

特許文書の多言語機械翻訳

Multi lingual machine translation for patent documents



元山梨英和大学教授 **江原 暉将**

1967年早稲田大学理工学部卒。同年NHK入局。2003年諏訪東京理科大学教授。2009年山梨英和大学教授。2015年退職。アジア太平洋機械翻訳協会（AAMT）／Japio特許翻訳研究会委員。

1 はじめに

知的財産への関心が高まるにつれ、特許出願の件数が世界的に増加している。WIPOが各国の出願件数を調査したデータを公表しており^[1]、それをもとに2014年の出願件数が100件以上の国を【付表1】に示す。そのような国は85カ国にのぼる。

特許出願の書類は各国の特許庁が指定する言語で記述されており、その国への出願はもちろん、その国の特許情報にアクセスするにも記述言語の理解が必要である。英語やフランス語などが記述言語として用いられる場合もあるが、各国の現地語で書かれた特許文書も多い。【付表1】には、文献[2]などを参考にして、各国の出願書類で用いられている言語名を示してある。用いられている言語数は51であり、出願件数の上位20カ国に限っても16言語を数える。

外国語で記述された特許情報に日本語でアクセスするためには翻訳が必要である。このように多くの種類の外国語で書かれ、しかも大量にある特許文書を人手で翻訳することは事実上不可能であり、機械翻訳が期待されている。すでに英語をはじめ中国語や韓国語など主要な言語と日本語の間では、機械翻訳が実用されているが^[3]^[4]、今後、これらの主要言語以外の言語にも機械翻訳の需要が広がると予想される。

2 ピボット翻訳

多くの言語間の機械翻訳を行うためには多くの翻訳システムを作る必要がある。例えば【付表1】に示す51言語間の相互翻訳を通常の2言語間の翻訳で実現しようとする^[5]とすると $51 \times 50 = 2550$ 方向の翻訳システムを必要とする。必要な翻訳システム数を少なくする方法としてピボット方式がある。ピボット方式とは、一つの言語を中心となる言語（ピボット言語）として定め、ピボット言語とその他の言語との間の翻訳システムだけを作り、すべての言語間の翻訳をピボット言語を介して行う方式である。51言語の中の一つの言語をピボット言語とすると、必要なシステム数は $50 \times 2 = 100$ です。このようなピボット方式は古くから提案されており^[5]、最近では統計的機械翻訳の枠組みの中で行われるようになった^[6]。

統計的機械翻訳では翻訳元言語（A言語とする）と翻訳先言語（B言語とする）の間で対訳関係にある文を大量に集めた「対訳コーパス」を元にして翻訳システムを構築する。しかしA言語とB言語との間で、大規模な対訳コーパスが得られない場合、ピボット言語（P言語とする）を介することで、上記問題を回避しようとするのが、統計的機械翻訳におけるピボット方式である。A言語とP言語との間には、大規模な対訳コーパスが存在し、P言語とB言語との間にも、大規模な対訳コーパスが存在する場合に、P言語を介してA言語とB言語を結びつける。「結び付け方」にはいくつか考えられるが^[6]、ここでは最も単純な方式について説明しよう。

まず、P 言語と B 言語の間の対訳コーパスを元に、P 言語から B 言語への翻訳システム (PB 翻訳システムと呼ぶ) を構築する。次に A 言語と P 言語の間の対訳コーパスの P 言語側の各文を PB 翻訳システムを用いて B 言語に機械翻訳する。この機械翻訳結果と対応する A 言語側の文をペアにすることで A 言語と B 言語との間の対訳コーパスが得られる。この対訳コーパスを用いて、A 言語から B 言語への翻訳システムが構築できる。

3 ピボット言語

ピボット方式を実現するためには何語をピボット言語とするかがまず問題になる。文献 [5] では「中間言語」と呼ばれる人工的な言語をピボット言語としている。しかし、大量な対訳コーパスを必要とする統計的機械翻訳では、人工言語のコーパスを得ることが難しいため、自然言語をピボット言語として用いることになる。特に、国際的な共通言語である英語をピボット言語とする場合が多い。

ここで、機械翻訳を単純にモデル化すると、機械翻訳システムの中では、翻訳元言語の「語順」と「語彙」を翻訳先言語のそれに変換する作業が行われている。語順変換と語彙変換が機械翻訳の 2 大要素技術である。そこで翻訳元言語と翻訳先言語で語順と語彙が類似しているか否かで翻訳精度が大きく異なる。英語とフランス語間あるいは日本語と韓国語間などのように、語順と語彙が類似している言語間では機械翻訳の精度は比較的高くなり、類似していない言語間では精度が低くなる。類似していない言語の代表例として英語と日本語がある。

