社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－SLP－54（23）
2004／12／21

# ［招待講演］多言語翻訳技術に関する公開性能評価
## －音声翻訳技術のための国際評価ワークショップ IWSLT2004－

中岩　浩巳[†‡]　秋葉　泰弘[‡]　Michael Paul[‡]

†日本電信電話（株）NTT コミュニケーション科学基礎研究所 〒619-0237 京都府相楽郡精華町光台 2-4
‡ATR 音声言語コミュニケーション研究所 〒619-0288 京都府相楽郡精華町光台 2-2
E-mail: †nakaiwa@cslab.kecl.ntt.co.jp, ‡{hiromi.nakaiwa, yasuhiro.akiba, michael.paul}@atr.jp

あらまし　2004 年秋，ATR で行われた音声言語翻訳技術に関する国際評価ワークショップ IWSLT2004 で実施した公開評価キャンペーンの概要を，著者らが同会議の評価委員会メンバとして議論した内容を踏まえて，説明するとともに，本ワークショップで翻訳技術の評価に関し指摘された課題について概説する．
キーワード　機械翻訳，評価キャンペーン，音声言語，自動評価，主観評価

# Open Evaluation Campaign of Multilingual Machine Translation Technology
## －International Workshop on Spoken Language Translation (IWSLT2004)－

Hiromi NAKAIWA[†‡]　Yasuhiro AKIBA[‡]　and　Michael PAUL[‡]

†NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corp.　2-4 Hikaridai, Seika-cho, Souraku-gun, Kyoto, 619-0237 Japan
‡ATR Spoken Language Translation Laboratories, Advanced Telecommunication Research Institute International　2-2 Hikaridai, Seika-cho, Souraku-gun, Kyoto, 619-0288 Japan
E-mail: †nakaiwa@cslab.kecl.ntt.co.jp, ‡{hiromi.nakaiwa, yasuhiro.akiba, michael.paul}@atr.jp

**Abstract**　This paper reports on an overview of International Workshop on Spoken Language Translation (IWSLT2004) held at ATR this autumn. This workshop mainly focused on an evaluation campaign of multilingual (Chinese-to-English, and Japanese-to-English) spoken language translation technologies. The summary of the open evaluation campaign and issues of evaluation techniques of machine translation systems are described.
**Keyword**　Machine Translation, Evaluation Campaign, Spoken Language, Automatic Evaluation, Subjective Evaluation

## 1. はじめに

インターネットの普及など IT 技術の進化や EU 圏の拡大などの様々な言語での会話の機会が増えるにつれ，言語を越えた会話を支援する技術に対する需要が急激に増加している，これに伴い，音声翻訳技術に関する研究が活発になっており，様々なプロジェクトが進められている（例えば，VERBMOBIL, C-STAR, NESPOLE!, BABYLON），これらプロジェクトのうち，C-STAR[1]を以外は，特定の言語対に対する音声翻訳プロトタイプの構築を目的にしているものである．これに対し C-STAR で現在進められているプロジェクトでは，多言語で共通のタスクを対象とする音声言語コーパスの共同開発を進めている．この最初のステップとして，ATR が構築した旅行会話文を収録した日英音声言語コーパスを C-STAR 参加機関によってそれぞれの母国語に翻訳することを進めている．本多言語コーパスは様々な音声翻訳技術開発と評価のための最初の言語資源として活用されている．

音声翻訳システムの効率的及び効果的な研究開発のためには，翻訳システムの翻訳品質評価がきわめて重要である．特に，C-STAR プロジェクトのように共通のコーパスを用いて開発されている翻訳システムでは，翻訳手法の有効性を検討するための評価が必要となる．このようなことから，多言語音声翻訳システムの評価法の確立が強く求められている．

この目的を達成するため，この C-STAR が共同開発したコーパスの一部を活用して，2004 年度の評価キャンペーンを実施した．本キャンペーンでは，中英および日英翻訳を対象に 500 種類のテスト文を用いて機械翻訳するタスクを設定した．参加機関が翻訳システム開発に活用できる言語資源に応じて 3 種類のトラック（Small Data (S), Additional Data (A), Unrestricted (U)）を設定した．各システムの翻訳品質は，人間による評

[1] http://www.c-star.org/

価（主観評価）と自動評価技術による評価（自動評価）により行われた．この評価結果については，本稿の 3 節で概説する．

本年の IWSLT のために提供されたコーパス，参照訳，参加した MT システムの機械翻訳結果，及びそれぞれの翻訳結果の評価結果は本ワークショップ後に一般公開される予定である．この公開されるコーパスは，将来の機械翻訳システム研究および機械翻訳評価法の研究のためのベンチマークとして活用されることを期待している．

今回実施した IWSLT2004 が音声多言語翻訳研究のための標準的評価法と標準的コーパスを確立するための第 1 ステップとなることを期待している．

## 2. 評価キャンペーン IWSLT2004

2004 年の評価キャンペーンは C-STAR パートナーが共同で構築した多言語コーパスの 1 部を用いて実施した（2.1 節参照）．タスクは，500 文の中英翻訳及び日英翻訳である．

