解　説

機械翻訳最新事情：（下）評価型ワークショップの動向と日本からの貢献

塚田 元[1]　永田 昌明[1]　隅田 英一郎[2]　黒橋 禎夫[3]

1 NTT コミュニケーション科学基礎研究所
2 情報通信研究機構／ATR 音声言語コミュニケーション研究所
3 京都大学

近年，統計的機械翻訳研究コミュニティが中心となって，コンテスト形式の評価型ワークショップが開かれており，（上）統計的機械翻訳入門で紹介した統計翻訳技術の急速な進歩を後押しする立役者となっている．本稿では，代表的な評価型ワークショップを紹介するとともに，これを背景に進展した自動評価などの技術動向を解説する．また，これらのワークショップに日本から参加している研究機関の翻訳システムを紹介することで，日本における統計的機械翻訳研究の動向も合わせて報告する．

はじめに

（上）統計的機械翻訳入門では，過去から現在に至る統計的機械翻訳（以下，統計翻訳）の主要な研究成果を解説した．（下）となる本稿では，この技術進展を強力に後押しした評価型ワークショップに関連する話題を紹介する．評価型ワークショップは，参加者に共通の学習データを提供し，そのデータを用いて作成したシステムを定量的に評価して競わせるコンテスト型のワークショップである．本稿では，評価型ワークショップというレース場に集まってくる統計翻訳システムたちを追うことで，試行錯誤しながら今まさに進められている研究の状況をお伝えしたいと考えている．最初に，世界の代表的な評価型ワークショップを紹介し，これを背景に進展した自動評価などの技術動向について解説する．最後に，評価型ワークショップに日本から参加している研究組織の翻訳システムを紹介することで，日本における統計翻訳研究の動向を報告する．

機械翻訳といえば，よく耳にする言葉に用例に基づく翻訳（用例翻訳）がある．用例翻訳，統計翻訳のどちらも，互いに翻訳になっている 2 つの言語の文の対（対訳または用例）を集めたデータ（対訳コーパス）に基づいて，機械翻訳を実現する技術である．前者が翻訳しようとする文に類似する対訳コーパス中の用例を見つけ出し，それを活用する手法として発展してきたのに対し，後者は用例を生成する統計モデルを活用する手法として発展してきた．このような生い立ちの違いから，前者が初期の段階から構文などの言語学的な情報に重きをおいて研究を進めてきたのに対し，後者は初期の段階では表層的な情報だけを用いていた．そのため，研究コミュニティが別々に分かれてしまう傾向が見られた．しかし，（上）統計的機械翻訳入門でも解説したように，近年，統計翻訳コミュニティでも構文情報を積極的に利用するようになっており，両者の違いはなくなりつつある．そこで本稿では，用例翻訳の最新成果も合わせて紹介することとしたい．

評価型ワークショップ

統計翻訳の研究は 1980 年代の終わりに IBM で始まり，約 20 年もの歴史を持っている．しかし，統計翻訳は計算量が大きく，大量の学習データが必要不可欠であるた

め最初の約 10 年間は IBM の後を追う研究はほとんどみられなかった．この様相が一変したのが，90 年代後半から 2000 年にかけてである．高性能な計算機が普及して計算量の問題が解決されるとともに，大量の学習データが利用可能になり，急速に研究が進展することとなった．この急進展を後押ししたのが，評価型ワークショップである．

評価型ワークショップの果たした大きな役割に，(1)統計翻訳の研究に必要な大量の学習データを提供したこと，および(2)共通タスクを設定したことの 2 つがある．前者のおかげで研究の基盤が整い，後者のおかげで翻訳手法の定量的な比較が厳密に行えるようになった．この相乗効果により，アルゴリズム研究が加速し，統計翻訳システムの性能が急速に向上することとなった．

本章では，統計翻訳の技術進展に貢献した代表的な評価型ワークショップを紹介する．紹介するものの中には，評価型ワークショップというよりは，コンテスト型の評価プロセスを採用した研究プロジェクトも含まれているが，本稿ではこれも評価型ワークショップと区別せずに同列に並べて紹介することにする．翻訳対象も単なるテキストだけでなく音声を対象としたものも数多く存在するが，これも特に区別することなく紹介したいと思う．

◆ NIST 主催評価型ワークショップ（MTE）

米国 NIST（National Institute of Standards and Technology）主催の評価型ワークショップ（NIST Machine Translation Evaluation，以下 MTE）は，2001 年に DARPA（Defense Advanced Research Projects Agency）の TIDES プロジェクト（2001 ～ 2005）の一部として始まり，TIDES 終了後も継続して開催されている．軍事的な背景もあり，言語としてはアラビア語から英語への翻訳（ア英翻訳）と中国語から英語への翻訳（中英翻訳）が主な対象である．分野としては新聞記事や放送ニュースが主な対象であり，膨大な学習データが LDC（Linguistic Data Consortium）より供給される．たとえば，2006 年の学習用対訳コーパスは，中英約 900 万文（約 2 億単語），ア英約 400 万文（約 1 億単語）もの規模を誇る．

本ワークショップは，世界最大の学習データ量を誇り，コンテストとしてもきわめて競争が激しい．本ワークショップのタスクは，統計翻訳手法を評価する共通タスクとして，事実上の業界標準となっている．