特許文書では専門用語の翻訳 (語彙変換) が重要なポイントであるが、対訳コーパスの大規模化とそこから抽出した専門用語辞書の充実によって、語彙変換の精度は高くなっている。一方、語順変換は、対訳コーパスの大規模化だけでは解決が難しく、課題として残っている。

語順の違いを特徴づける性質として最も重要なものは目的語 (O) と動詞 (V) の順序である^[7]。日本語は目的語が動詞に先行する OV 型であり、英語は動詞が目的語に先行する VO 型である。【付表 1】に示されている 51 言語を、文献 [8] などを参考にして、OV 型と VO 型で分類すると【表 1】のようになる。

表 1 目的語 (O) と動詞 (V) の語順型

O と V の語順	言語
OV 型	Armenian, Azerbaijani, Georgian, Hindi, Japanese, Kazakh, Kirghiz, Korean, Mongolian, Persian, Turkish, Uzbek
VO 型	Arabic, Belorussian, Bulgarian, Chinese, Croatian, Czech, Danish, English, Estonian, Filipino, Finnish, French, Greek, Hungarian, Hebrew, Icelandic, Indonesian, Irish, Italian, Latvian, Lithuanian, Malay, Maltese, Maori, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Spanish, Swedish, Thai, Ukrainian, Vietnamese, Welsh
その他	Dutch, German

VO 型言語が多いが、OV 型言語も一定の数使われている。OV 型言語と VO 型言語の間は語順の違いが大きく、それらの言語間での機械翻訳の精度を低下させている。

話をピボット翻訳に戻すと、世界で広く用いられている英語や中国語がピボット言語の候補として考えられるが、両言語とも VO 型であり、OV 型同士の機械翻訳に、これらの VO 型言語をピボット言語として用いることは、少なくとも語順の観点からは、望ましくない。例えばヒンディ語と日本語は共に OV 型であり、目的語と動詞の語順以外にも【表 2】に示すようなさまざまな語順の類似性がある^{[8][9]}。

表 2 日本語、ヒンディ語、英語の語順

語順特徴	日本語	ヒンディ語	英語
主語 (S) と動詞 (V)	SV	SV	SV
目的語 (O) と動詞	OV	OV	VO
斜格 (P) と動詞	PV	PV	VP
前置詞 (Pr) か後置詞 (Po) か	Po	Po	Pr
所有格 (G) と名詞	GN	GN	GN & NG
指示詞 (D) と名詞	DN	DN	DN
数詞 (Nu) と名詞	NuN	NuN	NuN
形容詞 (A) と名詞 (N)	AN	AN	AN
関係節 (R) と名詞	RN	NR & RN	NR
固有名詞 (Pn) と普通名詞 (N)	PnN	PnN	NPn & PnN
本動詞 (V) と助動詞 (X)	VX	VX	XV
副詞 (D) と動詞	DV	DV	DV & VD

例えば【図 1】に示す例文¹の単語対応を見ると、ヒンディ語（3 行目）と日本語（1 行目と 4 行目）の単語対応は、ヒンディ語と英語（2 行目）あるいは日本語と英語と比較して交差する単語対応が少なく、語順が類似していることが分かる。

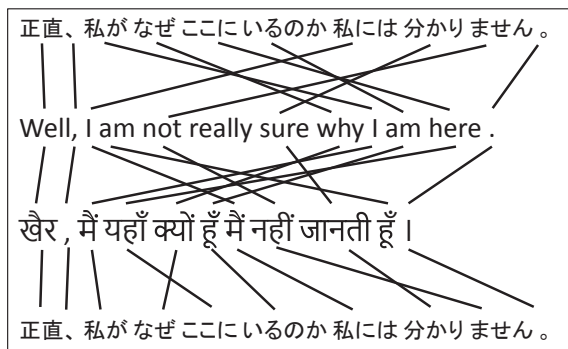


図 1 日本語、英語、ヒンディ語の単語対応例

一方、【表 1】に示す OV 型言語には世界的に広く用いられている言語が少なく、対訳コーパスが圧倒的に豊富な英語をピボット言語とする場合が多い。しかし、OV 型言語間の翻訳において VO 型言語をピボット言語として用いるのは望ましくない。このジレンマを解決する一つの手段として語順の並べ替えがある。

4 語順の並べ替え

語順の大きく異なる言語間の機械翻訳において原文の語順を訳文のそれに近づけてから翻訳する事前並べ替えが提案された^[11-13]。特に日本語の持つ主辞後置性に着目すると、原文を構文解析し、その結果得られた構文木の主辞を後方に移動させるという Head Finalization と呼ばれる方法で並べ替えが実現できる^[14]。その結果、特許文書の英日翻訳や中日翻訳などに対して、翻訳精度を大幅に向上させることができた^[15-17]。

