ISSN 1883-1818
No.77
June 2022

AAMT Journal

Asia-Pacific Association for Machine Translation

機 械 翻 訳

AAMT Journal
Asia-Pacific Association for Machine Translation 機械翻訳
No.77

目　次

C O N T E N T S

巻頭言

MT が浸透した世の中へ

出内　将夫
情報通信研究機構

今後、機械翻訳（以降、MT）の技術がさらに世の中に浸透し、当たり前に使われるようになるとしたら、どのような道を辿るのだろうか。

技術が成熟すると、様々な製品やサービスなどに組み込まれて世の中に浸透し、その技術の存在が当たり前になっていく。最終的には、使用者が技術を意識せずとも、技術がいつのまにか使われて利便性を享受できる状況になっていく。例えば、スマートフォンを使ってニュース記事を読む際には、サーバに保存された記事データをネットワーク経由で取得し、記事のテキストや画像をスマートフォンの画面に配置して表示する。このために、様々な技術が用いられているが、個々の技術が意識されることは少ない。将来、MT が利用者に意識されない形で利用される日も来るのではないだろうか。

MT や情報通信の技術がどう成熟してきたか振り返った上で、今後の MT の行く先を考える。AAMT ホームページから公開されている過去の AAMT ジャーナルを読むと、1990 年代前半の MT は高価であり、技術を使用するためのノウハウが必要だったことが伺える。その後、2000 年代にはインターネットの普及もあり MT を誰もが利用できるようになり、2010 年代後半に MT の品質が向上したことで、今や MT は様々な製品やサービスに組み込まれつつある。約 30 年間で使用者の幅が広がり、MT の翻訳品質が向上し、当たり前の技術に近づいてきたと言える。

今後、MT をさらに当たり前の技術とするには、何が必要だろうか。MT が普及した要因の一つであるインターネットの技術を参考例として考える。1990 年代のインターネット利用者は意識的にインターネットに接続していた。それが、2000 年代には常時接続が家庭やモバイル端末で普及し始め、現在はインターネット接続が当たり前となりつつある。インターネットが普及した一因に、ベストエフォートと呼ばれる考え方があると言われている。ベストエフォートとは、最低限の品質を保証せず、接続が集中する場所や時間帯では通信品質が下がることを許容する考え方である。この割り切った考え方により、回線価格が下がり、技術の普及や設備投資が進んだ結果、インターネットが当たり前になったと言われている。MT 技術でも翻訳品質の保証は難しいことから、ベストエフォートの考え方とは相性が良いと考えられる。MT 技術の普及を促進するには、ベストエフォートが通用する用途とベストエフォートでは困る用途を分けて考え、まずは前者をターゲットとすることで、技術の普及が進み、技術自体も洗練されていくと思われる。

上記では、インターネットのベストエフォートの考え方に着目し、最低限の品質を保証していないと述べたが、より高額だが高品質な、ビジネス用途向けの帯域保障付きの接続も存在する。また、インターネットでは通信品質を向上する様々な技術が使われている。例えば、ネットワーク上のデータ破壊を検知したり修正したりできる技術があり、それらが物理媒体、信号、パケットや順序、といった様々なレベルで幾重にも利用されることで、送信されたデータが正しく受信されたか確認できる。MT の分野でも、翻訳前の文の意味が、翻訳後の文に含まれているか、様々な観点で確認する技術を組み合わせれば、MT の品質保証技術が確立でき、用途の幅が広げられるかもしれない。

過去からの技術の進歩を踏まえ、これからさらに 10 年後を考えると、今後さらに MT が当たり前の技術となると見込まれる。MT 技術が適切な用途で利用されつつ、MT 技術自体も進歩を続ける良いスパイラルが続くことを願っている。

Towards the World where MT has become Ubiquitous
Masao Ideuchi
National Institute of Information and Communications Technology

第１７回ＡＡＭＴ長尾賞受賞記念

ニュースを対象とした日英機械翻訳システムの研究開発

後藤　功雄
日本放送協会

1. はじめに

NHK では外国人に向けた英語による迅速な情報発信を支援するために、ニュースを対象とした日英機械翻訳の研究開発を進めている。この度、研究開発したシステムが第 17 回 AAMT 長尾賞を受賞するという栄誉にあずかり、深く御礼を申し上げる。本稿では NHK での機械翻訳システムの利用、英語ニュースの特徴、NHK における機械翻訳システムの研究開発を紹介する。

2. NHK での機械翻訳システムの利用

NHK 日英機械翻訳開発プロジェクトで研究開発した日英機械翻訳システムは、国際放送向けの英語ニュースの制作支援と、災害時の特設ニュースのインターネットライブ配信への英語字幕付与に利用されている。それぞれの利用について説明する。

国際放送向けの英語ニュースは次の流れで制作している。(1)まずライターと呼ばれるニュース翻訳者が日本語原稿から英語原稿を作成する。固有名詞の訳の確認や適切な用語や表現の選定などが必要となり、一番時間が必要な行程である。 (2)次に、英語ニュースを監修するデスクが英語原稿の内容を確認する。正確な英語になっているか、ネイティブスピーカーにも分かりやすいニュース構成になっているかを確認する。(3)そして、文法、言い回し、ニュアンスなどをリライターと呼ばれる英語ネイティブのジャーナリストが修正し、(4)最後にもう一度デスクが確認し [1]、英語原稿が完成する。英語ニュース制作支援では、主に(1)の行程で下訳の作成やフレーズなど短い表現の参照で機械翻訳システムを利用して制作時間の短縮に取り組んでいる。特に夜間や休日での災害時など、人手が少ない中でも英語で多くの情報を迅速に発信する必要があるときに重要な役割を果たしている[1]。

2022 年 6 月から、災害時の総合テレビ特設ニュースに英語字幕を付与したインターネットのライブ配信を開始した。総合テレビの特設ニュースの日本語字幕を機械翻訳システムで英語に翻訳し、その結果を英語字幕として付与してインターネットで配信する[2]。このサービスは在留・訪日外国人に向けたもので、震度 5 弱以上の地震発生時、津波注意報・津波警報・大津波警報・大雨特別警報などの発表時に原則実施する。サービス提供時には、NHK ワールド JAPAN のサイトやアプリから、誘導バナーを通してアクセスできる。ライブ配信ページには、AI 翻訳のため、正確な表現ではない場合もある「おことわり」を掲載している。

表 1：　英語ニュースの書き方のポイント

1	文は短く。一文は一要素を原則に。特にリード。複数の要素がある場合は、文を分ける。
2	SVO の構文が基本
3	能動態を使い、受動態はなるべく避ける
4	繰り返しを避ける
5	なるべく人を主語に
6	修飾語は避ける
7	ソースを明確に（誰が言ったのか明示する）

[1] リライターの修正とデスクの確認が複数回繰り返されることもある。

Research and Development of Japanese–English Machine Translation System for News
Isao Goto
Japan Broadcasting Corporation (NHK)

3. 英語ニュースの特徴

英語ニュース原稿は日本語ニュース原稿の内容を元に制作される。英語ニュースは日本語ニュース原稿の直訳ではなく、英語ニュースに独特の書き方やスタイルが存在する[3–9]。NHK の英語ニュース制作で使われるポイントを表 1 に示す。

日本語ニュースと英語ニュースの例を図 1 に示す。対応していない部分があったり、文の順番が変わっているところもある。英語ニュースの書き方やスタイル[3–9]のうち、図 1 の例に見られる違いの理由として考えられるもののいくつかを示す。

① 逆ピラミッド型[10]になっている
ニュースはリード（記事の導入部分）[2]と本文（リード以降の文）からなり、リードではこのニュースの最も重要なことを伝え、リードに続く本文はリードの内容を詳しく伝える。

② 短く、シンプルに書く
同じ情報の重複がない（冗長でない）。詳細は重要度により省略することもある。

③ 同じ表現の繰り返しを避ける

ここで、①～③のそれぞれについて、NHK の日本語ニュースとの違いを図 1 の例を用いて確認する。

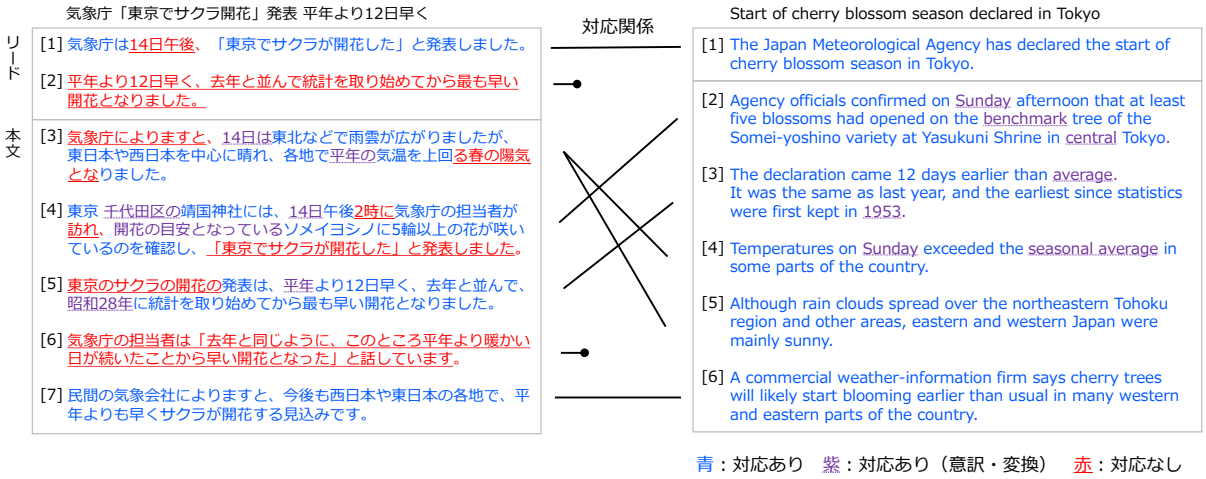
① NHK の日本語ニュースでは、リードはニュース全体の要約であることが多い。そして本文は背景の説明から始まることもある。例えば、図 1 (a)、(b)では、日本語のリードはニュース全体の要約になっているのに対して、英語のリードはニュースで最も重要なことだけを伝えている。また、(a)では、日本語ニュースではリードの次に背景情報である気温を説明しているが、英語ニュースでは、リードの次はリードで伝えた開花の根拠となった出来事を説明している。これによって、文の順番に違いが生じている。

② 前記の通り、NHK の日本語ニュースでは、リードがニュースの要約になっていることが多く、その場合に本文でリードの内容が再度述べられる。例えば図 1(a)のリードの 1 文目の内容は本文の 4 文目でも述べられており、リードの 2 文目の内容は本文の 5 文目でも述べられている。英語側では、重複する内容のうち、一方だけに英語が対応しており、英語ニュースでは内容が重複していないことが分かる。また、詳細な情報を省略することでも英語ニュースが短くなっている。例えば、(a)では 6 文目の内容、(b)では 7 文目の後半の内容などが省略されている。

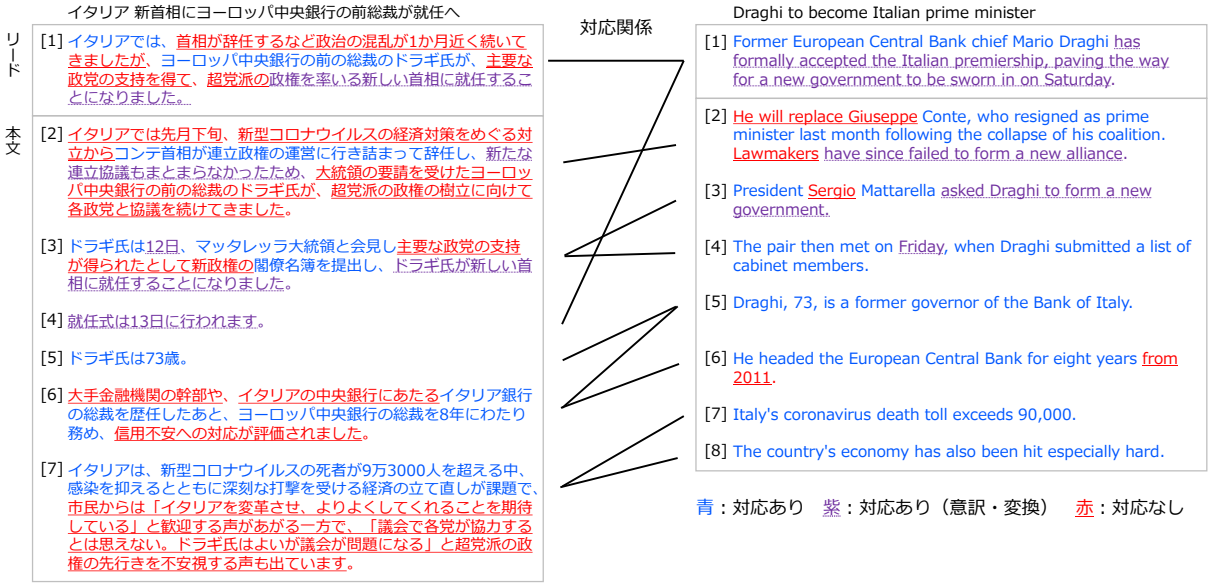
③ 日本語ニュースでは同じ表現の繰り返しがしばしば出現する。例えば、(a)では、「開花」という表現が 8 回出現している。英語ニュースでは、「開花」に相当する表現として、”the start of cherry blossom season”, “blossoms had opened”, “start blooming”という 3 種類の表現が 1 回ずつ使われており、同じ表現の繰り返しが避けられている。また、(b)では、日本語ニュースに「ドラギ氏」という表現が 5 回出現しており、英語原稿では、次の順番でドラギ氏を指す表現が出現している。”Mario Draghi”, “He”, “Draghi”, “Draghi”, “Draghi”, “He”。フルネーム、ファミリーネーム、代名詞を使って同じ表現の連続する繰り返しが低減されている。

