

# 5 非制限話し言葉翻訳に関する最近の技術進展

Marcello Federico / Diego Giuliani / Gianni Lazzari
(Fondazione Bruno Kessler - FBK-irst)

翻訳：奥村明俊（NEC）

話し言葉翻訳は，自然言語処理研究における最も困難なタスクである．話し言葉翻訳では，音声認識と機械翻訳のそれぞれの課題を解決し，さらにこの2つの技術を融合しなければならない．融合タスクの中心的課題は，音声認識から機械翻訳への誤り伝播の抑制と，話し言葉翻訳の探索における計算複雑性である[2]．過去の研究プロジェクトは，より現実的なタスクとするため，旅行情報の要請や予約のスケジューリングといったドメインを対象として，管理された状況で録音した音声を用いるという制限つきの話し言葉翻訳に取り組んできた．その結果，本特集号でも述べられているようにC-STARコンソーシアムやいくつかの国際研究プロジェクトによって，話し言葉翻訳システムが開発された．

近年，アメリカのGALEプロジェクトとヨーロッパのTC-STARプロジェクトという2つの主要プロジェクトが，いわゆる非制限話し言葉翻訳に取り組み始めた．従来の話し言葉翻訳のプロジェクトは，実験室で録音された話し言葉に注力していたが，TC-STARとGALEの両プロジェクトは，ニュース放送や政治演説のように実生活で録音された話し言葉，found speechの翻訳を対象としている．

我々が非制限話し言葉翻訳に注目するようになったのは，以下の理由からである．

- 機械翻訳における統計的アプローチが目覚ましい進展を遂げ，いわゆるルールベースアプローチによる精巧な従来の手法に対して十分対抗できるようになった．
- 機械翻訳の共通の評価指標と基準の採用により，評価キャンペーンが広まって研究所間で評価結果を共有できるようになった．
- 統計的機械翻訳システムをトレーニングするための大規模対訳コーパスが利用可能となり，性能向上と機械翻訳分野の改革をもたらした．
- 強力なオープンソースツールが利用可能となり，統計的機械翻訳に取り組む研究コミュニティが急速に拡大した．

本稿では，TC-STARプロジェクトの研究課題とプロジェクトで導入された評価の枠組みを紹介し，話し言葉翻訳のコア技術である音声認識や機械翻訳について，筆者らが所属するイタリア・トレントの研究機関FBK-irst（旧ITC-irst）の最近の成果について述べる．

## TC-STARプロジェクト

TC-STAR☆1プロジェクト（Technology and Corpora for Speech to Speech Translation, 2004-2007）は，欧州委員会第6次フレームワークプログラム（FP6）より助成を受け，音声翻訳のコア技術研究を進展させる長期的取り組みとして発足した．音声翻訳技術は，音声認識，話し言葉翻訳および音声合成の結合である．このプロジェクトの目的は，かなり意欲的なもので，音声翻訳におけるブレークスルーをもたらして人間と機械翻訳の性能の差を大幅に低減しようとするものである．また制限されない会話音声のドメイン–政治演説やニュース放送–，そしてヨーロッパ英語，ヨーロッパスペイン語，北京語の3言語をターゲットとして選択した．音声翻訳技術全域において目覚ましい進展をもたらすために，定期的に競争的評価が行われ，その結果は一連のオープンワークショップにて発表され議論された．このような場は，科学コミュニティや企業，特に技術移転やサービス領域で活動する企業の注目を集めた．

このプロジェクトは，音声認識技術，話し言葉翻訳技術，音声合成技術，そして，それぞれの技術の統合といった技術分野の主要メンバを集め，コンソーシアムには，イタリア・トレントのITC-irst（コーディネータ），ドイツ・アーヘンのRWTH，フランス・パリのCNRS-

☆1 http://www.tc-star.org

LIMSI，スペイン・バルセロナの UPC，ドイツ・カルルスエールの UKA，ドイツ・IBM GmbH，ドイツの Siemens AG，フィンランドの Nokia Corp.，ドイツの Sony Int'l GmbH，フランス・パリの ELDA，オランダ・ナイメーヘン KUN-SPEX らが参加し，研究，技術，インフラストラクチャの面でバランスよく貢献している．

