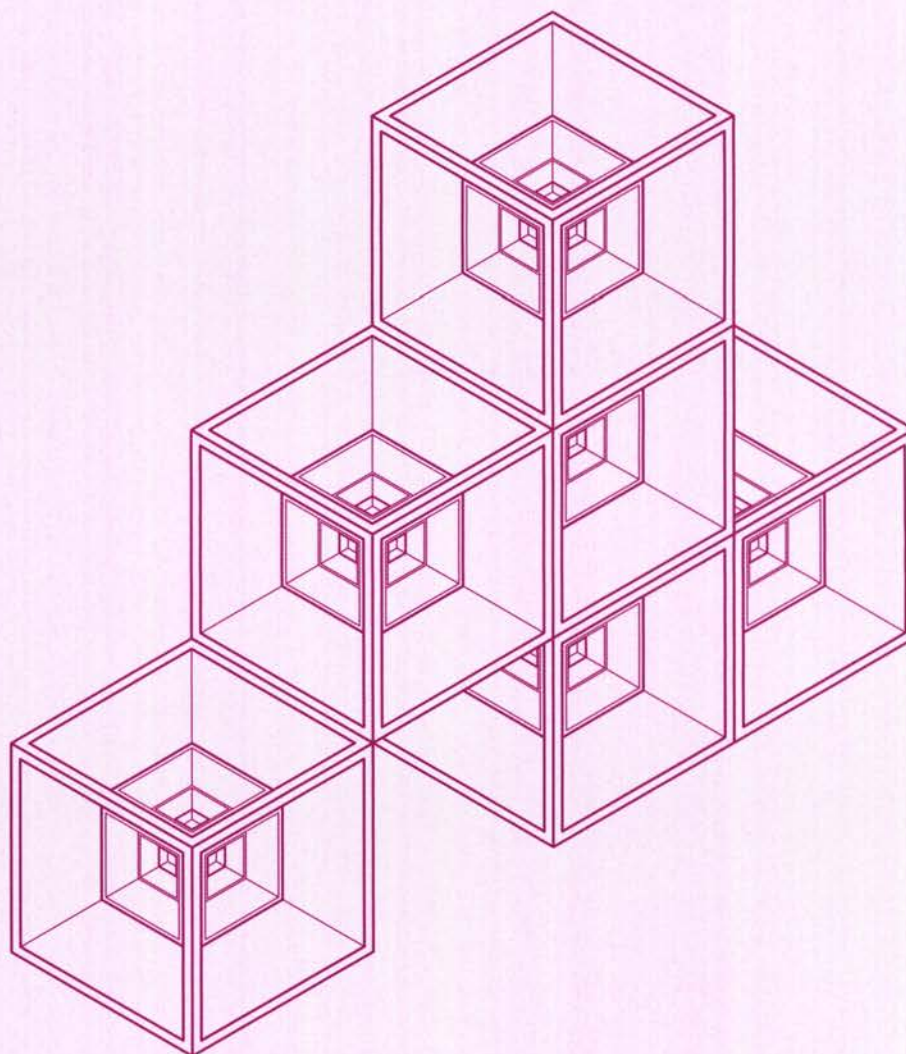


AAMT

Asia-Pacific Association for Machine Translation

Journal



June 2011

No.49

アジア太平洋機械翻訳協会

目 次

巻頭言：	翻訳に「意味」は必要？.....潮田 明.....1
シンポジウム報告：	第1回特許情報シンポジウムを開催して.....江原 暉将.....2
	第10回日中自然言語処理共同研究促進会議に参加して.....江原 暉将.....3
	特許翻訳：品質はなぜ必要か.....竹内 仁恵.....4
	AMTA 2010 報告 — 企業ユーザーの視点から.....立見 みどり.....6
	LISA Open Standards Summit 参加報告.....山本 ゆうじ.....10
プロジェクト報告：	平成21年度音声翻訳実証実験の概要.....安田 圭志、隅田 英一郎.....15
新サービス紹介：	翻訳 SaaS「Eiplaza®/MT」のご紹介.....東芝ソリューション株式会社.....17
新製品紹介：	「オフィス翻訳ピカイチ」.....クロスランゲージ株式会社.....19
翻訳サービス紹介：	「My サイト翻訳」.....株式会社 高電社.....20
MT 活用事例：	言語グリットを用いたスマート翻訳—京大翻訳！—.....
	稲葉 利江子、村上 陽平、田仲 正弘、林 冬恵、石田 亨.....23
	機械翻訳+ポストエディットの実証研究（その2）：英日翻訳での実験結果.....
	山田 優.....30
MT サミット XIII：	MT Summit XIII における「特許翻訳特別セッション」および The 4 th Workshop on
	Patent Translation の開催について.....江原 暉将、横山 晶一.....36
委員会活動報告：	AAMT 機械翻訳課題調査委員会 WG1.....37
	AAMT 機械翻訳課題調査委員会 WG2.....43
	AAMT 機械翻訳課題調査委員会 WG3.....51
活動報告：	協会活動報告（2010 年 11 月～2011 年 5 月）.....AAMT 事務局.....78
編集後記	80

C O N T E N T

Foreword:	Is "Semantics" Unnecessary for Translation?..... <i>A. Ushioda</i>	1
Symposium Report:	Report on The First Symposium on Patent Information Processing <i>T. Ehara</i>	2
	Report on The 10th China-Japan Natural Language Processing Joint Promotion Conference <i>T. Ehara</i>	3
	Patent Translation: Why Quality Matters <i>K. Takeuchi</i>	4
	AMTA 2010 – The Commercial Users' Perspective <i>M. Tatsumi</i>	6
	Report on the LISA Open Standards Summit <i>Y. Yamamoto</i>	10
Project Report:	An Overview of Speech Translation Field Experiments in Fiscal Year 2009 <i>K. Yasuda, E. Sumita</i>	15
New Services:	Machine Translation SaaS "Eiplaza®/MT" <i>Tbshiba Solutions Corporation</i>	17
New Product:	Office HONYAKU Pika-ichi..... <i>Cross Language Inc.</i>	19
Translation Services:	My Site HONYAKU..... <i>Kodensha Co., Ltd.</i>	20
Utilization of MT:	Smart Translation using the Language Grid "Kyodai HONYAKU!" <i>R. Inaba, Y. Murakami, M. Tanaka, D. Lin, T. Ishida</i>	23
	An Empirical Study on Machine Translation plus Post-editing: Experiment Report..... <i>Y. Yamada</i>	30
MT Summit XIII:	Special Session and The 4 th Workshop on Patent Translation in MT Summit XIII..... <i>T. Ehara, S. Yokoyama</i>	36
Committee Report:	Committee for seeking future direction of MT WG1.....	37
	Committee for seeking future direction of MT WG2.....	43
	Committee for seeking future direction of MT WG3.....	51
AAMT Activities:	AAMT Activities (from November, 2010 to May, 2011)	78
Editor's Note:	 <i>T. Utsuro</i>	80

翻訳に「意味」は不要？

株式会社富士通研究所

ソフトウェア&ソリューション研究所 潮田 明

英語学習で日々奮闘している中高生に「文法も知らず原文の意味も全く理解しないで英作文していただきます」と言ってもいったい何のことやら分らないでしょう。今世界中で盛んに研究開発が進められている統計ベース機械翻訳（SMT）の少なくとも初期の実態です。かつては機械翻訳において人手で作成された辞書や文法規則は必須の要素でしたが、SMT においては意味解釈はおろか、人手による辞書も文法規則も基本的に不要です。

SMT とは、大量の対訳コーパスから自動的に単語・フレーズの翻訳確率や単語・フレーズの並び順などに関する統計情報を獲得し、それらをパラメータとする統計モデルを用いて翻訳を行う機械翻訳方式です。1991 年に現在の方式の原型が米国で提案されましたが、当初はまだ対訳コーパスの量もごく限られており、また計算機の処理能力も今とは比較にならないほど低かったため、あまり実用的な方式であるとは受け止められませんでした。しかし 2000 年代に入ると、ネット上のテキスト量の爆発的增加も後押しして対訳コーパスの量が飛躍的に増大し、統計モデルも年々改良され、そしてまた計算機の処理能力も飛躍的に向上したため、SMT は瞬間に発展し、2005 年ぐらいには、「もうすでに SMT は従来のルールベース機械翻訳（RBMT）を完全に凌駕した」という意見が特に欧米の研究開発者の間で多く聞かれるようになりました。

確かに SMT の潜在能力は非常に高いのですが、世界中の研究者がブレークスルーを目指して SMT の研究を推し進めていくうちに、だんだん従来の RBMT が何十年もかけて解決の糸口を模索してきた様々な障壁に同じように突き当たるようになってきました。中でも最大の問題は、文の構造を捉え

ずに統計処理のみで翻訳を行うには限界があるということです。特に文の構造や語順の全く違う言語間の翻訳を行う場合にはこの問題は顕著に現れます。そこで、現在 SMT の研究開発現場では構文解析をはじめ RBMT がこれまで取り組んで来た様々な言語学的課題に形を変えて取り組んでいます。

考えて見れば、学生達が英文法で苦しめられているのに機械だけは文法が不要というのは虫のいい話のように思えます。同じ理屈で、機械は「意味」など気にしなくても翻訳できます、という（少なくとも一部の）専門家達の考えも、根本的にどこかおかしいのかも知れません。ある程度のところまでは行けます、というのと、最後まで要りませんというのは全く別の話です。

一方で、大量テキストから意味的な情報を統計的に自動抽出する研究も進められています。しかしそれだけでは限界があるでしょう。文章とは、表現したい意味や概念が紙の上に（あるいは画面上に）投影された射影物であって、文章を理解するのに必要な意味的な情報がすべて文章の中に織り込まれている訳ではないからです。機械翻訳精度の飛躍的向上のためには、計算機パワーと大量コーパスをフルに活用した統計的な意味情報の抽出と、コーパスへの意味タグ付けあるいはルール等によって人間が言葉の意味に関する情報を明示的に機械に教え込む手段とを、うまく組み合わせることが必要だと思われれます。当社もこれまで RBMT 開発で蓄積して来た技術をテコにそのような方向でのブレークスルーを目指しています。変化する技術潮流の中で、それぞれ異なるアプローチから出発した研究も、違う道のりを辿りながら結局は同じ終着点を見据えるようになって来ているように感じられます。

第1回特許情報シンポジウムを開催して

AAMT/Japio 特許翻訳研究会 山梨英和大学

江原 暉将

2010年11月10日(金)に東京大学で第1回特許情報シンポジウム(The first symposium on patent information processing)が開かれた。主催はAAMT/Japio 特許翻訳研究会である。特許情報は情報処理技術の応用分野の一つとして近年世界的に関心が高まっている。本シンポジウムは、特許情報処理技術の研究開発を促進することを目的としており、研究者、実務家、政府関係者が集まって、構想、方法論、将来展望、実務経験、政策などを議論する場として企画された。プログラムは、基調講演1件、招待講演3件、一般講演10件で構成された。参加者は中国と韓国から各3名ずつ、日本からは150名近くの参加があり、大変盛況であった。

基調講演は、東京大学教授 辻井潤一先生(AAMT/Japio 特許翻訳研究会委員長)による"Fine-Grained Information Access and MT for Patents"であり、Semantic-based information accessの重要性と困難性、それを可能にするDeep parsingおよびその特許翻訳への応用について講演した。

招待講演1は、特許庁普及支援課特許情報企画室調査班長 岡崎輝雄氏による「特許庁における機械翻訳の活用状況とその将来像」である。特許庁における英語での情報発信と機械翻訳の利用について講演した。その中で特許電子図書館(IPDL)と高度産業財産ネットワーク(AIPN)におけるMTの役割を説明した。前者は一般ユーザーに開放されており、後者は各国特許庁の審査官などが利用している。

特許情報の世界的活用については、日本、韓国、中国、欧州、米国の5庁による取り組みがなされており、その中で英語と独・仏・西・伊・中・韓・日の各言語との間でのMTが利用されている。

招待講演2は、China Patent Information CenterのDirector of Information Processing Dept., Mr. Wan Dan氏による"CPIC's MT Development: Current Status and Future Directions"である。CPICでは中英・英中のMTシステムを開発済みでありSIPO(State Intellectual Property Office of the P.R.C.)からサービス中である。また、日中・韓中のMTシステムは開発中である。本講演で英語を共通言語とする特許情報の流通システムを提案している。この方法にはMTシステ

ムの数を減らせるという利点がある。しかし日韓・韓日のように英語を介さないほうが高精度の翻訳ができる場合もあるので慎重な検討が必要である。

昼食をはさんで招待講演3が行われた。韓国のSiriussoft CorporationのDirector, Global Division, Ms. Minah Kim氏による"Current Status of Korea's Machine Translation for Patent Domain Users"である。Siriussoft社は韓国特許庁(KIPO: Korean Industrial Property Office)の機械翻訳システムを開発している会社である。英韓・韓英および日韓・韓日のシステムを開発済みであり、KIPOは英韓・韓英のサービスを一般に提供しており、日韓に関しては内部で利用している。本システムの特徴は約30万文という多量の文パターンを用いている点である。特許文書は定型表現が多く文パターンの利用は有効である。

招待講演3の後、一般講演が2つのセッションに分かれて行われた。セッション1は、特許情報のさまざまな処理や利用に関するものであり、セッション2は、特許文の機械翻訳に関するものである。セッション1では、(1)特許記述言語PML (2)特許請求範囲の読解支援 (3)特許の質の自動評価 (4)知財訴訟分析 (5)パテントマップの自動生成、に関する発表があった。セッション2では、(6)Terminologyを援用した統計的機械翻訳 (7)規則方式翻訳と統計翻訳の融合 (8)多言語特許検索 (9)連語辞書のMTへの利用 (10)翻訳者からのMTへの提案、に関する発表があった。基調講演、招待講演、一般講演ともに熱のこもった発表で、議論も盛り上がり、本分野への関心の高さを感じるものであった。

会後は懇親会が行われ、海外からの6名の参加者とともに、本会議での議論の続きなど話に花が咲いた。本シンポジウムは、AAMT/Japio 特許翻訳研究会がMT Summit時に開催している"Workshop on Patent Translation"を補完するものであり、隔年で開催する予定である。本シンポジウムの論文集はJapioの以下のページから読むことができる。

日本語版

<http://www.japio.or.jp/kenkyu/kenkyu03-02.html>

英語版

<http://www.japio.or.jp/english/kenkyu/kenkyu03-02.html>

第 10 回日中自然言語処理共同研究促進会議に参加して

山梨英和大学

江原 暉将

2010 年 11 月 5 日(金)から 7 日(日)にかけて中国蘇州市にある蘇州大学で第 10 回日中自然言語処理共同研究促進会議(The 10th China-Japan Natural Language Processing Joint Promotion Conference)が開かれた。日本側の参加者は、井佐原均 AAMT 会長、NICT の林弘監事はじめ 19 名、中国側の参加者は、蘇州大学の姚建民教授、中国科学院の劉洋教授、マイクロソフトアジア研究所の黄昌寧教授、精華大学の孫茂松教授など 29 名であった。

会議は 3 件の基調講演、19 件の一般講演で構成された。初日は、蘇州大学の張学光教授、孫楽教授および井佐原会長の挨拶のあと、黄昌寧教授の基調講演 "Combinatory Categorical Grammar and Beyond" と孫茂松教授による基調講演 "The Progress of NLP Research at Tsinghua University"があった。前者は黄教授が長年研究してきた Combinatory Categorical Grammar に関するものであり、後者は精華大学での NLP 研究を概括するものであった。その後、昼食をはさんで 8 件の一般発表があった。その中で筆者も "Machine translation for patent documents combining rule-based translation and statistical post-editing"と題して、NTCIR-8 での特許文機械翻訳について発表した。また NTCIR-9 でも特許翻訳に関して日英、英日、中英の 3 つのタスクが予定されていることを紹介した。

二日目は、劉洋教授による基調講演 "Joint Tokenization, Parsing, and Translation"とデモ "Multilingual Machine Translation System"が行われた。これは、Tokenization や Parsing と Translation を結合して扱う手法の提案である。デモでは中国内の諸言語と中国語の間の機械翻訳が示された。基調講演のあと、一般講演が 11 件行われた。

三日目は excursion であった。蘇州は世界遺産の庭園も多くあり、当日は、虎丘公園と寒山寺を訪れた。虎丘公園の高台にそびえる虎丘塔は、南京から蘇州を目指した旅人の目標となった塔とのことである。寒山寺は、張継の有名な漢詩「楓橋夜泊」(写真右側の詩)の舞台となった寺で、これ以外にも多くの漢詩が回廊を飾っている。写真左側の詩は、同

じ張継の「宿楓橋」(楓橋に宿る)である。畏友木村展幸氏が訓読と翻訳をつけてくれたので原文と合わせて紹介したい。

七年不到楓橋寺
客枕依然半夜鐘
風月未須輕感慨
巴山此去尚千重

七年到らず楓橋の寺
客枕依然、半夜の鐘
風月未だ軽々しく感慨するを須みず
巴山、此を去る尚ほ千重

七年(ななとせ)ぶりの楓橋の
旅の枕を敬(そばだ)てて
昔の響き懐かしむ
風の伝うる寺の鐘
往きし月日のまぼろしか
身は遠国(をんごく)の客(かく)にして
巴山は遙か幾千里



この会議は 10 回を数え、その大きな成果の一つである「日中・中日言語処理技術の開発研究」も終盤を迎えている。今後、自然言語処理技術の実用化において日中間の連携を深める必要がある。世界経済の中で東アジアの地位が高まりつつある今日、中国語、日本語、韓国語の重要性がますます大きくなっている。その中で本会議が果たす役割も一層増してきている。次回は 2011 年 10 月に日本の松江で開催される予定である。

Patent Translation: Why Quality Matters 「特許翻訳：品質はなぜ重要か」

Organized by the World Intellectual Property Organization (WIPO)

株式会社知財翻訳研究所

品質管理・教育センター 竹内 仁恵

2010年12月10日に東京渋谷の国連大学にて開催されたWIPO主催のシンポジウム“Patent Translation: Why Quality Matters”「特許翻訳：なぜ品質は重要なのか」についてレポートする。当日は約300名収容の国際会議場が6～7割埋まった状況でシンポジウムがスタートした。統一テーマである「翻訳品質の重要性」について、5名（WIPOは2名）のスピーカーがそれぞれの立場から内容の濃い講演を行い、また参加者からも鋭い質問がなされ、あっという間の4時間であった。なお講演は全て日本語で行われた。以下に順を追って各スピーカーの講演内容の概要を述べ、翻訳品質の向上についての示唆などをまとめる。

最初は、米国特許エージェントであるJames Judge氏による“Terms of Art in Translation as a Means of Assuring Patent Quality”「翻訳における特許の質を確保する手段としての専門用語」と題する講演であった。大阪で10年以上米国特許出願に携わるJudge氏は、特許明細書の日英翻訳において「生きている」すなわち当業者がわかる、さらに言えば「審査官が正しく技術を理解できる」用語を使用する必要性を分かり易い例をあげて説明された。特許の価値は、最も近似した先行技術文献が審査官の調査によって引例として挙がってくることと密接な関係があり、そのような調査を可能にするのが“terms of art”であるとの説明は、そうした「専門用語」の実際の調べ方の順を追った説明とともに大変参考になるものであった。特許の資産価値を高めるために「専門用語」を適切に用いた翻訳が望まれるのは言うまでもなく、使い易さの点で多少難点

はあるものの、USPTOのClassification Definitionsを活用して審査官が実際に使用する言葉の定義を知るという作業は試してみる価値が充分あるように思われた。

続いて、欧州及びドイツ弁理士であるFriedman Horn氏による“The Importance of Translation Quality”「特許出願手続きにおける翻訳の質の確保」と題する講演が行われた。Horn氏も日欧の特許出願と翻訳に関して長い経験をお持ちで、実例を挙げての特許翻訳全体に関わる問題の解説は、現場で日々悩む参加者に対する示唆に富んだものであった。強い特許のためには、「一つ一つの言葉」ではなく「それらを繋いで作った文＝意味」を翻訳することが重要である、すなわち求められるのは「実際のまたは意図された技術的な意味」を考慮した翻訳、悪意に誤解されることで出願人に不利益になることがない翻訳であるというものである。特許翻訳の常識としてまことしやかに語られることが多いいくつかの定説を鮮やかに覆し、翻訳品質の向上とはどのような方向性で考えるべきなのか、多くのヒントが得られる講演であった。

休憩後、シンポジウムの主催者であるWIPOのPCT翻訳部を代表してJames Philips氏とSally Young氏が講演を行った。WIPOのPCT翻訳部門の活動の全体的な紹介の後、この日の議題である「なぜ品質は重要なのか」についてWIPOならではの立場が説明された。具体的には、PCT翻訳部門による翻訳は、出願明細書自体の翻訳ではなく要約の翻訳に代表される“translation for information”であるという点である。アジア言語課

の Young 氏から、英訳された要約はさらに仏語に翻訳されるので「明解」であることが何より重要であり、さらにそれ以降様々な段階で行われる調査の際に信頼できる英訳でなければならないこと、そのため、ひとつの発明に複数の出願がある場合はターミノロジーに一貫性を持たせることが重要であることなどの説明があった。現在 8 つの言語で行われる PCT 出願における要約の英訳品質の確保が、PCT 加盟国の出願人全体にとっていかに必要かつ有益であるかについて、参加者は理解を深めることが出来たと思われる。さらにこうした目的を念頭においた、品質の管理・確保のための現状での WIPO の取り組みが紹介され、将来的には「翻訳ガイドライン」をウェブ上で無料で公開したいとの説明や、サポートツールは使用しているが、機械翻訳は現在も今後も使用する予定はないなどの説明もあった。最後に 2011 年の日英翻訳サービスの入札に期待する具体的な内容が明らかにされ、参加者からも現実的な質問が出るなど、活発な質疑応答が行われた。

最後に、(株)リコー知的財産センター、日本知的財産協会(JIPPA)特許第1委員会委員長、荳林啓氏による「特許制度ハーモナイゼーションに対するユーザの期待と翻訳が果たす役割について」と題した出願人サイドを代表する講演が行われた。荳林氏からは、グローバル化に伴い、国内の特許出願のうち良いものだけを海外に出願するのではなく最初から例えば日米同時に申請するのが現在の傾向であること、その結果、国際的状況としては同じ発明について複数の出願が重複して存在することになり、手続きやコストの面で出願人の負担が増してきた経緯が説明された。さらにそうした状況に対して、審査結果の相互利用を始めとする「特許制度の国際調和」が今後ますます求められていくこと、将来的には、各国の審査で異なる引例が出されたり特許権の行使について裁判所の判断が国ごとに異なったりすることが解消される期待が高まっている

るとの説明もあった。そうした傾向に伴い、出願人にとって翻訳品質は「権利の品質」を左右する重大な問題であり、現在求められるのは「稼げる権利」ではなく「つぶれない権利」であること、それを実現できる翻訳こそが必要であるとの結論であった。時間が押していたためにやや駆け足での説明となってしまったのが残念ではあったが、翻訳のエンドユーザーとも言える出願人サイドを代表しての荳林氏の講演は、翻訳品質の影響を直接的に受ける立場での高い品質を期待する切実な思いが伝わる講演であった。

以上のように、特許翻訳の品質確保の重要性について、米国及び欧州のパテントエージェント、WIPO、出願人という異なる立場から講演がなされた。「特許翻訳：なぜ品質は重要なのか」という議題に対する答えは決して1つではないが、敢えてまとめると、「強い翻訳」が「強い権利」を特許権者に約束するからと言えるのではないだろうか。

「強い翻訳」とは適切な専門用語を一貫性を持って使用し、意図された内容を正確に訳出することで実現され、それが出願から許可後に至るまであらゆる段階での調査、交渉、権利行使において「強さ」を発揮するということになるのだろう。

AMTA 2010 報告 — 企業ユーザーの視点から

Dublin City University

立見 みどり

2010 年秋、米国コロラド州デンバーのダウントウ
ンの中心にある Westin Hotel Tabor Center で開催
された The Ninth Conference of the Association
for Machine Translation in the Americas (AMTA
2010) に参加しました。プレコンファレンス・ワ
ークショップ (10 月 31 日)、メインコンファレン
ス (11 月 1~3 日)、ポストコンファレンス・ワー
クショップ (11 月 4 日) と全日程 5 日間の充実し
た内容のコンファレンスでした。コロラド州の州都
であるデンバーは、19 世紀中旬のゴールドラッシ
ュをきっかけに生まれた町です。海拔 1,609 メートル
(およそ 1 マイル) であるため「マイル・ハイ・
シティ」とも呼ばれ、年間 300 日が晴天だというデ
ンバー。コンファレンス期間中は夏が続く、気持
ちよく過ごすことができました。

AMTA の新しい試み

今回の AMTA の新しい動きとして、翻訳者との連
携を重視した点が上げられます。米国には、主にフ
リーランスの翻訳者で構成され 1 万人以上のメン
バーを擁する ATA (American Translators
Association) という組織があり、毎年秋に米国の
いずれかの都市でコンファレンスが行われていま
すが、今回初めて AMTA と ATA の相互協力により、
同都市において連続する日程での開催が実現し、
AMTA の参加者は ATA のセッションへ、ATA の参
加者は AMTA のセッションへの参加を呼びかけら
れました。わたし自身は残念ながら時間の関係で
ATA のセッションには参加できませんでしたが、
AMTA で ATA のメンバーの方と一緒にセッション
に参加して意見交換をする機会などがあり、貴重な
経験となりました。機械翻訳と人間翻訳に携わる

人々どうしの理解を深め、協力関係を構築しようと
いう試みは、博士課程の研究テーマとして機械翻訳
出力の人手によるポストエディティングを選んだ
わたしには特に興味深い動向でした。

この流れのひとつとして、メインコンファレンス初
日の基調講演は、1 つは ATA のプレジデントであ
る Nicholas Hartmann 氏、もう 1 つはフリーラン
ス翻訳者でありまた翻訳関連書の著者およびロー
カリゼーションと翻訳のコンサルタントでもある
Jost Zetzsche 氏によって行われました。

Hartmann 氏の講演には、多くの翻訳対象文書が機
械翻訳では処理不能であることを例を示しながら
紹介するなど、むしろ機械翻訳と人間翻訳の間の溝
を示唆するような内容も含まれていましたが、結論
としては、プロの翻訳者が機械翻訳の利用に際して
その品質基準を示したり、ポストエディティングに
役立つツールの開発に貢献するなど、両者の協力で
きる余地があることを強調していました。Zetzsche
氏はさらに、翻訳者と機械翻訳開発者が互いの仕事
を理解し合い協力し合うことで、機械翻訳+ポスト
エディティングという作業手順を有効な翻訳手法
として確立し認知させることができるという展望を
語りました。

コンファレンスの全体像

3 日間のメインコンファレンスは、前年カナダのオ
タワで開催された MT Summit XII と同様、
Research、Government Users、Commercial Users
と 3 つのカテゴリーに分類され、別々の会場で常に
3 つのセッションが同時進行するという構成でし
た。わたしは IT 企業と大学の共同研究でローカリ

セッション現場のワークフローに着目した研究を行っていた都合上、主に企業ユーザー (Commercial Users) のセッションに参加することになりましたが、ここに全体の傾向を簡単に紹介します。

Research カテゴリーでは、学术界および産業界から、最新の機械翻訳研究結果が発表されました。内容は、統計ベースの翻訳システム研究が半数以上で、他には規則ベースや用例ベースの翻訳システムと組み合わせたハイブリッドシステムに関する研究、品質評価、品質予測などに関する研究などがありました。また言語を特定した研究では、アラビア語に関する研究が目立っていました。口頭発表のほかに、ポスターセッションが2日にわたって行われました。

Government User カテゴリーでは、主に米国の政府機関をクライアントとする翻訳組織や情報セキュリティ会社、米国政府の支援を受けた非営利団体などから、機械翻訳システムを活用するに当たってのワークフローの開発、改善、機械翻訳システムの性能評価、翻訳者が機械翻訳システムを利用した場合の翻訳結果の評価、対訳コーパスの構築と利用などに関する発表がありました。

Commercial User カテゴリーでは、翻訳・ローカライゼーションベンダー、多言語へのドキュメント翻訳を必要とするグローバル企業、情報管理サービス会社、コンテンツ・文書作成支援ソフト会社などによる、機械翻訳システムの導入やカスタマイズに関する各社のアプローチ、収益性、品質、問題点 (原文の品質、ポストエディティングコストなど)、人手翻訳と機械翻訳の統合、機械翻訳品質向上の試みなどが主なトピックでした。

Commercial User カテゴリーの中で印象に残った発表を以下にいくつか挙げさせていただきます：

One technology, many solutions: MT at Adobe

Flournoy, R. & Rueppel, J. (Adobe)

Adobe 社から、過去2年間におよぶ、機械翻訳+ポストエディティングというワークフローの確立への取り組みについて発表がありました。コスト削減を目指してまずは限られた言語の少数の製品マニュアルの翻訳に導入、成果が上がったため、対象製品数、言語数ともに拡大しているそうです。今後は、マニュアルだけでなく、サポート文書や顧客とのコミュニケーションへの適用も検討中とのことで、現場での機械翻訳の幅広い活用状況と可能性が伺える内容でした。

Using Machine Translation for the Localization of Electronic Support Content: Evaluating End-User Satisfaction

Stewart, O., Lubensky, D., Macdonald, S., Marcotte, J. (IBM)

IBM 社のカスタマーサポート文書の機械翻訳結果について、ユーザーがその文書を読んだことで問題を解決できたか (コールセンターに電話する必要性がなくなったか) をベースに評価するという試みについて発表されました。従来の人手による評価は、翻訳出力を読んでその品質を直接評価してもらうという手法が主流でしたが、ここでは一歩進んで、その文書の目的をどれだけ達成できたかを基に機械翻訳結果の有用性を判断するというプラクティカルなアプローチが提案されました。

Where can MT be most successful and what are the best MT engines for various languages?

