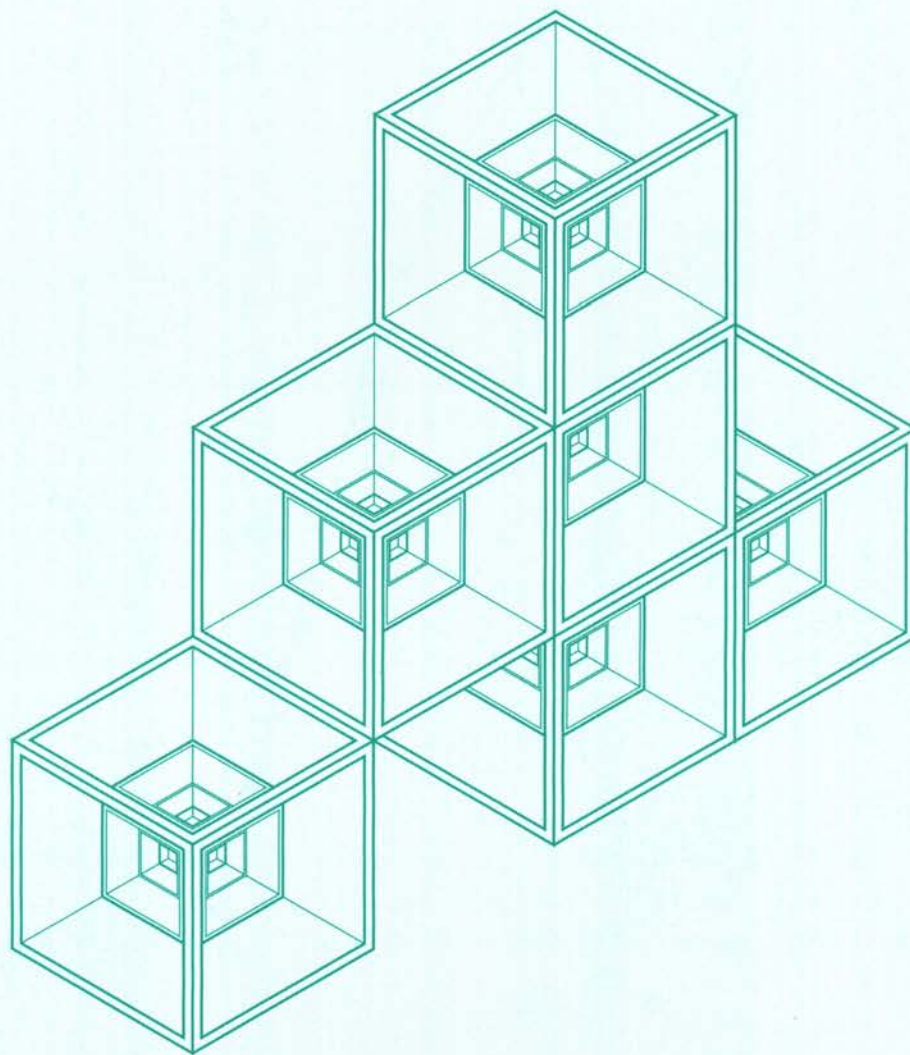


AAMT

Asia-Pacific Association for Machine Translation

Journal



June 2008

No.42

アジア太平洋機械翻訳協会

目次

巻頭言：	MT 研究雑感	梶 博行.....	1
新サービス紹介：	携帯電話を使った世界初の音声翻訳サービス「しゃべって翻訳」	隅田 英一郎.....	2
翻訳支援ツール：	翻訳工学の追及—MT+TM 翻訳ソリューション SATILA と業務翻訳が目指す道—	山本 ゆうじ.....	5
プロジェクト報告：	日タイ MT システムを利用した交流授業リポート	井佐原 均.....	9
会員紹介：	AAMT 会員のひろば（第 2 回）		10
委員会報告：	AAMT 機械翻訳テストセット構築に向けて	AAMT 機械翻訳課題調査委員会	18
	共有ユーザ辞書仕様 UTX の現状と今後の課題(日)	AAMT 機械翻訳課題調査委員会	25
	共有ユーザ辞書仕様 UTX の現状と今後の課題(英)	AAMT 機械翻訳課題調査委員会	32
活動報告：	協会活動報告（2007 年 12 月～2008 年 5 月）		44
編集後記：		宇津呂 武仁.....	46

CONTENTS

Foreword:	My Latest Thoughts on MT	<i>H. Kaji</i>	1
New service Report:	The world's first speech translation service on cell phone, SyabetteHonyaku	<i>S. Sumita</i>	2
MT systems:	Development of Translation Engineering - MT + TM workflow SATILA, and the future of MT	<i>Y. Yamamoto</i>	5
Report:	Report on The Thailand-Japan Intercultural Communication Project	<i>H. Isahara</i>	9
AAMT Members:	AAMT Agora		10
Committee Report:	Toward Building AAMT Test-Sets for Quality Evaluation of Machine Translation Sentence	<i>MT Committee for seeking future direction of MT</i>	18
	Introduction to UTX, a specification for a shared user dictionary (Japanese version)	<i>MT Committee for seeking future direction of MT</i>	25
	Introduction to UTX, a specification for a shared user dictionary (English version)	<i>MT Committee for seeking future direction of MT</i>	32
AAMT Activities	AAMT Activities (December 2007～May 2008)		44
Editor's Note:	Message from the Chair of the Editorial Committee	<i>T. Utsuro</i>	46

MT 研究雑感

静岡大学情報学部

教授 梶 博行

機械翻訳の研究に永年携わってきたという理由で巻頭言の執筆を依頼された。この分野の発展にどれだけ貢献できたかは甚だ心もとないが、確かに四半世紀以上もの間、機械翻訳に関わってきた。1980 年代にはメーカーで日英機械翻訳システムの研究開発を担当した。事業部と協力して世に出した製品は会社の業績に貢献することなく終わったが、1990 年代以降、次世代の機械翻訳につながるコーパスからの知識獲得に関する研究に取り組んできた。3 年前に大学に移ったが研究テーマは継続している。

近年、統計的機械翻訳を中心に、世界的にみると機械翻訳の研究はたいへん活発になっている。たいへん喜ばしいことであるが、日本の地位・競争力が相対的に低下していることも事実である。機械翻訳関連の国際会議における日本からの論文は以前ほど多くない。このような状況は、日本が機械翻訳の分野をリードした 1980 年代を経験した者としてはいささか残念である。何とか日本からの情報発信を増加させたいものである。

2006 年度から日中・中日機械翻訳の文科省プロジェクトが情報通信研究機構を中心として推進され、筆者も参画している。このプロジェクトでは、用例ベース機械翻訳を基本的な枠組みとしながら、ルールベース翻訳や統計的機械翻訳のよい点も取り入れる必要性が議論されている。日英や日中のような構造的に大きく異なる言語間の機械翻訳ではヨーロッパ系言語間の機械翻訳より高度な技術が要求される。このプロジェクトを通じて新しいアイデア・技術を生み出し、再び日本の存在感を高めることが期待される。

この日中・中日機械翻訳プロジェクトにおいて、筆者は主として訳語選択の課題を担当しているが、ソース言語とターゲット言語それぞれの単言語コーパスを組み合わせる方法を利用する方法を提案している。大規模なパラレルコーパスが利用できる言語対や分野は限られるので、技術的には困難であるが単言語コーパスを利用することが必要と考えるからである。統計的機械翻訳における翻訳モデルの学習も単言語コーパスの利用が必要になるのではないだろうか。このような研究を通じて情報発信に努めていきたい。

若い研究者・技術者の育成も日本の機械翻訳にとって大きな課題である。自然言語処理の研究者は増えているようだが、機械翻訳に限るとそうでもない。特にメーカーの若手研究者が少ないようである。1980 年代の投資が回収できなかったこともあって慎重な態度をとらざるを得ないのであろうが、メーカーも次世代機械翻訳に本格的に取り組む時期にきていると思われる。機械翻訳に興味をもつ学生は筆者の大学でも思っていた以上に多い。メーカーでも再び力を入れる日が来ることを期待し、機械翻訳システムの開発に携わることができる人材を育てていきたいと考えている。

携帯電話を使った世界初の音声翻訳サービス「しゃべって翻訳」

国際電気通信基礎技術研究所(ATR)

音声言語コミュニケーション研究所 隅田 英一郎

AAMT 会員にとっては自明なことではあるが、異なる言語を話す人とのコミュニケーションや異なる言語を話す集団への情報発信を行うことは、経済活動等のグローバル化に伴い極めて重要になってきている。特に、音声翻訳、人間が話した言葉をそのまま相手の言語に自動通訳する技術は、人類にとって夢の技術であった。この技術は、音声を認識する技術、話し言葉を翻訳する技術、相手の言語で音声を合成する技術で構成されており、長年の研究の結果、その基本的な部分が、最近、日本語、英語、中国語の旅行会話を対象に実用可能なレベルまで到達した。ATRはこの研究開発に中心となって貢献し、その研究成果に対して、2007年6月にAAMT長尾賞を頂いた。

一方で、昨今、携帯電話はもはや国内のみならず海外でも使えて当たり前の時代になりつつある。また、非常に多くの人が携帯電話を常に持ち歩く多機能の電子的ツールとして利用している。

この携帯電話で、音声翻訳が使えれば、その利便性は極めて高い。

ATRは、関連会社ATR-Trekを通じて、自ら開発した音声翻訳の技術を、2007年11月に日英双方向の携帯電話用アプリとして、DoCoMo905iシリーズの携帯電話で商用化した。

ここでは、同サービスの利用の実際と研究開発の経緯を紹介する。

1 「しゃべって翻訳」サービスの利用

本サービスでは、携帯端末に向かって発話すると、端末で音声の特徴量を抽出し、サーバで文として認識する。これを別のサーバが相手の言語に翻訳し、結果を端末に返すサーバ・クライアント型の音声翻訳システムである。

携帯電話アプリの標準的な利用料金である150円/月(税別)で使える。この他にポケット通信料がかかるが、音声の生データでなく特徴量を圧縮して送信するので通信料は少なく済む。ただし、海外で利用する場合は、日本と異なるポケット通信料が適用され定額割引もないので、使いすぎには注意が必要となる。海外で利用した場合、1回あたりにかかるポケット料金は、だいたい50円から100円程度、デコメールを1通送るのと同程度にわかる。

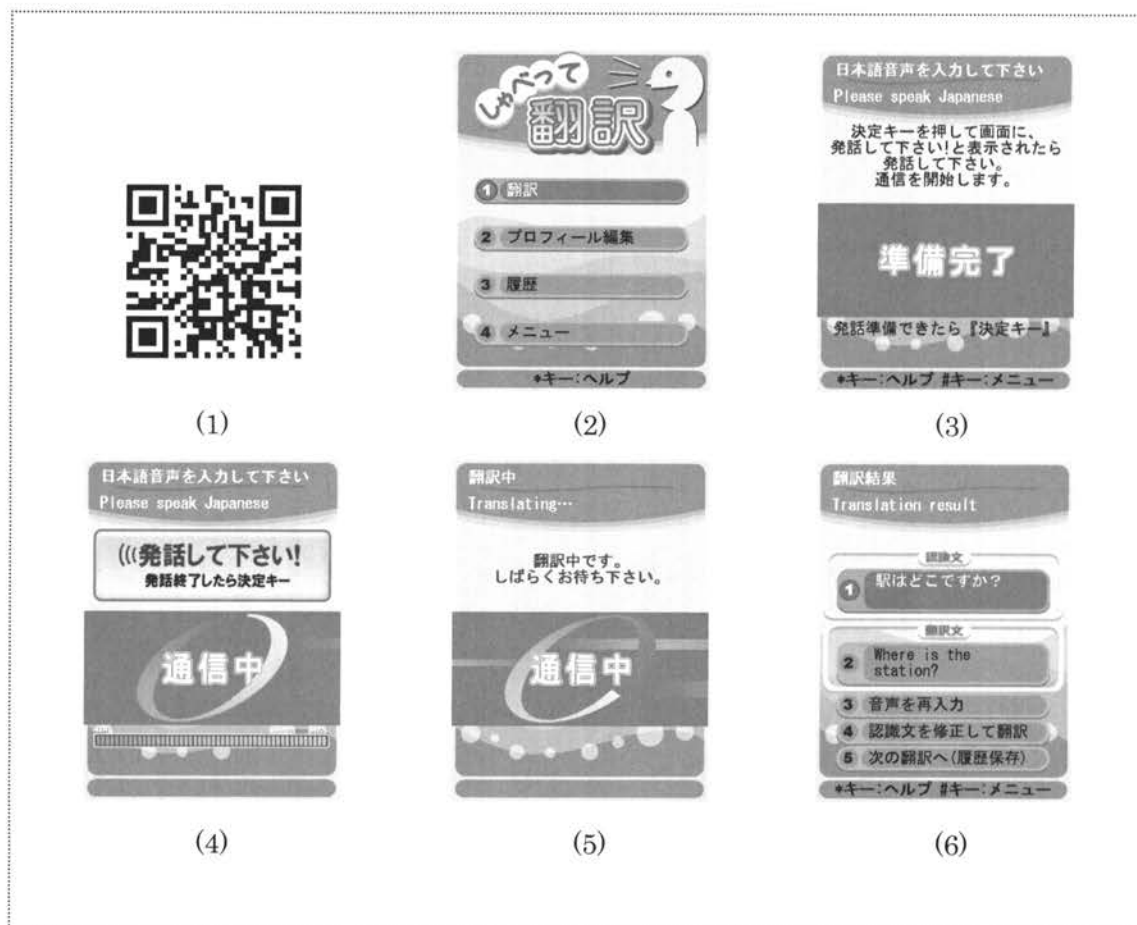
図1で利用イメージを説明する。

- (1) まず、しゃべって翻訳の携帯サイト(QRコードのURL)にアクセスして、アプリをダウンロードする(あるいは【iMenu】→「メニュー/検索」→「辞書/便利ツール」→「便利ツール」→「しゃべって翻訳」と順にメニューをたどる)
- (2) アプリを起動すると、TOPの画面が表示される。
- (3) しゃべる前に決定キーを押す。
- (4) 「発話してください」と出たら、マイクに向かって、音声を入力し、もう一度決定キーを押す。
- (5) 音声認識と翻訳が走る。
- (6) 認識結果と翻訳が表示される。

なお、【しゃべって翻訳】の詳しい取扱説明書はATR-Trekのサイ(<http://www.atr-trek.co.jp/>)、でご覧いただける。

2 音声翻訳技術の研究開発経緯

音声翻訳技術の実現が人類にもたらす価値は科学的、文化的、経済的に非常に大きい。An MIT Enterprise Technology Review誌の2004年2月号の特集「10 Emerging Technologies That Will Change Your World」が、Universal Translationを世界を変える10の技術のひとつとして取り上げており、翻訳技術の中でも特に、音声翻訳技術に



焦点をあてて紹介している。

音声翻訳の歴史は約 20 年と機械翻訳の歴史の約 60 年に比べ、まだ浅い。音声翻訳がはじめて提唱されたのはテレコム'83 であり、NEC がラボラトリーモデルとして音声翻訳のデモを行ない注目を集めた。音声翻訳実現のためには、長期的な基礎研究を行なう必要があるという認識のもとに、基盤研究円滑化法のもと、1986 年に ATR 自動翻訳電話研究所が設立され、国内外から様々な研究機関の音声言語研究者が参画した。1993 年には、ATR、CMU、シーメンスによる世界 3 地点を結んだ音声翻訳実験も行われた。ATR のプロジェクト開始の後、ドイツで Verbmobil プロジェクト、欧州で Nespole!, TC-Star プロジェクトが進められた。現在、米国で TransTac, GALE プロジェクトが進められている。以下、ATR で研究開発されてきた、旅行対話を対象とした音声翻訳の概要について紹介する。

3 ATR における音声翻訳研究

1986 年に音声翻訳プロジェクトをスタートして以来、音声翻訳の研究を時限プロジェクトとして進めてきた。特に、設立当初の 1986 年当時は現在と比べるとハードウェアの性能が極めて乏しく、ATR での音声翻訳の立ち上げは、かなり挑戦的なものであった。第 1 期から第 3 期までの特徴を表 1 に示す。音声翻訳は、一般に大きく分けて音声認識、機械翻訳、音声合成の 3 つのコンポーネントとこれらの統合部分から構成される。それぞれの技術の困難さから、ATR では、あらゆる会話を対象とするのではなく、特定の分野に対象を絞り込むことにより、認識・翻訳・合成の精度を利用可能なレベルまで向上させることを狙って研究開発を進めてきた。

