自然言語処理 89－9
（1992. 5. 15）

機械翻訳の評価基準について

野村 浩郷　　井佐原 均
九州工業大学情報工学部　電子技術総合研究所

日本電子工業振興協会の機械翻訳システム調査専門委員会技術動向調査ワーキンググループでは、機械翻訳評価基準の検討を進めてきたが、この度その第一版をまとめたので、ここに報告する。この機械翻訳評価基準は、客観的評価、数値的集計、視覚的判断という基本理念を貫いて作成したものであり、ユーザによって機械翻訳導入のための評価に使われるものと、専門的な研究者や技術者によって今後の研究・技術開発の検討のために使われるものとから構成される。ユーザのための評価基準は、さらに、経済的評価基準と技術的評価基準の二つの部分に分けられている。

JEIDA's Criteria on Machine Translation Evaluation

Hirosato NOMURA　　Hitoshi ISAHARA
Kyushu Institute of Technology　Electrotechnical Laboratory
Iizuka, 820, Japan　1-1-4 Umezono, Tsukuba, 305, Japan

This paper describes the Criteria on Machine Translation Evaluation developed by JEIDA Machine Translation Market and Technology Study Committee which is the subcommittee of the Machine Translation System Research Committee of JEIDA (Japan Electronic Industry Development Association). We have developed following three criteria to evaluate machine translation from different point of view, which contains the checklist(s) and the criteria to evaluate the user's answer for the list.

1) evaluation on economic issues to be evaluated by users
2) evaluation on technological issues also to be confirmed by users
3) evaluation on technological issues to be discussed by researchers

1．はじめに

日本電子工業振興協会(JEIDA)機械翻訳システム調査専門委員会(委員長：長尾真京都大学工学部教授)の技術動向調査ワーキンググループ（主査：野村浩郷九州工業大学情報工学部教授）では、数年来、機械翻訳評価基準の検討を進めてきたが、この度その第一版［参考文献参照］をまとめたので、その内容を整理し、ここに報告する。従来より機械翻訳評価基準の必要性がさけばれ検討が進められているが、具体的なものはまだ報告されていない。これまでの検討はどちらかというと精神論的なものであるか、または断片的なものが多く、そのままの形で直接的に適用するには不足である。したがって、ここで述べる機械翻訳評価基準は世界で最初のものであり、実際に適用できかつ客観的な評価が行えるものである。しかしながら、ここで示す機械翻訳評価基準は必ずしも完成されたものではない。今後はこの機械翻訳評価基準の適用実験を繰り返し、その結果をフィードバックして内容を充実していく予定である。ワーキンググループの委員を表１に示す。

表１　ワーキンググループ委員名簿

（敬称略、順不同、旧委員も含む）

野村 浩郷　（九州工業大学）
井佐原 均　（電子技術総合研究所）
伊藤 悦雄　（株式会社東芝）
大井 耕三、湯村 武、藤本 良二（三洋電機株式会社）
新納 浩幸、高橋 雅則　（松下電器産業株式会社）
鈴木 克志　（三菱電機株式会社）
千田 滋也、加登岡 隆　（株式会社リコー）
中岩 浩巳　（日本電信電話株式会社）
福持 陽士　（シャープ株式会社）
松平 正樹、日比 孝　（沖電気工業株式会社）
森本 康嗣、富永 雅介　（株式会社日立製作所）
八木沢 津義、藤田 稔　（キヤノン株式会社）
山端 潔、市山 俊治　（日本電気株式会社）

ワーキンググループでは、まず、機械翻訳を活用しようとしているユーザが実際に使える客観的な機械翻訳評価基準を設定することを考え、次に機械翻訳の開発にたずさわっている研究者・技術者が使える専門的な機械翻訳評価基準を考えることにした。

ここで示す機械翻訳評価基準は、個々の項目の評価ができるだけ客観的に行えるように配慮されており（客観的評価）、かつ個々の項目の評価の集計は数値的に行えるように工夫されており(数値的集計)、さらに 最終の判断はできるだけ視覚的に行えるように考慮されている（視覚的判断）。すなわち、客観的評価、数値的集計、視覚的判断という基本理念を貫いて作成したのがこの機械翻訳評価基準である。

機械翻訳の評価に関わる項目は多岐にわたるので、ここで示す機械翻訳評価基準にあらゆる項目が含まれているのではない。個々の部分はさらに詳細化しなければならないものも多い。しかし、そのような詳細化を進めていくと、項目数が爆発しそのため実際に適用しようとするときに大きな困難をもたらす。したがって、ここで示す機械翻訳評価基準は、できるだけ少ない項目数でできるだけ適切な評価を達成できるように配慮して作られている。

