解説

機械翻訳システム評価法の最前線

国際電気通信基礎技術研究所・音声言語コミュニケーション研究所

隅田 英一郎 eiichiro.sumita@atr.jp
佐々木 裕 yutaka.sasaki@atr.jp
山本 誠一 seiichi.yamamoto@atr.jp

機械翻訳に対する増大しつづける需要とその翻訳品質に対する期待に現時点の技術は応じきれていない．

しかし，最近，機械翻訳技術の研究は大きく変わった．1つは，翻訳品質を自動的に評価する手法が提唱され普及したこと．もう1つは大量の対訳から翻訳知識を学習する手法が盛んに研究され，短時間に低コストで機械翻訳システムを構築する技術が開発されたこと．この2つが相まって，機械翻訳システムが長足の進歩を遂げ社会の需要と期待に応える日は近い．

本稿では，特に，翻訳品質の評価に焦点をあてて，①代表的な人手評価の手法，②最も広く利用されている自動評価の手法，③評価型国際ワークショップの1つであるIWSLT，④自動評価の応用と展望について述べる．

機械翻訳システムと翻訳品質評価

情報通信技術（Information & Communications Technology, ICT）を基盤とするグローバリゼーションの勢いはとどまるところ知らない．必然的に多言語翻訳に対する需要も増大し続けている．アジアでは諸国の経済成長が，欧州ではEUの拡大が，米国では安全保障が契機となり，それぞれのニーズにあった多言語機械翻訳の研究プロジェクトが今強力に推進されている．次に焦点が当たるのはBRICs（ブラジル，ロシア，インド，中国の4カ国）の言語であろうか．日常生活に目を転じると，有名な検索エンジンは検索結果の翻訳サービスを提供しているし，Webやmailの翻訳サービスも多数ある☆1．しかし，残念ながら，その翻訳品質は十分ではないと感じる人が多い．つまり，機械翻訳に対する需要とその翻訳品質に対する期待に現時点の技術は応じきれていない．

一方，今の世の中，万事，評価は避けて通れない．評価は進歩を促すポジティブな道具である．たとえば，ダイエットを成功させるには，自らの体型や体重を第三者的に測定し，あるダイエット手法を試したときの変化を踏まえて，そのまま続けるか調整するかを判断しなくてはならない．同様に，機械翻訳システムの開発者にとってはシステムを改良するために，その利用者にとっては良い翻訳システムを入手するために，翻訳品質の客観的評価は重要である．

機械翻訳の研究は半世紀を超える歴史があるが，最近まで，翻訳品質評価は，人手評価が主流であった．翻訳品質評価の自動化，すなわち，自動評価はほとんど試みられることがなかった．

理由は「翻訳では1つの入力に対して複数の訳文が許されるのが普通であり，その多様性から機械で扱うのは困難と考えられていたからである」．**図-1**に，ある短い

原文
窓を開けてもいいですか．

訳文
1. May I open the window?
2. OK if I open the window?
3. Can I open the window?
4. Could we crack the window?
5. Is it okay if I open the window?
6. Would you mind if I opened the window?
7. Is it okay to open the window?
8. Do you mind if I open the window?
9. Would it be all right to open the window?
10. I'd like to open the window.

図-1　正しい翻訳は1つではない

☆1 http://www.aamt.info/japanese/mtweb-j.htm

日本語文に対する英語の訳文を 10 文だけ挙げた．すべての訳文に共通なのは「the window」だけであり，訳語や構文が異なっているのが分かる．

一般に，可能な翻訳文は膨大な数になりすべてを列挙するのは不可能である．また，外延的にではなく内包的に取り扱うことも難しい．異なる表現の内容等価性を扱う自然言語処理を「言い換え技術」と呼ぶが，これに関しては基礎研究に緒がついたばかりである [1]．

ところが，2001 年に BLEU（BiLingual Evaluation Understudy）[2] という名前の自動評価の手法が提案されると流れが変わった．BLEU は，単純な動作原理から当初多くの手厳しい批判を受けながらも，その意義が認められると急速に普及し，いまや，BLEU をはじめとする自動評価結果を示すことは機械翻訳関係の論文では当然のこととなっている．この動きと並行して，機械翻訳システムの研究にパラダイムシフトが起こっていた．従来，機械翻訳システムの研究は，大規模な辞書や汎用的な文法や翻訳規則の作成を前提としており，多数の研究者と年単位の時間を要する大事業であった．近年，大量の対訳から翻訳知識を学習する手法が盛んに研究され，その結実として，短時間に低コストで機械翻訳システムを構築する技術が開発された☆2．