表 1 ATR における研究開発の概略

	第 1 期 (1986.4～1993.3)	第 2 期 (1993.4～2000.3)	第 3 期 (2000.4～2006.3)
研究目標	音声翻訳の要素技術と実現可能性の確認	日常の話し言葉への展開と分野の拡張	実際の環境で利用可能な技術の構築
対象分野	会議予約 (日英独)	ホテル予約 (日英)	日常旅行会話 (日英中)
言語的特徴	文法的に正しい表現	日常的な表現(口語的表現, 非文等を含む)	広範囲な話題での日常的な表現
音響的特徴	明瞭な発声	不明瞭な発声を含む	雑音を含む発声

認識、翻訳、合成ともにコーパスを用いた手法を採用しており、特に翻訳に関しては、旅行会話分野の日英対訳 100 万文を構築し、ここから翻訳に必要なモデルを構築した。旅行会話に関しては、従来のルール・ベースの翻訳技術に比べて圧倒的な翻訳品質を誇っている。人間と比較すると TOEIC で 600 点以上の人間の翻訳能力と等価ということが明らかになっている。

4 携帯電話を使った音声翻訳の実装

図2に商用化したシステムの音声認識部の構造を示す。本システムは、分散型音声認識を基礎とした構造を持つ。携帯電話側(フロントエンド)において、パーティクルフィルタを用いた雑音抑圧及び音響分析、ETSI ES 202 050 に準拠した符号化が行われ、bit-stream データのみが音声認識サーバに送信される。音声認識サーバ側(バックエンド)では、受信した bit-stream を展開し、音声認識及び、単語信頼度の計算処理が行われる。このようなシステム構造を採用することの利点は、携帯電話の情報処理能力の限界に縛られず、大規模かつ精密な音響モデルや言語モデルが利用可能な点が挙げられる。更に、各々のモデルは携帯電話ではなくサーバ側に存在するため、それらの更新作業が容易であり、常に最新の状態が維持可能である。

5 今後の展開

これまでの音声翻訳の研究の過程と昨年末実現した

携帯電話を用いた分散型の音声翻訳システムの実用化について述べた。いまだ数多くの課題が残されていることも事実である。実際の旅行用音声翻訳サービスでは、実在するあらゆる地名、観光関係固有名詞などに対応する必要があるが、それらを一つ一つ登録するには限界があり、また、場面、状況に応じて訳し分けが必要なものもある。これらの問題を解決することで、さらに使いやすい音声翻訳を実現することができる。

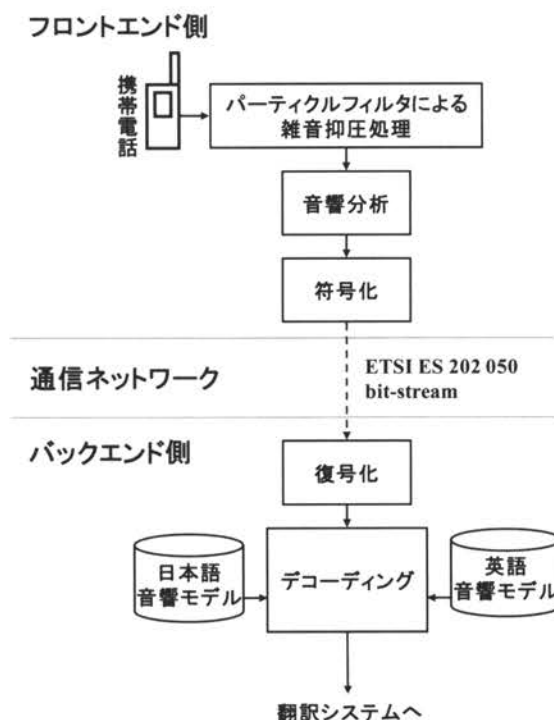


図 2 携帯電話を使った分散型の音声翻訳

翻訳工学の追究 —MT+TM 翻訳ソリューション SATILA と業務用翻訳ソフトがめざす道—

秋桜舎 代表 山本ゆうじ（言語・翻訳コンサルタント）

はじめに

以前に、本誌 37 号(2005 年 6 月)で SATILA を紹介させていただきました(本記事とあわせてご覧ください)。その際に旧・共有ユーザー辞書仕様 UPF について提案したことは、その後に筆者が参加したワーキンググループ 3 の各位のご協力のもと、新しい仕様 UTX に取り込まれ、着々と実現しつつあるのはうれしい限りです。

あれから日本の翻訳業界の流れも少しずつ変わってきました。2007 年には日本翻訳連盟(JTF)主催の「翻訳祭」パネル ディスカッションのテーマとして「MT+TM(翻訳メモリ)」が選ばれ、筆者はその司会を務めさせていただきました。また、海外でも「MT+TM」の試みが、さまざまな場所で行われており、ヨーロッパ語間では特に進展を遂げています。本稿では、前回の記事を踏まえううえで、最新の情報を含めてご紹介します。

改めて自己紹介しますと、私は言語・翻訳コンサルタントをしています。秋桜舎とその実務翻訳ブランドである transPC において、IT 翻訳とローカライゼーションを専門とする翻訳者として働く一方で、IT を活用したさまざまな翻訳、文書作成、文書処理の合理的にノウハウを「翻訳工学」として蓄積し、企業向けに提供するコンサルティングを行っています。現在、IT を活用した翻訳と文章作成に関する書籍を執筆中です。原始的な方法論から抜け出せない翻訳業界の現状に抗して「翻訳工学」と名づけたのですが、研究すべき課題は山積しています。各位のご指導をいただければ幸いです。

SATILA とは

筆者が開発・提供している翻訳ワークフロー SATILA とは、翻訳者支援を目的とした翻訳メモリ ツールと翻訳ソフトの統合ソリューションです。



翻訳メモリ ツールについては、拙作の動画デモがあるので、あわせて参照いただければ幸いです。

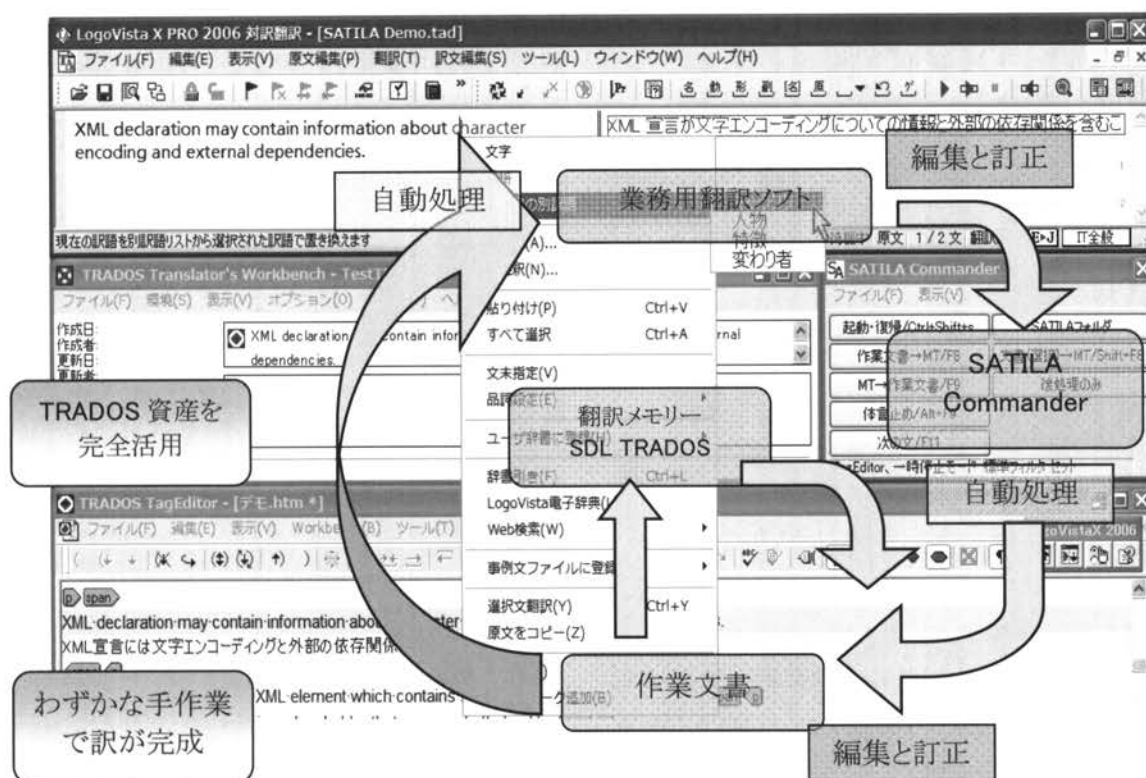
http://www.translationzone.com/jp/landing/micro-site/webinar_recordings.asp

図 1 は、SATILA での実際の作業画面を示しています。画面上に翻訳ソフト(LogoVista X PRO 2006)、中央に翻訳メモリ ツール SDL TRADOS 2007 の Workbench(翻訳メモリ検索ツール)、下に TagEditor(作業文書用のエディタ)、右に統合ツール SATILA Commander が表示されています。SATILA では、これらの各アプリケーションの特性を把握し、翻訳者の知識と技能に基づいて、その潜在能力を引き出して連携させ、最大限の効率と品質を実現します。

SATILA は、特定のソフトウェアや環境に依存せず、柔軟な組み合わせが可能です。SATILA は、翻訳者の視点から見てもっとも効率がよい方法により、翻訳メモリと翻訳ソフトをツールとノウハウで統合したソリューションです。具体的には、統合ツール SATILA Commander により、翻訳メモリと翻訳ソフトをツール群で結合することにより、能動的な用語統一と自動適用を行います。また、クライアントのスタイル ガイド、仕様を辞書に組み込み、用語や表記の厳密な適用が可能となります。SATILA Commander では、翻訳者のノウハウに基づいた各種の整形フィルタを組み合わせ、後編集を自動化します。

SATILA では、翻訳ソフトを通じて、カスタマイズした用語集を自動適用することにより、訳抜けがなく、確実に用

図 1 SATILA 画面



語が使われるなどのさまざまな利点があります。その結果、専門用語やスタイルのチェックの手間を大幅に減らすことができます。

SATILA では、条件がそろえば、翻訳作業効率を最大で 2 倍程度向上できます。習熟度が完全でない場合や、辞書作成の工程などをすべて含めた場合でも、総合的な作業能率を最低でも 2 割程度向上できます。

SATILA ソリューションのパッケージ

当方が提供する SATILA ソリューションのパッケージには、以下のものが含まれます。

- ライセンス
 - 基礎データ(辞書、学習データ、フィルタ)
 - SATILA Commander ライセンス
- カスタマイズ データの構築作業
- 講習とサポート

オンラインと紙ベースの教材 SATILA は、短期集中講習を含むパッケージとして提供しています。SATILA では、各ツールを「どのように使うか」という技能が要求されます。

特に業務用翻訳ソフトでの翻訳作業では、手作業や翻訳メモリのみを使用した場合とは、方法論が異なり、発想の切り替えが必要です。また、ユーザー辞書のカスタマイズが必須であり、カスタマイズ・調整済みの専門辞書の提供と、個々の顧客の環境に応じてあらたに辞書をカスタマイズする作業もパッケージに含まれています。

単体翻訳ソフトとの違い

SATILA が翻訳メモリ ツールと連携することにより、単体の翻訳ソフトを越える利点は、以下のとおりです。

- 既存の翻訳メモリ資産を 100%活かせる
- 業務翻訳で必須の、品質管理を翻訳メモリ側で行う
- 書式情報を翻訳メモリ側で管理する

翻訳メモリとは、単なる原文と訳文のペアではありません。いつ、だれに訳されたか、という管理情報も重要です。また、翻訳メモリ自体をすばやく参照し、検索できる必要があります。さらに太字などの書式がタグによりカプセル化され、文字列情報と隔離されている必要があります。

そうでないと書式がわずかに異なるだけで、翻訳メモリーが有効に再利用できなくなります。現時点の業務用翻訳ソフトに付属する「翻訳メモリー機能」でこの要件を満たす製品は、ひとつもありません。また、既存の業務用翻訳ソフトにこれらの要件を要求するのも、開発コスト的に困難と思えます。翻訳メモリーは、翻訳業界で実際に広く使われている実績があり、一朝一夕に追いつけるものではないからです。そうすると、業務用翻訳ソフトにとって現実的なのは、翻訳作業で重要なポイントを押さえたうえで、SATILA のように既存の翻訳ソフトと翻訳メモリーを統合することです。

翻訳者の経験とノウハウを取り込む

筆者が今、特に強調したいのは、「**翻訳支援システムの開発には、翻訳者のノウハウを要する**」ということです。翻訳ソフトが翻訳支援ツールとしての市民権を得るには、「**翻訳者の意見に耳を傾ける**」とことが重要です。

世間では、「翻訳は専門技能である」ということがよく理解されていないことがあります。「英語と日本語のそれぞれができれば翻訳なんてだれでもできる」と誤解しているむきもあります。

残念ながら、現在の業務用翻訳ソフトには、現場の翻訳者の意見が有効に活かされていません。消費者翻訳ソフトの単なる延長線上でしか捉えられていないのです。これは、翻訳ソフトが翻訳者にほとんど使われていない理由のひとつです。

なぜ翻訳者の言い分を聞くべきなのでしょう。翻訳者のノウハウと最善慣行(ベスト プラクティス)の積み重ねにより、「どう訳すべきか」という点が絞られてきます。ところが現時点ではこの視点がぼやけています。翻訳者の意見は、翻訳者にとってのみ役立つではありません。翻訳者の意見は、必然的に**一般消費者が使う翻訳ソフトの改善にもつながります**。

もちろん、翻訳者であればだれでもよい、というわけではありません。翻訳会社による MT の活用は繰り返し試みがされましたが、成功事例が限られています。これは MT の潜在的な実力が足りないのではなく、不適切な方法を使ったからです。実際に翻訳メモリー ツールと翻訳ソフトの両方使いこなしている翻訳者が、ばらばらにではなく、それらを統合して使った経験があつてはじめて、どの点を改

善すべきか適切に判断できます。

翻訳知識の機械学習には、大量かつ良質の訳文を必要とします。**低品質の翻訳はいくらあってもノイズとなるだけです**。ウェブページそのものをコーパスとして使おうとしても、雑多なウェブページからは文章としての品質を下げる要因が多々見つかります。誤字、脱字、誤用はもちろん、不適切な訳語、表記などです。ブログという手軽にウェブページとして公開できる表現手段が急速に普及した結果、ウェブ上の言語的なノイズもまた増加しています。気軽な情報発信ができること自体は望ましいですが、読み直しも推敲もしない乱雑な文章があふれている点には注意すべきです。さらに自動生成されたコンテンツ、特に自動翻訳されたコンテンツも言語的なノイズの一因となり、ハウリングのような効果をもたらします。良貨が悪貨を駆逐できる、**すなわちよい翻訳が悪い翻訳を駆逐できる仕組みを構築しなければ、この悪循環は止まりません**。

現時点での翻訳ソフトの専門辞書でも、「訳語の妥当性」については品質管理が不十分でよく吟味されているとはいえません。これでは、翻訳ソフトを適切に使っても、言語的な品質を低下させる可能性があります。

UTX について

「訳語の妥当性」に関連して、共有ユーザー辞書仕様 UTX についての私的見解を述べておきます。言語処理学会での論文も適宜ご参照ください。以下に述べることはあくまで私見であり、ワーキンググループ 3 で出された結論と符合することはあっても、必ずしもワーキンググループ 3 で出された結論そのものではないことにご注意ください。

まず UTX が主に扱うのは広義の専門語彙であり、基本的語彙辞書の構築は想定していません。接続詞など、文法の解釈に関わる基本的語彙は、MT のシステム辞書として提供されるものとしています。特定文脈で特殊な意味を持つ場合にのみ、基本語彙を登録します。その意味で、システム辞書とは異なる性質を持つことにご留意ください。タブ区切り形式である UTX-Simple に含まれるのは、最低限、原語、訳語、品詞の情報のみでかまいません。UTX-Simple では各社の独自性を打ち出すことよりも、利用者の便宜として翻訳資産を有効利用することに重点があります。

