# 特許と機械翻訳の新たなトレンド「統計翻訳」

株式会社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）研究マネージャ、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。2001年ごろから状況が変わり、統計翻訳が研究コミュニティで急速に広がった。新聞、旅行会話をはじめとする多様な分野で盛んに試みられ良い成果が報告されている。集中的な研究の結果、統計翻訳の解くべき課題と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、翻訳研究の現在である。

一方、特許は次の理由で機械翻訳が困難と考えられている分野である。①特許は厳密性が要求されるため、一文が非常に長い。文が長いと、文の構造解析失敗の可能性が高くなり、翻訳も解析に依存するので失敗する。②特許は新規性、進歩性が要求されるため、新しい概念や用語が次々に出てきて、辞書登録が追いつかず訳語が欠落してしまう。

難しい対象である特許に対して、先に述べた統計翻訳が有効であるか否かを検討することは意義がある。本稿では、Japio殿が作成しアジア太平洋機械翻訳協会のAAMT/JAPIO研究会に貸与した日英対訳コーパス（公開特許公報要約と英文要約PAJを1993年から2004年まで12年分の全件を対応させたもの）に対して、現時点で、標準的と想定される統計翻訳の手法を適用した結果を報告する。この対訳データは、統計翻訳を中心とする翻訳技術の世界規模の評価型ワークショップで利用されているデータ（中国語⇒英語が800万文）に匹敵し、統計翻訳を適用して実験する意味がある大規模データである。また、日英の対訳で大規模な公開データが存在しない現状から考えても意義がある。

以下では、統計翻訳の手法を簡単に説明し、特許文での実験結果について述べる。

## 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

※1 Kishore Papineni（Yahoo）による。

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.）とでどちらが大きいか？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。

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の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※2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 [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 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。

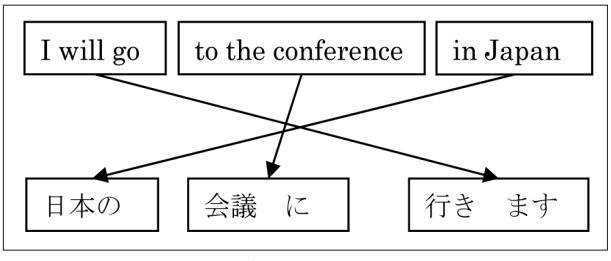


図1 フレーズベース統計翻訳の動作例

※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。

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。

### 2.3 統計翻訳の長所

統計翻訳が着目されている理由の一つに、次の性質がある。図 2に例示したように、統計翻訳に関して、様々な実験から経験的に、「学習に用いる対訳データ量を増やせば翻訳品質が改善する※3」ことが分かっている。（アルゴリズムの改善が仮になくても）データ量を増やせば翻訳品質があがるという極めて投資判断がしやすい性質である。

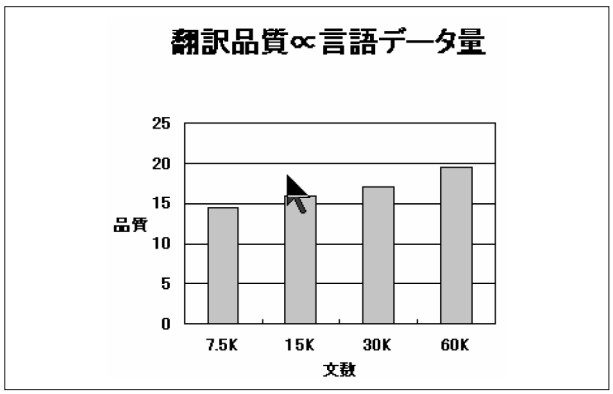


図2 統計翻訳の長所

## 3 特許データでの実験

### 3.1 実験条件

今回は、時間の制約から、公開特許公報要約／PAJ対訳データの一分野に限定して、英語から日本語への翻訳実験をした（表1）。

表1 諸元

| | 延べ文数 | 延べ語数 | 語彙 |
|---|---|---|---|
| 学習用データ | 770K | 24M | 110K |
| パラメタ調整用データ | 1K | 31K | 3.6K |
| テスト用データ | 0.5K | 20K | 2.8K |

※3 データ量が倍になると、翻訳品質が数ポイント（3.2.1節で述べるBLEU）改善する。

実験対象は、データ量が比較的多く、報告者にとって内容が分かりやかったのでG06の分野とした。

### 3.2 翻訳品質

#### 3.2.1 機械評価

現在、翻訳品質の評価手法[3]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、BLEU、WER、PERの3つである。

- BLEU 長さが1から4の単語列の一致度を測り、訳文が参考訳のどれかと完全に一致する場合に最大値の1になる。スコアが大きければ品質が良いことになる。
- 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。