英語ニュースでは、ニュース制作時から 1 週間未満の日付であれば曜日で、それ以外は日付には月を伴い表現する。また、日本語ニュースでは風速をメートル／秒で表すが、英語ニュースではキロメートル／時で表す。さらに、日本のニュースの場合に外国人にとってニュースの理解に必要な日本の背景知識が追加される場合もある。例えば、「震度 6 強」は”an intensity of upper 6 on the Japanese scale of zero to 7”と震度の説明が補足されたり、「青森県」は”Aomori Prefecture in northern Japan”と場所が補足されたりする。

[2] ニュース制作者により明示されることが多い。



(a)



(b)

図 1：NHK 日本語ニュースと英語ニュースの対応関係の例

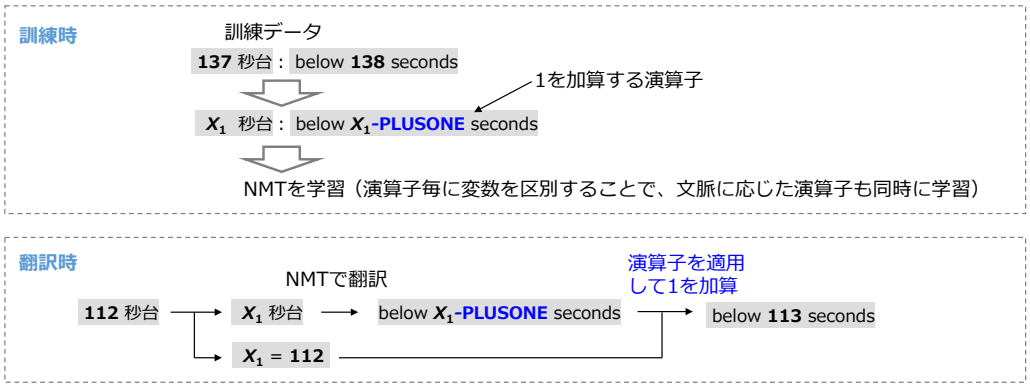


図 2：目的言語文脈に依存する値の変化への対応

4. 機械翻訳システムの研究開発

NHK では 1986 年から機械翻訳の研究開発に取り組んでいる[11]。NHK 日英機械翻訳開発プロジェクト[3]は、2017 年から開始しており、ニュースを対象とした日英機械翻訳の研究開発と利用を進めている。

近年、ニューラル機械翻訳（NMT）によって翻訳品質が大幅に向上し、本プロジェクトでも NMT を活用することで機械翻訳の品質を高めている。英語ニュースには前節で述べた特徴があるなど、ニュースを対象とした日英機械翻訳の開発にはさまざまな課題があり、NMT の技術だけで解決できるわけではない。以下、これまでに取り組んだ課題と対策について説明する。

- 対訳として高品質なデータの欠如

前節で説明したように、日英のニュース記事対は文レベルではあまり対訳になっていない部分も多い。このため、日英ニュース記事対から文対応を推定して抽出した文対は、対訳としての品質は高くなく、訳抜けや過剰訳が多く含まれるデータとなってしまう。この状況は新聞や通信社など他の報道記事でも日英記事では似た傾向がある。日英の報道記事対から抽出した対訳文対を用いて日英翻訳を学習すると、訓練データ中の訳抜けを学習してしまい、翻訳時に訳抜けが頻出する。この原因は訓練データの対訳としての質にある。そこで、我々は高品質な NHK ニュースの対訳を 100 万文対構築した。構築方法は、NHK NEWS WEB の日本語ニュースを人手で英語に翻訳する方法、機械翻訳で英語に翻訳してから人手でポストエディットする方法、NHK ワールド JAPAN の英語ニュースを機械翻訳で日本語に翻訳してから人手でポストエディットする方法の併用である。

このほか、NHK ワールド JAPAN の英語ニュースの機械翻訳による逆翻訳のデータや既存の対訳データなど複数の種類のデータを訓練データとして用いる。NMT の学習では英語ニュース文・人手翻訳文、日本語ニュース文・人手翻訳文・機械翻訳文など、複数の特徴を持つ各データに各特徴に対応するタグを付与して学習する。翻訳時には目標出力の特徴を表すタグを特徴の数だけ付与することで、出力の特徴を制御する[12]。

- 数字・日付表現

ニュースにおいて数字は重要であり、正確な翻訳が求められる。しかし、例えば「1 億 6000 万」は”160 million”で数の単位が異なることから、単語列として翻訳すると不正確な翻訳になる場合がある。正確に翻訳するために、大きな数字や小数点を含む数字は変数に置換して翻訳し、日本語数字表現（例：1 億 6000 万）はルールで認識して値（例：160000000）に変換し、値からルールで英語表現（例：160 million）を生成して後処理で英文中の変数部分に挿入する。

また英語側の文脈によって数字を変更する必要がある場合がある。例えば，「112 秒台」が“below 113 seconds”となる場合，数字を変更する必要がある。このように英語側の文脈に応じて数字を変更できるように、変数として演算操作毎に細分化したものを用いて[13]、後処理で変数を数字に置換する際に演算を適用する（図 2）。上記例の目的言語側で 1 を加算する演算子として PLUSONE を設定する。学習時に、訓練データの対訳文で同じ値同士を対応づけて変数に置換した後に、日本語側より英語側の値が 1 大きい数字を対応づけてこれらのペアも変数に置換し、この時に目的言語側の変数名に演算子を追加する。このデータで翻訳を学習することによって、文脈に応じた演算子も同時に学習する。翻訳時は、NMT の出力に演算子付きの変数が含まれていた場合に、対応する値に演算を適用してから変数部分に挿入する。

日本語ニュースでは風速はメートル／秒で表すが、英語ニュースではキロメートル／時で表すという違いを反映するため、風速は単位の変換を計算する。このほか、定型的な表現で例えば「107 円 53 銭から 55 銭」は英語では“107.53 to 107.55 yen”というように「55 銭」の部分は“0.55 yen”ではなく“107.55 yen”に訳出しなければならないが、こういった表現に対して

[3] NHK の放送技術研究所、国際放送局、技術局、報道局のメンバーにより構成

適切に値を認識するルールを整備した。

前節で述べた日付表現の違いについては、前処理で日本語ニュースの日付表現を曜日に変換、または月を挿入する。ここで、「1 日」といった表現は日にちを表す場合と期間を表す場合がある。期間を表す表現を前処理で変換してしまうと意味が変わって正しく翻訳できない。そこで、日本語側では表現が同じだが、英語側では意味の違いに応じて表現が異なることを利用して、対訳データを用いて日本語の日付表現の日付と期間を自動分類する。その結果から日本語表現の日付と期間の区別を学習し、入力文の日付表現の日付と期間を識別する[14]。これによって適切に前処理を実施できる。

- ユーザー辞書

新しい固有名詞や用語などの翻訳知識は学習していないため機械翻訳で翻訳できない。そこで、ユーザー辞書の機能を機械翻訳システムに追加した。辞書機能として、相補関係にある 2 つの方法を組み合わせて利用する。1 つ目は変数に置き換えて翻訳した後に出力中の変数部分を辞書の訳に置き換える方法[15]である。この方法は学習データに出現していない表現にも対応できる。2 つ目は入力に訳を挿入することで出力を制御する方法[16]である。訓練データの対訳文で辞書の対訳表現が出現する場合に、辞書の訳を訓練データの入力文に挿入して NMT を学習し、翻訳時に入力文に辞書に登録されている表現が含まれていればその訳を入力文に挿入して出力をガイドする。例えば、辞書に「富士山：Mt. Fuiji」という項目が登録されていて、入力文が「富士山に登る」の場合、入力文は「<t>富士山<d/>Mt. Fuji</t>に登る」のようになる。この方法は翻訳で語彙の情報を活用できるという特徴があり、訓練データで対訳の出現頻度が閾値以上の場合に利用する。それ以外および辞書登録後にまだ上記の学習をしていない場合は、1 つ目の方法を利用する。

- 文脈の利用

日本語では、主語が自明な場合では主語を省略することが多い。日本語ニュースでも主語はしばしば省略される。一方で、英語ニュースでは主語を明記する。このため、翻訳時に主語の情報を文脈から取得する必要がある。文脈を利用する手法としては、入力文の前文を入力文に追加する 2-to-1 と呼ばれる手法[17]があるが、長いニュース文の中から必要な情報を活用することは難しい。また、必要な情報が前文より以前の文にある場合もある。そこで、ニュース記事の文を述語項構造解析して、前文脈中で主語・主題を含む文のうち入力文に一番近い文の主語・主題を文脈情報として入力文に追加する。こうして文脈から翻訳に必要な文脈情報を選択して文脈を翻訳に利用する[18]。

また、前章で述べたように英語ニュースでは同じ表現の繰り返しを避けることが望ましい。また目的言語側で一貫性が必要な場合もある。このためには目的言語側の文脈を考慮する必要がある。目的言語側の文脈を利用する NMT として、直前の目的言語文を入力文に追加して利用する方法がある。既存手法では、学習時に追加する目的言語文として訓練データの目的言語文が使われている[19]が、翻訳時には直前の文の機械翻訳の出力を目的言語文脈として利用する。学習時に用いる訓練データの目的言語文と翻訳時に用いる機械翻訳の出力文では、訳質や translationese の有無といった違いがある。学習時と翻訳時でのこれらの違いが、翻訳品質に悪影響を及ぼす。そこで、学習時に訓練データの目的言語文と機械翻訳の出力の両方を文脈として用い、最初は学習しやすい訓練データの目的言語文の割合を多くし、次第に機械翻訳の出力の割合を多くする。これにより、学習時と翻訳時の文脈の特徴の違いによる悪影響を低減する[20]。

- 日英ニュース記事の翻訳知識の活用

日々制作される日英のニュースに含まれる翻訳知識を機械翻訳が自動で学習することが望ましい。しかし本節で述べたように、これらのデータから抽出できる対訳文対は、3 節で述べたような相違がある。そのため、このようなデータで学習すると、訳抜けも学習してしまう。そこで、訓練データ中の内容語で訳抜けしている部分と訳が存在する部分を推定する。そして、

これらの区別を表すラベルを原言語文の各単語に付与することでこの区別を NMT で学習する。翻訳時は訳が存在することを表すラベルを入力文の各単語に付与することにより、訓練データで訳が存在する部分から学習した知識を主に使って出力を生成する。これにより、訳抜けを含む訓練データで学習した場合に、翻訳時の訳抜けを低減させる[21]。

- 制作支援

ニュースは内容が正確であることが重要である。そのため、下訳に翻訳システムを使う場合に、翻訳が正しくない部分があれば修正が必要となる。翻訳誤りの中でも訳抜け箇所は目立たず探しにくい。この検出のため、出力文から強制的に入力文を生成する際の確率を利用して訳抜け部分を推定する[22]。

5. おわりに

本稿では、NHK での機械翻訳システムの利用、英語ニュースの特徴、機械翻訳の研究開発を紹介した。これまでの取り組みにより、ニュースの情報発信において機械翻訳システムが活用されるようになってきた。

現在の機械翻訳システムは入力文の翻訳を出力する。この翻訳機能についても新しい翻訳知識の自動獲得などの課題がある。さらに、英語ニュースの特徴で述べたとおり、英語ニュースは日本語ニュースの単なる翻訳ではないため、日本語ニュースからの英語ニュースの生成は、翻訳の範囲を超えた技術が必要となる。今後はこれらの課題にも取り組んでいく。

謝辞

本研究成果の一部は、国立研究開発法人情報通信研究機構の委託研究（課題 197 および課題 225）により得られたものです。

参考文献

[1] 後藤 功雄, NHK 技研 R&D 2022 年 春号 解説 02，機械翻訳技術の研究動向.

