語順の入れ替えに着目した特許の統計翻訳
—事前・事後並べ替えによる高精度な英日・日英翻訳—

Statistical Patent Translation Focusing on Word Reordering

日本電信電話株式会社　コミュニケーション科学基礎研究所研究主任　須藤　克仁

PROFILE: 2002 年京都大学大学院情報学研究科修士課程修了、同年 NTT 入社。統計的機械翻訳、音声言語処理の研究に従事。

日本電信電話株式会社　コミュニケーション科学基礎研究所主任研究員　鈴木　潤

PROFILE: 2001 年慶應義塾大学大学院理工学研究科修士課程修了、同年 NTT 入社。博士（工学）。自然言語処理、機械学習の研究に従事。

日本電信電話株式会社　コミュニケーション科学基礎研究所主任研究員　塚田　元

PROFILE: 1989 年東京工業大学理工学研究科修士課程修了、同年 NTT 入社。統計的機械翻訳、音声言語処理の研究に従事。

日本電信電話株式会社　コミュニケーション科学基礎研究所上席特別研究員　永田　昌明

PROFILE: 1987 年京都大学大学院情報工学研究科修士課程修了、同年 NTT 入社。博士（工学）。統計的自然言語処理、統計的機械翻訳の研究に従事。

国立大学法人総合研究大学院大学複合科学研究科情報学専攻　星野　翔

PROFILE: 2011 年総合研究大学院大学情報学専攻 5 年一貫制博士課程に入学。統計的機械翻訳の研究に従事。

大学共同利用機関法人情報・システム研究機構　国立情報学研究所コンテンツ科学研究系准教授　宮尾　祐介

PROFILE: 2001 年東京大学大学院情報理工学系研究科助手、2010 年より国立情報学研究所准教授。博士（情報理工学）。構文解析とその応用の研究に従事。

1 はじめに

機械翻訳は外国特許情報の検索や調査に有益な技術であり、現在も様々な形で機械翻訳ソフトウェアが利用されている。近年では大量の対訳文例を利用して機械翻訳の知識を統計的に記述し利用する統計翻訳の発展に伴い、西欧言語間のように語彙や語順の差が比較的小さい場合には高精度で翻訳できるようになってきた。しかし、日英間においては、既存の統計翻訳技術では翻訳精度が他言語間と比較して低いことが多く、実用に向けては依然改善を要するのが実情であった。

筆者らは NTCIR 特許翻訳タスク [1] 等を通して日英間の統計翻訳の問題、特に語順の大きな違いに起因する翻訳時の語順誤りを解決するための研究を行ってきた。本稿ではその研究成果を紹介する。

2 統計翻訳における語順誤り

機械翻訳では、対訳辞書のような情報を用いた語句の翻訳と、正しい語順の翻訳結果を得るための語順の入れ替えの二つの問題を同時に解く必要がある。語句の翻訳は対訳文例の数を増やすことである程度改善できること

が多いが、語順の入れ替えはモデル化の難しさや探索のための計算量の多さゆえに、特に日英のような語順の大きく異なる言語対では難しい問題である。

表 1 に統計翻訳による誤訳の典型的な例を示す。ここで用いた統計翻訳は、句に基づく統計翻訳を行うソフトウェア Moses[2] を、NTCIR 特許翻訳タスクで提供された日英特許対訳文（約 320 万文）で学習したものである。大量の対訳文例を利用することで個々の単語はほぼ妥当な翻訳を得ることができているが、語順の誤りが非常に大きく、元の文意を読み取ることは難しい。

表 1 標準的な統計翻訳における誤訳の例（英日）

原文	The outlet port 64 of the pump 60 connects to a pump passage defined between the casing 41 and the support plate 46.
参照訳	ポンプ60の吐出口64は、ケース41と支持板46との間に形成されるポンプ通路52に接続している。
句に基づく統計翻訳	ポンプ通路52に接続され、出口ポート64との間に形成され、ケーシング41とポンプ60の支持板46が設けられている。

語順の入れ替えの問題は、統計翻訳の進歩と対象となる言語の増加に合わせて、より注目を集めるようになってきている。日英はその中でも特に語順の違いが大きく、翻訳が難しいと考えられてきた。例えば、2008 年に開催された NTCIR-7 特許翻訳タスク [3] の人手での翻訳評価結果において、当時最先端の統計翻訳をもってしても規則ベース翻訳には及ばず、語順誤りが訳質に大きく影響していた。

