統計翻訳に構造制約を導入する新しいアプローチ

国際電気通信基礎技術研究所（ATR）
音声言語コミュニケーション研究所
自然言語処理研究室長
隅田　英一郎

PROFILE 規則・用例・統計翻訳、音声翻訳、eラーニングの研究に従事。（独）情報通信研究機構（NICT）グループリーダ、神戸大学大学院工学研究科連携教授、（株）ATR-Langue取締役副社長兼務。博士（工学）。

eiichiro.sumita@atr.jp　0774-95-1301

1 はじめに

1988年に、原文と訳文を大量に集めた対訳データと統計的な学習アルゴリズムだけで翻訳システムを構築する統計翻訳（Statistical Machine Translation, SMT）と呼ばれる手法[1]が提案されたが、余り発展することもなかった。2000年前後に、改良型のSMTであるフレーズベース統計翻訳（Phrase-Based Statistical Machine Translation, PBSMT）[2]が提案され、対訳データの蓄積の進展、計算機パワーやメモリ容量の増大など研究環境の変化にも後押しされ、SMTが研究コミュニティで急速に広がった。実際、新聞、旅行会話をはじめとする多様な分野に盛んに適用され良い成果が報告されている。また、文長が長く、新規の専門用語の頻出するという理由で特許は機械翻訳が困難と考えられてきたが、昨年より、JAPIOとNICTの協力で特許の対訳データが構築され、これにSMTを適用する実験が開始された。

集中的な研究の結果、非局所的な語順制御が最も重要な課題と認識され、有望な解決策も分かりつつあり、様々な新展開が起こらんとしているのが、SMTを中心に日々前進している翻訳研究の現在である。

本稿では統計翻訳の基本のアイデアとその改良型であるフレーズベース統計翻訳の簡単な紹介を行い、課題となっている非局所的な語順の制御に関して、構文制約を用いる方法[3]ITGとその改良法[4]IST-ITGについて、実験結果と翻訳例とともに紹介する。

2 統計翻訳

2.1 統計翻訳の基本

ここで統計翻訳を簡単に説明する。統計のココロは、「世の事象は不確実なので、その起こる可能性を見積もって可能性の高いものを選ぶ」ということである。翻訳を同じ原理で扱ったのが統計翻訳である。

簡単な例[†1]で説明しよう。フランス語「il croit」を英語に翻訳してみる。仏和辞書を引けば、「il」には「he」と「it」の二つの訳語があり、「croit」には「thinks」と「grows」の二つの訳語があることが分かる。従って、「il croit」は「he thinks」「he grows」「it thinks」「it grows」の4通りの訳文がありうる。訳質は順に、◎、×、×、○である。何故そんな判断が出来るかというと、「he thinks」が世の中の英文の中で他に比べて高頻度で出現するからである。例えば、「John F.」に続く単語として、「Keneddy」「pencil」の2通りあった場合どちらがより尤もらしいか？統計の記法で書くと確率P（Kennedy｜John F.）と確率P（pencil｜John F.）とでどちらか大きいか？計算してみると前者が0.49で後者は0.0000007である。圧倒的に「Keneddy」であり、直感に合う。先の例「il croit」の翻訳では、語順をフランス語のままにしたが、可能な全ての語順も考えて確率を計算しても、結局、一番良い翻訳は「he thinks」になる。

†1 Kishore Papineni（Yahoo）による。

対訳辞書と語順変更など、原文から外国語の単語の集合へ変換をつかさどるのが翻訳モデルと呼ばれ、対訳データから統計的に学習される。尤もらしい訳文を選ぶところが言語モデルと呼ばれ、翻訳の目的言語の大量のデータから学習される。

しかし、この基本的な方法は、語順の処理に計算時間がかかりすぎたり、日本語と英語のように、文法や語彙が著しく違う言語対の間では高品質の翻訳が出来なくて普及しなかった。世紀の変わり目前後に次に述べるフレーズを使った統計翻訳が提唱されて上記の課題の一部が克服されて、世界の研究者が統計翻訳を再評価し本格的に取り組み始めた。

2.2 「フレーズ」を使った統計翻訳

現在標準的な手法と考えられているフレーズベース統計翻訳（PBSMT）[2]について、ごく簡単に説明する。翻訳のプロセスは以下の通りになる。

①入力文をフレーズ[†2]に分割する。

②各フレーズを確率的に翻訳する。

③フレーズの順序を確率的に調整する。

図1にサンプルを示したように、四角で囲まれたフレーズ毎に翻訳され（例えば、I will goが行きますに）、翻訳されたフレーズの語順が調整される。フレーズの内部は固定されており語順は変わらない。