### ＊翻訳タスク

TC-STAR は，Voice of America ニュース放送の中国語–英語と英語–中国語の翻訳，および European Parliament Plenary Sessions（EPPS）で録音された政治演説のスペイン語–英語と英語–スペイン語の翻訳という 2 つの実生活のタスクで，非制限話し言葉翻訳の研究を行った（図 -1 参照）．目的は，以下の機能を連結した完全自動処理の開発である：

- 録音音声信号の自動分割
- 複数の音声認識候補（仮説）を表す，音声認識–話し言葉翻訳間インタフェース
- 自動的に挿入された句読点を含む話し言葉翻訳–音声合成間インターフェース

関連する研究課題を明らかにするために，話し言葉翻訳の動作環境を規定した．

- **音声認識誤り**：音声認識の誤りによる性能低下を明らかにするために，音声認識による音声認識結果の翻訳と人手による逐語的書き起こし（VBT）の翻訳を比較した．
- **言語スタイル**：EPPS の演説の翻訳は，欧州議会により発行された最終テキスト版（FTE）と呼ばれるポストエディットにより洗練されたテキストの翻訳結果と比較された．機械翻訳トレーニングのための対訳データは，主として最終テキスト版を基に構成されているので，この対照条件によってトレーニングテキストと評価テキストとの言語スタイルの違いによる性能低下を検証できる．
- **言語ドメイン**：他のドメインへの移行による性能低下を明らかにするために，スペイン語から英語への翻訳に関して，EPPS 演説とスペイン議会（Cortes）演説の翻訳を比較した．

すべての翻訳方向において，トレーニング条件は，参加者間で公平な比較が行われるように整備されている．一般に，対訳コーパスの使用に関する制限はあったが，公に利用可能な単言語コーパスやツールはすべて使用可能である．


図 -1　欧州議会演説を翻訳する TC-STAR デモシステム

### ＊技術評価

プロジェクトの意欲的な目標達成のために，比較評価という戦略的アプローチが導入された．定期的に，音声認識，話し言葉翻訳，音声合成という個別技術と連結システムを競争的に評価する基盤が構築された．年に 1 度の評価キャンペーンでは，共通の言語資源上でかつ同一条件のもと，共同研究者らによる進展を測ることになっている．その進展はプロジェクトで設定された最先端の参照基準に基づいて評価された．評価キャンペーンは外部からも参加可能で，キャンペーンで使用される評価パッケージは公に利用可能となっている．翻訳品質は，人間による判断と，その判断とかなり相関性の高い BLEU スコア [1] などによる自動正確性評価手法により評価された．以前のシステムと比べてどの程度進展したかを評価するため，共同研究者らはそれぞれの評価キャンペーン用に開発した話し言葉翻訳システムを凍結することが求められた．最終公式評価直後に，以前のシステムによる結果が提出された．

3 年間のプロジェクトでさまざまなタスクの翻訳品質は目覚ましい進歩を遂げた．プロジェクト開始時と終了時の BLEU スコアで性能を比較すれば，特定のタスクや入力条件にもよるが，相対的に 40％から 60％の改善が見られる．その性能は，翻訳専門家には到底及ばないが，欧州議会データに対するエラー率は，実生活タスクとしては驚くほど低いものであった．具体的には，最も優れた翻訳システムは，単語の位置を無視すれば約 70％の単語正解率を示した．次章では，このような性能向上をもたらした先端技術について考察する．