3 英語の主辞後置並べ替えを用いた英日翻訳

英日間の翻訳における語順の違いという問題に対し、NTT では日本語の「主辞後置性」に着目した高精度な統計翻訳手法を考案した [4]。

日本語は主辞、すなわち句の文法的役割の中心となる語（例えば動詞句における動詞）が、主辞を修飾する語の後方に置かれるという文法的特徴（主辞後置性）を持つ。英語においては主辞は他の語の前方に置かれる（主辞前置性）ことを基本としつつも、主語等例外も多く、日本語との語順の違いの大きな要因となっている。

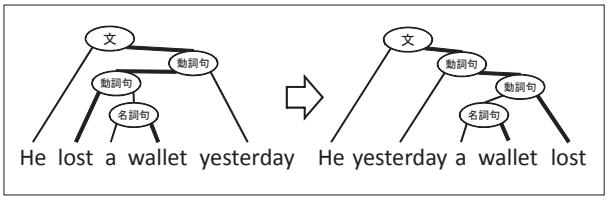

図 1　構文木上での主辞後置並べ替えの例。太線が主辞となっている要素への枝となっている。

表 2　主辞後置並べ替えと翻訳結果（ _ga は格助詞「が」に相当する仮想単語）

主辞後置並べ替え	pump 60 of outlet port 64 _ga casing 41 and support plate 46 between defined pump passage to connects.
翻訳結果	ポンプ60の出口ポート64は、ケーシング41及び支持板46との間に形成されるポンプ通路52に接続する。

ところが、この主辞後置性を利用すると、英語の主辞を後方に置くように語順を入れ替えることで、あたかも日本語であるかのような語順にすることができる（図 1）。このような並べ替えを「主辞後置並べ替え」と呼び、並べ替えた日本語語順の英語を「主辞後置英語」と呼ぶ。主辞後置並べ替えを翻訳の前に行うことによって、英日翻訳は逐語訳に近い簡単な翻訳の問題に変換することができ、効率的かつ精度の高い翻訳ができるようになる。翻訳を行う前に並べ替えを行うことから、このような方法を総称して「事前並べ替え」と呼んでいる。また、より日本語への翻訳をしやすくするために、日本語の格助詞「が」（主格）、「を」（目的格）に相当する仮想的な単語を挿入し、翻訳時の格助詞の脱落や誤訳を防ぐようにしている。表 1 の例文に対する主辞後置並べ替えの結果とそれを統計翻訳によって翻訳した結果を表 2 に示す。主辞後置並べ替えによってほぼ日本語と同じ語順に並べ替えができていること、またよい翻訳結果が得られていることが見て取れる。

主辞後置並べ替えによって英日翻訳の精度は従来と比べて大きく改善し、2011 年の NTCIR-9 では規則ベース翻訳を初めて上回り [5] [6]、2013 年の NTCIR-10 では構文解析を大規模な半教師あり学習 [7] により特許文の解析に適応させることで特許特有の表現に対する構文解析性能を向上させ、翻訳精度の差をさらに広げることに成功した [1] [8]。

表３　二段階並べ替えと翻訳結果の比較

原文	本実施形態に係る光走査型顕微鏡装置１は、図１に示される顕微鏡観察システム２において使用される。
参照訳	An optical-scanning microscope apparatus 1 according to this embodiment is included in a microscope examination system 2 shown in FIG.1.
二段階並べ替え	は 光 走査 型 顕微 鏡 装置 1 係る に 本 実施 形態 、れる 使用 さ おいて に 顕微 鏡 観察 システム 2 れる 示さ に 図 1 。
翻訳結果	The optical scanning type microscope apparatus 1 according to the present embodiment is used in a microscope observation system 2 shown in FIG. 1.
句に基づく統計翻訳	According to the present embodiment, the scanning optical microscope apparatus 1 shown in FIG.1 is used in a microscopic observation system 2.