MTE だけでなく他の評価型ワークショップでも状況は同じであるが，Google，IBM といった機械翻訳を得意とする企業系研究機関とまったく対等に CMU[☆1]，ISI[☆2]，RWTH[☆3] といった大学の研究室が成績を競い合っている．成績を上げるためには，組織力もさることながら「知恵」を出すことがより重要であることを物語っている．これは，従来の翻訳ルールを専門家が開発するアプローチでは考えられなかったことである．2005，2006 年の公式結果は，NIST の Web ページ[☆4] を参照されたい．

◆ GALE プロジェクト

GALE（Global Autonomous Language Exploitation）は米国 DARPA 主催の研究プロジェクトで，多言語のテキストおよび音声データを翻訳し，そこから軍事アナリストが必要とする情報を抽出することを目指している．プロジェクトは 2005 年に開始され，現在も継続して進められている．要素技術を競うというよりは，音声認識（Speech-to-Text），機械翻訳（Machine Translation），情報蒸留（Distillation）[☆5] を統合したシステムの性能を競うところに主眼がある．最終的には，仮想敵国の言語（基本的に中国語とアラビア語）の新聞，ニュースグループ記事，放送ニュース，放送会話をタイムリーに英語へ翻訳して蓄積し，アナリストが自由に情報を検索し，内容を把握，分析，判断することの支援を目指している．プロジェクトは，大学を含む多くの研究機関が SRI[☆6]，BBN[☆7]，IBM を代表とする 3 つのチームに分かれて性能を競い合う形式で進められる．年度ごとに最終的な翻訳結果と情報抽出結果の目標値が与えられており，そこへの到達度で各チームは評価される[☆8]．GALE は，EARS（音声認識）や TIDES（機械翻訳，情報抽出，自動要約）といったプロジェクトの後継として開始されたもので，年間 50M ドル近い研究資金が投入されている．非常に大きな研究資金源となっており，米国の統計翻訳研究はこの資金が後押ししているといっても過言でない状況にある．

◆ IWSLT

IWSLT（International Workshop on Spoken Language Translation）は，音声翻訳研究のコンソーシアムである C-STAR III のメンバが中心となり 2004 年から

☆1 Carnegie Mellon University.

☆2 University of Southern California's Information Sciences Institute.

☆3 Rheinisch-Westfälische Technische Hochschule Aachen.

☆4 http://www.nist.gov/speech/tests/mt/doc/index.htm

☆5 情報蒸留とは，与えられたクエリーに関係する情報を検索し，指定の長さに要約して提示する技術．

☆6 米国の独立系研究機関．スタンフォード大学と民間資本の共同で Stanford Research Institute として設立され，現在は同大学から独立している．

☆7 研究開発サービスを提供する企業．DARPA と関係が深く，インターネットの原型となった ARPANET の開発でも有名．

☆8 この秋にプロジェクト 2 年目を迎えたが，この節目の評価でア英，中英の両方の目標を達成したチームは 1 つもなく，現在再評価が進められている．早くもプロジェクト存続が危ぶまれている．

毎年開催している音声翻訳のワークショップである．日本からはNICT-ATRが中心メンバの1つとして，開催に貢献している．旅行会話を対象とした音声翻訳を目指しており，現時点では音声認識結果からテキストへの翻訳というタスクでコンテストを開催している．扱う言語は年によってやや異るが，2006および2007年は日本語，中国語，アラビア語，イタリア語の各言語から英語への翻訳という4つの言語対が設定された．

学習用対訳コーパスは2～4万文であり，NIST MTEなどと比較すると2桁小さな規模である．しかしながら，旅行会話というコンパクトなタスク設定のため，かなり精度の高い翻訳が可能である．音声認識結果の翻訳タスクとしてだけでなく，統計翻訳の入門用として，さらには，計算量の大きい挑戦的なアルゴリズムの基本検討用としてもふさわしいタスク設定となっている．

◆その他のワークショップ

以上紹介したもののほかに，ACL[☆9]やHLT-NAACL[☆10]主催の機械翻訳ワークショップ，それからヨーロッパの音声翻訳プロジェクトであるTC-Star[☆11]等でもヨーロッパ言語を中心とした（TC-Starは中国語も含む）共通タスクを設定し，コンテストを開催している．TC-Starは2004年に始まりヨーロッパのGALE的な存在であったが，残念なことにこの2007年にプロジェクトは終了した．2007年10月からは日本が主体となりNTCIR[☆12]主催の特許翻訳のコンテストも始まった．タスクは日本語と英語，双方向の翻訳であり，2008年12月にはワークショップが開催予定である．これで，ようやく日本語に関しても100万文を超える大規模学習データの共通タスクが設定されることとなった．

評価型ワークショップの技術動向

◆翻訳の自動評価

評価型ワークショップが後押しした研究分野に翻訳の自動評価がある．人間による主観評価は，流暢さ（fluency）や適切さ（adequancy）などのさまざまな要素を総合的に判断する．これは最も信頼できる評価方法であるが，時間もお金もかかる．その一方，評価スコアの一貫性を保つことが難しく，前回のコンテストの結果と今回のものを比較するのが困難であるという問題もある．このような人間による主観評価の欠点を補うために，近年，BLEU，NISTスコア，METEOR，TERなど，低コストかつ高速に計算でき，人間による主観評価との相関が高い自動評価尺度がいくつか提案されている．これらの自動評価尺度は，システムの開発と評価を短いサイクルで繰り返すことを可能にし，機械翻訳の研究開発に変革をもたらした．本稿では事実上の業界標準となっている評価尺度BLEU（Bilingual Evaluation Understudy）について，やや詳しく解説したいと思う．

