統計的機械翻訳

つかだ はじめ わたなべ たろう
塚田 元 /渡辺 太郎
すずき じゅん ながた まさあき
鈴木 潤 /永田 昌明
いそざき ひでき
磯崎 秀樹

NTTコミュニケーション科学基礎研究所

新しい機械翻訳技術，統計的機械翻訳（統計翻訳）を紹介します．本技術は，テキストデータから統計モデルを学習し，自動的に機械翻訳システムを構築するものです．アルゴリズムが言語に依存しないため，学習データさえあれば多言語化が容易であるとともに，短期間に低コストで頑健なシステムの構築が可能です．

統計翻訳の背景

一般に，機械翻訳は入力言語と出力言語の両方が分かる言語の専門家が翻訳ルールを記述することによって実現されます．このルールベース翻訳は，言語の専門家の確保が必要なため，マイナー言語への対応が難しく，多言語化が困難であるとともに，開発コストも高いという問題がありました．近年Webに代表されるように，大量の多言語データが利用可能になりつつあります．統計翻訳はこの大量の多言語テキストデータを利用して，言語の専門家なしに自動的に翻訳システムを構築可能にする技術です．

ルールベース翻訳との比較

2001年に米国DARPAのTIDESプロジェクトの一環として，アラビア語－英語，中国語－英語の機械翻訳コンテストが始まり，これを機に統計翻訳の技術は急速に進展しました．このコンテストでは，数百万文規模の対訳データ（後述）が学習用に提供されます．これくらい大規模な学習データがあれば，統計翻訳はルールベース翻訳を上回る性能を発揮するところまできています．しかし，統計翻訳といえども万能ではありません．学習データが少ない場合や，量はあっても実際に翻訳するテキストと大きく食い違う場合は，十分な性能が得られません．実サービスでは，学習データの条件が満たされず，従来のルールベース翻訳が有利になることもしばしばあります．学習データさえあれば，開発コストが低いことは統計翻訳の大きな長所ですが，翻訳時の計算コストが高いためルールベース翻訳が必要とする計算機より高価で高性能なものが必要になる欠点もあります．統計翻訳はルールベース翻訳にない，長所を数多く備えた次世代技術ですが，サービスで利用する際は各々の特性を見極めて使い分けることが必要になります．

統計翻訳の処理の流れ

■概要

統計翻訳技術の概要を**図1**に示します．学習データとして，対訳データと出力言語の単言語データを大量に用意します．対訳データは，入出力の各々の言語で同じ内容を表す文の組を集めたものです．各々のデータから，それぞれ翻訳モデルと言語モデルを学習します．翻訳モデルは従来技術の翻訳ルールや対訳辞書に相当するもので，翻訳としての確からしさを評価します．言語モデルは従来技術の文法に相当するもので，生成される単語列が出力言語としてどれだけ確からしいかを評価します．通常，n-gramと呼ばれるn個

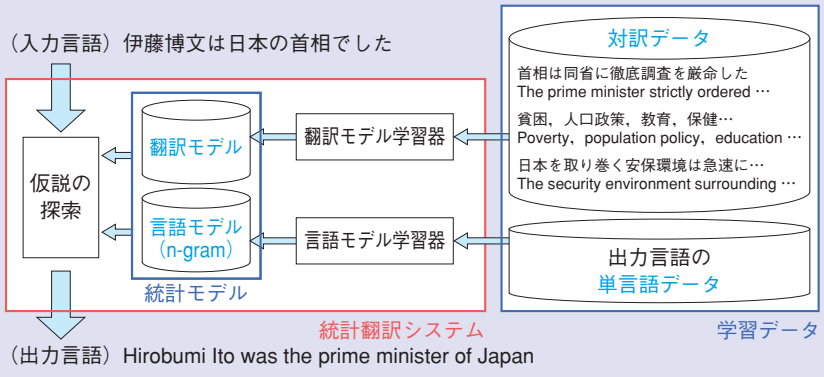


図1　統計翻訳技術の概要

の単語並びの統計量が用いられます．翻訳処理は，この2つの統計モデルを用いて，膨大な仮説の中から確からしい単語列を探索する処理として実現されます．

■翻訳モデルの学習

統計翻訳の中で，現在主流になっている句（部分単語列）に基づく翻訳モデル[(1)]について説明します．**図2**は，英日翻訳における学習の流れを示した例です．自動的に求めた単語対応を基に句の対応を求めます．そして，対応付けられた句の組を統計量でスコア付けします．句に基づく翻訳モデルは，文単位の対訳をその構成単位である句の対訳に分解しスコア付けしたものであると考えることができます．

■仮説の探索（翻訳処理）

英日翻訳の例を**図3**に示します．入力が与えられると，それをあらゆる句に分割します．入力文の中から，句を1つ選択し（必ずしも左から右に選択する必要はありません），その対訳を1つ用いて左から右に出力文を逐次的に生成します．こうして生成される膨大な仮説の中から，翻訳モデルのスコアおよび言語モデルのスコアを考慮して，もっとも確からしい解を探索します．