UTX では、有償辞書(公式辞書)と無償辞書(オープン辞書)の 2 種を想定しています。オープン辞書は、コミュニティで公開される場合、不特定多数の集合知の貢献により作成されます。この場合、特定の一個人が知的所有権を主張することは困難です。1 語でも貢献すれば、貢献者ということになるでしょう。数名ならともかく、数百人以上になるかもしれない貢献者全員を記載することは、あまり実用的ではありません。また「もっとも適切な訳語」が、使用回数などによって自動的に導き出された場合は、著作権の主張が困難です。さらに、UTX-Simple はテキスト形式であり、著作権保護のための有効な手段がありません。実際の作業では、特定の辞書を分割して、複数の辞書にしたほうがよいこともあります。そのためオープン辞書については、原則としてすべての貢献者に著作権を放棄してもらうのが現実的と考えられます。もしこれ以外によいアイデアがあればぜひお知らせください。

一方で、辞書管理委員会が認める、権威ある監修者によって監修された辞書は、しっかり保護すべきです。この場合は、暗号化による著作権保護機能を組み込んだ UTX-Secure(仮称)を使用します。不特定多数の貢献者による辞書の改変は不可能になります。

企業内での使用など、「閉じた環境」では、著作権保護機能のない UTX-Simple でも差し支えありません。しかし UTX の意義は、「共有すること」、相互に翻訳資産を出し合い、協力することにあると考えます。UTX-Secure のような形態、有償の公式辞書も重要な一方で、オープン辞書の促進・普及にも各位のご理解とご協力をお願いします。

いまユーザーが何を必要としているか

「翻訳ソフトのユーザーが何を必要としているか」については、ワーキング グループ 2 メンバー各位の活動によりアンケート調査が進められています。これまでに一定の成果はありましたが、まだユーザーと開発側の距離が埋まったわけではありません。さらなる調査と分析が期待されます。

各社の翻訳ソフトの操作性についても、進展が見られた点もありますが、まだおおいに改善の余地があります。特に辞書登録など、単純な目的を達するまでの手順が複雑すぎる人が多いようです。

「自分で作ったものを自分で使って改善してゆく」という

姿勢は、あらゆる物作りの根本であると考えます。自分の作ったソフトを使わないのは、味見もせずに料理を作るようなものかもしれません。たとえば、ある音声入力ソフトの開発元では、関係者が自分で使っている様子がほとんど見られませんでした。汎用のソフトであるため、その気になれば使えるにもかかわらずです。これでは、ユーザーの感じている不満が、作り手の側にも実感されないのも当然でしょう。

SATILA Commander は、大企業と比べるとささやかではありますが、開発者自身が実際に使っているツールでもあります。「自分で作ったものを自分で使う」というのは、企業が大きくなるほど困難かもしれませんが、可能な形で実践してもらいたいものです。開発者とユーザーのギャップを少しでも埋められるものと思います。

まとめ

SATILA は、大規模になればなるほどその効果を発揮します。クライアントや翻訳会社で一括導入することで、辞書や学習データの共有が可能になり、さらに大きな効果を上げることができます。合理的な翻訳支援ワークフローを研究されているメーカーの方、関心のあるメーカーの方ぜひご連絡ください。翻訳資産の「何をどこまでカスタマイズすべきか」という重要なポイントでお手伝いできます。

- 執筆者の連絡先

<http://cosmoshouse.com/mail.htm>

- SATILA の詳細

<http://transpc.cosmoshouse.com/satila/>

SATILA に関連するその他の情報

- 関連執筆記事

<http://transpc.cosmoshouse.com/news.htm>

本件に関連する、雑誌、メール マガジン、ウェブなどの
当方執筆の記事・コラム・講演など

日タイ機械翻訳システムで広がる異文化交流

独立行政法人 情報通信研究機構

井佐原 均

大規模コーパスの普及と計算機パワーの増大により、近年機械翻訳の研究開発に再興隆が見られます。その中でも、今後特に発展が期待されるエリアとして注目されているのがアジア言語を対象とするシステムの研究開発です。

2008年2月、独立行政法人情報通信研究機構（以下「NICT」）タイ自然言語ラボラトリと自然言語グループは、開発中の日本語-タイ語機械翻訳システムを利用して、日本とタイの高校間でオンライン交流授業を実施しました。京都府立東宇治高等学校（京都府宇治市）と Nawamintrachinee Mae Hong Son（タイ北部メイホンソン地域）の協力のもと、この機械翻訳システムを組み込んだインターネットのテレビ会議システム（図1）を通して、タイ側、日本側から文化紹介（プレゼンテーション）が行われました。



図1. 機械翻訳機能付きテレビ会議システムの画面イメージ

このテレビ会議システムでは、映像、音声、文字を送受信することができます。また、機械翻訳システムは京都大学黒橋禎夫教授の研究グループで開発された用例翻訳型の機械翻訳エンジンをNICTにおいてタイ語へ拡張したものです。音声認識機能は無いため、プレゼンテーションでの生徒たちによる

スピーチが人手でシステムに入力され、翻訳エンジンに掛けられました。

授業では、まず Nawamintrachinee Mae Hong Son によるプレゼンテーションが行われ、学校が位置するメイホンソン地域の民族衣装「タイヤイ」が紹介されました。次に、東宇治高校が日本の四季や京都の歴史、日本の学生たちの日々の生活について紹介し、着物姿のクラス代表が日本舞踊を披露すると、タイの生徒たちから鳴り止まない拍手が送られました。



図2. 交流授業の様子
（上段：タイ、下段：日本）

現在のシステムはプロトタイプであり、対訳コーパスに収められているタイ語は約 4,000 語程度です。今回の交流授業に際しては、プレゼンテーションで用いられるキーワードをあらかじめシステム辞書に組み込んでおくなどの措置が取られましたが、未知語が多く用いられる質疑応答等のリアルタイムの入力をうまく翻訳するレベルには至っていません。今後、対訳コーパスを 10 万語まで拡充し、様々な言い回しにも対応できるシステムの開発を目指します。

AAMT会員のひろば

現在、学者・研究者・翻訳者・流通業者など多岐にわたる AAMT 会員の交流は、年一回の総会や不定期のセミナー・研究会に限られている現状があります。もっと気軽にお互いを知ることができ、ニーズが合えば親密な交流を図ることが出来る場を AAMT Journal 誌面上で提供するべく、No.41 より「AAMT 会員のひろば」をスタートいたしました。おかげさまで、早くも会員の皆様よりご好評をいただいております。

会員の皆さまよりご寄稿いただいた、自己紹介・機械翻訳および翻訳業界に期待すること・AAMT へのご要望をご紹介します、会員間での新たな交流が生まれるきっかけとなることを期待しております。第二弾の今号では、法人会員 3 社、個人会員 3 名の皆さまよりご寄稿いただきました。次号以降も本企画を継続していくことを考えておりますので、会員の皆さまからのご寄稿を心よりお待ちしております。

本企画へのご寄稿・お問い合わせは AAMT 事務局(E-mail: aamt-info@aaamt.info)まで宜しくお願いいたします。

法人会員（敬称略・50 音順）

会員名

株式会社 インターグループ／INTER GROUP CORPORATION

自社紹介、機械翻訳および翻訳業界に期待すること、AAMT への要望

「AAMT に寄せる」

現在、国立国会図書館館長をされておられる長尾眞先生（アジア太平洋機械翻訳協会初代会長）が京都大学におられる時に、「Mu プロジェクト」に参加させていただいたご縁で、小生は協会の設立に関わらせていただき、AAMT 副会長の任に就かせてもらっております。

20 年近く前に、我が社が研究開発部門を抱えた大企業に混じって、協会の活動に参加させて頂いたことは、大変光栄と感謝しております。

弊社は、日本がまだまだ高度経済成長期にあった 1966 年に、やがて訪れる国際化の波に対応できる人材養成を目的に、財界の皆様の後援でその要望に応える形で設立された会社でございます。

まずは、コミュニケーションを通して国内外で活躍できるプロの通訳者養成をスタート。活躍できる人材が育ってくると各界より、言語の障壁を気にせずに国際間の交渉、交流をするために通訳、翻訳の業務が一気に増えてまいりました。

弊社の経営理念の一つに、「外交、ビジネス、科学技術、スポーツほか、各分野における異文化コミュニケーションにおいて、常に新しい手法を研究開発し、お客さまに提供する」ということがあります。

機械翻訳という言葉がまだよく知られていない頃に「Mu プロジェクト」に参加させていただいたのも、その理念に沿った動きです。当時からコンピュータの進化は目覚しく、想像以上に発達することは容易に感じ取れました。

しかしながら、コンピュータで翻訳ができるようになるとはまだ半信半疑であったのは確かです。それでも、

お手伝いできることは何でしよう、我社の赤坂のマンションを機械翻訳協会の事務局として提供し、社員の中から元気な2名を選び事務局長と事務局員として3年間に渡り活動のサポートをさせていただきました。

そのときの事務局長が元日航スチュワーデスだったことも懐かしい思い出です。

大型計算機で翻訳した出力分を修正し、読みやすい翻訳文に仕上げるポストエディットという手法で、大量の翻訳原稿を効率よく処理する方法を試してみたり、日本で初めて「ポストエディターコース」を立ち上げ新しい人材の育成にも取り組みました。

現在、マニュアル翻訳では主流となっているメモリ翻訳がまだ普及する前からいろいろなことを試みてきました。

新しいことはまず取り組んでみる我社にとって貴重な経験をさせていただいた長尾眞初代会長はじめ、田中穂積元会長、辻井潤一前会長にも大変お世話になり感謝しております。

井佐原会長が就任挨拶に「究極の機械翻訳システムは何でしょうか？そのひとつの方向は機械翻訳の存在を意識しない環境の実現でしょう」と書かれておられましたが、私が通訳者に「通訳者を介して話している人達が通訳者の存在を忘れるような通訳者を目指しなさい」と指導していることとまったく同じであることに、その目標の高さに驚くと同時に感心したのを覚えています。

話は変わりますが、最近弊社内でも、若い社員とのコミュニケーションがスムーズに行かないことがあります。日本語どうしてあってもです。

私は社員に、「いつも相手の立場に立って考えろ。どうすれば相手に判ってもらえるか工夫をしろ。相手が何を言いたいのか全身で理解しようと考えなさい」といつも言っています。

通訳には同時通訳という通訳手法がありますが、日本語と外国語の両方のことばの理解が完璧であっても同時通訳はできません。話者と同じスピードで通訳を行うために必要なのは、話者の意図を先に知り同時というより、むしろ先行通訳とも言うべき神経の集中力が求められる、まさに「一を聞いて十を知る」能力なのです。

日常生活においても、レベルは違ってもこの難しいコミュニケーションをとることによって、社会生活や仕事が成り立っています。

長年、コミュニケーションのサポートを業務としている立場にあって、その重要性をあらためてかみしめ、社員を指導しながら業務に誠実に取り組んでおります。

AAMTにおいても同様、井佐原会長をはじめ理事役員、会員の皆様と試行錯誤しながら機械翻訳システムの取組が活発になるよう盛り上げて行きたいと思っております。

先日、電車の中で、聞こえてきた高校生の会話の意味が分からなくて、社に戻ってから若い社員に意味をたずねたことがあります。言葉は昔からどんどん変わっていくもので、当然のことであり、それも文化の一つだと理解しています。しかし以前、犬の鳴き声を人間の言葉に翻訳するソフトが話題になりましたが、若い新入社員との会話用翻訳ソフトがヒットする事の無い様願っております。

株式会社インターグループ

代表取締役 小谷泰造

会員名

東芝ソリューション株式会社／TOSHIBA SOLUTIONS CORPORATION

自社紹介

東芝ソリューション株式会社は、2003年10月に、東芝グループでITソリューション事業を一手に担うことを目的として発足しました。あらゆる業界のIT戦略を担うことを使命とし、東芝グループの持つ豊富なノウハウをふんだんに活かした最先端のソリューションを提供しています。

会社発足前の東芝時代から自然言語処理技術の研究・開発に注力し、JAMTの発足時より法人会員として加入しています。30年に及ぶ機械翻訳の研究成果やノウハウは、現在では「The 翻訳」としてお客様にご提供しています。また、日英/英日翻訳で培った技術を応用し2007年に中日/日中翻訳の提供を開始しました。

The 翻訳シリーズは、お客様のニーズに合わせたラインナップをそろえています。パソコン用としては、ビジネス向けの「The 翻訳ビジネス」や翻訳専門家向けの「The 翻訳プレミアム」、更に、特許翻訳業務向けや学生向けの商品も提供しています。また、サーバ型翻訳システムとして「The 翻訳サーバ Enterprise Edition」も提供しており、翻訳機能を企業の各種システムへ組み込んで新たなソリューションを生み出すことができます。これらの商品は、お客様の翻訳業務の生産性を高めることをめざし、現在も改良を続けています。

これらの商品紹介のほか、お客様に役立つ情報を充実させた「The 翻訳」のホームページもぜひご覧ください。
The 翻訳ホームページ：http://pf.toshiba-sol.co.jp/prod/hon_yaku/

機械翻訳および翻訳業界に期待すること

「The 翻訳」シリーズは、翻訳精度向上のほか、翻訳業務効率の向上を重要課題として開発を行っています。そのために、ユーザの方々の生の声をうかがい、機能開発に生かすことを心がけてきました。今後も、翻訳業界を中心とした機械翻訳を利用する立場の方々との交流・連携の機会を増やすことにより、真に「使える」商品を今後も提供し続けていきます。利用者と開発者の仲立ち的な役割をAAMTに担っていただけることを期待しています。

AAMT への要望

AAMTの委員会活動は、開発者と利用者の双方の参加によるAAMTならではの成果を得られる、意義ある活動の場です。当社は、法人会員として委員会活動にも積極的に携わっています。当社の関わりが活動の活性化、スピードアップの一助となり、よりタイムリーな成果につなげられることを願っています。

会員名

株式会社ナビックス／NAVIX COMPANY, LIMITED

自社紹介

株式会社ナビックスは平成元（1989）年の創業以来、揺籃期にあった DTP システムを駆使した取扱説明書の多言語制作を主業務として参りました。

いわゆる多言語のローカライズベンダーとして世界のあらゆる言語への翻訳対応を国内で行うとともに、現在では英国、中国（上海）、タイ、ポーランドに海外拠点を持っております。

70 名ほどが勤務する東京のオフィスも国際色豊かで、外国籍の従業員が全体の 15%以上を占める状況です。

出身地もさまざま、中国、韓国、台湾、シンガポール、アメリカ、ブラジルなどとなっていて、アラビア語等の特殊言語を専門にする日本人社員もおります。

東京のオフィスでは、ここ数年は海外進出と同様にさまざまな事業にチャレンジをし続け、今ではドキュメンテーションサービスのみならず、留学支援事業に外国人の人材派遣や職業紹介なども手がけるに至っております。従来のマニュアル制作を専業としていた状況からグローバルなコミュニケーションを多面的にサポートする企業へと成長をしつつあります。

これからもオンリーワンのユニークな企業として、日本を含むさまざまな国や地域から集まった社員の情熱を、国際コミュニケーションを支援するサービスに変えて、世界中に弊社の想いが届くことを念願しております。

URL : www.navix.jp

機械翻訳および翻訳業界に期待すること、AAMT への要望

昨今の取扱説明書の制作における特筆事項の一つに、多言語化の進展があります。これは EU の拡大を例に挙げるまでもなく、従来対応して来なかった言語にも、西欧語などと同様に翻訳やドキュメントの制作ニーズが発生していることを意味します。しかしただそれだけではなく、製品を出荷する側も取説の現地語対応に踏み切ることで、売上を上げる効果が期待できるため、よりいっそうの多言語対応が一般的になってきているように思われます。

例えば英語とその他数言語程度の取説を作れば済んでいた話が、いきなり十数言語を必要とするようになるなど、輸出製品の取説の制作状況を見ると多言語化の流れはどんどん加速していますが、翻訳・制作業界は一般的に労働集約的で作業の標準化が困難な業界と考えられていることもあって、他業種に比べて機械化や自動化が遅れているとも指摘されているのが実態です。

それでは本当に機械化や自動化が不可能なのかと考えると、実は決してそうではないことも分かります。なぜなら取説の制作業務を例にとると、その本質は全面新規で翻訳や制作をするのではなく、飽く迄ベースモデルからの改訂作業であるためです。こうした事情もあって、逆に現場からは工程の機械化や自動化が強く望まれています。

実際には作業における細かなヒューマンエラーも後を絶ちませんし、誤りの有無をチェックしたり、修正する作業も最終的には人手で行います。このため大幅なコストの削減も期待薄ではありますが、だからこそブレイクスルーとなるべき技術を期待したいと考えております。

弊社では、その一つが機械翻訳であるとの認識を致しております。ただし機械翻訳一本槍でのソリューションというよりも、翻訳メモリや自動組版といった周辺技術を結集することで、新たなイノベーションに至るのが理想であり、また現実的であるとの考えです。