参照訳： 言語 は コミュニケーション の 道具 で ある
候補1： 言語 です ある 道具 の 通信
候補2： 言語 は 通信 の 道具 で ある

図-1　参照訳と翻訳候補

BLEUは，機械による翻訳はプロの翻訳者による翻訳（参照訳，reference）に類似しているほど良いと考え，類似度を0から1の間の数値で表す．具体的には，システムが出力した1つの翻訳候補と正解集合（複数の参照訳）を比較し，各長さの“単語n-gram”の適合率（precision）p_nの幾何平均を求め，短い文へのペナルティBPで補正したスコアとして定義される（式(1)）．

$$BLEU = BP \times \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log p_n\right) \qquad (1)$$

ここで単語n-gramとは連続するn個の単語からなる列であり，単語n-gramの適合率p_nとは候補に含まれるすべての単語n-gramのうち正解集合に含まれる単語n-gramと一致したものの割合である．原論文で，N＝4のときに人間による主観評価と相関が高かったと主張されていることもあり，Nは通常4が使われる．

たとえば，"language is a means of communication"という英文に対して，**図-1**の1つの参照訳と2つの翻訳候補を考えると，候補1では$p_1 = 4/6$，$p_2 = 0/5$，候補2では$p_1 = 6/7$，$p_2 = 4/6$となる．

一般に単語unigram（1-gramのこと）の適合率は適切さ（訳語の精度）に関連し，長い単語n-gramの適合率は流暢さ（語順の精度）に関連する．また同義語や言い替えに対応するために参照訳は4つ以上が望ましいとされる．単語n-gramの適合率の定義からも明らかなように，参照訳の数が増えると，BLEUの値は高くなる傾向がある．そのため，BLEUの値の良し悪しを判断する際には，参照訳の数を考慮する必要がある．

翻訳候補が正解より長い場合，単語n-gramの適合率は低下する．しかし，翻訳候補が短い場合，適合率では不適切さを評価できない問題がある．たとえば，不確かな訳語を一切出力しない極端に短い文は適合率の観点か

☆9 The Association for Computational Linguistics.
☆10 Human Language Technologies - The Annual Conference of the North American Chapter of the Association for Computational Linguistics.
☆11 Technology and Corpora for Speech to Speech Translation.
☆12 NII Test Collection for IR Systems.

ら有利になりがちである．これを補正するのが簡易化ペナルティ（brevity penalty）BP の役割である．翻訳候補が参照訳より長いときは 1 を，短くなればなるほど 1 より小さい値をとる．BLEU は大胆な簡略化に基づいており，1 文単位のスコアは決して信頼できるものではない．しかし，テストセット全体に対して算出したスコアは，人間の主観評価と相関が高くなる．

BLEU は類似の統計翻訳手法同士の優劣を評価するのにはそれなりに使える尺度であると感じている．しかし，言語学的な理由付けの乏しさから BLEU に対する批判も根強い．実際，統計翻訳システムとルールベース翻訳システムのように，根本的に異なる手法同士の比較はできないことが，評価型ワークショップにおいて BLEU 自動評価と人間による主観評価を比較することで明らかになりつつある．より厳密な評価のためには複数の異なった自動評価尺度を併用することが必要である．そして，言うまでもないが，さらに人間による主観評価を併用することが望ましい．

評価型ワークショップの副産物として，各システムの翻訳結果と人間による主観評価の対応データも集まりつつある．このデータをもとに，BLEU の改善を試みる研究も活発になっている．その結果，NIST スコア，METEOR，TER などの評価尺度が提案されている．NIST スコアは，BLEU と同じ考え方に基づいているが，幾何平均の問題点☆13 を改善するとともに，情報量の大きな単語（頻度の低い単語）をより重要視する尺度である．METEOR は，翻訳候補と参照訳との明示的な単語対応および同義語を考慮した尺度である．TER は，まとまった単語列の移動を 1 つのエラーとして数える編集距離で，翻訳結果を正しい翻訳に修正するコストを評価する尺度である．各々の尺度の原論文は，BLEU と比べてより人間の主観評価との相関が高いことを主張しているが，評価型ワークショップの結果を見るかぎり，あらゆる言語対，あらゆるドメインで BLEU を凌駕するところまでは至っていない．決定的な代案がないこともあり，不完全さを指摘されながらも，初期に提案された BLEU が自動評価尺度の事実上の業界標準となっている．機械翻訳の自動評価手法について，さらに興味のある読者は，文献 8）を参照されたい．