[2] NHK 広報局，災害時の情報をＡＩ英語字幕でより早く～在留・訪日外国人向けサービス拡充～，報道資料，https://www3.nhk.or.jp/nhkworld/upld/thumbnails/ja/information/info_ainews_disaster_20220725.pdf

[3] Robert A. Papper, Broadcast News and Writing Stylebook, seventh edition, Routledge, 2020.

[4] Mervin Block, Broadcast Newswriting: The RTDNA Reference Guide, A Manual for Professionals, CQ Press, 2011.

[5] Mervin Block, Top Tips of the Trade, https://mervinblock.com/top-tips-of-the-trade/

[6] David Ingram and Peter Henshall, The News Manual, https://www.thenewsmanual.net

[7] The Associated Press, The Associated Press Stylebook and Briefing on Media Law 2018, Basic Books, 2018.

[8] Rene J. Cappon, The Associated Press Guide to News Writing, 4th Edition, Peterson's, 2019.

[9] William Strunk and E. White, The Elements of Style, Fourth Edition, Longman, 1999.

[10] Inverted pyramid (journalism), Wikipedia, https://en.wikipedia.org/wiki/Inverted_pyramid_(journalism)

[11] 後藤 功雄, NHK 技研 R&D 2018 年 3 月号 解説 02，機械翻訳技術の研究と動向.

[12] Hideya Mino, Hideki Tanaka, Hitoshi Ito, Isao Goto, Ichiro Yamada, and Takenobu Tokunaga. Content-Equivalent Translated Parallel News Corpus and Extension of Domain Adaptation for NMT. In Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pages 3616–3622.

[13] 村上 聡一朗, 渡邉 亮彦, 宮澤 彬, 五島 圭一, 柳瀬 利彦, 高村 大也, 宮尾 祐介. 時系列株価データからの市況コメントの自動生成, 自然言語処理, 2020, 27 巻, 2 号, p. 299-328.

[14] Kazutaka Kinugawa, Hideya Mino, Isao Goto, Ichiro Yamada, Leveraging a Bilingual Corpus to Resolve Date–Duration Ambiguity in Japanese Numeric Day Expressions, Journal of Natural Language Processing, 2022, Vol. 29, No. 2, p. 638-668.

[15] Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. Training Neural Machine Translation to Apply Terminology Constraints. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pages 3063–3068.

[16] Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural Machine Translation. In Proceedings of the 3rd Workshop on Asian Translation, 2016, pages 47–57.

[17] Jörg Tiedemann and Yves Scherrer. Neural Machine Translation with Extended Context. In Proceedings of the Third Workshop on Discourse in Machine Translation, 2017, pages 82–92.

[18] Isao Goto, Hideya Mino, Hitoshi Ito, Kazutaka Kinugawa, Ichiro Yamada, and Hideki Tanaka. Neural Machine Translation Using Extracted Context Based on Deep Analysis for the Japanese-English Newswire Task at WAT 2020. In Proceedings of the 7th Workshop on Asian Translation, 2020, pages 72–79.

[19] Yunsu Kim, Duc Thanh Tran, and Hermann Ney. When and Why is Document-level Context Useful in Neural Machine Translation?. In Proceedings of the Fourth Workshop on Discourse in Machine Translation, 2019, pages 24–34.

[20] Hideya Mino, Hitoshi Ito, Isao Goto, Ichiro Yamada, and Takenobu Tokunaga. Effective Use of Target-side Context for Neural Machine Translation. In Proceedings of the 28th International Conference on Computational Linguistics, 2020, pages 4483–4494.

[21] 後藤功雄，美野秀弥，山田 一郎，訳抜けを含む訓練データと訳抜けのない出力とのギャップを埋めるニューラル機械翻訳，言語処理学会第 26 回年次大会，2020.

[22] 後藤 功雄, 田中 英輝. ニューラル機械翻訳での訳抜けした内容の検出, 自然言語処理, 2018. 25 巻, 5 号, p. 577-597.

第 9 回 AAMT 長尾賞学生奨励賞受賞記念

JParaCrawl v3.0: 大規模日英対訳コーパス

森下　睦
NTT コミュニケーション科学基礎研究所

1. はじめに

現在のニューラル機械翻訳モデルは、主に対訳コーパスを用いた教師ありの手法 [1, 2, 3, 4] で学習されている。学習時の対訳コーパスの質と量が翻訳精度に大きな影響を与えること知られているが、一般に公開されている対訳コーパスは多くの言語対で限られている。例えば、独英などの言語対ではすでに数億文の対訳文が公開されているものの、日英ではまだ同程度のものは存在せず、翻訳モデル学習時に大きな問題となっている。そのため、本稿ではさらに大規模なウェブベースの日英対訳コーパスを構築する。現在、日英で最大規模の対訳コーパスの 1 つは約 1000 万文の対訳文を含む JParaCrawl v2.0 [5] であり、これはウェブを大規模にクロールし対訳文を自動的に抽出することで構築されている。本コーパスは欧州言語対と比較すると小規模であり、2020 年を最後に更新が止まっているため、最新の情報を含んでいない。そのため、本研究ではウェブを全面的に再クロールし、対訳文を抽出することで JParaCrawl コーパスを拡大/更新する。本研究では、対訳文抽出手法を機械翻訳器を用いた手法へと変更し、新たに PDF や Word 文書も収集対象とすることで、対訳抽出数の向上を狙う。また、新たに作成した対訳コーパスを用いて、英日および日英の機械翻訳の精度がどのように向上するかを実験的に示す。本研究で作成した対訳コーパスは JParaCrawl v3.0 と名付け、今後の研究のために我々のウェブサイト [1]で公開する。本稿の貢献は以下のようにまとめられる [2]。

- 従来の JParaCrawl コーパスと合わせて 2100 万文対以上を含む大規模な日英対訳コーパスを構築した。
- 新たな対訳コーパスにより、幅広い分野で英日・日英機械翻訳の精度が向上することを実験的に確認した。
- 本コーパスを研究目的利用に限り無償で公開する。

2. 関連研究

対訳コーパスは様々な文書から対訳文を抽出することで作成されることが多い。代表的なものとして、国際機関が作成した対訳文書がある。例えば、欧州議会の議事録から作成された Europarl [6]、国連の翻訳文書から作成された UN 対訳コーパス [7] などがある。これらの文書は、通常プロの翻訳者が翻訳しており、文書 ID などのメタデータを持っていることもあるので、容易に対訳文を抽出することができる。しかし、これらの一般に公開されている対訳文書は多くない。

近年では、ウェブから対訳文を抽出する手法も多く提案されている。ウェブ上には 2 言語以上で書かれたウェブサイトが多数存在し、こういったウェブサイトから対訳文を抽出する。ウェブ上には、様々な言語や分野の対訳文が存在しており、大規模な対訳コーパスを作成するための有望な情報源である。ウェブから対訳文を抽出する初期の研究としては、大規模な分散システムを構築し対訳文を抽出したもの [8]、Common Crawl から対訳文を抽出したもの [9] などがある。最近では、多言語文埋め込みを用いた対訳文対応手法を用いて、Wikipedia や Common Crawl から大規模な多言語対訳コーパスを作成する研究 [10, 11] も報告

[1] http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/
[2] 本稿は言語処理学会第 28 回年次大会の発表論文「JParaCrawl v3.0: 大規模日英対訳コーパス」©森下睦, 帖佐克己, 鈴木潤, 永田昌明（CC BY 4.0）を基に一部改訂したものである。

JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus
Makoto Morishita
NTT Communication Science Laboratories

されている。

また、ParaCrawl プロジェクトはヨーロッパ言語の大規模な対訳コーパスをウェブから継続的に作成している [12]。我々は以前この活動にヒントを得て、大規模な対訳コーパスが存在しない日英向けの大規模な対訳コーパスを作成した [5]。このコーパスは JParaCrawl と名付けられ、1,000 万文を超える対訳文を含む日英における最大規模の対訳コーパスとなっている。しかし、JParaCrawl コーパスは、独英などの欧州言語対と比較するとまだ小規模であり、これをもとにした翻訳モデルの精度も欧州言語対と比較すると低精度である。ゆえに、さらに大きな日英対訳コーパスの作成が求められている。本研究では、ウェブを新たにクロールし対訳文をさらに抽出することで、JParaCrawl コーパスをさらに拡張することを目指す。

3. JParaCrawl v3.0

本研究では、さらにウェブをクロールして対訳文を抽出し、JParaCrawl v2.0 コーパスを拡張すること目指す。我々の手法は、以前の ParaCrawl および JParaCrawl プロジェクトに基づくものである。以下の節でその詳細な手順を述べる。

3.1. 対訳文を含むウェブサイトの発見

本研究では、ウェブから対訳文を抽出することで大規模な対訳コーパスを構築する。まず、CommonCrawl 上のテキストデータを CLD2 によって解析し、各ドメインの言語別データ量を得る。その後、英語と日本語が同量程度含まれるウェブサイトには対訳文が存在する可能性があるという仮説に基づき、クロール対象ウェブサイトを列挙する。本研究では、2019 年 3 月から 2021 年 8 月までに公開された Common Crawl のテキストアーカイブデータを分析対象とし、ウェブサイトの規模が大きく、英語と日本語の文章が同程度である 10 万件のウェブサイトを列挙した。2019 年 3 月以前に公開されたデータについては、JParaCrawl v2.0 作成時に既に分析済みであり、本収集の目標の一つである最新の情報を含むことをとも乖離しているため除外した。本手順で列挙されたウェブサイト一覧を確認したところ、前回の JParaCrawl v2.0 時に候補に挙がっていないウェブサイトが 7 割を占めていた。なお、本手順には ParaCrawl プロジェクトが提供する extractor を使用した。

3.2. ウェブサイトのクロール

次に、前節で列挙されたウェブサイト全体をクロールする。本研究では、 Heritrix を用いて、各ウェブサイトに対して最大 48 時間のクローリングを行った。これまでの JParaCrawl では、プレーンテキストのみを対象としていたが、今回はこれに加えて、PDF や Word 文書もクロールの対象とした。日本国内の官公庁や企業は PDF で文書を発信することがあり、これらも対訳抽出の対象とすることで対訳文数の増加に寄与すると考えられる。

3.3. 対訳文抽出

次に、クロールされたウェブサイトから対訳文を抽出する。本手順には、ParaCrawl プロジェクトが提供する Bitextor を日本語に対応させ使用した。対訳文書と対訳文の対応付けには、機械翻訳を用いた対応付けツール bleualign [13] を使用した。このツールでは、まず機械翻訳を用いて日本語文を英文に翻訳し、BLEU スコアを最大化する日英の文ペアを発見する。この際、日英翻訳には JParaCrawl v2.0 で学習した Transformer ベースのニューラル機械翻訳（NMT）モデルを使用した。なお、以前の JParaCrawl で使用した辞書ベースの手法 [14] よりも bleualign の方が高精度であることを予備実験により確認した。

3.4. ノイズ除去

最後の手順として、正しく対応付けられていない、翻訳が不正確など、学習時のノイズとなる文対をフィルタリングする。本手順には、Bicleaner [15] を使用し

た。ノイズ除去の後、クリーンな対訳文と JParaCrawl v2.0 を結合、重複文を削除した。以上の手順により、2100 万文以上を含む新しい大規模な JParaCrawl v3.0 を作成した。これは、以前の JParaCrawl v2.0 の倍以上の文数である。

表 1 に、これまでの JParaCrawl および今回作成した v3.0 の重複を削除した対訳文数および英語側単語数を示す。なお、これまでの JParaCrawl についても重複削除を行った文数を報告しているため、以前の論文で報告されている値とは異なることに注意されたい。PDF なども対訳抽出対象としたことにより、HTML のみを対象とする場合と比較して、約 10%対訳文抽出数が向上した。

バージョン	文数	単語数
v1.0	4,817,172	125,216,523
v2.0	8,809,771	234,393,978
v3.0	21,481,513	502,445,763

表 1 JParaCrawl コーパスに含まれる重複を取り除いた対訳文数および英語側単語数

4. 実験

本節では、新たに作成した JParaCrawl v3.0 の翻訳精度への影響を確認するために、NMT モデルを学習し様々なテストセットでその精度を評価した。以降では、使用したテストセットの詳細およびモデル学習時の設定について述べる。

4.1. 実験設定

4.1.1. テストセット

様々な分野で NMT モデルの精度を評価するために、15 種のテストセットでモデルを評価する。これらには、以前の JParaCrawl 発表時に報告した ASPEC [16] (科学技術論文), JESC [17] (映画字幕), KFTT [18] (Wikipedia 記事), TED [19] (講演) が含まれる。