機械翻訳は、まだまだ今後改良される必要のある技術とされますので、今後もユーザ企業として AAMT の活動には積極的に協力をさせて頂く所存です。

個人会員（敬称略）

会員名・所属

飯田 仁（東京工科大学メディア学部）／

Hitoshi IIDA (Tokyo University of Technology, School of Media Science)

自己紹介

現在、大学にて教育と研究に携わっています。

自己紹介を書くこの機会に自分の研究を振り返ってみました。

節目・出来事となる活動などを振り返って見ると、人工知能国際会議 IJCAI'79（東京・経団連会館）における日本語による質問応答システムの研究発表、COLING/ACL のジョイントとなった COLING'84（スタンフォード大）における英日機械翻訳方式とその実験システムの研究発表、そして 86 年 4 月に ATR（国際電気通信基礎技術研究所）での自動翻訳電話研究プロジェクトに設立当初から参加した(98 年 9 月まで)ことなどが思い出されます。

自身が自然言語処理に関わることになった 70 年代後半期、日本語処理を行うためには漢字フォントを自前で作成しなければならませんでした。計算機のメモリサイズも限られていたため、地名などを一語、辞書登録するだけでも使用メモリサイズを積算して知識ベース化しました。質問応答システムのエンジン部分では、DB 上の事実データを検索するだけでなく、データ登録がない事柄に対する知的な応答ができる仕組みを導入していました。例えば、「イランの米の生産量はどれだけか」などの質問に対して、事実データの情報が分からないことが分かると、コメの生産条件と検索対象地点の気候データとを見比べ、その生産条件を充たしていないことから生産データの登録がないであろうことを説明する応答を実現していました。非単調推論による検索に対する応答手法を事例的に可能にしていたといえます。

ATR においては、新たに音声翻訳を始めましたが、まずは通訳を介した音声対話の実態を探るべく、色々な形態のデータ収録に努めました。既に 20 年以上前のことになりますから、キーボード間の対話の収録も行いました。音声言語におけるポーズに対応する情報として、100 ミリ秒単位のキーストロークも同時に記録するように工夫しました。音声翻訳研究であることから、自ずと音声対話の処理・理解に関する追究もしました。

ATR における研究活動の中でとくに一つを挙げれば、用例翻訳の可能性を示し、具体的な翻訳手法として提示できたことでしょう。「用例に基づいた翻訳」というタイトルで、90 年春の情報処理学会大会で発表しました。その直後 6 月にテキサスで開催された TMI-90 において、EBMT という用語を初めて使って用例翻訳手法の可能性と具体的な手法について発表しました。

そして、91 年 4 月に設立された日本機械翻訳協会 JAMT に発足時から個人会員として参加し、引き続き現在ま

で AAMT 個人会員になっています。

その後は、色々な研究をする機会に恵まれたものの、未だ道半ばであります。幾つかを挙げれば、言語運用に基づくコミュニケーション・モデルの研究、海洋哺乳類のコミュニケーションの仕組み追究、マルチモーダル・コミュニケーションのモデル化とエージェントの実現、漫才に見られる「オープンコミュニケーション」のモデル化、汎用的教育用メディア・リソースの構築などであります。

機械翻訳および翻訳業界に期待すること

現在の MT 研究を見ると、ある程度形が見えている翻訳エンジンの下、大規模対訳コーパスの拡充と翻訳精度向上が中心課題になっているといえるでしょう。その一方で、翻訳支援技術の追究とシステム開発も着実に進められています。その中で、とくにテキスト翻訳における適切な表現の生成支援、あるいは音声翻訳におけるインタラクションを介した合意形成のための適切な解釈に基づく発話生成などについては、今後一層の研究と開発が求められると感じています。文書作成技術に目を向ければ、仮名漢字変換の行為を介したテキスト作成が一応定着し、誰もが当たり前のように使用しています。翻って翻訳については、機械翻訳の技術を核として、自然な行為として訳文を作成できる翻訳技術の利用環境や応用形態について一層の追究が望まれると思います。

会員名

井佐原 均 / ISAHARA Hitoshi

自己紹介

独立行政法人情報通信研究機構 (NICT) の上席研究員として、日中機械翻訳システムの開発、アジア言語の処理技術の研究開発、言語資源の開発に携わっています。

機械翻訳は学生時代から身近に感じていましたが、自分自身が直接に関わった最初のものは JEIDA (現 JEITA) の委員会活動として行った機械翻訳評価基準の作成でした。

AAMT との積極的な関わりは、シンガポールでの機械翻訳サミットのサポートに加わったのが、きっかけだったと思います。その後、プーケットでの機械翻訳サミットを経て、AAMT の会長を拝命し、現在に至ります。

機械翻訳および翻訳業界に期待すること

大量の言語資源と高速・大規模な計算機環境によって、高性能な翻訳システムが現実のものとなりつつあります。また、分野や文体への対応も速やかに行えるようになるでしょう。

このような状況の今こそ、機械翻訳の新しい対象、新しい使い方が見えてくるのだと思います。

そのような新しい動きは、研究者、開発者、ユーザがバラバラに居て、思いつくものではないでしょう。研究者、開発者、ユーザが一体となった機械翻訳協会の活動の中でこそ、見えてくるものがあるかもしれません。

AAMT の活動から、何らかの新しい動きの芽が出てくれば良いなと思っております。

AAMT の役員一同、今後とも努力してまいりますので、皆様のご協力をお願いいたします。

会員名

中川裕志（東京大学 情報基盤センター 教授）／NAKAGAWA Hiroshi

自己紹介

これまでの自分の研究を振りかえってみますと、研究者を始めたころは、ちょうど第 5 世代コンピュータプロジェクトが動き始めたところであり、私も人工知能の研究を始めました。特に論理型言語 Prolog は、従来のプログラム言語と様相を異にしており、興味をそそられ、Prolog 言語のプログラムにおける効率化を目指したプログラム変換の研究から始めました。当時、AIUEO という若手研究者、大学院生の定期的勉強会があり、そこで議論をしているうちに人間の知能に近い内容が顕在化した言語に興味を持ち自然言語処理の研究を始めました。1990 年にスタンフォード大学の CSLI に 1 年間滞在し、言語学について勉強していたのですが、どうもこれがよくなかったようです。というのは、90 年というのは、自然言語処理が言語学の知見を基礎にしたシステムから、統計的機械学習を中心にする時代への転換点だったからです。そんなわけで、私自身にとって 90 年代は失われた 10 年になりました。

1999 年に東京大学に異動して後、辻井先生をはじめとして統計的言語処理に強い研究者、大学院生と接する機会が増え、失われた 10 年を回復すべく努力を続けております。

現在の研究内容については私のホームページ <http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/>を訪ねていただければ幸いです。

機械翻訳および翻訳業界に期待すること

スタンフォード大学 CSLI に滞在していたとき、スタンリー・ピータース教授が、「言語学の原則は 60% のケースで正しければ上出来」とおっしゃっていたことをしばしば思い出します。この言明に続く問いは「では、残りの 40% は何の要因で決まるのか?」ということになりますが、意味に加え偏好や文脈知識によって決まってくると思われます。さて、言語学に依拠するシステムからコーパスと統計的機械学習によるシステムになってきても、この問題が解決されたわけではありません。言語外の知識への依存という責任の押し付けから、狙った言語現象が現れにくいというデータスパースネスの問題が変わったのですが、解決しているわけではありません。たとえば、構文解析アルゴリズムは語彙主導なものになってきているものの、HPSG においてすら外界の知識はうまく取り込まれているとは言い難いところです。そのような状態で、テストコレクションにおける正解率が 1% 上昇することに奔走する研究には、どうしても違和感を覚えます。また、機械翻訳の評価では BLUE 値がよく使われますが、これも人間の直感にあっていないという批判が付きまといまいます。ようするに、自然言語処理を自然科学のような整然とした学問にしようという目論みが現実と遊離しているのではないかという危惧を抱きます。

さて、自然言語処理に外部に身をおいてみますと、ビジネスなど実用的な用途で欲しがられている技術はもっとと即物的です。たとえば、企業が欲しがらる評判情報抽出技術では、大がかりな構文解析より、個々の語彙の実際の場での意味、つまり language in use が重要な役割をはたします。つまり、現実世界に立脚した意味がほしいのです。100 年ほど前にソシュールが言語学を外部から独立した体系性の研究とみなして以来、自然言語処理はまさにこの枠組みの中で成果をあげてきました。たとえば、自動要約は文書をより短い文書に変換する技術でほとんど外界とのインタラクションなく実現できます。統計的機械翻訳は、2 言語の並行コーパスさえあれば、後は一

切外界とのかかわりなく数学的操作だけで翻訳できます。そういった具合ですから、今までの自然言語処理はスケジュールの掌の中で旋回しているのです。しかし、一方で、外部の人々から自然言語処理や機械翻訳への期待はむしろ、現実世界とのインタラクションの部分にあるようです。

こういった事情で、自然言語処理も機械翻訳も根底的なパラダイムの境目にあるのではないのでしょうか？では、どうしたらよいか、という問いかけにドンピシャの答えを持ち合わせているわけではありません。ですが、すでに述べたように、自然言語処理への外界からのデマンドは常に個別の語彙を通じてやってきます。ですから、語彙あるいは辞書といった言語資源の強化が、理論的には地味であっても必要なテーマなのではないかと考えています。こういった研究で実世界への効果的な貢献をすることが、最近、世の中に蔓延するテキスト技術の軽視、無視の風潮、あるいは自然言語処理はもう終わった学問分野だという言説に一矢報いる道ではないのでしょうか。

AAMT 機械翻訳文テストセット構築に向けて

機械翻訳課題調査委員会ワーキンググループ 1

0. はじめに

現在、日本語・中国語を対象とした機械翻訳システムがインターネット上を始め、様々な形式で利用されるようになった。今後も日本語・中国語翻訳システムの利用者は増加するため、翻訳システムの一層の改良が必要と思われる。このような機械翻訳システムの中国語への関心を背景として、アジア太平洋機械翻訳協会の機械翻訳課題調査委員会（以下、AAMT 調査委員会）は、現状の日本語・中国語の翻訳システムによる翻訳文の達成度を評価し、また、現状の翻訳システムにおける問題点を調査することを目的として活動を推進してきた。本稿は、AAMT 調査委員会の機械翻訳評価に関する活動の基盤となる機械翻訳評価法に関する既存の研究を概観し、AAMT 調査委員会の活動を紹介する。

1. テストセットによる翻訳文の評価法：JEIDA テストセット

機械翻訳システムの技術的な問題点を検出することを目的とした評価法として、社団法人日本電子工業振興協会の機械翻訳技術分科会が提案したテストセット（以下、JEIDA テストセット）による評価法がある。JEIDA テストセットは、日本語・英語を対象とする評価法である。AAMT 調査委員会は、この JEIDA テストセットを日本語・中国語を対象とした評価法へ発展させた。本小節では、AAMT 調査委員会の評価法の基となった JEIDA テストセットを概観する。

JEIDA テストセットとは、JEIDA が提案した三種類の機械翻訳評価基準の一つである。JEIDA は機械翻訳評価における基準として、次の三種類を提案した（日本電子工業振興協会 1993, 野村・井佐原 1992, Nomura & Isahara 1992a/1992b, 高山他 1993, 中岩他 1993）。

- (1) 機械翻訳システムユーザー対象の経済的評価のための基準
- (2) 機械翻訳システムユーザー対象の技術的評価のための基準
- (3) 機械翻訳システム開発者対象の技術的評価のための基準

それぞれの基準の目的は次のとおりである。基準(1)の目的は、機械翻訳システム利用者が翻訳システムを導入する際に、考慮すべき経済的なメリット・デメリットを示すことである。この基準は、ユーザが想定する翻訳業務と翻訳システムの特徴（例えば、簡易版翻訳システムや本格的な翻訳システム）との関連から最適なシステムを選択する際の指針となる。基準(2)は、機械翻訳システム利用者の満足度を計ることを目的とする。特に、基準(1)により本格的な翻訳システムの導入が適切と判断された利用者を対象とする。基準(3)は機械翻訳システム開発者に翻訳文から推測される翻訳システムの技術的な問題点を示すことを目的とする。

AAMT 調査委員会における機械翻訳評価の目的の一つとして、現状の機械翻訳システムの問題点を明らかにすることが挙げられるため、基準(3)を日本語・中国語を対象とした評価法へと発展させることを検討した。以下、JEIDA テストセットの形式を紹介する。

JEIDA テストセットは機械翻訳システムを様々な技術的項目の達成度の観点から評価を行うだけでなく、各項目の達成度から評価対象となる翻訳システムの問題点を明らかにする。テスト文には翻訳システムが技術的に解決しておくべきと想定される言語形式が含まれている。この課題の達成度の判定は各テスト文に付与された設問を利用する。評価者は、テスト文の翻訳結果に対して、設問に回答するだけでよい。

このようにテスト文を設定する評価法は JEIDA テストセットに限ったことではない。例えば、池原・小倉（1990）、池原他（1994）、成田（1988）などが JEIDA テストセット開発時において提案されていた。但し、これらの評価法はテスト文のみであり、設問が付与されていない。そのため、翻訳出力を評価する際に、評価者の言語的直感に頼らざるを得ないという問題が残る。一方、JEIDA テストセットではテスト文に加え、設問を付与することにより評価者の主観評価が排除できると考えられた。

テストセットによる評価の信頼性をより高めるために、設問への回答は Yes/No の二値とされた。さらに、機械翻訳システムの特徴や問題点を客観的に評価するために、各項目に技術レベルも付与された。このテスト文と設問からなるテストセットの形式をとることにより、客観的に翻訳システムの問題点を検出できると考えられた。さらに、設問を付与することで、評価者間

にみられる言語現象に対する知識や気づきの差といった問題も回避できる。

この設問は、テスト文が評価対象とする翻訳処理と言語現象との関係を示すため、評価者が客観的に翻訳結果を評価できるだけでなく、開発者はその評価結果から容易に翻訳文の問題点を検出することができる。

この JEIDA テストセットは、機械翻訳システムが処理すべき言語現象を含む英語例文約 1400 文とその模範訳、および評価観点となる設問からなる。評価項目は表 1 の通りである。言語現象として、(1) 品詞、(2) 文の部分構造、(3) 文構造、(4) 構文、(5) コロケーションを扱う。各現象の会項目として、数種類の項目が設定されている。

各項目に対して、テスト文や設問が付与されている。実際の項目の書式は、表 2 が示すように、【番号】、【例文】、【訳文】、【質問】、【訳出例】、【関連文】、【参照項目】、【解説】から構成される。

部分名称	項目名称	項目数
品詞	冠詞	15
	名詞（固有名詞含む）	27
	代名詞	25
	形容詞	42
	副詞	54
	前置詞	38
	動詞類：動詞・準動詞	48
	動詞類：助動詞	37
	関係詞	19
	接続詞	26
	記号	16
文の部分構造	不定詞	26
	分詞・分詞構文	19
	動名詞	23
	時制	63
	数量表現	28
	慣用表現	8

文構造	文種（疑問文、命令文、感嘆文）	19
	否定	16
	特殊構文	19
	比較	21
	仮定法（条件法）	16
	態	10
	話法	4
	挿入	16
	省略	9
	倒置	7
	並列	25
構文	多品詞（品詞認定）	34
	係り先認定	27
コロケーション	コロケーション	27

表 1：設問項目

【番号】	例文の番号
【例文】	例文（1文のみ）
【訳文】	模範訳（例文の日本語訳）
【質問】	“A”が「B」のようにCとして訳されていますか？という形式の質問文。
	A：英語表現。“ ”で囲む。例文中のどの部分を翻訳することにポイントがあるのかを表す。文全体の場合、また明らかな場合などは省略する。 B：日本語表現。「 」で囲む。 C：内容や文法事項の補足説明（「習慣を表す表現」、「洗濯疑問文」など）を記す。記述が長くなるものや、原因に言及する場合は、【解説】に記述する。 *必ず○×（YES/NO）で答えられる形式にする。本テストセットは評価者が翻訳結果（訳文）を見るだけで○×を与えることを前提としており、解析の詳細に直接言及することは避ける。
【訳出例】	許容される訳出例や誤訳例を列挙。 正解例（YES）の文頭には○、誤例（NO）の文頭には×を付与する。 1行1文とし、原則として文全体を記述。