ここで示す機械翻訳評価基準は、ユーザによって機械翻訳導入のための評価に使われるものと、専門的な研究者や技術者によって今後の研究・技術開発の検討のために使われるものとに大別される。ユーザのための評価基準は、さらに、経済的評価基準と技術的評価基準の二つの部分に分けられている。ユーザのための経済的評価基準は、ユーザがどのような機械翻訳システムを導入すれば経済的な効果が期待できるかについて判断するためのものである。ユーザのための技術的評価基準は、機械翻訳システム導入を決定したユーザが、どのシステムが、もっとも自己の希望を満足するかを判断するための指針を与えるものである。

開発者のための評価基準は、機械翻訳に関わる技術的項目を階層的に整理して列挙し、個々の技術の間の関連を把握した上で個々の技術の達成度を評価できるようにしてある。以下の各章において、これらの評価基準のそれぞれの概要を述べる。

2．ユーザ側からの経済的評価基準

この評価基準は、機械翻訳システムを導入しようとするユーザに、経済的な側面から、導入が望ましい機械翻訳システムのタイプを提案するものである。

ここでいう、経済的側面とは、導入を検討する組織の管理部門がマクロ的に機械翻訳導入の是非を判断することができるように、主に経済面に着目して評価基準を考慮したことを意味している。したがって、技術的詳細に亙る評価基準は次章で述べる技術的評価基準に譲り、この評価基準では、比較的短時

間でさまざまな要因に基づく評価が可能なことを目的としている。

本評価の特徴として、２種類の機械翻訳システムのタイプを提案することがあげられる。

１つは、現状の翻訳体制に最も適した機械翻訳システムのタイプであり、もう１つは、ユーザの要求条件を最も満たせると考えられる機械翻訳システムのタイプである。

両者が同じ場合もあるが、両者が異なる場合は、ユーザがその差を、機械翻訳導入の際の現状の体制上の問題点や改善点として認識できる。すなわち、ユーザは現状と将来という２つの観点から、導入すべき機械翻訳システムを判断することができる。言い替えれば、本評価基準では、即時導入を計画しているユーザから、やや先の時点での導入を検討するユーザまでにわたり、意思決定を支援できることを目指している。

評価の枠組を、以下に示す。

（１）２種類のアンケートを行なう。

（２）アンケートの結果を評価し、ユーザに対するレーダーチャートを作成する。

（３）特徴グラフに対して、あらかじめ設定してある機械翻訳システムの典型的タイプの中から最も適合するタイプを提案する。

２種類のアンケートの１つはユーザの現状［翻訳業務の実態］を判断する設問群であり、もう１つはユーザの将来への希望［要求条件］を判断する設問群である。設問の例を図１に示す。

［Ｑ０１］現在の翻訳量は	Ａ１	Ａ６
Ａ（１）月に数百ページ程度またはそれ以上	大	大
（２）月に数十ページ程度	中	中
（３）月に数ページ程度	小	小

図１　設問の例

これらの設問に対する回答を定量的に評価するために、機械翻訳システムを特徴づける１４項目のパラメータ（Ａ１～Ａ１４）を設定した（図２）。設問の回答結果により、対応するパラメータの評価値を変動させる。設問とパラメータの対応を図３と図４に示す。例えば、図１の設問「現在の翻訳量は？」に対応するパラメータはＡ１の翻訳量属性とＡ６の期間属性である。

設問の内容がパラメータに強く関与する場合と、

	ＭＴ向き	ＭＴ不向き
パラメータ：	点数大	点数小
Ａ１：翻訳量	多い	少ない
Ａ２：文書の種類	翻訳容易	翻訳困難
Ａ３：翻訳の質	あまり問わない	問う
Ａ４：翻訳言語	単独ペア	多言語
Ａ５：専門分野	限定されている	不特定
Ａ６：期間	すぐ欲しい	急がない
Ａ７：機械化	進んでいる	遅れている
Ａ８：体制	分業化	非分業化
Ａ９：コスト	現在高い	現在低い
Ａ１０：前編集	少なくてすむ	多い
Ａ１１：後編集	少なくてすむ	多い
Ａ１２：設置条件	きびしくない	きびしい
Ａ１３：前修正	少なくてすむ	多い
Ａ１４：後修正	少なくてすむ	多い

図２　特徴パラメータ

間接的に弱く関与する場合とがあるので、重みを与える。この設問では、Ａ１の翻訳量属性を直接的に聞いているので、Ａ１の重みを大きくし７０ポイントとした。Ａ６の期間属性は、翻訳量が多ければ翻訳に要する期間にも影響を与えるであろうと判断し、この設問に対応づけはしたものの重みは小さくし２０ポイントとしている。

重みは、各属性Ａｉごとに合計点数が１００になるように正規化することにした。その根拠は、次の２つの前提である。

・Ａｉがなるべく多くの質問に分散していた方が、特定の質問への回答結果が評価に大きく影響を与えることがなくなり、多くの質問の回答結果が総合的に評価結果に寄与する。

・各Ａｉはシステム選定にあたって同じ重みをそれぞれ持つ。言い変えれば、特定のＡｉ（たとえばＡ２）が特に重視されることはない。

すなわち、アンケートに関する一般論として、必ずしもすべての設問に対して回答が得られるわけではないということと、機械翻訳システムを総合的に評価するためには、従来の評価基準において重視されていた訳文の質のみならず、さまざまな要因を考慮しなければならないという２点を前提としている。