さらに本実験では、会話文中心の対訳コーパスである Business Scene Dialogue コーパス (BSD) [20] や機械翻訳シェアードタスク WMT, IWSLT [21, 22, 23, 24, 25] のテストセットを追加した。これらには、ニュース記事、SNS 上のテキスト、Wikipedia 上のコメントなどが含まれる。なおシェアードタスク用テストセットの中には、特定の翻訳方向（英→日など）で使用することを前提としたものもあるが、参考までに英日、日英の両方向で評価した。

4.1.2. 学習設定

前処理として、対訳コーパスを sentencepiece [26] を用いてサブワード単位に分割した。この際、語彙数は 32,000 とした。翻訳モデルの学習には fairseq [27] を用い、small、base、big の 3 つの異なる大きさの Transformer モデルを学習した。以前の JParaCrawl の報告に基づき、TED (tst2015) では small モデルを、KFTT では base モデルを、その他のテストセットでは big モデルを使用した。なお、学習に関する設定は、公平な比較のために以前の JParaCrawl の報告とほぼ同じであるが、v3.0 モデルについては、対訳コーパスの大きさが原因で 24,000 ステップでは収束せず、更新回数を変更した。評価には sacreBLEU [28] を用い、BLEU スコア [29] を報告する。なお、以前の実験との整合性を保つために、すべてのテストセットを NFKC 正規化した。

4.2. 実験結果

表 2 に様々なテストセットにおける BLEU スコアを示す。JParaCrawl v3.0 で学習したモデルは、日英ではすべてのテストセットで、英日では 15 のうち 13 のテストセットで以前の JParaCrawl を上回る精度を達成した。この結果は、科学技術論文、ニュース、対話などの様々な分野において、新しい JParaCrawl が NMT モデルの精度を押し上げることを示している。特に、WMT21 ニュース翻訳タスクでは、JParaCrawl v3.0 モデルが大幅に精度を向上させることがわかった。これは、前回の JParaCrawl v2.0 が 2019 年のウェブデー

タをもとに作成されており、2021 年のニュースに頻出する単語が正しく翻訳できていないことが原因だと推測される。例えば、2021 年のニュース記事では、2019 年には存在しない COVID-19 に関連する単語が頻出する。こういった最新の単語を正しく翻訳できるようにするためには、対訳コーパスを常に更新し続けるなどの対応が必要だと考えられる。

4.3. 翻訳例

図 1 に JParaCrawl v1.0、v2.0、v3.0 で学習したモデルの翻訳例を示す。この例は、WMT21 ニュース翻訳テストセットから COVID-19 に関連するものを選んだ。この入力文には、「濃厚接触者」という日本語が含まれており、これは “close contacts” と翻訳されるべきである。しかし、v1.0 および v2.0 で学習したモデルは “strong contact person”と誤訳している。一方で、v3.0 で学習したモデルでは、これを正しく “close contacts”と翻訳できている。テストセットを目視で確認したところ、この例のように COVID-19 に関連する記事で多くの改善が見られた。この結果は、前節で述べた近年頻出する用語を正しく翻訳できているという仮説を支持するものである。

5. おわりに

本研究では、これまでの大規模日英対訳コーパス JParaCrawl をさらに拡張した JParaCrawl v3.0 を作成した。本対訳コーパスは、最新の CommonCrawl アーカイブを分析して対訳文が存在すると思われるウェブサイトを発見し、それらから対訳文抽出を行うことで作成した。新たな JParaCrawl v3.0 は 2100 万以上の対訳文を含んでおり、これは JParaCrawl v2.0 の倍以上の大きさである。本対訳コーパスを用いることで、様々な分野、特に最新のニュース記事の翻訳精度が向上することを実験的に確認した。今後の課題としては、継続的な JParaCrawl の更新や、より優れた対訳文抽出/フィルタリング手法の提案などが挙げられる。なお、本研究で作成した JParaCrawl v3.0 は我々のウェブサイトで研究目的に限り無償で公開している。

謝辞

本稿は第 9 回 AAMT 長尾賞学生奨励賞受賞記念として執筆したものです。このような大変名誉ある賞を授与していただき誠にありがとうございました。選考に関わってくださった委員の皆様に感謝致します。
また本稿は東北大学へ提出した博士論文 “Expanding the Applicability of Machine Translation” を一部抜粋し、改訂したものです。東北大学で私を熱心に指導してくださった鈴木 潤教授に感謝致します。

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS), pp. 3104–3112, 2014.
[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.
[3] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1412–1421, 2015.
[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS), pp. 6000–6010, 2017.
[5] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), pp. 3603–3609, 2020.
[6] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In Proceedings of the Machine Translation Summit X, pp. 79–86, 2005.
[7] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The united nations parallel corpus v1.0. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), pp. 3530–3534, 2016.
[8] Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. Large scale parallel document mining for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING), pp. 1101–1109, 2010.
[9] Jason R Smith, Herve Saint-Amand, M Plamada, P Koehn, C Callison-Burch, and Adam Lopez. Dirt cheap web-scale

parallel text from the common crawl. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1374–1383, 2013.
[10] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. arXiv preprint arXiv:1907.05791, 2019.
[11] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. CCMatrix: Mining billions of high-quality parallel sentences on the web. arXiv preprint arXiv:1911.04944, 2019.
[12] Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel
corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4555–4567, 2020.
[13] Rico Sennrich and Martin Volk. Iterative, MT-based sentence alignment of parallel texts. In Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA), pp. 175–182, 2011.
[14] Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. In Proceedings of the Recent Advances in Natural Language Processing (RANLP), pp. 590–596, 2005.
[15] Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In Proceedings of the 3rd Conference on Machine
Translation (WMT), pp. 955–962, 2018.
[16] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), 2016.
[17] R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz. JESC: Japanese-English Subtitle Corpus. arXiv preprint arXiv:1710.10639, 2017.
[18] Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.
[19] Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT3: web inventory of transcribed and translated talks. In Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT), pp. 261–268, 2012.
[20] Matıss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the business conversation corpus. In Proceedings of the 6th Workshop on Asian Translation (WAT), pp. 54–61, 2019.
[21] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo,
Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In Proceedings of the 5th Conference on Machine Translation (WMT), pp. 1–55, 2020.
[22] Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 conference on machine translation (WMT21). In Proceedings of the 6th Conference on Machine Translation (WMT), pp. 1–93, 2021.
[23] Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir K. Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan M. Pino, and Hassan Sajjad. Findings of the first shared task on machine translation robustness. In Proceedings of the 4th Conference on Machine Translation (WMT), 2019.
[24] Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. Findings of the WMT 2020 shared task on machine translation robustness. In Proceedings of the 5th Conference on Machine Translation (WMT), pp. 76–91, 2020.
[25] Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT), pp. 1–29, 2021.
[26] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 66–71, 2018.
[27] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), pp. 48–53, 2019.
[28] Matt Post. A call for clarity in reporting BLEU scores. In Proceedings of the 3rd Conference on Machine Translation (WMT), pp. 186–191, 2018.
[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics
(ACL), pp. 311–318, 2002.

	英日翻訳			日英翻訳		
テストセット	v1.0	v2.0	v3.0	v1.0	v2.0	v3.0
APSEC	24.7	26.5	26.8	18.3	19.7	20.8
JESC	6.6	6.5	6.5	7.0	7.5	8.4
KFTT	17.1	18.9	18.1	13.7	16.2	17.0
TED (tst2015)	11.5	12.6	13.1	11.0	11.9	12.0
Business Scene Dialogue Corpus	12.4	13.5	13.9	17.4	19.6	19.9
WMT20 News En-Ja	20.7	21.9	23.5	21.3	23.3	23.9
WMT20 News Ja-En	20.1	22.8	23.5	19.2	21.0	21.9
WMT21 News En-Ja	21.1	21.8	25.0	21.9	23.1	24.3
WMT21 News Ja-En	19.6	21.5	22.4	18.1	20.7	21.3
WMT19 Robustness En-Ja (MTNT2019)	12.4	12.5	14.4	15.6	16.8	17.3
WMT19 Robustness Ja-En (MTNT2019)	11.5	12.3	12.8	16.0	17.2	17.7
WMT20 Robustness Set1 En-Ja	15.2	15.8	18.7	20.0	20.6	21.6
WMT20 Robustness Set2 En-Ja	12.7	13.0	14.8	16.4	17.4	17.9
WMT20 Robustness Set2 Ja-En	7.9	8.2	8.6	12.0	12.6	14.0
IWSLT21 Simultaneous Translation En-Ja Dev	12.5	13.3	14.5	12.9	14.3	14.5

表 2　JParaCrawl v1.0, v2.0, v3.0 で学習した翻訳モデルの BLEU スコア

入力文	院内に「濃厚接触者」はいませんが、接触者全員に PCR 検査を実施し、女性が関係した病棟などを閉鎖して徹底的に消毒するということです。
参照訳	There are no known “close contacts” in the hospital, but all contacts will be subjected to PCR tests, and the wards and other areas where the women had been will be closed and thoroughly disinfected.
JParaCrawl v1.0	There is no “strong contact person” in the hospital, but a PCR test will be conducted for all the contacts, and women will close the wards and thoroughly disinfect them.
JParaCrawl v2.0	Although there is no “strong contact person”' in the hospital, PCR tests will be performed on all contact persons, and the wards related to women will be closed and thoroughly disinfected.
JParaCrawl v3.0	There are no “close contacts” in the hospital, but PCR tests will be conducted for all contacts, and the wards related to women will be closed and thoroughly disinfected.

図 1　WMT21 News Ja-En テストセットの翻訳例

第 9 回 AAMT 長尾賞学生奨励賞受賞記念

学習時と推論時における入力データの分布の異なりを考慮したニューラル機械翻訳モデルの学習手法

美野 秀弥
日本放送協会

1. はじめに

機械翻訳は，ある言語で書かれた文（原言語文）を入力としてこれを異なる言語の文（目的言語文）に変換する技術であり，異なる言語を扱う国や地域の人々とのコミュニケーションや異なる言語で書かれた情報の理解などの効率化のための重要な技術として注目されている．近年は，深層学習を用いたニューラル機械翻訳により旧来の統計的機械翻訳と比較して翻訳精度が飛躍的に向上し，盛んに研究が行われている．

ニューラル機械翻訳は，Long short term memory (LSTM)や Gated recurrent unit (GRU)などの回帰型ニューラルネットワークを用いたモデル [1,2]や，注意機構に着目したトランスフォーマーモデル [3]などによって構成されている．どちらもエンコーダとデコーダの２つのサブネットワークを用いたエンコーダ･デコーダモデルと呼ばれる系列変換モデルである．エンコーダは，入力文の各トークン（単語などに分割された情報）を入力とし，入力文の情報を有する文ベクトルを中間表現として出力する．デコーダは，エンコーダが出力した文ベクトルを入力とし，翻訳結果を出力する．

ニューラル機械翻訳は，旧来の統計的機械翻訳と同様に対訳データに基づく手法であり，学習のための対訳データのコーパス（対訳コーパス）の構築が必須である．学習に用いる対訳データは，一般的には多ければ多いほど翻訳精度が高くなるため，ウェブなどから大量の対訳データを自動で収集したり，単言語データを機械翻訳して擬似的にデータを増やすデータ拡張などを用いたりして対訳コーパスを構築する手法がある．

しかし，自動で収集したり機械翻訳を用いたりして構築した対訳コーパスの中には，誤訳などを含んだ，翻訳品質の悪い対訳データが混在していることがある．誤訳などのノイズが混在している対訳データをそのまま用いて翻訳モデルを学習すると，ニューラル機械翻訳器の翻訳精度は低くなってしまう．

本研究では，この翻訳精度低下の原因が翻訳時と学習時におけるデータの特徴の違いによるものであると仮定した．例えば，翻訳時は翻訳元の入力データと翻訳先の出力データが綺麗に翻訳されたデータであるべきなのに対し，学習時は出力データにノイズが含まれたデータを用いている場合，翻訳品質の違いがデータの特徴の違いに該当する．また，学習時のデータにノイズを含まない綺麗なデータが使えたとしても，ジャンル（ドメイン）の違うデータを用いると翻訳精度が低下することがあり，この場合はデータ間のドメインの違いがデータの特徴の違いに該当する．

学習データと翻訳時の入出力データには，下記の 4 種類の違いが生じる可能性がある．

1) 翻訳対象（入力データ）
2) 翻訳結果（出力データ）
3) 翻訳対象と翻訳結果との関係
4) 外部情報（翻訳対象とは別の入力データ）

Methods for Training a Neural Machine Translation Model Considering the Different Distributions of Input Data during Training and Inference
Hideya Mino
NHK Science & Technology Research Laboratories

