslation.com/			
i-Engineering		Text	File	英中日韓
	http://www.englinc21.com/i-eng/guest/10c10001.htm			

Infoseek	Web	Text		英中日韓
	http://www.infoseek.co.jp/Honyaku			
Internet Osaka		Text		英日
	http://ss2.inet-osaka.or.jp/			
石川新情報書府	Web			英日
	http://shofu.pref.ishikawa.jp/			
itranslator		Text		英仏独西伊葡日
	http://itranslator.bowneglobal.com			
J-Server	Web	Text	Chat, 掲示板	英中日韓
	http://www.j-server.com			
金沢 100 万石まつり	Web			英日
	http://www.100mangoku.net/			
韓国 Web 旅行	Web	Text	Chat, 掲示板	日韓
	http://korea.hanmir.com/			
関西コンシェルジェ			Chat, Mail, 掲示板	日韓
	http://www.my-concierge.org/kansai/index.htm			
国立健康・栄養研究所			健康関連ニュース	英中日
	http://humpty.nih.go.jp/			
Language Engineering	Web	Text	File, Mail	英仏独西伊葡中日
	http://www.lec.com/			
LogoMedia.net	Web	Text	File, Mail	英仏独西伊葡中日
	http://www.logomedia.net/			
LogoVista e-Trans	Web	Text		英仏独西伊葡日韓
	http://e-trans.logovista.co.jp/			
Lycos Travel		Text		英仏独西伊葡
	http://translation.lycos.com/			
祭りだ！高松コンシェルジェ			Chat, Mail, 掲示板	日韓
	http://www.my-concierge.org/shikoku/takamatsu/			
Microsoft Office Update Home		Text		英日
	http://software.fujitsu.com/jp/honyaku/eservice/index.html			
NAVER Japan	Web	Text	掲示板, 検索	日韓
	http://honyaku.naver.co.jp/			
Newstran.comm	Web	Text	News	英仏独伊西葡中日韓アラビア
	http://www.newstran.com/			
新居浜コンシェルジェ			Chat, Mail, 掲示板	日韓
	http://www.my-concierge.org/shikoku/niihama/			
Paran Go Japan!	Web			日韓
	http://gojapan.paran.com/index.html			
OCN	Web	Text		日英韓
	http://www.ocn.ne.jp/translation/			
Padideh		Text		英アラビア語
	http://www.padideh.org/trenglish.htm			

ParSit	Web	Text		英泰
	http://suparsit.com			
Poltran.com		Text		英葡
	http://www.poltran.com/			
ProMT	Web	Text	Mail	英仏独西伊露
	http://www.online-translator.com/text.asp?lang=en			
Q 州マイーコンシェルジェ	Web		Chat, Mail, 掲示板	日韓
	http://www.my-concierge.org/kyushu/			
ReaD		Text		英日
	http://read.jst.go.jp/			
Reverso	Web	Text		英仏独西
	http://www.reverso.net/			
Rustran.com		Text		英露
	http://www.rustran.com/			
Sakhr		Text		英アラビア
	http://mason.gmu.edu/~aross2/sakhr.htm			
瀬戸内・夢海道コンシェルジェ			Chat, Mail, 掲示板	日韓
	http://www.my-concierge.org/shikoku/syodoshima/			
Sister Cities International		Text		英仏独西伊葡日
	http://www.sister-cities.org/sci/resources/translation			
So-net	Web	Text		英日
	http://www.so-net.ne.jp/translation/			
SPORTS CHANBARA	Web			英仏独西伊葡中日韓露泰
	http://www.internationalsportschanbara.net/entry.html			
sScience.com		Text		英日
	http://www.sscience.com/index_J.htm			
Sysmeta		Text	Chat	日韓
	http://www.sysmeta.com			
SYSTRANBox	Web	Text		英仏独西伊葡中日韓露
	http://trans.voila.fr/			
Systranet.com	Web	Text	File	英仏独西伊葡希
	http://www.systranet.com/systran/net			
特許庁			特許公報	英日
	http://www.jpo.go.jp/			
TransLand		Text		英日
	http://www.brother.co.jp/jp/honyaku/demo/index.html			
Transtar Web	Web	Text		英中日
	http://www.transtar.com.cn/en/transtech/transonline.asp			
TransWalker JC		Text		日中
	http://www.kodensha.jp/wai2/pdamt/jpekin/data.html			
TransWalker JK		Text		日韓
	http://www.kodensha.jp/wai2/pdamt/jseoul/data.html			

Transwiz	Web	Text		英日韓
	http://transwiz.mobico.com/			
T-sail : Full & Magic	Web	Text		英仏独西伊葡日韓露
	http://www.t-mail.com/cgi-bin/tsail			
通連国際	Web	Text	Chat, File, Mail	英独中日露
	http://www.165net.com/jpindex.jsp			
VIL Net	Web	Text		英日
	http://www.vil-net.jp/hon/			
Yahoo Korea	Web			日韓
	http://kr.japan.yahoo.com/			
ToggleText	Web	Text		英インドネシア
	http://www.toggletext.com/			
金橋訳港世界通	Web	Text	Mail	英中日
	http://www.netat.net/translation/borwse.htm			
マジカルゲート	Web		Mail	英中日韓葡仏伊西独
	http://www.magicalgate.net/ja/index.html			
Al-Misbar	Web	Text		英アラビア
	http://www.almisbar.com/salam_trans.html			
中国馬鞍山		Text		英日中独露
	http://www.maanshan.gov.cn/zxfy.htm			
コネクト・インタナショナル			Mail	英日
	http://www.connect-i.co.jp/flextran/index.html			
the business mall			掲示板	日韓
	http://www.b-mall.ne.jp/bm0100/01_kcciBBS.aspx			
Tour2Korea	Web			日韓
	http://japanese.tour2korea.com/12Home/translator.asp?konum=7			
韓日文化交流	Web	Text	Mail	日韓
	http://www.kjclub.com/jp/trans/trans_kj.html			
alittera translation service		Text		英仏独伊西葡露
	http://www.alittera.com/Auto.asp			
Web-a-dex Internet Services	Web	Text		英仏独伊西葡露
	http://www.web-a-dex.com/translate.htm			
livedoor Translate	Web	Text		英日
	http://translate.livedoor.com/			

「CROSSROAD for Enterprise V2」

日本電気株式会社

1. はじめに

このたび弊社はイントラネット向け翻訳支援システム「CROSSROAD for Enterprise V2」を開発いたしました。本システムは企業内イントラネット上に設置するサーバー型の翻訳支援システムです。企業ユーザ殿による実績では、典型的な社員の翻訳に要する時間を約3割削減する効果が得られています。

2. 開発のねらい

近年の社会活動全体のグローバル化に伴い、一般の社会人が自分自身で英語を読んだり書いたりしなければならない機会が急増しています。本システムはこのような状況に対応した次のような特徴を備えています。