Lu, J. (CA Technologies)

CA Technologies 社からは、英語から多言語への翻訳に際して同社がとったアプローチが紹介されました。グローバル企業が機械翻訳システムを導入する場合のひとつのアプローチは、本社が音頭を取り、すべてのターゲット言語に共通の機械翻訳シス

テムを導入することでカスタマイズや教育の手間を省くというものですが、同社では反対に、機械翻訳システムの選択から運用まですべてを各国のローカリゼーションチームに任せ、各国チームがそれぞれの言語に最適なシステムを最適な方法で運用することで最大限にコストを削減するというアプローチがとられています。その結果、生産性が大きく向上しコストも大幅に削減されただけでなく、各国の翻訳者自身が自ら機械翻訳推進のイニシアチブをとったことで機械翻訳に対して肯定的な意識を持つことができ、仕事のやりがい向上にも貢献したと報告されました。機械翻訳の多くの導入例で翻訳者が不満を感じる結果になっている現実を考えると、CA Technologies 社のアプローチは、翻訳者と機械翻訳システムとの間に良い関係が築かれている例として学ぶところがありました。

ポストエディティングへの関心度

前年の MT Summit XII に引き続き、ポストエディティングへの注目度も高いコンファレンスでした。企業が機械翻訳システムを利用するにあたり、品質確保のために人手による後編集を行うケースは数多くあります。「機械翻訳は完璧でないから使えない」から、「完璧ではない機械翻訳をどう補って有効利用するか」という視点にシフトしたことが今日の機械翻訳の幅広い利用につながったことを思えば、ポストエディティングへの関心度の高さは自然な流れとも言えるでしょう。わたし自身が企業における機械翻訳+ポストエディティングワークフローをテーマに研究していたことから、ポストエディティング関連のセッションに多く参加しました。ポストエディティングに言及したトピックの多くは、当然のことながら Commercial User のカテゴリーに属していました。例を挙げますと、

Practical uses of MT at Global Language Translations and Consulting

Strock, D. (GLTAK)

機械翻訳+ポストエディティングは新しい市場であり、人手翻訳に代わるものとして考えるべきではない。高品質の翻訳は人手で、早くて安い翻訳は機械翻訳+ポストエディティングで提供し、両者の違いを明確にすべきである。またポストエディタには、言語スキルよりも分野ごとの専門知識が求められる、という考えが示されました。

Evaluating vendors for MT and post-editing at Avaya

Scott, B. (Avaya) & Beaton, A. (SDL)

機械翻訳+ポストエディティングが望ましいソリューションとされるのはどのような状況か（時間とコストの制約がある、短命の文書であるなど）またどのようなコンテンツに適しているか（技術文書や製品マニュアルなど、機械翻訳である程度の品質が期待できるタイプの文書）が論じられた後、Avaya 社での導入例が紹介されました。

Effective MT within a Translation Workflow Panopticon

Andrä, S. (Ontram) & Shütz, J. (BioLoom Group)

ポストエディティングのさまざまなシナリオ、フィードバック情報としてのポストエディタによる機械翻訳結果の評価、望ましいポストエディタのプロファイルなどについて論じられました。

さらに Commercial User カテゴリーでは、口頭発表のほかにも「Perspectives on Post-Editing」と題したパネルディスカッションも行われ、各企業から、機械翻訳+ポストエディティングワークフロー導入の背景、目的、導入による生産性向上率、問題点などが発表され、機械翻訳導入を新たなビジネスチャンスとして活用するには何をすべきかという

観点でディスカッションが行われました。

ポストエディティングはプレコンファレンス・チュートリアルでも大きく取り上げられ、午前と午後に渡って次の2つのセッションが行われました。

Introduction to Post-Editing: Who? What? How? And Where to Next?

O'Brien, S. (DCU)

このチュートリアルでは、ポストエディティング研究歴の長い O'Brien 氏により、ポストエディティングとは何か?といった基本的情報でありながら定義として確立されていない問題に始まり、各種ポストエディティングガイドライン、これまでのさまざまな組織によるポストエディティング効率化への取り組み、統計翻訳技術を利用したポストエディティングの自動化、翻訳者の視点から見たポストエディティングという仕事について、また関連する学術研究の成果などが紹介されました。チュートリアル参加者は、今後ポストエディティングに携わる可能性のある翻訳者から、ポストエディティングを長く手がける翻訳ベンダーの方々まで多岐に渡り、質問や忌憚ない意見交換が活発に行われたセッションとなりました。

MT Postediting Showcase

Gerber, L. (MITRE/NVTC) & Macklovitch, E. (Onscope Group, Inc.)

ポストエディティング・ショーケースでは、会場に設置された5つのブースで、経験豊富なポストエディタがそれぞれの組織で普段使用している環境・ツールでのポストエディティングデモを行いました。参加者は自由に見学したり質問したりでき、実践的な知識や技能を共有する機会となりました。

わたし自身も、ポストコンファレンス・ワークショップ「Bringing Machine Translation to the User」

で、原文の特徴（文法の正しさ、文章の複雑さ、単語の曖昧さなど）に基づいたスコアと、ポストエディティングの労力（機械翻訳結果とポストエディティング結果の間のテキストの変更量およびポストエディティングにかかった時間）を比較した調査の発表を行い、原文の特徴に基づくスコアは、ポストエディティングの時間よりもテキスト変更量との相関関係のほうが高い傾向があることを示しました。

まとめ

総合的には、前年の MT Summit XII でも感じられたことですが、ポストエディティングへの注目度も含め、機械翻訳について、研究・開発者だけでなく、利用者、ドキュメント製作者の視点からも多く論じられたコンファレンスでした。また個人的には、5日間と長いコンファレンスの中で、さまざまな分野の人々と知り合い意見を交換できる貴重な体験ともなりました。

メインコンファレンス2日目の基調講演で、翻訳支援ソフトウェア開発・販売会社であり翻訳ベンダーでもある SDL の CEO、Mark Lancaster 氏は、今後数年間のうちに最大 50%の翻訳が機械翻訳を利用したものになるだろうと予測しました。機械翻訳システム利用はビジネスの現場で現実的なオプションとして定着した感がありますが、品質を確保した上でのコスト削減や翻訳時間の短縮など、まだまだ課題が多く残されています。今後もこのようなコンファレンスを通じて、機械翻訳システムの開発・研究者と現場での利用者の間で情報交換が行われ、強固な協力関係が築かれることを期待します。

なお、この報告で紹介した内容を含め、AMTA 2010での発表論文・資料はすべて <http://www.mt-archive.info/AMTA-2010-TOC.htm> からダウンロードできます。

LISA Open Standards Summit 参加報告

秋桜舎代表 山本ゆうじ

AAMT 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 リーダー <<http://www.aamt.info/japanese/utx/>>

会議の概要

2011 年 2 月 28 日から 3 月 1 日の 2 日間、ボストンで開催された LISA Open Standards Summit に AAMT のワーキング グループ 3 から参加しました。この記事では、会議の内容と成果について報告し、関連する事項についてお知らせします。

この会議を主催した LISA (Localization Industry Standards Association) は、翻訳メモリの交換形式 TMX、用語集の交換形式 TBXなどを、策定していた団体です。LISA は支払い不能に陥ってしまい、残念ながら現在は活動を停止しています。おそらく、標準化活動は、今後、別の団体によって引き継がれていくものと思われます。

この会議に参加した目的は、UTX と AAMT を知ってもらうことです。参加者の所属は、グローバル企業の翻訳部門、翻訳支援ツール開発企業、機械翻訳開発企業、翻訳者コミュニティー、研究機関などでした。参加者の方は、技術的な知識を豊富にお持ちであることが印象的でした。以前から名前は存じ上げていた著名な方も何人かいらしており、直接お話しすることができました。アメリカを拠点とする組織や企業が大半でしたが、ロシア、イスラエル、アイルランドなど、複数の国にまたがることも多いようでした。ただ、欧米以外からの参加は少なく、日本からの参加は私 1 人でした。これは、翻訳の標準への関わりに関する縮図でもあったと感じました。

今回の会議に関して、私の立場は、AAMT で UTX 策定をする翻訳ソフト ユーザー、というものでした。ただし、これに加えて、私は現役の翻訳者でもあり、また SDL の公認講師として翻訳支援ツールを利用する立場もあります。

会議の形態は、いくつかの全体への基調講演や報

告に加えて、少人数のグループに分かれて、特定のテーマについて密接な意見交換をする、というものでした。このような少人数での意見交換も重要ですが、参加メンバーに UTX の概要を知ってもらうために、主催者に相談して、最終日に全員が集まったときに時間を頂いて、UTX についてのプレゼンテーションを行いました。特に、AAMT と UTX についてと、日本の翻訳ソフトの活用の現状といった背景も紹介しました。また、LISA が策定した TBX、TMX などの標準規格と UTX との相対的な関係についても説明しました。UTX は、高機能な用語形式である TBX を置き換えるものではありません。最低限の知識さえあれば、だれにでも使いやすく作りやすいことを紹介しました。

ユーザー辞書形式/用語集 UTX

私は、AAMT 機械翻訳課題調査委員会、共有化・標準化ワーキング グループのリーダーとして、翻訳ソフトのユーザー辞書および用語集のオープンな標準規格である UTX の策定に関わってきました。UTX については、2010 年末の JTF (日本翻訳連盟) 翻訳祭と、2011 年はじめの JTF 翻訳環境研究会でもご紹介しています (JTF から DVD が発売されていますのでご覧になれます <<http://goo.gl/xuRce>>)。最新版の UTX 1.11 の仕様書は、2011 年 6 月に公開される予定です。この仕様書を含め、UTX の詳細は、<<http://www.aamt.info/japanese/utx/>>からご覧ください。

なお、この仕様は、以前「UTX-Simple」と呼ばれていましたが、2011 年 4 月以降、単に「UTX」と呼ぶことになりました。UTX-Simple 1.10 の次バージョンは、UTX 1.11 となります。UTX-Simple は、XML 版の UTX と対比して、テキスト版の

UTX-Simple という位置付けでした。しかし、XML 形式の用語集はすでに複数あり、またテキスト版の形式が必要にして十分で、むしろ実用的かつ存在意義が大きいことが分かったためです。

UTX のサンプルは、本ジャーナルにも抜粋が掲載されています。より詳細なサンプルは以下の URL からご覧いただけます <<http://goo.gl/AeXgG>>。これは、UTX 関連の用語集です。

UTX は、翻訳ソフトのユーザー辞書として使われますが、翻訳の用語集に基づいて作成されるか、用語集そのものとして機能します。UTX ワーキンググループでは、この用語集を少しずつ育てながら、この用語集に基づいて、UTX 関連の仕様書や資料を、英語から日本語、日本語から英語に翻訳しています。UTX はタブ区切り形式なので、Excel で簡単に編集でき、また Google ドキュメントの表計算などでリアルタイムに共有・編集することができます。複数の翻訳者が同時に同じ用語集にアクセスして、翻訳で使う用語を参照・追加することができます。

会議での収穫

今回の会議では、世界中の翻訳業界のさまざまな方と直接お会いして、非常に活発で有意義な意見交換をすることができました。特に「コンテナー」と呼ばれる、翻訳に必要な情報をまとめてやり取りする形式の標準規格についての議論が活発に行われました。これについては、開発力を持つ、翻訳支援環境の十分に整った大企業では有意義と思われます。ただし、日本の場合は、後述の日本語表記の問題や、体系的な翻訳支援環境についての理解不足など、それ以前に解決すべき問題が多いのではと感じました。

UTX チームとして大きな収穫のひとつは、ブリガム ヤング大学の Alan K. Melby 教授により、UTX などの**変換ツール**が開発されていたことを知ったことです。これは、Translation, Training/education, and Testing

<<http://www.ttt.org/tbxg/>>のサイトで公開されています。TBX-Glossary (*.tbx)、UTX-Simple (*.utx)、GlossML (*.gml)、OLIF (*.olf) の各用語形式を相互に変換するツールです。

実はこのツールはしばらく前に発表されていたのですが、なんらかの手違いで、Alan さんご本人とお会いするまで私も知りませんでした。会議の合間を縫って、「このツールに問題がないか確認してほしい」と言われたので、UTX データで動作を確認しました。最初はうまくいかなかったのですが、「チームに連絡してすぐに修正する」と言われました。結局、問題点はすぐに分かり、会議中に動作の確認ができました。こちらは「あとで、メールでやり取りして修正してもいいのでは」と思っていたのですが、解決を先延ばしにせず、その場で問題を解決してしまう Alan さんの行動力とエネルギーには舌を巻きました。これほど**実行力のある方たち**が集まっていれば、物事の進展も迅速だろうな、と感じました。

日本語の品質評価と標準化

会議の主題は、ファイル形式の標準化ではありませんが、そのファイルの内容である翻訳の品質、また日本語の品質についても、考えさせられました。これは、同じ用語を繰り返し使う UTX とは密接に関係しています。以前から、日本語への翻訳の品質には、注意が払われていないことがあると、しばしば感じていました。会議の休憩中に、アメリカの某パソコン・周辺機器メーカーの方とお話したことが印象的でした。私もふだんからよく使うメーカーですが、かねてから日本語マニュアルの翻訳品質には疑問がありました。日本語だとはっきりしない点があるので、英語版のマニュアルを読む必要があるほどです。その方に疑問をぶつけてみると、その製品群については「翻訳の品質管理はやっていない」と明言されたのです。品質が問題なのは、海外企業の日本語翻訳だけではなく、ある高名な日本人翻訳者の方の訳文を拝見すると、基本的な表記の統一すらされていませんでした。今に始まった問題で

はありませんが、ファイル形式の標準化に加えて、その中身である、**翻訳の品質、日本語の品質について、客観的評価と標準化が必要**ではないかと感じました。

英語は、表記の整理という面では、日本語のはるか先を行っています。翻訳資産の共有と再利用が進んでいるのも、この点がほぼ解決されているからです。一例として、翻訳専門誌 *MultiLingual* では以下のような表記規則を定めています。

<<http://www.multilingual.com/editorialStylesheet.php>>

なお、次号の *MultiLingual* では、UTX の仕様が公開される予定です。

訳文や原文の表記がばらついていると、チェックのコストが増大し、また翻訳メモリーの一致率が下がることにより、翻訳コストが目に見えて増大します。対訳用語集の表記に一貫性がないと、翻訳者やチェッカーが表記に迷うことはもちろんですが、読者も混乱します。

現在は、翻訳顧客が、自社でスタイルガイドを策定せずに、他社企業のスタイルガイドを「流用」するように、翻訳会社に要求することもあるようです。このような事態が明るみに出ると、最悪の場合は、流用された企業から、翻訳顧客企業に、賠償金の支払いが要求されることがあるかもしれません。そのようなことにならないためにも、標準スタイルガイドは緊急の課題といえます。

私は、2011 年の 4 月から、JTF スタイルガイド委員会の委員として日本語の翻訳表記の標準化をお手伝いしています。スタイルガイド（表記）の標準化は、翻訳業務を合理的に見直し、不要なコストと労力を減らすのに役立ちます。また、TMX での翻訳メモリー、そして UTX での用語集の共有および再利用にも大きく貢献します。

海外と日本の翻訳ソフト

今回の会議では、用語についての討議にも参加しましたが、翻訳に関する課題の数が多すぎて、時間が足りませんでした。翻訳ソフトについても、もっ

と深く意見交換したいところでした。海外での（翻訳メモリーではなく）翻訳ソフトとその利用状況については、私自身もまだ知らない点があります。会議のあとで、アメリカで市販されている翻訳ソフトの状況も確認しておこうと、BestBuy というパソコン ショップを覗いてみました。翻訳ソフトに限らず、パッケージ ソフトは棚にほとんどありませんでした。ある程度予想はしていましたが、パッケージ ソフトの数の少なさには改めて驚きました。店員によると、ソフトは、最近ほとんどオンラインで売られている、とのことでした。この状況は、パソコン ショップや家電量販店などで、箱入りのパッケージ ソフトが棚にずらりと並べられて販売されている日本と比較して、興味深く感じました。

ボストンから帰国したあとで、AAMT のワーキング グループで、海外の有名な翻訳ソフト SYSTRAN を購入して、試してみました。以前から、日本の翻訳ソフトは、日本語の特性に関するチューニングという面に加えて、使いやすさの面でも、海外の翻訳ソフトよりは上ではないか、と予測していました。SYSTRAN はまだ徹底的に試したわけではありませんが、その予測はある程度裏付けられました。

業務用翻訳ソフトには、各言語固有の特性を理解した上での調整と、各言語固有のオプションが必要です。SYSTRAN のオプションは非常にシンプルで設定項目がほとんどありませんが、（完全と言えないまでも）日本の翻訳ソフトでは、比較的、オプションが充実しています。要するに、**日本の翻訳ソフトは、相対的に実用性が高い**、ということです。日本の翻訳ソフトは正しく使えば十分に業務に使えるものなのに、実際に使われていないのは実に惜しいことだと、改めて実感したわけです。**現状で不足しているのは、技術ではなく、翻訳ソフトを使いこなす翻訳者の技能と、それを広める講習なのだ**、ということも再確認しました。

海外の翻訳ソフトについては日本で知られていない一方で、日本の翻訳ソフトも海外でほとんどま

まったく知られていません。今後、日本の翻訳ソフトとその活用法を海外にもしっかり紹介した上で、UTX について紹介できればさらに有意義であるはずです。

まずは用語集としての UTX

UTX の一義的な存在意義は、翻訳ソフトの互換辞書であることです。これは、翻訳環境で用語集として使われることと密接に関連しています。用語を集中管理するという慣行が根付いていない組織では、翻訳ソフト導入の準備は不十分です。この場合、いきなり翻訳ソフトを導入するのではなく、まず **UTX を用語集として使う**のが適切でしょう。UTX の仕様は、シンプルでありながら、用語集の最善慣行（ベスト プラクティス）が多数盛り込まれています。UTX 仕様に沿って用語集を作るだけで、翻訳の品質を向上する道が開けます。その用語集による用語管理に基づいて、次のステップとして、翻訳のさらなる効率化と自動化を進めることがはじめて可能になります。

翻訳ソフトとはどういうものか。これは翻訳ソフトを実際に使わなければ理解することは不可能です。翻訳ソフトに長年関心を持っている方でも、実際には試していない方がまだ多数います。企業・組織の翻訳部門では、ぜひ、ユーザー辞書を作り込んだ上で、体験版などで、**実際に翻訳ソフトを試すこと**をお勧めします。ただ、誤った方法でむしろ効率が下がるという失敗例も多いので、どの企業・組織にもお勧めできるわけでもありません。用語集が使えるオープンソースの翻訳ソフトがあるとよいのですが、今はまだないようです。

翻訳者の意見を取り入れる

翻訳支援環境の開発者と翻訳者の間に大きなギャップがあるのは、日本で常々感じていたことでした。互いの意図や考えがまるで食い違っているのです。LISA の会議で、そのギャップは英語圏にもあるということを感じました（翻訳者の立場の人が少なかったこともあります）。まず、開発者は、「プロでない翻訳利用者」と「プロの翻訳者」の需要の方

向性の違いを認識し、区別する必要があります。その上で、開発者はもっと**翻訳者の意見に広く耳を傾け、翻訳の現場と需要について知る**ことで、適切な翻訳支援環境を開発できるはずです。環境が整ってくると、その環境をどう活用するかも変わってきます。一方で、翻訳者は、先入観を捨て、食わず嫌いをせずに、機会があれば、さまざまな方法を試してみることで、翻訳環境を向上できます。

現在の技術で、翻訳の品質と効率を大きく改善することは十分に可能です。問題は、そもそも**実務翻訳の体系的な研究と講習が不足している**ことです。アイルランドなどでは、実践的な翻訳学の研究が進んでいます。LISA の会議でもその関連の方とお会いしましたが、実にうらやましいと思いました。翻訳そのものの理論的研究は日本の大学でもされているようですが、実務翻訳とはほとんど関係ありません。また、実務翻訳については、企業での翻訳技能講習はあるものの、内容がばらついていて偏りがあります。オープンな議論が行われているというより、各社が個別に手探りしているのが実情ではないでしょうか。翻訳の重要性が高い日本で、翻訳者の地位が低く、翻訳の研究がされていないのは皮肉です。

日本の大学で、実務翻訳科を設けて、組織的に行うのがよいかもしれません。特に外国語大学で体系的に取り組んでもよいのではないのでしょうか。そのような基盤がしっかりできていれば、標準化の恩恵を最大限に得られるはずです。

今後は、AAMT が主宰し、2011 年 9 月アモイで開催される MT Summit で、富士通研究所の大倉委員が UTX について発表する予定です。また、地震で延期されましたが、来年、東京で開催される TAUS の国際会議には、私が参加する予定です。UTX は、オープンな仕様であり、どなたにもご活用いただけます。実際の翻訳業務で、広くご活用いただければ幸いです。

UTX用語集作成クイック ガイド — 4ステップでかんたん！

1. 表計算シートを新規作成する

- ・ Googleドキュメント、Microsoft Excelなど表計算ツールが編集しやすいでしょう。Unicode対応のテキスト エディター（Windows標準のメモ帳、秀丸、EmEditorなど）も使用できます。文書を保存するときは、UTF-8テキスト形式で。

2. 基本的な用語集情報を入力する

- ・ `#UTX 1.11; < source language >/< target language>;
<date created>; <creator>; <license>; #src
tgtsrc:pos (term status)`
- ・ 原語、訳語、原語の品詞を入力します。各列はタブで区切ります。用語ステータスは省略可能です。

3. 用語を追加する

- ・ 最良の訳語を1つだけ選びます。自信がなければ用語ステータスとしてprovisional（暫定）を指定します。適切だと確信したらapproved（承認）にします。
- ・ 「システム」、「パソコン」のような一般的な言葉は追加しないでください。用語集として意味があるのは、特定分野の「専門用語」です。

4. 再利用と共有

- ・ 訳語の知識を再利用、あるいは共有することができます。

UTXの詳細は<http://www.aamt.info/japanese/utx/>まで

平成 21 年度音声翻訳実証実験の概要

(独) 情報通信研究機構

安田 圭志、隅田 英一郎

平成 21 年度、日本国内において、大規模な音声翻訳実証実験が実施された[1]。本実証実験は、自動音声翻訳技術の翻訳精度の飛躍的向上及び訪日観光分野における同技術を活用したサービスの早期実用化を図ることを目的し、総務省が「地域の観光振興に貢献する自動音声翻訳技術の実証実験」(総事業予算額 9.85 億円)を民間法人等に委託して実施されたものである。

本稿ではこの実証実験について取り上げ、概要について述べる。

実証実験実施地区

実証実験は、日、英、中、韓の 4 カ国語を対象とし、表 1 に示す通り全国 5 地方で(北海道、関東、中部、関西、九州)観光施設など約 370 箇所に約 1700 台の端末を設置して実施された。

システム概要

図 1 に本実証実験でも用いられたシステム

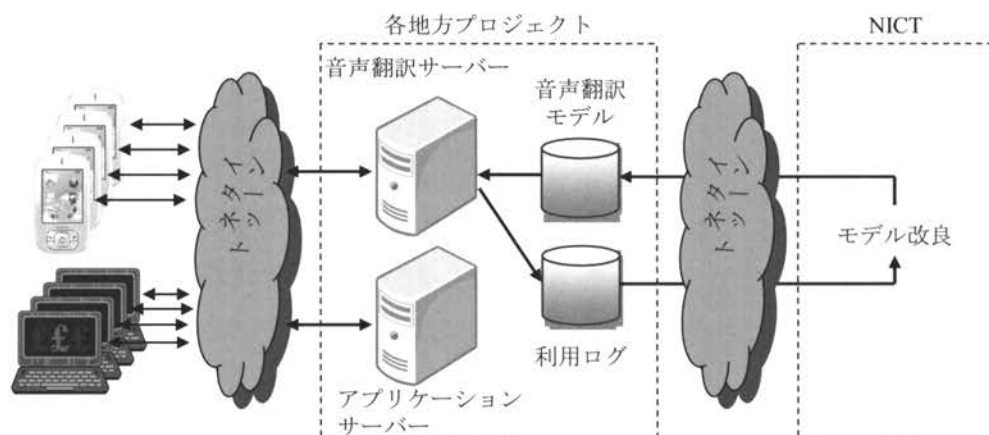


図 1 実証実験のシステム構成

の構成を示す。音声翻訳サーバは、音声認識、機械翻訳、音声合成用のサーバ群から構成される。スマートフォンやノート PC などの音声翻訳端末から入力された音声は、16kHz サンプリングの ADPCM 形式で音声翻訳サーバに送られる。

機械翻訳部については、BTEC コーパス[2]を用いて学習された統計的機械翻訳をベースにし、実証実験実施前に各地域ごとに収集した地域固有の表現や固有名詞を用いて適応化を行った翻訳エンジンを用いている。

翻訳結果は、テキストおよび合成音声の形で音声翻訳端末に送信される。サーバ側では、利用ログとして、入力音声、音声認識結果、翻訳結果、日時、端末 ID、言語指定等の情報を蓄積する。

アプリケーションサーバは、地方毎に設計されたナビゲーションや観光案内アプリケーション等のサービスを行うためのサーバである。

表 1 実証実験実施地方の詳細

実験実施地方	詳細実験地	事業者	実験期間	端末設置 施設数	端末数
北海道	札幌, 富良野, 知床	ドーコン [全体統括] NECソフトウェア北海道 [システム構築] NICT [音声翻訳中核技術]	12/28~2/22	92	403
関東	河口湖, 山中湖, 石和, 甲府	JTB-GMT [全体統括] NEC [システム構築] NICT [音声翻訳中核技術]	1/25~2/21	29	345
中部	金沢, 能登, 高山, 伊勢, 鳥羽	JTB中部 [全体統括, 実験運用] 駅探 [システム構築] NICT [音声翻訳中核技術]	1/5~2/22	120	310
関西	大阪, 奈良, 広島	ATR-Trek [全体統括, 実験運用, システム構築] NICT [音声翻訳中核技術]	12/14~2/28	106	343
九州	阿蘇, 博多, 天草, 長崎	熊本ソフトウェア [全体統括] 九州産交, JTB九州, 産学官ながさき [実験運用] NICT [音声翻訳中核技術]	12/20~2/28	26	313

アクセス数とアンケート結果

実証実験期間中, 5 地方全て含めると, 音声翻訳システムに対して約 20 万件のアクセスが記録された。

各地方プロジェクトでは, 一部の利用者に対して, 音声翻訳性能に関するアンケート調査を実施した。発話した内容が, 「半分程度以上伝わったあるいは理解できた」割合は, 日本語側で 5 割, 外国語側で 7 割程度であり, 今後更なるシステム性能改善の必要性が明らかとなった。

利用ログを用いたシステム改善

音声翻訳システムがフィールドで使われるようになると, 従来型の音声翻訳の要素技術の研究開発に加え, 利用ログを用いたシステム改善も重要な取り組みとなってくる。

本実証実験期間中においても, 利用ログを用いることによる, 音声翻訳性能の改善効果を調査した[2]。ここでは, 人手による書き起し, 対訳作成等を行い整備した数千文サイズの利用ログを, 音声翻訳システムの学習や適応化に用いることにより音声システムの性能が向上することが確認されている[1]。

むすび

本稿では, 平成 21 年度に実施された音声翻訳実証実験について扱った。実証実験やシステムの概要と, アンケート結果, 得られた利用ログを利用したシステム改善について述べた。

本実証実験で収集した利用ログについては, 研究利用への同意が得られている。人手による書き起し, 対訳作成等の利用ログの整備が非常に時間やコストがかかるという問題もあるため, NICTでは, 現在, 実証実験で得られたデータを用い, 人手による整備を要しない利用ログの利用方法に関する研究[3]を行っている。

参考文献

- [1] 河井恒他, H21 年度全国音声翻訳実証実験の概要, 日本音響学会 2010 年秋季研究発表会, pages 99-102.
- [2] Genichiro Kikui et al. 2006. Comparative study on corpora for speech translation. In IEEE Transactions on Audio, Speech and Language Processing, volume 14(5), pages 1674-1682.
- [3] 安田圭志他, 音声翻訳システム実利用データを用いた統計的機械翻訳のモデル適応, 言語処理学会第 17 回年次大会。