PBSMTにおいて、ステップ①②によって局所的語順の問題は大幅に改善されたが、ステップ③で取り扱われる非局所的な語順並び替えは、距離に依存した制約などが基本となっており、充分な制御ができず課題として残っている。

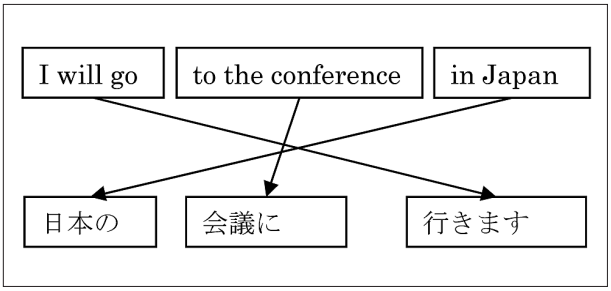

図1　フレーズベース統計翻訳の動作例

3 フレーズベース統計翻訳への構文の導入

前節で指摘した非局所的語順並び替えの問題に対しては様々な手法が提案されているが、大きく二つのタイプに分けることができる。

- 一つ目は構文情報を直接翻訳モデルの中に組み込むもので、語順並び替えに対する詳細な情報をモデル化できると考えられるものの、モデル構築には大量の訓練データを必要とする問題点がある。
- 二つ目は語順制約を翻訳モデルとは別に与えるもので、Inversion transduction grammar（ITG）制約[3]がこのタイプに属する。この手法の制約力は一つ目のタイプに比べると弱いと考えられるが、大量の訓練データを必要としないという利点がある。

本稿ではITG制約の拡張として翻訳元文の構文木情報を直接語順並べ替えの制約に組み込むImposing source tree on ITG（IST-ITG）制約[4]を紹介する。ITG制約では翻訳元文の木構造として特定のものを仮定していないのに対し、IST-ITGでは、翻訳元文を構文解析して得られる木構造を用いるため、より強い制約を与えることができる。

たとえば、4単語からなる翻訳元文 $\{a\ b\ c\ d\}$ に対しては可能なすべての並び替えは24（=4!）通りであるが、ITG制約は $\{C\ A\ D\ B\}$ と $\{B\ D\ A\ C\}$[†3]の二つの並びしか排除できないため、22（=4!-2）通りの語順を考えなければならない。これに対し、IST-ITG制約では、8（$=2^3$）通りの語順だけ考慮すれば済む。

3.1 ITG制約

並び替えに対して制約がなければ、一対一対応の単語モデルを仮定すると、N単語からなる翻訳元文に対し、$N!$通りの語順を翻訳先言語で考えなければならない。ITG制約では次のような制約をかけることで、考慮すべ

†2 文法的な意味の句でなく、翻訳する上で固定的に取り扱える単語列。このフレーズは単語の翻訳モデルに基づいて学習する。いくつかの方法があるが本稿ではPhilipp Koehnの方法に従った。

†3 以降、各大文字は小文字単語の対訳を表すものとする。

き語順を減らすことができる。

●*翻訳元文に対し、可能なすべてのバイナリ木を考える。*

●*翻訳先文の語順はバイナリ木の任意のノードを回転させることで得られる。*

N=4の場合、ITG制約は並び替えの組み合わせを4!=24通りから22通りに減らすことができる。この差はNが大きくなる程大きくなり、*N*=10では削減率は206,098/3,628,800=0.0568となり、大きな効果がある。ただし、ITG制約では、翻訳元文に対しバイナリ木を特定していない。従って、翻訳元文を構文解析することによってバイナリ木を特定できればより強い制約を与えることができるので、次節の改良手法が提案された。

3.2 Imposing Source Tree

改良手法（IST-ITG）では構文木から得られるbracketed sentenceを用いることにする。たとえば英日翻訳における翻訳元文「This is a pen」に対しては((This)((is)((a)(pen))))という構造を利用する。この構造はバイナリ木と等価である。IST-ITGを用いた場合、*N*=4の場合、ITGの22通りの語順に対し8通りの語順を翻訳先言語で考えればよいことになる。

たとえば、翻訳元文木((*a b*)(*c d*))であることが分かっていれば可能な翻訳先木は {*A B C D*}, {*B A C D*}, {*A B D C*}, {*B A D C*}, {*C D A B*}, {*C D B A*}, {*D C A B*}, {*D C B A*} の8通りとなる。同様に、翻訳元文木(((*a b*) *c*) *d*) に対しては, {*A B C D*}, {*B A C D*}, {*C A B D*}, {*C B A D*}, {*D A B C*}, {*D B A C*}, {*D C A B*}, {*D C B A*} の8通りとなる。一般的にはIST-ITGの下で考慮すべき語順は 2^{N-1} となる。