## 大語彙多言語音声認識

FBKの大語彙音声認識技術は，FBKで開発された隠れマルコフモデル（HMM）のツールキットに基づいている．混合ガウス分布（Gaussian mixture output densities）を持つCross-word triphone HMMが，音響モデルとして用いられている．音声認識システムは，多段階に動作する．まず音声区間を検出しそれを同類グループにクラスタリングして，入力音声ストリームを分割する．それぞれの音声区間に対する音声認識は，2パスデコーディングによって行われる．最初のパスでは，(i) 特徴空間の最尤線形回帰(MLLR)に基づく音響特徴量の正規化[3]と（ii）ガウス分布平均ベクトルのMLLR適応に基づく音響モデル適応に対し，単語レベルの教師データを提供する．次のパスでは，話者適応化されたモデルを用いて実際の音声認識処理を行う．どちらのデコードパスにおいても，4グラム言語モデルが用いられている．さらに，後段の話し言葉翻訳処理のために，最尤仮説だけでなく，その信頼度スコアと単語ラティスも出力される．

音響モデルと言語モデルのトレーニングデータはTC-STAR評価のオーガナイザにより策定・リリースされた．たとえば，英語–スペイン語翻訳のEPPSタスクに対し約101時間分の書き起こし付き音声データと200時間分の書き起こしなし音声データが利用可能になった．後者には，予備的なシステムによって自動的に書き起こしが付与された．合計約250時間分の音声データが音響モデルトレーニングのために使用された．言語モデルトレーニングでは最終テキスト版（FTE）の3千6百万単語のコーパスが利用可能となった．これらのデータは，音声データの人手による書き起こしとともに，1億6千万から6億7千4百万単語の広いドメインのコーパスでトレーニングされたバックグランド言語モデルをEPPSタスクに適応するために用いられた．同等量のデータがスペイン語–英語翻訳タスクにおいて音声認識システムのトレーニングのために利用可能となっている．

上記のベースラインシステムの進展は，主に，より優れた音響モデリングと音響モデル適応によるものである．

**改良された音響モデリング**：改良された音響特徴量抽出処理は文献3）で提案された話者適応学習アルゴリズムのテキスト非依存型の改良版により実現された．特に我々の最新鋭技術[6]が異分散線形判別分析による音響特徴量のプロジェクションと連結された結果，ベースラインシステムに対して単語の誤り率10％の低減がみられた．

**複数教師適応**：異なる音声認識システムは，ほぼ同等の単語誤り率を示しても，誤り方は異なることがある．この性質を活かして，認識結果の多数決によって認識誤りを低減させるROVER（Recognizer Output Voting Error Reduction）やコンフュージョンネットワークコンビネーションといったシステムを併用する手法が，認識性能を改善するためにしばしば用いられる．システム併用手法も，TC-STAR内で共同研究者により開発された音声認識システムを活用して研究されている．特にFBK/irstは，最終デコードパスを実行する前に複数の音声認識システムより生成された認識仮説を音響モデル適応のために活用する新しいシステム併用技術の実験を行った[6]．提案された音響モデル適応手法は，システムが異なれば認識誤りも異なるという事実に基づいて，認識仮説における誤りの影響を軽減することと補完的情報を教師に与えることを狙いとした．最終的に，認識結果は，複数教師適応が効果的でありROVERやコンフュージョンネットワークコンビネーションにとって代わることを示した．

## 大語彙話し言葉翻訳

機械翻訳に対する統計的アプローチは，単言語テキストとその対訳テキストから得られる観測と確率を取り込んだパラメトリックモデルに基づいている．機械翻訳の現在の最先端手法は，いわゆるphrase-based approachと呼ばれるもので，翻訳単位を1単語から単語の組に拡張した手法である．その中核となる要素は，phrase-pairsの確率を含む翻訳モデル，n-gram単語の確率を取り入れた言語モデル，翻訳元と翻訳先の言語間での単語の並び替えをモデル化したディストーションモデルである．

テキスト翻訳のための機械翻訳システムは，1つの入力仮説のみ処理するよう設計されており，その入力中の誤りに対し脆弱である．TC-STARにおいて研究は，音声認識システムの出力結果が入力となる話し言葉翻訳に注力している．最近では，複数の入力仮説の処理によって翻訳品質を改善するアプローチが提案されている[2]．特に，N-ベストリスト[8]，単語ラティス[7]，コンフュージョンネットワークを用いて，より優れた翻訳性能が得られることが報告されている．コンフュージョンネットワーク[5]用の新しいデコーダが2006年度ジョン・ホプキンズ大学サマーワークショップでTC-STARのメンバによって実装された．探索アルゴリズムはMosesデコーダに統合され，現在一般に公開されている統計的機械翻訳用のツールキットの中で最も人気が高い[4]．