研究開発の取り組み状況

統計翻訳の翻訳処理は膨大な仮説を扱うことから計算コストが高く，その高速化は重要な課題です．また，日英など語順の大きく変わる言語対を扱うためには，語の並び替えのモデル化が精度向上に重要な役目を果たします．前者に関して，音声認識分野で培われた技術を拡張した手法を開発しました[(2)]．さらに，後者に関して，句の並び替えのパタンに応じたモデル化

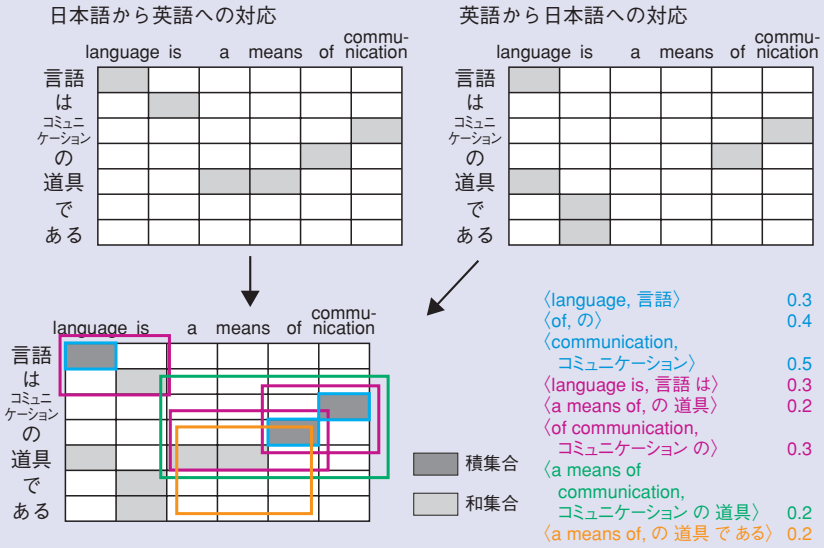


図2　句に基づく翻訳モデル

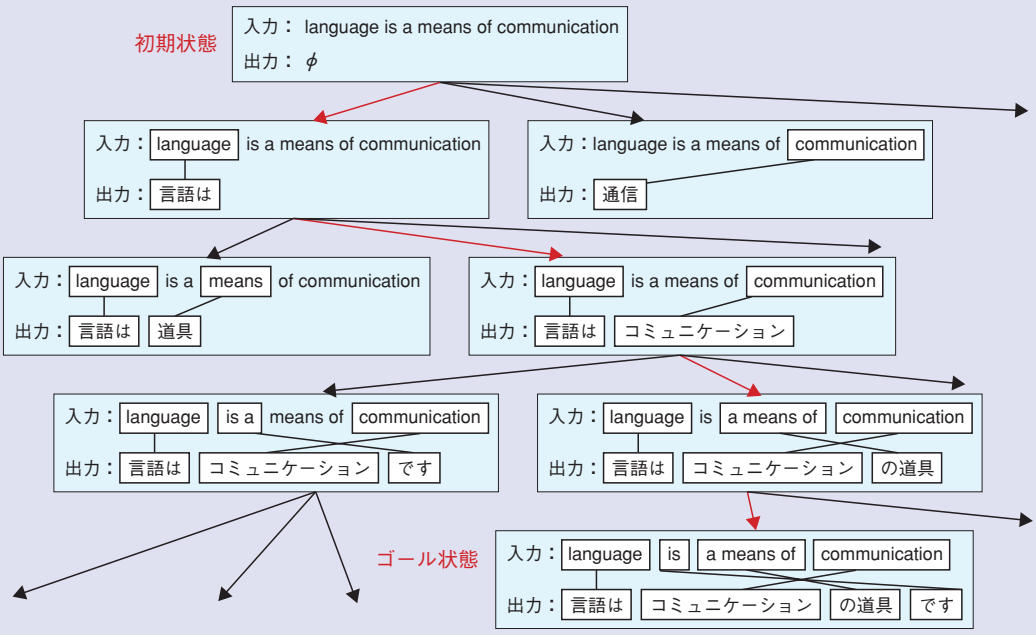


図3　仮説の探索（翻訳処理）

手法[3]や，「階層的な句」と呼ばれる文法規則を対訳データから自動獲得する翻訳手法[4]を開発しました．ここでは最新の研究成果である階層的な句に基づく翻訳手法を紹介したいと思います．