これらの評価技術と構築技術の進歩が相まって，機械翻訳の研究は大きく変貌を遂げている．自動評価を活用して翻訳を最適化問題としてとらえ直す手法，共通のデータを用いて異なる翻訳技術を比較するワークショップの開催☆3，必要となる対訳データの共同開発，など多方面にわたって活発に研究開発が進んでおり，また，最新の成果の商用化も始まっている．

機械翻訳システムの評価には，翻訳品質，処理速度，移植性など多様な観点☆4 からの評価が必要になるが，本稿では，翻訳品質の評価について検討する．次章以降で，①代表的な人手評価手法，②最も広く利用されている自動評価手法，③評価型国際ワークショップの 1 つである IWSLT（International Workshop on Spoken Language Translation），④自動評価の応用と展望について述べる．

人手評価

人手評価とは，何人かの評価者によって翻訳品質を評価することである．評価の観点，評価の基準，評価のレベル分け（N 段階）☆5 はさまざまであるが，ここでは，代表的手法として，評価型国際ワークショップ IWSLT で採用された方法について説明する．流暢さ（Fluency）と適切性（Adequacy）の 2 つの側面について翻訳品質を 5 段階で評価する．

- 流暢さ（Fluency）は，翻訳文がその言語を母国語とする人にとって，その言語の文章としてどの程度「自然な」表現であるかを評価する．5 段階の目安として，以下の「言葉」が（英語への翻訳では英語で）与えられ，当該言語の母語話者である評価者の直感によって判定する．

5	まったく問題ない
4	良い（Good）
3	非母国語的（Non-native）
2	不自然（Disfluent）
1	理解不能（Incomprehensible）

翻訳文と選択肢のボタンを持った簡易なインタフェースを提供し，評価者は深く考えずに次々評価するよう指示される．

- 適切性（Adequacy）は，どの程度，原文の情報が翻訳文に含まれているかを評価する．5 段階の目安として，以下の「言葉」が与えられ，評価者の直感によって判定する☆6．

5	すべての情報（All of the information）
4	ほとんどの情報（Most of the information）
3	多くの情報（Much of the information）
2	少しの情報（Little information）
1	情報なし（None of it）

適切性の評価は，原文の情報との比較が必要となるが，IWSLT では，原文の代わりに，原文の情報を過不足なく表現した参照訳を用意し，これとシステムの翻訳文との比較を行っている．目的言語の母語話者でかつ入力言語を正確に理解できる人材を確保することがなかなか困難であるためにとられる便法である．

上記のような「言葉」による目安だけで評価する場合，少し詳しい基準の説明があり，何回か試行して慣れさせる場合などがある．しかし，人間はその特性からいって，「機械のように」は作業できず，気まぐれとも思える判

☆2 「Machine translation in a day」というキャッチフレーズが存在するほどである．
☆3 音声認識分野において研究の活性化に成功した DARPA プロジェクトに触発されている．
☆4 http://www.isi.edu/natural-language/mteval/ に機械翻訳の評価に関する網羅的な報告書が公開されている．
☆5 N=2，4，5，10 などさまざまな提案がある．
☆6 この基準は情報の不足のみに着目しており，情報の過剰について頓着していない点は問題がある．

原文
窓を開けてもいいですか.

訳文
1. May I open the window?【良】
2. Do you mind if I open the window?【良】
3. I like to open the window.【良】
4. It is possible to open the window is?【悪】
5. Do you mind if I open the window can I pick it up?【悪】
6. Where is the ticket window?【悪】
7. The the the the the the.【悪】

※ 1 ～ 6 は実際のシステムの翻訳文，7 は作例.

図 -2 良い訳と悪い訳のサンプル

断をしがちである．判定にバラツキが生じるのは避けられない．複数の被検者に評価させる，結果を統計的に処理する，などの工夫が不可欠である．

人手評価は，人間が翻訳文をどの程度理解できるかを実際に判定できるという利点があるが，判定にはコスト（時間と労力）がかなりかかる．そこで，次章の自動評価が必要となる．

自動評価

自動評価では，まず，テストデータの原文に対する典型的な翻訳文を複数☆7 の参照訳として用意する．参照訳とシステムの翻訳文の双方を単語列☆8 として見て，あらかじめ決めた類似度により，翻訳文の品質を評価する．