その他の技術動向

（上）統計的機械翻訳入門では，句に基づく翻訳モデルおよび構文に基づく翻訳モデルを解説した．現時点では，大多数の参加システムは句に基づく翻訳モデルを採用しており，競争の激しい NIST MTE で 1 位のシステムも，句に基づく翻訳モデルに基づいている．その一方，徐々に構文に基づく翻訳モデルを採用するものも上位の成績を収めるようになってきている．構文に基づくシステムには，(1)構文を対訳コーパスから自動獲得する手法，(2)目的言語側（翻訳先言語）または原言語側（翻訳元言語）を汎用の構文解析器で解析した結果に基づき，翻訳モデルを作成する手法，さらには(1)と(2)のハイブリッド手法などが研究されている．（上）で説明した Chiang の階層的な句に基づく手法や後述する NTT システムは (1) に分類される．また，後述する京大システムは (2) に分類される．

統計翻訳では翻訳モデルに言語モデルを併用するが，評価型ワークショップで成績を競う過程でこの言語モデルの貢献が非常に大きいことが分かってきた．2007 年の NIST MTE では Google 社が提供する全世界の Web（1T トークン）から学習した英語 5-gram を利用した競争が繰り広げられることになっている．このような巨大な n-gram は単純な実装では計算機の主記憶に載らないため，分散実装や圧縮実装の研究が活発になりつつある．

音声認識のコンテストにおいては，複数のシステムの結果を統合して成績を上げることが広く行われてきた．統計翻訳においても同様のアプローチが，近年盛んに研究されるようになりつつある．

日本からの貢献

これまで日本からは NICT-ATR，NTT が 2004 年から 2006 年まで毎年 NIST MTE に参加している．また，NICT-ATR, NTT, OKI，東大（現在は京大にて研究を継続），奈良先端大（NTT と共同），長岡技術大学（NTT と共同），鳥取大学が IWSLT に参加している．さらに NICT-ATR は，2007 年に TC-Star にも参加をしている．本稿ではこれらの研究機関の中から，統計翻訳／用例翻訳手法に基づいて最も活発に研究を続けている NICT-ATR，NTT，京大の 3 組織の研究を紹介することで，日本の研究動向を紹介したいと思う．

NICT-ATR は，句に基づく翻訳手法をベースに，ドメイン適応などモデル学習に関する研究を進めている．NTT は対訳コーパスから自動獲得される構文に基づく手法を中心に研究を進めている．京大は，汎用の構文解析から得られる構文情報を最大限活用する翻訳手法を研究している．句か構文かという観点で見た世界の統計翻訳の研究状況が，ちょうど日本の中にも投影されているかのようである．

☆13 1 つの p_n が 0 であるだけで平均値が 0 になってしまう問題．

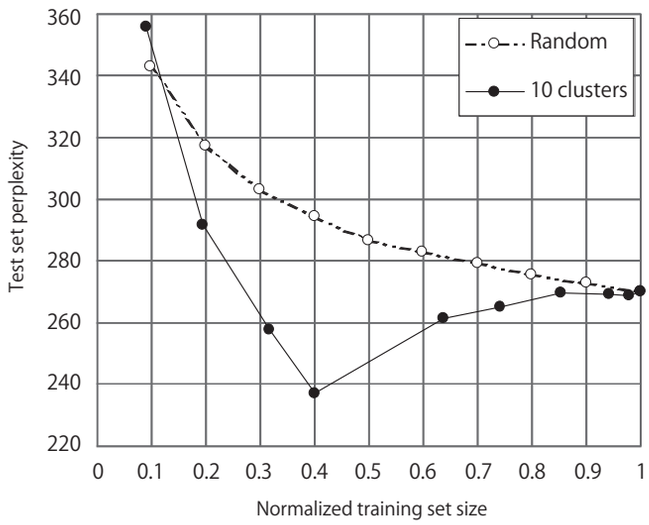

図 -2　コーパスの大きさとテストセットパープレキシティ

◆ NICT-ATR システム

NICT-ATR が，2006 年，2007 年に開催された評価型ワークショップ（IWSLT，NIST MTE，TC-Star）に参加した際のシステムについて解説する．一連のシステムは，句に基づく統計翻訳の典型的な枠組みである対数線形モデルをベースとしており，7 つの素性（句翻訳確率，逆句翻訳確率，単語翻訳確率，逆単語翻訳確率，句ペナルティ，言語モデル確率，句歪み確率，単語ペナルティ）を用いている．また，翻訳の実行は，内製したデコーダ（CleopATRa と名づけた）を活用している．以下，NICT-ATR による新規な試みを 2 つ紹介する．

訓練文選択

近年，非常に大規模なコーパスが利用可能となってきているが，コーパスの大規模化により，統計モデルの性能が向上するというメリットがある反面，学習に要する処理時間やメモリ量が増大するという問題が生じている．

NICT-ATR ではこの問題を解決するため，大規模なコーパスの中から，対象とずれたデータなど雑音的なデータを除去することにより，得られる統計モデルの性能を担保しつつ，学習データの量を減らし，統計モデルの学習に要する計算機的負担を軽減させる手法を提案した [6]．

提案手法では，統計モデルを用いるアプリケーションにおいて対象とするドメインに属する文を集めた小規模なコーパスを用いる（開発用セットと呼ぶ）．一方，大規模コーパスは，ある特定のドメインに属する文だけではなく，種々のドメインに属する文からなる多種多様なコーパスである．