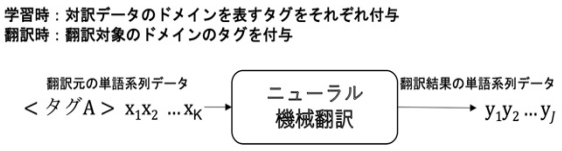


図 1: ドメインタグを用いたニューラル機械翻訳

1)の翻訳対象の特徴の違いは，ニュース文や特許文書など，翻訳対象の入力データの語彙，体裁、内容などの違いなどが該当する．2)の翻訳結果の特徴の違いは，「平易な語彙で翻訳する」,「専門用語などは固定訳を用いる」など翻訳結果の出力データの語彙や文法を制限する翻訳スタイルの違いが該当する．3)の翻訳対象と翻訳結果との関係の違いは,「翻訳対象のデータをコンパクトに翻訳する」,「詳細に翻訳する」など翻訳結果の語彙や文法を直接制限しない翻訳スタイルの違いが該当する．翻訳品質の違いも 3)に該当する．4)の外部情報は，（翻訳対象，翻訳結果）のペアである対訳データとは別に翻訳を補助する入力データのことである．外部情報には翻訳対象の周辺の文の情報や翻訳時に利用可能な辞書情報などが考えられ，これらの外部情報のスタイルの違いや品質の違いなどが該当する．

本研究では，上記４種類のデータの特徴の違いにより翻訳精度が低下する課題を持つ「ドメインタグを用いたニューラル機械翻訳」と「目的言語側の前文を用いた文脈情報考慮型ニューラル機械翻訳」の２つのニューラル機械翻訳のタスクにおいて新たな手法を提案し，翻訳実験により手法の有効性を示す．

2. 複数のドメインタグを用いたニューラル機械翻訳

まず，学習時と翻訳時との間の翻訳対象文の特徴の違いに着目した，ドメインタグを用いたニューラル機械翻訳のタスクを扱う．このタスクでは，1 章で定義したデータの特徴の中の 1), 2), 3)の違いに該当する．

ニューラル機械翻訳は，翻訳モデルの学習に対訳データを必要とする．しかし，対訳データの構築にはコストがかかることから，翻訳誤りなどのノイズを含むデータ（3)に該当）や翻訳対象とは違う分野のデータ（1)，2)に該当）など，実際に翻訳するデータの特徴とは違う特徴を有する対訳コーパスのデータを学習データとして利用することが多い．このようなドメインの違う対訳コーパスを混在して学習すると翻訳精度は低下することが知られている．

この問題を解決するために，対訳データの原言語側が有する特徴，目的言語側が有する特徴，および原言語側と目的言語側との対訳関係に関する特徴の３つを明示的に区別するドメインタグを用いたニューラル機械翻訳の新たな手法を提案する．ドメインタグを用いたニューラル機械翻訳は，翻訳対象のドメイン（インドメイン）のデータと翻訳対象ではないドメイン（アウトオブドメイン）の違いを考慮して学習することで翻訳精度を向上させる手法である．図１のように，コーパスの特徴を表すタグを原言語側のデータ（翻訳元の単語系列データ）に付与して学習，翻訳を行う．

本研究では，時事通信社のニュースを基とする特徴の違う下記の日英ニュースコーパス [4]を用いた．

- 日本語側のデータを１文ごと内容が等価になるように英訳したコーパス（内容等価コーパス）
- 日本語側と英語側のデータ同士を文単位で自動対応づけしたコーパス（自動アライメントコーパス）
- 目的言語である英語のデータを機械翻訳で日本語に翻訳したコーパス（逆翻訳コーパス）

このうち，内容等価コーパスを適切な翻訳と定義してこの対訳データをインドメインのデータとした．逆翻訳コーパスは，内容等価コーパスで学習した機械翻訳器を用いて作られるコーパスと自動アライメントコーパスで学習した機械翻訳器を用いて作られるコーパスの２種類に分けられる．

従来手法には，単一ドメインタグを付与する手法（コーパスタグ） [5]，逆翻訳であることを示すドメインタグを付与する手法（ノイズタグ） [6]，２種類のドメインタグを付与する手法（２つのタグ） [7]，があり，それぞれ表１に示すタグを付与して学習する．

ドメインタグ	内容等価コーパス	自動アライメントコーパス	逆翻訳コーパス（内容等価）	逆翻訳コーパス（自動アライメント）
コーパスタグ	<CE>	<AA>	<CE>	<AA>
ノイズタグ	<CE>	<AA>	<BT-CE>	<BT-AA>
２つのタグ	<CE>	<AA>	<BT>, <CE>	<BT>, <AA>
提案手法	<NS-S>, <CE-T>, <NO-BT>, <NO-AN>	<NS-S>, <NS-T>, <NO-BT>, <AN>	<CE-S>, <NS-T>, <BT>, <NO-AN>	<NS-S>, <NS-T>, <BT>, <AN>

表 1: 各手法がコーパスに付与するタグ

しかし，これらの従来手法では，本研究で用いた日英ニュースコーパスが持つ特徴を十分に区別することができない．例えば，表 1 において，内容等価コーパスで学習した機械翻訳器によって構築した「逆翻訳コーパス(内容等価)」に着目すると，コーパスタグでは内容が等価であることを示す<CE>タグを付与しているが，逆翻訳による機械翻訳特有のノイズが含まれているという特徴を反映できていない．ノイズタグでは，機械翻訳特有のノイズが含まれていることを示す<BT-CE>タグを付与しているが，1 つのタグだけでは逆翻訳に「内容同等コーパス」のドメインに合わせて翻訳する機械翻訳モデルを用いたという特徴が反映できない．２つのタグでは，<BT>と<CE>の２つのタグを付与することで，ノイズが含まれており，かつ，逆翻訳に「内容同等コーパス」のドメインに合わせて翻訳する機械翻訳モデルを用いたという特徴が反映されているが，目的言語側の英語文が時事通信社の元の英語文であるという特徴が反映できていない．

本研究では，従来手法の課題を解決するために，各コーパスの特徴を全て表現できるドメインタグを提案した．1 章で定義したデータの特徴の中の 1)，2)，3)の違いに着目し，下記の４つのタグを提案した．

- 原言語側のデータの特徴を表すタグ（翻訳対象）
- 目的言語側の特徴を表すタグ（翻訳結果）
- 逆翻訳による機械翻訳特有のノイズの有無を表すタグ（翻訳対象と翻訳結果との関係）

ドメインタグ	BLEU スコア	Δ
タグなし	20.36	-
コーパスタグ	22.41	+2.05
ノイズタグ	24.19	+3.83
２つのタグ	24.25	+3.89
提案手法	24.56	+4.20

表 2: 日英ニュース機械翻訳の実験結果

- 自動アライメントによるノイズの有無を表すタグ（翻訳対象と翻訳結果との関係）

表 1 の 4 行目に提案手法が各コーパスに付与するドメインタグを示す．４つのタグは，それぞれ，翻訳対象の特徴を示す<*-S>，翻訳結果の特徴を示す<*-T>，翻訳対象と翻訳結果との関係を示す<BT>または<NO-BT>，翻訳対象と翻訳結果との関係を示す<AN>または<NO-AN>に該当する．

提案手法の有効性を評価するために，本章で示した日英ニュースコーパスを用いた日英ニュース機械翻訳実験を行った．学習データには日英ニュースコーパスの合計 1,573,589 文対を用いた．テストセットには，内容等価コーパスからあらかじめ学習データとは別に取っておいた 2,000 文対を用いた．比較手法には，ドメインタグを付与しない手法と，ドメインタグを付与する３つの従来手法を用いた．翻訳性能の比較には，ランダムシードの異なる５つのモデルを学習してそれぞれのモデルが出力した翻訳結果の BLEU スコア [8]の中央値を用いた．

表 2 の実験結果より，ドメインタグを付与しない手法と比較してドメインタグを付与する手法の効果を確

認した．また，最も翻訳精度が高かったのは提案手法であり，提案手法の効果を確認した．

今後の課題として，翻訳精度の向上に寄与するドメインタグの選定手法を検討する必要があると考えている．本研究では，翻訳元の言語側（原言語側）の特徴，翻訳先の言語側（目的言語側）の特徴，および原言語側と目的言語側のデータ間の特徴に分けてドメインタグを付与したが，適切なドメインタグの付与は利用する対訳コーパスに依存すると考えられる．例えば，原言語側と目的言語側のデータ間の特徴の違いは，訳抜け，過剰訳，誤訳，意訳など，コーパスにより複数の特徴の組み合わせで区別できる場合がある．その場合，データ間の特徴の違いによって複数のタグを付与した方がよい可能性があり，他のデータセットを用いた実験を行う必要がある．また，翻訳品質の低いアウトオブドメインのデータセットを用いた実験を行い，アウトオブドメインのデータの翻訳品質と提案手法の効果との関連性を明らかにする必要がある．

3. 目的言語側の前文を用いた文脈情報考慮型ニューラル機械翻訳

次に，目的言語側の前文を用いた文脈情報考慮型ニューラル機械翻訳を扱う．１つ目のタスク「ドメインタグを用いたニューラル機械翻訳」と異なり，学習時と翻訳時との間の外部情報の特徴（1 章で定義したデータの特徴の中の 4)に該当）の違いに着目する．本研究が扱う外部情報は文脈情報である．

ニューラル機械翻訳は原言語一文を入力して目的言語への翻訳結果を出力するシステムが多いが，さらに翻訳精度を上げるために翻訳対象の周辺の文を文脈情報として利用する手法がある．しかし，従来手法では，文脈情報に原言語側の周辺の文を用いることで翻訳精度が向上するが，文脈情報に目的言語側の周辺の文を用いると翻訳精度が低下することが報告されている．目的言語側の文脈情報を用いた機械翻訳の従来手法では，以下のように学習，翻訳を行う．

【学習】

目的言語側の文脈情報の**参照訳**と原言語側の翻訳対象の文を入力とし，翻訳対象の文の参照訳を出力するように翻訳モデルを学習する．

【翻訳】

目的言語側の文脈情報の**機械翻訳結果**と原言語側の翻訳対象の文を学習済みの機械翻訳器に入力し，機械翻訳結果を得る．

すなわち，目的言語側の周辺の文として学習時は誤訳のない綺麗な正解訳を，翻訳時は機械翻訳誤りを含む機械翻訳結果を用いている．

学習時と翻訳時で特徴の異なるデータを用いることで生じるモデルの偏りは exposure bias [9]と呼ばれ，翻訳品質の低下に繋がることが報告されている[10]．特徴の異なる参照訳と機械翻訳結果を学習時と翻訳時で別々に用いる従来手法では，文脈情報を考慮した翻訳モデルが効果的に学習されずに精度低下を引き起こす可能性がある．

そこで，本研究では，目的言語側の周辺の文を用いる従来手法の問題が翻訳モデルの学習時と翻訳時で用いる周辺の文のデータの特徴の違いにあると仮定し，学習時と翻訳時で用いる目的言語側の文の特徴の違いによる影響を緩和するための学習データ制御方法を提案する．カリキュラム学習 [11]とスケジュールサンプリング [9] のアプローチを参考にし，下記の手順により，学習エポック（訓練データ全体の学習の単位）ごとに，参照訳を文脈情報として用いるデータの割合(サンプリングレート p）と機械翻訳結果を文脈情報として用いるデータの割合（サンプリングレート 1-p）を変更して学習データを再構成し，翻訳モデルを学習する．

(1) 1 エポック目の学習には，目的言語側の文脈情報が全て参照訳（サンプリングレート p=1）である学習データを用いる．

手法	ニュースコーパス		TED トークコーパス			
	EN-JA	JA-EN	EN-JA	JA-EN	EN-DE	DE-EN
Single-Sent.	41.99	24.23	15.69	8.48	24.04	28.46
Trg-Context GT	42.40	24.80	16.15	8.98	25.96	30.25
Trg-Context Pred.	42.45	24.31	16.27	9.15	25.98	30.23
提案手法	**42.79**	**24.86**	**16.37**	**9.66**	**26.01**	**30.60**

表 3: 実験結果

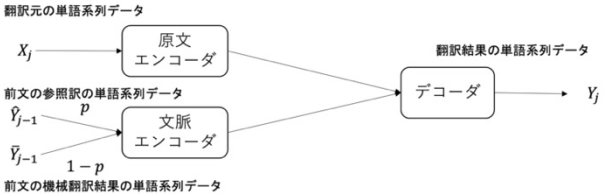


図 2: マルチエンコーダ文脈考慮型
ニューラル機械翻訳

(2) 2 エポック目以降の学習には, 変更したサンプリグレート p に従い，参照訳の割合と機械学習結果の割合を変えて再構成した学習データを用いる．

サンプリングレート p は，逆シグモイド関数を用いて計算し，学習エポックごとに変更する．学習が進むにつれて p の値は小さくなり，段階的に参照訳から機械翻訳結果へと学習データ内の文脈情報を変更することで，翻訳誤りに頑健で翻訳調の特徴に対応した翻訳モデルの学習を実現する．