回答の選択肢と各属性との間には、回答結果が属性に与える影響度を「大中小」の３段階で定義した（あまり細かく分けても意味がないと考えた）。図１の設問例では、回答として（１）を選んだ場合、

アンケート1. 翻訳業務の実態	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
1. 1 翻訳の実態														
現在の翻訳量は？	70					20								
定型文書が多いか？		20												
技術文書が多いか？		20												
事実伝達型文書が多いか？		15												
書き言葉が多いか？		15												
専門語が多いか？		8			20									
専門分野が限定されているか？					50									
言語対は？				100										
文書の再翻訳・再利用は？						20								33
文書の利用目的は？		7	20											
同一種類の文書の翻訳量が多いか？		15												
専門語の対訳表が蓄積されているか？					30									
1. 2 翻訳の体制														
翻訳専門の人員は何名か？	30													
外注する場合があるか？								13						
年間翻訳量の何％を外注するか？								13						
複数人数による分業を行うか？*								13						
用語や文体の統一の対処は？			15					13						
1. 3 コスト														
翻訳のための予算が占める割合は？									34					
外注コストが占める割合は？								12						
外注翻訳コストは？			10						33					
自部門内の翻訳コストは？			10						33					
1. 4 質														
翻訳品質は？			30											
訳語の整合性を重視するか？			15											
1. 5 期間														
即時性は要求されるか？						20								
完訳までの期間は？						20								
初訳の期間は？						20								
前編集に時間がかかっているか？										100				
後編集に時間がかかっているか？											100			
前修正に時間がかかっているか？													50	
後修正に時間がかかっているか？														34
1. 6 機械化の現状														
原文はどのような形で供給されるか？							20					34		
原文は一旦電子化するか？							20					33		
翻訳した文書はどのような形で提供？							20					33		
電子的に文を入力する装置を使うか？							20							
電子化された資料を使うか？							20							
1. 7 翻訳の工程														
文書のチェックを行う必要があるか？								12					50	
部門内の翻訳作業は主に誰が行うか？								12						
訳文の修正、チェックを行なうか？								12						33

図３　特徴パラメータと設問との対応

他の選択肢よりも翻訳量が相対的に多いことになるので、Ａ１に「大」を対応させている。評価値の計算は、以下の式で行なう。

「大」　の場合　　重み　＊　１
「中」　の場合　　重み　＊　０．５
「小」　の場合　　重み　＊　０．２

すなわち、図１の設問例に（１）を回答すると、属性パラメータＡ１の点数に７０＊１＝７０ポイントが加算されることになる。

アンケート2. 要求条件	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
2. 1 翻訳コスト低減														
文書を電子化していく要求があるか？							100		34					
翻訳工程の確立の要求があるか？								50	33					
翻訳時間を短縮する要求があるか？						70								
2. 2 質の向上														
どんな面で翻訳の質を上げたいか？			100											
2. 3 対象の拡大														
今後の翻訳量は？	100					30								
専門分野拡大の要求があるか？					100									
文書の種類の増加の要求があるか？		100												
言語対の種類を増やす要求があるか？				100										
2. 4 導入条件														
導入のために確保できる予算は？									33					
導入のために確保できる人員は？								25						
2. 5 運営条件														
導入後試行期間をおくことが可能か？								25						

図４　特徴パラメータと設問との対応

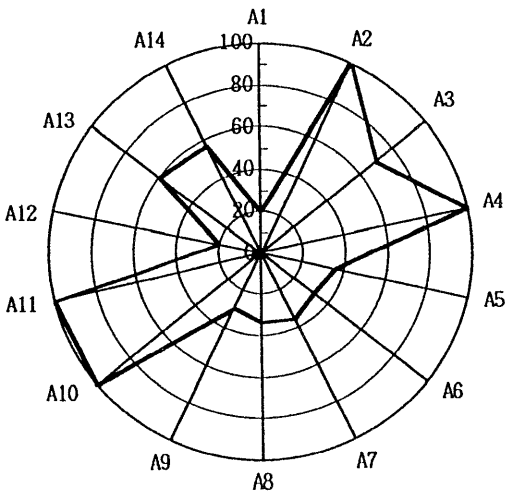

図５　回答結果に対するレーダーチャートの例

すべての設問を終了すると、２つのアンケートそれぞれに対して、１４項目のパラメータに点数が与えられた図５のようなレーダーチャートが作成される。一方、機械翻訳システムをあらかじめ図６に示すような７つのタイプ（Ｓ１～Ｓ７）に分類し、各タイプのシステムを利用するのに最も適したユーザを想定し、その想定したユーザに対してのアンケート結果である図７のようなレーダーチャートを用意しておく。