(1) 情報漏洩の心配なし

社内イントラネット上の翻訳サーバーに資源が集中するため、情報が社外にもれる心配がありません。最近ではWEB上に無料の翻訳サイトも存在しますが、組織の活動に関わる書類を社外サイトに無防備に提供することは情報漏洩の危険を伴います。本システムはこの点で安全です。

(2) 迅速に情報の受信・発信

本システムのもつ機械翻訳、定型ビジネス文例、翻訳メモリなど多彩な機能を使うことで迅速に情報の受信・発信が行なえます。日常的に必要な情報を一刻も早く入手するためには、利用者自身で英文を読んだり書いたりする必要があります。本システムはその負担の軽減に役立ちます。

(3) 自分たちの分野に適応

機械翻訳の品質を高めてシステムを有効活用するには、自分たちの分野に特有な単語を辞書登録する必要があります。しかし訳語まで含めて辞書に単語を登録する作業にはかなりのコストがかかります。本システムでは利用者間で部門辞書を共有できるので、ひとりひとりが作成した辞書を統合して利用部門全体の辞書を充実させてゆくことが可能です。

(4) 手軽で安価

本システムではサーバーだけを管理すれば良いためメンテナンスの手間を低く抑えられます。また導入部門全員でお使いいただけるため、ひとりあたりの費用を低く抑えられます。

3. システムの主な機能

本システムは、次のような多彩な機能で利用者の英文読解・作成を支援します。

(A) 機械翻訳

本システムは、英日・日英の機械翻訳機能を装備しています。まずざっと内容を翻訳して内容を把握する目的に適しています。機械翻訳の種類としては、直接入力した文を翻訳する「テキスト翻訳」、WEBページを翻訳する「WEB翻訳」、文書ファイルのまま翻訳する「ファイル翻訳」の各機能をご利用いただけます。

ファイル翻訳では次のファイル形式に対応しています。

- ・テキストファイル形式 (.txt)
- ・HTMLファイル形式
- ・Microsoft Office ファイル形式
(Word, Excel, PowerPoint)
- ・Acrobat ファイル形式 (.pdf)

(B) 日英対訳ビジネス例文集

英文を書くときに最も時間がかかるのは、単語や表現の使い方の確認です。そのコストを低減するために、典型的なビジネス状況で使う英文レターとその日本語訳の対訳文書例を豊富に用意しました(ホエールズウェブ・ドット・コム社提供)。また1文ごとの対訳文例も保持しています(研究社提供38万対訳を含む40万対訳)。日英・英日辞書も引けます。

対訳文書・対訳文の選択方法としては、木構造をたどる方法、キーワードを入力して検索する方法、書きたい文に近い類似文を得る方法、の3通りを用意しました。書きたい主題に近い対訳文書例を選択し、それをひな型として修正利用することで、従来に比べずっと短時間で英文レターが作成できます。

(C) 翻訳メモリ

マニュアルの改版がその典型ですが、翻訳済み文書を再利用し、それとの差分だけを翻訳することで新規翻訳コストを削減できる場合があります。この状況に対応するため、本システムには翻訳メモリ機能を搭載しました。新しく翻訳すべき文に対して、それと類似した文を既存の翻訳文書の中から検索します。次にその既存類似文に対する訳文を、新規翻訳対象文の下訳とします。類似した文がない場合に機械翻訳の結果を下訳に利用することもできます。本システムでは、原文と下訳とがエディタの左右に表示され、前後の流れを見ながら翻訳作業を効率的に進められます。

4. システム構成

本システムは翻訳サーバーと翻訳エンジンの2つの部分から構成されます。翻訳エンジンを複数台用意して処理速度を上げることが可能です。翻訳サーバーのOSの種類によって次の2つのタイプがあります。

(A) Windows 版

全体を1台に納めることができます。翻訳サーバーと翻訳エンジンを分離した構成も可能です。

(B) Linux-Windows 版

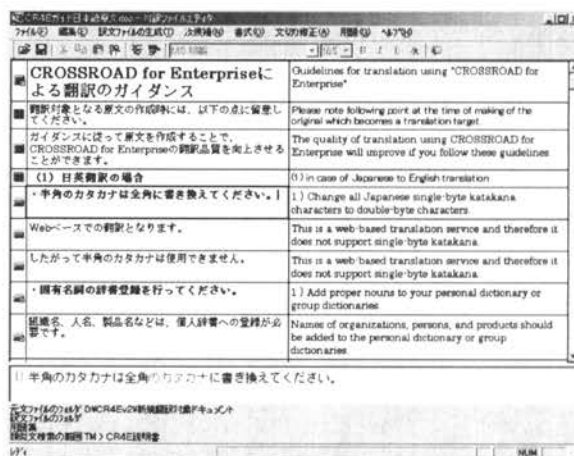
翻訳サーバーをLinux、翻訳エンジンをWindowsとした複数台構成です。

詳細は以下の紹介ページをご参照ください。

<http://www.sw.nec.co.jp/soft/crossroad-enterprise/>

お問い合わせ先：

support@cross.jp.nec.com



【 翻訳メモリのエディタ画面 】

「The翻訳オフィス™ V7.0」

東芝ソリューション株式会社

1. はじめに

東芝ソリューション株式会社は、ビジネス向け英日／日英翻訳ソフトの新商品として、ビジネス文書解析力を向上させるとともに、ビジネス辞書も増強し、総計182万語を標準搭載した「The翻訳オフィスV7.0」を、6月25日に発売しました。この商品は、翻訳精度の決め手となる翻訳エンジンの文書解析力の向上と、辞書語数の増強により、特にビジネス分野での翻訳精度をアップさせています。翻訳エンジンでは、海外からの技術レポートや製品情報などビジネスシーンに登場する表現の解析力を向上させることにより翻訳精度を従来商品より一層高めています。また、ビジネス用語を中心に辞書語数を従来商品に32万語追加した、182万語（英日93万語、日英89万語）を標準搭載しています。

さらに、新機能の「対訳エディタ」では、文単位・単語単位での原文と訳文の対応表示のほか、翻訳結果の確認や新語の登録、語句の学習など翻訳に必要な作業を一つの画面で行うことができます。

また、このような豊富な機能の中から使いたい機能に簡単に到達できるよう「The翻訳シンプルナビ」を搭載しました。「The翻訳シンプルナビ」は利用者の習熟度に従って画面を変更できるため、初めて使う方から習熟者まで、幅広いユーザ層に優れた操作環境を提供します。

2. The翻訳オフィス V7.0の主な機能

(1) 使いやすい「対訳エディタ」

対訳エディタは、原文と訳文を対比表示しながら翻訳する作業が、一画面上で切り換えずにできる機能です。読み込んだテキストの1文ずつの対応表示、原文と訳文を見比べながら1文ずつ翻訳、あるいは全文まとめたの翻訳、原文と訳文を対応させた形式での印刷もできます。さらに、翻訳結果の確認や新語の登録、語句の学習なども実行可能です。また、翻訳結果を原文と訳文の対応を保ったままMicrosoft® Excelへコピーできるため、対訳形式での文書配布にも適しています。