提案手法の考え方は，文クラスタリングにより大規模コーパスを特性の近いものごとのサブセットに分け，次に，各サブセットと開発セットの類似性をパープレキシティ☆14 により測定し，開発セットに類似したサブセットを統計モデルの学習セットとして用いるというものである．

LDC コーパスと TC-Star のデータを用いた実験の結果を**図 -2** に示した．横軸は類似性の高い方から順にサブセットを追加していったときの，コーパス全体に対する割合を示し，縦軸はモデルのテストセットパープレキシティを示している．実線が提案手法の性能を示しており，40% のところで，全コーパスを学習したときより低い最良値を達成している．この後の性能悪化は，異なるドメインのデータや雑音に帰着できる．この実験では，提案手法により学習セットのサイズを 60% 程度削減することが可能となり，統計モデル学習に必要となる処理時間を短縮するだけでなく，統計モデルの性能を改善することも可能であった．

文のドメイン推定とモデルのドメイン適応

統計モデルに基づくシステムの性能はソースとモデルがずれていると劣化する．逆にソースのドメインに合致したモデルを利用することで性能を向上できることが知られている．しかし，ドメインは，未知であったり，一定でなかったりする．したがって，ドメインを動的に推定し，かつ，推定したドメインにあったドメイン依存モデルを用いる必要がある．

提案法 [5] は，2 つのプロセスからなる．(1) オフラインプロセス：訓練データであるバイリンガルコーパスを

☆14 通常は言語モデルの評価に用いる尺度で，1 単語当たりの平均分岐数を表す．ここでは，2 つのコーパスの近さを測るため，一方で言語モデルを作成し，これを使ってもう一方のコーパスのパープレキシティを求める．この値が近いほど，2 つのコーパスは近いものだと考える．

	BLEU
ベースライン	52.38%
依存言語モデル	53.66%
依存翻訳モデル	54.30%
依存言語モデル＋依存翻訳モデル	55.09%

表 -1　統計翻訳のドメイン適応による性能

原言語側，目的言語側

$X \rightarrow \langle\ X_{[1]}$ **は** $X_{[2]}$, **The** $X_{[1]}$ $X_{[2]}\ \rangle$

$X \rightarrow \langle$ **国際** $X_{[1]}$, **international** $X_{[1]}\rangle$

$X \rightarrow \langle$ **テロ**, **terrorism**$\rangle$

$X \rightarrow \langle\ X_{[2]}$ **も** $X_{[1]}$, **also** $X_{[1]}$ $X_{[2]}\ \rangle$

図 -3　獲得される同期文脈自由文法

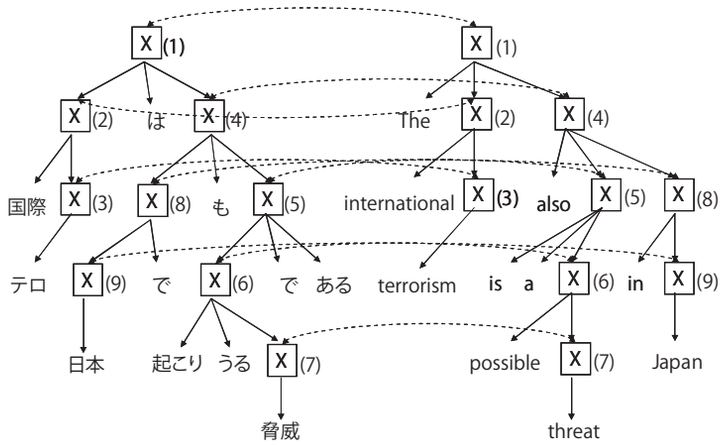

図 -4　翻訳例

利用して，その部分コーパスとしてドメインを定義する．部分コーパスは，エントロピーを基準としたクラスタリング手法によって自動構築する．各クラスタごとに，ドメイン依存の翻訳モデルと言語モデルを構築しておく．(2) オンラインプロセス：翻訳のソース文が与えられると，その文に対して最も高い確率を与えるクラスタを選択することで，ドメイン推定を行う．推定されたドメインに依存した言語モデルと翻訳モデルを使ってデコードする．

提案法を IWSLT2006 の日英オープントラックの旅行対話データ（参照訳の数は 16）で評価した（**表 -1**）．ドメイン依存モデルはドメイン非依存モデルと線形補間してデコードした．ベースラインのドメイン非依存モデルから，依存言語モデル，依存翻訳モデルの利用で，それぞれ独立に BLEU 値の向上が認められ，さらに，併用した場合には，ベースラインの 52.4% から 55.% へ 2.7 ポイントの向上が得られた．さらに，異なる言語対，音声認識結果，NIST MTE のニュースデータのいずれを用いた実験でも性能改善が確認できた．また，従来法である，クラスタ言語モデルや文混合モデルとの比較においても，より高い性能が確認でき，提案法が有効であることを検証できている．

◆ NTT システム

NTT ではこれまで重みつき有限状態トランスデューサに基づくデコーディング手法，大局的な句の並び替えのモデル化（長岡技術科学大学との共同研究），述語項構造に基づく語句の並び替え手法（奈良先端大との共同研究）などについて研究を進めてきた．本稿では，NTT において最近一番高い翻訳精度を達成している翻訳手法について解説する．