**コンフュージョンネットワークデコーディング**：コンフ

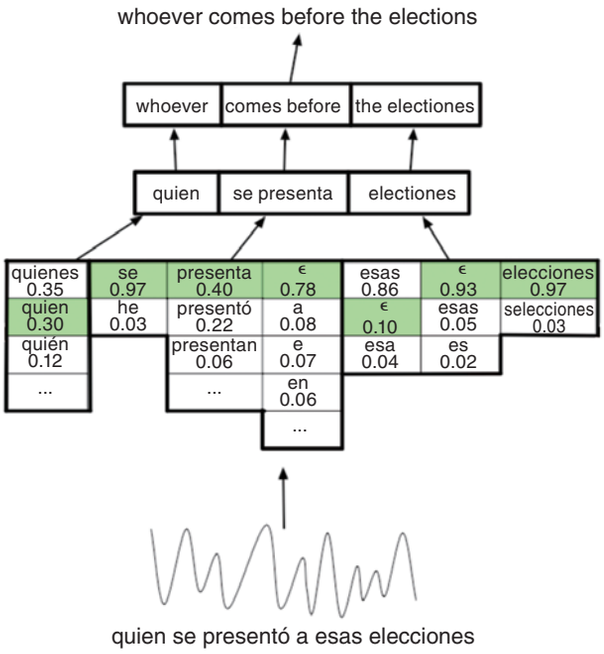


図 -2　コンフュージョンネットワークを利用した翻訳

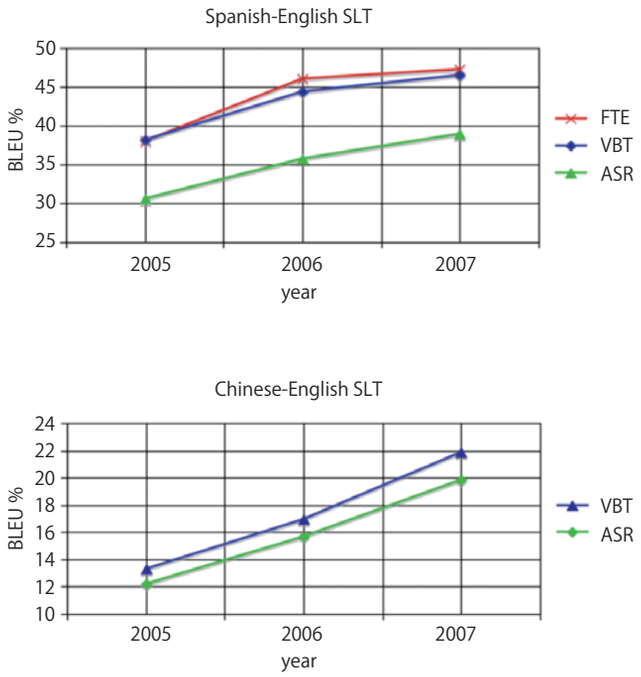


図 -3　FBK/ irst で 2 つの翻訳タスクにおける性能向上

ュージョンネットワークは，複数の音声認識の仮説をコンパクトに表現した形で音声認識システムが生成する．音声信号は数多くの音声認識候補を持つコンフュージョンネットワークにデコードされる．探索アルゴリズムは，入力となるすべての音声認識仮説の中から最も確率の高い翻訳を探索する．コンフュージョンネットワークからの翻訳結果を図 -2 に示す．