この他、音声ツールとの連携による文字入力、翻訳結果の読み上げや、翻訳結果をメールエディタに出力したり、メールを原文として取り込むことも可能で、様々な翻訳作業に利用する事ができるツールになっています。



図1：対訳エディタ画面

(2) 使いたい翻訳機能がすぐ使える「The翻訳シンプルナビ」

The翻訳の豊富な機能から使いたい機能を初心者でも容易に選べるようにThe翻訳シンプルナビ機能を搭載しました。行いたい作業の概要が書かれたガイドボタンを押すことで、必要な機能がすぐ使える状態で起動され、機能の使い方も同じ画面に表示されるようになっています。

また、使い慣れた人向けに、よく使う機能だけを選択しコンパクトなボタンで表示できるような設定も可能で、初めて使う方から習熟者まで幅広い利用者が個人にあった操作性を得ることが可能になっています。



図2: The翻訳シンプルナビ画面

(3) ビジネス辞書を増強し、総計182万語を標準搭載

翻訳エンジンのビジネス文書解析力を向上、ビジネス辞書を増強することにより特にビジネス分野での翻訳精度を高めています。搭載辞書語数は、従来のV6.0に比べて、英日を76万語から93万語に、日英を74万語から89万語に増強しました。

(4) 翻訳したい分野の文書登録で、翻訳精度を高める「セレクトコーパス翻訳機能」

翻訳したい分野の文書をコーパス^{*1}に登録すると、登録文書で使用されていた語句が優先して使用されるので、その分野に適切な翻訳結果を得ることができます。例えば、日本語の証券関連の文書を登録してから、本機能を使うと、「The stock of this brand brings profit.」という原文を、「この銘柄の株式は利益をもたらします。」と、初めから適切に翻訳することが可能です。

このように本機能は、登録する文書が英日のペアである必要がなく、英日翻訳ならば日本語、日英翻訳ならば英語の文書だけの登録でよいため、参考にする文書を集めやすいという特長があります。

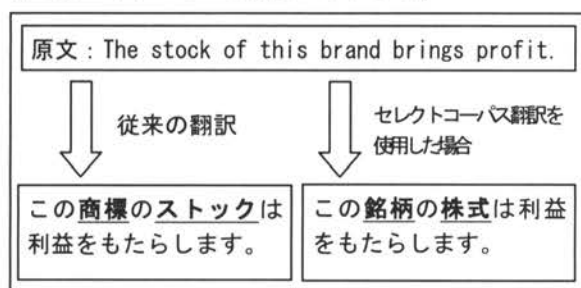


図3：従来の翻訳との比較

(5) Microsoft® Office2003など、さまざまなビジネスソフトウェアの最新版に対応

さまざまなビジネスソフトウェアの最新版に対応しました。Microsoft® Office2003をはじめ、Adobe® Acrobat®6.0/Adobe Reader® 6.0、Opera7~7.20、一太郎2004などに翻訳機能が組み込まれます。

(6) 自動アップデートサービスの提供

インターネット経由でThe翻訳オフィスのレビジョンアップを行うサービスパッチ検索とダウンロード・セットアップを実行する自動アップデートサービス機能を提供しています。これにより、ユーザは簡単な操作で最新のレビジョンに更新することができます。

(7) 中国語の日本語への訳振りに発音表示を実現

ホームページ全体やホームページ上の指定範囲の中

国語に日本語訳振り機能を搭載しています。更に、マウスカーソルで文字をポイントして自動的に訳振りをする際には、ピンイン^{*2}の表示が可能です。

(8) Microsoft® Wordからシームレスに呼び出せる「翻訳OCR機能」

Microsoft® Word画面上に組み込まれたボタンをクリックするだけで、「翻訳OCR機能」を呼び出し、原稿のスキャン・認識・翻訳という一連の作業を実行、結果を直接Microsoft® Wordに出力することができ、印刷原稿の翻訳が簡単に行えます。

(9) 研究社新英和・和英中辞典の標準搭載

翻訳家を始め広く使われている研究社新英和・和英中辞典の標準搭載を電子辞書として搭載しており、翻訳実務でニーズの高い辞書引きが可能です。

3. The翻訳オフィス V7.0の動作環境（全て日本語版）

対象機種：PC/AT互換機

CPU：Pentium®以上/Celeron™以上推奨

メモリ：1ツールあたり64MB以上

ディスク：最小300MB～フルインストール710MB

対応OS：Microsoft® Windows® 98/98SE/Me/

2000 Professional/ XP Home Edition/ XP Professional

連携可能なソフトウェアおよびバージョンにつきましては、製品情報URLをご覧ください。

■ 製品情報URL

<http://hon-yaku.toshiba-sol.co.jp/>

- Pentium、Celeronは米国インテル社の商標または登録商標です。
- Microsoft、Windowsは米国Microsoft Corporationの米国およびその他の国における登録商標または商標です。
- Adobe、Adobe Acrobat、Adobe Readerは、Adobe Systems Incorporated（アドビシステムズ社）の米国およびその他の国における登録商標または商標です。
- The翻訳、The翻訳オフィスは東芝ソリューション株式会社の商標です。
- その他、本文中の商品名称はそれぞれ各社が商標として使用している場合があります。
- 本製品の仕様は予告なく変更される場合があります。

^{*1} ここでは、狭義の意味で電子化文書を収集し、データベース化したものを指す。広義には資料文献の集成、全集の意味。

^{*2} 中国語の発音記号

「専門分野 翻訳これ一本 ビジネス・科学技術編 V2」

シャープ株式会社

1. はじめに

シャープは、原文と訳文を一組にした 26 万の文例に加え、新たに翻訳した文例を登録しデータベースとして活用できる「翻訳メモリ」を搭載した英日・日英翻訳支援ソフト「専門分野 翻訳これ一本 ビジネス・科学技術編 V2」を 2004 年 7 月 23 日に発売しました。

シャープヨーロッパ研究所（英国オックスフォード）との共同開発により、翻訳したい文と類似した文例を「翻訳メモリ」から自動的に検索・引用する“類似文検索”と、その際、原文と検索された文例の異なる箇所についても自動的に適切な訳語に変換する“非マッチ部訳語自動変換”を実現し、翻訳精度と作業効率をさらに高めることができました。また、本ソフトを使用しているパソコンの「翻訳メモリ」データは、LAN環境で共有利用できるため、社内や学校など、グループ内の複数人で共通の翻訳環境を構築できるようになりました。

2. 主な特長

(1) 対訳文例をデータベースとして活用できる「翻訳メモリ」を搭載

最初からメモリとして登録してある 26 万対訳文例に加え、新たに翻訳した文例を登録することができ、このデータベースを活用すれば、次回以降の翻訳作業の負担を大幅に軽減できます。