ユーザのアンケート結果からの２つのレーダーチャートのそれぞれに対して７つのタイプ別のレーダーチャートを比較し、最も類似しているタイプを提案することになる。

Ｓ１：粗翻訳用、高速翻訳、大量、バッチ主体（ＷＳ）
システムを社内に設置する。チューニングができる。
Ｓ２：高品質、分野限定、チューニング多、制限言語
（ＷＳ、専用、スタンドアロン）
分野を限定し、チューニングができる。
Ｓ３：翻訳支援、辞書引き、対話主体
（パソコン、ＷＳ、その中間タイプ）
辞書引き、翻訳支援機能が充実している。
Ｓ４：粗翻訳、低速翻訳、少量、安い（パソコン）
パソコン上で動く、安いシステム。
Ｓ５：ターミノロジーバンク（パソコン）
翻訳は行わず、辞書引き主体。用語の辞書引きができ、その結果を電子的に取り込む。
Ｓ６：英文ＷＰ
通常の英文ワードプロセッサ。
Ｓ７：粗翻訳用、高速翻訳、大量、バッチ主体
（ＬＡＮ、サーバー、ネットワーク）
システムをＬＡＮを経由して利用する。チューニングがしにくい。翻訳サービス（社外）。

図６　機械翻訳システムタイプ

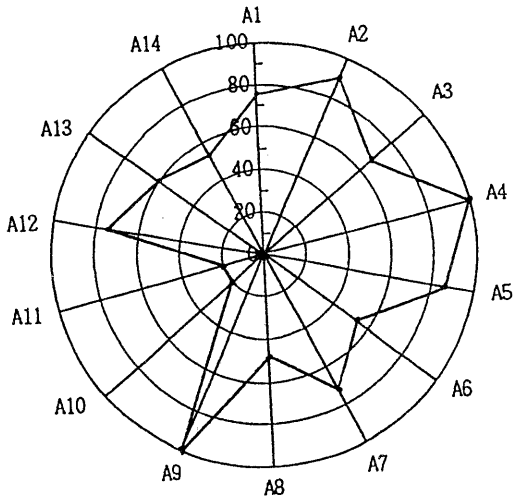

図７　システムタイプのレーダーチャート例

本評価基準は、シミュレーションを行ない、その結果を設問内容や重みの変更などに反映させることにより得られた。このシミュレーションでは、複数の作業者があらかじめ想定した７種類のＭＴシステムの各々について、そのシステムを使うのが好ましいと考えられる４タイプのユーザ（計２８ユーザ）を想定し、各ユーザごとにアンケートの設問に回答し、回答結果から想定したＭＴシステムが提案できるかどうかを自己評価した。

例えば、Ｓ１システムのユーザの一つとして次のプロフィールを持つユーザを想定し、シミュレーションを行なった結果、２種類のアンケート［翻訳業務の実態］および［要求条件］をもとに提案するシステムタイプがどちらもＳ１となることが確認された。

Ｓ１の仮想ユーザ：
「大企業の翻訳部門で、海外の工場で作る製品作りのマニュアルや、社内向けの技術伝達文書などの翻訳（日英）を行なっている。翻訳量は大量であるが、契約している外注会社があり、そこでほとんどの翻訳を行なう。翻訳品質としては高品質の訳は求めないが、用語が統一されていることが望ましい。」

このように、シミュレーションを通じて、本評価基準の目的である「機械翻訳導入時に、比較的短時間でさまざまな要因に基づく定量的評価を可能とする」ことがほぼ確認されたと考えているが、若干の検討項目として、属性分類の見直し、アンケート設問項目の見直しや特徴パラメータと設問との対応の見直しなどの余地が残されている。今後、これらの問題点を考慮しながら改良を重ねていく必要があり、本枠組は、そのためのたたき台としての役割を果たすものと考える。

３．ユーザ側からの技術的評価基準

この評価基準の目的は、機械翻訳システムを導入しようとするユーザにシステム選定の指針を与えることであり、その最大の特徴は、機械翻訳システムを単体として評価するのではなく、ユーザの要求と組み合わせて「機械翻訳システムに対するユーザの満足度」を評価する形式となっている点にある。

機械翻訳システム導入の際には、翻訳品質、導入コスト、維持コスト、前編集・後編集の機能、基本語辞書の充実度等、さまざまな項目について検討・評価する必要がある。これらの評価項目の重要度は、ユーザがどのような体制で翻訳システムを使用したいと考えているかに依存する。たとえば、外国語文書の概要把握のために粗翻訳結果のみ欲しいような場合には、生の翻訳品質が高いことが何より重要であろう。一方、翻訳業務に使用する場合には、翻訳