翻訳 SaaS「Eiplaza®/MT」のご紹介

東芝ソリューション株式会社

1. 情報活用 SaaS Eiplaza®

当社は、業務に必要な文書やテキスト情報を、効率良く管理・活用できる SaaS^{*1)}「Eiplaza®」（イーアイプラザ）のサービスを開始しました。

企業は、社会的責任を果たし、企業価値を向上させるため、ガバナンス、リスク、コンプライアンス対応が必須となっています。そのため、業務プロセスの明確化と成果物の記録などが求められています。この要求に対しては、業務で使われる情報を効率よく蓄積・管理し、全社で共有することができる情報活用システムが有効です。しかし、全社レベルでのシステム一括導入は大きな初期投資となるため、初期コストを抑え、段階的に導入したいという要望があがっています。

当社は、様々な情報活用のための機能をソリューションとして提供しておりますが、今回は SaaS の形態で提供開始しました。これにより、お客様は、ハードウェア資産などの初期コストを抑え、必要な規模で段階的に導入が可能となります。



図 1 : Eiplaza® の概念図

「Eiplaza®」は、当社が長年培ってきた日本語処理・自然言語処理技術、文字・画像認識技術、XML 技術などのコア技術を基にして、企業の情報活用で必須となるファイルサーバ、文書保管・検索、コミュニティ機能や、より高度な翻訳やテキスト分類、テキスト分析機能など、お客様の多様な要求にワンストップで応えられる情報活用機能をラインアップしています。この豊富なラインアップから、業務の流れの中で必要なものを自由に選択し、組み合わせご利用いただけます。これにより、コンプライアンスに対応した文書の生成・管理・活用などが容易に実現できます。

また、Eiplaza®は上記の情報活用 SaaS の他、規程・業務マニュアル管理（文書間の参照関係を自動抽出し、連鎖的な閲覧や漏れのない変更を支援）、契約書管理業務（内部統制に対応、契約の更新時期を自動で通知）等の個別業務を「業務 SaaS」として提供しています。

本稿では、「Eiplaza®」の内、翻訳 SaaS「Eiplaza®/MT」について紹介します。

2. 翻訳 SaaS「Eiplaza®/MT」のサービス

Eiplaza®/MT は、英日/日英、中日/日中翻訳を提供する翻訳 SaaS で、The 翻訳サーバと同一の翻訳エンジンを使用しています。また、表 1 にある全種類の専門用語辞書や翻訳メモリ、特許文書・製薬業界文書向けの翻訳知識を選択して使用することができます上、ユーザ辞書を適用できます。翻訳する文書にあわせた知識を使用できるため、精度の高い翻訳結果が得られます。

翻訳要求はブラウザ画面から行える（図 2）ほか、Microsoft® Internet Explorer® / Word / Excel / PowerPoint® の画面に組み込んだ翻訳

表 1 : Eiplaza®/MT で使用できる専門用語辞書(分野により使用できる翻訳方向が異なります)

インターネット	機械	バイオ・生物	数学・物理	トラベル
住所	自動車用語	化学	繊維・衣料	エンターテインメント
固有名詞(標準)	航空	運輸・物流	JIS用語	オークション
固有名詞(専門)	情報・通信	学術用語	日外/医学(関連)	キャンパスライフ
ビジネス	電気・電子	軍事・防衛	日外/医学(基本)	グルメ
ビジネスレター	医学・薬学(第一・第二)	建築・土木	日外/科学技術(基礎科学)	ショッピング
金融・法律	医学・薬学(規制用語)	原子力	日外/科学技術(工学系)	スポーツ
自然・社会科学	製薬・薬学	産業技術	日外/科学技術(理・医学系)	チャット
自然科学	医学(基礎・臨床)	生産・品質	特許	カルチャー

用ツールバーを用いて行えます。従って翻訳する文章をブラウザ画面にコピーすることなく、Web ページやWordの文書等をレイアウトを保ったまま翻訳できます*2)。

また、ユーザ辞書の構築支援のため、翻訳した文章中に出現した未知語をその未知語が出現した文と対にしたレポートを毎月一回ご提供いたします。

3. 翻訳 SaaS 「Eiplaza®/MT」の効果

Eiplaza®/MT により以下の効果が得られます。

- 1) 機械翻訳の特長である、均一な翻訳結果を短期間で得られる上、ユーザ辞書を使用できるため、企業内で定められた表現で訳語を統一可能である。
- 2) インターネットに接続できれば作業場所に関わらず、同一の翻訳環境で翻訳できる。クライアントマシンとサーバ間は SSL で暗号化されているためセキュリティも確保できる。
- 3) ブラウザに組み込まれたツールバーで翻訳できるため、Web アプリケーションなどの画面に手を加えずに翻訳でき、同一アプリケーションを日中英の3ヶ国語で使うことができる*2)。
- 4) SaaS なので翻訳ソフトや翻訳ソフト用の機器の運用コスト、初期導入コストが軽減できる。

4. 今後の Eiplaza®

Eiplaza®/MT では、2011 年には中国語の特許や中国からの旅行客向けの辞書強化を行う他、翻訳精度・機能強化を計画しています。さらに、今後提供予定の Eiplaza®のサービス「日本語文書診断(理解しにくい表現をチェック)」「文書分類・分析(大量のテキスト情報から、自動的に分類カテゴリを作



図 2 : Eiplaza®/MT の画面

成し分類)」と翻訳を組み合わせることにより、翻訳前に原文をチェックし翻訳精度を向上することや、多言語の文書を横断的に分類することなどができます。このように、当社は今後も情報価値を高めるサービスの提供を続けてまいります。

Eiplaza ホームページ :

<http://eiplaza.toshiba-sol.co.jp>

- * 1) SaaS (Software as a Service) : ソフトウェアの必要な機能を必要な分だけサービスとして利用する形態
- * 2) 翻訳できない部分、レイアウトが保たれない部分、翻訳後動作しないアプリケーションもあります。

- ・ Eiplaza、The 翻訳サーバ、KnowledgeMeister、KnowledgeMeister Succeed、TX1 は東芝ソリューション株式会社の登録商標または商標です。
- ・ Microsoft、Internet Explorer、PowerPoint は Micorsoft Corporation の米国およびその他の国の商標または登録商標です。

「オフィス翻訳ピカイチ」2011年6月17日（金）新発売

株式会社クロスランゲージ

グローバル化で海外企業との仕事や人間関係の規模範囲が著しく世界へ広がる現代、オフィス文書の翻訳は需要が高くなってきています。また、海外の取引先や友人とのやり取りに必要なコミュニケーションツールであるEメールの翻訳も同様です。このたびは、このようなニーズに応え、オフィス文書、Eメールをワンタッチで自動翻訳する、低価格のコンシューマー向け日英・英日翻訳ソフトを発売します。

クロスランゲージが開発・提供する翻訳システムは、日本を代表する大手企業、政府機関、学術研究機関をはじめ、日本国内 No.1 のポータルサイトにも採用され、日本で利用率 No.1 の翻訳システムです。同価格帯製品では最大級となる 303.0 万語の翻訳辞書を搭載しています。

【本製品の主な特徴】

■オフィス文書を翻訳

Word/Excel/PowerPoint に翻訳機能を追加する、アドイン機能を装備しています。オフィスソフトで開いた文書を、各ソフトに追加されるアドインボタンを押すだけで、レイアウトを保って英語→日本語、日本語→英語の翻訳を実行します。

＜文字装飾も反映する優れたレイアウト保持能力＞
オフィス翻訳機能は、単に原文をレイアウトどおりに翻訳するだけではなく、文中の文字装飾(斜体や下線など)や文字色の単語単位での再現が可能です。また、PowerPoint 翻訳では、翻訳時の日本語と英語の文の長さを判定してフォントサイズを自動調整。テキストをテキスト枠にしっかりと収め、修正の手間を省きます。

■メールを翻訳

Outlook Express, Windows メール, Windows Live メール対応

Windows XP の標準メールソフト「Outlook Express 6」と、Vista 標準の「Windows メール」、また「Windows Live メール "2009"」に翻訳機能を追加します。受信したメール、作成したメールを、メールウィンドウを開いた状態でそのまま翻訳します。原文/訳文を左右に振り分けて読みやすく表示する機能もサポートしています。

■ワンポイント翻訳機能

さまざまなアプリケーション内で、文章や単語を範囲指定して、設定したホットキー(Ctrl+T などキーの割り付け)を押すと、翻訳結果をウィンドウに表示します。

■クラス最大級の総見出し語数 303 万語！

同価格帯（標準価格 5,000 円前後）の翻訳ソフトとしてはクラス最大級となる 303 万語（英日 118.0 万語／日英 185.0 万語）の辞書を搭載。

【動作環境】

対応OS	Windows 7 (32bit/64bit) / Windows Vista SP2(32bit) / Windows XP SP3(32bit) ※各日本語版OS ※インストールには管理者権限が必要です。
必要メモリ	Windows 7/Vista:512MB以上 Windows XP:256MB以上（推奨 512MB以上）
必要ハードディスク容量	500MB以上
メール翻訳	Outlook Express 6.0/Windows Live メール 2009/ Windowsメール/ ※Windows Live メールでは、受信したメールの翻訳を参照するだけの機能に限定されます。翻訳結果の保存、印刷、メール送信／転送は実行できません。(Outlook Express、Windows メールでは可)
オフィスアドイン対応ソフトウェア	MS-Word , Excel , PowerPoint 2000/2002/2003/2007/ 2010 ※64bit版 Office2010は未対応

翻訳サービス紹介

リアルタイムに多言語情報の発信を可能にする、ウェブサイト自動翻訳サービス

「My サイト翻訳」

株式会社 高電社

■「My サイト翻訳」とは

株式会社 高電社（本社：大阪市阿倍野区、代表取締役 岩城 陽子）は、既存の日本語サイトに翻訳ボタンを貼付するだけで、リアルタイムにウェブサイトの自動翻訳が実現できるASP型サービス「My サイト翻訳」を開発・提供しています。

当サービスは、お客様側に特別な作業を要せず、最短5営業日で既存の日本語サイトと同様のレイアウトで多言語サイトが実現可能となり、既に数百の企業、自治体、観光サイト、教育機関などでご利用いただいております。対応言語は、英語、中国語（簡体字・繁体字）、韓国語、欧州5言語（フランス語、イタリア語、ドイツ語、ポルトガル語、スペイン語）です。

■東北地方太平洋沖地震に対する支援

このたびの東北地方太平洋沖地震におきまして多く

の尊い命が奪われ、未曾有の被害が発生いたしましたことは誠に痛ましく遺憾の思いでございます。

犠牲者並びにご家族、関係者の皆様に謹んでご冥福をお祈り申し上げますとともに、被災者の皆様へ心よりお見舞い申し上げます。当社の技術・サービスを活かし、被災された皆様や日本の復興のため何かお役に立てることがないかと検討いたしました結果、特に被害の大きい東北地方4県（青森県・岩手県・宮城県・福島県）の自治体、観光協会へ、当社のウェブサイト自動翻訳サービス「My サイト翻訳」を2012年3月末日まで無償でご提供させて頂くことといたしました。この取り組みは、各方面より多くの反響をいただき、2011年4月末時点で新規を含む5か所の自治体・観光協会よりお申し込みを頂戴しております。

公開順	自治体名	都道府県名	導入開始日
1	倉敷市	岡山県	2006年11月1日
2	越谷市	埼玉県	2008年4月1日
3	佐世保市	長崎県	2008年4月1日
4	綾瀬市	神奈川県	2008年7月1日
5	喜多方市	福島県	2008年10月1日
6	東海村役場	茨城県	2008年12月1日
7	所沢市	埼玉県	2008年12月1日
8	蕨市	埼玉県	2009年1月30日
9	日立市	茨城県	2009年2月1日
10	豊島区	東京都	2009年2月12日
11	鹿嶋市	茨城県	2009年3月2日
12	岸和田市	大阪府	2009年3月3日
13	大田区	東京都	2009年3月31日
14	北区	東京都	2009年4月1日
15	浦安市	千葉県	2009年4月27日
16	南砺市	富山県	2009年5月1日
17	一宮市	愛知県	2009年6月1日
18	市川市	千葉県	2009年7月1日
19	春日部市	埼玉県	2009年7月1日
20	足立区	東京都	2009年7月1日
21	小金井市	東京都	2009年8月1日
22	江東区	東京都	2009年8月3日
23	豊米市	宮城県	2009年9月1日
24	板橋区	東京都	2009年9月1日
25	つくば市	茨城県	2009年9月30日
26	寝屋川市	大阪府	2009年10月29日
27	蒲郡市	愛知県	2009年11月1日
28	君津市	東京都	2009年12月1日
29	宇都宮市	栃木県	2010年2月1日
30	芦屋市	兵庫県	2010年2月1日
31	岐阜市	岐阜県	2010年3月1日
32	大田区子育て支援課	東京都	2010年4月1日

公開順	自治体名	都道府県名	導入開始日
1	喜多方市観光協会	福島県	2008年10月1日
2	北名古屋国際交流協会	愛知県	2010年4月1日
3	佐賀県CSO推進機構	佐賀県	2011年4月中公開予定

※印 政令指定都市

公開順	自治体名	都道府県名	導入開始日
33	清瀬市	東京都	2010年4月1日
34	荒川区	東京都	2010年4月1日
35	東広島市	広島県	2010年4月1日
36	観音寺市	香川県	2010年4月1日
37	さいたま市 ※	埼玉県	2010年4月15日
38	浜松市 ※	静岡県	2010年4月19日
39	伊東市	静岡県	2010年5月12日
40	墨田区	東京都	2010年7月1日
41	福生市	東京都	2010年7月23日
42	石岡市	茨城県	2010年8月1日
43	茨木市	大阪府	2010年8月1日
44	三原市	広島県	2010年8月9日
45	帯広市	北海道	2010年9月1日
46	宮若市	福岡県	2010年9月1日
47	武蔵村山市	東京都	2010年9月30日
48	千代田区	東京都	2010年11月1日
49	半田市	愛知県	2010年12月1日
50	昭島市	東京都	2010年12月1日
51	熊谷市	埼玉県	2010年12月1日
52	台東区	東京都	2010年12月22日
53	文京区	東京都	2011年1月1日
54	岡山市 ※	岡山県	2011年2月2日
55	東村山市	東京都	2011年2月15日
56	狭山市	埼玉県	2011年3月1日
57	町田市	東京都	2011年3月1日
58	米子市	鳥取県	2011年3月24日
59	岡崎市	愛知県	2011年4月1日
60	四街道市	千葉県	2011年4月1日
61	砺波市	富山県	2011年4月1日
62	四日市	三重県	2011年4月15日
63	本宮市	福島県	2011年5月1日公開予定
64	十和田市	青森県	2011年5月1日公開予定

公開順	自治体名	都道府県名	導入開始日
1	葛飾区観光ポータルサイト	東京都	2010年10月1日
2	熊本城 桜の馬場 城彩苑	熊本県	2010年10月5日
3	福岡天神サイト	福岡県	2010年11月1日
4	やまがたロケーションガイド	山形県	2011年3月1日
5	べっぴん八瀬の絆	大分県	2011年3月1日

■最新情報をリアルタイムに翻訳

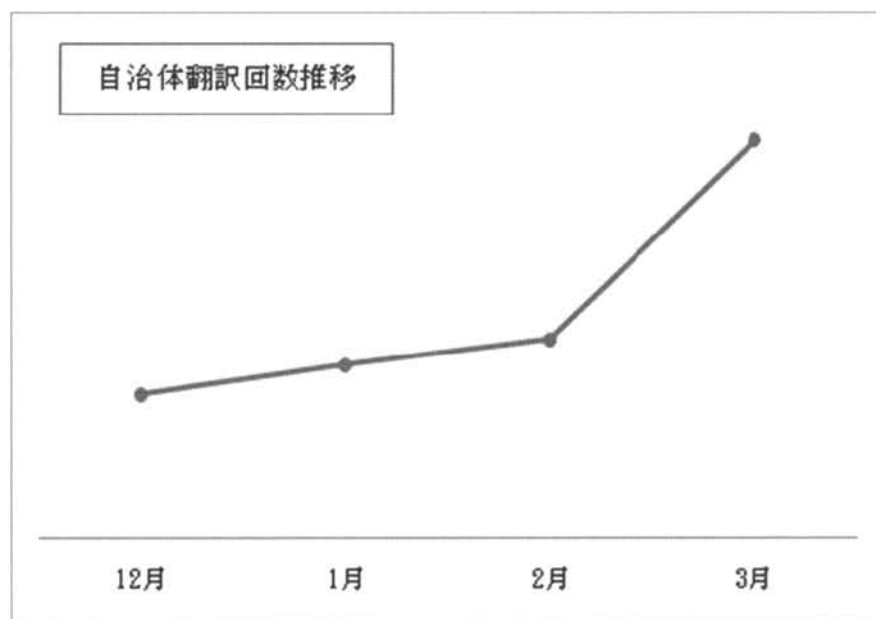
被災地には多数の在日外国人及び訪日観光客の方がおられ、日本人の方同様にリアルタイムに正確な情報を求めておられます。

「My サイト翻訳」は、サイト訪問者自身が翻訳開始のリンクボタンを押下し、また、その後のサイト内遷移の度に翻訳処理を行う為、最新情報をリアルタイムに翻訳することが可能です。これは緊急時の情報基盤として非常に適していると言えます。

また、PC と併せてモバイルサイトにも対応してい

るため、常にお手元で正しい情報を確認することができます。ライフラインが滞る状況下でのモバイルの役割は大きく、実際、震災発生後モバイルサイトでの自動翻訳の導入をご検討されている自治体も多く、多数のお問合せを頂戴いたしました。

参考情報として、震災以降、全自治体の翻訳回数の合計がこれまでより 2 倍以上増加し、被災地の自治体等では、通常の 20 倍以上となったところもございます。



“外国語による情報提供”という役割を当社の機械翻訳技術にて担わせていただきたく、ご活用いただければ幸いです。

「My サイト翻訳」は、お客様独自の専門用語や固有名詞をご登録いただく「オンラインユーザー辞書機能」を備えており、この機能により、更に高精度な翻訳結果が得られます。またこの機能は、単語・文章単位だけでなく、HTML ソースでの登録が可能のため画像や Flash 等の箇所を事前に作成しておいた外国語版の画

像や Flash 等に差し替えて表示することが出来、より完成度の高い多言語サイトが実現します。

当社では、この「オンラインユーザー辞書機能」の精度・利便性の追及をはじめ、多言語化されたウェブサイトが、海外検索サイトにヒットしやすくなる、SEO (検索エンジン最適化：サーチ・エンジン・オプティマイゼーション) 対策の組み合わせなどにより、更なる機械翻訳の可能性や多言語サイトの意義を検証・提案し続けております。

■翻訳エンジンの改良と当社の今後の取り組み

当社の翻訳エンジンは、数十年に渡る研鑽と改良が重ねられ、これまで翻訳方法の主流とされていた文法中心のルールベース翻訳に加え、現在は、文例翻訳や統計的手法を取り入れたハイブリッド翻訳へと進化して

います。

機械翻訳の利用用途も幅広くなってきており、特に昨年頃から、経済成長を続ける中国市場を視野とした EC サイトに機械翻訳が利用されるケースが増えて参

りました。

人力翻訳に対する機械翻訳の優位性は、大量のデータをリアルタイムに安価で翻訳できること、が挙げられます。すなわち、商品点数が多く、且つ情報が刻々と更新される EC サイトの多言語化には、機械翻訳は大変適していると言えます。

EC サイトへのサービス提供を機に、新たな機能も開発いたしました。例えば、中日翻訳においては、新機能「タイトル翻訳」の開発により、タイトルと本文

の訳し分けが可能となり、飛躍的な翻訳精度の向上が認められています。

具体例を申し上げますと、商品タイトル『金扣夾克』を通常モードで翻訳すると、『金はジャケットを掛ける』となりますが、これをタイトル翻訳モードにて翻訳すると、『金ボタンジャケット』というように、本来の正しい表現へと変わります。また、タイトルと本文が混在する場合は、「自動識別モード」を選ぶことにより、自動で訳し分けを行います。

原文	通常モードで翻訳	タイトル翻訳モードで翻訳
时尚中腰休闲直筒中裤 饰拉链裤子	流行するレギュラー丈のレジャーなストレートスタイルのクロップトパンツファスナーのズボンに扮する	流行レギュラー丈カジュアルストレートスタイルハーフパンツ 装飾ファスナーズボン
春装2010新款/ 贴皮装饰磨边皱褶牛仔短裤	春服の2010新作/ 皮を貼ってエッジングのしわのデニムのショート・パンツを飾る	春服2010新作/ 皮装飾エッジングしわデニムショートパンツ
2010 新款 秋装 双层荷叶边 女装 长袖衬衫 格子立领金扣衬衫	2010 新作 秋ファッション 2層のフリル レディースファッション 長袖のワイシャツ チェックのスタンドカラーの金扣はワイシャツを掛ける	2010 新作 秋ファッション 2層フリル レディースファッション 長袖ワイシャツ チェックスタンドカラー金ボタンワイシャツ

このように、当社は常に「お客様が買いたくなるサイト」を視点に既存技術の改良に努め、クライアントの真の目的が達成できるようサポートしております。

高電社は今年、創立 33 年目を迎えました。「言語の

壁を超え、世界中の人々の心と心をつなぐ」を理念に、これからも常に進化し続ける翻訳エンジンとサービスの開発・提供を目指し、真摯に取り組んで参ります。

言語グリッドを用いたスマート翻訳 — 京大翻訳！ —

京都大学社会情報学専攻¹, (独) 情報通信研究機構²稲葉利江子¹, 村上陽平², 田中正弘², 林冬恵², 石田亨^{1,2}

1. はじめに

京都大学社会情報学専攻では多言語による異文化コラボレーションを目標に、(独) 情報通信研究機構 (NICT) により技術開発されてきたインターネット上の多言語サービス基盤「言語グリッド」を運営してきており、現在、18ヶ国 138 組織が参加し言語資源の共有・利用を行っています。

「言語グリッド」は辞書や機械翻訳など多種多様な言語資源を Web サービスの形態で蓄積し共有することができるため、自由に組み合わせて使うことができ、各々のコミュニティが作った言語資源を追加し、国際交流や多文化共生活動のための言語サービスも容易に作り出すことができます。

また、京都大学と NICT は、言語グリッドの連邦制運営モデルを構築し、複数の言語グリッドを接続可能とする言語グリッド連携基盤を開発しました。タイ国立電子・コンピュータ技術研究所 (NECTEC) において、言語グリッドの運営が開始されたことにより、2011 年 1 月から言語グリッドの連邦制運営も開始しました。これにより、京都大学が運営する言語グリッドと NECTEC が運営する言語グリッドの双方のユーザは、それぞれが共有されている多言語サービスを相互に利用することができますようになります。これにより、利用できるアジア言語も拡がり、国内の多文化共生活動や国際交流活動の幅も広がっています。

今回は、これまでの「言語グリッド」における研究結果を基盤として、不特定多数のユーザを対象にし、特定のコミュニティ内で利用することを想定したスマート翻訳について紹介します。その後、スマート翻訳を適用したご当地翻訳の第一弾である京大翻訳！について紹介します。

2. コミュニティに特化した翻訳

機械翻訳を用いた多言語コミュニケーションツールの需要が高くなっている一方、多くのポータルサービスにおいても機械翻訳サービスが利用できるようになってきています。しかし、一般的なコミュニケーションに利用できる翻訳精度があるかという疑問です。将来的には、Web 上の膨大なテキストデータを解析し、対訳コーパスを準備することで機械翻訳の精度は向上していくと考えられますが、現状の新聞記事などを基にした統計翻訳などでは、多言語コミュニティにおけるコミュニケーションに適用できるかは未知数です。

その理由としてあげられるのが、コミュニティに特化した辞書の必要性和、多義性の問題です。

1) コミュニティに特化した対訳辞書が必要

コミュニティの中では、独自のコミュニケーションの中で、日々新たな言語が生まれ育っています。コミュニティ独自の特殊な用語が誕生する場合や、一般的に利用されている用語にコミュニティ独自の意味が付けられ使われている場合もあります。これらのコミュニティ用語は、コミュニティ内でのコミュニケーションおよびコラボレーションには無くてはならない用語です。

2) 単語の多義性

機械翻訳を連携し利用することを考えた場合、複数のサービスの訳語選択が一貫しないことにより、翻訳途中で言葉の意味が変わってしまうという問題があります。その解決方法として、翻訳対象の文章や文章を含むテキスト全体から文脈を検出し、機械翻訳サービス間で伝搬していくことにより連携を実現するという研究もなされています。

また、利用範囲をコミュニティ内に限定し、コ

コミュニティ独自の辞書を作成し、機械翻訳と連携することにより、上記の問題点を解決するという方法もあります。そこで、言語グリッドを用いた取組みとして、特定のコミュニティに特化したスマート翻訳の実現を目指しました。

3. スマート翻訳

特定のコミュニティに特化したスマート翻訳は、NICTが開発し、オープンソースで提供されている言語グリッド Toolbox（以下、Toolbox）をカスタマイズし、構築しました。

3.1 言語グリッド Toolbox

Toolbox は、オープンソースの CMS である XOOPS を拡張して、言語グリッド上のサービスを利用可能とした Toolbox フレームワークとその上に登録された各種の多言語コミュニケーション支援モジュールからなります。これらは全て PHP で開発されています。多言語掲示板機能、翻訳支援機能、多言語 Web 作成機能、言語資源作成機能などの各基本機能は、この多言語コミュニケーション機能モジュールとして実装されています。

各モジュールと Toolbox フレームワークは、新

規機能の開発や、既存機能の拡張のための API を提供しています。API はいずれも PHP の関数として実装されていますので、他の多言語コミュニケーション支援モジュールから利用することができます。基本機能の API を用いることで、辞書検索インタフェースを加えた翻訳支援機能などの複数の機能を組み合わせた新規機能を開発することもできます。また、Toolbox フレームワークの API は、言語グリッドで Web サービスとして提供される言語サービスを、通常の PHP の関数呼び出しとして実行することもできます。このため、Web サービスの知識を持たない開発者でも、言語サービスを容易に利用することができます。新規機能は、その他の機能と独立した多言語コミュニケーション支援モジュールとして開発ができ、Toolbox フレームワークとは別に配布することができます。この仕組みを利用して、今回、必要な機能を抽出し、新たな機能を組み込み、ご当地翻訳の基盤となるスマート翻訳モジュールを開発しました。

3.2 スマート翻訳モジュール

スマート翻訳モジュールとは、言語グリッドを用いて不特定多数のユーザ間での異文化コラボ

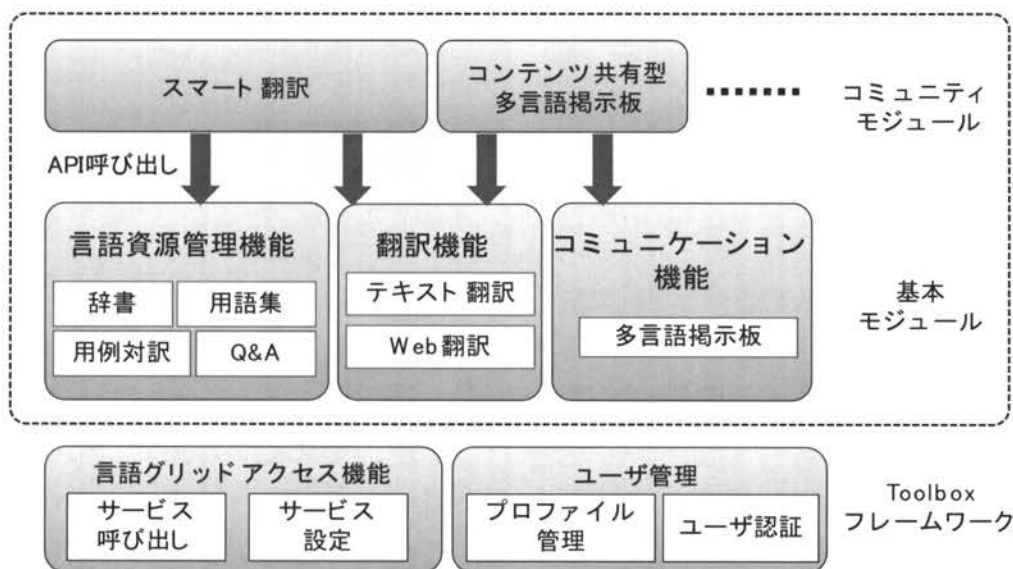


Fig.1 スマート翻訳モジュール

レーションの支援を目的とし、開発したモジュールです。不特定多数のユーザが訪れ、利用できることを想定しているため、コミュニティ用のオープンな Web サイトに、Toolbox の翻訳機能と翻訳作業を支援するための用例対訳・辞書の検索機能を外付けするための機能を配置しました。また、そのコミュニティに特化した翻訳を行うため、Toolbox が従来から備える辞書連携翻訳機能に加え、Toolbox 上に登録された用例を学習させた用例ベース翻訳を、その他の翻訳サービスと併用する機能も利用できるようにしました。用例ベース翻訳では、京都大学情報学研究科で研究開発がなされた KyotoEBMT を利用しています。