図 -2 を下から上に見ると，スペイン語発声は音声認識システムで処理され，システムはコンフュージョンネットワークを表形式で生成する．それぞれのコンフュージョンネットワークのエントリは単語と事後確率を含んでいる．したがって音声認識システムより与えられた多数の仮説は，単純にそれぞれの列の 1 つのエントリを選ぶだけで生成される．空語 (?) に応じたエントリが，異なる長さの仮説を生成するために導入されている．探索アルゴリズムは，すべての可能な入力パスを探索して最も確率が高い翻訳を見つける．1 つの入力仮説の翻訳に対し，コンフュージョンネットワークによる翻訳は，原則としてグラフにあるすべての可能な入力パスの探索を必要とする．ここで鍵となる知見は，線形構造のおかげでコンフュージョンネットワークのデコーディングがテキストのデコーディングにかなり類似していることである．デコーディングの間，探索処理は，区間ごとの翻訳選択肢，つまり元の位置に隣接するシーケンスを調べなければならない．テキストのデコーディングには 1 つの区間に対しちょうど 1 つの語句が存在するのに対し，コンフュージョンネットワークのデコーディングには 1 つの区間に対し複数の語句があるという点で，コンフュージョンネットワークデコーディングとテキストデコーディングは大きく異なる．

**大規模言語モデル**：言語モデルは，音声認識と機械翻訳システムの基本的構成要素である．また，大規模な n-gram 言語モデルを用いれば性能面で大きな成果があることが実験的に示されている．そこで，話し言葉翻訳の活動として，大規模な言語モデルを推定しアクセスするための効率的なデータ構造とアルゴリズムの開発にも注力している [9]．

## 翻訳性能

3 年にわたるプロジェクトの結果，2 つの翻訳タスクに関して FBK/ irst によって達成された話し言葉翻訳の性能向上を図 -3 に示す．翻訳タスクは，EPPS の演説のスペイン語から英語への翻訳と Voice of America のニュースの中国語から英語への翻訳である．BLEU スコアは，例年の評価のために開発された話し言葉翻訳システムを起動して 2007 年の評価セットで計算したものである．これらのスコアは音声認識の数年間の進展を含んでいないことに注意すべきである．EPPS タスクに対してのみ最終テキスト版 (FTE) の結果もあるが，逐語的書き起こし (VBT)，音声認識 (ASR) といった異なる入力条件のもとでの結果を示している．BLEU スコアは，2005 年から 2007 年の間に開発されたブラインド評

中国語からの英語訳

| | |
|---|---|
| Human | Primakov made the above announcement after a meeting held by government on Saturday. |
| MT 05 | Primakov on Saturday at a time when the Government after the meeting as well as making the announcement. |
| MT 06 | Primakov on Saturday held a meeting made following the government announced. |
| MT 07 | Primakov on Saturday at the first meeting after the government made the above announcement. |

スペイン語からの英語訳

| | |
|---|---|
| Human | What is happening with the adopted plan to prevent and combat trafficking in human beings? |
| MT 05 | What is happening with the plan was adopted for preventing and combating trafficking in human beings ? |
| MT 06 | What is happening with the plan adopted for preventing and combating the trafficking of human beings ? |
| MT 07 | What is happening with the plan adopted to prevent and combat the trafficking of human beings ? |

**図 -4　2005 年から 2007 年に FBK/irst で開発されたシステムによる翻訳例**

価用話し言葉翻訳システムによって 2007 年テストセットで計算されたものである．

数年にわたり大きな進展が見られたが，最も目覚ましい改善は，最も困難な翻訳タスクである中国語から英語への翻訳においてであった．逐語的書き起こし（VBT）と音声認識（ASR）という条件下で，BLUE スコアは 2005 年から 2007 年にかけて，それぞれ 64.6%，62%相対的に改善された．EPPS タスクのスペイン語から英語への翻訳タスクでは，BLUE スコアが 2005 年から 2007 年にかけて音声認識（ASR）で 27%，逐語的書き起こし（VBT）で 21%，最終テキスト版（FTE）で 24%相対的に改善された．BLUE スコアは，2 つの翻訳方向の間で性能レベルが大きく異なることを明確に示している．実際，スペイン語からの翻訳は平均的にかなり可読性の高いテキストであるが，中国語からの翻訳については必ずしもそうではない．図 -4 に両方の言語対の翻訳例を示す．