出力の品質が高いことはもちろんだが、加えて前・後編集の支援機能やユーザ辞書整備支援機能が充実していることが重要な評価のポイントとなろう。

したがって、技術的内容を評価する場合でも、各種の機能のうちどれを重視し、どれを考慮に入れないか、といった重みづけは、ユーザの要求にしたがって変更する必要がある。

このために、本技術的評価基準は、［A．ユーザの要求・状況分析］［B．システム仕様］［C．システム評価］の３つのシートと、［D．ユーザ希望システム構成］［E．ベンチマークテスト評価］の２つの補助シートから構成されている。

評価作業の流れを図８に示す。まず、ユーザがシートＡに回答することにより、機械翻訳に対するユーザの要求が明確化される。シートＡにおける質問項目を図９に示す。次に、ユーザはシートＡの内容を、指示にしたがってシートＤに転記する。これにより、システム構成に対するユーザの希望が整理され、実際に必要とされるハードウェア構成が明確になる。また、ユーザが機械翻訳で処理しようとする典型的な文書をベンチマークテスト文として、実際にそのシステムで翻訳してみる。結果に対し簡単な評価を行なってシートＥに記入しておく。

シートＢは機械翻訳システム提供者が回答するシートであり、機械翻訳システムの機能全般を明らかにするのが目的である。シートＢでの質問項目を図１０に示す。なお、システムのハードウェア構成についてはシートＤを、翻訳速度についてはシートＥをそれぞれ参照して回答を記入する。

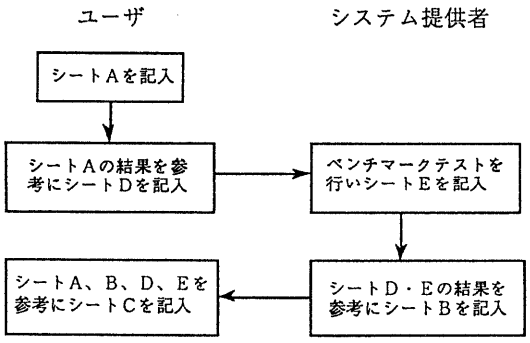

図８　技術的評価の流れ

○対象（何をやりたいか）
（翻訳言語対、翻訳期間、一次出力品質、最終品質、１分野当りの翻訳量、文書の特徴）
○体制・運営（どういう風にやりたいか）
【１】システム構成（処理形態、使用開始時の注意事項、専門用語辞書の作成、複数の人間による使用）
【２】運用形態（原文の供給形態、文字列の抽出、訳文の供給形態、レイアウト）
【３】人員
【４】導入（予算、重視事項、電源、使用する機器）

図９　シートＡの質問項目

○システム
【１】概要（翻訳言語対、翻訳方式、翻訳速度）
【２】ハードウェア（接続できる機器、標準構成）
【３】翻訳ソフトウェア（本体ソフトウェアの価格・機能、専門辞書の種類・語数・価格）
【４】周辺ソフトウェア（ＯＳ、日本語環境、ウインドウシステム、エディタ）
【５】導入時の合計コスト
【６】保守費用（ハード保守、ソフト保守）
○運営
【１】利用形態　【２】教育
【３】導入時・バージョンアップ時のインストール
○翻訳過程
【１】文書互換性　【２】文書管理機能
【３】文章抽出・前編集
（前編集支援機能、原文の知的な編集）
【４】翻訳中の機能
（翻訳できる文書および文の大きさの制限、部分翻訳の可否、未知語の処理、原文の処理、特別な表現の処理、訳文の処理、失敗した時の処理、学習機能）
【５】翻訳後の機能
（後編集の方法・機能、レイアウト、印刷）
○辞書
【１】辞書の種類　【２】翻訳時の複数辞書の使用
【３】一般辞書の記述内容
（複数の語義・訳語、慣用句・熟語への対応、品詞細分類の分類数、共起関係・意味素性の利用）
【４】辞書のチューニング
○操作性
【１】操作の一貫性　【２】操作方法
【３】マニュアル・エラーメッセージ
○その他
【１】システムのバージョンアップ
【２】文法のチューニング　【３】バックアップ機能

図１０　シートＢの質問項目

以上の準備が終ると、評価者は、シートCによって、システムがユーザの希望に合うかどうかを評価する。ここでは、主にシートA、シートBを参照しながらシートCの各評価細目（全部で３１ある）に対する評価点を計算していく。