階層的な句に基づく翻訳手法

翻訳モデルは，(上)で解説した階層的な句に基づくアプローチを採用する．対訳コーパスから統計量でスコア付けされた同期文脈自由文法を自動的に学習するというものである．このアプローチの問題は，翻訳処理の過程で n-gram 言語モデルと階層的な句に基づく翻訳モデルの統合が容易でないことにある．翻訳処理はビーム探索等により仮説を枝刈りしながら進めるが，正確な枝刈りのためには，各仮説に対して翻訳モデルと言語モデルの両スコアを正しく反映することが必要である．言語モデルは文頭から文末方向に文が逐次的に生成されるときに適用しやすいモデルであるが，階層的な句による翻訳モデルをそのような順序で文が生成されるように制御するアルゴリズムは自明ではない．

そこで獲得される同期文脈自由文法の形式に制限を加え，文頭から文末方向に文が逐次的に生成されることを保証する翻訳手法を考案した [4]．本手法では，**図 -3** のような同期文脈自由文法が獲得され☆15，それを用いて**図 -4** のように日本語が英語へと翻訳される．本手法では，獲得される文法の目的言語側(この例では英語側)が必ず終端記号（すなわち単語）で始まるように Greibach 標準型と同じ制限を加える．翻訳処理では，この制限を加えた文法をトップダウンに図 -4 の X の添字の順序に展開する．これは，原言語側（この例では日本語側）を Earley のアルゴリズムにより構文解析する際に，目的言語が文頭から文末に生成されるように制御することで実現できる．本アルゴリズムにより，文法を 1 段展開するごとに目的言語を文頭から文末にかけて逐次的に生成することが保証されるため，n-gram 言語モデルとの融

☆15 ここで示したものは獲得される文法のごく一部である．

合が容易となる．その結果，適切な枝刈りが可能となり，効率的かつ高精度な翻訳処理を実現できる．

本手法は日英など語順の大きく異なる言語対での効果を期待して考案したものであるが，IWSLT 2006 における評価では日本語–英語はもとよりアラビア語–英語，イタリア語–英語，中国–英語のすべての言語対で，従来の句に基づく手法よりも高い翻訳精度を達成した．

膨大な素性の活用

（上）で解説したように，近年式 (2) のような対数線形モデルを使ったモデル化が一般的になっている．

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} P(\boldsymbol{e}|\boldsymbol{f}) = \arg\max_{\mathbf{e}} \sum_{m=1}^{M} \lambda_m h_m(\boldsymbol{e}, \boldsymbol{f}) \quad (2)$$

ここで，**f** は原言語（翻訳元言語），**e** は目的言語（翻訳先言語），h_m (**e**, **f**) は素性関数（または単に素性），λ_m は素性に対する重みを表す．重みは BLEU 等の目的関数を最大化するように推定される．計算量や学習データ量の兼ね合いもあり，これまで，素性としては翻訳モデルや言語モデルなどのサブモデルを使い，サブモデルの重み付けを学習する程度の使われ方にとどめられてきた．IWSLT 2006 における検討で，原言語と目的言語の対応する単語ペア☆16（さらにそのバイグラム）といった膨大な素性を用いたモデルで，翻訳結果を再順位づけたところ，劇的に性能を向上できることが確認できた．そこでこの考えを押し進めて，このような膨大な二値素性を階層的な句に基づく翻訳の中に取り入れた手法を考案した [3]．NTT の手法では，従来用いられてきた素性に加えて，以下の膨大な二値素性（数百万から一千万素性）を活用する．

(1) 原言語と目的言語の単語ペア
(2) 挿入単語と各原言語側単語のペア
(3) 目的言語の bigram
(4) 原言語側の木において，上位の階層の各終端記号と下位の階層の各終端記号のペア

学習手法には，構文解析等でも用いられているマージン最大化学習法 MIRA（Margin Infused Relaxed Algorithm）を採用した．NIST MTE や IWSLT の共通タスクを用いた実験で，テストセットと条件の近い開発セット（対訳コーパス）を用いることで，従来の素性だけを用いた手法と比べて大きく翻訳精度を改善できることを確認した．

◆京大システム

言語は本来的に構造を持つ．文の構造を求める構文解析の研究は近年飛躍的に進展しており，日本語，英語などでは高精度な構文解析器が利用できる．また，近い将来，多くの言語でそのような状況が生まれると考えられる．そこで，中長期的にみれば機械翻訳においても構文情報を十分に利用することが妥当であると考え，京大では構造的言語処理に基づく用例ベース翻訳の研究をすすめている [1]．ここで利用するものは，対訳コーパス，対訳辞書，両言語の構文解析器である．

用例翻訳ではできるだけ大きな翻訳例を用いることで文脈を安定させ，翻訳を適切にする．これによって，言語構造が大きく異なる言語対であっても，その複雑な翻訳関係を 1 つの用例として直接的に扱うことが可能となり，高精度な翻訳につながる．このとき，大きな翻訳用例を利用しようとすれば，語列としては不連続であっても構造的につながっている用例を扱う必要があり，構文情報の利用が必須となる．また，（用例翻訳であれ統計翻訳であれ）対訳文内の語句の対応を正確に行うことがきわめて重要であるが，ここでも，言語構造が大きく異なる言語対においては構文情報の利用が有効である．なお，統計翻訳における構文情報の利用は，まず構文情報を用いずに統計的に語アライメントを行った後で構文を利用しはじめるという方式が主流であり，京大システムはアライメント段階でも構文情報を積極的に用いる点に特徴がある．