次章では、スマート翻訳の適用第 1 弾として今回取り組んだ「京大翻訳！」を例に、各機能の説明をします。

4. 京大翻訳！

京都大学では、2008 年に文部科学省の国際化拠点整備事業（グローバル 30）に採択され、学部・大学院において 11 の国際コースが 2010 年 10 月より順次スタートしています。グローバル 30 は、高等教育の国際協力の強化および留学生などに高度な水準の教育を提供するとともに、留学生と切磋琢磨する環境の中で国際的に活躍できる高度な人材の育成を図ることを目的としています。学内では、その準備の一環として、文書の多言語化活動も進められていますが、物理的な整備の他、日本人学生と日本語を母語としない学生および教職員のコミュニケーションなどを円滑に行いサポートするコミュニケーション基盤も重要となってきています。

一方、情報学研究科の学生が中心となり、多言語での情報共有を目的とした多言語コミュニティサイトの運営がなされ、留学生と日本人学生のコミュニケーションの場も提供され始めています。ただし、このコミュニティサイトは、研究科提供とされているためそのセキュリティポリシ



Fig.2 京大翻訳

ーより、情報学研究科ドメインのメールアドレス所持者のみがアクセスできるという制限が設けられています。

今回は、コミュニティサイトのコンセプトとは異なり、不特定多数のユーザが京大に関する文書の翻訳ができることを目的に『京大翻訳！』プロジェクトを立ち上げました。そのため、本サービスの提供も、不特定多数のユーザが利用できる環境を想定し、京都大学生協同組合（以下、京大生協）の公式サイトにて公開することとしました。

4.1 京大辞書

京都大学では、国際交流推進機能が中心となり、多言語文章を提供しています。留学生や外国人研究者向けハンドブックなどが日本語・英語・中国語（繁体字・簡体字）・韓国語にて提供されはじめています。さらに、京都大学でのグローバル 30 プロジェクトである K.U.PROFILE (Kyoto University Programs for Future international Leaders)が、京都大学規程類、入学手続き、学内支援、教務関連などの文書の英語化を進めており、日英文書を学内限定で公開しています。一方で、京大生協も留学生向けのハンドブックを英語、中国語、韓国語で公開し、留学生の生活支援を行っています。

我々は、まず、これらの多言語文書から京都大

学に特化した京大辞書の作成を行いました。辞書の言語は、アジアからの留学生が多いことを考慮し、日本語、中国語（繁体字・簡体字）、韓国語、英語を基本としました。次に、京大関係の不特定多数のユーザの利用を想定した以下のシナリオに基づき、京大辞書の増強を図りました。

1) 学内案内文書の翻訳

教務から様々なプロジェクトが主催する特別講演やイベントなどの情報が日常的に送付されてきますが、ほとんどが日本語のみの通知となっています。日本語が不得手な留学生はその案内の内容を理解することができません。そこで、それらの通知文書を翻訳し、内容が理解できることを目的とすることを想定しました。この場合、日時、講演タイトル、内容、場所などの情報が基本となります。そこで、主要な学内の建物、講義室などの辞書作成を進めました。また、建物や講義室などは、略して利用されることも多いため、それらの略語も網羅するような配慮が必要となりました。

2) 講義関連などの学生間での情報交換

学生にとっては、講義情報は重要となります。そのため、講義関連の翻訳が可能とすることを想定しました。京大では、学部生に提供されている全学共通科目及び各部局（研究科等）が提供する科目等、全学での一括管理がなされていません。そこで、教育推進部が管理する全学共通科目 5000 程度を対象に、日英での科目を提供いただき利用しました。科目名に関しては、中国語・韓国語にすることで科目名のニュアンスが異なる可能性を考慮し、中国語・韓国語でも英語表記がなされるようにしました。また、人名の翻訳が機械翻訳では困難であることは自明のため、京大教員名簿（日本語・英語）を作成し、登録しました。

3) 外国人訪問者向け案内の翻訳

大学という機関柄、外国人の訪問者も多く、大学および研究室等のホームページにおいても、

アクセス案内などが提供されています。そこで、各所から大学への案内にも対応できることを想定しました。京都は、その土地柄一般的な機械翻訳だけでは、地名の翻訳が困難な場合がおおいです。例えば、某ポータルサイトが提供する翻訳サイトでは、京都大学が面する東大路通（ひがしおおじどおり）は、“Toukyoudaigaku michi tsuu”と訳されてしまいます。このため、京都大学周辺の地名および、校舎などの辞書の作成をおこないました。

このように作成した辞書は、日英 15000 語以上であり、韓国語・中国語に関しても 5000 語程度（科目名・人名の英語表記を含めると約 15000 語）の京大辞書を作成しました。これらの京大辞書と機械翻訳を連携した翻訳結果例を Table1 に示します。辞書内に登録され、利用された用語を下線で示します。某ポータルサイトで提供されている翻訳結果と比較し、辞書連携することで翻訳精度が向上していることがわかります。

4.2 京大翻訳！の機能

京大翻訳！の各機能について、説明します。

1) テキスト翻訳機能

フォームに入力したテキストを、指定の言語に翻訳する機能です。オプション機能として、「折り返し翻訳」、「原語表示」、「引用符付きフレーズの翻訳回避」の3つを用意しました。

「折り返し翻訳」は、翻訳結果を再度翻訳元言語に翻訳し直し表示する機能です。

「原語表示」は、辞書から抽出され利用された用語が括弧（ ）付きで表示される機能です。連携している辞書により、どの用語が京大辞書から利用されているかを確認することができます。

「引用符付きフレーズの翻訳回避」は、ダブルクォーテーション（半角“）で囲んだ箇所は翻訳されずに表示される機能です。例えば英日翻訳で URL やメールアドレスが文章中に含まれている場合でも、翻訳されてしまう場合があります。そこで、”

Table1. 京大辞書連携による京大翻訳！の翻訳結果例

(a) 京大への道案内

	京都駅北出口のバス乗り場から市バス#206 に乗車してください。 京大正門前でバスを下車して、東一条通を東へ進んでいただくと、左手に正門が見えてきます。
京大翻訳！	Please get on city bus # 206 from a bus stop in Kyoto Station North exit. When I get off a bus in <u>Kyodai Seimon-mae</u> , and you advance <u>Higashi-Ishijo St.</u> towards the east, a front gate is being seen as a left hand.
某翻訳 サイト	City bus stop from Kyoto Station Kitade mouth please take the # 206. Get off the bus at the front gate of Kyoto University, and east you go through Itizyou Azuma, you will find the main gate on your left.

(b) 講義に関する説明

	京大で開講している「IT 時代のヒューマンライフ (A 群)」は、台湾大学との遠隔講義です。
京大翻訳！	<u>Human Life in ICT Era</u> which has started by Kyoto University is remote lecture with Taiwan University.
某翻訳 サイト	At Kyoto University has offered "IT era Human Life (A group)" is the Distance Education and National Taiwan University.

From: 日本語
A- A+

ブックセンタールネは、百万遍を下がったところにあります。
京大生協のホームページは、これです。
"http://www.s-coop.net/"

☒ 原語表示
☒ 引用符付きフレーズの翻訳回避
☒ 折り返し翻訳

翻訳
消去

To: 英語
A- A+

Book Center Renais (ブックセンタールネ) is at the place where I went down Hyakumanben (百万遍). A home page of Kyoto University Co-op (生協) is this.
http://www.s-coop.net/

翻訳時間 21 sec.

Powered by Language Grid / J-Server / Kyoto EBMT

Fig.3 テキスト翻訳機能

で囲むことにより翻訳が回避できるようにしました。これらの機能はオプションであるため、利用目的によりユーザが選択できる機能になっています (Fig.3)。

2) 用例対訳・辞書の検索機能

京大辞書作成と同様に、日英文書などから用例対訳の抽出も行いました。教務などの文章が中心となりますがキーワードを入力することで、対訳の検索が可能になりますので、文書作成時に利用が可能です。検索には、部分一致、完全一致の他、文書の類似度を計算し検索を行う手法も導入しました。

一方、辞書検索機能も、京大辞書からの検索が可能となっています。施設名や科目名等正式にどのように訳されるのか調べたい場合などに便利な機能

といえます。

3) 用例ベース翻訳を用いた最良翻訳選択機能

本サービスでは、京都大学情報学研究科（知能情報学専攻 黒橋禎夫教授）が科学技術振興調整費の支援を受けて開発した KyotoEBMT (Example-based Machine Translation) を導入し、利用しています。KyotoEBMT は日英・日中のように構造が大きく異なる言語間の翻訳を正確に行うために深い構造解析を行っており、さらに用例対訳を用いて機械翻訳の精度を向上させることができるため、正確な対訳文が蓄積されるに従い機械翻訳が学習し正確になっていきます。京大関連文書から抽出した 2000 文を学習させ KyotoEBMT を動作させています。通常時では、



Fig.4 KyotoEBMT を用いた翻訳

J-Server により翻訳結果が表示されますが、KyotoEBMT の翻訳結果の方が精度が高いと計算結果より判断した場合には、KyotoEBMT の翻訳結果も表示されるような設計になっています。例えば、「2月以降はできるだけ、生協で買ってください。」という大学事務から教員や研究員への連絡がなされることがあります。これを通常の機械翻訳と辞書連携では、“As much as possible, please buy me after February in CO-OP.”という翻訳結果となります。「生協」が辞書により「CO-OP」と翻訳されていますが、訳文の意味が異なります。この時、KyotoEBMT が同時に翻訳を行い、機械翻訳と辞書連携の結果よりも最良と判断した場合に、『別の翻訳結果を表示する』という文字が画面上に表示されクリック可能になります。クリックすると、EBMT の翻訳結果“As much as possible, please buy your items after February from the co-op.”が表示されます。最良と判断された箇所が緑背景となります (Fig.4)。これは、京大の日英文書から抽出し、登録されている「できるだけ、大学の生協 (COOP) で買ってください。」・“As much as possible, please buy your items from the University Co-op.”という用例対訳が学習されたことによる結果です。

5. おわりに

言語グリッドを用いて、特定のコミュニティに特化し、不特定多数のユーザに提供を目的としたスマ

ート翻訳について説明し、第1弾である京大翻訳！について紹介しました。今回は、京大をドメインとしましたが、スマート翻訳を用いることで、他大学を含め、各自治体においても「ご当地翻訳」の展開が可能だと考えています。京大翻訳！を足がかりに、京都翻訳、○○大翻訳などと広がっていくことを期待しています。

公開サイト：

以下関連の公開サイトになっています。

・言語グリッド

<http://langrid.nict.go.jp/>

プロジェクト全体のポータルサイトになっています。様々な活動やニュースの情報を得ることができます。

・言語グリッド Toolbox

<http://langrid.nict.go.jp/langrid-toolbox-wiki/>

Toolbox に関してのサポートサイトです。ソースコードのダウンロードおよびインストール手順等の情報を提供しています。ユーザ組織専用の Toolbox を提供するホスティングサービスも NICT が提供しています。

・京大翻訳！

<http://www.s-coop.net/smarttrans.html>

謝辞：

スマート翻訳および京大翻訳！は、多くの皆さまに支えられて進めております。ここに感謝の意を表します。

京都大学情報学研究科 黒橋禎夫先生, 中澤敏明氏には、KyotoEBMT を提供いただくとともに、Smart 翻訳への導入において、多くの技術協力をいただきました。

京都大学生生活協同組合の中森一朗氏はじめ関係者の皆さまには、京大翻訳！の運用に関して、多大な協力をいただいております。

本研究の一部は、京都大学グローバル COE プログラム「知識循環社会のための情報学教育研究拠点」並びに、総務省戦略的情報通信研究開発推進制度 (SCOPE) から助成を受けて行われています。

参考文献：

- [1] T. Ishida, "Language Grid: An Infrastructure for Intercultural Collaboration," in Proceedings of the IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06), pp.96-100, Keynote address, 2006.
- [2] Masahiro Tanaka, Yohei Murakami, Donghui Lin and Toru Ishida, Language Grid Toolbox: Open Source Multi-language Community Site. 4th International Conference on Universal Communication (IUCS2010), 2010.
- [3] Yohei Murakami, Donghui Lin, Masahiro Tanaka, Takao Nakaguchi and Toru Ishida. Language Service Management with the Language Grid. 7th International Conference on Language Resources and Evaluation (LREC2010), pp. 3526-3531, May, 2010.
- [4] Toshiaki Nakazawa, Sadao Kurohashi, Fully Syntactic EBMT System of KYOTO Team in NTCIR-8, In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-8), pp.403-410, Tokyo, Japan (2010.6).

機械翻訳＋ポストエディットの実証研究（その2）： 英日翻訳での実験結果

立教大学大学院 異文化コミュニケーション研究科 博士後期課程
山田 優

前回のAAMTジャーナル(No.48)では、機械翻訳のポストエディット(以下MT+PE)を翻訳支援ツールとして活用した場合の、翻訳品質と作業効率に関する先行研究を概観した。これらの文献で検証対象とされていた欧州言語の組合せでは、翻訳の分野や目的を限定すれば、MT+PEは実務レベルの品質／効率にほぼ達していることが実証データから分かった。

Bowker & Ehgoetz (2007)が示したように、ユーザーが求める品質相応の納期とコストを評価条件に加えることにより、MT+PEによって生み出されたプロダクト（訳出物）が商品（実務）としての価値を持つようになる。すなわち、MT+PEは、例えば、出版翻訳には向かないとしても、日常業務に支障ないビジネス翻訳としては十分耐えうる品質とスピードを兼ね備えているということである。また、O'Brien (2006b)の実験により、MT+PEの作業が時間だけでなく、実際の作業者の認知負荷のレベルにおいて、翻訳メモリ(TM)85%マッチの分節を処理する負荷とほぼ同等相当だということも示された。

一方で、MT+PEのポジティブな検証結果はいずれも西欧言語の組合せでのみ実証されているように見受けられ、言語間の「距離」のある組合せ、すなわち英語と日本語のように統語や語彙の非似なるペアでも同様の結果が得られるかどうかは不明である。ということで、今回のレポートでは、英語→日本語でのMT+PEの検証を行う。筆者が行った一連の実験のうち、断片的にはあるが、MT+PEの作業負荷、効率性、品質に関する結果を本紙面で報告したい^(注1)。

実験の目的と評価方法

本実験の主な目的は、英日翻訳のMT+PE作業におけるMT訳の受容レベルを検証することである。平たく言えば、MT+PEは実務に耐えうるのか、効率は上がるのか、品質はどうなのか、といった点を検証するのが

目的だ。以下の3つの観点から検証を行った。

まず、翻訳者にとってMT+PEという作業は、普通に翻訳をするよりも楽なのか、という点である(ALPAC, 1966を参照)。被験者に作業後のアンケートを実施し、MT+PEの体感作業負荷を数値化してもらった。普通に翻訳する場合の体感工数を100(基準)として、MT+PEはどうだったか、例えば、MT+PEによって通常よりも2割程度の工数軽減を感じたならば80と記入(100マイナス20)する、逆に2割くらい増えたと感じた場合は120(100プラス20)という具合に回答してもらった。

主観的データなので信頼性に乏しく、またAPLAC報告書(ibid.)でも指摘されたように、翻訳者は概してMT+PEの作業を過大評価し過ぎるという危険性があるものの、その一方で、結局のところ、翻訳者自身がMT+PEをどう感じているのか、それが実務では一番重要な要素だとも言えるだろう。

2つ目の評価点は、修正量である。修正量(テキスト類似度)とは、PEを行う前のそのまのMT訳と、PE後の最終訳出物とが、どの程度似ているかということである。MT訳の品質が高ければ、PE量(すなわち修正量)が少なくなるだろう、という考えに基づいている。

修正量の測定には、MTの自動評価法のひとつであるGTM(General Text Matcher)を用いた。BLEUやWE Rなど広く使われているツールに比べて、比較的マイナーなGTMを選択した理由は、GTMではセンテンス単位での評価が可能だという点と、またTatsumi (2009)のPEの処理時間と自動評価点との相関性を調査した先行研究において、GTMの点数が最も高い相関を示していたことによる。

3つ目の評価ポイントは品質だ。効率UPや作業負荷の軽減が達成されても、最終品質が下がっては仕方ない。MT+PEの実践導入を検討する翻訳会社等にとっては、最も関心ある事柄であろう。ということで、品質面も限定的ではあるが評価対象とした。

実験デザイン

実験では、2種類の原文（英語）を用意した。それぞれ約40ワード、内容は工業機械のマニュアル文書の抜粋を使用した。被験者には、両方のテキストをポストエディットしてもらった。用意した2つのテキストは、内容こそ似ているが、構文の複雑性などが異なっている。ひとつのテキストは、いわゆる制限言語(CL=controlled language)で書かれた原文で、ASD-STE 100 (Ishikawa, 2005)に準拠している（以降、こちらのテキストを「CLテキスト」と呼ぶ）。これに対して、もうひとつの原文は、通常のライティングによる英文テキストである（こちらを「Non-CLテキスト」と呼ぶ）。これら2つのテキストを、それぞれ機械翻訳にかけてから、PEを行った。

2種類のテキストを比較する実験デザインはO' Brien (2006a)と似ているが、ここでは厳密に制限言語を用いたMT+PEの作業効率を評価するのが目的ではなく、実務翻訳で言うところの「ローカリ分野」の対象となるマニュアル・テキストと、そうでない原文との違いをみるために用意したに過ぎない。つまり、CLテキストは「マニュアル文書」、Non-CLテキストは「マニュアルではない文書」といった状況を想定した。

それぞれのテキストが40ワードと少ない点については、実験自体がパイロット・レベルであり、本結果に基づいて追試等を別途行っている点是否めないが、近年の翻訳実務においてTMS（翻訳マネージメントシステム）等の導入により、プロジェクトが分割化され、翻訳者に一度に割り当てられるワードカウントが数十から数百という状況も珍しくない現状があり、このような点を考慮すれば、用意したテキストのワードカウントは、翻訳現場の現状を反映しているともいえよう。

被験者は大学院で通翻訳を学ぶ学生（セミプロ）8名に依頼した。翻訳スキルや分野の背景知識では、プロのレベルに達していない点は、本実験の限界点として挙げておく。しかしながら先行研究でも、翻訳学科の修士課程の大学院生を「プロ」と看做すケースも多々ある（Lörscher, 1996等参照）。いずれにしても、本実験結果の解釈には、筆者がプロに対して行った別実験の結果とも照合しながら補足的に考察するように心がけた。

実験作業には、Google Translator Toolkit(以下GTT)を使って作業をした(Garcia, 2010参照)。上記2種類

のテキストを同サイトにアップロードし、GTTのインターフェース上でPE作業を行った。これにより、実験で使ったMTエンジンは同社のSMT（統計的機械翻訳）ということになる。最近のMTの動向を考慮した。

GTTは割合と使い易いツールではあるが、実験参加者にツールに慣れてもらうために、実験前に練習用テキストを用意して、1時間程度のトレーニング時間を設けた。

作業は、教室環境で全員一斉に行った。最初にNon-CLテキスト、続いてCLテキストの順でPEを行い、作業後にアンケートに回答してもらった。

仮説（予測される結果）

検証ポイントとなる体感工数、修正量、品質の各項目について予測される結果を以下に述べておく。

仮説1：先行研究の実験では、MT+PEにより、実務に耐えうる効率や品質が得られることが観察されたものの、英語→日本語での使用に関しては、筆者自身はまだ懐疑的である。翻訳者が、普通の翻訳よりもMT+PEの作業の方が容易だと感じるかどうか、正直分らない。ということで、英日のPEでは、通常の翻訳と比較した場合の作業負荷の軽減はごく僅かだろうと予測する。具体的な数字としては、Krings (2001)が時間的に20%程度の短縮を達成できたという結果を受けて、英日でも10～20パーセント程度の負荷軽減であれば現実的な数字だろうと仮定する。

仮説2：Non-CLとCLテキストの比較では、おそらくCLテキストのPE作業のほうが楽に感じられるだろう。O' Brien (2006a)の結果でも、CLテキストの方が工数的には軽減されていた。ということで、CLテキストのMT+PEの主観評価点の方がNon-CLテキストよりも低くなる（すなわち、作業が楽になる）と予測される。

仮説3：上記の仮説2の結果に準じて、実際の修正量にも違いが見られると思われる。CLテキストのPE作業の方が楽に感じるのであれば、実際の修正量も減っていると予想される。ということで仮説3は、CLテキストのMT+PEの修正量は、Non-CLテキストよりも減少するとしておく。

仮説4：翻訳者は、出力されたMTの品質にかかわらず、各自の最終訳出物に対しては一定の責任を負うはずである。CLとNon-CLのMT訳の品質に差があったとしても、PE後の両テキストの最終品質に差があるとは

考えられない。むしろその差は、上記で予想したように、体感工数や修正量の差として現れるはずである。仮説4は、CLおよびNon-CLテキストの最終訳出物の品質の差はない、とする。

結果

それでは実験結果を、上記の仮説順に見ていくことにしよう。まず初めは、被験者の主観的工数負荷の評価点である。

主観的工数負荷の軽減

通常の翻訳(100)を基準としたNon-CLテキストのMT+PE作業の評価は、平均86.9であった(表1を参照)。単純比較では、普通の翻訳に比べ約13%程度の感覚的負荷の低減がもたらされたと言える。ただし標準偏差(17.5)を考慮すると、必ずしも優位な差だとは言い切れない。それでも、8名中6名の被験者がMT+PEは普通の翻訳よりも楽だと評価しており、多少なりとも工数軽減を感じているようだ。残りの2名(翻訳者AとC)は、普通の翻訳と変わらない(Cのポイントは100)、または普通よりも大変であった(Aは110)と回答している。Non-CLテキストの方の結果なので、筆者自身はもう少し悪い結果を覚悟していたが、MT+PEという作業そのものが邪魔になってしまうほどではないようである。

次にCLテキストの結果を記す。平均点は73.1と、Non-CLテキストよりも下がっている。Non-CLとの差は13.8ポイントで、統計的にも有意差がある($t=2.986$)。被験者全員が、普通の翻訳よりも作業が楽になったと回答していることも含めて考察すると、CLテキストの方はNon-CLの場合よりも事実として作業が楽になっており、普通の翻訳との比較でも明らかに工数軽減を体感していると言える。

まとめると、MT+PEでの作業は、CLテキストのように、原文の構文構造をシンプルにしておけば、通常の翻訳作業よりも10〜25%程度の作業負荷低減が実感できるようである。よって仮説1は証明されたといえよう。またNon-CLとCLテキストの比較でも、作業負荷の差を体感出来たので、仮説2も同様に支持されたことになる。

表1：体感工数低減結果

翻訳者	Non-CL テキスト	CL テキス ト	Non-CL と CL の差
A	110	80	-30
B	80	80	0
C	100	70	-30
D	85	85	0
E	90	80	-10
F	50	30	-20
G	90	70	-20
H	90	90	0
平均	86.88	73.13	-13.75
標準偏差	17.51	18.70	
t			-2.986
有意差 = 5% ($t=2.365$)			

修正量 (テキストの類似度)

では、実際の修正を見る。繰り返しになるが、修正量(テキスト類似度)とは、PE前のMT訳と、PE後の最終訳との差分をGTMを使って測量したものだ。1.0が最高点で、それに近ければ近いほど類似度が高い、すなわち修正量が少ないといえる。

Non-CLとCLテキストの平均点数は、0.217 vs. 0.483であった。CLテキストの点数が、Non-CLの約2倍であった。これらの点数と、先の工数負荷の結果とを合わせて解釈すると、Non-CLとCLテキストとの体感工数の差は、GTMで0.266点(0.483-0.217)の差がもたらしたとも換言できる。GTMスコアと工数負荷軽減との間に相関関係があるとまでは断言しないが、工数負荷の差と同様に、GTMスコアにも有意差を観察できた。

表2：GTM スコア (PE 修正量)

翻訳者	Non-CL テキスト	CL テキス ト	Non-CL と CL の差
A	0.290	0.515	-0.225
B	0.197	0.493	-0.296
C	0.182	0.508	-0.326
D	0.269	0.492	-0.223
E	0.271	0.594	-0.323
F	0.179	0.500	-0.321
G	0.189	0.441	-0.252
H	0.156	0.317	-0.161
平均	0.217	0.483	-0.266
標準偏差	0.052	0.079	
t			-12.450
有意差 = 5% ($t=2.365$)			

では、ここでGTM点数自体の意味について考えてみたいと思う。補足データも交えながら、GTMスコアについて解釈を加えてみる。

まず、GTMの詳細アルゴリズムは無視して、便宜的に、CLテキストの0.483という点数を、そのままのMT訳の完成度が訳48%であったと読み替えてみると、翻訳者はPE作業中に100%に足りない52%分のテキストを修正したことになる。同様にNon-CLのケースでは、 $1.000 - 0.217 = 0.783$ 、つまり約78%分の書き直したと考えることができる。Kring (2001)の実験では、修正量の換算方法が異なるものの、大雑把に言えば、PEでの修正量が60%で、その作業時間は普通の翻訳に比べて20%減少したと報告されている。これを基にすると、本実験のCLテキストのPEでは、52%が修正され、体感工数が10〜20%軽減したというのは妥当な数字と言えるかもしれない。他方、Non-CLの場合も、78%の修正量で普通の翻訳と比較した場合の工数低減に有意差がなかったというのも、それはそれで信憑性のある結果であった。

また、筆者はこの実験とは別に、翻訳メモリ(TM)に関する検証を行い、そこでTMのマッチ率とGTMスコア(TMで表示されるFuzzyマッチ訳と最終的な訳文とのテキスト差)、それに各マッチ率の処理時間等を測定した。一般的に、TMを使った翻訳の場合は70%マッチ以下になると、TM上に既存のFuzzyマッチ訳を表示させるメリットがないと言われているが、その理由として、Fuzzyマッチを見比べながら作業するよりも、最初から翻訳し直したほうが速いから、というものである。しかし、筆者の実験では、50〜60%マッチくらいまでは、Noマッチよりも翻訳速度的に有意差があった。逆に言うと、TM50〜60%マッチまでは、Fuzzyマッチ訳をベースに作業をしたほうが効率性は上がるということなのだ。

この50〜60%マッチ付近を、効率性の分岐点と考えて、そのFuzzyマッチと最終訳出物との類似度をGTMで調べると0.460であった。この数字と本実験結果を比較してみると面白い。CLテキストはGTM0.483であり、若干ではあるがTMのFuzzyマッチの分岐点0.460を上回っていた。つまりCLテキストのMT+PE作業は、TMの50〜60%Fuzzyマッチよりも少し高いマッチ率の作業と同等だったと考えることができ、速度的にも普通の翻訳よりも速かったと想像できるのである。これに

対してNon-CLテキストの方は、GTM0.217であり、これは上記の分岐点よりも低いために、作業負荷的にも翻訳速度的にも普通に翻訳した時よりもメリットがほとんどなかったと解釈できるのだ。

品質

品質については、「正確性(accuracy)」のみを対象とし、筆者自身が評価を行った。結果は、作業工数および修正量の少なかったCLテキストの方には、特に目立ったエラーは見つからなかったが、Non-CLテキストには、1箇所致命的な誤訳が見つかった。驚くことに、8名中半分の4名が同じミスをしていた。原文のワード数が40と少ないにもかかわらず、5割の確率で同じミスが発生したのは、やや驚きの結果であった。以下に詳述する。

原文:

Gain access to blade.
After removing old blade, new blade may be fitted by proceeding in reverse order, using gloves to avoid injuries by teeth of blade.

