句レベルを用いた統計的後編集の精度向上

諏訪東京理科大学システム工学部
電子システム工学科教授
江原　暉将

PROFILE 1967年早稲田大学理工学部電気通信学科卒。同年NHK入局、技術現業局を経て1970年より放送技術研究所に所属。情報検索、音声認識、機械翻訳の研究などに従事。2003年より現職。
アジア太平洋機械翻訳協会（AAMT）Japio特許翻訳研究会副委員長

eharate@rs.suwa.tus.ac.jp

1 はじめに

筆者らは、特許文の日英機械翻訳の中で統計的後編集という手法を研究してきた[*1,*2]。これは、【図1】の右側に示すように、まず入力の日本語文を既存の規則方式日英機械翻訳システムによって機械翻訳英語文に変換する。その英語文に統計的後編集を加えて、より精度の高い後編集済英語文にするというものである。

従来の統計的機械翻訳では、語順が大幅に異なる日本語と英語間で精度良く翻訳することが困難であったため、本手法では、語順の変更を規則方式日英機械翻訳にゆだね、もっぱら訳語選択の精度向上を統計的手法で実現しようとするものである。統計的後編集は、英語から英語という単一言語内での統計的機械翻訳ということもできる。

統計的後編集のためには【図1】中央に示すように、翻訳モデルと言語モデルが必要であり、それらは学習データから機械学習の手法で構成される。翻訳モデルの学習データとしては、後編集前の英語文と後編集後の英語文のペアが大量に必要である。筆者らの場合、後編集前の英語文として公開特許公報の規則方式日英機械翻訳結果を用い、対応する後編集後の英語文としては、当該公報を人手で翻訳したPAJ（Patent Abstract of Japan）の英語文を利用した。また、言語モデルの学習データとしてはPAJの英語文を用いた。

従来の研究では、翻訳モデルとして単語レベルのモデルを用いてきたが、本報告では、近年発展が著しい句レベルのモデルを利用する。正確な翻訳を行うには出来るだけ広い範囲の文脈を利用するほうが良く、単語より範囲の広い句[注1]を用いることで精度の向上が期待できる。

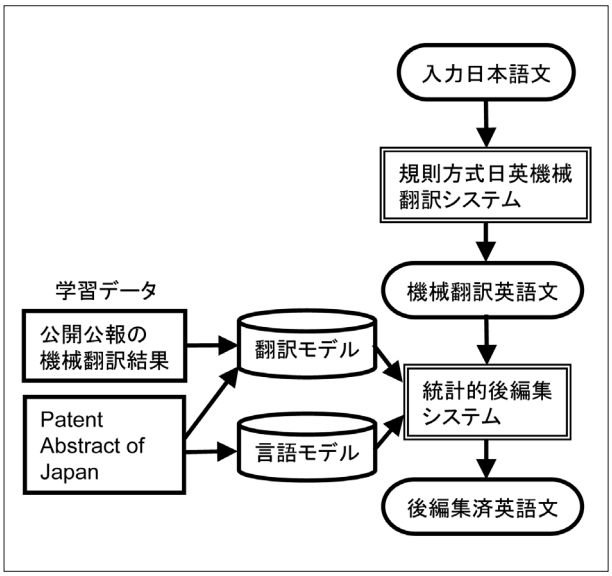

図1　統計的後編集システム

2 実験方法

句レベルのモデル学習機や統計的後編集を実行するデコーダは、既存の統計的機械翻訳システムであるMoses[*3]を利用した。実験に用いた学習データや試験データはこれまでの実験[*1,*2]とほぼ同様のものを用いた。

注1　ここで言う句とは、言語学での句の定義とは異なり、単に単語連続といった意味である。

ただし、Mosesではパラメータ調整のための開発データを必要としているため、これまでの実験とデータの内容が若干異なっている。具体的には公開特許公報の要約部分の中の「課題」の箇所および対応するPAJの“problem to be solved”の部分を利用した。学習データ、開発データ、試験データの規模を【表1】に示す。

表1　使用データの規模

データ種類	データ量
言語モデル学習データ	33万文
翻訳モデル学習データ	93420文
開発データ	188文
試験データ1（closed test）	188文
試験データ2（open test）	188文