このような事前並べ替えの技術をピボット翻訳に応用することが考えられる。ヒンディ語から日本語への翻訳において英語をピボット言語として用いる場合を例に説明しよう。ヒンディ語と英語の対訳コーパスおよび英語と日本語の対訳コーパスの英語側の語順を OV 型言語に合った語順に並べ替えておく。そのような英語を

「OV 型英語」と名づける。これによってヒンディ語と OV 型英語および OV 型英語と日本語の対訳コーパスが得られる。次に OV 型英語と日本語の対訳コーパスを用いて OV 型英語から日本語への機械翻訳システムを構築する。この構築は上述した事前並べ替えによる英日システムの構築と同様に行える。次にヒンディ語と OV 型英語の対訳コーパスの OV 型英語部分を OV 型英語から日本語への機械翻訳システムで日本語に機械翻訳する。この機械翻訳結果と元のコーパスのヒンディ語部分とからヒンディ語と日本語の対訳コーパスが得られる。最後に、このコーパスを用いてヒンディ語から日本語への機械翻訳システムを構築する。ヒンディ語と日本語は語順が類似しているため、通常の句ベースの統計的機械翻訳を用いても精度の良いシステムを構築できることが期待できる。

上記のアイデアに基づいて、Workshop on Asian Translation 2016 (WAT2016) のヒンディ語から日本語への翻訳タスク^[18]において実験を行った。ここでは、英語をピボット言語とし、語順の並べ替えを用いたシステムを実験している。このタスクは特許文書を対象としたものではないが、特許においても同様の手法が可能であると考えられる。システム内容と実験結果の詳細については Workshop で発表予定であるが、テスト文に対する自動評価を行った結果、【表 3】に示すように、語順並べ替えを用いた手法は用いない手法に比較して BLEU 値で 0.19 (2.5%) 向上している^[19]。

表 3 自動評価結果

語順の並べ替え	BLEU
なし	7.47
あり	7.66

1 WIT³(Web Inventory of Transcribed and Translated Talks) のデータから引用改変した^[10]。

5 あとがき

多言語機械翻訳を実現する一つの方法としてピボット翻訳を紹介し、特に英語をピボット言語として採用したOV型言語間のピボット翻訳の場合、語順の並べ替えが有効であることを示した。語順並べ替えを含むピボット翻訳は、表1に示すOV型言語間において英語などのVO型言語をピボット言語とする場合にも適用可能である。

今後の課題として、同じOV型言語間でも表2や図1に示すように語順の違いがあり、この違いを吸収する手段が必要になる。

いずれにしても、ここで述べた語順の並べ替えを含むさまざまな工夫を用いて、今後需要が増すと考えられる多言語機械翻訳に取り組んでゆくことが望まれる。

参考文献

- [1] WIPO : Statistical Country Profiles, IP Filings (Resident + Abroad, Including Regional), *WIPO Statistics Database*
http://www.wipo.int/ipstats/en/statistics/country_profile/ (2016年6月21日アクセス).
- [2] IP-COSTER : IP-Guide, *IP-COSTER Home Page*,
<https://www.ip-coster.com/IPGuide.aspx> (2016年6月21日アクセス)
- [3] 岡崎輝雄 : 特許庁の特許情報提供における機械翻訳の活用と今後の展望、*AAMT/Japio 特許翻訳研究会シンポジウム資料*, Nov, 2009.
- [4] 檀本英吾 : 中韓文献翻訳・検索システム、*Japio YEAR BOOK 2014 寄稿集*, pp.62-65, Dec, 2014.
- [5] 市山俊治、村木一至 : 機械翻訳システム PIVOT の中間言語、*情報処理学会第38回全国大会講演論文集、人工知能及び認知科学*, pp.345-346, Mar, 1989.
- [6] 内山将夫、伊佐原均 : 統計的機械翻訳におけるピボット翻訳の比較、*言語処理学会第13回年次大会論文集*, pp. 187-190, Mar, 2007.
- [7] 江原暉将 : 多次元尺度構成法を用いた語順パラメータの関係付け、*言語処理学会第1回年次大会発表論文集*, pp.173-176, Mar., 1995.
- [8] Matthew S. Dryer : Word Order, *The World Atlas of Language Structures*, Chapter F, *Oxford University Press*, pp. 330 – 397, 2005.
- [9] 角田太作 : 世界の言語と日本語、*くろしお出版*, p.273, April, 1991.
- [10] Mauro Cettolo, Christian Girardi, Marcello Federico : WIT³: Web Inventory of Transcribed and Translated Talks, *Proceedings of the 16th EAMT Conference*, pp.261-268, May 2012.
<https://wit3.fbk.eu/> (2016年8月24日アクセス)
- [11] Chi-Ho Li, Dongdong Zhang, Mu Li, Ming Zhou, Minghui Li, Yi Guan : A Probabilistic Approach to Syntax-based Reordering for