必要ならば正／誤の理由（説明）も示す。

*各例は実際の機械翻訳システムの訳を参考にして作成した。

【関連文】 当該の例文と関連する言語現象を含んだ例文を挙げる。

文の一部だけの記述は認めない。必ず文全体を記述する。

例文の後に、「」で囲んだ訳文を記述する。

捕捉事項（may/might での丁寧度の違いなど）がある場合は、訳文の後ろに（ ）で囲んで記述する。

【参照項目】 本テストセット内の関連項目への参照ポインタ。原則として、相互参照とする。

*文番号を明示するのみで、文そのものは記述しない。

【解説】 その他の捕捉事項。フリーフォーマット。

表 2：各項目の書式

2. JEIDA テストセットの自動評価化

評価者による主観評価に代わる評価法として、これまで数多くの自動評価法が提案されてきた (Niessen et al. 2000, Akiba et al. 2001, Papineni et al. 2002, NIST 2002, Leusch et al. 2003, Turian et al. 2003, Babych & Hartley 2004, Lin & Och 2004, Banerjee & Lavie 2005, Giménez et al. 2005, Corston-Oliver et al. 2001, Kulesza & Shieber 2004, Gammon et al. 2005, Albredcht & Hwa 2007)。自動評価法は評価者による主観評価と異なり、評価結果の信頼性低下や評価コストといった問題が少ない。AAMT 調査委員会でも、日本語・中国語 JEIDA テストセットの自動化を将来的な課題としている。本小節では、日本語・英語 JEIDA テストセットの自動化を試みた AAMT 技術動向調査委員会と情報通信研究機構との共同研究（内元他 2006/2007, Uchimoto et al. 2007）（以下、AAMT-NICT）を概観する。

一般的な機械翻訳自動評価法は、翻訳出力とその正解となる参照訳や言語モデルとの類似性を自動的に判定することで評価を行う。このような類似性に基づく

評価法であれば、正解との類似性が低い文ほど何らかの問題があると考えられる。これらの自動評価法におけるテスト文として JEIDA テストセットのように翻訳処理と翻訳結果の関係が明らかな文例を用いることで、翻訳の良し悪しを相対的に評価するだけでなく、翻訳結果にどのような問題点が含まれているかを検出できると考え、AAMT-NICT は JEIDA テストセットに基づいた自動評価法を構築した。

JEIDA テストセットの自動化に際し、AAMT-NICT は、JEIDA テストセットの妥当性を検証した。JEIDA テストセットの評価観点は、計算言語学における翻訳処理に関する研究に加え、理論言語学や比較言語学における日本語と英語における様々な言語的特徴に関する研究に基づき考案されたものである。したがって、設問の理論的根拠は確かなものと考えられる。しかし、各項目に対して付与される例文の妥当性に疑問が残るため、AAMT-NICT は JEIDA テストセットの例文の妥当性を検証した。その結果、一つの例文中に複数の項目が含まれているなど例文の改

善が必要であることがわかった。さらに、JEIDA テストセットの例文に対して、複数の項目が含まれている例文を調査した。その結果、150 例文中 103 文において複数の項目が含まれていることがわかった。そこで、設問と例文の整合性に留意し、AAMT-NICT は新に 148 設問を追加し、JEIDA テストセットの改良を行った。この改良を通じて、1 文に対して一問であった設問が最大五問まで増えた。

追加された設問の例を表 3 に示す。追加 1 に示される設問は ID1.15.1.3-5 として JEIDA テストセットに設定されていた設問である。この設問の項目が表 3 に提示されている例文に当てはまるとして追加されている。また、追加 2 に示される設問は JEIDA テストセットにおいて設定されていなかった設問である。この設問が新に追加設定されている。

修正前	ID	1.17.1.3-1
	例文	The percentage of stomach cancer among the workers appears to be the highest for any asbestos workers.
	訳文	労働者の胃癌の割合は、アスベスト労働者のために最高となるようだ。
	設問 0	“appear to”が“ようだ”のように訳されていますか。
追加 1	ID	1.15.1.3-5
	訳出例	“for”が適切に訳されていない。
	設問 1	“for”が原因・結果を示す「... で」と訳されていますか。
追加 2	ID	追加-1
	訳出例	訳されていない表現がある。
	設問 2	全ての英語表現が日本語に訳されていますか。

表 3 : JEIDA テストセット修正例

この修正により、JEIDA テストセットの評価基準としての妥当性が向上することを AAMT-NICT は確認した。JEIDA テストにおいて、各例文に対して設問が一問の修正前テストセットと修正後のテストセットを用いて機械翻訳文を評価し、その評価結果を評価者による評価結果と比較した。その結果、修正後テストセット ($r=0.65$) が修正前テストセット (0.50) よりも評価

者の結果と相関が高いことがわかった。

この修正後 JEIDA テストセットを基に、AAMT-NICT は評価の自動化を試みた。自動評価の対象となるのは、テスト文の翻訳結果とその正解との類似度である。この類似度は最小二乗法 (Draper & Smith 1981) により式 (1) から求めた。

$$A = \sum_{i=1}^m \lambda_{S_i} \times S_i \quad (1)$$

$$+ \sum_{j=1}^n (\lambda_{Q_j} \times Q_j + \lambda_{Q'_j} \times Q'_j) + \lambda_e$$

$$Q_j = \begin{cases} 1 & : \text{if subgoal is accomplished} \\ 0 & : \text{otherwise} \end{cases} \quad (2)$$

$$Q'_j = \begin{cases} 1 & : \text{if subgoal is unaccomplished} \\ 0 & : \text{otherwise} \end{cases} \quad (3)$$

Q_j は i 番目の設問の達成率である。 λQ_j は設問達成率に対する加重和であり、 $\lambda Q'_j$ は設問非達成率の加重和である。 n は設問数を示す。 S_i は参照訳との類似性を示し、 λS_i は類似性の加重和を示す。

テストセットにおける文例の翻訳正解例と翻訳システムの出力結果の類似度を自動判定することにより、JEIDA テストセットを用いた評価法の自動化がなされた。JEIDA テストセットに基づく自動評価法と既存の自動評価法に得られた評価結果を主観評価との相関関係により比較を行った。その結果、JEIDA テストセッ

トに基づく自動評価法が高い相関を示すことが確認された。AAMT-NICT は既存の自動評価手法において参照訳数が一種類、三種類、五種類の場合と比較した。その結果、表 4-6 からわかるように、既存自動評価法が評価結果の妥当性が参照訳の数に依存しているのに対し、JEIDA テストセットに基づく自動評価は参照訳数への依存はみられない。このことから、JEIDA テストセットに基づく自動評価は参照訳にかかるコストを軽減できるといえる。

手法	流暢さ		適確さ	
	訓練データ	評価データ	訓練データ	評価データ
既存手法 1 (WER)	0.43	0.48	0.42	0.48
既存手法 2 (複合)	0.52	0.51	0.49	0.47
提案手法 (自動評価)	0.90*	0.59*	0.89*	0.62*
提案手法 (手動評価)	0.90*	0.62*	0.90*	0.68*

表 4：自動評価と人間による評価との相関係数：参照訳 1 種類

手法	流暢さ		適確さ	
	訓練データ	評価データ	訓練データ	評価データ
既存手法 1 (WER)	0.47	0.51	0.45	0.51
既存手法 2 (複合)	0.54	0.54	0.51	0.52
提案手法 (自動評価)	0.90*	0.60*	0.90*	0.64*
提案手法 (手動評価)	0.90*	0.62*	0.90*	0.69*

表 5：自動評価と人間による評価との相関係数：参照訳 3 種類

手法	流暢さ		適確さ	
	訓練データ	評価データ	訓練データ	評価データ
既存手法 1 (WER)	0.49	0.53	0.46	0.53
既存手法 2 (複合)	0.56	0.56	0.52	0.54
提案手法 (自動評価)	0.90*	0.60*	0.90*	0.63*
提案手法 (手動評価)	0.91*	0.63*	0.90*	0.69*

表 6：自動評価と人間による評価との相関係数：参照訳 5 種類

3. おわりに

AAMT 調査委員会は、日本語・英語が対象であった JEDIA テストセットを日本語・中国語対象の評価セットへと改良を終えた。この改良版テストセットを用いて、インターネット上で無償利用できる日本語・中国語翻訳システムを対象とした評価実験を行なった。

現在、その評価結果の分析を行っている。また、AAMT 機械翻訳文テストセットの公開にむけ準備を行っている。

共有ユーザー辞書仕様 UTX に関する論文とスライドについて

機械翻訳課題調査委員会ワーキンググループ 3

言語処理学会論文

26 から 31 ページは、機械翻訳課題調査委員会ワーキンググループ 3 のメンバーが執筆し、富士通研究所の大倉委員が言語処理学会で発表した、UTX を紹介する論文を掲載しています。この論文は、現時点までで、ワーキンググループ 3 が公式に合意し決定した事項から構成されています。32 から 38 ページは、本論文の英訳です。ご協力いただいた NICT・ボンド委員のお名前が追加されています。

言語グリッド講演のスライド

39 から 43 ページは、NICT の言語グリッドプロジェクト研究会で、秋桜舎の山本委員が講演をした際に使用したスライドを参考までに掲載しています。この講演は、UTX と秋桜舎の活動を紹介するものでした。

これらのスライドでは、ワーキンググループ 3 として正式に決定した事項以外にも、山本委員の個人的な見解が含まれている点にご留意ください。これらのスライドでは、言語処理学会論文に至るまでの経緯と背景、基本的な考え方について紹介しています。

UTX に期待すること

ワーキンググループ 3 の委員から、UTX に期待することについて、抱負を頂きました(掲載順は順不同)。

——「パベルフィッシュ」は永遠な夢かも知れませんが、UTX の普及によって「パベルフィッシュ」に近い翻訳装置の開発が一步進むはずだと思います。少しでも多くの人たちに外国語で書かれている文書や話されている言葉の理解を支援するツールとして活躍することを期待しています。」(加藤委員 ※パベルフィッシュとは『銀河ヒッチハイク・ガイド』に登場する、言語を一瞬で翻訳する架空の動物(編注))

——翻訳ソフト制作に携わる者として、翻訳ソフトの質の向上と普及を常に願っていますが、UTX によって多くの人が翻訳ソフトの利便性を共有できるようになることを期待しています。(クロスランゲージ・鈴木委員)

——OKI では、機械翻訳をより多くの分野で高品質に使えるように、インターネットユーザの皆様がご自身でそれぞれの分野辞書を作成できるコミュニティ型機械翻訳サイト「訳してねっと」を開発、運営してきました。UTX は、その考えをさらに押し進めて、すべての機械翻訳システムやさまざまなソフトで使える辞書を自由に作成／共有できるしくみを作るものです。言語の壁を超えるための大きな動きが起こることを期待しています。(沖電気・村田委員)

——機械翻訳にとって、辞書の質こそが要だと思えます。UTX の充実と普及により、翻訳支援ツールの開発が、また一步前進することを期待します。(インターグループ・原委員)

——UTX は、MT メーカーはもちろん、ユーザーにとっても翻訳精度の向上という面できわめて重要です。また集合知としての翻訳資産を有効に統合・再利用することにより、オープンソースのローカライゼーション支援など、大きな応用の可能性を秘めています。翻訳支援用途からはじまって、自動翻訳・概訳用途にも大きく貢献できます。UTX 自体が日本から発信するプロジェクトとして意義深いばかりでなく、日本からの情報発信を支援するプロジェクトとしても意義が大きいと考えます。引き続き関係各位のご協力・ご賛同をいただければ幸いです。(秋桜舎・山本委員)

——成功の鍵は 4c(contents, community, conciseness, compatibility)。今度こそ、三度目の正直で世界標準へ！(NICT・内元委員長)

共有ユーザー辞書仕様 UTX の現状と今後の展開

大倉清司[†]、山本ゆうじ^{††}、村田稔樹[‡]、内元清貴^{‡‡}、加藤マイケル孝仁⁺、島津美和子^{*}、鈴木次良^{**}、
Francis Bond^{‡‡}

(機械翻訳課題調査委員会 共有化・標準化ワーキンググループ)

[†]富士通研究所 ^{††}秋桜舎 [‡]OKI ^{‡‡}情報通信研究機構
⁺ラーニングコンサルタント ^{*}東芝ソリューション ^{**}クロスランゲージ

1. はじめに

機械翻訳システムを実用的に使用するには、ユーザー辞書が必須だが、辞書の仕様が異なると相互利用できない。そのため、AAMT（アジア太平洋機械翻訳協会）[1]はどの機械翻訳システムでも共通に利用できる、共有辞書の仕様を策定している。より具体的には、1995年にIPAの支援を受けて策定されたUPFをベースに、その後の技術や利用方法の変化を反映し、2006年から新しい仕様策定を開始した。2007年には「UTX (Universal Terminology eXchange)」という新名称に改め、現在はそのシンプルな形式であるUTX-Simpleの仕様策定を行っている。2008年度は、実際の翻訳分野を選定し、その分野の辞書データの作成・収集を行い、翻訳プロジェクトと連携しつつ、UTX仕様の修正・確定を行う予定である。また、辞書を継続的に作成・共有・蓄積するためのコミュニティの創設も構想している。本稿では、UTX-Simpleの仕様を紹介するとともに、我々が想定している辞書の共有・再利用の過程と方式について述べる。

2. UTXについて

翻訳ソフトなどの機械翻訳システムを実用的に使用するには、ユーザー辞書の利用が重要である。翻訳対象文書で使われている専門用語や人名・地名などは、基本辞書に登録されていないことが多いため、適切に翻訳でき

ない場合が多い。これらは専門用語辞書を使っても適切に翻訳できないことがある。そこでこのような用語をユーザー辞書に登録すれば、機械翻訳システムの精度向上を見込むことができる。例えば、以下の原文をある翻訳ソフトで翻訳したとする。

XML declaration may contain information about character encoding and external dependencies.

結果は次のようになる。

「XML 公表が文字符号化と外部の属国についてのインフォメーションを含んでいるかもしれません。」

この訳文では、公表、符号化、属国、インフォメーションなど、原文の文脈と分野に合致しない訳語が使われている。しかし、適切な用語を辞書登録すれば、以下のように、より適切な訳文が生成される [2]。

「XML 宣言が文字エンコーディングについての情報と外部の依存関係を含むことがあります。」

特に“XML declaration”のような複合語をユーザー辞書に登録することにより構文解析の失敗も防げるようになることが知られており [3]、統計翻訳の精度向上にもつながるため、

辞書登録は重要である。

しかし、個人が大量の辞書登録を行うのは時間がかかり、辞書登録による効果もわかりにくい。また現状では個人の辞書を集めてもフォーマットが不統一であり、共通化・共有化できない。

翻訳辞書の標準化された仕様としては、LISA[4]のTBX[5]が挙げられるが、仕様が複雑で広く普及していないのが現状である。これらの問題を解決するために、AAMTではUTXを策定中である。

UTXの特徴は以下の通りである。

1. 分かりやすく使いやすい「ユーザーのための」辞書：複雑な仕様はむだにユーザーの負担を増やし、結局は使われない。ユーザーの立場に立ち、実際に使われるように単純かつ実用的な仕様を

心がける。

2. 「専門用語」という観点：辞書の分野を明確化し、「一語一義」とする。無意味に訳語を増やさず、使い分けを厳密に規定する。該当分野で訳語が一義的に定まる語を登録対象とする。
3. 翻訳精度の向上：UTX辞書が普及すれば共有も簡単にできるため、翻訳精度向上に欠かせない辞書の構築がより効率的になるようになる。
4. 多言語辞書、単言語辞書としても有効：二言語間だけでなく、多言語にも対応する。また単言語辞書として、用語の統一など校正支援ツールなどの入力としても使えるように仕様を策定している。
5. 共有のための情報も保持：XML形式においては辞書作成者、エントリー作成日時やMTシステム固有の情報などの詳細情

- 1つの「ヘッダ」と1つ以上の「エントリー」からなる
- 文書の最初の行に「ヘッダ」を記述する
- 2行目以降、1行に1エントリー記述
- #ではじまる行はコメント行

【ヘッダ】半角文字でそのファイルに関する記述を入れる。区切り文字は半角スペース。

#UTX-S バージョン番号 原言語/目的言語 最終更新日付時間 4カラム目以降の情報

例)#UTX-S 0.9 en-US/ja-JP 2007-12-03T14:28:00Z+09:00

※UTX-SのSは簡易版の意味。言語名(ISO 639)と日付時間(ISO 8601)はISO準拠[2,3]。

単言語辞書の場合、目的言語を省略する。

4カラム目のフォーマット例：

source:plural/3sp/past/pastp/presp/comparative/superlative/target:..../optionXXXX/optionYYYY