提案手法の効果を，時事通信社ニュースをベースにしたニュースコーパス [4]を用いた英日・日英機械翻訳実験と，IWSLT2017 [12]で公開されている TED トークコーパスを用いた英日・日英，および英独・独英機械翻訳実験で確認した．比較手法として，下記の 3 種類の手法を用いた．

手法1. 文脈情報を利用しないニューラル機械翻訳手法（Single-Sent.）

手法2. 学習時の文脈情報として目的言語側の参照訳のみを用いたニューラル機械翻訳手法（Trg-Context GT）

手法3. 学習時の文脈情報として目的言語側の機械翻訳結果のみを用いたニューラル機械翻訳手法（Trg-Context Pred.）

手法 1 はトランスフォーマーモデル [3]のエンコーダとデコーダを用いて実装した．手法 2, 3 は文脈情報を用いたニューラル機械翻訳手法であり，図 2 に示すトランスフォーマーモデルをベースとしたマルチエンコーダ文脈考慮型ニューラル機械翻訳を用いて実装した．マルチエンコーダ文脈考慮型ニューラル機械翻訳は，２つのエンコーダと１つのデコーダで構成され，文脈情報と翻訳対象の文を別々のエンコーダに入力するアーキテクチャである．翻訳性能の比較には，2 章と同様に，ランダムシードの異なる 5 つのモデルを学習してそれぞれのモデルが出力した翻訳結果の BLEU スコア [8]の中央値を用いた．

実験結果を表 3 に示す．まず，文脈情報を用いない Single-Sent.の手法と比較して，文脈情報を用いた手法は全てのタスクにおいて BLEU スコアが改善しており，文脈を利用することで翻訳精度が向上していることを確認した．次に，学習時と翻訳時で異なる文脈情報（学習時は参照訳，翻訳時は機械翻訳結果）を用いた Trg-Context GT と，学習時と翻訳時ともに機械翻訳結果を文脈情報として用いた Trg-Context Pred. との比較では，ニュースコーパスの JA-EN タスク，TED トークコーパスの DE-EN タスクを除いて，Trg-Context Pred.の翻訳精度が高かった．翻訳タスクによっては，学習時と翻訳時の両方の文脈情報に機械

翻訳結果を使ってギャップを解消することで翻訳精度が向上する可能性がある．最後に，Trg-Context GT，及び Trg-Context Pred. と提案手法との比較では，全タスクにおいて提案手法が BLEU スコアで上回っており，参照訳と機械翻訳結果の両方のデータを制御して用いる提案手法の効果を確認した．

今後の課題として，コーパスを変えた場合に，提案手法の各タスクにおける最適なハイパーパラメーター値がどのように変わるかを調査し，提案手法の効果の範囲を確認する必要がある．また，本研究では計算機資源の関係で目的言語側の前の１文のみを対象として文脈考慮型ニューラル機械翻訳を用いたが，提案手法は目的言語側の周辺の複数文を用いることも可能である他，原言語側の周辺の複数文を併用することも可能である．文脈情報の適用範囲を広げた際の効果を明らかにする必要があると考える．

4. おわりに

本研究では，ニューラル機械翻訳の学習時と翻訳時において特徴が違うデータを用いることで翻訳精度が低下する課題に着目した．学習時と翻訳時におけるデータの特徴を，翻訳対象である入力データ，翻訳結果である出力データ，入出力データ間の関係，翻訳対象とは別の外部情報の４種類に分類し，これらの特徴の違いにより翻訳精度が低下する２つのニューラル機械翻訳のタスクに取り組んだ．

１つ目は，翻訳対象である入力データ，翻訳結果である出力データ，入出力データ間の関係，の特徴の違いを扱うドメインタグを用いたニューラル機械翻訳のタスクである．従来手法では扱えなかった複数の特徴を区別して学習できる複数のドメインタグを用いた手法を提案し，４つのニュースの対訳コーパスを用いた日英機械翻訳の実験で，従来手法と比較して提案手法の翻訳精度が向上することを確認した．

２つ目は，翻訳対象とは別の外部情報として目的言語側の前文を用いる文脈考慮型ニューラル機械翻訳のタスクである．従来手法は学習時と翻訳時とで特徴が違う目的言語側の前文を用いていたために翻訳精度が低下していたが，学習時に用いる目的言語側の前文を学習の進行に応じて参照訳から機械翻訳結果へ段階的に切り替える手法を提案し，英日・日英，および英独・独英機械翻訳の実験で，従来手法と比較して提案手法の翻訳精度が向上することを確認した．

今後の課題として，本研究が扱った２つのタスク以外のタスクにおいて，学習時と翻訳時とのデータの特徴の違いにより従来手法の翻訳精度が低下していないかを検証することが考えられる．学習時と翻訳時におけるデータの特徴の違いは大量の学習データを必要とするタスクで起こりやすい現象であると考えられ，従来手法が学習時と翻訳時におけるデータの特徴の違いを考慮した手法となっていないタスクがまだあると考えられる．

謝辞

本稿で紹介した研究を行うにあたり，徳永健伸教授に多くのご指導とご助言を賜りました．深く感謝いたします．本研究成果の一部は，国立研究開発法人情報通信研究機構の委託研究（課題 197 および課題 225）により得られたものです．

参考文献

[1] Bahdanau, D., Cho, K., and Bengio, Y. (2015). “Neural Machine Translation by Jointly Learning to Align and Translate.” In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

[2] Luong, M.-T. and Manning, C. D. (2015). “Stanford Neural Machine Translation Systems for Spoken Language Domain.” In International Workshop on Spoken Language Translation, Da Nang, Vietnam.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). “Attention is All you Need.” In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc.

[4] 田中英輝, 中澤敏明, 美野秀弥, 伊藤均, 後藤功雄, 山田一郎, 川上貴之, 大嶋聖一, 朝賀英裕 (2021). 時事通信社ニュースの日英対訳コーパスの構築-第3報. 言語処理学会第27回年次大会, pp. 228-231.

[5] Kobus, C., Crego, J., and Senellart, J. (2017). “Domain Control for Neural Machine Translation.” In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pp. 372–378, Varna, Bulgaria. INCOMA Ltd.

[6] Vaibhav, V., Singh, S., Stewart, C., and Neubig, G. (2019). “Improving Robustness of Machine Translation with Synthetic Noise.” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

[7] Berard, A., Calapodescu, I., and Roux, C. (2019b). “Naver Labs Europe’s Systems for the WMT19 Machine Translation Robustness Task.” In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 526–532, Florence, Italy. Association for Computational Linguistics.

[8] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318.

[9] Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. (2015). Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28, pp. 1171-1179. Curran Associates, Inc.

[10] Zhang, W., Feng, Y., Meng, F., You, D., and Liu, Q. (2019). Bridging the Gap between Training and Inference for Neural Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4334–4343, Florence, Italy. Association for Computational Linguistics.

[11] Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In Danyluk, A. P., Bottou, L., and Littman, M. L. (Eds.), ICML, Volume 382 of ACM International Conference Proceeding Series, pp. 41-48. ACM.

[12] Cettolo, M., Federico, M., Bentivogli, L., Niehues, J., St¨uker, S., Sudoh, K., Yoshino, K., and Federmann, C. (2017). “The IWSLT 2017 Evaluation Campaign.” In Proceedings of the 14th International Workshop

解説記事

特許庁における機械翻訳に関する取組

名和　大輔
特許庁　総務部総務課特許情報室

1. はじめに

特許庁は、特許、実用新案、意匠及び商標に関する制度に基づき、特許権をはじめとする産業財産権の付与を行う機関であり、その過程で生じる情報を特許情報プラットフォーム（J-PlatPat）等を通じて対外提供することにより、権利の確認、重複した研究開発の防止、最新の技術情報の入手等を可能としている。

産業財産権に関する対外的な情報提供と当該情報の翻訳との関係を整理すると、まずは、日本の公報や審査情報を英訳することが、（1）日本ユーザーの外国での円滑な権利取得、及び（2）海外ユーザーの日本での出願の促進につながる。（1）について、産業財産権は、外国で取得される権利とは基本的に独立して存在するものの、各国特許庁の審査官は一般的に他国の特許庁の判断内容を参照するところ、日本の審査情報等を外国の審査官に理解できるように発信することで、適切に当該国で権利化されることが期待される。（2）について、海外ユーザーが日本での権利範囲や審査運用を知ることが、権利取得にあたっての予見性を高め、積極的な出願につながると考えられる。

次に、外国での特許公報等を日本語に翻訳して提供することにより、日本ユーザーの（1）海外でのビジネス展開、及び（2）日本国内での安定的な権利取得を支援できる。世界では日本語や英語以外の言語で記載される出願は多く、特に中国への出願が急増する状況下で、（1）について、日本ユーザーにとっては、中国公報等の確認がビジネス上のリスクの回避につながり、（2）について、日本のユーザーや審査官にとっては、中国公報等への容易なアクセスが適切な出願、そして妥当な審査判断をもたらす。

特許庁では、図1に示すような機械翻訳プラットフォーム（MTP: Machine Translation Platform）と呼ぶシステムを構築している[1][2][3]。当該システムが有する日英機械翻訳機能により、特許情報プラットフォーム（J-PlatPat）及びワン・ポータル・ドシエ

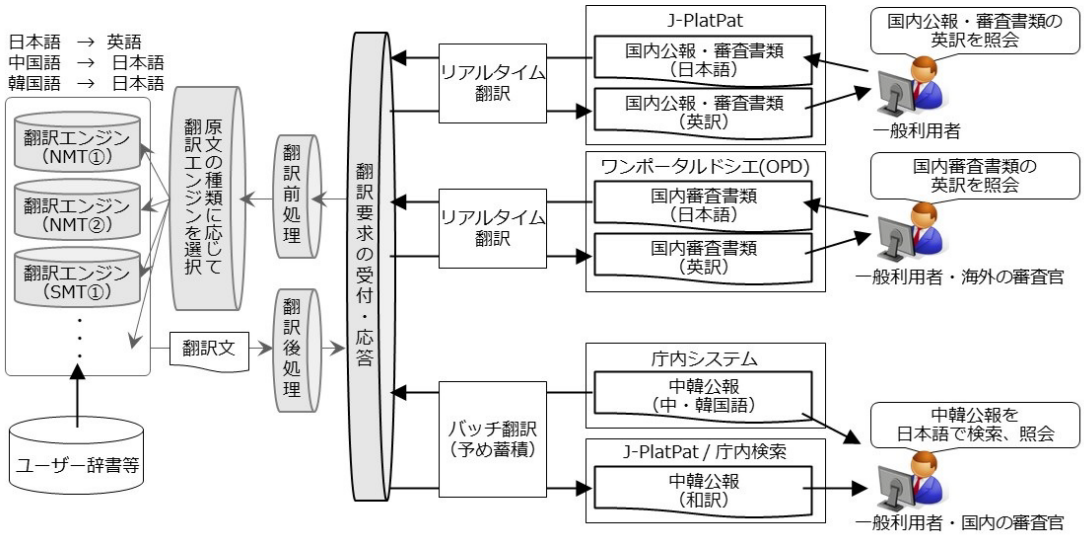


図1　機械翻訳プラットフォームの概要

Initiatives of machine translation by Japan Patent Office
Daisuke Nawa
Patent Information Policy Planning Office, Japan Patent Office

（OPD）において、国内の各種公報、審査書類等の英訳が即時的に提供されている。また、当該システムが有する中日／韓日機械翻訳機能を用いて得られる中韓公報の和訳は、J-PlatPat での検索対象として、又はダウンロードするバルクデータとして提供されている[4]。

現行の MTP においては、国立研究開発法人情報通信研究機構（NICT）で開発されたニューラル機械翻訳（NMT）エンジン及び統計的機械翻訳（SMT）エンジンが採用されている他、翻訳対象に応じて機械翻訳（RBMT）を組み合わせる、特許文献や審査書類に特有の表現に対する辞書を用いる等、全体としての翻訳品質を最適化している。

翻訳品質は、翻訳エンジン自体の性能に加えて、学習データやそれを用いた学習手法によって決まるものであり、特に NMT では学習データとして大量の対訳コーパスが必要であるといわれている[5][6][7]。特許庁では特許公報、審査書類等の翻訳品質向上に資する対訳コーパス・辞書を作成し、主にこれらの効果的な利用に関する調査研究事業を行ってきた[8]。以下では令和３年度末に終了した三つの事業を紹介する。

2. 機械翻訳の品質向上や翻訳対象の拡充に向けた最近の取組

2-1. 審査書類・審決の日英機械翻訳に関する取組

日本の審査・審判情報は、J-PlatPat や OPD を通じて特許庁外に提供されており、１．で述べたような目的の下、MTP による日英機械翻訳機能がこれらに実装されている。適切な対訳コーパスを数多く学習させることで NMT による機械翻訳が高品質になると期待されるところ、日本の審査書類・審決を元に、日英対訳コーパスの作成を行うとともに、当該対訳コーパスを用いて翻訳エンジンを学習させる際の日英機械翻訳における課題分析を実施する事業を行った。