図2の句に基づく翻訳モデルの例では，〈communication, コミュニケーション〉と〈of communication, コミュニケーション の〉の2つの句の組が獲得されています．前者は後者に含まれていることから，前者をXで表すと後者は〈of X, X の〉というパタンで表現できます．この処理を続けると，日英対訳コーパスから例えば**図4**のような階層的な句に基づく翻訳モデルを学習することができます．Xの添え字は，入力言語側と出力言語側の句の対応を表しており，これによって句の並び替えを適切にモデル化することができます．階層的な句を用いた日英翻訳の例を**図5**に示します．例えば，X(4)の下のX(5)とX(8)の順序が入れ替わることで，日本語と英語の語順が適切にモデル化されていることが見て取れます．我々の手法は，類似手法[5]と比べて言語モデルとの融合が容易で，より効率的な翻訳処理が実現できる特徴があります．

入力言語	出力言語	
X →	〈X_1 は X_2, The X_1 X_2〉	0.2
X →	〈国際 X_1, international X_1〉	0.4
X →	〈テロ, terrorism〉	0.5
X →	〈X_1 も X_2, also X_2 X_1〉	0.6

図4　階層的な句に基づく翻訳モデル

技術の適用分野

入力を単語分割する処理を除けば，統計翻訳のアルゴリズムは言語に依存しません．そのため，対訳データさえあれば多言語翻訳を容易に実現できます．対訳データの確保の仕方を工夫する必要がありますが，多言語のWebサービスとは非常に相性のよい技術です．魅力的なコンテンツさえあれば，Web 2.0的アプローチでユーザに翻訳してもらうことで対訳データを確保できるかもしれません．

統計翻訳は多言語だけでなく特殊分野の翻訳も得意です．新聞，マニュアル，特許等の公文書はすでに大量に翻訳されています．コンテンツホルダと提携し，これらを対訳データとして用いることで分野に特化した翻訳サービスも実現できます．

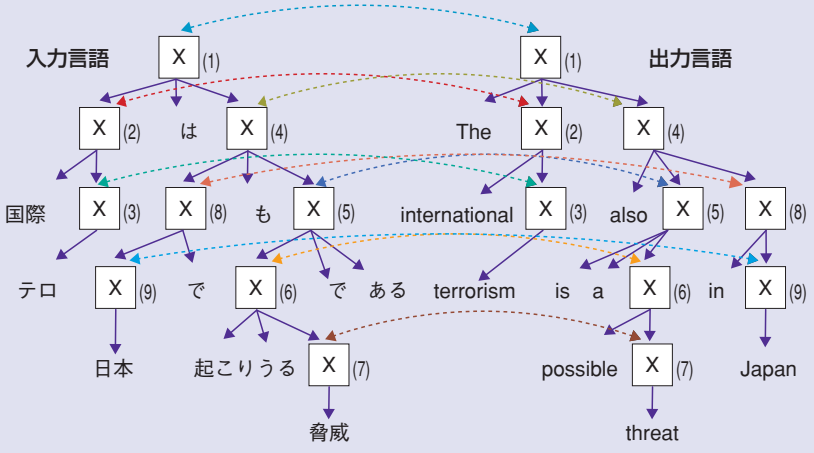


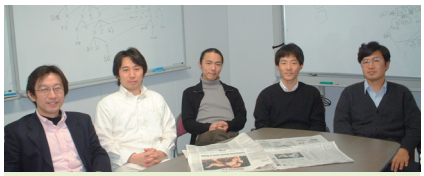
図5　階層的な句に基づく翻訳例

今後の取り組み

統計翻訳は駆け出しの技術です．翻訳精度の向上，翻訳処理の高速化，学習用の対訳データの確保の方法など，まだまだ技術的な課題は山積みです．これらの問題の基礎研究に引き続き取り組んでいきたいと考えています．一方，現状の技術でも適用可能なサービスはあるかもしれません．基礎研究と合わせてサービスへの適用の検討も進めていきます．

■参考文献

(1) P.Koehn, F.J.Och, and D.Marcu: "Statistical Phrase-Based Translation," Proc. of HLT-NAACL 2003, pp. 127-133, 2003.
(2) H.Tsukada and M.Nagata: "Efficient Decoding for Statistical Machine Translation with a Fully Expanded WFST Model," Proc. of EMNLP 2004, pp. 427-433, 2004.
(3) M.Nagata, K.Saito, K.Yamamoto, and K.Ohashi: "A Clustered Global Phrase Reordering Model for Statistical Machine Translation," Proc. of COLING-ACL 2006, pp. 713-720, 2006.
(4) T.Watanabe, H.Tsukada, and H. Isozaki: "Left-to-right Target Generation for Hierarchical Phrase-based Translation," Proc. of COLING-ACL 2006, pp. 777-784, 2006.
(5) D.Chiang: "A Hierarchical Phrase-Based Model for Statistical Machine Translation," Proc. of ACL 2005, pp. 263-270, 2005.



（左から）塚田　元/ 鈴木　潤/
渡辺　太郎/ 永田　昌明/
磯崎　秀樹

革新的な自然言語処理技術の研究開発に取り組んでいきます．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
TEL　0774-93-5372
FAX　0774-93-5385
E-mail　tsukada@cslab.kecl.ntt.co.jp