## 結論

話し言葉の翻訳は，過去数年において進展が見られたが，いまだ困難なタスクであることに変わりはない．TC-STAR プロジェクトは，欧州において，系統的かつ組織的に手ごわい研究課題に取り組む類のない機会である．FBK-irst や他の共同研究者によるプロジェクトの重要な成果は，評価基準やオープンソースソフトとして研究コミュニティで入手可能となっている．話し言葉翻訳発展のため，将来もこのようなプロジェクトが続くことを願ってやまない．

**謝辞**　本成果の一部は，European Commission の TC-STAR プロジェクト Technology and Corpora for Speech to Speech Translation Research（IST-2002-2.3.1.6, http://www.tc-star.org）と，2006 年度ジョン・ホプキンズ大学サマーワークショップの支援によるものである．

**参考文献**

1) Papineni, K., Roukos, S., Ward, T. and Zhu, W. : BLEU : A Method for Automatic Evaluation of Machine Translation, IBM Thomas J. Watson Research Center, Technical Report RC22176 (2001).
2) Casacuberta, P., Federico, M., Ney, H. and Vidal, E. : Recent Efforts in Spoken Language Translation, IEEE Signal Processing Magazine (to appear)(2008).
3) Gales, M. J. F. : Maximum Likelihood Linear Transformations for HMM-based Speech Recognition, Computer Speech and Language, 12 (2), pp.75-98 (1998).
4) Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. : Moses : Open Source Toolkit for Statistical Machine Translation, Proc. of ACL ? Demos & Posters, pp.177-180, Prague, Czech Republic (2007).
5) Bertoldi, N., Zens, R. and Federico, M. : Speech Translation by Confusion Network Decoding, Proc. ICASSP, pp.1297-1300, Honolulu, USA (2007).
6) Giuliani, D. and Brugnara, F. : Experiments on Cross-System Acoustic Model Adaptation, Proc. IEEE ASRU Workshop, pp.117-122, Kyoto, Japan (2007).
7) Mathias, L. and Byrne, W. : Statistical Phrase-based Speech Translation, Proc. ICASSP, pp.561-564, Toulouse, France (2006).
8) Zhang, R., Gikui, G., Yamamoto, H., Watanabe, T., Soong, F. and Lo, W. K. : A Unified Approach in Speech-to-speech Translation : Integrating Features of Speech Recognition and Machine Translation, Proc. COLING, pp.1168-1174, Geneva, Switzerland (2004).
9) Federico, M. and Cettolo, M. : Efficient Handling of N-gram Language Models for Statistical Machine Translation, Proc. ACL Workshop on Statistical MT, pp.88-95, Prague, Czech Republic (2007).

（平成 20 年 4 月 14 日受付）

**Marcello Federico**
federico@fbk.eu

1987 年ミラノ大学コンピューターサイエンス学科卒業．Fondazione Bruno Kessler 科学技術研究所の Human Language Technology 研究ユニットを統括．統計機械翻訳，話し言葉翻訳，統計言語モデル，情報検索，音声認識の研究に従事．

**Diego Giuliani**
giuliani@fbk.eu

1986 年ミラノ大学コンピューターサイエンス学科卒業．Fondazione Bruno Kessler 科学技術研究所の上級研究員．音声認識，話者適応，マイクロフォンアレイなどの研究に従事．

**Gianni Lazzari**
lazzari@fbk.eu

1977 年ボローニャ大学電子工学科卒業．Società Consortile Distretto Tecnologico Trentino の CEO．NESPOLE! プロジェクト推進責任者．音声翻訳など話し言葉に関する研究に従事．

**奥村　明俊（正会員）**
a-okumura@bx.jp.nec.com

1986 年，京都大学大学院工学研究科修士課程修了．同年，NEC 入社．機械翻訳や情報抽出など自然言語処理，音声翻訳，ロボットエージェントの研究開発に従事．現在，共通基盤ソフトウェア研究所にてメディアプロセシング，情報センシング，音声言語，情報セマンティクスの研究グループを統括．工学博士．