本事業は、（１）審査書類対訳コーパスの作成、（２）審決対訳コーパスの作成、及び（３）機械翻訳の調査・分析に大別される。（１）では、特許審査書類の様々な類型から日本語原文を合計約 200 万文選定し、これを人手で英訳した。（２）では、特許のみならず、実用新案、意匠、及び商標という特許庁での審決対象から全 300 案件を選定し、人手翻訳によって約 10 万文対を作成した。（３）では、こうして得られた対訳コーパスを NMT に学習させて、その前後での翻訳品質を評価した。評価には、自動評価指標 BLEU 及び RIBES による自動評価と、評価者が目視で行う人手評価とを併用した。

対訳コーパスの学習では、これを分割して段階的に行い、その都度、翻訳品質の評価を行った。初回の学習後は、対訳コーパスの種類や翻訳対象に応じて、BLEU は無学習時の約 0.27~0.34 に対して約 0.18~0.23 の上昇、RIBES は無学習時の約 0.73~0.77 に対して約 0.08~0.10 の上昇が見られ、対訳コーパスの品質改善効果が確かめられた。学習を重ねるごとに、自動評価指標で示される全体的な翻訳精度の向上は限定的になる一方で、人手評価で示される、審査書類・審決に特有の用語や表現の誤訳数は継続的に減少し、翻訳品質が改善する傾向が見られた。審査書類・審決では、特許公報等と同様に、全体的には登場頻度が少なくとも各案件では重要な意味を持つ技術用語が存在することが多く、さらに技術の進歩に伴って新技術用語が現れるため、適切な対訳コーパスを準備する重要性は高い。

2-2. 中国審決の中日機械翻訳に関する取組

中国における特許出願件数は増加を続け、同国での特許審判事件の処理件数も同様に増加している[9]。中国で事業を展開する日本企業にとって、当該事業に関連する特許審判事件の内容は、特許権が付与されている発明の範囲を正確に把握すること、同国での審判における判断傾向を調査すること等の観点から需要が大きい。中国審決情報を機械翻訳で提供するためには、中国公報等と同様に、その品質を高めるための対訳コーパスを用いた学習が求められることから、中国審決の全文人手翻訳を行い、文単位の対訳データ形式に

編集した対訳コーパスを作成する事業を行った。本事業ではさらに、作成した対訳コーパスをニューラル機械翻訳エンジンに学習させた際の翻訳品質の向上具合等を調査した。

具体的には、中国審決約 2 万件から、対訳コーパス約 200 万文対を作成し、これを分割して翻訳エンジンに段階的に学習させ、その都度、学習効果を確認した。２－１．で述べた日英対訳コーパスを用いた結果と同様に、初回の学習後は、対訳コーパスの種類や翻訳エンジンの種類に応じて、BLEU は無学習時の約 0.28~0.36 に対して約 0.32~0.39 の上昇、RIBES は無学習時の約 0.80~0.82 に対して約 0.11~0.14 という上昇が見られた。追加学習の効果は自動評価指数において限定的であったものの、人手評価による各種翻訳エラー数で測る効果は比較的持続した。

同事業で作成された中国審決の人手翻訳文は、特許庁が提供する外国特許情報サービス「FOPISER（フォピサー）」に蓄積され、図２に示すような画面で、審決の根拠条文・日本語テキスト等による検索・照会が可能となっている[10]。さらに令和４年度は、当該事業の対訳コーパスで学習させた翻訳エンジンを用い、中国審決の機械翻訳文を作成することにより、上記「FOPISER（フォピサー）」での提供範囲を広げる予定である。

2-3. 中国公報の中日機械翻訳に関する取組

中国における特許出願件数の増加に伴い、世界の特許文献での中国特許文献の占める割合は上昇し続けている[3]。中国公報を必要に応じて調査しなければ、例えば、日本企業にとっては、中国で事業を行う際のリスクの増大をもたらし、審査官にとっては、不十分な先行技術文献調査による審査品質の低下が起こり得る。特許庁では、これらの課題に対応するため、MTP に導入されている中日翻訳機能で中国公報の和訳を作成・蓄積することにより、J-PlatPat や審査官用端末での日本語検索を可能にするとともに、民間事業者等に対する当該和訳のバルクデータ提供を行っている。

MTP による中日翻訳において、公報に記載される新技術用語等、翻訳エンジンの学習データに含まれていない用語を正確に翻訳することは難しく、技術の進展に応じて生じる新語を収集した上で、当該新語が含まれる対訳コーパスを翻訳エンジンに定期的に学ばせ

図２　外国特許情報サービス（FOPISER）の検索・照会画面

ることが、翻訳品質の向上に必要となる。そこで、MTP の翻訳エンジンの学習データに含まれていない用語（未知語）を検出する翻訳ログに基づき、未知語を含む 6 千文の対訳コーパスを作成するとともに、未知語を効率的に選定する方法について調査分析を行った。

当初の想定としては、未知語には、技術革新等により登場した新語が含まれ、このような新語を辞書登録することによる翻訳品質の向上を目指したが、翻訳エンジンの学習データに含まれていない一般的な用語、及び形態素解析で本来の文字として分割されない単語の存在を主な理由とした、上記新語を効率的に抽出することの難しさが明らかとなった。対訳コーパスによって翻訳エンジンを学習させる際にも、適切な単語分割は翻訳品質の向上に重要であると考えられ、学習データの拡充のみならず、翻訳プロセスの改善が、MTP の性能向上への一つの方向性として示された。

3. 今後に向けて

機械翻訳の近年の進展は、NMT の登場に続いて、自己注意機構と呼ばれる仕組み等、NMT をベースとしてさらなる翻訳品質の向上を実現する技術開発に支えられている。現行の MTP で採用されている翻訳エンジンの提供元である NICT においても、第２世代 NMT が開発され、第１世代に比べて翻訳品質が向上したことが報告されている[11][12]。

特許庁では、２．で述べたように、翻訳エンジンへの学習データの構築や学習手法の調査に関する取組を継続的に行っているが、当然ながら翻訳エンジンの選択が翻訳品質に大きく影響する。アクセス頻度の高い翻訳システムでは翻訳品質のみならず、翻訳速度や安定した応答性能も要求されることを考慮しつつ、翻訳対象に応じた適切な翻訳エンジンと学習データの組合せを探り、ユーザーニーズに即した翻訳サービスの提供に努めたい。

注・参考文献

[1] 西本俊之、「特許庁におけるニューラル機械翻訳の実用化」、AAMT Journal No.73
[2] 園尾聡、「ニューラル機械翻訳による特許機械翻訳システムの開発」、Japio YEAR BOOK 2019 寄稿集
[3] 中西聡、「特許庁における特許情報の翻訳に係る取組み」、Japio YEAR BOOK 2020 ミニ特集
[4] 「特許情報の一括ダウンロードサービスについて」 https://www.jpo.go.jp/system/laws/sesaku/data/download.html
[5] 張津一 松本忠博、「ニューラル機械翻訳における長文分割によるコーパスの拡張」、言語処理学会第 25 回年次大会発表論文集（2019 年 3 月）
[6] 米田昌弘 鶴岡慶雅、「教師なし英日ニューラル機械翻訳の検討と文の潜在表現の分析」、言語処理学会第 25 回年次大会発表論文集（2019 年 3 月）
[7] 須藤克仁、「ニューラル機械翻訳の進展－系列変換モデルの進化とその応用」、人工知能 第 34 巻 4 号（2019 年 7 月）
[8] 「機械翻訳に関する調査報告書」 https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/kikai_honyaku.html
[9] 中国国家知的財産権局、「2021 年国家知識産権局年次報告書」
[10] 外国特許情報サービス「FOPISER（フォピサー）」収録内容追加のお知らせ（中国審決和訳文） https://www.jpo.go.jp/support/fopiser/tsuika-20200622.html
[11] 内山将夫、「自動翻訳技術の概要と情報通信研究機構の取組」、日経エレクトロニクスセミナー、 https://www2.nict.go.jp/astrec-att/member/mutiyama/pdf/nikkei-seminar-20191203.pdf
[12] 本間奨、「特許翻訳におけるドメイン適応型機械翻訳」、パテント 2022 Vol.75 No.2

解説記事

人間の翻訳と機械の翻訳（６）：翻訳中核プロセスの一つの性質

影浦　峡
東京大学・大学院教育学研究科

1. これまでのまとめと今回の話題

本連載では、第１回から第５回までで、翻訳の対象、翻訳の対象である文書の性質、翻訳のプロセス、MT研究と翻訳論の翻訳に関する認識の違い、翻訳者のコンピテンスについて見てきました。少し傾向の違う第４回を除く４回で、対象（第１・２回）と人（コンピテンスを扱った第５回）、プロセス（第３回）を一通りみたことになります。人間の翻訳（という表現は冗長ですが）から見たとき、残る大きな領域は社会、そして翻訳の中核プロセスです。また、機械翻訳についても、そのものとしては正面から扱っていませんでした。

今回は、翻訳の中核プロセスの位置付けを考えます。翻訳の中でも特にこの、中核の部分は、プロの翻訳者は「できる」けれども、「わかる」かたちでの認識の対象としては立ち遅れているので[1]、明確なことはそれほど言えません。この記事では、本連載の第１回と第２回で述べたこと、すなわち翻訳は文書を対象とすること、文書は歴史の中で固有の位置をもつ存在であることが、翻訳の中核的なプロセスの位置付けとどう関係してくるかに関わる、一つの点を見ます。

2. 翻訳と読むこと

翻訳者・通訳者の深井裕美子氏は、「深井裕美子の読解ゼミ」案内で、次のように書いています。

> 翻訳は日本語の力が大事などと言いますが、実際にはまず原文の完璧な理解が必須です。よく言われる通り**「読めていないものは訳せない」**のです。読めるけれど、訳すのが難しくて…という人も、実際には**「読めているつもりが読めてない」**ことが多々あります[2]。

当たり前といえば当たり前ですが、重要な指摘です（私は深井氏のゼミを受講したことがないので、ここで言われている「読む」ことが具体的に何を指すのかについて私が理解していない可能性はありますが、大枠としては問題ないのではないかと思います[3]）。

「読む」という行為は、翻訳と結び付けずにそのものとして独立して捉えても（それが普通ですが）、基本的に文書を対象とします。「本を読む」、「小説を読む」、「解説書を読む」「詩を読む」が通常の読む行為を指すのに対して、「文を読む」「段落を読む」は表現としては妥当ですが少し有標性が高い感じがします[4]。なお以下で「理解する」ことが主題化されますが、上の引用で下線部に用いられた「理解」と言う言葉と「読めて」という言葉の重なりを踏まえてのことです。

起点言語文書を読むことは翻訳における必須の要素であること、そして読むことは翻訳することと同様に文書を対象とすることから、読むことの検討は、個別にいちいち翻訳と関連付けなくても、翻訳の検討の一部となります。

3. X文Y訳と語学

3.1 X文Y訳をめぐるいくつかの事例

英文和訳や和文英訳が中学校や高等学校の英語の授業の一環としてなされることが示すように、一般にX文Y訳と呼ばれるものは、語学の一環としてなされます。そして、当たり前ですが、語学は、文書ではなく言語を扱うものです[5]。では、文書を扱う翻訳と語学の一環としてのX文Y訳の違いは何か。説明の視点は複数ありえますが、ここでは、特にX文Y訳側の位置付けを明確にすることができ、またある程度の一般化

An aspect of the core translation process
Kyo Kageura
Graduate School of Education, The University of Tokyo

ができる、一つの点を扱います。まず、次の事例を考えます。

That thing is in the way[6].

例えば高等学校の英語の授業でこの文を訳す、いわゆる英文和訳で注目される点は、典型的には、be 動詞のあとに“in”が続く文の構造が把握できるかどうか、そして“in the way”が「邪魔になっている」という意味を熟語的に持つことを知っているかどうかなどであり、訳としては、例えば

あのものは行く手を塞いでいる。

といったものができればよいことになります。

She is not what she was.

であれば、

彼女は以前の彼女ではない。

と、過去形が意味することを取れるかどうかということになります。

これらの点を展開すると、例えば、文の構造については、構造を複雑にして、

Be patient toward all that is unsolved in your heart and try to love the questions themselves like locked rooms and like books that are now written in a very foreign tongue[7].