- Statistical Machine Translation, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 720-727, June, 2007.
- [12] Jason Katz-Brown, Michael Collins : Syntactic Reordering in Preprocessing for Japanese → English Translation: MIT System Description for NTCIR-7 Patent Translation Task, *Proceedings of NTCIR-7 Workshop Meeting*, pp. 409-414, Dec, 2008
- [13] Peng Xu, Jaeho Kang, Michael Ringgaard, Franz Och : Using a Dependency Parser to Improve SMT for Subject-Object-Verb Languages, *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pp. 245-253, June 2009.
- [14] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, Kevin Duh : Head Finalization: A Simple Reordering Rule for SOV Languages, *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pp. 244-251, July, 2010.
- [15] 須藤克仁、鈴木潤、塚田元、永田昌明、星野翔、宮尾祐介 : 語順の入れ替えに着目した特許の統計翻訳, *Japio YEAR BOOK 2013 寄稿集*, pp.292-297, Dec, 2013.
- [16] 内山将夫 : 英日・中日特許 SMT システムの実用化と課題, *Japio YEAR BOOK 2014 寄稿集*, pp.252-255, Dec, 2014.
- [17] 須藤克仁、鈴木潤、秋葉泰弘、塚田元、永田昌明 : 英中韓から日本語への特許文向け統計翻訳, *Japio YEAR BOOK 2014 寄稿集*, pp.262-267, Dec, 2014.
- [18] WAT2016 organizer : The 3rd Workshop on Asian Translation, *WAT2016 Home Page*, <http://lotus.kuee.kyoto-u.ac.jp/WAT/index.html>, 2016.
(2016年8月24日アクセス)
- [19] WAT2016 organizer : The 3rd Workshop on Asian Translation Evaluation Results, *WAT2016 Home Page*, <http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html>, 2016.
(2016年8月24日アクセス)

付表1 各国の出願件数（2014年）と出願書類の記述言語

Country name	# of filings	Language name
China	837,897	Chinese
United States of America	509,622	English
Japan	465,987	Japanese
Republic of Korea	230,556	Korean
Germany	179,535	German
France	72,369	French
United Kingdom	52,612	English, Welsh
Switzerland	44,417	German, French, Italian
Netherlands	37,738	Dutch
Italy	29,298	Italian
Russian Federation	28,515	Russian
Canada	24,715	English, French
Sweden	23,858	Swedish, English
India	22,458	English, Hindi
Finland	14,075	Finnish
Austria	13,789	German
Iran (Islamic Republic of)	13,768	Persian
Israel	13,437	Hebrew, English
Denmark	12,547	Danish, English
Belgium	12,188	Dutch, French, German
Australia	11,743	English
Spain	10,928	Spanish
Brazil	6,717	Portuguese
Turkey	6,496	Turkish
Poland	6,172	Polish
Singapore	5,930	English
Norway	5,877	Norwegian, English
Ireland	4,780	English, Irish
Saudi Arabia	4,123	Arabic
New Zealand	3,429	English, Maori
Luxembourg	3,142	French, German, English
Ukraine	2,990	Ukrainian
Malaysia	2,664	Malay, English
Kazakhstan	2,453	Russian, Kazakh
South Africa	2,329	English
Mexico	2,187	Spanish
Czech Republic	2,181	Czech
Belarus	1,781	Belorussian, Russian
Hungary	1,434	Hungarian, English
Thailand	1,405	Thai
Portugal	1,333	Portuguese
Greece	1,253	Greek
Romania	1,252	Romanian

Country name	# of filings	Language name
Chile	998	Spanish
Egypt	883	Arabic, English
Argentina	791	Spanish
Indonesia	771	Indonesian
Philippines	608	Filipino, English
Viet Nam	561	Vietnamese
Azerbaijan	542	Azerbaijani
Cyprus	493	Greek
Malta	476	Maltese, English
Barbados	474	English
Senegal	472	English, French
Bulgaria	468	Bulgarian
Colombia	461	Spanish
Cameroon	460	English, French
Slovakia	455	Slovak
Côte d'Ivoire	397	English, French
United Arab Emirates	392	Arabic
Uzbekistan	374	Uzbek, Russian
Morocco	368	French
Iceland	302	Icelandic
Serbia	289	Serbian
Estonia	278	Estonian
Croatia	259	Croatian
Lithuania	254	Lithuanian
Pakistan	202	English
Latvia	193	Latvian
Cuba	189	Spanish
Tunisia	176	English, Arabic, French
Qatar	175	Arabic, English
Kyrgyzstan	173	Kyrgyz, Russian
Mali	163	English, French
Niger	163	English, French
Congo	162	English, French
Kenya	161	English
Monaco	159	French
Armenia	156	Armenian
Bahamas	144	English
Mongolia	140	Mongolian
Georgia	131	Georgian
Benin	109	English, French
Peru	103	Spanish
Algeria	101	French, Arabic