MT訳:

ゲインのアクセスはブレード。
古い刃を削除した後、新しいブレードを装着逆の順序で進めて、ブレードの歯でけがを避けるために手袋を使用することがあります。

最終訳 (被験者Bの例):

刃を取り替えます。
古い刃を取り除いた後、刃先でけがをしないように手袋を使用しながら、新しい刃を装着時とは逆の手順で装着してください。

原文(英語)は、new blade may be fitted by proceeding in reverse order、つまり「取外時とは逆の順序で新しい刃を装着します」という意味である。しかし、MT訳が「装着逆の順序で進めて」となっているため、これに引きずられるようにPE後の訳が「装着時とは逆の手順で装着してください」となってしまう。「装着時と逆の順序で装着する」のは理論的に不可能であり、取扱説明書の翻訳としては致命的な誤訳だ。なぜこのようなエラーが半数の被験者で発生したのだろうか。

MT+PE作業のように、提示された訳を修正するよう

な作業の場合、提示訳（この場合はMT訳）にエラーが含まれていると、翻訳者（作業者）はそのエラーに気づかないことが多々ある。翻訳メモリの作業でも、TM内にミスがあれば、それが新規訳に残ってしまう、いわゆる「エラーの伝播(error propagation)」という現象が起きる。上記の誤訳も、伝播の一種と考えられるだろう。

Bowker (2005)やRibas (2007)では、TM作業でのエラー伝播を検証しており、100%マッチで残ってしまうのならともかく、マッチ率が低い場合でもエラーが伝播する点を指摘している。本実験をTMのマッチ率で考えた場合、50～60%マッチ相当の作業であったと先述したが、結果的にエラーは修正されていないなかったことになる。

Guerberof (2008)は、MTとMTの内在エラー修正率を比較した実験を行い、TM作業の方がMT+PEよりもエラー伝播率が高いことを示した。その理由として、TMの場合は、人間の翻訳者による訳文がベースとなっているため、訳文が「自然な訳」であり、翻訳者は細かなミスに気づきづらい点を挙げている。また、Fiederer & O'Brien (2009)は、品質の3要素のうち、翻訳者は、styleの要素をaccuracyやclarityよりも重要視してしまう傾向を挙げている。つまり、本実験でPE作業をした翻訳者は、品質のstyleの部分、すなわち訳文がマニュアルとして正しい文体になっているかという側面に気を取られたために、accuracyの要素を見過ごしてしまったと考えることができるのだ。

更に、今回使用したMTエンジンがSMTであったということからも、従来のルールベース訳と比較して「自然な訳」であった、という可能性も見過ごせない。今後、SMTを用いたPEを行う場合には、考慮しなければならない品質的なデメリットがはっきりとした形で露呈された結果となった。

品質と作業負荷評価の関係

最後に、観察レベルでの分析でしかないのだが、品質に関する興味深い結果が見られたので記載しておく。エラーに気付いた翻訳者とそうでない翻訳者との違いを、彼らの作業負荷評価の結果から考察してみた。

ミスをした4名の翻訳者（A、B、C、D）のうち、AとBは、MT+PE作業の評価を、普通の翻訳と変わらないか、むしろ大変だったとしていた。彼らのNon-CLの

点数は、それぞれ110と100であった。これに対して翻訳者BとDは、MT+PE作業を過大評価しすぎている感があった。彼らの点数は80と85だった。翻訳者Fが50という評価をしているのを例外とすれば、残りの翻訳者（ミスをしなかった翻訳者）は90という妥当な評価をしていた。

つまり、エラーをした翻訳者のMT+PEに対する評価は両極端であり、作業が楽ではなかったというように悲観的な見方をしている者か、または非常に楽であったと楽観的な評価をした者だということが分かる。MT+PEに対する苦手意識が強くても、得意気すぎた態度でも、品質的には悪くなる可能性があるもので、より冷静な評価が必要であるということが示されたようである。今後、MT+PEの実践へ導入を考えている翻訳会社等にとって、翻訳者（ポストエディター）の選定の際に考慮したい結果であろう。

まとめ

以上、実験デザインとして不完全な部分は残るものの、限定的な状況における英日翻訳のMT+PEの受容性についての実験結果を述べてきた。被験者のサンプル数、テキストのワード数、原文テキストの分野等に関しては、今後の研究での精緻化を目指したい。

ということで、最後に現時点における英日のMT+PEの結果をもう一度まとめておく。CLテキストの結果で見たように、一般的なマニュアルのようなシンプルな文体で書かれた文章のMT+PEであれば、翻訳者（作業者）は、普通の翻訳と比較して10～20%の作業負荷軽減を体感できるようである。その根拠として、PEでの実際のテキスト修正量が、翻訳メモリでの50～60%Fuzzyマッチを若干上回るレベル（GTM0.460以上）に達していることが確認できた。しかし一方で、SMTエンジンを使うとMT訳が「自然」になるために、それに翻訳者がとらわれてしまい、accuracy等の基本的な誤訳に気づきづらくなる可能性が増すようだ。MT+PEの実践への応用には、以上のような点を考慮して検討していきたい。また研究面でも、より確実なデータを提供していけるように努めたいものである。

【註】

（注1）本稿は、2011年3月発行予定の『異文化コミュニケーション論集』（立教大学大学院 異文化コミュニケーション研

究科)に掲載予定の筆者の論文(英語)に加筆および修正を行い日本語で書き改めたものである。

【参考文献】

- ALPAC. (1966). Languages and machines: Computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council. Retrieved from http://www.nap.edu/openbook.php?record_id=9547&page=R1
- Bowker, L. (2005). Productivity vs quality: A pilot study on the impact of translation memory systems. *Localisation Focus*, 4(1), 13-20.
- Bowker, L., & Ehgoetz, M. (2007). Exploring user acceptance of machine translation output: A recipient evaluation. In D. Kenny and K. Ryou (Eds.), *Across boundaries: International perspectives on translation* (pp. 209-224). Newcastle-upon-Tyne: Cambridge Scholars Publishing.
- Fiederer, R., & O'Brien, S. (2009). Quality and machine translation: A realistic objective? *The Journal of Specialised Translation*, 11, 52-74.
- Garcia, I. (2010). Is machine translation ready yet? *Target*, 22(1), 7-21.
- Guerberof, A. (2008.) Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus*, 7(1), 11-21.
- Ishikawa, S. (2005). Potential application of XML [XML no motsu potential o kangaru]. TC symposium, Japan Technical Communicators Association.
- Krings, H. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes* (G. S. Koby, Ed). Ohio: Kent State University Press.
- Lörscher, W. (1996). A psycholinguistic analysis of translation processes. *Meta* 41(1), 26-32.
- O'Brien, S. (2006a). Controlled language and post-editing. *MultiLingual*, October/November, 17-19. Retrieved from <https://216.18.156.115/multilingual/downloads/screenSupp83.pdf>
- O'Brien, S. (2006b). Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, 14 (3), 185-205.
- O'Brien, S. (2008). Processing fuzzy matches in translation memory tools: An eye-tracking analysis. In S. Gopferich, A. Jakobsen, & I. Mees (Eds.), *Looking at eyes: Eye-tracking studies of reading and translation process*. (pp. 79-102). Copenhagen: Copenhagen Business School.
- Ribas, C. (2007). *Translation memories as vehicles for error propagation: A pilot study*. Minor Dissertation. Tarragona. Universitat Rovira i Virgili.
- Tatsumi, M. (2009). Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. MT Summit XII proceedings. Retrieved from <http://www.mt-archive.info/MTS-2009-TOC.htm>

MT Summit XIII における「特許翻訳特別セッション」および The 4th workshop on patent translation の開催について

AAMT/Japio 特許翻訳研究会

江原 暉将、横山 晶一

AAMT/Japio 特許翻訳研究会では、本年 9 月に中国のアモイで開催される MT Summit XIII に合わせて The 4th workshop on patent translation (第 4 回特許翻訳ワークショップ) を主催する。さらに MT Summit XIII の本大会の中で「特許翻訳に関する特別セッション」を提案し採択された。特許翻訳ワークショップは MT Summit にあわせて開催してきており、今回は第 4 回になる。

特許の分野では機械翻訳の実利用が進んでいる。知財に関する主要な五庁（日本、韓国、中国、欧州、米国）の間で機械翻訳を用いた特許文書の国際的流通が図られている。このプロジェクトでは、一般に公開されている特許文書の翻訳と審査情報など部内のみで利用する文書の翻訳の双方に機械翻訳を利用するものである[1]。

一方、世界知的所有権機関(WIPO: World Intellectual Property Organization)でも機械翻訳を PATENTSCOPE [2]という検索システムの中で実用しており、こちらは統計方式機械翻訳である Moses をベースにしたシステムである[3]。

研究面では NTCIR-9 の中で日英・英日および中英の特許翻訳タスクが行われており、多くの研究者の注目を集めている[4]。

このような背景の下で、本特別セッションとワークショップは開かれる。沢山の興味ある発表がなされると期待されるので是非ご参加いただきたい。この原稿を執筆している段階ではプログラムは未定であるが、主催者としては以下のような招待講演を企画中である。

- ・ KIPO (Korean Industrial Property Office) による五庁プロジェクトの概要と韓国での特許機械翻訳

利用の現状 (KIPO は五庁プロジェクトの中で機械翻訳部分の幹事庁である)

- ・ WIPO による PATENTSCOPE での機械翻訳利用状況
- ・ CPIC (China Patent Information Center) による中国での特許機械翻訳利用の現状
- ・ JPO (Japan Patent Office) による日本での特許機械翻訳利用の現状
- ・ 研究者側からの基調講演 (Microsoft Research Asia の辻井潤一氏を予定)

詳細は MT Summit XIII のホームページ[5]を参照していただきたいが、ワークショップの主な日程は以下のとおりである。

ワークショップ日程

発表申し込み締め切り	2011 年 7 月 22 日
採否の通知	2011 年 8 月 12 日
カメラレディ締め切り	2011 年 8 月 26 日
ワークショップ当日	2011 年 9 月 23 日

特別セッションの日程は現時点では未定である。

参考文献・サイト

- [1] 岡崎輝雄：特許庁における機械翻訳の活用状況とその将来像、第 1 回特許情報シンポジウム発表資料
[http://www.japio.or.jp/kenkyu/files/kenkyu03/AAMT_Japio_sympo\(20101210\)-02.pdf](http://www.japio.or.jp/kenkyu/files/kenkyu03/AAMT_Japio_sympo(20101210)-02.pdf)
- [2] <http://www.wipo.int/patentscope/translate>
- [3] http://www.wipo.int/patentscope/en/news/pctdb/2011/news_0001.html
- [4] <http://ntcir.nii.ac.jp/PatentMT/>
- [5] <http://mt.xmu.edu.cn/mtsummit/>

AAMT 機械翻訳文テストセット第2版作成報告

機械翻訳課題調査委員会ワーキンググループ 1

1. はじめに

昨今、日中間の貿易や人的交流が盛んになるにつれ、日本語と中国語の間の機械翻訳への期待が高まっている。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）では、今後益々重要性が増すことが予想される日中機械翻訳システムの翻訳品質を評価する手法の確立を目的として活動を推進してきた。具体的な活動内容は、日本語および中国語の文法的特徴（評価項目）を翻訳システムが正しく処理することができるかどうかをチェックするための例文と設問のセット（AAMT 機械翻訳文テストセット）の開発である。評価方法に Yes/No での回答が可能な設問形式を取り入れることにより、客観性が高く、かつ、人間による主観評価に近い評価を低コストで実現することを狙ったものである。

2009 年度に行った「AAMT 機械翻訳文テストセット第1版」の公開に引き続き、2010 年度は、設問による評価の信頼性と客観性をより高めるために、評価項目の追加と設問内容の見直しを行い「AAMT 機械翻訳文テストセット第2版」を作成した。本稿では、AAMT 機械翻訳文テストセット第2版の概要と、これを用いて代表的な日中システムを対象に評価を行った結果について紹介する。

2. AAMT 機械翻訳文テストセットの強化内容

AAMT 機械翻訳文テストセット（以下、テストセット）は、日本語原文、中国語訳文（参照訳）、設問の組から構成される。以下では、第2版に向けて行った評価項目の追加と設問内容の見直しについて説明する。

2.1 設問項目の追加

テストセット第1版は以前に作成された日本語と英語の間の評価用例文をベースに作成したため、中国語の主要な文法現象であるにもかかわらずテストセットでカバーされない項目が存在していた。たとえば、英語の“not”という単語に対して中国語では2つの単語（「不」と「没」）が対応するが、これらを翻訳システムが区別して訳出することができるかどうかについての評価がテストセット第1版ではできなかった。テストセット第2版に向けては、中国語文法の専門家による評価項目の見直しを実施し、不足していた項目について例文と設問を追加した。テストセット第2版で新たに追加した文法項目は39、合計251文である。文法項目名と件数の詳細を表1に示す。

表1 第2版で追加した文法項目と文数

文法項目	文数	文法項目	文数
「の」と「的」	20	「～ている」が無標になる場合	17
「Vた」と「的」	3	「過」（～たことがある）	2
「もの」と「的」	3	自他動詞	4
方位詞	6	～で	12
疑問詞	8	のだ文	3

否定文（否定副詞の位置）	5
「不」と「没」	5
「在」文	5
「有」文	14
形容詞述語文	16
兼語文	10
二重目的語をとる動詞	4
できる・可能（「会・能・可以」）	13
許可（可以）	4
願望	4
前置詞	7
比較文	6
「Vたばかり」と「刚刚」	2
「～ている」と「了」「着」	8
「了」（変化）	3

「把」字文	5
補語	9
使役文	11
受身文	6
被を用いない受身文	2
可能	2
自発	1
尊敬	4
敬語	4
ようだ・そうだ	6
「う・よう」と「吧」	9
「～ば～ほど」と「越～越～」	1
介詞	2
決まり文句	5
計	251

一方で、テストセット第1版の中に、日本語と中国語の翻訳の評価では意味のない設問が散見されたので、これらの例文と設問の削除を行った。テストセット第1版と見直し後の第2版の評価例文数を図2に示す。

表2 テストセット中の評価文数

テストセット	評価文数
第1版	325
第2版	576

2.2 設問内容の見直し

設問への回答は Yes または No のどちらかであり、誰が評価しても同じ回答が導かれるべきものである。ところが、テストセット第1版を使って実際に2人の評価者で評価を行ってみると3割以上の設問で評価結果に食い違いが生じた。多数の食い違いが生じた原因として「設問文が明確でない」ことが考えられた。たとえば、第1版には下記のような設問が含まれている。

「慣用表現の訳が適切ですか？」

このような設問では、適切かどうかの判断は評価者の言語知識に依存するため、実際に評価者によって意見が分かれる場面が少なくなかった。そこで、第2版で追加する項目については、設問文の書き方を「指定した表記が訳文中に含まれるか（含まれないか）否か」を問う内容とした。たとえば、下記のような設問を基本としている。

「否定文が正しく訳されていますか。「不」が「没」になっていたら、×（No）」

3. 評価実験の実施

テストセット第2版（テストセット第1版および第2版の追加部分）を用いて、現在 Web からアクセスすることができる代表的な日中翻訳システムについて実際に評価を行った。評価の方法は以下のとおりである。

① 評価対象：

6 システム（Web 上で無償で使えるもの）

提供元の内訳：日本企業 3、米国企業 2、中国企業 1

② 評価者：

2 名（日本語検定 1 級を持つ中国語ネイティブ）

③ 評価観点：

－正確さ（Adequacy）：

原文と訳文を見て 1～5 で評価

－流暢さ（Fluency）：

訳文のみを見て 1～5 で評価

－設問：

訳文と設問を見て Yes/No 評価

なお、設問評価の結果との比較のために、統計に基づく客観評価 4 種（BLEU、NIST、WER、PER）についても評価を行った。

4. 評価実験の結果

今回の実験結果について、以下の 3 つの観点から検討を行った。

- ・ 設問評価の内容を見直したことによる評価者間の評価の異なりの変換
- ・ 設問評価と従来の主観評価との関係
- ・ 設問評価と統計ベース評価（客観評価）との比較

4.1 評価者間の評価の異なり

2 人の評価者による設問回答の結果は、4 通りの組み合わせ（Yes-Yes, No-No, Yes-No, No-Yes）のいずれかである。2 人の評価者による設問回答の組み合わせの件数と割合を、テストセット第 1 版、第 2 版追加項目に分けて、表 3、表 4 にそれぞれ示す。

表 3 第 1 版での設問評価内訳

評価者 1	評価者 2	件数	割合	
Yes	Yes	701	36.0%	74.3%
No	No	748	38.4%	
Yes	No	37	1.9%	25.7%
No	Yes	463	23.8%	

表 4 第 2 版追加項目の設問評価内訳

評価者 1	評価者 2	件数	割合	
Yes	Yes	456	30.3%	83.9%
No	No	806	53.6%	
Yes	No	72	4.8%	16.1%
No	Yes	170	11.3%	

第1版では2人の評価者で全体の25.7%の評価が異なっていたのに対して、第2版の追加項目を対象にした評価では異なりが16.1%まで減っている。このことから、第2版に向けて行った設問内容の見直しは、設問評価の客観性を高めるという点に関して大きな効果があったといえる。

しかしながら、第2版の追加項目においても全体の10%以上で異なりが生じており、原因の調査が必要である。たとえば、次の例では、評価者1がYes、評価者2がNoと異なった回答をしている。

原文：私が住んでいる部屋。 （参照訳：我住的房间。）

設問：「住んでいる部屋」が「住的房间」になっていますか。

＜例1＞機械翻訳文：我住在的房间。

（設問回答） 評価者1⇒YES 評価者2⇒NO

＜例2＞機械翻訳文：我是住的房间。

（設問回答） 評価者1⇒YES 評価者2⇒NO

例1は設問の回答としてNOとすべきであるが、評価者1は評価の過程で「在」の一字を見落としのために設問回答を誤った可能性がある。例2では、設問回答としてYESが正しいが、文全体で誤訳になっているために、評価者が先入観でNOと回答してしまった可能性がある。このように、評価者によって評価が異なる原因として、人間のミスや勘違いといったことも考慮する必要がある。

4.2 設問評価と従来主観評価の相関

設問評価と従来主観評価（正確さ（Adequacy）と流暢さ（Fluency））との関係を調べるために、相互の相関係数を求めた。2人の評価者別の第1版と第2版追加項目での相関係数を、表5および表6に示す。

表5 第1版での主観評価相互の相関

評価者 1				評価者 2			
	正確さ	流暢さ	設問		正確さ	流暢さ	設問
正確さ	1			正確さ	1		
流暢さ	0.656385	1		流暢さ	0.643233	1	
設問	0.695812	0.482246	1	設問	0.622676	0.369789	1

表6 第2版追加項目での主観評価相互の相関

評価者 1				評価者 2			
	正確さ	流暢さ	設問		正確さ	流暢さ	設問
正確さ	1			正確さ	1		
流暢さ	0.700983	1		流暢さ	0.613221	1	
設問	0.659782	0.44162	1	設問	0.655224	0.344173	1

表5、表6を見ると、どちらの評価者においても、また、第1版と第2版追加項目のどちらにおいても、設問評価は

正確さの評価と高い相関を示している。正確さの評価に比べて流暢さの評価との相関は高くない。設問評価は文法項目についてのチェックであるため、流暢さよりも正確さとの相関が高くなるだろうと予想したが、それを裏付ける結果となった。

4.3 統計的客観評価との比較

設問による評価と一般的に使われている統計的客観評価との比較を行うために、4つの代表的自動評価手法と2つの主観評価（正確さ、流暢さ）、そして、設問評価の間の相互の相関係数を求めてみた（表7～表10）。表中の英字表記の意味は以下のとおりである。なお、各自動評価の詳細についてはここでは触れない。

BLEU	: 自動評価	BiLingual Evaluation Understudy
WER	: 自動評価	Word Error Rate
PER	: 自動評価	Position independent WER
NIST	: 自動評価	National Institute of Standards and Technology
ADQ	: 主観評価	正確さ
FUL	: 主観評価	流暢さ
Q&A	: 設問回答	

表7 各指標間の相関係数（第1版 評価者1）

	BLEU	WER	PER	NIST	ADQ	FUL	Q&A
BLEU	1						
WER	-0.9850	1					
PER	-0.9687	0.9874	1				
NIST	0.9784	-0.9892	-0.9977	1			
ADQ	0.9596	-0.9302	-0.9282	0.9366	1		
FUL	0.8937	-0.9092	-0.8750	0.8875	0.8974	1	
Q&A	0.9906	-0.9814	-0.9737	0.9813	0.9817	0.9263	1

表8 各指標間の相関係数（第1版 評価者2）

	BLEU	WER	PER	NIST	ADQ	FUL	Q&A
BLEU	1						
WER	-0.9850	1					
PER	-0.9687	0.9875	1				
NIST	0.9785	-0.9892	-0.9977	1			
ADQ	0.9878	-0.9940	-0.9875	0.9896	1		
FUL	0.3690	-0.4599	-0.3967	0.4080	0.4188	1	
Q&A	0.9993	-0.9861	-0.9701	0.9778	0.9905	0.3606	1

表 9 各指標間の相関係数（第2版追加項目 評価者1）

	BLEU	WER	PER	NIST	ADQ	FUL	Q&A
BLEU	1						
WER	-0.9834	1					
PER	-0.9794	0.9824	1				
NIST	0.9840	-0.9894	-0.9987	1			
ADQ	0.9186	-0.9649	-0.9519	0.9608	1		
FUL	0.9577	-0.9523	-0.9299	0.9431	0.9334	1	
Q&A	0.9957	-0.9893	-0.9804	0.9871	0.9391	0.9561	1

表 10 各指標間の相関係数（第2版追加項目 評価者2）

	BLEU	WER	PER	NIST	ADQ	FUL	Q&A
BLEU	1						
WER	-0.9835	1					
PER	-0.9794	0.9824	1				
NIST	0.9840	-0.9894	-0.9987	1			
ADQ	0.9677	-0.9901	-0.9847	0.9908	1		
FUL	0.9082	-0.8437	-0.8098	0.8234	0.8039	1	
Q&A	0.9713	-0.9807	-0.9947	0.9956	0.9932	0.8049	1

表 7～10 を見ると、いずれの条件においても、すべての指標間で高い相関を示した。正確さ評価（ADQ）との相関に注目すると、いずれの条件においても、設問評価（Q&A）との相関係数が高い値になっており、設問評価の手法が主要な自動評価に比べて遜色ない信頼性を持っていることが期待できる。

5. おわりに

2010 年度の活動報告として、テストセット第2版に向けての評価項目の追加と設問内容見直しについての概要と、設問評価と従来の主観評価との相関、また、主要な自動評価手法と相関を比較した結果について紹介した。

テストセット第2版については、評価結果について分析がすべて終わりしだい、AAMT のホームページ上で公開する予定である。

設問回答による評価手法は、特に第2版に向けた設問内容の見直しにより、設問と訳文だけを見て評価ができるようになり、評価者の負担軽減につながったと考えられる。しかし、設問評価を人間が行う限り、評価コストの観点では自動評価が圧倒的に有利である。そこで、設問評価に人間ではなく機械が回答することができるように、設問の正規表現への書き換えを進めている。2010 年度中にテストセット第2について設問評価の自動化を実現し、2011 年度は既存の自動評価手法と組み合わせることも検討しながら、低コストで高信頼の評価テストセットの改良を進めていく。

英日・日英機械翻訳に関するアンケート結果報告

機械翻訳課題調査委員会ワーキンググループ 2

1. はじめに

AAMT 機械翻訳課題調査委員会 WG2（調査・広報・啓蒙）では、その前身である市場動向調査委員会の時代から調査活動の一環として、年一回インターネット上で一般の機械翻訳ユーザを対象としたアンケートを行っている。AAMT の Web サイトに特設ページを設け、機械翻訳に関するいくつかの設問に回答してもらうという形式である。本稿では、2010 年（2009 年度）のアンケート調査結果について報告する。

2010 年の調査は、定点観測と言える従来の基本項目（アンケート回答者自身の属性情報、機械翻訳の必要性、利用頻度、利用目的、分野など）に加え、利用者辞書登録の利用方法をテーマにアンケートを実施した。2009 年（2008 年度）アンケートでは携帯電話における機械翻訳の利用をテーマにしたため、アンケート回答方法も通常の PC での回答方法に加え、携帯電話用の回答ページも設ける工夫をしたが、今回は従来通りの方法に戻している。

以下に 2010 年のアンケート実施概要を示す。

- アンケート期間
 - 2010 年 1 月 27 日～2 月 9 日
- アンケート進呈品
 - 抽選で 10 名に機械翻訳ソフト、20 名にクオカード（3,000 円分）を進呈。
 - 回答者は進呈品として機械翻訳ソフトか図書券かを指定可能。
- アンケート回答者数
 - 1,022 人（うち、機械翻訳を必要としていたと回答した人は 745 人）
 - なお、過去のアンケート回答総数は、2003 年 1,413、2004 年 1,720、2005 年 1,018、2006 年 1,061、2007 年 835 人、

2008 年 1,495（946）人、2009 年 1,811 人（1,059 人）。

※括弧内は機械翻訳を必要とする人数。

2. 設問項目

主に以下に示す項目について質問した。

- 日英・英日機械翻訳ソフト・サービス全般に関して（日英・英日機械翻訳の必要性とその理由、機械翻訳の利用形態・用途・利用頻度・対象分野、機械翻訳に対する満足度等）
- 利用者辞書の利用方法（利用者辞書を利用しているか否か、利用者辞書の登録語数、利用者辞書を利用することによる作業効率向上の度合い等）
- 回答者の基本属性について（年齢層、性別、職業、英語力）

3. 機械翻訳全般に関する調査結果

本節では、機械翻訳全般に関する調査結果として、日英機械翻訳の必要性の有無とその理由、機械翻訳の利用形態・用途・利用頻度・対象分野、機械翻訳に対する満足度等についての調査結果を示す。

3.1. 日英・英日機械翻訳の必要性の有無とその理由についての調査結果

図 3-1 に（有償・無償か、また、Web 上のサービスかパソコンソフトかを問わず）日英・英日機械翻訳が必要か否か、また、必要・不要それぞれの理由を下記に示す。機械翻訳は、73%（745 人）の人が必要と答え、27%（277 人）が不要と回答している。

日英・英日機械翻訳の必要性は、2008 年（2007 年度）

から実施しているアンケート項目だが、2008 年調査では 63%（回答者総数 1,495 人）、2009 年調査では 53%（回答者総数 1,811 人）だった。回答者数が多いほど、「不要」と答える人の割合が高くなる傾向がある。

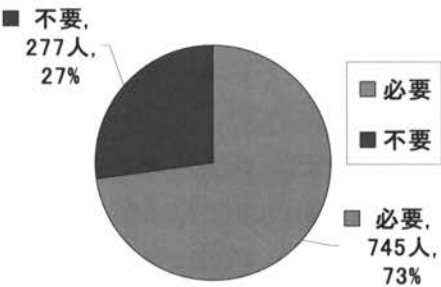


図 3-1 日英・英日機械翻訳が必要か

日英・英日機械翻訳の必要・不要それぞれの理由を図 3-2 と図 3-3 に示す。いずれも、理由は複数選択可能な回答とした。日英・英日機械翻訳が必要な理由としては、英語の能力不足をカバーする場合と、英語の能力はあるが生産性向上に使う（概要把握、下訳作成など）場合がほぼ同程度あった。また、自由記述の回答では、英語以外の他言語との翻訳の仲介・支援のために利用するという回答が 2、3 あった他は、選択肢以外の回答はなかった。一方、機械翻訳が不要な理由としては、そもそも必要性がない（「英語を話す機会がない」）という理由が多く、英語の能力があるから機械翻訳を必要としないという理由は少なかった。

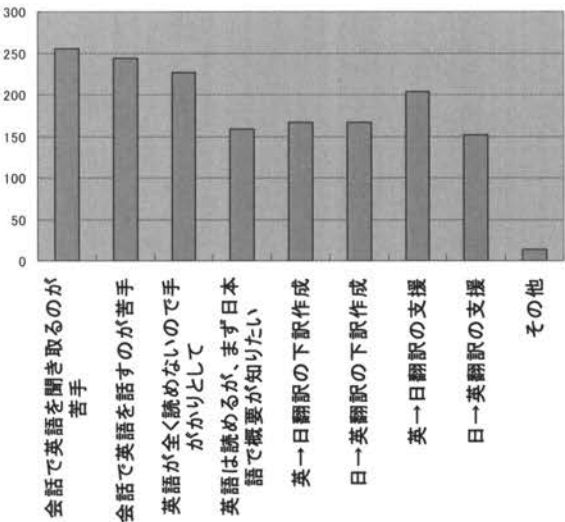


図 3-2 機械翻訳が必要な理由（複数選択可）

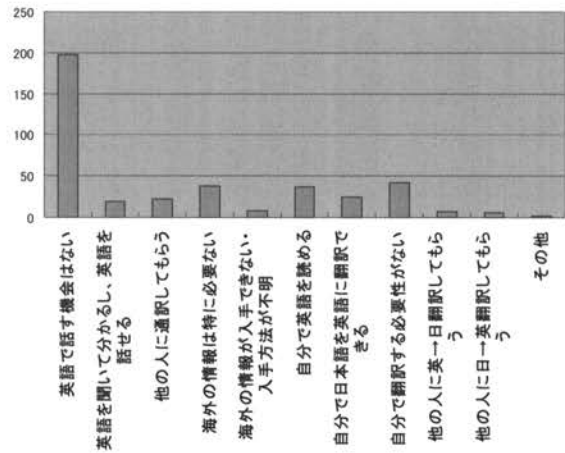


図 3-3 機械翻訳が不要な理由（複数選択可）

英語力と機械翻訳の必要性の関係を果を図 3-4 に示す。これを見ると、英語力が低い層と高い層で機械翻訳を不要とする割合がやや高いが、大きな違いではない。日英・英日機械翻訳は、利用者の英語力を問わず、おおむね一般に一定の価値を認められたと考えてよいであろう。

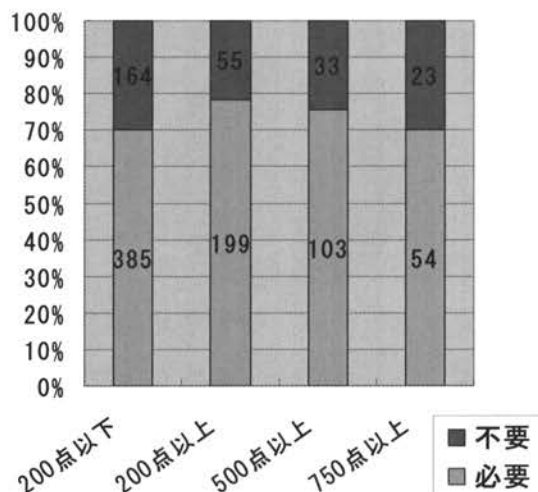


図 3-4 機械翻訳の必要性和回答者の英語力(TOEIC)の関係

3.2. 日英・英日機械翻訳ソフト・サービスの利用状況

以後、機械翻訳が必要という人の機械翻訳ソフト・サービスの利用状況を示す。

図 3-5 に、利用したことがある機械翻訳ソフト・サービス(複数選択可)、図 3-6 に最もよく利用する機械翻訳ソフト・サービスを示す。いずれも、無料の翻訳サイトの利用が圧倒的に高く、「機械翻訳と言えば翻訳サイト」という利用スタイルが広く浸透していることが窺われる。