利用可能なデータ量によって 3 種類の言語資源条件のトラックを設定した．各参加者は個々のトラックに 1 システムのみの登録を受け付けたが，1 トラックに対して複数の翻訳結果を投稿することは許した．

今年のワークショップには 14 の研究機関が参加し，中英には 20 翻訳システム，日英には 8 翻訳システムの翻訳結果の投稿があった．

翻訳結果は人手による主観評価と自動評価技術を活用した自動評価の 2 種類により評価した．主観評価は英語母国語者により行った．翻訳品質は，流暢さ (fluency)と適切性(adequacy)の 2 種類の観点から評価した．自動評価には 5 種類の自動評価手法(BLEU, NIST, mWER, mPER, GTM) を用いた．すべての翻訳結果の投稿に対して自動評価は行ったが，主観評価に関しては，そのコスト面での理由から，各参加者のトラック毎に参加者自身が選んだ 1 種類の翻訳結果の投稿のみを対象とした．投稿のあった機械翻訳システムの概要は 2.6 節に示す．

### 2.1. 多言語音声言語コーパス

旅行会話基本表現コーパス（Basic Travel Expression Corpus; BTEC）は海外旅行の際に様々な旅行の場面において外国語で会話をするときに役立つと判断した対訳文をバイリンガルの旅行専門家が選択した用例集である．評価キャンペーン IWSLT2004 では，英語，日本語および中国語の BTEC の 1 部を用いた．IWSLT2004 での提供コーパスの詳細を表 1 に示す．各参加者には BTEC からランダムに選択された 2 万文対が訓練セットとして提供された．中英及び日英の訓練セットの英語部は異なったものであり，両者間の重なりは無い．

最大 16 種類の複数翻訳結果を持つ 506 文からなる開発セットも，翻訳システムのチューニングのために参加者に提供された．

| タイプ | 言語 | 文数 | | 平均長 | 単語数 | 単語種類 |
|---|---|---|---|---|---|---|
| | | のべ | 異なり | | | |
| 訓練セット | 中 | 20,000 | 19,288 | 9.1 | 182,904 | 7,643 |
| | 英 | | 19,949 | 9.4 | 188,935 | 8,191 |
| | 日 | 20,000 | 19,046 | 10.5 | 209,012 | 9,277 |
| | 英 | | 19,923 | 9.4 | 188,712 | 8,074 |
| 開発セット | 中 | 506 | 495 | 6.9 | 3,515 | 870 |
| | 日 | 506 | 502 | 8.6 | 4,374 | 954 |
| | 英* | 8,089 | 7,173 | 7.5 | 67,410 | 2,435 |
| テストセット | 中 | 500 | 492 | 7.6 | 3,794 | 893 |
| | 日 | 500 | 491 | 8.7 | 4,370 | 979 |
| | 英* | 8,000 | 6,907 | 8.4 | 66,994 | 2,496 |

* 複数参照訳は自動評価で活用した

表 1　IWSLT2004 BTEC コーパス

最終的な翻訳性能評価のために BTEC からランダムに選択された 500 文からなるテストセットも最後に提供された．提供されたデータのうち中国語部は日英の対訳文から翻訳されたものであり，テストセットに関しては同じ英文に対して日本語文と中国語文が対応付けられている．個々の対訳文については，原言語となる日本語文および中国語文が英語訳の元の文として適切なものか（例えば，原言語文からは生成できない表現が英語文中に含まれていないか）の訳語としての過不足チェックを行い必要に応じて修正した．

日本語文及び中国語文の単語分割情報も各言語の単語分割ツールを持たない機関のために提供した．しかし，各参加者は，トラック毎に設定した使用可能な言語資源の範囲内では，各自が持つ様々な言語資源を活用することを許した．

### 2.2. データトラック条件

利用できる言語資源に応じて 3 種類のトラックを設定した．Small Data トラックでは提供されるコーパスのみの利用に限定した．Additional Data トラックでは中英のみに設け活用できる対訳言語資源（LDC より入手可能なもの）のみに制限を加えた．Unrestricted トラックでは制限を設けなかった．活用できるデータの概要は表 2 のとおりである．

### 2.3. 評価仕様

本ワークショップにおける評価は，最終的に音声翻訳結果であることを考慮して，投稿する翻訳結果としては，下記の条件の文が投稿されることを前提とした．

- 小文字のみ

| 言語資源 | データトラック | | |
|---|---|---|---|
| | Small Data | Additional Data | Unrestricted |
| IWSLT04 | ○ | ○ | ○ |
| LDC | × | ○ | ○ |
| Tagger | × | ○ | ○ |
| Chunker | × | ○ | ○ |
| Parser | × | ○ | ○ |
| 外部対訳辞書 | × | × | ○ |
| 他の資源 | × | × | ○ |