ここでは，最も有名な手法である BLEU[2] について説明する．BLEU は，『品質が良い訳文と（複数の）参照訳とは文中の単語列が頻繁に一致し，品質が悪い訳文ではそうはならない』という性質を根拠にしている．**図 -2** の機械翻訳システムの翻訳文と，（参照訳として）図 -1 の正解の翻訳文とを比べるとこの性質が理解できる．

これを定量化するために，まず，*n*-gram☆9 の適合率を考える．*n*-gram 適合率とは，訳文中の *n*-gram がいずれかの参照訳中の *n*-gram に一致する度数（分子）を訳文中の *n*-gram の総度数（分母）で除した割合である．

次に，*n*-gram 適合率には図 -2 の 7 のような悪い訳の値が大きくなってしまう問題があり，これを回避するため，修正 *n*-gram 適合率（modified *n*-gram precision）を導入する．修正 *n*-gram 適合率では，*n*-gram 適合率の分子を，参照訳ごとの当該 *n*-gram の度数の最大値を超えないように *n*-gram が一致する度数を刈り込み（clip）したもの $\mathrm{count}_{\mathrm{clip}}$ に修正する．さらに，テストセットの全文に対する修正 *n*-gram 適合率は次式で求める．

$$p_n = \frac{\sum_{C \in \text{テスト セット の全訳文}} \sum_{n\text{-}gram \in C} count_{clip}(n\text{-}gram)}{\sum_{C \in \text{テスト セット の全訳文}} \sum_{n\text{-}gram \in C} count(n\text{-}gram)}$$

上式には「翻訳文が短いとスコアが高くなる」という性質があり，これを補償するため，BLEU は，次式のように，短い翻訳文に対するペナルティ項 BP（Brevity Penalty）と，修正 *n*-gram 適合率の $n = 1 \ldots N$ についての加重幾何平均との積で表される☆10.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

また，*BP* は以下のように定義される．ただし，訳文の長さの総和を *c*，訳文に最も類似した参照訳の長さの総和を *r* とする．

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{(1-r/c)} & if \quad c \le r \end{cases}$$

BLEU は，*n* の値が小さいときは単語に近い単位での内容の伝達を測定することになるため適切性を評価しており，*n* の値が大きいときは単語の並びで表現を評価しているため，流暢さを評価していると解釈できる．

BLEU は，新聞記事を対象とした翻訳システムを使った評価実験では，人手評価と高い相関（0.96 ～ 0.99）を達成している[2]．したがって，BLEU スコアを改善するような翻訳システムの変更は人手評価も改善できる．簡便に利用できる BLEU で頻繁に気軽に評価し，大きな改善を実現した後でコストのかかる人手評価を行えばよい．このようにして機械翻訳システムの変更・評価・フィードバックの開発サイクルを効率的に回すことができ，短期間に大幅な改良を達成できる．

このほかに NIST，mWER，mPER，GTM など多数の指標が提案され百家争鳴状態である（IWSLT の総括論文☆11 に個々の参考文献があるので参照されたい）．

また，BLEU をはじめとする機械評価手法はその実装したソフトウェアが公開されている☆12 ことが多く，誰でもすぐに利用できる．

ここまでで説明した，人手評価と自動評価の長所・短所を**表 -1** に対比して示した．次にこれらの評価手法を使って，音声翻訳のための技術を比較したワークショップ IWSLT について説明する．

☆7 先に述べたように一般に翻訳の正解は多様である．したがって，参照訳を多くすると，より評価は安定する．通常，4 ～ 16 通り用意する．
☆8 定義を変えると複数のシステムを比べた場合の順位が変わることもある．単語単位か文字単位か，句読点の有無，複合語の扱いなどをあらかじめ決めておく必要がある．
☆9 *n* 個の単語の連鎖
☆10 通常 *N* は 4，重み w_n は均等とされ，値 1/4 が使われる．
☆11 http://www.slt.atr.jp/IWSLT-2004/proceedings/WO_tsujii.pdf
☆12 http://www.nist.gov/speech/tests/mt/resources/scoring.htm

	長所	短所
人手評価	• 5 段階評価など，直感的に理解しやすい. • 参照訳なしで評価可能.	• 評価者によるバラツキがある. • コスト（時間と労力）が大. • 評価者の資質要求が厳しく，特に話者人口の少ない言語の場合，評価者集めが大変.
自動評価	• 評価結果に揺れがなく，常に一定のスコア. • 参照訳さえあればコスト（時間と労力）がほぼゼロ.	• スコアが直感的でない. • 複数の参照訳の準備が前提.