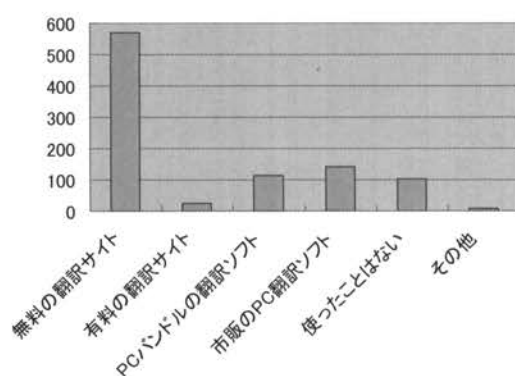


図 3-5 利用経験のある機械翻訳ソフト・サービス(複数選択可)

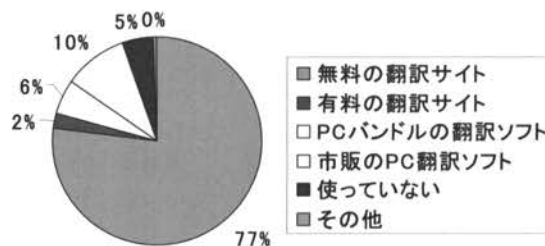


図 3-6 最も利用する機械翻訳ソフト・サービス

また、図 3-7 に最もよく利用する機械翻訳ソフト・サービスに対する満足度を示す。翻訳精度と機能について厳しい評価の一方、無償の翻訳サイトの利用者が圧倒的に多いことから価格面での満足度は高い。機能については、無償翻訳サイトではかなり機能が制限されることからきていると思われる。総合的に見ると、不満か満足かで二分すると、60%以上の人が不満と感じている。

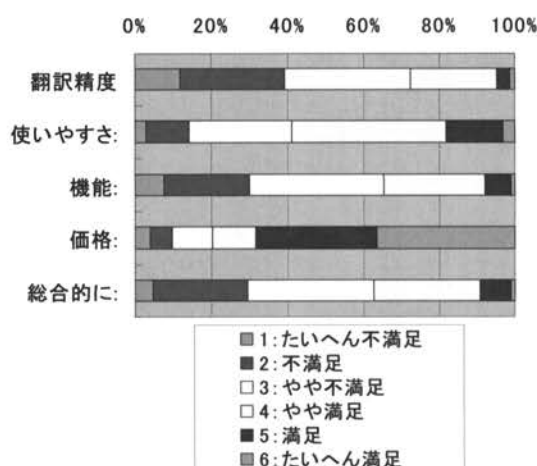


図 3-7 最もよく利用する機械翻訳ソフト・サービスに対する満足度

図 3-8 に最もよく利用する翻訳ソフト・サービスの利用頻度、図 3-9 にそのソフト・サービスを利用することによる作業時間短縮の割合、図 3-10 に最もよく利用する翻訳ソフト・サービスの用途を示す。総合的に見てかなり不満があるといいつつも、月数回以上、つまり定期的に機械翻訳ソフト・サービスを利用している人が90%近くいること、また、90%以上の人が20%以上の、75%以上の人が40%以上の作業時間短縮の効果を認めてい

ることがわかる。

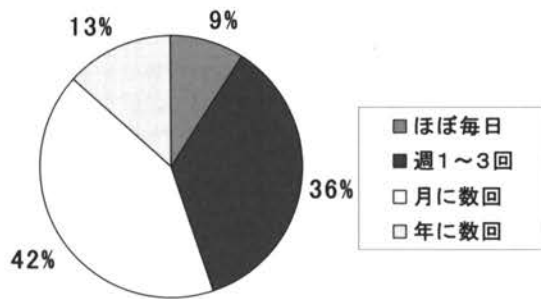


図 3-8 最も利用する機械翻訳ソフト・サービスの利用頻度

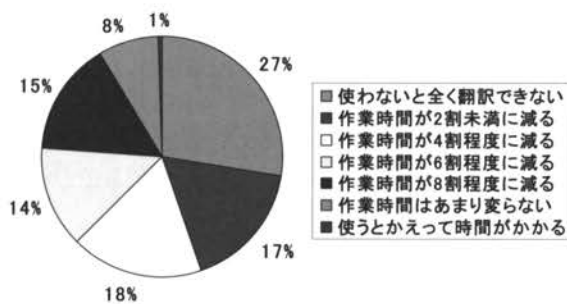


図 3-9 最も利用する機械翻訳ソフト・サービスによる作業時間短縮

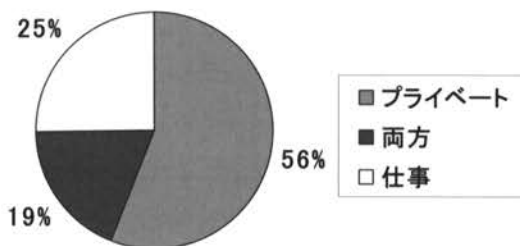


図 3-10 最も利用する機械翻訳ソフト・サービスの用途

図 3-11 から図 3-14 まで、最もよく利用する日英・英日機械翻訳ソフト・サービスについてのさまざまな利用状況の回答を示す。

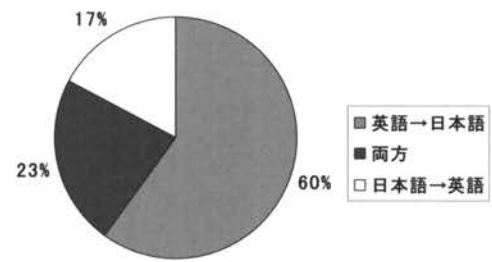


図 3-11 最も利用する機械翻訳ソフト・サービスで利用する翻訳方向

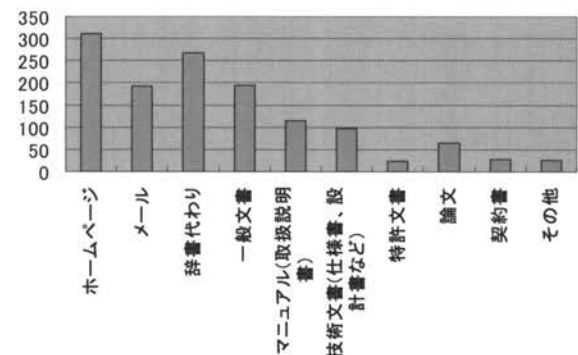


図 3-12 最も利用する機械翻訳ソフト・サービスにおける翻訳対象 (複数選択可)

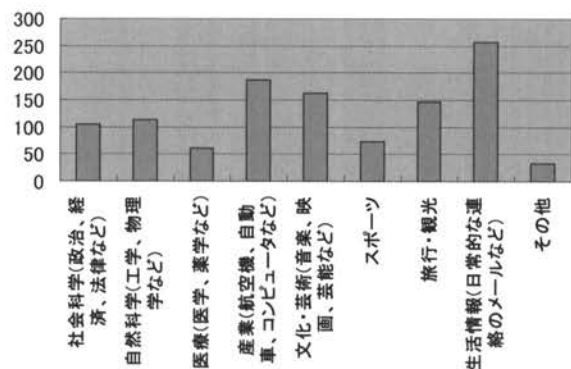


図 3-13 最も利用する機械翻訳ソフト・サービスにおける翻訳分野 (複数選択可)

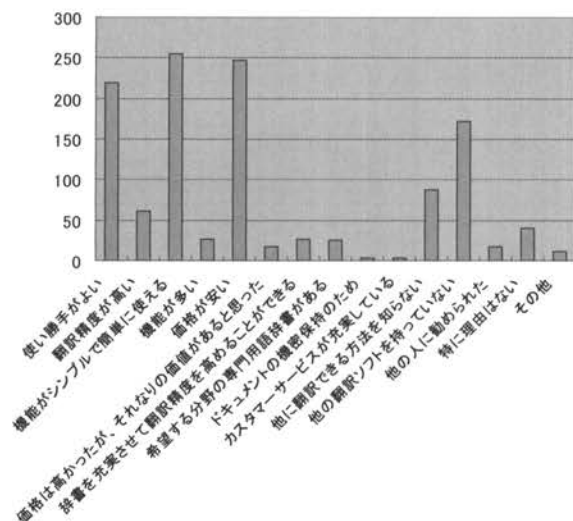


図 3-14 最も利用する機械翻訳ソフト・サービスを利用する理由 (複数選択可)

4. 利用者辞書の利用についての調査結果

本節では、利用者辞書の利用についての調査結果を示す。

図 4-1 に、利用者辞書に語 (文) を登録しているかどうかの割合、また、図 4-2 に利用者辞書に語 (文) を登録していない人についてその理由を聞いた結果を示す。無償翻訳サイトでは基本的に利用者辞書登録機能が提供されていないため、無償翻訳サイトの利用が多い状況では、全体としては利用者辞書への登録を行っている利用者は少なくなる。しかしながら、主に無償翻訳サイトを利用する人が 77% (「3.2 日英・英日機械翻訳ソフト・サービスの利用状況」参照) であることを考慮すると、登録している人が 18% に登るということは、利用者辞書登録は、その機能が使える場合には利用する人がかなり多いとも言える。

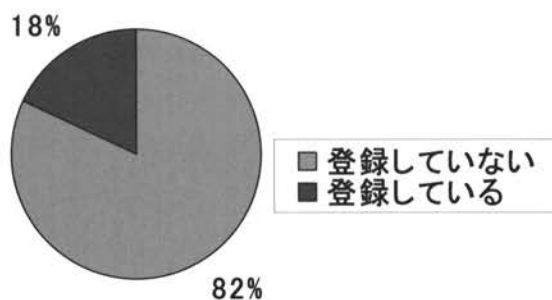


図 4-1 利用者辞書に語(文)を登録しているか

一方、利用者辞書を利用していない人にその理由を聞いた結果を示す。無償翻訳サイトの利用者の割合の多さを考えると「登録機能がない」という回答が圧倒的に多いことが予想される設問だが、「登録作業が分からない」「何を登録すればよいかわからない」「登録が面倒」という回答がかなり多い。機械翻訳において、利用者辞書への語 (文) 登録は、利用者側でできる数少ない翻訳品質向上策の一つだが、機能の使いやすさという点で工夫すべき点がある機能ということがわかる。

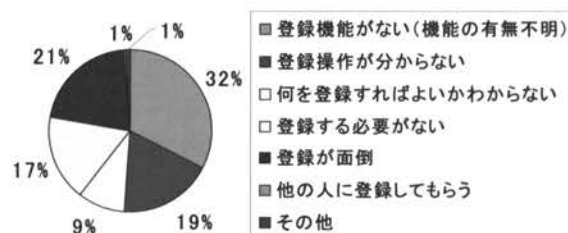


図 4-2 利用者辞書に語(文)を登録していない理由

次に、利用者辞書を利用している人はどのくらいの語 (文) 数を登録しているかを図 4-3 に示す。おおむね数十～数百程度の語数を登録しているユーザーが多い。なお、この設問の回答は、多い方では 1 万、2 万、また数十万という回答もあった。

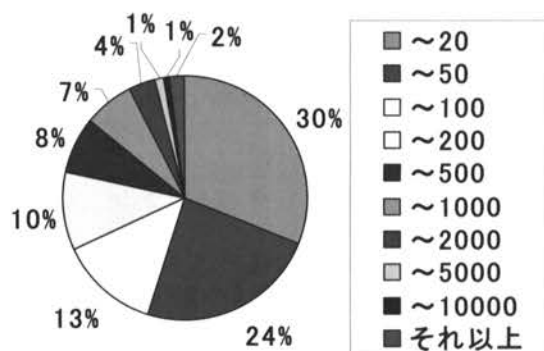


図 4-3 利用者辞書に登録している語(文)数

次に、図 4-4 と図 4-5 に、利用者辞書の利用と作業効率の向上の割合について示す。図 4-4 に示すように、利用者辞書に語 (文) 登録している人全体で見ると、4 割程度の作業向上が認められるというのが一般的なよう

である。また、利用者辞書の利用によって「かえって悪くなった」という回答は皆無であった。図 4-5 では利用者辞書への登録語（文）数別に作業効率の変化を示しているが、語（文）数別に見ると、特に登録語数の高いところで回答人数が少なくなってしまうため、傾向が捉えにくい。しかし、強いていうと、登録語数が 100、200 語近辺を超えるあたりから作業効率の効果が大きくなる傾向が多少はありそうである。

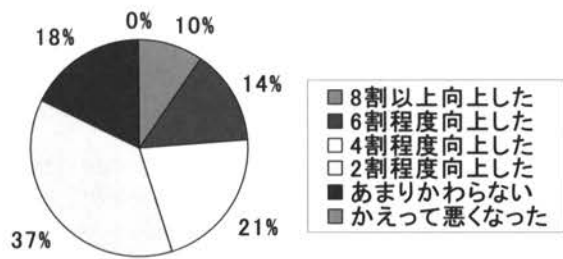


図 4-4 利用者辞書登録前と後での作業効率の変化

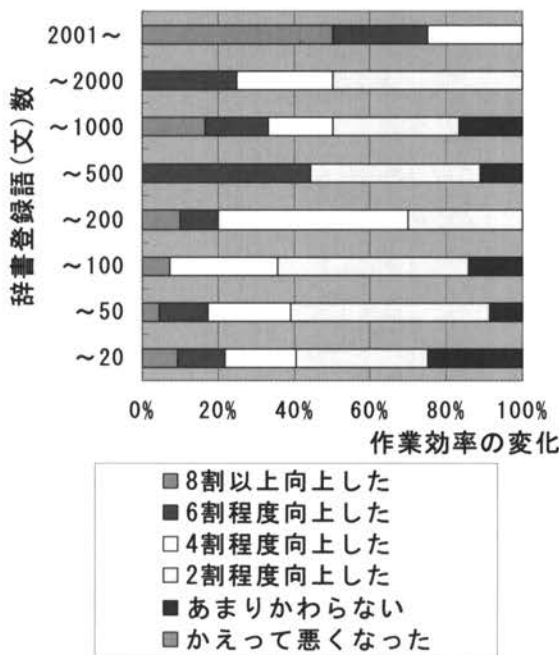


図 4-5 利用者辞書登録語(文)数と作業効率の変化

次に、図 4-6 に利用者辞書への登録語（文）数と翻訳精度に対する満足度との関係を示す。こちらは、登録者全体で見ても、登録していない人も含めた満足度とあまり変わらない。また、登録語（文）数と満足度の関係についても、やはり特に登録語数が多いところで回答人数

が少ないこともあり、はっきりした傾向がつかめなかった。

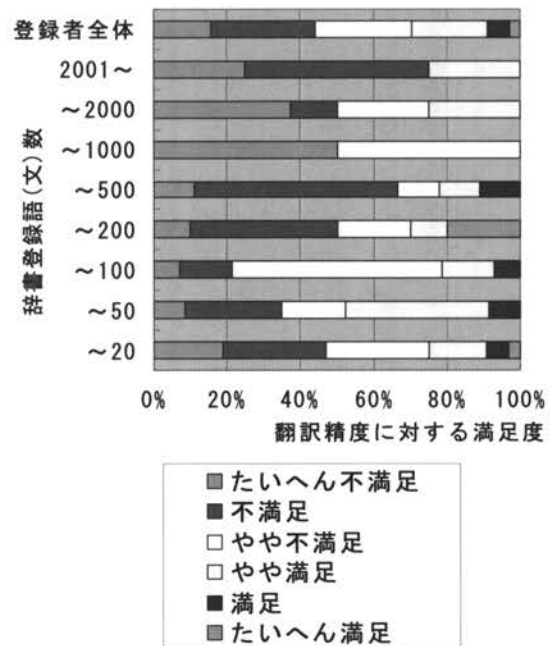


図 4-6 利用者辞書登録語(文)数と翻訳精度満足度

以上のことから、利用者辞書の利用についての傾向をまとめると、利用者辞書の利用によってある程度の作業効率向上が認められると言ってよい。ただし、それが、利用者の翻訳精度についての満足度にまでは反映されていないと言える。

5. 回答者に関する調査結果

本節では、回答者の基本属性（年齢、性別、職業、英語力）を示す。

図 5-1 に年齢、図 5-2 に性別、図 5-3 に職業、図 5-4 に英語能力（横軸は TOEIC スコア）毎の回答者の割合を示す。前回は携帯電話向けサイトでもアンケートの広報を行ったが、今回は行っていない。しかし、全体傾向として回答者の属性に大きな変化はなかった。

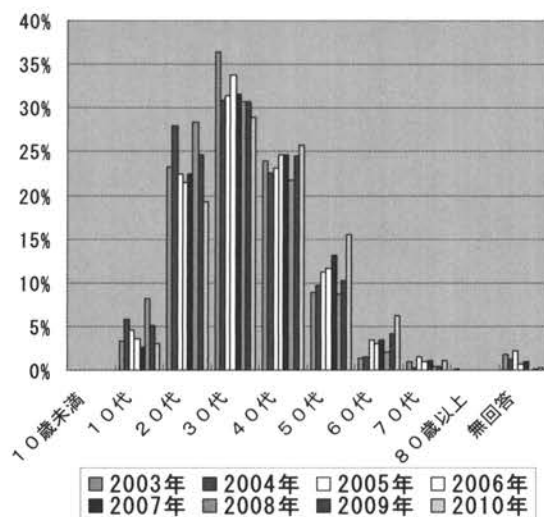


図 5-1 年齢

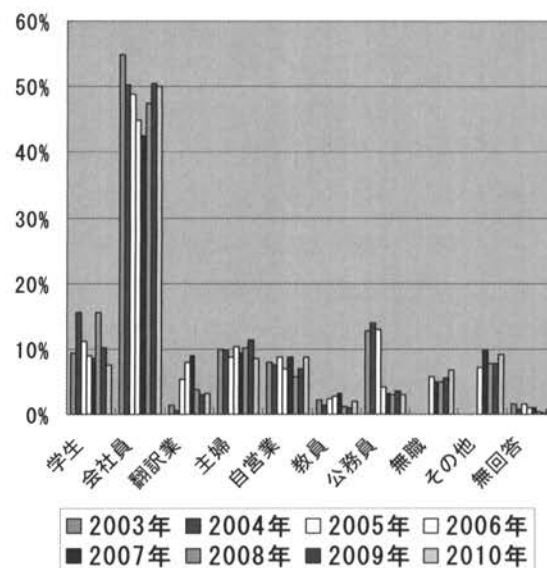


図 5-3 職業

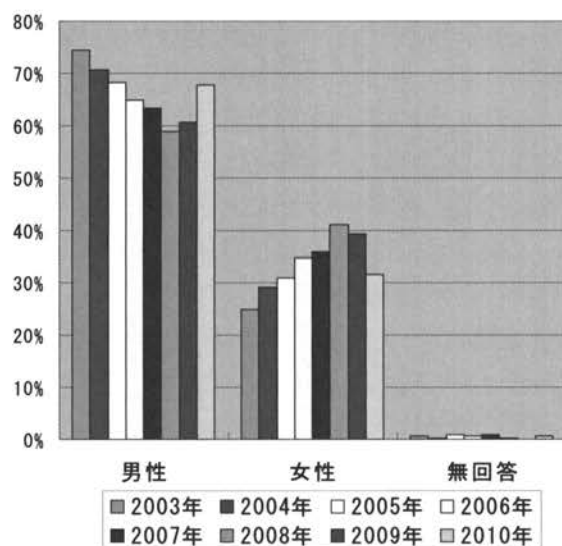


図 5-2 性別

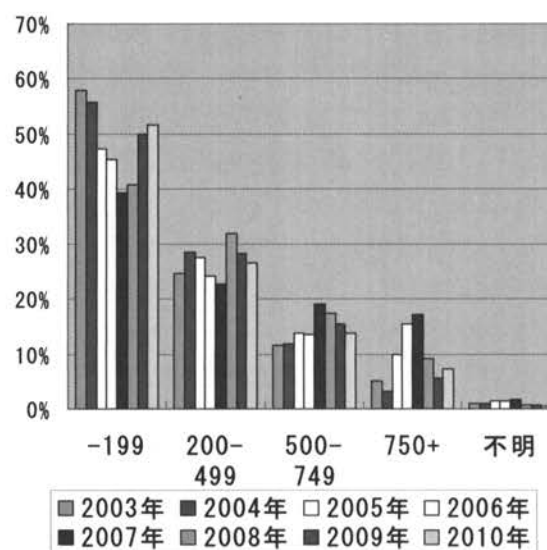


図 5-4 英語能力

6. まとめ

機械翻訳に関するアンケートを実施した。本年は新しい設問項目を追加し、下記のような新しい知見を得ることができた。

- 翻訳ソフト・サービスの利用は、無償翻訳サイトの利用が広く一般に浸透している。
- 翻訳ソフト・サービスについては、翻訳品質

など、不満が多々あるものの、仕事に、また、プライベートに、定常的に利用する人も多く、作業時間の短縮については一定の効果を認められている

- 利用者辞書への語（文）の登録については、作業効率向上の効果が認められていると言ってよい。ただし、無償翻訳サイトでは機能自体がないことが多い。また、使うにしても、利用者にとってのハードルの高さなどの面で課題があると思われる。

今後、機械翻訳研究者／ベンダーに上記のアンケートに基づいた研究・開発・販売を期待すると共に、今後も機械翻訳技術利用の現状やニーズに係る調査を続け、機械翻訳をより役立つものとしていきたい。

機械翻訳課題調査委員会ワーキンググループ 3 活動報告(2010/7-12)

機械翻訳課題調査委員会ワーキンググループ 3

2010 年 7 月 5 日（月）第 4 回会合

双方向識別子/term state 関連

（双方向識別子は、ある辞書を逆の翻訳方向にも使えることを示す符号）

山本委員、双方向識別子を“term state”として拡張することを提案。Banned/prohibited, new, approved（この状態が「双方向使用可能」に相当するか。継続議論が必要）、pending, alternative（双方向では使用できず、代替としてのみ含める語）などの term state が考えられる。禁止語は monolingual, bilingual（訳し方として不可）の両方でありえる。

例：差別語、過剰なカタカナ語、あいまいな語など、企業利用者が企業用語において、一括して使用を禁止する用語。（launch の訳語を「ローンチする」ではなく「発表する」とするなど。モノリンガルでは、「ローンチする」という語を禁止する、ということになる。

2010 年 8 月 6 日（金）AAMT 機械翻訳課題調査委員会第 4 回

TBX メンバーとの情報交換

・Bond 元委員を介して、TBX（用語データの XML ベースでオープンな標準。ISO と LISA により承認されている。

<http://www.lisa.org/TBX-Basic.926.0.html>）のメンバー 3 名から山本委員にコンタクトがあり、TBX チームと山本委員とで用語管理についてメールで議論のやりとりが続いている。山本委員が TBX チームに UTX の特徴と意義を TBX チームに説明。TBX と WG が当初構想していた UTX-XML とは重なりがあるため、WG3 では、UTX-S 1 つに絞り、XML 版は TBX を採用することを山本委員が提案。本日出席の他の委員も同意。→合同会議でのコメン

トあり。TBX は用語管理が主眼、UTX は機械翻訳が主眼。競合しないようにうまく連携を図ることが重要。

・TBX チームより、AMTA2010 (2010/10/31-11/5) に投稿する原稿(ドラフト)(10 ページ)が山本委員に送られ、内容は UTX-S を含む 5 種の用語フォーマットを相互に利用する仕組み。1 ページ分ほど、UTX に関する記述があり、WG3 は原稿のコメントを求められている。

・用語フォーマットの中で UTX は唯一ライセンスの考慮がされていることが原稿で評価されていた。

UTX における用語ステータス (term status)

次の 4 つの区分に決定

An entry can optionally have one of the four term statuses:

1. provisional
2. approved (=approved)
3. alternative(<alternative)
4. banned (>banned)

分かりやすさのため、2,3,4 については方向性を示す括弧内の表記を用いてもよい。

ソートするとどれがどれに対して banned なのか、alternative なのかが不明となるため、concept ID を設けることに決定。

<concept ID>

When there are multiple possible entries, use a “Concept ID.” A “Concept ID” is applicable to a source term. If a term should be banned entirely in a particular language, then that term should be defined as such in a monolingual language dictionary, rather than in a UTX dictionary.

- ・UTX という禁止語リスト相当が TBX にあるか。
- ・approved の term があれば機械翻訳の辞書情報

として十分。banned の term は機械翻訳以外の用語チェックの用途をもつ。これを仕様にも明記しておく必要がある。

<Concept ID の付与方法>

If a source has only one term status, Concept ID is not required. A Concept ID consists of unique serial number in a dictionary up to a limit of 10 digits. When multiple dictionaries are merged, the same concept IDs are distinguished by their dictionary IDs...

<Dictionary ID> (optional)

A dictionary ID is a unique identifier to a dictionary. It consists of four alphanumeric characters.

ID が重なった場合は AAMT の方で連番を管理するかについては継続審議。

2010 年 9 月 8 日 (水) AAMT 機械翻訳課題調査委員会 5 回議事録

・WG1・WG3 合同-プレゼンテーション「機械翻訳は使えるのか-Post-editing の観点から検証する」

山田優氏 (立教大学大学院異文化コミュニケーション研究科博士後期課程)

要旨：現在、機械翻訳に下訳をさせて、修正(=後編集)を行うワークフローが確立している。本発表では、後編集と人間による翻訳を比較したこれまでの実証研究 (いずれも日本語以外の実験) を概観し、日本語の場合をみるために、英日機械翻訳と人間翻訳を比較する実験を行い、修正量と被験者に対するアンケートをもとに、現状の機械翻訳のレベルを後編集の観点から分析した。

-委員からの質疑応答：エラーを埋め込んだ実験、被験者に事前に与える情報による違い、Philip Koehn(2009)の IWSLT での講演“Human translation and machine translation”の紹介、SDL の AutoSuggest Dictionary の紹介、専門用語のみ示した場合の影響、タイプライター上での翻訳作業とコンピューター上での翻訳作業の違いなど。

[1] UTX における term status と dictionary administrator

前回、term status について次の 4 つの区分 (provisional, approved, alternative, banned) を決定したが、(すでに英語圏で実際に使用されている) SDL の用語区分を参考に、banned は forbidden(上からの押し付けというニュアンスがない)、alternative は non-standard に変更。alternative は「場合によっては使ってもよい」と誤解される可能性があるが、non-standard とすれば、「誤って使われた原語」であるということが明確になる。

また、TBX-Basic: Data categories(http://www.lisa.org/fileadmin/standards/tbx_basic/TBX_Basic_datacategories.pdf)では、usage status を preferred, admitted, not recommended, obsolete の 4 つに区分。

日本語仕様では、provisional は「暫定」、approved は「承認」、non-standard は「非標準」、forbidden は「禁止」とする。

・dictionary administrator

4 つの term status について決定を下す dictionary administrator (辞書管理者) を設け、仕様に盛り込む。文案を練る。

本日時点の最終案：“While a dictionary may have multiple contributors, a dictionary administrator is ultimately responsible for a dictionary and defines the framework of a dictionary.”

Term status の説明を dictionary administrator に関連づける記述に変更する。

・上記の変更・追加を UTX の仕様書に反映する。

ボーイスカウト関連

加藤委員とは別途に山本委員がボーイスカウトの関係者と MT の使用について話をした。加藤委員より、現状について報告。

ボーイスカウトにはボーイスカウト固有のテクニカル用語が豊富にある。また、海外の団員とのコミュニケーションを支援するシステムを活用するために必要な用語 (掲示板で使う用語) もある。また、TM もコミュニケーションに有用。具体的には、

人が作った辞書がすでにあるので、それを UTX 形式にして、言語グリッドのようなプラットフォームに入れて、そのプラットフォームを使用できるようにするといったことが考えられる。

ニーズは存在するが、ボーイスカウトの中に ICT 推進に向けて積極的に動く人がいない。組織的な問題。加藤委員が働きかけを継続中。

・山本委員、スマートフォン用の自動翻訳ソフトを試用してみた。日中翻訳で従来のものと比較してある程度実用的な精度が得られるとの感想。

2010 年 10 月 13 日（水）AAMT 機械翻訳課題調査委員会 6 回議事録

[1] UTX-Simple における dictionary ID

- ・ dictionary ID

次のようにまとめた。

(Optional) A dictionary ID is a unique identifier for a dictionary. It may be required to distinguish dictionaries when multiple dictionaries are merged. If two dictionaries happen to be share entries with the same concept ID, related entries could be grouped into one group of entries. Unique dictionary IDs can help avoid such a situation. A dictionary ID is not mandatory. It can be added when multiple dictionaries are merged with IDs to distinguish a dictionary.

-Example:

	Entries	Concept ID
Dictionary A	outlet	76531
Dictionary B	instantiate	76531

With dictionary IDs, entries which happen to have the same concept ID can be distinguished even when the originating dictionaries are merged.