たとえば、ユーザが翻訳作業に使える期間の長短により、ユーザの機械翻訳システムに対する要求は異なってくる。翻訳要求を受けてから翻訳結果を納入あるいは利用するまでの期間について、ユーザがシートAの翻訳期間についての設問に、「即時」、「１日以内」、「１週間以内」といった比較的短期間の納期を答えた場合、一般的に辞書の整備等のシステムのチューニングに充分な時間をかけることが困難であり、粗翻訳を翻訳システムで行ない、その出力を直接エディタで後編集する作業工程が良いと考えられるため、（１）システムが高速であること、（２）生翻訳出力が高品質であること、（３）後編集機能が優れていることが重視される。故に、短期間の納期を選択した場合は、シートBから得られる「翻訳速度」、「後編集の機能」、「一次出力の品質」の得点が合算され翻訳期間についての評価点となる。一方、「１カ月以内」、「１カ月以上」といった比較的余裕のある納期が与えられる場合には、システムのチューニングを行う時間的余裕があるため、「一次出力の品質」に加えてユーザによるシステムの辞書・文法拡張能力が重視されると考えられることから、「一次出力の品質」と「辞書文法チューニング能力」が合算され翻訳期間についての評価点となる。すなわち、システムが十分なチューニング能力を備えていることが重要なポイントとなる。翻訳期間に関する視点から得られたこの得点は、レーダーチャート上の各独立した軸である「対象」・「翻訳速度」・「編集能力」・「一次出力の品質」・「チューン後の品質」に適切な重み付けの上、カウントされる。

評価点の計算がすべて終ったら、１０本のレーダーチャートの軸に対応する評価細目の点数を、シートCからレーダーチャートシートに転記し、レーダーチャートを完成させる。ユーザは、複数のシステムのレーダーチャートを比較検討し、もっとも自分に適したシステムを選択する。

レーダーチャートは本評価基準による評価作業の最終生産物である。これは、システムの性能およびユーザのシステムに対する満足度を、対象・翻訳速度・一次品質・チューン後の品質・編集能力・システム構成・体制・運用形態・人員・導入・システムの１０の軸についてそれぞれ１０点満点で評価し、グラフとして視覚的に表示したものである。以下、各評価軸について説明する。

（１）対象

シートAに対する回答で明らかになる「ユーザが機械翻訳によってやりたいこと」に対し、システムがどの程度対応できるかを総合的に評価する。

（２）翻訳速度

翻訳速度のカタログ値およびベンチマークテスト結果を総合して、翻訳速度に対する絶対評価を行なう。この項目は、ユーザの要求に対する満足度ではなくシステム単体に対する絶対評価項目である。ユーザの要求を考慮した相対評価は、「対象」を評価する際に行なわれる。

（３）一次品質

辞書や文法をユーザに合わせてチューンアップする前の、機械翻訳システムの出力品質に対する評価を行なう。この項目も、機械翻訳システム単体としての絶対評価である。

（４）チューン後の品質

辞書や文法に対するチューニングを行なった後の一次出力品質を評価する。一次出力品質に対するユーザの期待の程度が高ければ高いほど厳しい評価が行なわれるようになっており、その意味でユーザのシステムに対する満足度を評価する形になっている。

（５）編集能力

後編集を支援する機能に対する評価項目である。ここで後編集機能とは、一旦システムを落した後でも文を再翻訳することなく起動できるすべての編集機能を指す。このように定義するのは、翻訳と後編集を分業する場合に有用である機能とそうでない機能を区別したいためである。

（６）システム体制

システム使用形態（後編集をシステムで行なうか、システム外で行なうか、全く行なわないか）、辞書・文法のチューニングの必要性の有無、専門用語辞書作成の必要性の有無、といった、システムの大枠に対するユーザの要求をシステムが満足できるかどうかを評価する。

（7）運用形態

原文や訳文の供給形態に対するユーザの要求をシステムが満足しているかどうかを評価する。評価は、「原文の供給形態」「訳文の供給形態」「文字列抽出機能の必要性」「原文レイアウトの訳文への反映の必要性」の4つに対して行なわれる。

（8）人員

ユーザが想定している人員配置に対してシステムが対応できるかどうかを評価する。

（9）導入

導入時の要求に対する満足度を評価する。予算をオーバーしないこと、電源工事など導入に必要な工事が許容範囲内であることを評価するのに加えて、シートAで明らかになるユーザの機械翻訳システムの「導入のポイント」をシステムが満足するかどうかをも評価する。

（10）システム

システムのハードウェア構成に対するユーザの満足度を評価する。ユーザの希望構成はシートAおよびシートDに示されている。また、シートBにはユーザ希望構成に対応するシステム仕様／価格がシステム提供者により記入されている。評価者はこの二つを見比べて満足できる構成かどうかを判断し、システム適合度として数値化する。

本評価基準の最大の特徴は、ユーザの要求分析の結果とシステム分析の結果を組み合わせて「機械翻訳システムに対するユーザの満足度」を評価する形式になっている点にある。本評価基準の妥当性を検証するために、ユーザ3者とシステム6つを組み合わせてシミュレーションを行なった結果、予算や機器構成といった、ユーザの要求に対応する評価が容易な項目はもちろん、人員配置や翻訳業務の形態などの項目についても、システム間の細かな差異を定性的に反映した評価結果を得られることがわかった。シミュレーション結果の一部を図11に示す。この例のユーザは「納期は1週間程度。一次出力品質、最終品質に対する要求は読んで誤らない程度の正確さと適切な文体である。数式や定型ヘッダなどを多く含むマニュアルを扱うことが多い。予算は300万円程度である。」というもので（もちろん、実際のシミュレーションでは、シートAを用いて、これよりはるかに詳細にユーザの希望を調査している。）