(見出し語の情報は、source:の後に記述。plural=複数形、3sp=三人称単数、past=過去形、pastp=過去分詞、presp=現在分詞、comparative=比較級、superlative=最上級、訳語の情報はtarget:の後に記述)

【エントリー】1行のフォーマットは、タブ区切りでカラムに分ける。各カラムに記述する内容は以下の通り：

1カラム目	2カラム目	3カラム目	4カラム目以降
原言語の単語	目的言語の単語	品詞	言語により依存

図1 UTX-Simpleの仕様

報も維持できるため、情報を切り捨てずに済み、また正確に管理・共有できる。

6. ソフトウェアのローカライゼーションの促進：特にオープンソース系のローカライゼーションにおいては、翻訳を個人単位で行っているために訳語の統一が難しい。UTX の導入により、共通の辞書が普及して翻訳作業の効率化が期待できる。

海外の例では、欧州委員会（EC）が 2008 年 1 月に、約 100 万の文章を 22 言語に翻訳した翻訳データを無料で公開すると発表している。日本では、依然として厚い言語の壁に阻まれて、ソフトウェアやコンテンツの翻訳がなかなか進まず、多くの潜在的機会が埋もれ、日々莫大な損失が発生している。翻訳資産の共有化がすべての業界で推進されることが、今後の産業振興の鍵となろう。

UTX のメリットとして、ユーザーにとっては、分かりやすく作りやすいフォーマットであること、個々の特定分野で翻訳精度の向上が期待できること、インターネット、LAN 上のコミュニティを通じた辞書の共有と再利用ができることなどがある。翻訳ソフト開発メーカーにとっても、ユーザー辞書利用の促進により機械翻訳が活性化し新規需要の掘り起こしが期待できること、またメーカー固有の辞書項目も完全に保持することなどが挙げられる。

UTX の仕様策定にあたって、以下の 2 つの作業が重要である。

1. 仕様の策定（UTX-Simple および UTX-XML）
2. 実際の辞書データの作成と、辞書を継続的に作成・共有・蓄積するためのコミュニティの創設

本稿では、上に挙げた 2 つの作業内容について説明するとともに、今後の展望・課題についても述べる。

3. UTX-Simple の仕様

UTX では、以下の 2 種の仕様に定める。

- UTX-XML（XML 形式）
- UTX-Simple（タブ区切りテキスト形式）

UTX-XML（XML 形式）は、辞書に必要な情報を全て含む。UTX-Simple（タブ区切りテキスト形式）は、最低限、原語、訳語、品詞の 3 種の情報だけで使用可能な仕様を規定する。仕様策定に関しては、最終的には XML 形式を策定するが、普及という観点から、シンプルな形で共有できる UTX-Simple をまず策定している。

標準化し、さらに普及させるためには仕様をユーザーに分かりやすく使いやすい形式にする必要がある。また、機械翻訳システムでユーザー辞書をフルに活用するには、現状のシステムよりも簡単に登録できなくてはならない。UTX-Simple 形式の辞書は仕様が単純で作成の手間が減り、辞書の共有が促進される。なるべく記述する情報を少なく、しかし実用的になるように仕様を検討した。現在のバージョンは 0.9 である（2008 年 5 月の最新版は 0.91）。基本的な仕様は図 1 の通りである。

UTX-Simple の仕様策定にあたり、以下のことが論点として上った。

- (1) 人間にとっての読みやすさを優先するか、機械にとっての処理のしやすさを優先するか。

何を必須の項目にするかについて議論した。最小限原語、訳語、品詞情報があればよいことにした。例えば英語では名詞の複数形の情報が必要な場合がある。しかし複数形は全て

の言語の全ての品詞で必要なわけではない。確かに複数形の情報がないと機械翻訳がうまく訳せない状況も少なくないが、人間が記述すること、また普及の観点から、オプションとして記述することにした。

(2) XML 版 (UTX-XML) と 簡易版 (UTX-Simple) の位置づけをどうするか。

UTX-Simple は日付、作成者などの情報がなく、恒久的・継続的な辞書の構築には適さない。しかし、一方で使いやすさから、容易に作成、共有することができ、普及が期待できる。XML 版の策定にはまだ時間を要する。

(3) カラムの内容を固定にするか (X カラム目=Y、と決めてしまうか)。

カラムは、言語共通の情報と、言語固有の情報を入れる枠組みとした。共通の情報は 1~3 カラム目に、実際翻訳において必要とされている言語依存の固有情報は 4 カラム目以降に記述する。

(4) 言語固有のカラムを言語対ごとにあらかじめ決めておくか否か。

当面、UTX では日本語、英語、中国語を中心に扱うが、本来は ISO に従い存在する言語全てを扱えるようにしたい。多数の言語について必要なカラムを現時点ですべて策定するのは現実的ではない。そこで、ユーザーにカラム名を指定できる幅を持たせる。言語対によっては、特定のカラムに決まった情報を入れることを推奨する。実際の利用からフィードバックを得て、

各言語で最低限必要な情報を見極めていく。

(5) ガイドラインの作成について

個々のエントリーについてどう記述すればよいか、表記などのガイドラインを決める必要がある。UTX-Simple を読み込むツールはそのガイドラインに従い、警告を出すようにする。例えばツールの警告出力オプションを指定しておけば、必要な項目がなければ警告を出す。基本的にフォーマットエラーはなしで、フォーマットが違っていた場合、その行は無視することにする。

基本的なガイドラインとしては以下を取り決めた。

- デフォルト値と「該当なし」との区別
「該当なし」を示す値: - (ハイフン)
それ以外は、自動処理 (ツールの判断)
- 辞書に保持されている情報をツールがどこまで使うかはツール自体の仕様に依存する。
- ユーザーが定義したカラムは保持するのみ。
- 英語の名詞句は語頭を小文字で書く。
複数形の場合は、見出し語を- (ハイフン) にする。

UTX-Simple の具体例を図 2 に示す。

4. 辞書を継続的に作成・共有・蓄積するためのコミュニティの創設

辞書データの作成と収集の基本方針として、

#UTX-S 0.9 en-US/ja-JP 2007-12-03T14:28:00Z+09:00 source:plural/3sp/past/pastp/presp/comparative/superlative /target:									
syllable	音節	noun	syllables						
new	新規の	adj						newer	newest
go	行く	verb		goes	went	gone	going		
prosody	韻律	noun							

図 2 UTX-Simple 0.9 版の基本記述例

漫然とした分野の辞書は作らず、スポーツ分野、IT 分野など、辞書ごとに特定の使用分野を限定する。

現状では個人単位の翻訳において、特にオープンソースの翻訳に関して辞書が共有できないことが問題となっている。辞書はさまざまな提供元に拡散しており、それぞれにフォーマットや使用ライセンスが異なる。言語資源を集約できれば、各国語間のローカライゼーションを一挙に加速できる。

UTX は仕様がオープンであり、UTX の辞書を真に普及させるために誰でも参加できる「みんなで作る辞書」共有辞書コミュニティを構築する。そのためには以下の取り組みが必要である。

- 辞書を作成・共有・蓄積するための辞書コミュニティと流通インフラの確立
- 公式辞書コミュニティでは、品質を保証した辞書を有償提供（AAMT または関連組織が管理する）
- オープン辞書コミュニティでは、オープンソース的な許諾による自由かつ無償で相互利用（AAMT または関連組織はホスティングのみを行う）

多様な需要に対応するため、公式辞書と無料辞書を作り、それぞれを管理・共有するためのコミュニティを作る。公式辞書は有償だが、専門家が監修して信頼できる内容を保証する。無料辞書は、内容の正確さが保証されないが無料で利用できる。

すでに共有化・標準化ワーキンググループでは、Mozilla 24 などのイベントに参加して UTX について紹介するなど、オープンソースのコミュニティとも意見交換を行っている。

UTX のフォーマット自体は単純だが、分散した言語資源を集約し、「集合知」として活用できる。そのために参加者にとって辞書作

成の継続的な動機付けを維持し、参加者が不公平にならない枠組みの作成が必要になる。

5. 今後の展望と課題

仕様や指針の策定と同時に、辞書データの作成および収集、各種ツールの開発と活用、UTX の実効性の評価と確認、そして各方面との協力・連携を進めていく。最終的には UTX の ISO 標準化を目指す。

5.1 ライセンス体系の策定

有償・無償のそれぞれの辞書形態について、辞書の配布を単純かつ明快で単一のものとする。特に無料・オープン辞書では相互利用可能とし、改変と商業利用が許諾される必要がある。使用権の問題が解決できるところから作成していく。

5.2 UTX-XML の仕様策定

UTX-Simple に基づき、より高機能な UTX-XML の仕様の詳細を決定していく。

5.3 表記指針の策定

辞書で音引き、中黒などの表記が統一されていないと、異なる辞書を統合するたびに修正が必要になる。翻訳の現場では、各社の表記の違いを吸収するだけで多大な労力が日々浪費されている。そのため、辞書表記の統一指針を定めて遵守する必要がある。

5.4 辞書データの作成・収集

実際に翻訳が必要とされている分野をいくつか選定し、UTX-Simple、UTX-XML の仕様によって具体的な辞書データ作成・収集を行う。その辞書を使った翻訳を行うことにより、UTX の仕様の修正・確定を行っていく。

5.5 各種ツールの開発と活用

以下のツール開発が必要になる。

UTX 変換ツール (正規化ツールを含む)

翻訳ソフトや翻訳サイト独自の形式と、UTX 形式を相互に変換するツール。UTX 形式の仕様が正しく実装されているか検証する正規化ツールを含む。また、UTX-XML と UTX-Simple のコンバーターも必要である。

用語抽出・辞書作成ツール

一語一語登録するのではなく、原文を解析して一括して必要な辞書登録を行うためのツール[6,7,8]。

辞書検索ツール (用語集検索ツール)

辞書や用語集を直接検索して参照できるツール。

5.6 UTX の実効性の評価と確認

UTX の普及のためには、UTX を使って翻訳精度が向上し、翻訳効率が改善することを実証する必要がある。評価に関しては、AAMT でもテストセットを作成し、UTX 辞書の効果を確認する予定である。

5.7 各方面との協力・連携

共同執筆者が運営する、ユーザが専門用語を登録できるコミュニティ型機械翻訳サイト「訳してねっと」[9]との連携をはじめ、共同執筆者以外にも、UTX にはすでに各方面からの協力が得られている。今後は、Edict、多数の辞書やシステムをつなぐ他のプロジェクト(言語グリッド)、言語資源を収集・管理・配布する組織(GSK)などとの協力・連携も考えている。実際に UTX を使ってもらい検証していくとともに、ツール開発面でも協力・連携していきたい。

UTX の趣旨に賛同している組織および個人(順不同、敬称略)

富士通研究所、秋桜舎、OKI、情報通信研究機構、東芝ソリューション、クロスランゲージ、NHK、シャープ株式会社、NEC、坂本義行

参考文献

- [1] <http://aamt.info/>
- [2] <http://tran.blog.shinobi.jp/Entry/316/>
- [3] 富士秀. 英日機械翻訳文の読解に関する評価実験. 言語処理学会第2回年次大会論文集, 1996.
- [4] <http://www.lisa.org/>
- [5] <http://www.lisa.org/standards/tbx/>
- [6] <http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/atr.html>
- [7] 小山, 影浦, 竹内. 日本語専門分野テキストコーパスからの複合語用語抽出, 情報処理学会, 自然言語処理研究会, 176-NL-2006, pp.55-50, 2006.
- [8] 日野, 佐々木, 宇津呂, 土屋, 中川, 佐藤. ウェブからの関連語収集手法を用いた専門用語の訳語推定. 言語処理学会第11回年次大会論文集, 言語処理学会, pp.21-24, 2005.
- [9] <http://yakushite.net/>

問い合わせ先

AAMT (アジア太平洋機械翻訳協会) では、UTX の仕様策定や辞書作成、評価にご協力いただける方を募集しております。現時点では、日本語、英語、中国語を優先しています。興味のある方は下記ページをご覧ください。

- UTX についての説明

<http://www.aamt.info/japanese/utx.htm>

- UTX メーリングリスト (参加自由)

<http://groups.yahoo.co.jp/group/UTX>

- Francis Bond 氏の UTX 辞書 (言語学)

<http://www2.nict.go.jp/x/x161/en/member/bond/data/>

Introduction to UTX, a specification for a shared user dictionary

Okura Seiji†, Yamamoto Yuji††, Murata Toshiki‡, Uchimoto Kiyotaka‡‡,
Michael Konin Kato+, Shimazu Miwako*, Suzuki Tsugiyoshi**, Francis Bond ‡‡

Sharing/Standardization Working Group, MT Research Committee

Asia-Pacific Association for Machine Translation

†Fujitsu Laboratories, ††CosmosHouse, ‡Oki Electric Industry, +Learning consultant

‡‡National Institute of Information and Communications Technology

*Toshiba Solutions, ** Cross Language

1. Foreword

In order to use a machine translation system for computer aided translation, careful tuning of user-created dictionaries is indispensable. These user dictionaries, however, may not be always compatible between different MTs, often rendering the effort to create such dictionaries futile. To address this issue, AAMT (Asia-Pacific Association for Machine Translation) [1] has undertaken to establish a specification of sharable dictionaries, which can be used across different machine translation systems. AAMT created its first version of specification, UPF, with support from IPA (Information-technology Promotion Agency, an institute in Japan) in 1995. In 2006, AAMT started to create a new specification to reflect and incorporate the subsequent advancement of technology and the changing usage of MT. In 2007, the new format received a new name "UTX," short for universal terminology exchange. As of 2008, AAMT is working to establish UTX-Simple, which is the simple, stripped-down version of UTX before building the full XML version. As for the future plan in 2008, AAMT plans to advance the project by revising and expanding UTX specifications, selecting the actual translation domain, collecting and compiling the terminological data of the domain, and

seeking collaboration with translation-related projects led by other organizations or individuals. In addition, AAMT is conceiving of creating a user community for producing, sharing, and accumulating user dictionaries in a sustainable way. This article introduces the specification of UTX-Simple, and describes our process and methods of sharing and reusing dictionaries.

2. Basic of UTX

As I mentioned earlier in the foreword, careful tuning of user-created dictionaries is indispensable in order to use a machine translation system for computer aided translation. When we use commercial high-end translation software for computer-aided translation workflow, specialized terminology, names of persons, and place names in the target document are often not included in basic system dictionaries, and they are not translated as well as one would expect. These terms may be missing from optional specialized dictionaries. It is known that if these terms are registered into a user dictionary, the precision of a machine translation system can be improved. For example, let's assume that the following English sentence is translated into Japanese with translation software.

XML declaration may contain information about character encoding and external dependencies.

The result would be:

「XML 公表が文字符号化と外部の属国についてのインフォメーションを含んでいるかもしれません。」

In this translation, certain terms are used, such as 公表, 符号化, 属国, and インフォメーション, but these are not suitable for its context and domain. If the suitable terms are properly registered to a user dictionary, however, a more suitable translation is generated as follows: [2]

「XML 宣言が文字エンコーディングについての情報と外部の依存関係を含むことがあります。」

User dictionary compilation is important, because addition of compound words such as "XML declaration" into a user dictionary decreases failures in syntax analysis[3], and it also improves the accuracy of statistics translation.

However, user dictionary compilation is a very time-consuming process for an individual user, and its effect is also not immediately clear for relatively small translation projects. In addition, the formats of individuals' dictionaries are varied if they do not share the same MT, and sharing these dictionaries can be difficult.

TBX [5] of LISA [4] is an existing standardized specification for terminology, but it is intended to be used by professional terminologists only. Its complex nature denies commitments by average MT users, and it is not widely in use. In order to address these problems, AAMT is establishing a new standard

called UTX.

The features of UTX are as follows:

1. **"Dictionary for the user" - simple and easy to use:** A complicated specification merely increases users' burdens, and it will be forgotten eventually. A specification must be simple and practical to reflect and include actual MT users' needs, viewpoints, and scenarios.
2. **Entry as a "technical term":** UTX clearly defines the domain of a dictionary, and adheres to the principle of "one word, one meaning." The number of entries in a dictionary should be well-chosen, and differences of word usages must be clearly defined. An entry must be a unique term within an applicable domain.
3. **Improvement in translation accuracy:** When UTX dictionaries become widespread, sharing dictionaries can be drastically simplified. Users can compile user dictionaries more efficiently by exchanging existing data, and UTX will eventually contribute to improvement in translation accuracy.
4. **Multilingual and monolingual dictionary:** UTX is designed for dictionaries not only between two languages but also among multiple languages. In addition, UTX also includes a specification for a monolingual dictionary, which can be used for proofreading tools for terminological standardization, etc.
5. **Inclusion of the information to support managing and sharing dictionaries:** The XML version of UTX includes detailed information, such as the dictionary creator, entry creation timestamp, information intrinsic to discrete machine translations, etc., and it enables users to manage and share UTX dictionaries effectively.
6. **Promotion of localization of software:** Especially in open source localization projects,

translation is carried out individually, and terminological standardization can be difficult. By exchanging and sharing translation resource systematically through the use of UTX, more and more common dictionaries become available, and they increase the efficiency of translation exponentially.