	Entries	Concept ID	Dictionary ID
Dictionary A	outlet	76531	AD64
Dictionary B	instantiate	76531	5d32

[2] Microsoft Terminology Collection

Microsoft の localization の用語が TBX 化された。

<http://www.microsoft.com/language/en-US/terminology.aspx>

2010 年 11 月 17 日（水）AAMT 機械翻訳課題調査委員会 7 回議事録

[1] UTX-S 仕様書（英語）検討

原稿の前回からの主な変更は構成（目次）：

・ 1. の下位項目として Background と Goal を既存の英語版 UTX ウェブページから追加。

・ 6. Header の下位項目を Dictionary information と Column Definitions とし、具体的な Column Definitions は 7. Body に記述する形をとった。

今回の主なポイント

- ・ 禁止語（7.4.4 forbidden）

禁止語は翻訳システム全体ではなく、特定の辞書のみ有効とする。

7.4 の "If a forbidden term is included innon-standard" を以下に差し替え。"It is possible that a term is forbidden in one dictionary, but if the same term exists in another dictionary, it may not be forbidden. In other words, when using a set of dictionaries in a translation tool, the forbidden status may differ."

- ・ 用語ステータスと逆翻訳方向の関係

承認された語以外は逆方向にすると用語ステータスが異なってくる。→逆方向で用語ステータスにより用語管理をするときは別の辞書を新たに作成してもらうことにする。ただし、承認された語のみは、そのまま逆翻訳方向でも使える。

2010 年 12 月 15 日（水）AAMT 機械翻訳課題調査委員会 8 回議事録

[1] JTF 翻訳祭報告(山本委員)

700 名ほどの参加があり、盛況。

・ 山本委員担当の分科会「平均時速 650 語の翻訳支援技術—機械翻訳を活かして効率と高品質を両立する秘密」

山本委員担当分科会で時間の 1/3 は UTF の紹介

を行った。この分科会の申し込み者は多かった。本日の委員会で山本委員が当日発表に使用したスライドを紹介 2011/1 中旬、JTF 翻訳環境研究会で、この分科会と同じテーマで、さらに具体的かつ詳細な発表を山本委員が行う予定。

- ・山本委員が聴衆として参加した分科会

- 「統計的機械翻訳の理論と実装」河野弘毅氏(アイタス シニアローカリゼーションスペシャリスト)

- 「ポストエディット入門 2010」斎藤玲子氏(日本オラクル、WPTG - Worldwide Product Translation Group - Japan, Senior Language Specialist)

[2]論文発表

1)2011/3/7-8 LISA Open Standards Summit 2011

(Washington D.C.)

<http://www.lisa.org/LISA-Open-Standards-Summit-2011.1628.0.html>→山本委員が参加予定

2)2011/3/7-11 言語処理学会年次大会(NLP2011)

(豊橋技術科学大学)→大倉委員が参加予定

発表内容

1)は翻訳祭での山本委員担当の分科会の UTF 関連部分を英訳したもの

2)は、タイトル「シンプルで相互運用性が高くオープンで専門分野にフォーカスした用語集形式

UTX-Simple による実地的な翻訳支援」

発表概要(案)を作成(12/28 締め切り)

[3]Trados デモ(山本委員)

以前バージョンの UTX の仕様(英語・日本語)を翻訳メモリ(TM)として使用して、SDL Trados Studio 上で UTX の V1.10 の英語仕様を翻訳する様子をデモ。SDL Trados Studio と連携する統計機械翻訳サービスとして SDL Trados Studio では現在、SDL 自動翻訳、Google 翻訳、Language Weaver の3種を提供(なお、これらを使用すると、秘密保持契約についての確認画面が表示される。実際の翻訳業務では、秘密保持に違反する可能性もある)。なお、山本委員は SDL Trados の公認講師。

SDL Trados Studio に付属する、MultiTerm という独立した用語管理ツールがある。用語の取り込みには別ツールの MultiTerm Convert を使用する。事前設定には XML の知識が前提で、使用方法も煩雑。その点、UTX はシンプルで使いやすい。この点を今後、UTX の利点として強調できる。

(議事録要約はここまで)

なお、次のページに、UTX ページ <<http://www.aamt.info/japanese/utx/>>でダウンロードできる、UTX 辞書のサンプルを掲載する。

さらに、統計機械翻訳とルールベース機械翻訳での UTX の効果の比較を紹介する。

UTX形式辞書サンプル 1/法令用語日英標準対訳辞書「(英日として使用可能)、抜粋

#UTX 1.11; en-US/ja-JP; 2011-05-13T01:37Z+09:00; copyright: Japan (2009); license: CC-BY-ND 3.0; bidirectional							
#description: This is a legal dictionary. You may use this dictionary only when you agreed that you and you alone are fully responsible for the results of its use. AAMT and its members do not guarantee the contents of this dictionary including, but not limited to its accuracy./ このデータは法律辞書です。このデータは使用者が使用結果に対して全責任を持つ場合にのみお使い頂けます。データの著作者、AAMTおよびAAMTのメンバーは正確性を含め、辞書の内容について一切の保証はいたしません。							
#src	tgt	src:pos	yomi	nuance	comment	example (Japanese)	example (English)
a case of violation	違反事件	noun	いはんじけん		【注】(出典)出入国管理及び難民認定法31条		
abandon	遺棄する	verb	いきする				
abandon	廃棄する	verb	はいきする				
abandonment	遺棄	noun	いき				
abandonment	廃棄	noun	はいき				
abatement	減殺	noun	げんさい				
ability	能力	noun	のうりよく	一般的な能力			
ability to participate in oral argument	弁論能力	noun	べんろんのうりよく				
ability to pay	支払能力	noun	しはらいのうりよく				
abolish	廃止する	verb	はいしする				
abolition	廃止	noun	はいし			認可事業者がその事業を廃止しようとするときは、	When the approved services operator intends to abolish the services,
abridged copy	抄本	noun	しょうほん				
absence from work	休業	noun	きゅうぎょう				
absence from work benefit	休業給付	noun	きゅうぎょうきゅうふ				
absentee	不在者	noun	ふざいしゃ				
absorption-type merger	吸収合併	noun	きゅうしゅうがっぺい				
absorption-type split	吸収分割	noun	きゅうしゅうぶんかつ				
abstract	要約	noun	ようやく				
abstract	要約書	noun	ようやくしょ				
abuse of right	権利の濫用	noun	けんりのらんよう				
academic research	学術研究	noun	がくじゅつけんきゅう				
accelerate	促進する	verb	そくしんする				
acceleration	期限の利益を喪失する	verb	きげんのりえきをそうしつする	名詞形(期限の利益の喪失)			
acceleration	促進	noun	そくしん	速度を上げる場合			
accept	受け入れる	verb	うけいれる				
accept	受理する	verb	じゅりする				
accept	譲り受ける	verb	ゆずりうける				
acceptance	受入れ	noun	うけいれ				
acceptance	受理	noun	じゅり	原則			
acceptance	承諾	noun	しょうだく				

UTX形式辞書サンプル2/医学用語集 (北里大学 医療情報学研究室)、抜粋

#UTX 1.11; en-US/ja-JP; 2011-05-13T01:37:00Z+09:00 copyright: Medical Informatics, School of Allied Health Sciences, Kitazato University (2009); license: CC-BY 3.0

#description: This is a medical dictionary. You may use this dictionary only when you agreed that you and you alone are fully responsible for the results of its use. The author of the original data, AAMT and its members do not guarantee the contents of this dictionary including, but not limited to its accuracy. Some part of speech properties are indicated as noun, even when they are not. / このデータは医学辞書です。このデータは使用者が使用結果に対して全責任を持つ場合にのみお使い頂けます。データの著作者、AAMTおよびAAMTのメンバーは正確性を含め、辞書の内容について一切の保証はいたしません。品詞が名詞でない場合も名詞扱いになっていることがあります。

#src	tgt	src:pos
1/2 T vector	1/2Tベクトル	noun
1/2FF	1/2拡張分画	noun
131I-hippurate	131I-ヒプル酸塩	noun
17 α-hydroxycorticosteroid	17αヒドロキシコルチコステロイド	noun
4-hydroxybutylic aciduria	4-ヒドロキシブチル酸尿症	noun
50% lethal dose	50%致死量	noun
5-aminolevulinic acid	δアミノレブリン酸	noun
A band	A帯	noun
A behavior pattern	A型行動様式	noun
A fiber	A線維	noun
A mode echocardiography	Aモード心エコー法	noun
A standard light	A標準光源	noun
a wave	a波	noun
abasia	失歩	noun
abdominal	腹部	noun

c	tgt	src:pos
ominal	腹部放射菌症	noun
inomycosis	腹部大動脈瘤	noun
ominal aortic	腹部浮球感	noun
urysm	腹部膨満	noun
ominal	腹式呼吸	noun
ottement	腹部血管雑音	noun
ominal bandage	腹帯	noun
ominal bloating	腹困	noun
ominal breathing	腹部けいれん	noun
ominal bruit	腹水	noun
ominal	腹膜炎顔貌	noun
umference	腹部型ホジキン病	noun
ominal cramp	腹筋反射	noun
ominal dropsy	腹腔穿刺	noun
ominal face	腹腔内圧	noun
ominal Hodgkin's		
ase		
ominal muscle		
ex		
ominal		
acentesis		
ominal pressure		

UTX形式辞書サンプル3/Computational Linguistics Term List (by Francis Bond)、抜粋

#UTX-S 0.91; en-US/ja-JP; 2008-05-21; copyright: Francis Bond (2008); license: CC-by 3.0

#src	tgt	src:pos
1-to-1 mapping	1対1の写像	-
AAP	照応的一致原理	-
AAVE	アフリカ系アメリカ人日常英語	-
ABF	平均分岐数	-
ADV	副詞	noun
ARP	項具現化原理	-
ATN parser	ATNプログラム	-
ATNG	拡張遷移ネットワーク文法	-
AVM	属性値行列	-
Across-the-Board exception	一律の例外	-
Act disallowed	行為は却下可能	-
Action Aider	行為補助部	-
Adjacent Feature Principle	隣接素性原理	-

	tgt	src:pos
Pair Grammar	アークペア文法	-
iment Realization	項具現化原理	-
ciple	アロー論理	-
w Logic	拡張遷移網	-
nented	オートマトン理論	-
osition Network	束縛継承原理	noun
omata Theory	ブール代数	-
	制約に基づく語彙主義	-
	文脈自由文法	-
	クロスリンガル情報検索	-
rasure	CP削除	-
	等位構造制約	-
i	文脈依存文法	-

UTX と統計機械翻訳 (Google Translator Toolkit <http://translate.google.com/toolkit>)

以下は、Google Translator Toolkit で UTX-Simple 1.10 仕様書英語版を機械翻訳した例である。UTX 形式から変換した用語集を指定できるものの、「辞書管理者」、「項目」など「使用すべき」と翻訳者が指定した用語が完全に無視されている。翻訳者は、目視して確認し、手作業で修正する必要がある、後編集（ポストエディット）に多大な労力を必要とする。全般的な翻訳精度も低く、まったく意味を成さない文が多い。

The screenshot shows the Google Translator Toolkit interface. The top bar indicates the project is 'UTX_Simple_1.10_English' and is 13% complete with 1214 words. The source text is in English, and the target text is in Japanese. The glossary section lists terms and their translations.

原文

While a dictionary may have multiple contributors, a dictionary administrator is ultimately responsible for a dictionary and defines the framework of a dictionary. The dictionary administrator would decide whether an entry is approved or not, as explained in "7.4.2 approved." If the dictionary is compiled by a single individual, the dictionary administrator is also the contributor.

0.2. UTX converter

UTX converter is a generic name for a tool that converts a UTX dictionary to/from other formats.

0.3. Translation direction

A UTX dictionary typically has a source language and a target language (monodirectional). A complete bidirectional dictionary is a dictionary where all terms are known to be used for the opposite translation direction. In this case, a UTX converter may convert the entire dictionary to a monodirectional dictionary whose translation direction is reversed. When only some entries have approved status as explained below, only these entries may be exported as a reversed dictionary to be used for the opposite translation direction.

Example

Consider a Japanese-English UTX dictionary. When some

翻訳: 英語 → 日本語 13% 完了 1214 ワード

複数の貢献者を持っている可能性があります辞書が、管理者が最終的には辞書の責任者であり、辞書のフレームワークを定義して辞書。

[0] 辞書に説明したように、あるいはエントリが承認されていないかどうか、管理者が決定すると「承認7.4.2。」 [1] [2] [2]

HTML タグを挿入

単一の個人によってコンパイルされている辞書の場合は、管理者は、貢献しています辞書。

0.2. UTXはコンバータ

UTXはコンバータはUTXは、他の形式へからの辞書に変換するツールの総称です。

0.3. 変換方向

UTXは、通常、ソース言語とターゲット言語(単方向)が辞書です。双方向の完全な辞書のすべての条件が逆の翻訳方向に使用されることが知られている場所を辞書です。この場合において、UTXはコンバータは、その翻訳方向を逆辞書は全体が単方向に辞書を変換することができます。逆は逆の翻訳方向に使用する辞書としてのいくつかのエントリは、以下に説明するようにステータスを承認をしたときは、これらのエントリだけがエクスポートされることがあります。

自動翻訳の検索 **カスタム翻訳の検索**

翻訳検索結果
過去訳はありません。

機械翻訳
辞書に説明したように、あるいはエントリが承認されていないかどうか、管理者が決定すると "承認7.4.2。"

提案を使用する

用語集 (4)

1. approved
承認 *adjective*
用語ステータスの一つ。
2. entry
項目 *noun*
3. dictionary
辞書 *noun*
4. dictionary administrator
辞書管理者 *noun*

UTX とルールベース機械翻訳

以下は、LogoVista 2008 の対訳エディターで UTX-Simple 1.10 仕様書英語版を機械翻訳した例である。前ページと同じ、(UTX 仕様を翻訳するための) UTX 形式の用語集をインポートした。「辞書管理者」、「項目」など、「使用すべき」と指定した用語は、翻訳結果に適切に反映されている。そのため、翻訳者やチェッカーは改めて用語適用を確認する必要がなく、後編集 (ポストエディット) は必要最低限で済む。また全般的な翻訳精度も高い。統計機械翻訳と比較すると、LogoVista 以外の翻訳ソフトでも、UTX を使う限り、同様の精度向上が望める。

青字は、ユーザー辞書として定義されている語である。ここでは、全体の半分程度を占めている。図では「フレームワーク」という語が選択されているため、対応する原語 **framework** とともに赤字で表示されている。このように語を選択した状態で、右クリックメニュー、もしくは右端に表示されている[別訳語]から別訳語を選択することで、不適切な用語は入れ替えることができる。また品詞の判定が誤っている場合は、語句を辞書登録するか、別解釈をさせて他の訳の可能性を確認することができる。さらに語句を辞書登録すると、次回以降は確実に誤りを減らすことができる。統計機械翻訳ではこのようなことはできない。

The screenshot displays the LogoVista 2008 translation editor interface. The main window is titled "対訳翻訳 - [UTX-AAMT dictionary sample.tad]". It features a menu bar with options like "ファイル(F)", "編集(E)", "表示(V)", "原文編集(P)", "翻訳(T)", "訳文編集(S)", "ツール(L)", "ウィンドウ(W)", and "ヘルプ(H)". Below the menu is a toolbar with various icons for editing and navigation.

The main content area is divided into three panes. The left pane shows the English text, the middle pane shows the Japanese translation, and the right pane shows a list of alternative translations for the selected word "framework".

Left Pane (English Text):

- 0. Definitions
- 0.1. Dictionary administrator

While a dictionary may have multiple contributors, a dictionary administrator is ultimately responsible for a dictionary and defines the framework of a dictionary.

The dictionary administrator would decide whether an entry is approved or not, as explained in "7.4.2 approved."

If the dictionary is compiled by a single individual, the dictionary administrator is also the contributor.
- 0.2. UTX converter

UTX converter is a generic name for a tool that converts a UTX dictionary to/from other formats.
- 0.3. Translation direction

A UTX dictionary typically has a source language and a target language (monodirectional).

A complete bidirectional dictionary is a dictionary where all terms are known to be used for the opposite translation direction.

In this case, a UTX converter may convert the entire dictionary to a monodirectional dictionary whose translation direction is reversed.

When only some entries have approved status as explained below, only these entries may be exported as a reversed

Middle Pane (Japanese Translation):

- 0. 定義
- 0.1. 辞書管理者

辞書が複数の寄稿者を持っていることがある間に、辞書管理者が最終的に辞書に対して責任があって、そして辞書のフレームワークを定義します。

項目が承認であるか否かにかかわらず、辞書管理者は、「7.4.2の承認」で説明されるように決定するでしょう。

もし辞書がコンパイルされているなら一人の個別、辞書の管理者に貢献者がまっています。
- 0.2. UTX変換ツール

UTX変換ツールはその他の形式に / からUTX辞書を変換するツールに汎用の名前です。
- 0.3. 翻訳方向

UTX辞書が一般的には起点言語と目標言語 (片方向) を持ちます。

辞書がすべての期日が既知である辞書である完了する双方向が反対の翻訳方向に使用されます。

この場合、UTX変換ツールがその翻訳方向性が反転される片方向辞書に全体の辞書を変換することがあります。

以下を説明されるように、ただ一部のエントリだけが承認状態を持っているとき、ただこれらのエントリだけが反対の翻訳方向に使用される反転された辞書としてエクスポートさ

Right Pane (Alternative Translations):

The right pane shows a list of alternative translations for the selected word "framework". The list is titled "別訳語" and includes the following entries:

別訳語	辞書
フレームワーク	S
機構	S
骨組み	S
枠組み	S

The word "枠組み" is currently selected, and a mouse cursor is pointing at it.

At the bottom of the window, a status bar indicates the current state: "待機中 訳文 3 / 13 文 翻訳済み 英語から日本語へ AAMT".

翻訳支援のためのシンプルでオープンな辞書仕様 UTX-Simple 1.10

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

※この記事は、第17回言語処理学会年次大会に投稿した論文をベースにしています。“UTX-Simple”の正式名称は論文投稿後、“UTX”に変更されました。

1. はじめに

機械翻訳を活用した翻訳支援の研究において、人間による実務的な翻訳の知見が十分に活かされているとはいえない。特に用語管理については、翻訳品質が問題となるプロ翻訳者による翻訳では必須といえるが、その研究は十分でない。

機械翻訳のルールベースと統計ベースの方式では、翻訳支援の方法論も異なる。ルールベース機械翻訳においては、ユーザー辞書に登録した用語を翻訳結果にほぼ反映することができるため、用語管理が必要な翻訳には向いている。一方、統計的機械翻訳においては、用語が確実に反映されないため、厳密な用語管理を必要とする翻訳には不向きである。この場合、後編集作業として用語の統一作業が発生する。いずれの機械翻訳の方式においても、用語管理が重要である。「機械翻訳では十分な訳質は得られない」という考えもあるが、用語管理を行うことで、機械翻訳精度の向上、そして翻訳支援による翻訳効率の向上が可能であると考えられる。

本稿では、用語管理の観点から AAMT（アジア太平洋機械翻訳協会）が策定した用語集形式 UTX-Simple 1.10 について説明する。UTX-Simple は翻訳現場での相互運用性が高く、オープンで、対象を専門用語に限定した仕様が特徴である。2010 年末に策定した UTX-Simple の新バージョン 1.10 では「暫定」「禁止」「承認」「非標準」の 4 つの用語ステータスを導入し、用語管理における実用性を高めた。過剰な情報をそぎ落とした良質の用語集は、機械翻訳の精度を効果的に上げるだけでなく、翻訳支援にも使える。今後、UTX-Simple は、ローカライゼーション、オープンソース、教育、行政、医療、法律などのさまざまな分野で活用が期待される。

2. UTX-Simple について

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[2]。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、2008 年にそのシンプルな形式である UTX-Simple 1.00 の仕様を策定し、公開した[2, 3, 4, 5]。これまでに、医学辞書、法律辞書、言語学辞書を UTX-Simple 1.00 形式に変換し、公開している[2]。

翻訳辞書の標準化された仕様としては、LISA(Localization Industry Standards Association) [6]の TBX(Term Base eXchange) [7]が挙げられるが、仕様が複雑であるため作成と管理の手間がかかり、広く普及していないのが現状である。より軽量で分かりやすい形式が求められている。AAMT は LISA と連携して、標準化をすすめている。

UTX-Simple の主な特徴は以下の 4 つである。

1. シンプル：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、使いやすく分かりやすい、シンプルで実地的な仕様とする。
2. オープン：UTX-Simple の仕様は公開されている。仕様に基いた辞書を誰でも自由に作成して使用できる。
3. 特定分野内で高密度：「専門用語」という観点から、辞書の分野を明確化し、「一語一義」とする。その分野内で、網羅的に密度の高い辞書を目指す。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。

4. 汎用的：既存のテキスト エディターや表計算ソフトで迅速に共有・再利用ができる。翻訳精度向上に欠かせない辞書の構築がより効率的にできるようになる。

この他、今回策定した UTX-Simple 1.10 から、一方の翻訳方向だけでなく、双方向にも対応している。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。特にオープンソースの翻訳は個人で個別に行っているために訳語の統一が難しい。UTX-Simple の導入により、共通の辞書を共有・再利用することで、翻訳作業の効率化が期待できる。

3. UTX-Simple 1.00 における課題

UTX-Simple 1.00 は、原則的に一方向の辞書フォーマットの仕様であり、例えば英日の辞書を日英として使うことは想定されていなかった。

また用語管理の観点からも課題があった。

UTX-Simple 1.00 においても、翻訳支援の要素として一語一義の原則があったが、登録せざるをえない訳語と正式な訳語が、同一概念に含まれることを示す手段がなかった。また、単語を管理する手段もなかった。例えば、複数の翻訳者から用語が提示された場合、どれが暫定的で、どれが最終的な用語かが判断できなかった。また、業務としての翻訳では、政治的要因、社会的要因、企業のブランドイメージ、用語統一などのさまざまな理由から「使ってはならない語」がある。このような語を適切に管理できないと、用語集としての価値は低くなる。

これらの用語管理に関する課題を解決するため、仕様の拡張が行われた。

4. UTX-Simple 1.10 の仕様策定

UTX-Simple 1.10 での最大の向上点は、「用語ステータス(term status)」を導入し、用語管理を通じた翻訳支援での実用性を高めたことである。これは必須項目ではない。

UTX-Simple の機能向上にあたり、当初は、英日辞書を日英辞書としても使用するために、各項目にフラグを付けることが提案された。後に、この目的は、用語管理の機能に含められることが分かった。

用語管理のために、UTX-Simple 1.10 では、「暫定」「禁止」「承認」「非標準」の4つの用語ステータスを導入した。それぞれの用語ステータスを持つ語を、暫定語、禁止語、承認語、非標準語と呼ぶ。また、辞書管理の観点から、辞書管理者と用語提出者の概念を導入した。辞書管理者は、辞書の責任者であり、辞書の枠組みを定義する。用語提出者は、新しい用語を辞書に追加する。辞書管理者は、用語提出者が追加する用語が適切かどうかを判断し、用語ステータスを決定する。用語提出者が1人の場合はその人が辞書管理者となる。

- **暫定(provisional)**：その項目がまだ辞書管理者（後述）によって承認されていないことを意味する。辞書管理者は、可能な限り、暫定語を以下で説明する「禁止」「承認」「非標準」のいずれかに決定するか、削除することが望ましい。
- **禁止(forbidden)**：用語管理の観点から、項目の中に使ってはならない訳語が含まれていることを意味する。翻訳ツールが複数の辞書間の優先順位を適切に扱わない場合は、異なる分野別辞書との競合を避けるため、目標言語の用語を抑制する必要があることもある。例えば ICT の文脈では、英語の用語である“window”は日本語の単語である「窓」と翻訳されることは極めて少ない。この「窓」という訳語を機械翻訳システムが適切に扱えないのであれば、この訳語は明示的に抑制する必要があるだろう。禁止用語は、翻訳ツールの外で用語チェックに使用するために抽出することができる。
- **承認(approved)**：項目が辞書管理者により承認されていることを意味する。その用語が必ず使われなければならないことを示す。何か明らかな理由がある場合、起点言語の1つの用語に対して複数の目標言語の用語を割り当てることは可能であるが、用語ステータスが承認となるのはそのうちの1項目だけである。承認語は常に双方向性をもつ。つまり、辞書に定義されている翻訳方向と反対方向にして使える。同一概念（後述）に属する複数の項目が存在する場合、翻訳方向を逆にしたときに、唯一有効な項目である。

- **非標準(non-standard)**：1つの項目に対し、1つ以上の非標準の起点言語の用語があることを意味する。非標準用語は、単に起点言語の用語の異形に対応するためにのみ許可される。非標準用語は目標言語の用語としては使われてはならない。UTX 辞書が翻訳のためではなく、文書のオーサリングのための用語集として使用される場合、非標準用語は用語として使ってはならない。非標準用語が辞書に登録されているのは、起点言語の文書の筆者が承認されていない不適切な語を使用した場合でも、自動翻訳が翻訳できるようにするためにすぎないからである。

複数の訳語を管理する観点から、概念 ID、辞書 ID も定義した。

- **概念 ID (concept ID)**：同じ概念を共有する複数の項目を、10 桁までの数値で表す。つまり、単語の訳が複数ある場合、概念 ID が必要となる。概念 ID は辞書内で他と重ならない通し番号からなる。表 1 は、用語ステータスと概念 ID の例である。

表 1 用語ステータスと概念 ID

- **辞書 ID (dictionary ID)**：辞書のための

src	tgt	term status	concept ID
outlet	コンセント	approved	73
outlet	アウトレット	forbidden	73
power point	コンセント	non-standard	73
PowerPoint	PowerPoint	approved	
outlet store	アウトレットストア	approved	
plugin	プラグイン	approved	245
plug-in	プラグイン	non-standard	245

一意的な識別子で、4 文字の半角英数字（大文字・小文字を区別しない）で辞書管理者が定義する。複数の辞書を統合するとき必要である。2つの辞

書に同じ概念 ID を持つ項目がある場合に区別できる。

例えば、辞書 ID がないと、違う概念を表すが概念 ID が同じ単語が統合されてしまう（表 2）。この場合、辞書 ID があれば、概念 ID が同じ項目があっても区別できる（表 3）。

表 2 同じ概念 ID が違う辞書に存在する例

	Entries	Concept ID
Dictionary A	outlet	76531
Dictionary B	instantiate	76531

表 3 辞書 ID の例

	Entries	Concept ID	Dictionary ID
Dictionary A	outlet	76531	AD64
Dictionary B	instantiate	76531	5d32

5. 今後の展望と課題

UTX-Simple の最大の利点のひとつは、編集が簡単にできることである。高機能な用語集形式は、複雑で扱いにくく、翻訳現場での翻訳知識を効率よく収集することができない。今後、UTX-Simple と既存の用語データベースを自動で同期できる仕組みが開発されれば、翻訳の内容に即した多数の用語を、翻訳の現場から収集して、自動翻訳の精度向上に活かすことができる。UTX-Simple と各種用語集形式との変換ツールの開発も期待される。

アジア太平洋機械翻訳協会の共有化・標準化ワーキンググループでは、オープンな辞書コンテンツの収集と公開も行っている。すでに公開中の医学、法律、計算言語学に加え、その他の分野の辞書も追加したい。複数の用語提出者からの用語を、辞書管理者が管理する仕組みができたことから、CGM（消費者生成メディア）的手法による辞書作成が現実的になった。辞書をオンラインコミュニティで共同作成する場合は、動機付けまたはインセンティブの問題がある。有効な動機付けまたはインセンティブの仕組みの研究は、今後の辞書コンテンツの充実に必須であるだろう。

用語集作成での最善慣行の周知も今後の課題である。現状では、英語と比較して、日本語の標準的な表記にはばらつきが多い。テクニカルライティングの訓練も不足しており、難解な書

き方がむしろ奨励されていることもある。論理的で明快な書き方は、日本語として読みやすいのはもちろんだが、機械翻訳の精度も大きく向上させる。原文としての日本語作文の最善慣行には、日本語表記の整理と統一も含まれる。例えば、日本翻訳連盟では、翻訳における表記の整理が検討されている[8]。

UTX の次バージョンでは、用語ステータスが適切に機能しているかを検証し、翻訳分野の管理、多言語対応などについて検討する。実際の翻訳プロジェクトでの使用実績を積んで、フィードバックを反映させていきたい。

参考文献

- [1] <http://aamt.info/>
- [2] <http://www.aamt.info/japanese/utx/>
- [3] 大倉清司他(2008)「共有ユーザー辞書仕様 UTX の現状と今後の展開」言語処理学会第 13 回年次大会 (東京)
- [4] Francis Bond et al. (2009) “Sharing User Dictionaries Across Multiple Systems with UTX-S” in Second International Workshop on Intercultural Collaboration (IWIC2009), Stanford

[5] <http://www.aamt.info/japanese/utx/utx-simple-1.00-specification-j.pdf>

[6] <http://www.lisa.org/>

[7] <http://www.lisa.org/standards/tbx/>

[8] 日本翻訳連盟の表記検討フォーラム SINAPS
<http://jtf-forum.jp/>

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からお気軽にご連絡ください。

UTX についての説明 URL:

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

UTX メーリングリストについて:

<http://groups.yahoo.co.jp/group/UTX/>
(どなたでも参加できます)