(2) 入力原文と検索された文例の異なる箇所も自動的に適切な訳語に変換する“類似文検索+非マッチ部訳語自動変換”を搭載

「翻訳メモリ」内に原文と類似した対訳文例があれば自動的に検索・引用します。

また、原文と検索された対訳文例との間に部分的に異なる箇所（非マッチ部）があった場合でも、そこに対応する訳語を可能な限り適切な訳語に自動変換します。

原文・訳文が一目でわかる「対訳エディタ」により、翻訳手段に応じた識別表示ができます。

さらに、非マッチ部の識別表示もできるため、訳文の見直し作業も効率的です。

(3) LAN環境で「翻訳メモリ」の共有が可能

LANによって接続されたグループで「翻訳メモリ」のデータを共有・蓄積することも可能です。このとき、共有PCにサーバOSは不要です。個々のPCにインストールされた「翻訳メモリ」は、メンバー各自が個人管理できる部分と、メンバー全員が共有できる部分を持つ構成となり、検索は個人管理の翻訳メモリ部が最優先されます。共有PCの翻訳メモリデータの編集（登録・削除・修正）については、管理者PCによる制限運用もできます。

(4) 最新版 英日・日英翻訳エンジンを搭載

当社特長技術のマルチレイヤー翻訳方式で自然に読みやすい訳文を出力します。「英英変換→英日翻訳→日英変換」の自動3段階処理により、きめ細かく翻訳し、事実関係のよくわかる読みやすい文章に仕上げます。

最新版の翻訳エンジンを搭載し、かつ「翻訳メモリ」検索エンジンと併用することによって翻訳これ一本の翻訳精度がさらにアップしました。

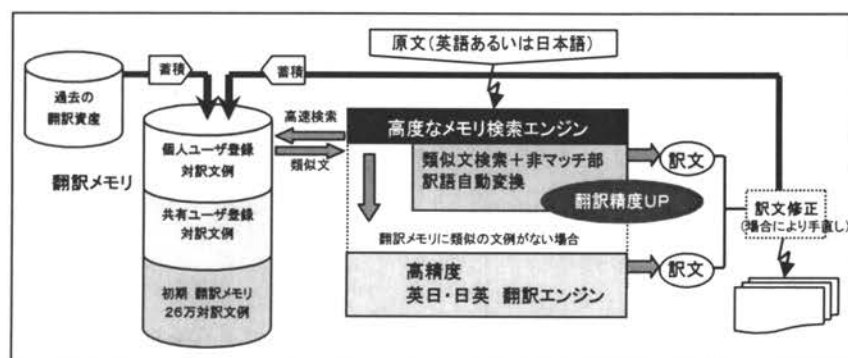


図1 翻訳メモリを利用した翻訳動作の概念図

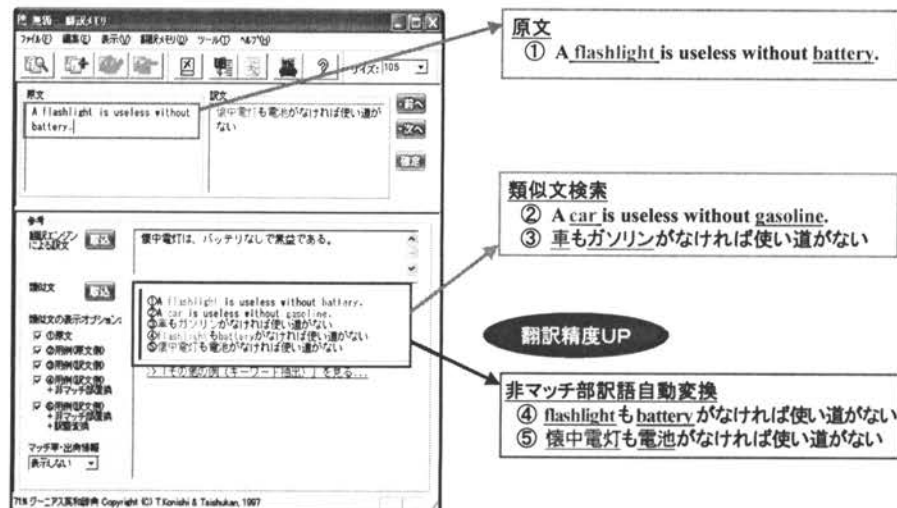


図2 【翻訳メモリ】の検索/編集画面例

(5) 専門辞書など約5万語を増強した253.8万語の翻訳辞書を搭載

専門用語を中心に、旧製品から英日・日英合計で約5万語を増強し、合計253.8万語の翻訳辞書を搭載しました。これにより、従来以上の翻訳精度を実現します。

	旧製品 (ビジネス・科学 技術編)	新製品 (ビジネス・科学 技術編V2)
英日翻訳辞書	1,070,000語	1,110,000語
日英翻訳辞書	1,410,000語	1,428,000語
合計	2,480,000語	2,538,000語

(6) 英文速読や英語学習を支援するシャープの特長技術「おまかせ訳振り」

(シャープヨーロッパ研究所との共同開発)

英文中の単語や連語の下にルビのように訳語を表示する「おまかせ訳振り」機能は、英語の下調べや英文速読に便利で、従来より好評をいただいております。本機能では、英日翻訳用のユーザ辞書に登録された単語や連語の訳語表示ができ、さらに使用している画面から直接ユーザ辞書に単語や連語を登録することもできます。ここで新規登録した語彙は、英日翻訳でも利用できます。

(7) Word 2003 や Adobe Reader 6.0 などの最新定番ソフトとの連携翻訳も可能

「Word 翻訳」や「PDF 翻訳」の機能をアップし、最新版 Microsoft Word 2003 や Adobe Reader 6.0 (あるいは Acrobat 6.0)に対応しました。さらに、「Word 翻訳」では、今回あらたに上下対訳表示が行えるようになりました。

3. 専門分野 翻訳これ一本 ビジネス・科学技術編 V2の動作環境

対応パソコン	下記のOSが動作し、CPUがPentiumプロセッサ133MHz以上のパソコン (WindowsXPの場合はPentiumIIプロセッサ300MHz以上)
OS	Microsoft WindowsXP/2000/Me/98 または Microsoft WindowsNT4.0 operating system(日本語版)
メモリ	・128MB以上推奨 (Microsoft Windows 2000/Me/98、Microsoft WindowsNT4.0の場合) ・256MB以上推奨 (Microsoft Windows XPの場合)
ハードディスク	空き容量250MB(最小)~650MB(最大)、分割インストール可能
ディスプレイ	SVGA(800×600ドット)以上/65,536色(ハイカラー)以上の解像度をもつカラーディスプレイ

■ 翻訳これ一本 ホームページ

翻訳これ一本シリーズの商品情報に加え、サポート情報など、翻訳これ一本に関する総合的な情報をご提供しています。<http://www.sharp.co.jp/ej/>