3.3 非バイナリ木への拡張

前節では、翻訳元文の構造としてバイナリ木を仮定した。しかしながら、構文解析の結果が常にバイナリ木になるとは限らない。このような場合、構文木のノードには3つ以上の枝を持つものが存在することになる。このような場合、同一ノードを持つ枝の間での語順の入れ替えは自由であるとするが、ITG制約は適用するものとする。たとえば、非バイナリ木(*a*(*b c d*))に対しては、(*b c d*) 内の入れ替えで6通り、*a*と(*b c d*)の入れ替えで2通り、全体で6*2=12通りの語順が許される。

一般的な語順の数は次式で表される。

$$\prod_{i=1}^{n} S_{B_i}$$

ここで、S_kは*N*=kのときのITG制約のもとでの可能な語順の数を表し、*N*=3では6通り、*N*=4では22通りである。また、B_iはi番目のノードの枝の数である。

3.4 IST-ITGのフレーズベースモデルへの拡張

前節では単語モデルにおけるIST-ITGについて述べた。本節ではこれをフレーズベースモデルに対して拡張する。構文解析木のノードは単語でなければならない。一方、フレーズベースモデルではフレーズを単位として翻訳が行われるため、制約の単位との不一致が生じる。この不一致を埋めるために、次の制限を導入する。

●フレーズが二つ以上に分割されるような語順入れ替えは許されない。

この制限を新たに導入することによって語順入れ替えは必ずフレーズ単位で行われるようになるため、単語モデルの場合と同じように語順入れ替えに制約をかけることができるようになる。たとえば、木((*a b c*)((*d e*)(*f g*)))に対し、*b*, *c*, *d* がフレーズ*ph*をなしている時(図2参照)、ノード1に対しては語順入れ替えを行うことができない。理由はこのノードはフレーズ*ph*の一部*b*, *c*を含んでいるため、入れ替えを行うと*ph*が二つに分割されてしまうためである。同様の理由でノード2と4に対しても入れ替えを行うことはできない。またノード3はフレーズの一部を含んでいないため、ノード5はフレーズ全体を含んでいるため語順入れ替えを行うことができる。結果として、この木から許される翻訳先文の語順は {*A PH E F G*}, {*A PH E G F*}, {*G F E PH A*}, {*F G E PH A*} の4通りである。

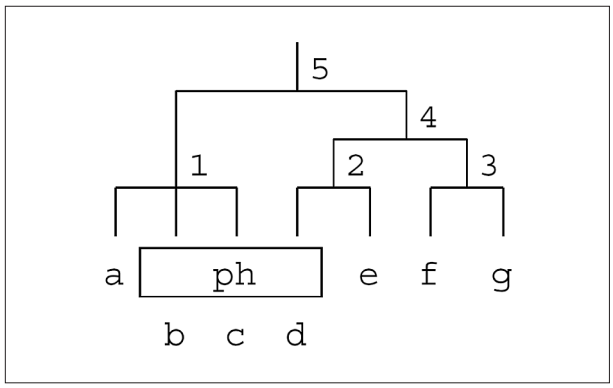

図2　フレーズのある構造木のサンプル

4 評価実験

4.1 評価コーパス

特許コーパスを用いた評価実験を行った。コーパスの詳細を表1に示す。このコーパスは対訳文の自動アライメント[5]を用いて作成されたもので、アライメントスコアの高いほうから900文を評価セットに、次の1,000文を開発セットに、残りを訓練セットとして用いた。このコーパスはNTCIR-7ワークショップ[6]の特許翻訳タスクにおける訓練コーパスのサブセットとなっている。

表1　日英特許コーパス

	文数	述べ語数（E/J）	異なり語数（E/J）
評価	900	29K/32K	3,967/3,683
開発	1,000	39K/37K	4,971/4,614
訓練	10M	273M/257M	797K/282K

4.2 翻訳品質の評価手法

現在、翻訳品質の評価手法[7]として、機械評価と呼ばれる手法が多数提案され利用されている。人間による評価は時間と費用が嵩むため、その代替手段として編み出されたものである。基本は、各テスト文に対して複数の参考訳を用意して、訳文と参考訳の一致・不一致の度合いを測る。代表的なのは、WER、PERである。

- WER　語順を考慮した単語の不一致率である。スコアが小さければ品質が良いことになる。
- PER　WERと違って、語順を無視し文を単語の集合として考えた場合の単語不一致率である。

WERやPERと、人手評価のスコアには相関があることが知られているので、これらの機械評価は広く使われている。

4.3 英日翻訳実験

まず、英日方向の翻訳実験を行った。フレーズベース翻訳モデルの学習にはGIZA++ツールキット[8]を、言語モデルの学習にはSRI language model ツールキット[9]を用いた。言語モデルは5gramであり、Kneser-Ney平滑化[10]を用いている。翻訳デコーディングのためのパラメータチューニングには1,000文の開発セットを用いてminimum error training[11]行った。英語の構文解析はCharniakパーザ[12]を、日本語のセグメンテーションには茶筅[13]を用いた。デコーダは我々が独自に開発したMoses[14]上位互換デコーダであるCleopATRaを用いた。CleopATRaにはIST-ITG制約下でデコーディングを行うアルゴリズムが実装されている。表2に評価結果を示す。