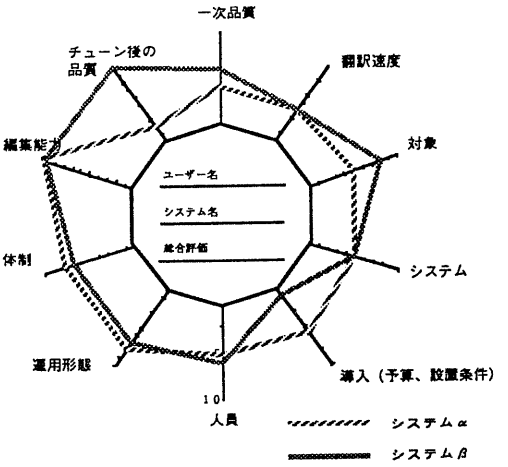

図11　シミュレーション

システムαはパーソナルコンピューター上の翻訳ソフトウェアであり、システムβはワークステーション上の翻訳ソフトウェアである。

評価基準作成の過程において、あるいはシミュレーションにおいて様々な問題点が明らかになった。まず、同一のシステムタイプを持つシステムの間の差異をもっと目に見えやすい形にすることが必要である。ユーザの要求分析を行なうための設問も、今後さらに改訂が必要であると考える。また、評価点の付与においては、例えば「納期が短いならば辞書作成支援機能よりも後編集支援機能を重視する」など、ユーザの業務形態と機械翻訳システムに対する要求の関係としていくつかの仮定を行なっている。これらの仮定の正しさも、具体的な翻訳業務の分析を通じて今後検証していく必要があろう。さらに、もっとも基本的な評価項目である翻訳品質評価も、さらに高精度化することが必要であろう。

4．開発側からの評価基準

本章では、機械翻訳システムの研究開発に従事している研究者、開発者、プログラマ、及び研究開発プロジェクトを管理している管理者自身が、自らの開発しているシステムを評価する際にチェックすべき評価的項目について示す。本評価基準は、そのシステムが技術的にどのレベルまで達成しているか、開発目的に合致したシステムとなっているかを、内部的に評価するために作成した。機械翻訳システム

を開発者側の立場から評価する場合、評価項目は多岐にわたり、それらを個別に評価することは困難であるため、本評価基準では2種類の評価軸を設け、各々の軸の観点から評価し、システムの特徴や弱点を明確にすることを試みた。

本評価基準では、システム全体をさまざまな角度から評価するために以下の6つ軸を設けて、関連する設問、回答にその軸を示す記号を付与している。

・汎用性／分野依存性：システムが汎用的なものか分野に特化したものかを示す
・機能網羅度：システムが基本的な機能を網羅しているかどうかを示す
・緻密度：システムが細かなところまで考慮して作成されているかどうかを示す
・オリジナリティ度：システムに用いられている要素技術にオリジナル性があるかどうかを示す
・システム解放度：システムがユーザーに解放されているかどうかを示す
・使い勝手度：システムが使い易いかどうかを示す

同時に、各設問、回答に技術的な難易度を示す記号を付与している。

・難易度大（A）＝3点
・難易度中（B）＝2点
・難易度小（C）＝1点

これは、以下のように分類したシステム内部の各処理を技術的に評価するためのものである。

・辞書
・解析技術
・中間表現と加工
・生成
・非文法的な現象の処理
・カスタマイズ、学習機能
・環境、操作

本評価基準での評価方法は、

1）まず、機械翻訳システムの開発者などがそのシステムに関する評価基準の設問に回答する。
2）次に、各評価軸（汎用性／分野依存性、機能網羅度、…）ごとに、軸の記号が回答に付与されている場合に1点として合計点数を求める。また、各処理（辞書、解析技術、…）ごとにA＝3点、B＝2点、C＝1点として技術的な難易度の合計点数を求める。
3）各評価軸ごとに満点を１００として評点を求める。また、各処理ごとにも満点を１００として評点を求める。
4）各評価軸の点数をもとに目標としているシステムとのギャップを認識する。また、各処理の点数から技術的に進んでいる部分および遅れている部分を認識する。

以下では、開発者側の評価基準の項目について示す。本資料では、個々の評価項目について解説するスペースがないので、それらを大きく8種類に分類して解説を加えた。

（1）システムの特徴

システム名、開発者名、動作するコンピュータ、言語対、システムの目的などシステムの概略を知るための質問を行なう。これらは開発者自身がシステムを評価する際には分かり切ったことかも知れないが、第三者にシステムを説明する際には重要な項目である。この項目によってシステムがどういったコンセプトで作られているか、どのような用途で使われることを想定して作られているかがはっきりする。