4 日本語の文節間・文節内二段階事前並べ替えを用いた日英翻訳

先に述べた通り、英語の主辞の位置は一貫していないため、主辞後置並べ替えと同様の方法によって日本語を英語の語順に並べ替えることは難しい。そのため、主辞に代わる特徴として、主語・目的語と述語の関係、及び係り受け関係に着目する。

日本語と英語の語順の違いとして広く知られているものとして、日本語が主語、目的語、述語の順（SOV）であるのに対し、英語は主語、述語、目的語の順（SVO）であることが挙げられる。この特徴に合致するように日本語を SVO の順に並べ替えることを考える。

並べ替えの単位としては日本語の文節を利用する。日本語の文節、英語の文節に相当する句を考えたとき、この基本単位を跨ぐような語順の入れ替えはあまり発生しないとされている。そこで、図２に示すように、文節の順序を入れ替え（文節間の並べ替え）、さらに文節内の語順も入れ替える（文節内の並べ替え）ことによって、英語に近い語順に入れ替えるようにする [9]。

文節間の並べ替えでは、日本語の主語、目的語、動詞にそれぞれ相当する文節を、主語、動詞、目的語の順に並べ替える [10]。具体的には、動詞を主語の直後または目的語の直前に移動することで、SOV の語順を SVO に変形することができる。さらに、主語も目的語も無い場合に動詞を最終文節の直前に移動することで、日本語の SOV からより英語の SVO と似た語順にすることができる。

文節内の並べ替えでは、文節内の語を内容語と機能語の２種類に分類し、それぞれの分類内の語順は保持したまま、機能語を内容語の直前に移動する規則を提案した [9]。これは文節末尾の日本語の助詞が英語側では前置詞に対応することを利用している。

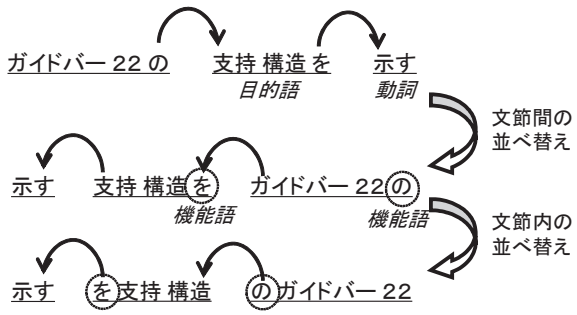

図２　日本語の並べ替え例（実線は係り受け関係を示す）

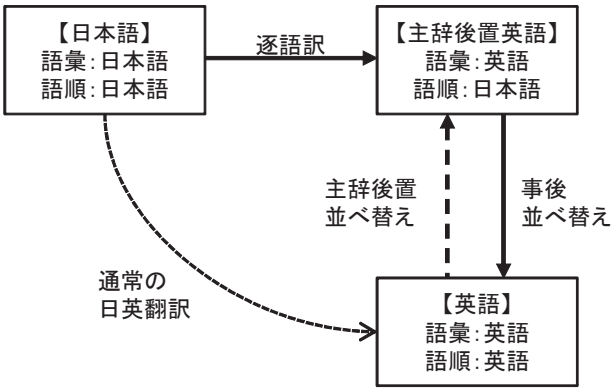

図 3　日英翻訳における主辞後置英語を介した事後並べ替え

文節内の単語の並べ替えに関しては、文節内の語順を逆転させる方法が NTCIR-7 の日英翻訳で利用されている [11] が、固有名詞など並べ替える必要のない語句も逆順となってしまう問題点があった。

この文節間・文節内の二段階の並べ替えを行う、日英事前並べ替え手法を適用することで、日英翻訳の精度が改善することを確認した。また、次に説明する事後並べ替えを用いた手法との翻訳システム組み合わせの結果、NTCIR-10 において、規則ベース翻訳には及ばなかったものの、統計翻訳システムの中では最も高い翻訳精度となった [1] [8]。

5 主辞後置英語を介した事後並べ替えを用いた日英翻訳

日英翻訳における事前並べ替えが英日翻訳における主辞後置並べ替えと比較して容易でなく、精度向上の度合いも若干小さいものであったため、筆者らは「事後並べ替え」という新しいアプローチを考案し、日英特許翻訳において有効性を確認した [12]。