表 2　使用が許された言語資源

- 句読点（'.', ',', '?', '!', '"'）なし
- "-"による単語の繋ぎなし（空文字とする）
- 数字はスペルで表記

評価する側では投稿結果に対するこれらの加工はまったく行わず，そのままで評価作業を進めた．もし英訳中に非 ASCII 文字（日・中の文字）が現れた文が 1 文でも含まれた場合には，その翻訳結果の投稿全体を受け付けなかった．

### 2.4. 主観評価

90 年代に行われた DARPA による機械翻訳評価等のコンテスト形式の機械翻訳評価で採用された人手評価の基準である流暢さ(fluency)と適切性(adequacy)の 2 種類を本ワークショップでも採用した．流暢さ(fluency)は，翻訳文が目的言語で文法的な正しさを示し，適切性(adequacy)は，入力文に記述された情報が翻訳結果で伝わるかを示す．

流暢さ(fluency)と適切性(adequacy)は表 3 のとおり人間によりそれぞれ 5 段階で評価し，翻訳システムが訳文を生成できなかった場合には 0 とした．また，個々の翻訳結果は 3 人の評価者による評価を行い，評価者による評価基準の揺れの排除を目指した．また，評価件数を削減するため，異なるシステムの同じ原文に対する翻訳結果が同じ英訳となった場合には，人手による評価結果も同じものであることから，それらは 1 回のみ評価した（pooling）．今回は最終的に 10 人の英語母国語話者が主観評価作業を行った．

### 2.5. 自動評価

自動評価には 5 種類の自動評価手法（BLEU, NIST, mWER, mPER, GTM）を用いた．詳細を表 4 に示す．翻訳システムの出力は最大 16 種類の参照訳と比較した．比較する前に翻訳出力は，2.3 節で記述した基準で書かれているかをチェックした．また，自動評価の前に翻訳システムの出力と参照訳は広く入手可能な品詞タガーにより品詞情報が付与され，品詞も加味した形で自動評価した．

### 2.6. 評価キャンペーンの参加機関

1 4 の研究機関が評価キャンペーン IWSLT2004 に参加した．翻訳方式のタイプ別システム数は統計翻訳 SMT が 7，用例翻訳 EBMT が 3，ルールベース翻訳 RBMT が 1，複数の翻訳タイプの翻訳エンジンを活用したもの HMT が 4 であった．言語対に関しては，中英には 13 機関が 20 翻訳システム（Small Data:9, Additional Data:2, Unrestricted:9）で，日英には 6 機関が 8 翻訳システム(Small Data:4, Unrestricted:4)で参加した．主観評価はこの 28 システムからの pooling 後の翻訳結果 11,134 文に対し行った．

### 2.7. 評価キャンペーンのスケジュール

評価キャンペーンのスケジュールは表 5 のとおりである．訓練コーパスは公式な翻訳結果投稿期限の約 3 ヶ月前にリリースした．参加者は公式な翻訳結果投稿期限の 1 週間前に開発コーパスを用いた自動評価サーバーを活用して開発したシステムの性能確認が出来るようにした．公式な翻訳結果投稿の期間は 3 日間とし，この 3 日間は自動評価結果を自動評価サーバーから投稿者に電子メールで自動返送する機能をオフにし，この期間の自動評価結果を受けた改善が行えないように配慮した．また，公式な翻訳結果投稿の締め切り後に，その後の改良の評価とその結果の論文への反映が出来るように，評価サーバーは立ち上げたままとし，自動評価結果は自動返送することとした．

## 3. 評価結果

### 3.1. 主観評価結果

本節では，以下の観点で実施した主観評価の結果について報告する．

- 評価者は翻訳結果をどのくらいコンシステントに評価しているか
- 3 評価者からなるグループはどのくらいコンシステントに評価しているか
- 主観評価結果に基づき翻訳システムをどのように評価するか

主観評価結果のバラツキには，各評価者内でのバラツキと評価者間でのバラツキがある．本節では，これ

| Fluency | |
|---|---|
| 5 | Flawless English |
| 4 | Good English |
| 3 | Non-native English |
| 2 | Disfluent English |
| 1 | Incomprehensible |

| Adequacy | |
|---|---|
| 5 | All Information |
| 4 | Most Information |
| 3 | Much Information |
| 2 | Little Information |
| 1 | None |

表 3　人間による評価基準

| | |
|---|---|
| mWER | Multiple Word Error Rate:システム出力と最も似ている参照訳の間の編集距離 |
| mPER | Position independent mWER: mWER の変化形で，語順情報を無視したもの |
| BLEU: | 参照訳に対するシステム出力の n-gram 適合率の幾何平均 |
| NIST: | BLEU の variant. 重み付き n-gram 適合率の算術平均 |
| GTM: | ユニグラムベースの F 値を用いた類似度 |