In January, 2008, the European Commission announced that it will release the translation of about 1 million sentences into 22 languages for free. In Japan, the thick language barrier hinders the translation of software and other contents, resulting in immense economic educational, and cultural losses every day. Sharing of translation assets in all industries will be the key to future industrial development.

sharing and reuse of dictionaries is possible through user communities on LAN or the Internet. The major merits of UTX for translation software manufacturers are: the entire market of machine translation is enlivened by the promotion of user dictionaries; thus new demands and applications of MT can be explored. UTX-XML format retains entry properties that are proprietary to manufacturers, and no data is lost during the conversion.

The following two tasks are important for the specification establishment of UTX.

- Establishment of specifications (UTX-Simple and UTX-XML)
- Creation of actual dictionary data and the community for sustainable creation, sharing,

- A dictionary file consists of one header and one or more entries.
- The header is located in the first line of the file.
- One entry is stored in each subsequent line.
- The line starting with a sharp "#" is a comment line (which only occurs at the beginning of the file).

[Header] The description of the dictionary file. The delimiter is an "en" space.

#UTX-S <version number> <source language>/<target language> <last updated> <other properties>

Example: #UTX-S 0.9 en-US/ja-JP 2007-12-03T14:28:00Z+09:00

* UTX-S means "UTX-Simple". The language name (ISO 639) and the date/time (ISO 8601) formats conform to ISO[2, 3].

In case of a monolingual dictionary, the target language is omitted.

Example of the fourth column:

source:plural/3sp/past/pastp/presp/comparative/superlative/target:..../optionXXXX/optionYYYY

(The information on a source word is described after "source:".) Abbreviation denote as follows: the plural = plural form, 3sp = third person singular, past = past form, pastp = past participle, presp= present participle, comparative= comparative degree, superlative= superlative degree, the translation is described after target: .

[Entry]One line is delimited to columns by tabs. The contents described in each column are as follows:

First column	Second column	Third column	Fourth column and more
Word of source language (Headword)	Word of target language (Translation)	Part of speech	Other properties (optional)

Fig. 1 Specification of UTX-Simple 0.9

The major merits of UTX for MT users are: it is a collection of plain formats that are easy to create; it improves the translation accuracy in various domains;

and accumulating of dictionaries

In this article, we illustrate these two tasks, as well

as the future plan of UTX and its issues.

3. Specification of UTX-Simple

The following two formats of specifications are defined in UTX.

- UTX-XML (XML format)
- UTX-Simple (tab-delimited text format)

UTX-XML (XML format) includes comprehensive information required for a user dictionary. UTX-Simple (tab-delimited text format) is a specification that requires only three types of information: a source word, its translation, and the part of speech of the source word. All other properties are optional. AAMT will eventually establish XML format specification, however, it has started by creating UTX-Simple, which is highly practical and can be easily shared to expand and sustain the community of UTX.

In order to make a specification usable, it is necessary to make it plain and easy to use. In order to encourage the use of user dictionaries with a machine translation system, entry addition must be simpler and more straightforward than the present system. As the UTX-Simple dictionary is simple, it can significantly reduce the time and effort necessary for compilation, and the sharing of dictionaries is accelerated. The specification is designed so that it carries only essential information while keeping its usability. The current version is 0.91. The basic specification is shown in Fig.1.

The following issues were discussed in the process of establishing UTX-Simple:

- (1) Should priority be given to readability for human, or to ease-of-processing for machines?

We also discussed about what kind of information to include, and reached to the conclusion that specification requires only three essential types of information: a source word, its translation, and the part of speech of the source word. For example, the

information on the plural form of a noun may be required in the translation from/to English. However, the plural form is not an essential piece of information for all the parts of speech of all languages. In some situations, information of the plural forms could improve the translation result, but it also significantly increases the burden of the creator of the dictionary if it were obligatory. In order to ensure smooth creation, we concluded that the information of the plural forms as optional, but not essential.

- (2) Comparison of the XML version (UTX-XML) and the simple version (UTX-Simple)

As UTX-Simple does not retain detailed information of its creator, the timestamp for each entry, etc., it is not suitable for a compilation of a permanent, versatile dictionary. However, it is practical and easy to create, use, and share. It may be used by a variety of users regardless of their background knowledge in advanced linguistics. The establishment of the XML version may require some more time.

- (3) Should we fix the order of columns (Column X always corresponds to Property Y etc.)?

A column accommodates both information common to all languages, as well as information unique to a specific language. The first to third columns are assigned to three essential properties, while language-specific properties are added after the forth column as they are necessary.

- (4) Should we define language-specific columns for every language?

The main languages of focus are currently on Japanese, English, and Chinese. We plan to include all existing languages defined in ISO in the future. It is not realistic or practical to define all columns required for all languages altogether at this time. Therefore, we left the liberty of defining necessary

optional columns to users. Feedbacks from actual usages will be incorporated to decide the optimal definitions of properties in the future. We would recommend using certain name for certain property for particular language pairs.

(5) About creation of guidelines

It is necessary to establish guidelines for notation, i.e. how to describe each entry, for each target language. The tools which read UTX-Simple would issue warnings following the guideline, if they find problems in the dictionaries. For example, if one or more mandatory properties are missing, a warning is issued according to the corresponding option of the tool. To cope with human errors, the tool could be set to skip an entire line, if the format has problems.

The following is the basic guideline:

- Difference between a missing value (default) and "not applicable"

Other missing values (defaults) are subject to handling by MT systems.

To what extent a tool uses the information contained in a dictionary depends on the specification of the tool itself.

- The user-defined columns which are not identified by the tool will be kept intact.
- An English noun phrase starts with a lower-case letter (except for proper nouns etc.).

An example of UTX-Simple is shown in Fig.2.

4. Creation of a user community

As the basic policy of creating and collecting dictionary data, the dictionary must belong to a specific domain, such as sport, IT, medicine etc.No

"generic" dictionary should be permitted.

In open source localization projects, translation is carried out individually, and dictionaries are not shared as they should be. Dictionaries are scattered across various providers, and their licenses and formats are also varied. If these scattered language resources are centralized, the localization between different languages is significantly accelerated.

In order to spread UTX dictionaries in which anyone can participate in creation, and in order to realize "open dictionaries for everyone," a shared dictionary community should be established. For that purpose, the following efforts are required:

- AAMT will establish two types of dictionary communities for producing, sharing, and accumulating dictionaries. AAMT will also establish a framework for distributing UTX dictionaries.
- The official dictionary community (managed by AAMT or its delegate) offers supervised dictionaries with guaranteed quality for a fee.
- The open dictionary community offers free dictionaries with open source license and promotes mutual exchanges. AAMT or its delegate provides hosting service only, but no management or guarantee.

In order to meet various needs, official dictionaries and free dictionaries should be distinguished. The corresponding community for managing and sharing each type of dictionary must also be created. The official dictionary community (managed by AAMT or its delegate) offers supervised dictionaries with a

#UTX-S 0.9 en-US/ja-JP 2007-12-03T14:28:00Z+09:00 source: /target:plural/3sp/past/pastp/presp/comparative/superlative									
syllable	音節	noun	syllables						
new	新規の	adj						newer	newest
go	行く	verb		goes	went	gone	going		
prosody	韻律	noun							

Fig.2 Example of the basic notation of UTX-Simple

guaranteed quality for a fee. Free dictionaries can be used for no fee, although the correctness of the contents is not guaranteed.

Sharing/Standardization Working Group of AAMT has already introducing UTX to open source community by participating in events such as Mozilla 24 (hosted by Mozilla). We exchanged opinions on various topics concerning UTX with open source contributors.

Although the format of UTX is simple, it effectively centralizes dispersed language resources through a common format, to be used as a collective intelligence. Therefore, we must maintain and stimulate the motivation for the dictionary creation among the community participant, and create a framework to ensure fairness for the participants with different degrees of contribution.

5. The prospect and issues

We are currently committed to the following tasks: the establishment of specifications and guidelines for creating dictionaries, production, and collection of dictionary data, development of various tools and how to use them, evaluation of the effectiveness of UTX, and collaboration and cooperation with other parties. Eventually, our goal is to make UTX into one of ISO standardizations.

5.1 License system

Make the distribution system of dictionaries simple and clear, both for paid and free dictionaries. Especially, a free/open dictionary must allow mutual use and changes, and its commercial uses need to be permitted. We will start with available royalty-free dictionaries.

5.2 Establishment of UTX-XML

Based on UTX-Simple, the details of the specification of more sophisticated UTX-XML will be determined.

5.3 Establishment of a notation guideline

Notation of dictionaries must be standardized through guidelines for each language and its dialects, if applicable. If notation is not standardized among dictionaries, we must unify them each time we merge one dictionary to another, and this process would be time-consuming. For instance, in case of Japanese, the notation of prolonged sound marks, "middle dots," etc. must be standardized. Equivalent in English would be variations of British and American spellings, punctuation rules, hyphenation etc. In the actual projects of translation, significant time is wasted only to absorb the differences of the notations defined by different parties. Therefore, it is necessary to define and observe the standardization guideline of dictionary notation, if necessary, for different specialized areas.

5.4 Production and collection of dictionary data

We will select some domains for which translation is needed, and build and collect actual dictionary data in accordance with the specification of UTX-Simple and UTX-XML. By carrying out translation with the dictionary and collecting feedbacks, the UTX specification will be further improved.

5.5 Development and use of various tools

We will need to develop following tools:

UTX converters (including parsers)

We will need tools which convert from a format unique to a translation application or a translation site to UTX format, and vice versa. A parser which verifies the conformity to the UTX specification must also be included. In addition, a converter between UTX-XML and UTX-Simple is also required.

Term extraction and dictionary building tools

Tools to analyze multiple documents, extract terms,

and add them to a user dictionary [6, 7, and 8] or make a new one instantly, not by building a dictionary one-word-at-a-time.

Dictionary search tool (glossary search tool)

We will need tools to perform a direct search of a dictionary and glossary to see the translation of a word.

5.6 Evaluation of the effectiveness of UTX

In order to promote UTX, it is necessary to prove that the performance and accuracy of translation are improved by the use of UTX. As for evaluation, the effectiveness of a UTX dictionary is due to be checked with test sets created by AAMT.

5.7 Collaboration and cooperation with other parties

Besides the co-authors (which include most of MT developers in Japan) of this article, collaborations with other parties has already been began for UTX, including a cooperation with Oki's community-oriented machine translation site "Yakushite-net" [9] in which a user can add his/her own technical terms. In the future, collaboration and cooperation with Edict, the other projects which connect many dictionaries and systems (such as Language Grid), and organizations (GSK) which collect, manage, and distribute language resources etc. are also envisioned. While providing UTX to other parties and verifying its effectiveness, we are also interested in collaborating in terms of tool developments.

The organizations and individuals who support the UTX project

(In no particular order, titles omitted)

Fujitsu Laboratories, CosmosHouse, OKI, National Institute of Information and Communications Technology, Toshiba Solutions, Cross Language, NHK, Sharp Corporation, NEC, Sakamoto Yoshiyuki

6. Bibliography

- [1] <http://aamt.info/>
- [2] <http://tran.blog.shinobi.jp/Entry/316/>
- [3] Fuji, "Evaluation experiment to reading comprehension of an English-Japanese machine translation sentences," the collected papers of the 2nd annual meeting of Association for Natural Language Processing, 1996.
- [4] <http://www.lisa.org/>
- [5] <http://www.lisa.org/standards/tbx/>
- [6] <http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/atr.html>
- [7] Koyama, Kageura, and Takeuchi, "The compound term extraction from a Japanese specialized-area text corpus," Information Processing Society of Japan, Natural Language Processing Study Group, 176-NL-2006, pp.55-50, 2006.
- [8] Hino, Sasaki, Utsuro, Tsuchiya, Nakagawa, and Sato, "Translation inference of technical terms using the related term collection technique from the Web," The collected papers of 11th annual meeting of Association for Natural Language Processing, pp.21-24, 2005.
- [9] <http://yakushite.net/>

AAMT (Asia-Pacific Association for Machine Translation) invites organizations and individuals who could collaborate with us to establish the specification of UTX to provide and/or build dictionaries, and to perform evaluation. At present, priority is given to the Japanese, English, and Chinese languages. If you are interested, please do not hesitate to contact us from the following page:

About UTX:
<http://www.aamt.info/japanese/utx-e.htm>

UTX mailing list:
<http://groups.yahoo.co.jp/group/UTX>
(Anyone can participate in this mailing list, but the correspondence is mostly in Japanese. We are planning to start another mailing list in English.)

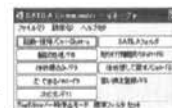
UTX-Simple dictionaries by Francis Bond
<http://www2.nict.go.jp/x/x161/en/member/bond/data/>

共有ユーザー辞書仕様UTXと 翻訳ソリューションSATILAの概要

発表者：秋桜舎 代表 山本ゆうじ
(AAMT 機械翻訳 課題調査委員会ワーキング グループ3)

秋桜舎の活動紹介

- 言語・翻訳コンサルティング
 - 実務翻訳
 - 翻訳工学
 - 翻訳ワークフローSATILA
 - 実務日本語



SATILA Commander

CosmosHouse.com

2

翻訳工学

- 合理的な方法論
- 個人プレーではない、チームワークとしての実務翻訳
 - 用語集
 - 「実装可能なスタイルガイド」
- 翻訳資産の共有と再利用

3

MTとTM

- 機械翻訳（翻訳ソフト）
MT (machine translation)
 - 自然言語処理を行う
 - 単語単位で再利用
 - 用語集（ユーザー辞書）が翻訳資産
- 翻訳メモリー ツール
TM (translation memory)
 - 文単位で再利用
 - 翻訳メモリーが翻訳資産

Reports configuration information of the installed product.