■ バージョンアップ版のご案内

オンラインショッピングサイト「SHARP PC ONLINE」では、従来の翻訳これ一本シリーズからのバージョンアップ版を発売しています。

<https://www.sharppconline.jp/>

■ 企業・団体ユーザ様向け追加ライセンス版のご案内

企業・団体ユーザ様向け追加ライセンス版の販売については、最寄りのソフト販売店様にご相談ください。

「コリヤ英和！一発翻訳 Ver. 6.0 マルチリンガル」

ロゴヴィスタ株式会社

1. はじめに

ロゴヴィスタは、Windows 対応 8ヶ国語翻訳ソフト「コリヤ英和！一発翻訳 Ver. 6.0 マルチリンガル」を、2004年8月6日より発売しました。

本製品は低価格翻訳ソフトのベストセラーとして累計出荷190万本を超える「コリヤ英和！」シリーズの8ヶ国語翻訳ソフトの最新バージョンです。ニュース、スポーツ、ファッション、トラベルなど、世界8ヶ国語のWebサイトやメールをワンタッチで日本語に翻訳できるマルチリンガル対応翻訳ソフトです。

「英語・日本語 合計 231 万語」＋「ヨーロッパ言語・韓国語 合計 143 万語」の豊富な翻訳辞書を収録し、「英語・日本語・フランス語・ドイツ語・イタリア語・ポルトガル語・スペイン語・韓国語」の8ヶ国語を自在にクロス翻訳する「LogoVista X 翻訳エンジン」を搭載。言語の組み合わせにより、合計 46 通りの翻訳方向を実現できます。また、8ヶ国語の「翻訳辞書引き」や使いやすい「操作パネル」、さらには英日・日英翻訳専用ツールのフル搭載と、機能面も充実。海外の最新情報の入手に最適なマルチリンガル・コミュニケーション・ツールです。

また、本製品には姉妹品の英日・日英翻訳ソフト「コリヤ英和！一発翻訳 Ver. 6.0」の全ての機能を搭載しています。

2. 「コリヤ英和！一発翻訳 Ver. 6.0 マルチリンガル」の主な特徴

(1) 8ヶ国語をクロスする LogoVista X 翻訳エンジン

世界の 8ヶ国語を自在にクロス翻訳する「LogoVista X 翻訳エンジン」を搭載。「英語・日本語・フランス語・ドイツ語・イタリア語・ポルト

ガル語・スペイン語・韓国語」に対応し、翻訳エンジンを組み合わせることにより、合計 46 通りの翻訳方向を実現。Ver. 6.0 では、英日・日英翻訳エンジンの改良とヨーロッパ言語翻訳エンジンのブラッシュアップにより、さらに精度が向上。

●翻訳辞書：英語・日本語 合計 231 万語 ＋ ヨーロッパ言語・韓国語 合計 143 万語

●ユーザ辞書：英日、日英専用（複数作成可能）

●専門辞書（別売）：英日、日英専用

（英日 5 分野 87 万語、日英 24 分野 239 万語）

(2) 世界の Web サイトをワンタッチで日本語に Internet Explorer に追加された翻訳ツールバーからボタンひとつで海外のホームページを日本語に翻訳・辞書引きすることができます。Ver. 6.0 では日本語キーワードで外国語の Web サイトを検索し、自動翻訳する「一発サーチ翻訳」機能も新搭載。

(3) メール・ニュースの翻訳もマルチリンガル対応メール・ニュース、その他テキストもマルチリンガル翻訳が可能。翻訳したい文章を選んで「操作パネル」のボタンをクリックするだけで翻訳する「クイック翻訳」や、キーボード入力により FEP 感覚で翻訳変換できる「タイピング翻訳」を搭載。

(4) 8ヶ国語の翻訳辞書引き機能がより使いやすく 8ヶ国語をクロスして合計 46 通りの辞書引きを行うことができる「翻訳辞書ビューワ」を搭載。辞書引きの操作も簡単になり、いっそう使い勝手が向上。

(5) 本格的な翻訳に最適な「翻訳エディタ」

文章をじっくり翻訳するのに最適な「翻訳エディタ」を新搭載。完全左右対訳の表示なので、各センテンスを比較しやすく、さまざまな編集補助機能を利用することが可能。

(6)Microsoft Wordに翻訳機能を組み込み

マルチリンガルの翻訳ツールバーを Microsoft Word にアドイン（組み込み）することができます。普段お使いの Word でダイレクトに翻訳・辞書引き機能を利用できます。

(7)8カ国語を音声読上げ

音声合成エンジンによる「英語・日本語・フランス語・ドイツ語・イタリア語・ポルトガル語・スペイン語・韓国語」の8カ国語に対応した音声読み上げが可能です。発音のチェックもOK。

(8)他にも多彩な機能を搭載

●はじめてでもわかりやすい

ボタン式「操作パネル」

●Netscape や Opera で

Web 翻訳「HTTP プロキシ翻訳」

●「LogoVista 電子辞典シリーズ」との

連携辞書引き

●海外旅行や出張に便利な

「世界時計/度量衡/カレンダー」

(9)英日・日英翻訳専用ツールも搭載

マルチリンガル翻訳機能に加えて、英日・日英翻訳専用の豊富な機能を搭載。姉妹品の「コリャ英和！一発翻訳 Ver. 6.0」(メーカー希望小売価格：9,800円税込)のすべての翻訳機能を利用できます。

3. 製品仕様

(1)価格

14,800 円 (税込価格)

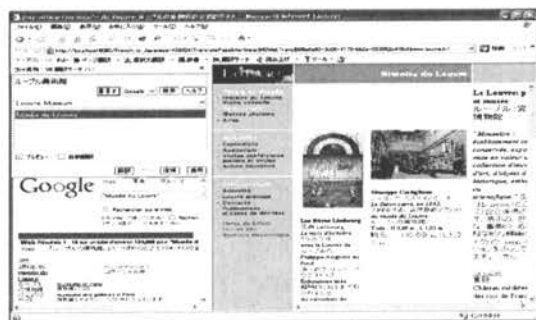


図-1 一発サーチ翻訳

(2)動作環境

●対応 OS： Microsoft Windows XP/Me/98SE/2000 Professional 日本語版

●対応機種： PentiumII 400MHz 以上の CPU 搭載機種を推奨

●必要メモリ： 64MB 以上 (128MB 以上を推奨)

●必要ハードディスク容量(フルインストール時)： 600MB 以上

●対応ブラウザ・対応メールソフト：

Microsoft Internet Explorer 6.0/5.5、Netscape 7.1/6.2/4.7、Opera 7.23、POP3 対応メールソフト (すべて日本語版)

●アドイン対応アプリケーション：

【マルチリンガル翻訳の場合】

Word 2000～2003 (日本語版)