表 4　自動評価手法

| イベント | 年月日 |
|---|---|
| 評価仕様公開 | 2004 年 2 月 15 日 |
| 参加申し込み期限 | 2004 年 4 月 15 日 |
| 採否の告知 | 2004 年 4 月 30 日 |
| サンプルコーパスリリース | 2004 年 5 月 7 日 |
| 訓練コーパスリリース | 2004 年 5 月 21 日 |
| 開発コーパスリリース | 2004 年 7 月 15 日 |
| 評価サーバー運用開始 | 2004 年 8 月 1 日 |
| テストコーパスリリース | 2004 年 8 月 9 日 |
| 公式な翻訳結果投稿期限 | 2004 年 8 月 12 日 |
| 評価結果のフィードバック | 2004 年 9 月 10 日 |
| 最終原稿の提出期限 | 2004 年 9 月 17 日 |
| ワークショップ | 2004 年 9 月 30 日—10 月 1 日 |

表 5　評価キャンペーンのスケジュール

らバラツキについて分析した後，MT システムをこの分析に基づいてラング付けした．

### 3.1.1. 各評価者内でのバラツキ

本節では各評価者内での主観評価のバラツキについて示す．このために，評価者にはランダムに選択した 100 訳文を 2 度評価させた．同じ文の 2 度目の評価文は 1 度目と連続しないように提示した．バラツキは 1 度目と 2 度目の評価結果の差の平均により評価した．各評価者の 2 度の評価結果の差の期待値を表 6 に示す．また，同じ文に異なる評価を与えた場合の割合（誤り率）を表 7 に示す．評価結果の差の期待値は平均で 0.4 程度あることがわかった．よって，翻訳システムの訳文品質に差があると言うためには 0.8 以上の差がある必要があることがわかった．

表 7 の誤り率は予想以上に高い結果となった．最も誤りの少ない評価者でも 20%程度の誤りがあった．

以上のように表 3 に示した 5 段階評価では，主観評価における多くの誤りが含まれているため，5 段階ではなく，5 段階の内のあるランク以上か未満かの 2 種類に分類する形で評価を行うことを考える．具体的には，(1) 5 と 5 未満，(2) 4 以上と 4 未満，(3)3 以上と 3 未満，(4)2 以上と 2 未満，である．

表 8 に 2 種分類評価において同じ文に異なる評価を与えた場合の割合（誤り率）を示す．このとおり，表 7 の結果に比べて，2 種分類評価では誤り率が低下していることがわかる．特に，5 と 5 未満の評価が最も誤り率が低くなっていることがわかる．

### 3.1.2. メジアン評価のバラツキ

本節では，3 人の評価者からなる各評価チームのメジアン評価のばらつきについて述べる．これは，同じ 100 訳文をすべての評価者に評価させることで調査した．4 種類の評価チーム間のメジアン評価のバラツキの期待値を表 9 に示す．これより，グループ間のメジアン評価のバラツキは平均 0.55 存在することがわかった．

この結果から，2 種類の翻訳システムに有意な品質差があると言うためには，両者間に 1.1 以上の差が必要であることがわかる．

### 3.1.3. 翻訳システムのランキング

3.1.1 節及び 3.1.2 節での議論に基づき，本節では 2 種類のランキングを試みた．

**1. 標準ランキング**

ただ単に，各システムの翻訳結果に対する各チームの 3 人の評価者による評価結果のメジアンをすべての文で平均することで得られる[0,5]の間の評価結果により，翻訳システムをランキングする．今までの議論のとおり，評価者や評価チーム間で評価にバラツキがあるので，これは参考データとして提示する．

**2. 改良型ランキング**

3.1.1 の表 8 の結果のとおり 5 段階評価にくらべて 2 種分類評価の場合はバラツキが少なく，その中でも最もばらつきの少ない「5 か 5 未満か」の割合により翻訳システムをランキングする．さらに，3 人の評価者からなる 4 チームのうち，メジアンによる評価ではなく，最も誤り率の低い評価者の結果を採用した．このように，各チームから最も誤り率の低い評価者を採用した場合の各チームの誤り率を表 10 に示す．本表のとおり，表 8 に示す平均誤り率よりかなり低い値となることがわかる．なお，この評価は 5 の評価が与えられた割合となるので，[0,1]間の評価結果となる．

表 11 に標準ランキングと改良型ランキングによる評価結果を示す．本評価キャンペーンの目的は，研究機関ごとの翻訳性能を競うのではなく，有望な翻訳

| 評価者 ID | 流暢さ | 適切性 |
|---|---|---|
| G0 | 0.21 | 0.33 |
| G1 | 0.37 | 0.39 |
| G2 | 0.35 | 0.44 |
| G3 | 0.49 | 0.38 |
| G4 | 0.34 | 0.34 |
| G5 | 0.22 | 0.44 |
| G6 | 0.77 | 0.64 |
| G7 | 0.29 | 0.44 |
| G8 | 0.44 | 0.44 |
| G9 | 0.46 | 0.55 |
| 平均 | 0.39 | 0.44 |