UTX-Simple 仕様 Version 1.10

目次

0.	はじめに	65
0.1.	背景.....	65
0.2.	目的.....	65
1.	対象ユーザー	66
2.	対象分野	66
3.	定義.....	66
3.1.	辞書管理者	66
3.2.	UTX変換ツール	66
3.3.	翻訳方向	67
3.4.	用語.....	67
3.5.	項目	67
4.	UTX-Sファイル	68
5.	ヘッダー.....	68
5.1.	辞書情報	68
- 5.1.1.	ライセンス（省略可能）	68
- 5.1.2.	双方向（省略可能）	69
- 5.1.3.	辞書ID（省略可能）	69
5.2.	列定義	70
6.	本文.....	70
6.1.	起点用語（必須）.....	70
6.2.	目標用語（必須）.....	70
6.3.	品詞（必須）.....	70
6.4.	用語ステータス（省略可能）.....	70
- 6.4.1.	暫定	70

- 6.4.2.	承認	71
- 6.4.3.	非標準	71
- 6.4.4.	禁止	71
6.5.	概念ID(省略可能)	72
6.6.	その他の省略可能な列	74
7.	望ましいMTの機能	75
8.	UTX-Simple作成ガイドライン	76
8.1.	一般的なガイドライン	76
8.2.	英語特有のガイドライン	76
8.3.	日本語特有のガイドライン	76
9.	改版履歴	77

文書情報

作成者/所有者：AAMT 共有化・標準化ワーキンググループ

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

状態：最終版

更新日付：2011年4月15日

文書の言語：日本語

All Rights Reserved, Copyright (C) AAMT, 1996-2011

0. はじめに

0.1. 背景

CAT（翻訳支援）ワークフローでルールベースの翻訳ソフトウェアを使用するとき、原文文書の専門用語、人名、地名は、基本的なシステム辞書には一般的に含まれておらず、期待するほど適切には翻訳されない。しかし、用語が吟味され、特定の分野で適切であれば、追加の詳細情報がなくても、中核的な用語情報（原語および訳語）のみで、機械翻訳の適切性と正確さを向上するには十分であることは広く知られている。だが、ユーザーによって作成された辞書（ユーザー辞書）は、異なるMTシステム間でしばしば互換性がなく、辞書の労力が報われないことが多い。この問題に対処するため、AAMT（アジア太平洋機械翻訳協会）www.aamt.infoは、異なるMTシステム間で使用できる共有可能な辞書に関する、一連の仕様を策定した。1995年に、AAMTは、IPAの支援を受けてUPF (Universal PlatForm) と呼ばれる最初の仕様を策定した。その後、機械翻訳の技術や利用方法のさまざまな変化を反映するために、2006年から新しい仕様の策定を開始した。新しい仕様は、2007年8月に「UTX (Universal Terminology eXchange)」と呼ぶことが正式に決定された。UTXはオープン標準であり、仕様を利用するにあたってAAMTが課金することはない。2009年に、AAMTは、単純なタブ区切り形式のUTXであるUTX-Simpleを策定した。より複雑な機能は、既存の用語集の仕様であるTBXとTBX-Basicで実現できることが分かったため、AAMTはUTX-Simpleの開発に重点を置くようになった。

また、AAMTは、専門分野のオープンなユーザー辞書データの作成と収集も行う。また、持続可能な方法で、ユーザー辞書を生成、共有、蓄積するユーザー コミュニティーを作成することを計画している。

0.2. 目的

UTX-Simple（以下、UTX-S）の目的は、ユーザーの立場から、簡単に作成でき、簡単に使える機械翻訳用辞書の形式を提供することである。UTX-Sは、高度な管理性や可逆変換よりも、使いやすさと簡潔さを重視している。

同じUTX-S辞書を、さまざまなメーカーの翻訳ソフトで利用できる。さらに、UTX-S辞書は機械的処理に適した形式でありながら、人間にとっても読みやすい形式であるため、翻訳ソフト用途以外の一般的な用語集として使うこともできる。

翻訳ソフトのユーザーがユーザー辞書を作成するとき、個人として各自、分散して辞書を作成するのでは効率的でない。また、辞書形式が標準化されていないと、単純なプレーンテキスト ファイルでさえも、共有したり再利用したりすることが難しい。しかし、UTX-S形式が採用されれば、違うメーカーから提供される翻訳ソフトなどさまざまなツールで共有辞

書が広く使われるようになり、再利用性が非常に高くなる。

UTX-Sは、たとえば、複数の情報源から用語集をすばやく作成する場合や、さまざまなアプリケーション間で使用される用語集を配布し再利用する場合に適切である。より詳細かつ長期の用語管理では、TBXが適切である。

1. 対象ユーザー

UTX-Sは、特に、翻訳ソフトのエンド ユーザーあるいは翻訳者向けに設計されている。UTX-S辞書の作成、編集および使用にあたり、言語学、文法、XML、翻訳ソフトなどの特別な知識は必要ない。UTX-S辞書は、起点言語（原文言語）と目標言語（訳文言語）についての最小限の知識があれば作成することができる。

2. 対象分野

UTX-S辞書はどの翻訳分野でも作成し、使用できるが、たとえばICT、医学、法律、エンジニアリングなど、専門性の高い分野で特に効果が高い。理想的には、「ICT分野の中のRuby（スクリプト言語）辞書」、「医学分野の中の心臓外科辞書」などのように、それぞれの専門分野の中でも特定の領域ごとに個別の辞書を作成することで効果が高まる。1つの辞書に明らかに異なる複数の分野が含まれる場合、それぞれの分野について個別の辞書を作成する必要がある。こうすることで、辞書を管理することが簡単になり、また、その他の辞書と組み合わせることによって、容易に再利用することができる。

UTX-Sは、専門性がない一般的・汎用的な内容の翻訳には適していない。このような一般的な内容で使用した場合、UTX-Sの期待される効果は限られる。UTX-Sのフレームワークは、使用するMTシステムが充実したシステム辞書を持っていることを想定している。UTX-Sは、既存のシステム辞書のみでは不可能なほど、翻訳の適切性を向上することができる。非専門用語は、それらが分野内で特定の意味と訳語を持つときのみ、UTX辞書に含めることができる。

3. 定義

3.1. 辞書管理者

1つの辞書に対して複数の用語提出者がいる可能性はあるが、最終的には辞書管理者が辞書の責任者であり、辞書の枠組みを定義する。辞書管理者は、「6.4.2 承認」で説明されるように、ある項目がその辞書用語として承認されるか否かを決定する。辞書が個人によって作成される場合、辞書管理者と用語提出者は同一となる。

3.2. UTX変換ツール

UTX変換ツールとは、UTX辞書をその他のファイル形式に変換、あるいはその他のファイル

形式をUTX辞書に変換するツールの総称である。

3.3. 翻訳方向

UTX辞書は、基本的には、1つの起点言語と1つの目標言語を持つ（すなわち翻訳方向は片方向となる）。完全な双方向辞書とは、すべての用語が逆の翻訳方向でも使用できることが確認されている辞書である。この場合、UTX変換ツールは、その辞書全体を、翻訳方向を逆にした片方向辞書としてエクスポートすることができる。以下で説明する「承認」ステータスが、すべての項目にではなく、一部の項目にのみある場合、それらの項目だけを逆の翻訳方向の辞書としてエクスポートすることができる。

例

日英のUTX辞書を例とする。いくつかの項目が「承認」ステータスを持つとき、適切なUTX変換ツールを使うことで、これらの項目を英日辞書としてエクスポートすることができる。

3.4. 用語

用語とは、起点あるいは目標言語の見出し語である。用語は、辞書の見出し語のような単語の基本形（レンマ）である必要がある。「8 UTX-Simple作成ガイドライン」を参照。UTXの観点からは、用語とは、ある特定分野の専門用語である。それに対して、単語とは、UTX辞書に必ずしも含まれない、概念を表す文法的な単位を指す。

3.5. 項目

UTX辞書では、項目は1行の論理行で表される。UTX辞書は、「一語一義」の原則に従う必要がある。すなわち、辞書は「1つの用語は、その辞書の特定分野においてはただ1つの意味しか持たないということである。したがって、1つの原語は、理想的には1つの訳語のみを持ち、これが1つの項目（1行）を構成する。訳語が複数必要である場合（つまり複数の意味がある場合）、なぜそれらが複数必要なのか、その根拠を明確に示す必要がある。これにより、体系的な方法で用語集を作成することが促進される。複数の訳語が必要な場合は、同じ概念を示すために概念IDを指定することができる（6.5 概念ID（省略可能）を参照）。

UTXは、組織的なオーサリングおよび翻訳のプロジェクトで使用される場合に、最も効果的である。テクニカルライターのような、起点言語の文書の作成者は、スタイルガイドに沿って一貫した用語を使うよう務めていることが前提となる。UTXが組織的でないプロジェクトで使われる場合、その効果は限定される。

4. UTX-Sファイル

UTX-Sファイルの文字コードは、UTF-8（BOMなし）である。改行コードは、“`¥r¥n`”（CR+LF）である。拡張子は、“`.utx`”である。

UTX-Sファイルは、以下の要素から構成される。

1. 辞書情報。「5.1 辞書情報」を参照。
2. 列定義。「5.2 列定義」を参照。
3. 項目から構成される本文（タブ区切りテキスト）。「6 本文」を参照。

1と2から、ヘッダーが構成される。

項目は、行頭に“`#`”を記述することによりコメントアウトすることができる。

5. ヘッダー

UTX-Sヘッダーは、“`#`”から始まる必須の2行によって構成される。辞書の説明のために、必須の2行の間に、省略可能な行として、1行以上のコメント行を追加することができる。

5.1. 辞書情報

ヘッダーの1行目はUTX-Sファイルについての必要な情報を、セミコロンを区切り記号として次のように記述する。

```
#UTX-S <version>; < source language >/< target language>; <date created>;  
<creator>; <license>; <bidirectional (optional)>; <dictionary ID (optional)>;  
<other optional fields>
```

- 起点言語/目標言語：ISO 639、3166に準拠
- 作成日付：ISO 8601に準拠

例：

```
#UTX-S 1.10; en-US/ja-JP; 2011-04-10T14:28:00Z+09:00; comment: This is an example  
of UTX-S header.
```

5.1.1. ライセンス（省略可能）

辞書のライセンスは、クリエイティブ コモンズ、パブリック ドメイン、あるいは他のライセンスの形式で宣言することができる。ユーザーが辞書をどのように共有し、使用することができるかを明確にすることが強く奨励される。

5.1.2. 双方向（省略可能）

辞書中のすべての用語を逆方向でも使うことができる場合は、辞書のヘッダーに「双方向」フラグを含めることができる。この場合、すべての用語が承認されたとみなされるため、個々の用語ステータスは不要である。

例：

```
#UTX-S 1.10; en-US/fr-FR; 2011-04-15T10:00:00Z+09:00; copyright: AAMT (2011);  
license: CC-by 3.0; bidirectional
```

5.1.3. 辞書ID（省略可能）

辞書IDは、辞書の一意な識別子である。4文字の英数字（大文字小文字の区別なし）からなり、辞書管理者が決定する。複数の辞書を統合するとき、個々の辞書を区別するために必要となる。2つの辞書が同じ概念ID（6.5 概念ID（省略可能）を参照）の項目を持つ場合、関係のない項目が1グループにまとめられてしまうことが起こりうる。一意な辞書IDにより、このような状況を回避することができる。辞書IDは、必須ではない。複数の辞書を統合する必要が発生したときに、辞書IDを追加することができる。

例：

統合されたとき、辞書を区別するIDがないと、関係のない項目がグループ化されてしまうことがある。

	項目	概念ID
辞書A	outlet	76531
辞書B	instantiate	76531

辞書IDを使用すると、辞書が統合されても、同じ概念IDを持つ項目を区別することができる。

	項目	概念ID	辞書ID
辞書A	outlet	76531	AD64
辞書B	instantiate	76531	5d32

5.2. 列定義

UTX-Sヘッダーの2行目、あるいは、追加の説明行がある場合はヘッダーの最後の行は、列の定義を含む（いずれの行も“#”で始まる）。列定義は、タブで区切られ、3つの必須の列に加えて、ユーザー定義の列を必要に応じて追加できる。列定義と本文は密接に関連するため、詳細は「6 本文」で説明する。

6. 本文

UTXの本文は、1項目ごとの1論理行の集合から構成される。

6.1. 起点用語（必須）

1列目のsrcは、原語（起点言語の用語）を含む。

6.2. 目標用語（必須）

2列目のtgtは、訳語（目標言語の用語）を含む。単言語辞書の場合、目標用語は空白にする。

6.3. 品詞（必須）

3列目の、src:posは原語の品詞を含む。

UTX-Sでは次の品詞が定義されている。

noun / properNoun / verb / adjective / adverb / sentence

品詞が不明な場合は、空白にする。

sentenceは必要な場合のみに使う。文のペアの項目は、辞書ではなく翻訳メモリーに入れる必要がある。

6.4. 用語ステータス（省略可能）

項目は、用語の状態を示すために、4つの用語ステータス（暫定、承認、非標準、禁止）のうちの1つを持つことができる（省略可能）。用語ステータス（term status）は、辞書の主な翻訳方向についてのみ管理される。逆方向のために用語ステータスを管理する必要がある場合は、別の辞書を作成する。

6.4.1. 暫定

用語ステータス「暫定（provisional）」は、用語提出者によりその項目が追加されたが、

辞書管理者によってまだ確認されていないことを意味する。暫定の用語ステータスは一時的なものであり、辞書管理者は、用語が以下に示す「承認」「非標準」「禁止」のどれに該当するかすみやかに決定する必要がある。辞書管理者は、必要に応じて、その用語を辞書から除外（削除）することもできる。

6.4.2. 承認

用語ステータス「承認（approved）」は、項目が辞書管理者により承認されていることを意味する。承認ステータスはその用語が必ず使われなければならないことを示す。その根拠はさまざまだが、その用語が、特定の分野における専門用語であったり、ある組織の用語集であったりする場合である。その用語の単語表記に異形がある場合（たとえば“plug-in”と“plugin”）、その中で承認済みの表記のみが使われるべきである。

明らかな理由がある場合、原語に対して複数の訳語を割り当てることは可能であるが、用語ステータスが承認となるのはそのうちの1項目だけである。

用語ステータスが「承認」であるとき、用語を「逆」にし、辞書に定義されている翻訳方向と反対方向にして使うことができる（5.1.2 双方向（省略可能）を参照）。

承認語は常に双方向である。同一の起点言語に対して複数の訳語があり、それらを逆方向にして使う場合、承認語は、それらの項目のうち、唯一有効な用語になる。

用語提出者が1名である場合、（同時にその辞書の辞書管理者でもある）その用語提出者は、辞書に項目を追加した直後に承認ステータスを割り当ててもよい。あるいは、その用語提出者は、該当する翻訳プロジェクトにおいてその新規の項目が適切に翻訳されることを確認できるまで、用語ステータスを空白にしておくか、暫定ステータスを割り当てることができる。

6.4.3. 非標準

用語ステータス「非標準（non-standard）」は、非標準の原語であることを意味する。非標準語は、単に原語の異形に対応するためにのみ許可される。非標準語は、訳語としては使われてはならない。UTX辞書が翻訳のためではなく、文書のオーサリングのための用語集として使用される場合、非標準語は用語として使ってはならない。非標準語が辞書に登録されているのは、起点言語の文書の作成者が承認されていない不適切な語を使用した場合でも、機械翻訳システムが翻訳できるようにするためにすぎないからである。

6.4.4. 禁止

用語ステータス「禁止（forbidden）」は、項目に使ってはならない訳語が含まれているこ

とを意味する。そのような語は、言語的、社会的、用語的、その他の観点や、企業のブランドイメージの観点から明示的に禁止される。翻訳ツールが複数の辞書間の優先順位を適切に扱わない場合は、異なる分野別辞書との競合を避けるため、訳語を抑制する必要があるが生じることもある。

例

ICTの文脈では、英語の“window”は、たとえば日本語の「窓」と翻訳されることは極めて少ない。この「窓」という訳語を、機械翻訳システムが適切に扱えないのであれば、この訳語は明示的に抑制する必要があることがある。

禁止語は、逆方向の項目では承認ステータスとなることがある。また、非標準語は、逆方向の項目では禁止ステータスになることがある（たとえば概念IDが同一である場合）。

翻訳ツールは、禁止語の使用を抑制できることが望ましい。用語が、ある辞書で禁止されていて、別の辞書では禁止されていないということもある。言い換えると、翻訳ツールにおいて複数の辞書を使う場合、禁止のステータスは辞書間で異なることがある。翻訳ツールやUTX変換ツールは、そのような競合を検出する仕組みを持つことが望ましい。

禁止語は、翻訳ツールの外で用語チェックに使用するために抽出することができる。

6.5. 概念ID（省略可能）

異なる用語ステータスを持つ複数の項目がある場合、それらが同じ概念を共有していることを示すために「概念ID」を使用する。1つの原語に対し、1つの訳語しかない場合、概念IDは不要である。

概念IDは、同一辞書内で一意の通し番号からなる。概念IDは、10桁までの整数値である。複数の辞書を統合するとき、同一の概念IDを持つ項目はその辞書IDで区別できる。

例（英語から日本語）：

(用語番号)	src	tgt	term status	concept ID
1	outlet	コンセント	approved	73
2	outlet	アウトレット	forbidden	73
3	power point	コンセント	non-standard	73
4	PowerPoint	PowerPoint	approved	
5	outlet store	アウトレット スト ア	approved	
6	plugin	プラグイン	approved	245
7	plug-in	プラグイン	non-standard	245

（ここでの用語番号は単に説明の目的で使われている。これらはUTX辞書には存在しない。またUTX辞書では品詞が必須であるが、ここには挙げていない。）

上記の例では、用語番号1、2、3は同じ概念を指している。用語番号6と7に関しても同じである。4と5は、「一語一義」の原則に従い、概念IDの欄が空白になっている。したがって、区別が必要な項目は他に存在しない。なお、“outlet”の訳語としては「アウトレット」は禁止されているが、「アウトレット ストア」という語の一部としては承認されていることに注意されたい。また、複数の項目が同一の概念IDを共有している場合、承認用語は1つしかないことにも注意されたい。その他の用語はすべて禁止か非標準のいずれかとなる。

例（日本語から英語）

src	tgt	term status	concept ID
コンセント	outlet	approved	73
アウトレット	outlet	approved	98
コンセント	power point	forbidden	73
PowerPoint	PowerPoint	approved	
アウトレット ストア	outlet store	approved	

6.6. その他の省略可能な列

5つ目以降の列もすべて省略可能である。ユーザーは必要に応じて情報を定義することができる。単語の活用について、追加の列は以下のように記述する。

- 原語の活用形を定義する場合：src:<活用形>
- 訳語の活用形を定義する場合：tgt:<活用形>

英語用には、<活用形>の種類として次の表記があらかじめ定義されている。

- plural：複数形
- 3sp：3人称単数形
- past：過去形
- presp：現在分詞形
- pastp：過去分詞形
- comparative：比較級
- superlative：最上級

半角ハイフン“-”は、単語に対するその情報がないことを明示的に表す。たとえば、英語の単語“information”には複数形がないため、これを明示的に示す場合は“-”を使う。

あらかじめ定義されている記号

```
#src tgt src:pos src:plural src:3sp src:past src:pastp src:presp src:compa
```

rative src:superlative

実際の例

#UTX-S 1.10; en-US/ja-JP; 2011-04-15T10:00:00Z+09:00; copyright: AAMT (2011); license: CC-by 3.0				
#src	tgt	src:pos	term status	src:plural
early adopter	アーリー アドプター	noun	approved	early adopters
fast	高速な	adjective	provisional	
optional	省略可能な	adjective	approved	
optional	オプションな	adjective	forbidden	
save	保存する	verb	approved	

7. 望ましいMTの機能

機械翻訳システムが以下の機能を持つ場合、UTX-Sは効果的に機能する。

- 非専門用語用の高品質のシステム辞書
- 複数のユーザー辞書が使用可能である
- 短い語よりも長い複合語を優先する
- 禁止項目の使用が抑制可能である

8. UTX-Simple作成ガイドライン

8.1. 一般的なガイドライン

一般的に、UTX辞書は専門分野の専門用語のみを含むべきである。多くの場合、UTXの項目は名詞、特に複合名詞である。翻訳ソフトの基本辞書に含まれない、よく調整された辞書を集め、共有し、再利用することにより、翻訳精度が向上できる。なお1文単位の翻訳は、一種の単語とみなすのが妥当だと思われる場合を除き、UTX辞書に含めない。原則として、UTXは対訳文のデータベースである翻訳メモリーと区別されるべきである。

たとえば「XML declaration」のような語は、辞書に登録するだけで「XML宣言」などと正しく訳せる。一方で、windowのような基本語彙は、すでに翻訳ソフトの基本辞書に含まれているため、含めるべきではない。

- それぞれの項目には、1つの訳語だけを含める。
- 基本辞書に含まれるような基本的な単語は除外する。
- 辞書の分野を明確に定義する。
- 単語の基本形を記述する（例：市販辞書の見出しの形式のように、名詞なら単数形、動詞なら基本形）。
- 原語、訳語そのものの以外のコメント情報は、コメント欄に記す。
- 1つの原語に対して、1つの最適な訳語を記述する（一語一義の原則）。合理的な使い分けが必要な複数の訳語がある場合、それらは別の項目として作成する。
- 特定の機械翻訳システムの処理方法に依存するような単語は含めない。
- アルファベットおよび数字は、半角の文字で記述し、全角で記述してはならない。
- 項目の中で、訳語「…研究所」の「…」のように、単語の一部の入れ替えを前提とする表記は避ける。
- コメントを追加するには、ユーザー定義でコメント用列を定義するか、“#”で始まるコメント行に記述する。

8.2. 英語特有のガイドライン

- 固有名詞を除き、1文字目は常に小文字とする。
- 冠詞（a、an、the）は、それが固有名詞の一部である以外は記載しない。

8.3. 日本語特有のガイドライン

- 半角カタカナなど機種依存文字は使用しない。
- サ変動詞は「する」で終わる。例：強調する

- 形容動詞はadjectiveとして示す。
- 形容動詞は「な」で終わる。例：静かな
- 音引きは省略しない。例：ユーザー、セキュリティー、コミュニティー
- 中黒(middle dot)は省略しない。もしくは半角スペースで代用する。
- すでに定着しているものを除き、カタカナ語、和製英語は避ける。

9. 改版履歴

UTX 1.00 (2009年11月10日)

- 初版。

UTX 1.10 (2010年11月22日)

- 全体的な改版。
- 用語ステータス、辞書管理者、双方向、概念ID、辞書ID、および望ましいMT機能の項目を追加した。

UTX 1.10 (2011年4月15日)

- 日時、字句をわずかに修正。

協会活動報告

(2010 年 11 月～2011 年 5 月)

理事会 (E-mail にて議論)

2010 年 10 月 14 日

2011 年度の MT サミット日程調整について

2011 年 4 月 1 日

2010 年度予算理事会

機械翻訳課題調査委員会

2010 年 11 月 17 日 (2010 年度 第 7 回)

- | | |
|------------------------------|--------------------|
| ①今年度事業計画及び来年度活動計画について | 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 |

2010 年 12 月 15 日 (2010 年度 第 8 回)

- | | |
|------------------------------|--------------------|
| ①今年度事業計画及び来年度活動計画について | 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 |

2011 年 1 月 21 日 (2010 年度 第 9 回)

- | | |
|------------------------------|--------------------|
| ①今年度事業計画及び来年度活動計画について | 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 |

2011 年 2 月 25 日 (2010 年度 第 10 回)

- | | |
|------------------------------|--------------------|
| ①今年度事業計画及び来年度活動計画について | 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 |

2011 年 4 月 8 日 (2011 年度 第 1 回)

- | | |
|------------------------------|--------------------|
| ①今年度事業計画及び来年度活動計画について | 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 |

2011 年 5 月 20 日 (2011 年度 第 2 回)

- | | |
|------------------------------|--------------------|
| ①今年度事業計画及び来年度活動計画について | 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 |

AAMT/Japio 特許翻訳研究会

2010 年 11 月 19 日 (2010 年度 第 6 回)

- ①前回議事録の確認 ②今年度の研究内容について ③次回の開催について

2010 年 12 月 10 日 (2010 年度 第 1 回特許情報シンポジウム)

- ①主催者挨拶 守屋敏道 (財) 日本特許情報機構専務理事 特許情報研究所所長)
②基調講演 辻井潤一教授 (AAMT/Japio 特許翻訳研究会委員長)
③招待講演 1 岡崎輝雄氏 (特許庁 普及支援課特許情報企画室 調査班長)
④招待講演 2 Mr. Wang Dan (China Patent Information Center, Director of Information Processing Dept.)
⑤招待講演 3 Ms. Minah Kim (Siriussoft Corporation, Global Division, Director)
⑥セッション 1 ⑦セッション 2 ⑧まとめ ⑨懇親会

2011 年 3 月 4 日 (2010 年度 第 8 回)

- ①前回議事録の確認
②年次報告書・MT サミットでの特許翻訳 WS 関連今年度の研究内容について
③次回の開催について

2011 年 4 月 8 日 (2011 年度 第 1 回)

- ①前回議事録の確認 ②昨年度活動報告
③今年度活動について I, II ④AAMT 総会について
⑤次回の開催について

2011 年 5 月 13 日 (2011 年度 第 2 回)

- ①前回議事録の確認 ②今年度の研究内容について ③次回の開催について

編集委員会

2010 年 11 月 22 日 (2010 年度 第 2 回)

- ①AAMT Journal No.49 進捗確認 ②No.50 の企画について
③その他 ML にて最新号の進捗打合せ及び次回委員会の開催日調整。

2010 年 12 月 AAMT Journal No.48 発行

2011 年 4 月 4 日 (2011 年度第 1 回)

- ①AAMT Journal No.49 進捗表 ②No.49 追加記事・企画 ③No.50 企画について

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

東日本大震災にて亡くなられた方々のご冥福をお祈り申し上げますとともに、被災された皆様に心よりお見舞い申し上げます。被災地の一日も早い復興を心よりお祈り申し上げます。

AAMT ジャーナル 49 号をお送りします。

今号は、もともとは、2011 年 3 月に発行予定でございましたが、東日本大震災の影響もあり、3 月発行を見合わせ、総会号として発行致することとなりました。また、あわせて、震災後、4 月に編集委員会を開催し、49 号への追加記事、ならびに、50 号の編集企画について検討を致しました。以上のような経緯で、今号には、もともと、2011 年 3 月号に掲載する予定でご寄稿頂きました記事、ならびに、総会号として企画されました記事が掲載されている関係で、例年になく大部となっています。

今号の巻頭言は、富士通研究所の潮田明様よりご寄稿頂きました。

今号では、国際・国内会議の参加報告を、多数ご寄稿頂きました。第 1 回特許情報シンポジウム、第 10 回日中自然言語処理促進会議、第 9 回米国機械翻訳協会大会、WIPO(世界知的所有権機関)主催シンポジウム”Patent Translation: Why Quality Matters”、LISA(ローカライゼーション業界の標準化に関する協会) Open Standards Summit といった多彩な会議への参加報告を掲載しています。

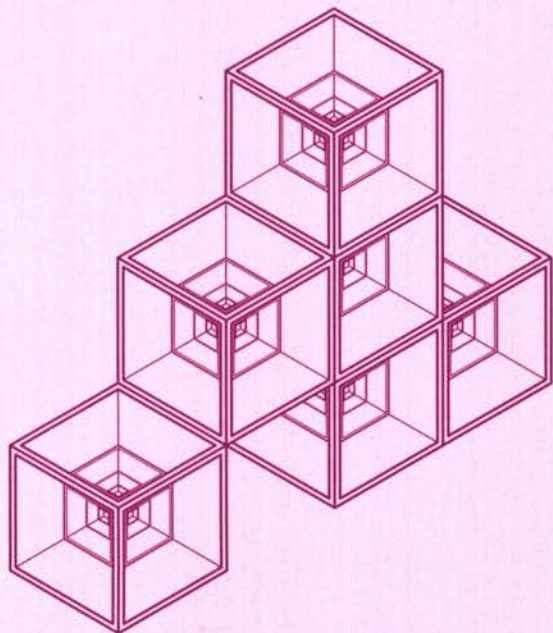
また、プロジェクト報告として、NICT の安田圭志様、村上陽平様より、それぞれ、音声翻訳実証実験についての報告、ならびに、言語グリッドプロジェクトによる京大翻訳の報告をご寄稿頂きました。

また、AAMT の MT 委員会からは、WG1、WG2、WG3 の各 WG から、委員会活動報告をご寄稿頂きました。

Memo

Memo

AAMT



AAMTジャーナル No.49

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：〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘1-1

豊橋技術科学大学 情報メディア基盤センター内

phone：0532-44-6620 fax：0532-44-6620

編集委員会：宇津呂 武仁 小谷 克則 大倉 清司

鈴木 博和 三浦 貢 村上 嘉陽

事務局：神崎 享子 服部 絹依

印刷所：株式会社ナビックス

Asia-Pacific Association for Machine Translation

c/o TOYOHASHI University of Technology Information and Media Center

1-1 Hibarigaoka, Tempaku-cho, Toyohashi-City, Aichi, 441-8580 Japan

Phone:+81-532-44-6620 FAX:+81-532-44-6620

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