【英日・日英翻訳の場合】

Adobe Reader 6.0、Adobe Acrobat 6.0/5.0、Word 2000～2003、Excel 2000～2003、PowerPoint 2000～2003、Outlook 2000～2003、一太郎 2004/13 (すべて日本語版)

●音声読上げに必要な環境： Sound Blaster 互換 16 ビットサウンドボード

4. お問い合わせ先

ロゴヴィスタ株式会社

サポートセンター

TEL 03-5690-9167

<http://www.logovista.co.jp>



図-2 パッケージ画像

「MAC-Transer V2」

株式会社クロスランゲージ

本製品は、Mac OS X 版業務用翻訳ソフトの最新版で、最新用語を増強し、英日 64 万語・日英 84 万語の基本語辞書と、改良した翻訳エンジンを搭載しています。

新機能として、翻訳に使用する各辞書に、表示の色を指定することが可能になりました。たとえば、コンピュータ辞書に「赤」を指定して、翻訳を実行すると、コンピュータ辞書からの訳語は赤色で表示されます。ひと目見るだけで、訳語の出典が明らかになります。また、用語集などがあれば、一括してユーザー辞書に登録できる辞書ツール機能が追加されました。

本製品は、PDF ファイルや Microsoft Word の doc ファイル、RTF ファイルを直接、読み込み翻訳することができます。また新たに Acrobat 6 の PDF ファイルにも対応しました。ホームページ翻訳では、Mac OS X 標準 Web ブラウザの『Safari』に対応しています。

さらに、辞書引きソフトの「ロボワード for トラランサー」をバンドルしており、デジタル辞典として、研究社新英和中・新和英中辞典を収録しています。追加辞典として『MAC-Transer V2 スタンダード with リーダーズ』と、『MAC-Transer V2 プロフェッショナル』には、「研究社リーダーズ+プラス V 2」を収録しています。

【本製品の特長】

◆英日 64 万語、日英 84 万語合計 148 万語の翻訳辞書で翻訳精度が向上【機能アップ】

見出し語で英日 64 万語、日英 84 万語の基本語辞書を搭載しました。改良した翻訳エンジンとともに高精度の翻訳を実現。

◆新改良の日英翻訳エンジンで、日本語から英語の翻訳精度が向上【機能アップ】

日英エンジンの日本語文の区切り解析（形態素解析）に新たなロジックを盛り込み、全面改良を加えました。また、カタカナ表記の「ゆらぎ」（例えば、インデクス／インデックス）といった違いを翻訳エンジン内部で自動補正する機能を新しく搭載しています。

◆訳語カラー表示機能付き高機能対訳エディタ【新機能】

基本語辞書、ユーザー辞書、学習辞書、専門語辞書などの使用する辞書に表示色を指定でき、訳語の出典がひと目でわかります。訳語対応や別訳語をその場で表示するポップアップ訳語メニュー、フレーズ指定機能を備えた高機能対訳エディタを英日・日英ともに搭載しています。

◆PDF ファイル、Word ファイル、RTF ファイルをダイレクトに読み込み翻訳【機能アップ】

PDF ファイル、Word ファイル、RTF ファイル、テキストファイルを直接、対訳エディタに読み込み、翻訳できます。Adobe Acrobat 6 対応の PDF ファイルにも対応しています。

◆プロフェッショナルには翻訳メモリ機能と 24 分野の専門語辞書を搭載

基本翻訳メモリ 33,000 例、科学技術翻訳メモリ 54,000 例、ビジネス翻訳メモリ 6,000 例合計 93,000 例搭載しています。新しく「会計」「マーケティング」「生産工学」を加えた 24 分野の専門語辞書を同梱。



株式会社クロスランゲージ 営業部

〒169-0051 東京都新宿区西早稲田2-20-9

ウィン西早稲田ビル7F

TEL03-5287-7588 FAX03-5287-7589

<http://www.crosslanguage.co.jp/>

<http://www.transer.com/>

事務局だより

事務局長交代のお知らせ

平成 16 年 4 月 1 日付で、村田勇造氏の後任として、^{なかせまこと}中瀬真氏が事務局長（社団法人電子情報技術産業協会(JEITA)からの派遣による非常勤職員）に着任しました。

引き続きご支援、ご鞭撻を賜りますようお願い申し上げます。

「ATLAS V11」

富士通株式会社

1. はじめに

「ATLAS」は、予てより企業・官公庁・教育機関の“実務翻訳”の現場より圧倒的な支持を得ている翻訳ソフトである。最新版「ATLAS V11」では、翻訳資産を効果的に利用することが可能な「Cliche(クリシェ)ディスプレイ」機能を標準搭載。

業界最高水準の翻訳品質

「ATLAS V11」では、翻訳ソフトの要でもある翻訳品質を向上させるため、「基本辞書」と「翻訳エンジン」の両面で改善を図った。「基本辞書」には、英日 91 万語・日英 89 万語を収録。基本の翻訳エンジンについては、一般的な文章解析・訳文生成の品質向上を行い、メール文、特許文の翻訳精度向上を実現。

2. 機能

Cliche ディスプレイ

翻訳メモリの検索結果が一覧で見やすく表示され、原文と訳文の単語の一致・不一致が視覚的に判断できる。これにより、例文を参考にして訳文を簡単に作成することが可能。

番号	原文	訳文	類似度
1	サービスの詳細については、「お客様ガイド」をご覧ください。	For the details of the service, see the "Service Guide".	97%
2	弊社は製品がお客様に役立つことを願っています。	We hope that our products will help to you.	85%

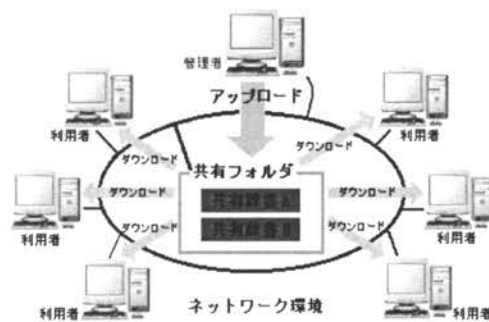


番号	原文	訳文	類似度
1	サービスの詳細については、「お客様ガイド」をご覧ください。	For the details of the service, see the "Service Guide".	97%
2	弊社は製品がお客様に役立つことを願っています。	We hope that our products will help to you.	85%

Cliche ディスプレイ

共有辞書ツール

ネットワーク上に構築された共有フォルダ内に共有辞書をアップロードすることにより、グループ内の管理者や利用者が同一の辞書を利用できる。この機能を利用することで、翻訳品質の均質化を実現するだけでなく、知的財産を利用者全員で共有することが可能。



辞書共有の概念図

共有辞書がアップロードできるのは一人の管理者のみであるが、各利用者がある日時以降の単語を抽出し、管理者がそのデータを共有辞書にマージすることができるツールも用意している。