（2）辞書

機械翻訳においては辞書にどのような情報が書かれているかが直接翻訳性能に影響する。ここでは辞書の各側面について明らかにするための質問を用意した。具体的には、辞書の種類、規模、記述内容、利用方法、拡張性など、辞書にかかわる全般について質問項目を挙げた。近年その重要性が叫ばれているコーパスに関する質問項目もここに含まれている。

（3）解析技術

ここでいう解析技術とは入力方式、形態素解析、構文解析、意味解析、文脈解析の5つである。システムによっては構成としてこの5つにわかれているとは限らず、前の2つは構文解析の部分として扱っているシステムが多いと思われるが、ここでは機能としてこれらの処理を行なっているかを質問している。これらの処理がどれぐらい深く行なわれているかが一般的には翻訳性能に大きく関わる。

（4）中間表現と加工

翻訳方式とその内容に関しての質問を行なう。翻訳方式については、トランスファ方式の場合はどのような変換規則があるかを、中間言語方式やその他

の方式の場合はその特徴や規模をチェックする。

（5）生成

訳語選択、自然な訳語を生成するための処理などについての質問を行なう。具体的には、訳語選択処理がどの程度行われているか、訳文が非文かどうかのチェックを行なっているか、生成処理についてどの程度行われているか等についてチェックする。

（6）非文法的な現象の処理

一般的な文法現象とは異なる言語現象を含むテキストの処理について質問する。これらの機能は、機械翻訳システムの本質的な性能とは直接関係ないが、ユーザの使い勝手の面からは、非常に重要である。具体的には、箇条書、タイトル文、引用句、挿入句、図、表、数式、手紙などの定型文、日付、数詞、記号、文法範疇外の文等が取り扱えるかを質問する。

（7）カスタマイズ・学習機能

ここでは、システムをユーザ毎にカスタマイズする機能について質問する。現在の機械翻訳システムは、どのような文種・分野でも翻訳できるレベルには、達していない。よって、実用レベルの訳文を得るためには、ユーザ毎にシステムをカスタマイズする必要がある。これをコスト的・時間的に容易に行えるかどうかは、そのシステムが実用となるかどうかを決めるものであり、非常に重要である。

（8）環境・操作

ここでの質問は、機械翻訳システムの本質的な性能（訳文の質、翻訳速度）とは関係なく、主に使い勝手に関するものである。ただし、このことはこの質問が重要ではないことを意味しない。現在の機械翻訳システムは、完全自動翻訳のレベルには達しておらず、何らかの形で人手の介入を必要とする。そのため、その機械翻訳システムを利用することにより、翻訳コストの削減、翻訳時間の短縮などが実現できるかどうかは、本節で質問されている機能がサポートされているかどうかに強く影響される。

以上のように、開発者自らが機械翻訳システムを評価するための、2つの評価軸による評価基準を作成した。これにより、

1．評価軸ごとのシステムの総合評価
2．各項目ごとのシステムの評価

が得られる。これにより、総合、又は、個別のシステムの長所、短所が浮かび上がるはずである。たとえば、自社のシステムの使い勝手はどのくらいの評価なのか、または、解析系は弱点となっているのか、といった評価である。これらの評価項目、および評価軸、難易度を作成するについては少なくとも日本では一般的と思われる機械翻訳システムを想定し、主観的に決定している。したがって、できるだけ、客観的な評価基準となることを心掛けてはいるつもりだが、項目によって細かい部分と粗い部分がある可能性がある。また、特殊なタイプのシステム、たとえば、Meteoのような（天気予報の翻訳だけを行うといった）システムでは全体に低い評価しかえられないはずだが、そういう場合、システムの実用性と本評価とが対応しなくなる。こういったシステムを評価することは本評価では充分ではない。

つまり、この評価は、それぞれの翻訳システムが基本的な機能をどれくらい備えているかを知るものであって、そのための指標である。

今後の課題として、さらなる評価項目の詳細化、評価軸の追加等の見直し、難易度の細分化、見直しなどがあげられる。また、開発者から見た翻訳システムの評価として、このような設問形式によるもの以外の評価方法も検討する余地がある。

5．おわりに

世界で最初の機械翻訳評価基準を示した。ここで示した機械翻訳評価基準は大きく分けて二つの部分に分けられている。第一の部分はユーザが機械翻訳を導入するときの機種の選定や体制の準備のための評価基準であり、第二の部分は機械翻訳の開発に携わっている研究・技術者が技術の達成度と今後の問題点を明確にするためのものである。ここで示した機械翻訳評価基準は最終的なものではない。したがって、今後適用実験を繰り返しその結果を取り入れて改善していく。さらに、機械翻訳の翻訳品質に関わる評価基準部分をとりだし、今後新たに、ユーザによる機械翻訳品質評価基準および開発者による機械翻訳品質評価基準をまとめるつもりである。

参考文献

機械翻訳システムの実用化に関する調査研究
（社）日本電子工業振興協会　平成4年3月