以下では，京大システムの構成を簡単に紹介する．翻訳システムは大きく分けて 2 つの部分からなる．与えられた対訳コーパスから翻訳知識を学習する**アラインメント部**と，学習された知識を用いて新たな文を翻訳する**翻訳部**である．

アラインメント部において重要な点は 2 つある．1 つは対訳文中の対応候補を十分に見つけ出すことであり，2 つ目は，見つかった対応候補の中から適切な対応を選択することである．対応候補の検出は，対訳辞書，字訳関係の編集距離による解析，標準化した数字のマッチング，対訳コーパス全体から学習される文字列共起度などの情報を利用して行う．このようにして見つかった対応候補の中には，曖昧性を持つ対応や誤った対応などが含まれるため，対訳文全体の整合的対応という尺度を定義し，これに基づいて候補の選択を行う．この際，文の構造の利用が有効であり，日英新聞記事コーパスを用いた実験では，構造を用いない場合のアラインメント精度が 70.3% であるのに対し，構造を用いた場合は 76.5% という結果が得られている [2]．このようにして得られた対応と，その依存構造木上のすべての連続的組合せを用例データベースに登録する．

翻訳部では，入力文を依存構造木に変換し，木構造上の各部分木について翻訳用例をデータベースから検索す

☆16 ある特定の単語ペアが存在すれば 1 を返し，そうでなければ 0 を返すような素性関数．以降，同様の表現を用いる．

る．このようにして得られた用例集合から，できるだけ大きな用例を優先しながら，入力文全体をカバーする用例集合を選択し，その対訳部分を結合することにより，翻訳を生成する．この枠組みによる実際の翻訳例を**図 -5** に示す．

IWSLT2005 および 2006 に参加した京大システムの成績は中程度であったが，その原因は，既存の構文解析器が新聞などを対象に開発されており，会話文を正確に解析できないことであった．現在は，新聞，科学技術文などを対象として研究を進めており，NTCIR での特許翻訳コンテストにも参加予定である．

また，2006 年から，このような用例ベースの枠組みで，京都大学，情報通信研究機構（NICT），東京大学，静岡大学，科学技術振興機構が共同して日中機械翻訳のプロジェクトを推進している[7]．このプロジェクトでは，科学技術文書を対象として，新たに大規模な日中対訳コーパスを構築するとともに，専門用語対訳辞書を自動構築し，それを用例ベース翻訳に組み込むことにより，実用に近い日中翻訳システムの構築を目標としている．また，北京オリンピックに向けて，北京観光の多言語情報サービスでの利用も検討されている．

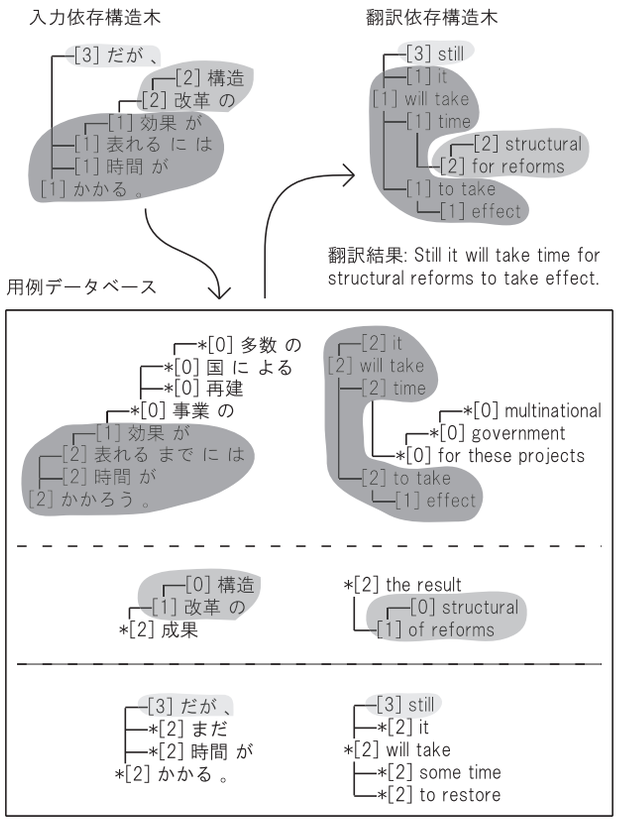

図 -5　用例ベース翻訳による翻訳例

おわりに

評価型ワークショップにも負の側面はある．予算の関係から人間による主観評価を十分に行うことができず，批判の多い BLEU による評価がメインになりがちである．そのため，BLEU で不利になりやすい翻訳手法（言語学的な規則に基づく手法など）の研究が不当に低く評価されてしまう危険性がある．しかし，この問題の根源は評価型ワークショップにあるのではなく，現在の不完全な自動評価尺度にある．自動評価尺度を改善していくためにも，評価型ワークショップは人間による主観評価をちゃんと行い，システムの翻訳結果と人間による主観評価の対応データを蓄積する責任があるように感じる．