翻訳メモリ

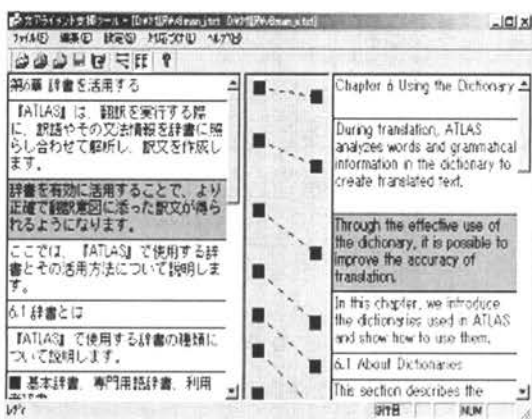
原文と訳文を対にした翻訳例を翻訳メモリとして登録できる。蓄積されたデータ内に、翻訳したい文と類似の文が検出された場合、その類似度（一致率）の順番で、表示することができる。10 万文対の翻訳メモリデータを標準搭載している。

翻訳メモリ変換ツール

翻訳メモリデータ交換用の標準フォーマットである TMX 形式のファイルと、ATLAS の翻訳メモリのデータを相互に変換することができる。また、翻訳メモリのツールとしては一般的な TRADOS 社製「TRADOS」のデータとも相互変換が可能である。

文アライメント支援ツール

過去に翻訳した文書の原文と訳文を読み込み、翻訳メモリに変換するための支援ツールである。文書形式はテキストまたは Word 文書を読み込むことができる。入力文書を読み込むと、自動的に文切りし、更に、自動的に対応付けを行うこともできる。このツールによって、過去の翻訳資産を有効に利用することができる。



文アライメント支援ツール

対訳エディタ

対訳エディタとは、原文と訳文を左右に表示し、編集作業を行いながら翻訳を可能とする機能である。編集と翻訳を同一画面で行うことができ、各機能を簡単に呼び出せることから、効率の良い作業を実現できる。

メール翻訳 (Outlook Express、LotusNotes、Outlook)

メールソフトに、ATLAS のメニューコマンドまたはツールバーを追加し、受信したメール、これから送信するメールを翻訳することができる。

ホームページ翻訳

ホームページ翻訳は、Internet Explorer または Netscape と連携して機能する。日本語のホームページを英語に、英語のホームページを日本語にクリックひとつでレイアウトを保持したまま翻訳を行うことができる。

おまかせ記文

対話形式で問いに答えるだけで、英語・日本語のレターを自動作成するツールである。

Internet Update

富士通の公開ページより、最新の時事用語やスポーツ、音楽等のエンターテインメント辞書が無償でダウンロードできる。

連携翻訳 (Word、Excel、PowerPoint、Acrobat、OASYS、一太郎)

原文ファイルのレイアウトそのままに翻訳する。Microsoft Office では、テキストボックスなどのオブジェクトや表の内容も翻訳できるので効率的に訳文ファイルを作成できる。Microsoft Office や Adobe Acrobat などのアプリケーションから直接対訳エディタを呼び出し、そのまま翻訳、編集して元の文書へレイアウトを保持したまま結果を反映することができる。

元の文書 (Word 文書)



対訳エディタで翻訳&後編集



対訳エディタで翻訳&後編集



Excel からの単語一括登録

Excel ワークシートから辞書へ直接、単語一括登録が可能。



Excel からの単語一括登録

専門用語辞書

専門分野向けに 25 分野、英日 217 万語・日英 221 万語の専門用語辞書をオプションで用意している。

専門用語辞書フリーパックの中から医学・薬学系の辞書をパックした「医学・薬学専門辞書」(9 分野、173 万語)および IT・製造業系の辞書をパックした「IT・製造業専門辞書」(13 分野、202 万語)も用意。

また、ATLAS V11 本体に収録されている「英和・和英電子辞書」用の専門用語データも 5 分野 227 万語を収録している。

専門用語辞書一覧

分野	英日	日英
1. 情報処理	138,000 語	139,000 語
2. 電気・電子	96,000 語	98,000 語
3. 物理・原子力	166,000 語	167,000 語
4. 機械	78,000 語	80,000 語
5. 工業化学	135,000 語	137,000 語
6. プラント	62,000 語	64,000 語
7. 土木・建築	52,000 語	53,000 語
8. 金属	44,000 語	45,000 語
9. 地学・天文	71,000 語	73,000 語
10. 輸送	60,000 語	61,000 語
11. 自動車	42,000 語	43,000 語
12. 軍事	25,000 語	26,000 語
13. 農林水産	68,000 語	69,000 語

分野	英日	日英
14. 生物	159,000 語	161,000 語
15. 医学 (生化学)	146,000 語	147,000 語
16. 医学 (薬学)	64,000 語	65,000 語
17. 医学 (解剖学)	56,000 語	58,000 語
18. 医学 (疾患・症状)	180,000 語	182,000 語
19. 医学 (精神医学)	31,000 語	32,000 語
20. 医学 (医療機器)	52,000 語	53,000 語
21. 金融・経済	63,000 語	64,000 語
22. 法律	8,000 語	10,000 語
23. ビジネス	131,000 語	132,000 語
24. 人名・地名	209,000 語	211,000 語
25. 環境	34,000 語	35,000 語

3. 動作環境

動作 OS: 日本語 Windows XP/Me/2000/98/NT4.0

必要メモリ: 32MB 以上 (Windows XP では 64MB 以上)

必要ディスク容量: 翻訳ダブルパック 445MB

4. 価格 (税込)

ATLAS 翻訳ダブルパック V11	92,400 円
ATLAS 専門用語辞書フリーパック V11	81,900 円
ATLAS 翻訳スーパーパック V11	134,400 円
ATLAS 医学・薬学専門辞書 V11	9,900 円
ATLAS IT・製造業専門用語辞書 V11	39,900 円

「ATLAS」の旧版ユーザ向けには、バージョンアップ製品を販売している。詳しくは下記の URL を参照のこと。

<http://software.fujitsu.com/jp/atlas/versionup.html>

■お問い合わせ先

富士通株式会社

プラットフォームソリューションセンター

プロダクトマーケティング統括部ミドルウェア部

〒105-7123

東京都港区東新橋 1-5-2 汐留シティセンター

T E L : 03-6252-2661

受付時間: 9:00~12:00/13:00~17:00 (土日祝日を除く)