表 6 評価者別同じ文の評価差の期待値

| 評価者 ID | 流暢さ | 適切性 |
|---|---|---|
| G0 | 0.19 | 0.23 |
| G1 | 0.33 | 0.34 |
| G2 | 0.32 | 0.34 |
| G3 | 0.47 | 0.33 |
| G4 | 0.26 | 0.32 |
| G5 | 0.14 | 0.40 |
| G6 | 0.53 | 0.55 |
| G7 | 0.28 | 0.39 |
| G8 | 0.37 | 0.37 |
| G9 | 0.33 | 0.37 |
| 平均 | 0.322 | 0.364 |

表 7 評価者別の誤り率

| 評価者 ID | 流暢さ | | | | 適切性 | | | |
|---|---|---|---|---|---|---|---|---|
| | 5 or 5 未満 | 4 以上 or 4 未満 | 3 以上 or 3 未満 | 2 以上 or 2 未満 | 5 or 5 未満 | 4 以上 or 4 未満 | 3 以上 or 3 未満 | 2 以上 or 2 未満 |
| G0 | **0.01** | 0.07 | **0.06** | 0.07 | 0.08 | 0.10 | **0.07** | 0.08 |
| G1 | 0.05 | 0.10 | 0.15 | 0.07 | 0.17 | 0.10 | 0.08 | **0.04** |
| G2 | 0.08 | **0.03** | 0.13 | 0.11 | 0.07 | 0.13 | 0.13 | 0.11 |
| G3 | 0.12 | 0.06 | 0.23 | 0.08 | **0.05** | 0.13 | 0.16 | 0.04 |
| G4 | 0.07 | 0.07 | 0.09 | 0.11 | 0.09 | **0.06** | 0.08 | 0.11 |
| G5 | 0.05 | 0.09 | **0.06** | **0.02** | 0.11 | 0.11 | 0.10 | 0.12 |
| G6 | 0.11 | 0.17 | 0.30 | 0.19 | 0.07 | 0.22 | 0.27 | 0.08 |
| G7 | 0.04 | 0.09 | 0.13 | 0.03 | 0.08 | 0.10 | 0.16 | 0.10 |
| G8 | 0.09 | 0.06 | 0.19 | 0.10 | 0.11 | 0.13 | 0.15 | 0.05 |
| G9 | 0.11 | 0.06 | 0.18 | 0.11 | 0.13 | 0.13 | 0.15 | 0.14 |
| 平均 | 0.073 | 0.080 | 0.152 | 0.089 | 0.096 | 0.121 | 0.135 | 0.087 |

表 8 2 種分類評価における誤り率

| | 流暢さ | | | 適切性 | | |
|---|---|---|---|---|---|---|
| | T2 | T3 | T4 | T2 | T3 | T4 |
| Team 1 (T1) | 0.49 | 0.75 | 0.47 | 0.54 | 0.61 | 0.34 |
| Team 2 (T2) | — | 0.68 | 0.66 | — | 0.59 | 0.48 |
| Team 3 (T3) | — | — | 0.44 | — | — | 0.51 |
| 平均 | 0.58 | | | 0.51 | | |

表 9 チーム間の 3 評価結果メジアンの差の期待値

| | 流暢さ | 適切性 |
|---|---|---|
| Team 1 | 0.01 | 0.07 |
| Team 2 | 0.05 | 0.09 |
| Team 3 | 0.05 | 0.05 |
| Team 4 | 0.01 | 0.05 |
| 平均 | 0.03 | 0.07 |

表 10 各チームで誤り率最小の評価者の誤り

手法や評価方法に関する議論が目的で実施しているので，本稿では参加研究機関名は記せず，研究機関毎の番号で区別するのみとした．標準ランキングの表中のシステム評価結果間に挿入された太線は，最も高い評価を与えられた翻訳システムの評価結果と，表７で示したチーム間の誤り率の期待値（流暢さ：0.58，適切性：0.51）が２倍以上の差がある境界を示す．また改良型ランキングでも同様に，最も高い評価を得た翻訳システムの評価結果と，表８で示した誤り率の期待値（流暢さ：0.03，適切性：0.07）の差が２倍以上ある境界を示す．すなわち，この境界より上にあるシステムは，評価者による誤り率を加味すると，最も良いシステムと明確な差があるとはいえないことが推測できることを示している．また，表中の＊は，最も良いシステムの性能と有意な差が無いことを示す．これは5-fold cross validationに基づくT検定により検証した．

標準ランキングの結果を見ると，流暢さにおいては明確な品質差が見受けられる箇所もあるが，適切性においては，各トラックのシステムの評価結果間には太線がほとんど挿入されていないことからも，明確な品質差があるとは言いにくいことがわかった．