その他よく聞く批判として，あまりに短期的に成績向上を目指すため，長期的にブレークスルーを起こすであろう技術の研究開発が疎かになることが挙げられる．これに対しては，同様の研究スタイルをとる音声認識の研究経緯から楽観的に考えている．連続音声認識は DARPA 主催のコンテストにより，90 年代に急速に技術が進展し，あっと言う間に夢の技術から実用的な技術へと変貌した．しかし，コンテストの弊害で技術が停滞したという話は聞かない．統計翻訳技術は，約 10 年遅れて連続音声認識技術が歩んだ道と同じ道を辿っているように思える．2000 年以降，評価型ワークショップの後押しもあり，統計翻訳の性能は急激に向上した．最近では Web ページの翻訳サービス（Google Translate BETA）や携帯電話の音声翻訳サービス（ATR-Trek「しゃべって翻訳」）といった商用サービスも始まりつつある．今後も技術は着実に進歩していくだろう．

NIST MTE など英語を中心とするワークショップに参加して痛感するのは日本語を中心とする対訳コーパスの少なさである．NIST MTE では数百万文は当り前で中英コーパスなどは一千万文に届く勢いである．それに対し，日本語に関しては比較的集めやすい日英対訳コーパスであっても，100 万文集めるのは容易でない．その結果，皮肉なことにさっぱり内容が分からずに作っているアラビア語から英語への翻訳システムの方が，日本語から英語への翻訳システムより遥かに性能が高かったりする．少ないデータ量で性能を上げるのが研究だという考え方もあろうが，情報が日々増大している時代，ありあまる膨大なデータを使いきる技術の研究開発は急務であろう．日本で統計翻訳の研究を活性化させるためには，共通タスクとなる日本語の膨大な対訳コーパスをどうやって整備していくかが鍵となる．そのためには，日本にも TIDES や GALE に匹敵する予算規模の国家プロジェクトが必要なのかもしれない．米国に比べればいろいろ不利な研究状況の中にありながら，それでも NTCIR

特許翻訳では100万文を超える学習データの作成に成功した．今後，日本の統計翻訳研究の起爆剤になることを多いに期待している．

謝辞　本稿の執筆にあたっては，科学研究費補助金(特定領域研究，情報爆発IT基盤)の助成を受けた．

参考文献

1) Nakazawa, T., Yu, K. and Kawahara, D. and Kurohashi, S. : Example-based Machine Translation based on Deeper NLP, Proceedings of International Workshop on Spoken Language Translation (IWSLT'06), pp.64-70 (2006).
2) Nakazawa, T., Yu, K. and Kurohashi, S. : Structural Phrase Alignment Based on Consistency Criteria, Proceedings of Machine Translation Summit XI, pp.337-344 (2007).
3) Watanabe, T., Suzuki, J., Tsukada, H. and Isozaki, H. : Online Large-Margin Training for Statistical Machine Translation, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), pp.764-773 (2007).
4) Watanabe, T., Tsukada, H. and Isozaki, H. : Left-to-Right Target Generation for Hierarchical Phrase-Based Translation, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006), Sydney, Australia, Association for Computational Linguistics, pp.777-784 (2006).
5) Yamamoto, H. and Sumita, E. : Bilingual Cluster Based Models for Statistical Machine Translation, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp.514-523 (2007).
6) Yasuda, K., Yamamoto, H. and Sumita, E. : Method of Selecting Training Sets to Build Compact and Efficient Statistical Language Model, Proceedings of the MT Summit XI Workshop Using Corpora for Natural Language Generation: Language Generation and Machine Translation (UCNLG+MT), pp.31-37 (2007).
7) 井佐原均，黒橋禎夫，辻井潤一，内元清貴，中川裕志，梶　博行，中村　徹：科学技術文献を対象とする日中機械翻訳システム開発プロジェクト，言語処理学会第13回年次大会，pp.83-86 (2007).
8) 安田圭志，隅田英一郎：機械翻訳の研究・開発における翻訳自動評価技術とその応用，人工知能学会誌 小特集「テキストの自動評価」，Vol.23, No.1 (2008).

(平成19年12月15日受付)

塚田　元（正会員）
tsukada@cslab.kecl.ntt.co.jp

1989年東京工業大学大学院理工学研究科修士課程修了．現在，NTTコミュニケーション科学基礎研究所主任研究員．統計的機械翻訳および音声言語処理の研究に従事．

永田　昌明（正会員）
nagata.masaaki@lab.ntt.co.jp

1987年京都大学大学院工学研究科修士課程修了．現在，コミュニケーション科学基礎研究所主幹研究員．工学博士．統計的自然言語処理の研究に従事．

隅田英一郎（正会員）
eiichiro.sumita@atr.jp

1982年電気通信大学大学院電気通信学研究科修士課程修了．博士(工学)．現在，ATR室長．NICT研究マネージャ，神戸大学大学院連携教授，ATR-Langue副社長兼務．機械翻訳，eラーニングの研究に従事．

黒橋　禎夫（正会員）
kuro@i.kyoto-u.ac.jp

1994年京都大学大学院工学研究科電気工学第二専攻博士課程修了．博士（工学）．2006年より京都大学大学院情報学研究科教授．自然言語処理，知識情報処理の研究に従事．