<http://software.fujitsu.com/jp/atlas>

日本の機械翻訳の歴史

AAMT インターネットワーキンググループ

当協会独自調査による日本の機械翻訳の歴史年表です（同一年内の事項の前後関係は考慮されていません）。

1959	機械翻訳専用機「やまと」（電気試験所）、翻訳実験用計算機「KT-1」（九大）完成
1964	「MT 日米セミナー」開催[東京]
1978	電子協機械翻訳システム調査委員会発足
1980	電子協MT調査団ヨーロッパ、アメリカ、カナダ訪問
1981	アナロジーに基づく翻訳方式を世界で初めて提唱(京都大学 長尾眞教授)
1982	Mu プロジェクト開始
1985	科学万博に翻訳システム出展(富士通) 初の商用翻訳ソフト発売(ブラビス)
1986	自動翻訳電話研究所発足(ATR)
1987	「第1回 MT SUMMIT(電子協主催)」開催[箱根] 「近隣諸国間機械翻訳システム」開発プロジェクト開始(CICC)
1989	海外ニュース自動翻訳放送試行開始(NHK) 「翻訳技術国際フォーラム(電子協主催)」開催[大磯]
1990	「自動翻訳フェア(日本翻訳協会主催)」開催[東京]
1991	日本機械翻訳協会設立
1992	アジア太平洋機械翻訳協会設立 翻訳ソフト初の the Software Products of the Year 受賞(東芝) 「MT World 92(JAMT 主催)」開催[東京] 「機械翻訳に関するシンポジウム(電子協主催)」開催[大阪] 「機械翻訳フェア(日本科学技術情報センター主催)」開催[筑波]
1993	「第4回 MT SUMMIT(AAMT 主催)」開催[神戸] 「翻訳フェア'93(AAMT 主催)」開催[東京]
1994	1万円を切る翻訳ソフト登場(カテナ) 「多言語機械翻訳国際シンポジウム'94(CICC 主催)」開催[東京]
1995	初のブラウザ連動型翻訳ソフト発売開始(沖) 「パソコン翻訳ショウ(日刊工業新聞主催)」開催[東京]
1996	MT プレインストールパソコン発売開始(シャープ、東芝等)
1997	機械翻訳ユーザ辞書共通フォーマット設定(AAMT) 日本人初の MT Award 受賞(京都大学 長尾眞教授)
1999	機械翻訳による日本特許英文検索サービス開始(特許庁) 「第7回 MT SUMMIT(AAMT 主催)」開催[シンガポール]

2000	機械翻訳白書「機械翻訳 21世紀のビジョン」発行(AAMT) 「MTセミナー(AAMT主催)」開催[東京] インターネット博覧会で翻訳ソフト正式採用(日本電気) 翻訳ソフト初のプロレジ大賞受賞(ノヴァ)
2001	翻訳ソフト初のグッドデザイン賞受賞(IBM)
2002	5,000円を切る翻訳ソフト登場(ホロン) FIFAワールドカップ公認翻訳ソフト登場(東芝) ヒット商品番付東前頭6枚目に翻訳ソフト選出(SMBCコンサルティング)
2003	2,000円を切る翻訳ソフト登場(ソースネクスト)
2004	成田空港で旅行者向け自動通訳機実験(国土交通省)

編集後記

AAMT ジャーナル編集委員会委員長 井佐原 均

猛暑と台風の続いた今年の夏でしたが、皆様はいかがお過ごしでしょうか。ようやく秋めいてきた季節と共に、AAMT ジャーナル 36 号をお送りいたします。発行が遅れたことを、お詫びいたします。

本号は、従来にも増して充実した内容になっておりますので、ご満足いただけることと存じます。まず、AAMT の初代会長（現理事）の長尾真先生（前京都大学総長、現情報通信研究機構理事長）から、巻頭言として、機械翻訳の今後についてのお話をいただきました。また、技術解説として、東京大学中川裕志先生と、岐阜大学池田尚志先生に、それぞれの研究を分かりやすく解説していただきました。

日本以外のアジア圏のアクティビティの紹介として、本年 3 月に中国海南島で開催された IJCNLP の報告とインドにおける機械翻訳の活用法の記事を載せております。新製品紹介には 6 つのシステムが紹介されています。

インターネットワーキンググループの活動としてまとめた「日本の機械翻訳の歴史」を前ページに載せておりますが、より網羅的な年表とするために、皆様からのご意見や情報をお待ちしております。事務局宛にご連絡ください。

最後に、この右のページにありますように、来年の 9 月に第 10 回機械翻訳サミットがタイ（プーケット島）にて開催されます。1987 年に第 1 回が箱根で開催されてから、20 年 10 回目の節目の回であり、1999 年以來のアジアでの開催ですので、多くの皆様に発表・ご参加いただきますよう、お願いいたします。

MT Summit X

**September 12-16, 2005
Phuket, Thailand**

The tenth Machine Translation Summit, organized by the Asia-Pacific Association for Machine Translation (AAMT), will be held at Phuket Island, Thailand in September 2005.

MT Summit X will bring together the scientists, developers and users in the area of machine translation. The frontier research and demonstration from leading academic institutes, business and government sectors will show up their latest results. In the main conference, MT Summit X will provide various features including research papers, reports on users' experiences, invited talks and panel discussions. Surrounding the main conference the expanded program will cover the sessions of tutorial, workshop and exhibition. This wide coverage features in the machine translation arena will facilitate participants from the sectors of academy, research, business and government to get together to explore the machine translation technology for conquering the problems in language barriers and digital divide.

AAMT cordially invites you all to MT Summit X, to contribute and witness the efforts in machine translation technology.



**2005年9月12日(月)～16日(金)
タイ王国、プーケット**

アジア太平洋機械翻訳協会(AAMT)主催による第10回機械翻訳サミット (MT Summit X) が2005年9月、タイ王国のプーケット島で開催されます。

MT Summit X は、機械翻訳分野の研究者・開発者・ユーザ等の幅広い関係者を対象としています。主要な学術、民間、政府機関等による最新研究発表やデモンストレーションが予定されています。サミットの本会議では、研究論文発表、ユーザレポート、招待講演、パネル討論等が用意されており、本会議前後の拡張プログラムとしては、チュートリアル、ワークショップ、デモセッション等が予定されています。このように機械翻訳に関わる幅広い層の参加者が一堂に会することにより、言語障壁やデジタルディバイドを克服するための議論が活発に行われるものと期待しております。

当協会は、MT Summit X への皆様のご参加をお待ちしております。



Welcome to Phuket, the paradise on earth.

地上の楽園、プーケットでお会いしましょう！



AAMTジャーナル

No. 36



Asia-Pacific Association for Machine Translation (AAMT)

Mitsui Sumitomo Kaijo Bldg., Annex 3F

3-11, Kanda-Surugadai, Chiyoda-ku,

Tokyo 101-0062 JAPAN

TEL:+81-3-3518-6418 FAX:+81-3-3518-6472

URL:<http://aamt.info> Email:aamt@aamt.info



発行：アジア太平洋機械翻訳協会 (AAMT)

住所：〒101-0062 東京都千代田区神田駿河台3-11

三井住友海上別館ビル3F

TEL：03-3518-6418 FAX：03-3518-6472

ホームページ：<http://aamt.info>

Eメール：aamt@aamt.info

編集委員会：井佐原 均 大倉 清司 熊野 明

松川 淑子 村上 嘉陽

事務局：中瀬 真 高田 佳代子

印刷・製本：株式会社ナビックス