改良型ランキングについては，流暢さ適切性ともに，明確な品質差があると判断できる箇所が多い．このことから品質差を明確化するという観点からは，改良型ランキングのほうが適しているといえる．ただ，改良型ランキングは，ほぼ完全な翻訳を行っているか否かだけによりシステム評価を行っていることに相当するため，様々な品質の翻訳結果を総合的に加味した評価とはなっていない．これに関しては，例えば，事前の試行評価作業での誤り率の低い評価者のみによる評価や，評価文の選択基準の見直しなどの工夫が必要となると考えられる．

### 3.2. 自動評価結果

5種類の自動評価手法（BLEU, NIST, mWER, mPER, GTM）を用いて自動評価した結果を表１２に示す．なお，表中の＊は表１１と同様に1位システムと有意な品質差が無いことを示している．

表から，SMTとHMTは高い評価を得る傾向にあること，またSMTが最も高い評価を受ける傾向にあることがわかった．

### 3.3. 主観評価と自動評価の相関

表１３に表１１の主観評価結果と，表１２の自動評価結果の相関係数を示す．結果のうち「改良型ランキング対自動評価（部分）」とは，誤り率の２倍以上の差がある翻訳システムのみに対する相関を示す．全体的傾向として，流暢さに関してはBLEUが，適切性に関してはNIST，mPER，BLEU及びmWERが主観評価と高い相関を示した．また部分的な結果に対する相関値は，予想通り，全てを用いた場合より高い相関が得られることがわかった．

標準ランキング

| トラック | | 流暢さ スコア | 流暢さ MT_ID | 適切性 スコア | 適切性 MT_ID |
|---|---|---|---|---|---|
| 中英 | Unrest-ricted | 3.776 | SMT1 | 3.662 | SMT2 |
| | | 3.776 | SMT2＊ | 3.526 | SMT1＊ |
| | | 3.400 | HMT3 | 3.254 | HMT14 |
| | | 3.036 | SMT4 | 3.188 | EBMT7 |
| | | 2.954 | SMT5 | 3.082 | EBMT6 |
| | | 2.934 | HMT14 | 2.996 | SMT4 |
| | | 2.718 | EBMT6 | 2.960 | RBMT8 |
| | | 2.648 | EBMT7 | 2.800 | HMT3 |
| | | 2.570 | RBMT8 | 2.784 | SMT5 |
| | Additional Data | 3.256 | SMT1 | 3.110 | SMT1 |
| | | 2.846 | SMT5 | 2.724 | SMT5 |
| | Small Data | 3.820 | SMT9 | 3.338 | SMT10 |
| | | 3.356 | SMT10 | 3.088 | SMT1 |
| | | 3.332 | SMT2 | 3.084 | SMT5 |
| | | 3.120 | SMT1 | 3.056 | EBMT7 |
| | | 3.074 | SMT5 | 3.048 | SMT2 |
| | | 2.948 | SMT4 | 3.022 | SMT12 |
| | | 2.914 | HMT11 | 2.950 | SMT9 |
| | | 2.792 | SMT12 | 2.938 | HMT11 |
| | | 2.504 | EBMT7 | 2.906 | SMT4 |
| 日英 | Unrest-ricted | 4.308 | HMT9 | 4.208 | HMT9 |
| | | 4.036 | SMT10 | 4.066 | SMT10 |
| | | 3.650 | EBMT13 | 3.316 | EBMT13 |
| | | 2.472 | RBMT8 | 2.602 | RBMT8 |
| | Small Data | 3.484 | SMT9 | 3.412 | SMT10 |
| | | 3.480 | SMT10 | 3.086 | SMT5 |
| | | 3.106 | SMT4 | 2.990 | SMT4 |
| | | 3.102 | SMT5 | 1.942 | SMT9 |

改良型ランキング

| トラック | | 流暢さ スコア | 流暢さ MT_ID | 適切性 スコア | 適切性 MT_ID |
|---|---|---|---|---|---|
| 中英 | Unrest-ricted | 0.558 | SMT1 | 0.446 | SMT2 |
| | | 0.532 | SMT2＊ | 0.394 | SMT1 |
| | | 0.406 | HMT3 | 0.294 | HMT14 |
| | | 0.326 | SMT4 | 0.254 | EBMT6 |
| | | 0.296 | HMT14 | 0.250 | SMT4 |
| | | 0.286 | SMT5 | 0.228 | HMT3 |
| | | 0.224 | EBMT7 | 0.226 | EBMT7 |
| | | 0.222 | EBMT6 | 0.178 | SMT5 |
| | | 0.180 | RBMT8 | 0.164 | RBMT8 |
| | Additional Data | 0.410 | SMT1 | 0.316 | SMT1 |
| | | 0.284 | SMT5 | 0.212 | SMT5 |
| | Small Data | 0.582 | SMT9 | 0.338 | SMT10 |
| | | 0.420 | SMT2 | 0.296 | SMT9＊ |
| | | 0.390 | SMT10 | 0.290 | SMT2＊ |
| | | 0.356 | SMT1 | 0.284 | SMT1＊ |
| | | 0.344 | SMT5 | 0.268 | SMT5 |
| | | 0.314 | SMT4 | 0.257 | HMT11 |
| | | 0.278 | HMT11 | 0.250 | SMT12 |
| | | 0.246 | SMT12 | 0.232 | SMT4 |
| | | 0.186 | EBMT7 | 2.906 | EBMT7 |
| 日英 | Unrest-ricted | 0.698 | HMT9 | 0.600 | HMT9 |
| | | 0.608 | SMT10 | 0.564 | SMT10 |
| | | 0.506 | EBMT13 | 0.360 | EBMT13 |
| | | 0.170 | RBMT8 | 0.120 | RBMT8 |
| | Small Data | 0.520 | SMT9 | 0.358 | SMT10 |
| | | 0.440 | SMT10 | 0.304 | SMT5 |
| | | 0.368 | SMT4 | 0.262 | SMT4 |
| | | 0.334 | SMT5 | 0.126 | SMT9 |