表2　英日特許翻訳における性能評価

	WER	PER
ITG	80.29	39.15
IST-ITG	74.90	38.93

4.4 日英翻訳実験

続いて、同じコーパスを用いて日英方向の翻訳実験を行った。日本語の構文解析には、係り受け解析器[15]を用いて文節間の係り受け関係をまず抽出し、これをbracketed sentenceに変換した。

表3に評価結果を示す。

表3　日英特許翻訳における性能評価

	WER	PER
ITG	80.43	40.50
IST- ITG	76.62	40.67

4.5　実験のまとめ

まず、PERに着目してITGとIST-ITGを比較すると、ほとんど差がなく、訳語選択は差異がないことがわかる。

一方WERでは、IST-ITGはITGよりも、英日方向で5.4%、日英方向では3.8%性能が向上している。このように、IST-ITGはWERにおいて大きな性能向上が見られ、非局所的な語順入れ替えのための制約としてうまく働いていることが確認できた。

5 おわりに

本稿ではフレーズベース統計翻訳（PBSMT）における語順入れ替えの制約として、翻訳元文の木構造情報を利用する手法（IST-ITG制約）を紹介した。IST-ITG制約は、翻訳元文の木構造情報を取り込むことによって、強い制約を与えることができて速度と精度が改善される、本稿では、特に、後者について実験結果を紹介した。

参考文献

1. Brown, P.F.；Cocke, J.；Della Pietra,S. A.；Della Pietra, V. J.；Jelinek, F.；Mercer, R. L.；and　Roossin, P. S.. "A statistical approach to language translation. "In Proc. of COLING-88.
2. P. Koehn, F. J. Och, D. Marcu, "Statistical phrase-base translation, "Proc. HLT-NAACL, pp. 127-133, 2003.
3. Dekai Wu, "Stochastic inversion transuduction grammars and bilingual parsing of parallel corpora,"Computational Linguiatics, 23（3）, pp.377-403, 1997.
4. Yamamoto, Hirofumi and Okuma, Hideo and Sumita, Eiichiro, "Imposing Constraints from the Source Tree on ITG Constraints for SMT," Proceedings of the ACL-08：˜HLT Second Workshop on Syntax and Structure in Statistical Translation（SSST-2）, 2008, p.1-9.
5. Masao Utiyama and Hitoshi Isahara, "Reliable Measures for Aligning Japanese-English News Articles and Sentences", ACL-2003, pp. 72-79, 2003.
6. http://ntcir.nii.ac.jp/
7. 隅田　英一郎，佐々木　裕，山本　誠一，"機械翻訳システム評価法の最前線"（2005）-情報処理，Vol.46，No.5，通巻483号，pp.552-557.
8. F. J. Och, H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," Computational Linguistics, No. 1, Vol. 29, pp. 19-51, 2003.
9. http://www.speech.sri.com/projects/srilm/
10. R. Kneser, H. Ney, "Improved backing-off for m-gram language model, "Proceedings of the IEEE International Conference of Acoustic, Speech, and Signal processing. Vol. 1, pp. 181-184, 1995.
11. F. J. Och, "Minimum error rate training for statistical machine trainslation, "Proc. ACL, 2003.
12. E. Charniak, "A Maximum-Entropy-Inspired Parser," Proc. NAACL-2000, pp.132-139, 2000.
13. http://chasen-legacy.sourceforge.jp/
14. http://sourceforge.net/projects/mosesdecoder/
15. http://chasen.org/˜taku/software/cabocha/

サンプル翻訳（英日）

原文：a sealant 7, which serves as a seal for cutting gas 9, also serves as a guide for the moving holder 3.

参照訳：7はシール材であり，後述の加工ガス9のシールと移動ホルダ3のガイドを兼ねたものである。
ITG：シールの役目も兼ねているため，シール材7をガイドするための加工ガス9，可動保持部3を備えている。
IST-ITG：封止剤7を加工ガス9シールの役目を兼ねて，可動ホルダ3をガイドする。

サンプル翻訳（日英）

原文：4は送油路を示し，本体1の中空部をもって構成してある.
参照訳：an oil passage 4 is formed as a hollow portion in the main body 1.
ITG：4 is a hollow portion of the body 1 with an oil supply passage is shown.
IST-ITG：4 is an oil supply passage, with a hollow portion of the main body 1.