表 -1　人手評価 vs. 自動評価

機械翻訳の評価型ワークショップ

近年，対訳コーパス（原文とその訳文のペアを大量に集めた翻訳の模範とすべきデータ）から自動的に翻訳システムを構築するコーパスベースの翻訳技術の研究・開発が盛んになってきた.

その結果，共通の学習データを用いて異なるシステムを開発し，共通のテストデータと評価手法を用いて，どのような翻訳技術が有効かを比較検討することが可能となり，このような技術評価を目的とするワークショップが頻繁に開催されるようになりつつある（**表 -2**）.

ここでは，昨年秋開催された評価型国際ワークショップ IWSLT-2004 を概観し評価結果について紹介する.

回 国際ワークショップ IWSLT-2004 の概要

2004 年 9 月 30 日～ 10 月 1 日にかけて，ATR[☆13] で開催された評価型国際ワークショップ IWSLT では，旅行会話に関するコーパス BTEC（Basic Travel Expression Corpus）を用いた評価キャンペーンと音声翻訳関連の技術論文を集めた一般セッションが設けられた.

翻訳に関する評価型ワークショップを開催するには，国際的な協力が不可欠となる．IWSLT では，C-STAR[☆14]（Consortium for Speech Translation Advanced Research）の協力により，対訳コーパスの準備，および人手評価作業を各機関が分担するというかたちで国際協力が行われた.

回 評価キャンペーンの概要

IWSLT-2004 では，14 団体が 5 種類のトラックに参加した（**表 -3**）．small トラックは，提供されたコーパスのみの利用が許される．additional トラックは，LDC[☆15] より入手可能な対訳コーパスの利用は許される．unrestricted トラックは利用する言語資源に一切制限がない.

	IWSLT	TIDES
対象	音声翻訳	文書翻訳
言語対	日本語→英語 中国語→英語	アラビア語→英語 中国語→英語
主催	ATR	（米国）標準技術局
期間	2004 ～	2001 ～ 2005
公開性	一般公開	非公開（参加者のみ）
URL	http://www.slt.atr.jp/IWSLT-2004	http://www.nist.gov/speech/tests/mt/index.htm

表 -2　評価型ワークショップ

言語	トラック	延べ参加団体数
中英	small	9
	additional	2
	unrestricted	9
日英	small	4
	unrestricted	4

表 -3　IWSLT-2004 における各トラックへの参加団体数

原文

1. 航空券を家に忘れてしまいました.
2. オペラ座はどこですか.
3. このフィルムの現像と焼き付けをお願いできますか.
4. 背中のマッサージはいかがですか.
5. 次のかたどうぞ．パスポートと申告用紙を出してください．何か申告する物はありますか.
6. 2 つ目の角にあります.
7. 海側の部屋に替えてください.
8. お勘定をおねがいします.
9. 玉ネギをお願いします.
10. 搭乗開始時刻は何時ですか.

訳文

1. i left my ticket at home
2. where is the opera
3. can i develop and print this film
4. would you like a facial massage on my back
5. please give me your passport and next person form do you have anything to declare
6. it 's on the second corner
7. i 'd like to ocean view room
8. i 'll be arriving i have the bill please
9. i 'd like onion
10. what time does boarding

図 -3　統計翻訳システムのテスト文の翻訳（一部）[☆16]

回 評価結果

IWSLT に参加したシステムの実力を実感していただくために，好成績をあげた統計翻訳システムの訳文を**図 -3** に示す．下線で示したように変なところもあるが，平均的にはかなり高品質といえる.

評価キャンペーンでは，人手評価（流暢さ，適

☆13 http://www.atr.co.jp/index_j.html
☆14 http://www.c-star.org/
☆15 http://www.ldc.upenn.edu/
☆16 文頭の大文字や句読点など英語の正書法に従っていないのは，訳文の評価が lower-case only, no punctuation marks という条件で行われたためである.