表１１　主観評価結果

（表中でSMT，EBMT，RBMT，HMTはそれぞれ統計翻訳，用例翻訳，ルールベース翻訳，複数タイプの翻訳エンジンによる翻訳を示す．また，番号はそれぞれの開発研究機関を示す．）

| トラック | | mWER | | mPER | | BLEU | | NIST | | GTM | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | スコア | MT_ID | スコア | MT_ID | スコア | MT_ID | スコア | MT_ID | スコア | MT_ID |
| 中英 | Unrist-ricted | 0.379 | SMT2 | 0.319 | SMT2 | 0.524 | SMT2 | 9.56 | SMT2 | 0.748 | SMT2 |
| | | 0.457 | SMT1 | 0.393 | SMT1 | 0.440 | SMT1 | 7.50 | HMT14 | 0.684 | SMT4 |
| | | 0.525 | SMT4 | 0.427 | HMT14 | 0.350 | SMT4 | 7.36 | SMT4 | 0.671 | SMT1 |
| | | 0.531 | HMT14 | 0.442 | SMT4 | 0.311 | HMT3 | 7.24 | SMT1 | 0.666 | SMT4 |
| | | 0.573 | SMT5 | 0.487 | EBMT7 | 0.275 | HMT14 | 6.13 | EBMT7 | 0.611 | HMT14 |
| | | 0.578 | HMT3 | 0.499 | SMT5 | 0.243 | EBMT7 | 6.00 | RBMT8 | 0.602 | EBMT7 |
| | | 0.594 | EBMT7 | 0.531 | HMT3 | 0.243 | SMT5 | 5.92 | HMT3 | 0.584 | SMT5 |
| | | 0.658 | RBMT8 | 0.542 | RBMT8 | 0.162 | RBMT8 | 5.42 | SMT5 | 0.563 | HMT3 |
| | | 0.846 | EBMT6 | 0.765 | EBMT6 | 0.079 | EBMT6 | 3.64 | EBMT6 | 0.386 | EBMT6 |
| | Additional Data | 0.496 | SMT1 | 0.420 | SMT1 | 0.351 | SMT1 | 7.39 | SMT1 | 0.655 | SMT1 |
| | | 0.572 | SMT5 | 0.480 | SMT5 | 0.311 | SMT5 | 5.82 | SMT5 | 0.632 | SMT5 |
| | Small Data | 0.455 | SMT10 | 0.390 | SMT10 | 0.454 | SMT9 | 8.55 | SMT10 | 0.720 | SMT10 |
| | | 0.469 | SMT9＊ | 0.404 | SMT2＊ | 0.414 | SMT2 | 8.34 | SMT9＊ | 0.694 | SMT9 |
| | | 0.471 | SMT2＊ | 0.420 | SMT9 | 0.408 | SMT10 | 7.85 | HMT11 | 0.685 | HMT11 |
| | | 0.488 | SMT5 | 0.425 | SMT5 | 0.374 | SMT5 | 7.74 | SMT5 | 0.672 | SMT5 |
| | | 0.507 | SMT1 | 0.430 | SMT1 | 0.349 | SMT1 | 7.48 | SMT9 | 0.670 | SMT9 |
| | | 0.532 | HMT11 | 0.451 | HMT11 | 0.346 | SMT4 | 7.12 | SMT4 | 0.665 | SMT4 |
| | | 0.538 | SMT4 | 0.452 | SMT4 | 0.338 | HMT11 | 7.09 | SMT1 | 0.647 | HMT12 |
| | | 0.556 | HMT12 | 0.465 | HMT12 | 0.278 | HMT12 | 6.77 | HMT12 | 0.644 | SMT1 |
| | | 0.616 | EBMT7 | 0.500 | EBMT7 | 0.209 | EBMT7 | 5.95 | EBMT7 | 0.601 | EBMT7 |
| 日英 | Unrest-ricted | 0.263 | HMT9 | 0.233 | HMT9 | 0.630 | HMT9 | 11.25 | SMT10 | 0.824 | SMT10 |
| | | 0.305 | SMT10 | 0.249 | SMT10＊ | 0.619 | SMT10＊ | 10.72 | HMT9＊ | 0.796 | HMT9 |
| | | 0.485 | EBMT13 | 0.420 | EBMT13 | 0.397 | EBMT13 | 7.88 | EBMT13 | 0.672 | EBMT13 |
| | | 0.730 | RBMT8 | 0.597 | RBMT8 | 0.132 | RBMT8 | 5.64 | RBMT8 | 0.568 | RBMT8 |
| | Small Data | 0.418 | SMT10 | 0.337 | SMT10 | 0.453 | SMT10 | 9.49 | SMT10 | 0.764 | SMT10 |
| | | 0.484 | SMT5 | 0.379 | SMT5 | 0.440 | SMT5 | 8.46 | SMT5 | 0.732 | SMT5 |
| | | 0.527 | SMT4 | 0.430 | SMT4 | 0.366 | SMT4 | 7.97 | SMT4 | 0.698 | SMT4 |
| | | 0.614 | SMT9 | 0.570 | SMT9 | 0.364 | SMT9 | 3.41 | SMT9 | 0.539 | SMT9 |