4

翻訳支援と自動翻訳

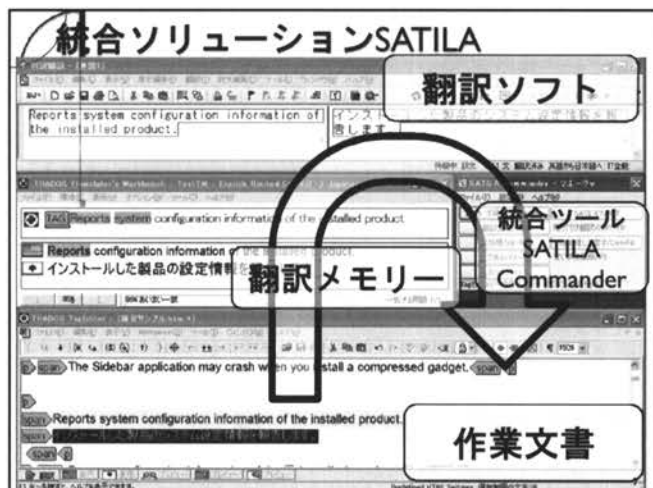
- 自動翻訳
 - 一般ユーザー向け
 - ボタン一つでソフトにお任せ
- 翻訳支援
 - プロの翻訳者向け
 - 予測不可能な翻訳ソフトを「指定した通りに」動かす
 - →ユーザー辞書
 - UTX（新しい共有ユーザー辞書仕様）

5

翻訳ソフトの利点

- 自動適用による用語の正確さ・効率化
- 言い換え表現の蓄積—再利用と共用
 - 教育的効果
- 専門辞書の単語数
 - 不適切な訳語はユーザー辞書でカバーする

6



実務日本語

- 合理的な規則に基づいて実用文書を作成するための、日本語の表記と作文の指針
- 人間に分かりやすく機械に処理しやすい（制限言語ではない）
- 自動・手動での日英翻訳の改善に必須
- 技術・法律・学術文書、ビジネス文書作成・共有・再利用のために

拡張指針
基本指針 仕様

読者にとって読みやすく簡潔
読者にとって読みやすく簡潔
読者にとって読みやすく簡潔
読者にとって読みやすく簡潔
読者にとって読みやすく簡潔

重要なのはどっち？

- 開発者の視点「自然言語処理を使うこと」

ギャップ

- ユーザーの視点「正確に訳すこと」
 - 「最初から最後まで自然言語処理」にこだわらない
 - プロの翻訳者の実際的な視点

「標準化・共有化」WGの活動

UTX仕様策定

辞書共有コミュニティの創設と維持

辞書データの作成と収集

ユーザー辞書の意義

- 業務用翻訳ソフトでは900万語に近い（語彙力デモ）
- しかしそれでも不十分
 - 「辞書に載せる基準に達しない語」であっても、登録しないと訳せない
- ↓
- ユーザー辞書が必要

意識フレーズの翻訳知識

"in the same boat"	「同じような状況にある」
"be shuffled to the sidelines"	「指をくわえて見ている」
"for one"	「たとえば」
"take an active part"	「積極的に参加する」

UPF

- 共有辞書の最初の試み（SGML形式）
 - →普及せず
- 失敗の理由
 - ユーザー辞書自体の意義、使い方、作り方が理解されていない
 - 共有のためのインフラがない
 - ユーザーにも翻訳ソフトメーカーにも強い動機がない

13

UTXの特長

(universal terminology exchange)

- シンプル
- 分かりやすく、使いやすい「ユーザーのための辞書」
 - 最低限、見出しと訳語だけでも使用可能→作りやすい
 - 多言語にも対応
- 共有のための情報も保持
 - 辞書作成者／作成日時
- 「専門用語」という観点
- 分野を明確化
- 一語一義
 - 1分野内では、1語1訳の方針
 - 特定の文脈（分野）で訳語が一義的に定まる語

14

明確なメリット

- ユーザーにとって
 - 個々の特定分野での翻訳精度向上
 - インターネット、LANなどで他の人と辞書を共有
 - コミュニティを通じた辞書の共有と再利用
- メーカーにとって
 - ユーザー辞書利用の促進による活性化
 - 新規需要の掘り起こし
 - メーカー固有の辞書項目も完全に保持

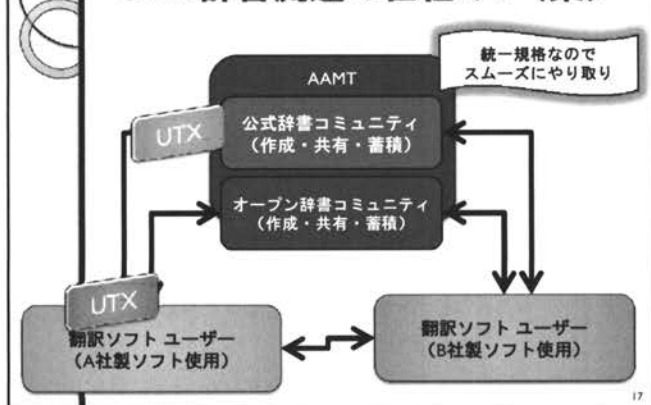
15

「みんなで作る辞書」共有辞書コミュニティの構築

- 辞書を作成・共有・蓄積するための辞書コミュニティと流通インフラの確立
- 公式辞書コミュニティ
 - 品質を保証した辞書を有償提供（AAMTが関連組織が管理）
- オープン辞書コミュニティ
 - オープンソース的な許諾による自由かつ無償で相互利用（AAMTが関連組織はホスティングのみ）
- 「訳してねっと」との連携も検討

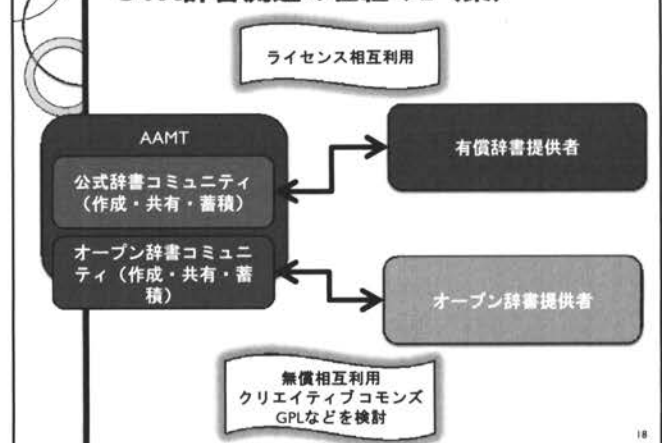
16

UTX辞書流通の仕組み1（案）



17

UTX辞書流通の仕組み2（案）



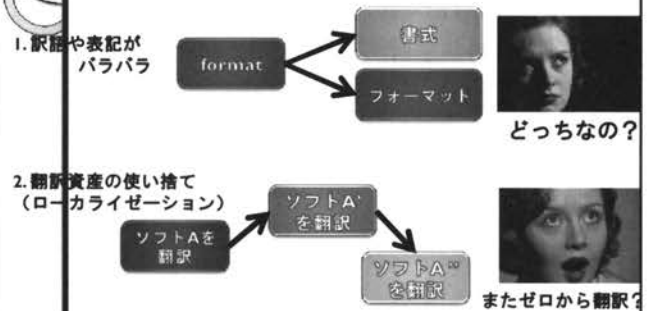
18

さまざまな活用

1. オープンソース翻訳
2. 専門用語辞書・社内用語集として
3. 個人間での異文化間コミュニケーション
4. 開発途上国向けの翻訳支援

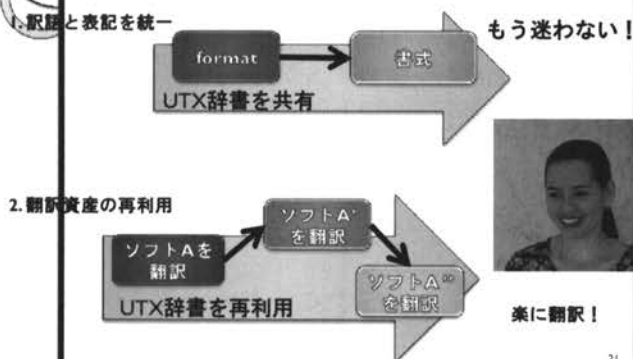
19

オープンソース翻訳の課題



20

UTX辞書で解決！



21

専門用語辞書・社内用語集として

- 問題点
 - 社内文書作成用の用語集と翻訳用の用語集がバラバラ。
 - 特定分野の専門知識を社内で蓄積したいが、形式がバラバラで生かせない。
- →UTXの利点
 - 単一言語の用語集としても柔軟に利用できる。
 - 統一された規格なので、さまざまなツールとのデータのやり取りが簡単。
 - 既存の用語集からの取り込みも容易にできる。

22

個人間での異文化間コミュニケーション

- 問題点
 - 海外の作家や、映画の俳優にファンレターを出したいが固有名詞が分からない。
 - スポーツやゲームについて、海外の友人と気軽にチャットしたいが、選手名やゲームの用語が分からない。
- →UTXの利点
 - 市販の専門辞書には載っていないニッチな用語集が機械翻訳で利用できる。

23

開発途上国向けの翻訳支援

- 問題点1
 - NPOの活動などで、人手や資金が不足している。
- →UTXの利点1
 - 翻訳支援や自動翻訳を低コストで行える。
- 問題点2
 - 対訳集や辞書の少ない言語がある。
 - 医療など特定分野の辞書が少ない場合もある。
- →UTXの利点2
 - 辞書コミュニティに集中して蓄積がされるため、対訳集の少ない言語でも徐々に辞書が構築できる。

24

汎用辞書形式として

- 読みがなのデータ
 - →音声認識
 - →音声合成
- 「正しい表記」の情報
 - →文章校正（デモ）

25

UTXの種類

- UTX-Simple
 - タブ区切り形式。必要最低限の情報のみで簡便にやり取り。現状の各社のテキスト書き出し形式を置き換える。
→必要最低限とは？
- UTX-Main
 - XML形式。すべての機能を備える。

26

用語決定の過程

- トップダウン方式（用語管理者方式）
 - 企業・組織内で、用語管理者が用語を決定する。
 - 利点：厳密な用語統一ができ、バラツキがない。用語調停機構が不要。
 - 欠点：用語管理者に負担が集中する。
- ボトムアップ方式（コミュニティ方式）
 - オープンソースなどで、個々の参加者が提出した用語を（自動的な）用語調停機構により評価し、決定する。
 - Cf. Digg, Slashdot.orgのカルマ
 - 利点：多数の意見を反映できる。
 - 欠点：全体を見渡す人がいないので、バラツキが出る。
- 考察
 - トップダウンでもボトムアップでも、承認用語と未承認用語は区別する必要がある。

27

各種ツールの活用・開発

- 用語抽出・辞書作成ツール
 - 一語一語登録するのではなく、原文を解析して一括して必要な辞書登録を行うためのツール。
- UTX変換ツール（正規化ツールを含む）
 - 翻訳ソフトや翻訳サイト独自の形式と、UTX形式を相互に変換するツール。UTX形式の仕様が正しく実装されているか検証する正規化ツールを含む。
- 辞書検索ツール（用語集検索ツール）
 - 辞書や用語集を直接検索して参照できるツール。

28

何が必要とされているか

- UTXのための辞書コンテンツそのもの
- 辞書コンテンツを育てる仕組み
 - トップダウン
 - ボトムアップ
- ツール開発
 - 特に用語抽出・辞書作成ツール

29

まとめ

- 翻訳ソフトはユーザー辞書のカスタマイズにより大きく精度が向上する
- UTXはそのための鍵になる
- 普及には辞書共有のためのインフラが重要
- UTXは日本からの情報発信を促進する

発表者の連絡先：

<http://cosmoshouse.com/mail.htm>

30

協会活動報告

(2007 年 12 月～2008 年 5 月)

予算理事会

2008 年 3 月 (E-mail にて議論)

- ①第 1 号議案 2007 年度事業報告 (案) ② 第 2 号議案 2007 年度決算報告 (案)
- ③第 3 号議案 2008 年度事業計画 (案) ④ 第 4 号議案 2008 年度収支予算 (案)
- その他

監事会

2008 年 4 月 21 日 平成 19 年度会計報告

機械翻訳課題調査委員会

2007 年 12 月 3 日 (2007 年度 第 7 回)

- ①WG③Mozilla DocFest M24 参加報告
- ②各 WG の活動について (各 WG に分かれて議論) ③活動内容の報告 (各 WG から)
- ④活動内容についての議論 ⑤まとめと次回委員会について

2008 年 1 月 25 日 (WG②個別委員会)

MT ユーザアンケートのための最終調整

2008 年 2 月 4 日 (2007 年度 第 8 回)

- ①前回委員会の議事録の確認 ②各 WG の活動について (各 WG に分かれて議論)
- ③活動内容の報告 (各 WG から)
- ④活動内容についての議論 ⑤まとめと次回委員会について

2008 年 2 月 5 日～2 月 25 日 AAMT ホームページ上にて「AAMT MT ユーザアンケート」実施

2008 年 3 月 10 日 (2007 年度 第 9 回)

- ①ボンド氏によるレクチャー ②前回委員会の議事録の確認
- ③各 WG の活動について (各 WG に分かれて議論) ④活動内容の報告 (各 WG から)
- ⑤活動内容についての議論 ⑥まとめと次回委員会について

2008 年 4 月 5 日 (2008 年度 第 1 回)

- ①前回委員会の議事録の確認
- ②各 WG の活動について (各 WG に分かれて議論) ③活動内容の報告 (各 WG から)
- ④活動内容についての議論 ⑤まとめと次回委員会について

2008 年 5 月 14 日 (2008 年度 第 2 回)

- ①前回委員会の議事録の確認
- ②各 WG の活動について (各 WG に分かれて議論) ③活動内容の報告 (各 WG から)
- ④活動内容についての議論 ⑤まとめと次回委員会について

AAMT/Japio 特許翻訳研究会

2007 年 12 月 12 日

- | | |
|------------|---------------|
| ①前回記事録の確認 | ②今年度の研究内容について |
| ③次回の開催について | |

2008 年 1 月 25 日

- | | |
|------------|---------------|
| ①前回記事録の確認 | ②今年度の研究内容について |
| ③次回の開催について | |

2008 年 3 月 14 日

- | | |
|---------------|---------------|
| ①前回の議事録の確認 | ②年次報告書作成経過報告 |
| ③今年度の研究活動について | ④来年度の研究活動について |
| ⑤次回の開催について | |

編集委員会

2007 年 12 月 AAMT Journal No.41 発行

2008 年 1 月 21 日

- | | |
|--------------------------|------------------|
| ①AAMT Journal No.42 進捗確認 | ②No.42 以降の企画について |
| ③その他 | |

2008 年 5 月 14 日

- | | |
|--------------------------|----------------|
| ①AAMT Journal No.42 進捗確認 | ②No.43 の企画について |
| ③その他 | |

インターネット WG

- ①AAMT ホームページの更新（毎月）
- ②AAMT Forum メイリングリストの管理
- ③AAMT Forum への情報発信（月数回）
- ④委員会活動におけるメイリングリストの管理
- ⑤会員専用ホームページ開設にむけた準備
- ⑥その他事務局ネットワークのインフラ管理全般

AAMT ジャーナル編集委員会委員長
筑波大学大学院システム情報工学研究科
知能機能システム専攻
宇津呂 武仁

AAMTジャーナル42号をお送りします。

今号では、静岡大学の梶博行先生に巻頭言をご執筆頂きました。

梶先生は、前所属の日立製作所の頃から機械翻訳の研究に取り組んで来られ、静岡大学に移られてからは、JST、NICT等の機関とともに、「科学技術文献を対象とする日中機械翻訳システム開発プロジェクト」にも参加しておられます。

また、今号では、機械翻訳課題調査委員会の活動成果を何件か紹介しております。特に、同委員会委員ならびに編集委員を掛け持ちされておられます大倉氏・山本氏・内元氏におかれましては、大変なご苦勞をいただきまして、「共有ユーザー辞書仕様UTXの現状と今後の展開」の和英文原稿・紹介スライドをご執筆いただきました。

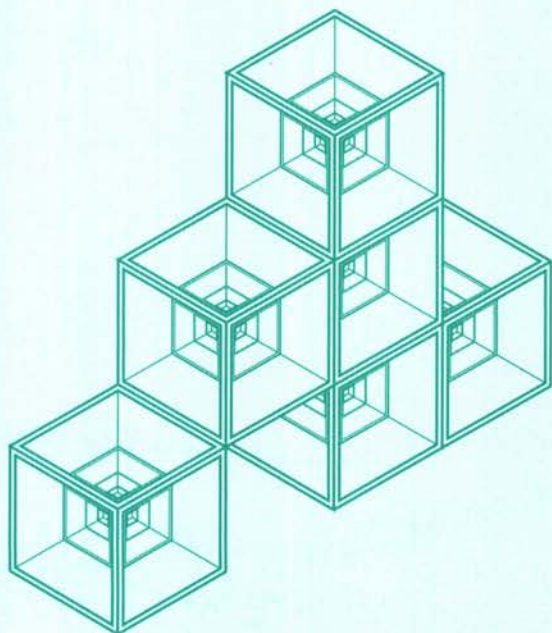
さらに、前号に引き続きまして、AAMT会員に向けたやや個別的な取り組みとして、「AAMT会員のひろば」の企画におきまして、会員自己紹介文を順次掲載しております。今号にはその第二弾として、法人会員の紹介文3件、個人会員の紹介文3件の計6件を掲載しました。本企画に際しましては、昨年度、会員の皆様に自己紹介ページ用の原稿を依頼させて頂きましたが、原稿の集まりのほどは十分ではございませんでしたので、今年度も、順次再度の依頼を継続してさせていただいております。

会員の皆様におかれましては、前号、ならびに、今号の企画を参考にさせていただいて、是非、原稿のご寄稿のほどをお願い致します。

MEMO

MEMO

AAMT



AAMTジャーナル No.42

発行：アジア太平洋機械翻訳協会（AAMT）

ホームページ：<http://www.aamt.info>

住所：〒619-0289 京都府相楽郡精華町光台3丁目5番地

独立行政法人 情報通信研究機構 言語翻訳グループ内

phone：0774-93-4625 fax：0774-93-4627

編集委員会：宇津呂 武仁 内元 清貴 山本 ゆうじ 大倉 清司

熊野 明 三浦 貢 村上 嘉陽

事務局：神崎 享子 喜田 あゆみ

印刷所：株式会社ナビックス

Asia-Pacific Association for Machine Translation

c/o Computational Linguistics Group, NICT

3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0289

Phone:+81-774-93-4625 FAX:+81-774-93-4627

URL:<http://www.aamt.info>