事後並べ替えは、日本語文をまず逐語訳した結果を、正しい英語の語順に並べ替える処理を指す。このとき、日本語文を逐語訳したものは日本語語順の英語、すなわち主辞後置英語であり、日英翻訳における事後並べ替えは、英日翻訳における主辞後置並べ替えの逆問題に相当する（図 3）。この問題は主辞後置英語から英語への翻訳の問題と捉えることができ、統計翻訳の技術を利用して解決できる。

英語への翻訳においては英語の構文情報を利用した翻訳手法が正しい英語の語順を得るために有用であることが知られているが、一般によい翻訳精度を得るためには語句の翻訳と語順の入れ替えの膨大な候補を探索する必要があり翻訳速度が非常に遅くなってしまう。事後並べ替えにおいては語句の翻訳を考慮することなく語順の入れ替えのみを行えばよいため、翻訳精度を落とさずに高速に翻訳を行うことができる。

6 おわりに

本稿では、英日・日英翻訳における筆者らの研究成果について紹介した。語順の入れ替えは機械翻訳の重要な問題であるが、語順の違いの大きい日英の翻訳であることに加え、文が比較的長い特許の翻訳においてはその重要度はさらに大きいと言える。本稿では日英間の翻訳について紹介したが、筆者らは中日翻訳への応用についても効果を確認している [13]。今後は日英翻訳の更なる高精度化を含めた他言語への展開と合わせ、特許で多く見られる長い名詞句の頑健な翻訳についても検討したい。

参考文献

[1] I. Goto, K. P. Chow, B. Lu, E. Sumita and B. K. Tsou, "Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop," in *Proceedings of the 10th NTCIR Conference*, 2013.

[2] "Moses," [Online]. Available: http://www.statmt.org/moses/.

[3] A. Fujii, M. Utiyama, M. Yamamoto and T. Utsuro, "Overview of the Patent Translation Task at the NTCIR-7 Workshop," in *Proceedings of the 7th NTCIR Workshop Meeting*, 2008.

[4] H. Isozaki, K. Sudoh, H. Tsukada and K. Duh, "Head Finalization: A Simple Reordering Rule for SOV Languages," in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, 2010.

[5] I. Goto, B. Lu, K. P. Chow, E. Sumita and B. K. Tsou, "Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop," in *Proceedings of the 9th NTCIR Workshop Meeting*, 2011.

[6] K. Sudoh, K. Duh, H. Tsukada, M. Nagata, X. Wu, T. Matsuzaki and J. Tsujii, "NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT," in *Proceedings of the 9th NTCIR Workshop Meeting*, 2011.

[7] J. Suzuki, H. Isozaki, X. Carreras and M. Collins, "An Empirical Study of Semi-supervised Structured Conditional Models for Dependency Parsing," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009.

[8] K. Sudoh, J. Suzuki, H. Tsukada, M. Nagata, S. Hoshino and Y. Miyao, "NTT-NII Statistical Machine Translation in NTCIR-10 PatentMT," in *Proceedings of the 10th NTCIR Conference*, 2013.

[9] S. Hoshino, Y. Miyao, K. Sudoh and M. Nagata, "Two-Stage Pre-ordering for Japanese-to-English Statistical Machine Translation," in *Proceedings of the 6th International Joint Conference on Natural Language Processing*, 2013.

[10] M. Komachi, Y. Matsumoto and M. Nagata, "Phrase Reordering for Statistical Machine Translation Based on Predicate-Argument Structure," in *Proceedings of International Workshop on Spoken Language Translation*, 2006.

[11] J. Katz-Brown and M. Collins, "Syntactic Reordering in Preprocessing for Japenese-to-English Translation: MIT System Description for NTCIR-9 Patent Translation Task," in *Proceedings of the 7th NTCIR Workshop Meeting*, 2008.

[12] K. Sudoh, X. Wu, K. Duh, H. Tsukada and M. Nagata, "Syntax-Based Post-Ordering for Efficient Japanese-to-English Translation," *ACM Transactions on Asian Language Information Processing*, vol. 12, no. 3, 2013.

[13] D. Han, K. Sudoh, X. Wu, K. Duh, H. Tsukada and M. Nagata, "Head Finnalization Reordering for Chinese-to-Japanese Machine Translation," in *Proceedingssss of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2012.