切性）と自動評価（BLEU，NIST，mWER，mPER，GTM）が行われた．システムごとの詳細については，IWSLT-2004 の Web サイトの論文集をぜひ参照していただきたい．アルゴリズムと性能の関係が分かる．ここでは，IWSLT-2004 で得られた評価全体に関する知見を紹介する．

人手評価

- 同一の 100 文を評価者に 2 度評価させたときの評価結果の差は平均 0.4 であった．このことから，2 つの翻訳システムに品質の差があると言うためには，少なくとも 0.8 以上の差が必要である．
- 人手評価基準を，5 か 5 未満かの 2 クラス分類に設定すると，評価結果が安定する．

自動評価と人手評価の相関

翻訳システムのランキングに関して自動評価と人手評価の相関について説明する．

- 5 種類の自動評価指標の中では，流暢さに関しては BLEU が人手評価と中英 0.85，日英 0.94 という最も高い相関を示した．適切性については，NIST が最も高く，中英 0.53，日英 0.97 という相関を示した．
- 評価値 5 か 5 未満かの 2 クラス分類による評価では，流暢さに関しては，BLEU が最も高い 0.86（中英），0.91（日英）を示した．適切性については，BLEU が中英については 0.74 で最も高かったが，日英については mWER が −0.97 という最も高い相関[☆17]を示した．

このように，自動評価は人手評価との相関があると言えるが，最も相関の高くなる評価手法は，条件によって異なっていて，オールマイティな手法はないのが現状である．

翻訳評価の短期的課題

人手評価は，前章で述べたように，個々の評価者の評価のバラツキと評価者間の評価のずれが大きく，安定した評価の実施方法について，さらなる研究が必要である．

自動評価は，参照訳と訳文との単純な類似度を測っていて，原文に含まれている単語の意味的な重要度については考慮していない．たとえば，「手数料が必要ない」を「手数料が必要」と訳してしまった場合，実用場面では非常に大きな問題が生じるが，現在の自動評価手法では単純に 1 単語の欠落として評価される．また，コーパスベースの翻訳技術では，「右」を「左」と誤訳したり，原文にない情報を付加して訳出したり（図 -3）といった誤りが，生じることがあるが，自動評価はこの種の誤りに必ずしも感度が良くない．もちろん，人手評価はこれらの点には敏感で厳しい．

以上は，今後改良していくべき課題である．

自動評価技術の応用と展望

◙ 評価スコアを TOEIC に換算

ここまでに紹介した翻訳品質の評価手法では，2 つのシステムを比較してどちらがより良いかは分かるが，あるスコアを達成したシステムが実際どの程度有用なのかという問いには答えられない．

そこで，英語能力の検定試験として有名な TOEIC[☆18] スコアに着目し，TOEIC スコアが既知の人間をモノサシとして用いる翻訳一対比較法[3] が提案されている．TOEIC スコアが既知の複数の日本語母語話者（ここでは TOEIC 被験者と呼ぶ）に，テストデータの日本語文を英語に翻訳させる．TOEIC 被験者の翻訳文と機械翻訳システムの翻訳文とを対にして，日英バイリンガルの評価者が比較し，優れた方を選択する．すべての一対比較が完了した段階で，回帰分析により機械翻訳システムの TOEIC スコアを計算する．回帰分析には，各 TOEIC 被験者のスコアと次式で定義する被験者勝率を用いる．

$$\text{被験者勝率} = (N_{human} - 0.5 \times N_{even}) \,/\, N_{total}$$

ここで，N_{total} はテストデータに含まれる文数を，N_{human} は各 TOEIC 被験者による翻訳がシステムの翻訳よりも優れていた文数を，N_{even} は翻訳の品質が同等であった文数を表す．均衡する点，すなわち，被験者勝率が 0.5 に対応する TOEIC スコアをシステム能力として求める．

翻訳一対比較法により，システムの TOEIC スコアを得ることができるようになった．しかし，システムの翻訳と TOEIC 被験者による翻訳とを評価者が文単位で比較しているので，膨大なコストがかかる．そこで，複数の参照訳を持つテスト文のセットを用意して，BLEU などの自動評価手法で，各 TOEIC 被験者およびシステムによる翻訳結果を評価し，その評価値を用いてシステム TOEIC 換算点を計算する方法が提案されている[4]．この方法を使えば大幅なコスト削減ができる．**図 -4** に BLEU の値と被験者の TOEIC スコアとの相関を示す．