表１２　自動評価結果

（表中で SMT，EBMT，RBMT，HMT はそれぞれ統計翻訳，用例翻訳，ルールベース翻訳，複数タイプの翻訳エンジンによる翻訳を示す．また，番号は開発研究機関を示す．）

| 比較対象 | 言語対 | 流暢さ | | | | | 適切性 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mWER | mPER | BLEU | NIST | GTM | mWER | mPER | BLEU | NIST | GTM |
| 標準ランキング　対自動評価（全て） | 中英 | -0.7124 | -0.5830 | **0.8505** | 0.5995 | 0.5132 | -0.4324 | -0.4404 | 0.4376 | **0.5318** | 0.3711 |
| | 日英 | -0.8867 | -0.7836 | **0.9404** | 0.5995 | 0.6387 | -0.8978 | -0.9376 | 0.7884 | **0.9701** | 0.9401 |
| 改良型ランキング　対自動評価（全て） | 中英 | -0.7214 | -0.6010 | **0.8600** | 0.5950 | 0.5214 | -0.6427 | -0.5779 | **0.7407** | 0.6820 | 0.5136 |
| | 日英 | -0.8252 | -0.7030 | **0.9070** | 0.4871 | 0.5383 | **-0.9690** | -0.9641 | 0.9157 | 0.9176 | 0.9152 |
| 改良型ランキング　対自動評価（部分） | 中英 | -0.8734 | -0.6743 | **0.9548** | 0.5736 | 0.5454 | -1.0000 | **-1.0000** | 1.0000 | 1.0000 | 1.0000 |
| | 日英 | -0.8376 | -0.7223 | **0.9288** | 0.5089 | 0.5632 | -0.9894 | **-0.9984** | 0.9195 | 0.9907 | 0.9977 |

表１３　主観評価結果と自動評価結果の相関係数

（表中で太字は同じ比較対象・言語対において最も相関の高いものを示す）

## 4. まとめ

2004 年秋，ATR で行われた音声言語翻訳技術に関する国際評価ワークショップ IWSLT2004 で実施した公開評価キャンペーンの概要を説明するとともに，本ワークショップで翻訳技術の評価に関し指摘された課題について概説した．今後は，本ワークショップでの議論なども含めた形で論文をまとめるとともに，今回の評価キャンペーンで活用した言語資源の公開を行っていきたい．IWSLT は単発の会議ではなく 2005 年にも開催される予定である．次回は，真の音声言語翻訳の評価キャンペーンとなるように，音声認識部も含めた評価に展開する予定である．


## 謝辞

本研究は，情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」（ATR），the Province of Trento の"the project FUPAT Web Feq"（IRST），the Nature Science Foundation of China の資金番号 60175012 及び 60375018（NLPR）により実施したものである．また，本稿の論文は IWSLT2004 評価委員会の議論に基づきまとめたものであり，本委員長の東京大学辻井潤一教授，同委員の情報学研究所の神門典子教授，及び，伊 IRST の Marcello Federico 博士に感謝する．また，主観評価作業を行っていただいた C-STAR パートナーに感謝する．


## 文　献


[1] Y. Akiba, et. al, Overview of the IWSLT04 Evaluation Campaign, Proc. of IWSLT2004, pp.1-12, Kyoto, Japan. Sep.30-Oct.1 2004.[2]


[2] スペースの制約から，本稿に関連する参考文献は割愛した．本ワークショップの詳細は，下記 URL に掲載している論文，特に，参考文献[1]を参照のこと．
http://www.slt.atr.jp/IWSLT2004/