なお、参考までに統計的機械翻訳も実施した。このとき用いたデータの規模は【表1】と同一である。統計的機械翻訳の場合は、当然、日本語から英語への直接翻訳となる。

3 実験結果

翻訳実験の結果を以下の（a）から（f）の自動評価指標で評価する。（a）～（e）についての詳細は、文献[*4]で説明されている。また（f）は英語の構文解析器であるCharniak parser[*5]が出力するスコアを評価対象英文に含まれる単語数で除した値である。

（a）BLEU（Bilingual Evaluation Understudy）
（b）NIST（National Institute of Standards and Technology score）
（c）NMG_REF（Normalized Mean Grams by the Reference）
（d）NMG_COR（Normalized Mean Grams by the Corpus）
（e）METEOR（Metric for Evaluation of Translation with Explicit Ordering）
（f）Charniak score

実験結果を【表2】に示す。いずれの評価指標も正の方向に大きい値が評価値が高い。規則方式機械翻訳結果と比較して統計的後編集結果のほうが、いずれも評価値が高く、統計的後編集の有効性が示されている。また、単語レベルの統計的後編集と比較して句レベルのそれは評価値が高く、今回の手法で翻訳精度を向上できることが分かった。ただ、句レベルの統計的後編集と句レベルの統計的機械翻訳を比較すると、差が小さく、いくつかの指標では、統計的機械翻訳のほうが精度が良いという結果になった。付録に示す翻訳例を目視しても、両者の差は少ない。

表2　実験結果

試験データ	システム	BLEU	NIST	NMG_REF	NMG_COR	METEOR	Charniak score
1 (closed test)	規則方式機械翻訳	0.1049	4.3964	-0.1930	0.9237	0.4883	-10.4108
1 (closed test)	統計的後編集(単語レベル)	0.2071	5.3806	0.1237	0.9002	———	-11.3057
1 (closed test)	統計的後編集(句レベル)	0.4551	8.2645	0.8987	1.0723	0.7381	-10.0308
1 (closed test)	統計的機械翻訳(単語レベル)	0.1314	4.5194	-0.1494	0.7150	———	-12.4396
1 (closed test)	統計的機械翻訳(句レベル)	0.4547	8.3648	0.9544	1.0480	0.7359	-10.3565
2 (open test)	規則方式機械翻訳	0.1081	4.4032	-0.1463	0.9288	0.5026	-10.4562
2 (open test)	統計的後編集(単語レベル)	0.1728	4.7893	0.0533	0.8693	———	-11.4846
2 (open test)	統計的後編集(句レベル)	0.2912	6.3398	0.4412	1.0989	0.6419	-9.9857
2 (open test)	統計的機械翻訳(単語レベル)	0.0940	3.8980	-0.3000	0.7004	———	-12.4223
2 (open test)	統計的機械翻訳(句レベル)	0.2821	6.5346	0.4244	1.0766	0.6444	-10.3069

4 結果の考察 ー産業日本語との関連ー

前節に示したように句レベルの統計的後編集を行うことで機械翻訳の精度を向上できた。しかしながら、翻訳例を見ても分かるとおり、構文的に崩れた翻訳結果となる場合がある。特許文書は長文であり、係り受け関係が複雑になる。翻訳例に示した、規則方式機械翻訳結果を見ると、規則方式システムの構文解析部で、係り受け関係の解析に誤りがあるようである。

第1の翻訳例では、「帯状土塊を破砕するとともに、」の正しい係り先は「土塊が自動的に落下する」であるが、規則方式の構文解析では「培土作業を停止すると」に係っていると解釈しているようである。本システムは語順の変更を規則方式機械翻訳にゆだねており、規則方式での構文解析に誤りがあると、誤訳につながってしまう。第2の翻訳例でも、「未然に防止するとともに、」の係り先が「実現させることのできる」であるべきなのにシステムは「制御装置を提供する。」に係っていると解釈しているようである。