◙ 人間の翻訳能力を評価

前章とは逆に翻訳システムの自動評価のアプローチを人間の能力測定に適用する研究が始まっている．日本語文を英語にする翻訳文の評価は，英語によるコミュニケ

☆17 mWER は誤り率で値が小さい方が良い．したがって負の相関になる．
☆18 http://www.toeic.or.jp/

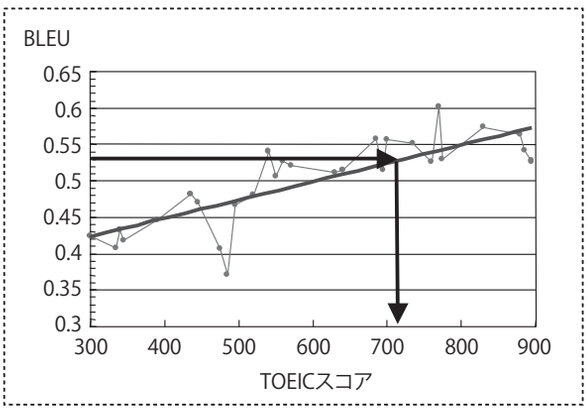

図 -4　自動評価を利用した TOEIC スコアの推定

ーション能力のベースである文生成能力を測定する上で重要である．翻訳課題と『読む』『書く』『聴く』『話す』の 4 技能を計測するさまざまなテストとの関係，被験者の負担を減らすための問題数の削減手法などの研究が行われている．

現在，いわば国際共通語になっている英語によるコミュニケーション能力を高めることは，国家の戦略課題の 1 つであると言われている．ICT を活用して英語教育の改善を図る e-Learning に期待が集まっている．英語の e-Learning にはさまざまな利用形態があり，それに応じて各種の ICT の利用が考えられるが，その中核となる技術の 1 つに，ここで述べた英語能力の自動評価技術がある．

◎ 自動評価の展開と深化

機械による表層的な品質評価が多数のテスト文を使えば人手評価と一定の条件下で高い相関を持つことが発見されてから，これを活用して機械翻訳の研究が急速に進展していることを紹介した．

しかしながら，現在の自動評価は 3 つの点で改良が必要である．

① 現在の評価は，1 文単位の評価に関して高い精度を実現できていない（良い訳と悪い訳をうまく識別できない）．

② ある文を評価するときに，その文が置かれた文脈が考慮されていない．根本的な課題と言えよう．

③ 本稿で説明した自動評価は参照訳を必要としたオフラインの評価である．参照訳を必要としないオンラインの評価は応用範囲も広く期待されていて有望な成果が出はじめたところである [5]．

機械翻訳以外の関連分野への展開も重要である．すでに，要約の分野では BLEU に影響を受けた自動評価手法が使われはじめている [6]．

今後このような課題を克服した評価方法が確立できれば，機械翻訳システムを始めとする自然言語処理システムの品質も【自動的】に改善できると期待され，夢は膨らむ．

IWSLT-2005 への招待

IWSLT は，2005 年秋にはピッツバーグで，2006 年秋には再び京都で，開催される予定である．次回の IWSLT-2005 ☆19 は，今回行われたテキスト入力の評価に加えて，音声認識結果を入力とした評価を予定している．

IWSLT-2004 に参加した 14 団体のうち日本からの参加は東大（と ATR）だけであり，この方面での日本の消極性が危惧される．評価型ワークショップは共通のデータに基づいた議論ができ非常に有用である．また，日本の力を世界に宣伝するという意味もあるので，多くの日本の研究機関の参加を期待する ☆20．

謝辞　本研究は，情報通信研究機構（NICT）の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」，および，科学研究費補助金（課題番号：16300048）により実施したものである．また，本解説の原稿にコメントをいただいた ATR 音声言語コミュニケーション研究所のメンバに感謝する．

参考文献

1）乾健太郎，藤田　篤：言い換え技術に関する研究動向，自然言語処理，11(5), pp.151-198 (2004).
2）Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp.31-318 (2002).
3）菅谷史昭，竹澤寿幸，横尾昭男，山本誠一：音声翻訳システムと人間との音声翻訳能力評価手法の提案と比較実験，信学論，J84-D-II, 11 (2001).
4）Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S. and Yanagida, M.: Applications of Automatic Evaluation Methods to Measuring a Capability of Speech Translation System, In Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics, pp.371-378 (Apr. 2003).
5）Ueffing, N., Macherey, K. and Ney, H.: Confidence Measures for Statistical Machine Translation, In Proceedings of MT SUMMIT IX (2003).
6）Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25-26 (2004).

（平成 17 年 3 月 24 日受付）

☆19 http://www.is.cs.cmu.edu/iwslt2005/

☆20 IWSLT-2004 で使われた，学習用，開発用，テスト用のコーパス，参照訳，各システムの翻訳結果のすべてが GSK（言語資源協会，http://www.gsk.or.jp/）を通じて一般に公開される．これらは評価手法の研究や任意のシステムと IWSLT-2004 に参加したシステムとの比較検討に役立てることができる．