BLUEや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。

#### 3.2.2 評価結果

各原文に対して参考訳1個という条件でBLEUスコア0.27を得た（表2）。テストセットや参照訳数が異なる場合、あるいは、翻訳方式が根本的に異なる場合に、BLEUスコアを相互に比較することは間違った判断をする可能性があって危険であるが、あえて、ここでは他の結果と比較してみる。特許ほどではないが比較的長文になる新聞やニュース放送を題材とした評価型ワークショップの公開されている最近の統計翻訳の結果のBLEUスコアと比べる。ここでは日本語から英語への翻訳のように難しいタスクである中国語から英語へのタスクに着目する。2007年3月のTC-STAR※4のVOAの中国語放送を対象としたタスクのBLEUスコア（参照訳2個）が最大0.25程度であり、2006年9月に行われたMT2006※5（米国NIST主催）のBLEUスコア（参照訳4個）が最大0.35程度であることから、今回の0.27は見込みのあるスコアといえるだろう。

また、特徴的なのは、WERがPERの2倍近いことである。本実験の特許翻訳では、訳語選択より語順の決定に問題があると解釈できる。要約をタイトルと本文に分けて、スコアを見ると、タイトルが圧倒的に性能がよく、入力の長さが短く、語順の影響が小さいためだと考えられる。

**表2 機械評価**

| テストデータ | | BLEU | WER | PER |
|---|---|---|---|---|
| 要約全体 | | 0.27 | 0.84 | 0.45 |
| 個別 | タイトル | 0.43 | 0.29 | 0.21 |
| | 本文 | 0.24 | 0.83 | 0.44 |

翻訳のサンプルを1文（図3）示す。上で述べたように、語順は明らかに問題だが訳語の選択は概ね良い。

●原文●the diagnostics server can start the automatic repair sequence of the electronic system 200 by a control signal .
●訳文●自動 診断 サーバ 、 修理 の シーケンス を 開始 する こと が できる 、 電子 システム 2 0 0 の 制御 信号 によって 。
●正解●診断 用 サーバ は 制御 信号 により 電子 システム 2 0 0 の 自動 修理 シーケンス を 起動 する こと が できる 。

**図3 サンプル翻訳**

さらに、この訳文の語順を変えていくとほぼ正解になる様子を図 4に示した。

※4 http://www.elda.org/tcstar-workshop/
※5 http://www.nist.gov/speech/tests/mt/

①自動 診断 サーバ 、 修理 の シーケンス を 開始 する こと が できる 、 電子 システム ２ ０ ０ の 制御 信号 によって
②自動 診断 サーバ 、 電子 システム ２ ０ ０ の 制御 信号 によって、修理 の シーケンス を 開始 する こと が できる
③自動 診断 サーバ 、 制御 信号 によって、電子 システム ２ ０ ０ の修理 の シーケンス を 開始 する こと が できる
④診断 サーバ 、 制御 信号 によって、電子 システム ２ ０ ０ の自動修理 の シーケンス を 開始 する こと が できる

**図4　語順を変えれば正解訳に近づく**

### 3.2.3　訳し分け精度

さらに、訳し分けに注目して分析してみる。英語の多義語の研究でよく用いられる単語「line」について調べた。調べた246件の「line」のうち214件が正しく、正解率は86.7%と高かった。

● 原 文 ●　a development | group 11 | communicates | **through a public line** | such as the internet | with a | development | community | , and distributes | prepared | software | . |
●訳文● 開発 | 集団 １ １ | は | 、**公衆 回線 を通して** | 、インターネット 等 | で | 作成 され た | ソフトウェア | 開発 | コミュニティ | 配布 | する 。 |
●正解● 開発 集団 １ １ は 、インターネット 等 の **公衆 回線 に より** 、 開発 コミュニティー と の コミュニケーション を 行い 、 作成 した ソフトウェア の 配布 等 を 行って いる 。

**図5「line」の訳し分け例**

実際「line」には色々な訳がある。直線や線分や回線や視線の「線」、「伝送路」の「路」、「行列」「改行」の「行」など。統計翻訳はこれらをかなりよく訳しわけている。図 5に例を示す。太字が「line」に関わる部分である。訳し分けが高精度なのは、フレーズによって局所文脈が捉えられているためと考えられる（「｜」はフレーズの境界を示している）。

## 4　おわりに

今回の実験は、特許文の大規模な対訳コーパスを用いて標準的なフレーズベース統計翻訳の手法に従ってシステムを構築した場合の翻訳性能を確かめた。

① 各原文に対して参照訳1個という条件で見込みのあるBLEUスコア0.27を得た。

② 訳語選択は良好であるが、語順の選択はかなり問題があると考えられる。

報告者は、今回の実験を出発点として、今後、様々な工夫を積み重ねていけば、性能の大幅な改善が可能と考えている。

**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] Koehn, Philipp and Och, Franz J. and Marcu, Daniel, Statistical Phrase-Based Translation, In Proc. of HLT/NAACL03, pp.127-133.

[3] 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.