産業日本語プロジェクトでは、係り受け関係の明確な日本語表現を目指しており、また、書き手が考える係り受け関係をタグとして明示的に表示することも考慮されているようである。原文の係り受け関係が明確になれば、構文解析の精度が上がり、訳文としての英語の語順の正確度も上がる。その上で、統計的後編集による訳語選択の精度向上が加わることでトータルの翻訳精度を向上させることができる。このように産業日本語プロジェクトの成果を利用することで機械翻訳の精度向上が実現できると考えるが、合わせて、機械翻訳システム自体の性能向上も図られるべきである。規則方式、統計方式、用例方式、いずれの方式をとるにしても、あるいは、筆者らのシステムのようにこれらの方式を組み合わせて用いるにしても、システム自体の性能向上が基本であり、それと産業日本語など原文の分かりやすさを向上させることで、機械翻訳の利用が一層促進されるであろう。

5 おわりに

規則方式日英機械翻訳に句レベルの統計的後編集を組み合わせたシステムを用いて、特許文書の日英機械翻訳実験を行い、精度を評価した。従来の単語レベルのシステムと比較して精度の向上が得られた。

今後の課題として、規則方式システムの構文解析精度の改善が上げられる。そのための一方法として、日本語と英語の対訳文書から日本語構文解析のための知識を自動的に学習する手法が考えられる。

これらシステムの改良と合わせて、産業日本語プロジェクトの成果を利用することで、構文解析精度を一層向上させ、高精度な機械翻訳システムの実現を目指したい。

参考文献

＊1 江原暉将：特許文の日英機械翻訳結果とPAJを比較して翻訳知識を抽出する研究，創立20周年記念誌，日本特許情報機構，pp.172-175（2005）。

＊2 江原暉将：規則方式機械翻訳と統計的後編集を組み合わせた特許文の日英機械翻訳，Japio 2006 Year Book，日本特許情報機構，pp.184-187（2006）。

＊3 Philipp Koehn et al.：Moses：Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics（ACL）, demonstration session, Prague, Czech Republic, June （2007）.

＊4 江原暉将：新しい機械翻訳自動評価基準NMGの提案，Japio 2007 Year Book，日本特許情報機構，pp.238-241（2007）。

＊5 Eugene Charniak：
A Maximum-Entropy-Inspired Parser, Proceedings of the 1st Meeting of the North American Chapter of the Association

for Computational Linguistics, pp.132-139, (2000).

付録　翻訳例

[closed test]

公報原文：帯状土塊を破砕するとともに，培土作業を停止すると抱き込まれていた土塊が自動的に落下する構造の培土器を提供する。

PAJ：to provide a ridger having a structure capable of crushing zonal clod and automatically dropping caught clod when ridging operation is stopped .

規則方式機械翻訳：to provide structural Tstica that it wins to over one's side and the clod falls automatically when a zonal clod is crushed , and the Tstica work is stopped .

統計的後編集（単語レベル）：to provide a ridger which is ridger to side and a clod falls down automati when a band-shaped screened is ignitionability and a drip-watering operation is stopped .

統計的後編集（句レベル）：to provide a ridger having screened and to automatically dropping when ridging operation is stopped .

統計的機械翻訳（単語レベル）：to up and crushing , belt-like screened , &ge；1.2 clods work stop and formed by ridger a screened is right-side from falling to the structure .

統計的機械翻訳（句レベル）：to provide a ridger having a structure capable of crushing zonal clod and the operation is stopped and automatically dropping clod .

[open test]

公報原文：電動機や制御回路が過負荷を受けるのを未然に防止するとともに，スムーズな乗り心地を実現させることのできる電気自動車の制御装置を提供する。

PAJ：to provide a control device for an electric vehicle , which prevents overload on an electric motor and a control circuit , and to realize smooth ride .

規則方式機械翻訳：to prevent the electric motor and the controlling circuit from receiving the overload beforehand , and to provide the controller of the electric vehicle that can achieve smooth riding comfort .

統計的後編集（単語レベル）：to prevent a electric motor and a control circuit from receiving an overload in advance , and to provide a transmission） of a motor-driven vehicle which can realize planarizing riding comfortableness .

統計的後編集（句レベル）：to prevent an electric motor and a control circuit from being overload in advance , and to provide a control device for an electric vehicle that can realize smooth ride comfort .

統計的機械翻訳（単語レベル）：to obtain a smooth ride comfort the electric vehicle controlling device which can realize preventing the occurrence of electric machine and a control circuit is excessive load influence to other and <s> .

統計的機械翻訳（句レベル）：to provide a control device for an electric vehicle capable of realizing smooth , preventing overload of a motor and a control circuit for receiving and ride comfort .