について、主語がないとか、like 以下はどうかかるかといったことを取れることが語学力として問われることになるでしょう。熟語については、頻出するものから、徐々に、あまり用いられないもの、書き言葉でのみ用いられるもの、少し古いものなど、英文和訳の素材に取り入れられていくことになります。

ところで、次の文

A syllogism, as defined by Aristotle, is ‘a discourse in which, certain things being stated, something other than what is stated follows of necessity from their being so’[8]

は、Google 翻訳[1]も DeepL[2]も、それぞれ

アリストテレスが定義した三段論法とは、「特定の事柄が述べられている場合、述べられていること以外の何かが、そのようなことから必然的に生じる言説」です。（Google 翻訳）

アリストテレスの定義によれば、三段論法とは「ある事柄が述べられているとき、述べられていることとは別のことが、そうであることから必然的に導かれる」論法である。（DeepL）

と、いずれもそれなりに意味が通る日本語にしていますが、大学初年時レベルの英語学習者には必ずしも簡単ではないようです。“in which”、“being stated”、“of necessity”、“what is being”等、文法で学ぶ事項が満載で、さらに“,”も 4 つあって複雑です。

3.2 （言）語学の対象と「理解する」

ここで、“That thing is in the way. ”に戻ります。ダメットは、ムーアの論を起点としてこの例を導入し、次のように論じています[10]。

日本語［英語］の教師が学生に対して、「あのものは行く手を塞いでいる（That thing is in the way）」という文を学生の自国語（ママ）に翻訳するよう求めるとき、**その教師はこの文を単なるタイプとみなしている。実際、このとき「どんなものが？」とか「何の行く手を？」と尋ねたりするのは馬鹿げている。**この文が単にタイプとして考察されている限りは、「問題となっている当の物」などといったものは存在しない。

さらに、次のように続けます。少し長くなりますが、引用します。

だがこれに対し、ムーアが語っていたのは、**この文のある特定の発話の理解ということについてである。**ここでの発話とは、文の本物の発話、つまり、何事かを言うために使用される発話のことであって、単に文のタイプに言及するだけのつもりで行われる発話のことではない。言うまでもなく、ある特定の発話（いま説明したような［本物の発話という］意味合いでの）を理解するためには、発話された文タイプを理解していることが必要である。しかし明らかにムーアは、当の発話を理解するためには、さら

[1] https://translate.google.com/ (2022 年 9 月 18 日)
[2] https://www.deepl.com/ (2022 年 9 月 18 日)

にそれ以上の何かが要求されねばならないと見なしていた。**こうして彼は、「理解する」という動詞に対して二重の意義を認めていたわけである。**

実際、That thing is in the way という文を「本物の発話」において理解するためには、「どんなものが？」とか「何の行く手を？」が定まる必要がありますが、英文和訳でそれは問われません。国語や英語の授業で、言語表現が「本物の発話」として扱われないことは、例えば国語で「月は太陽の周りを回っている。」という文が出てきたとき、その真偽が国語の問題としては問われないことが、示しています[11]。

4. 翻訳

4.1 「本物の発話」と文書

ここまでで図式的には自明でありかつ明確ですが、翻訳は「本物の発話」を扱います。実際、それは、連載の第 1 回と 2 回で述べた、翻訳は言語ではなく文書を対象とすること、そして文書は歴史的に固有の位置付けをもつことと対応しています。言語哲学系列の議論と異なるのは、ダメットの例もそうですが、それらが直接のその場での発話を取り上げるあるいは想定することが多いのに対して、文書には話者や受け取り手が共有する「その場」が存在しない点です。このことは、翻訳においては明示的に読み手を想定するといったことが必要であることと関係してきますが、今回はこれ以上述べません。

翻訳が、文書、「本物の発話」を扱うのであれば、起点言語文書の表現を前にしたとき、それを「理解する／読める」必要があるといったときの「理解する／読める」は、ムーアが認めていたとダメットが述べた理解の二重性の中で、タイプの理解に加えて、第二の、「本物の発話」の理解を指すことになります。あまりに当たり前なのですが、とはいえこの区別が、単に個々の現象に適用したときに境界が明確に確定しないという問題だけでなく、概念として共有されていないことが、「言語」をめぐる混乱の一因となっているように思われます[12]。

さて、改めて、今度は翻訳の観点から

That thing is in the way.

を考えてみます。口頭でその場でこの文が語られる場合とは異なり、文書に現れた場合には、That thing や the way が直接外部に対応付けられることはありません。それにも関わらずタイプとしての理解では不十分であるということは、「どんなものが？」「何の行く手を？」を、翻訳に伴う理解のプロセスで確定しなくてはならないということです。ここで大きく二つのケースが考えられます。

第一のケース。例えばこの文がある歴史書に出てきた表現であるならば、まず、“That thing”は文書の別の箇所を指し、その箇所はある年に発生した何らかの事態の指示あるいは記述として具体的な外部の出来事に結びつけられます。そこで「どんなものが？」が固定されます。「何の行く手を？」も同様です。そして、これは、文書が記述する対象が現実に存在するような領域では広く当てはまります。分野としては大きく異なる物理学の論文や図書でも、「どんなものが？」「何の行く手を？」の確定プロセスは同様になるでしょう。こうした文書においては文のレベルでの真偽は重要です。ノンフィクションを含むこうした文書の翻訳プロセスで「ファクトチェック」は、執筆時ほどではないにせよ、なされる習慣が（一部では）あります。

第二のケース。フィクションが典型ですが、その世界そのものは存在しないため、“That thing”が文書の別のところで外部の参照ではなく対象の記述として与えられる場合。この場合は、一般に、アイテムとしては実在する外部との関係はつけられないけれども、属性のレベルで外部を参照して記述が理解されることになります。興味深いことに、硬い領域の中で、抽象数学は、アイテムとして実在する外部が一般には存在しないという点で、フィクションと近い位置付けにあると考えることができます。実際、3.1 であげたリルケからの引用を理解して翻訳するときに、“rooms”は具体的にどのような部屋なのだろう、“books”はどのよう

なものだろう、“very foreign tongue” は具体的に何語なのだろうという問いに確定した答えはなさそうですし、三段論法の説明において“certain things”、“something”って何だろうという問いに具体的に答えることはできませんし、「本物の発話」としてこの発話を捉えたとしても、その理解においてまたそのような問いに答えることはそもそも不要と思われます。

4.2 文書の歴史的固有性と概念

三段論法の定義における“certain things”や“something”は、そもそもそれ自体が抽象概念でありタイプ的であるから、抽象的なレベルで認識されたものを書き留めた文書では、語学レベルでのタイプの扱いと「本物の発話」の区別は解消されるのではないかとの疑問が生じます。この点について少し検討しましょう。

（言）語学の観点も意識します。“crimes against humanity”は法的概念で、抽象的な犯罪の類型（まさにタイプ）を示す概念です。ではその概念規定は何かについては、言語学ではなく法学が扱う領域です。そして一般の市民がそれを扱うときも、それが「本物の発話」を前にした理解の場面であるならば、言語学的な意味としてではなく、素人であるとはいえ法的な概念として、この言葉が指す概念を理解しようとします。しかし、それとは別に、(言) 語学的に“crimes”の意味、“against”の意味と役割、“humanity”の意味がわかっていて、(言) 語学的に対応する日本語の表現がわかれば、英文和訳として「人道に対する罪」「人道に対する犯罪」「人間性に対する犯罪」「人間性に対する罪」などと、「それなりに言葉としての意味がわかる」表現を作ることができそうです。けれども、“crimes against humanity”が起点言語文書でそのものとして（つまり「本物の発話」に相当する意味を持つものとして）用いられている限り、「人間性に対する犯罪」「人間性に対する罪」は、“crimes against humanity”が表す概念を表すことに失敗していますし、現在では、「人道に対する罪」もその意味で不適切であることになります。これがどう決まるかというと、言語における意味としてではなく、現在では“crimes against humanity”の概念がICC規程において確立しており、他に条件がなければ第一の参照先はICC規程であり、そしてICC規程については日本語の正訳が存在するという、この世界における事実に基づいて決まっています。

理論的には、このレベルを経由しないと翻訳にはなりませんし、実際、翻訳者は、こうしたことを訳出プロセスで当たり前に行っています。前回紹介したISO 17100が示すコンピテンスのカテゴリは、二重の理解を明確にはしていませんが、例えば「分野に関するコンピテンス」は、“crimes against humanity”を（言）語学的なものとしてではなく、概念として把握することが翻訳において重要であると認識されていることに対応しています[13]。

5. 理解せず訳すこと

以上で、翻訳が言語ではなく文書を対象とするという点に関わる重要な点の一つは、言語表現をタイプとして見るのではなく「本物の発話」として見る、そのため、(言) 語学が扱う言語の意味や理解ではなく、その場の発話とも違うけれど「本物の発話」である文書を理解することが必要となる、そのために、「どんなものが？」や「何の行く手を？」をその文書において固定し、また、抽象概念もこの世界で固定するといった作業は当然している、そうしないと理論的には翻訳にならない、ということを確認しました。

理解の観点から言うと、文タイプの理解は必要条件であるけれども、EMT 2017にあるように、本来それは翻訳の前提であって[14]、翻訳はその先にある、そのそこから始まる理解が、ダメットが言うムーアの理解の第二番目に対応することになります。

そう考えると、機械翻訳は、現在のところそのものとしては“crimes against humanity”に対してICC規程とその日本語正訳を参照するという手続きを踏まないし踏めないのだから、理論的には、翻訳はそもそも

できない、ということになります。ところで、上で見た三段論法の例では、MT の出力でそれなりに（かなり？）意味がわかります。リルケの例を MT にかけていただけるとわかりますが、文体は適切とは思えないものの、やはり意味としては通じます。

この一つの理由として、翻訳でも英文和訳でも、少なからぬ場合に、結果としては同じになる、ということがあげられます。文書において現れた

That thing is in the way.

を、英文和訳ではなく翻訳する際には、「どんなものが？」、「何の行く手を？」について、その文書が扱うまさにその世界において固定する必要がありますが、ところで、その上で、この部分に対応する日本語はどのようになるかを考えて見ると、特別な場合を除けば、その手続きを経た上で、最終的に、

あのものは行く手を塞いでいる。

あれが邪魔をしている。

といった表現になるでしょう。それならば、「どんなものが？」、「何の行く手を？」を固定するプロセスは経なくても見かけ上は変わらないことになります。

もう一つ、“crimes against humanity”を、Google 翻訳も DeepL も多くの文では「人道に対する罪」と誤って訳しますが、「人道に対する犯罪」と適切に訳すこともあります。これをめぐって二つのことが考えられます。第一は、抽象概念は（言）語学的なタイプではないとはいえ、最終的な固定が記号表現のレベルでなされることです。“crimes against humanity”については ICC 規程に到達し、その概念規定を理解すれば、それに該当する事象と結びつけなくても、一応、第二の理解の段階に入っている。第二は、社会としては危機的なことですが、「本物の発話」として表出され流通している言語表現も、少なからぬ場合に、本来、ムーアの第二段階の理解に必要な固定作業を行っていないので、MT の曖昧さを総体として許容するという点です。日本語でも英語でも、それなりに確立したメディアの中で、「人道に対する犯罪」や“crimes against humanity”の誤用は少なからず目にしますし「人道に対する罪」が現在でも使わるのを目にもします。

6. おわりに

議論が紆余曲折し、話が拡散しましたが、まとめると、本稿では、(1)（言）語学は文（あるいは言語表現一般）をタイプとして扱うこと、それに対して「本物の発話」はタイプとしての理解を前提としつつ別のレベルでの理解が必要になること、(2) 翻訳が文書を対象とする以上、翻訳においては「本物の発話」の理解に相当する理解が求められること、(3) 抽象概念は概念としてはタイプであるが文書における存在としては（言）語学が扱う対象としての言語表現のタイプとしての理解では理論的には扱えないこと、を見てきました。そして、それにもかかわらず、(4)「本物の発話」の理解はできていないと考えられる機械翻訳がそれなりに訳してしまうことについても、いささか粗くですが、考察しました。

本稿では翻訳のコアプロセスの要素である「理解」を中心に考えましたが、（言）語学が想定する「理解」の位置付けはそれなりに明確にできたものの、翻訳における「本物の発話」の理解については十分に踏み込めませんでした。これについては、今後あらためて取り上げ、もう少し明確にできればと思います。

謝辞

本記事の一部は、科研研究費基盤(S)「翻訳規範とコンピテンスの可操作化を通した翻訳プロセス・モデルと統合環境の構築」（19H05660）に関わっています。この問題に触れる討論を翻訳プロセスの分析について重ねてきた阿辺川武・藤田篤・本田友乃・宮田玲・山本真佑花の各氏、この問題に関連する重要な問題提起を行った、国語科で学ぶとされていることの明確化を試みている名倉早都季氏に感謝します。

注・参考文献

[1] Miyata, R., Yamada, M. and Kageura, K. (2022) "Introduction," Miyata, R., Yamada, M. and Kageura, K. (eds) *Metalanguages for Dissecting Translation Processes: Theoretical Development and Practical Applications*. London: Routledge.

[2] 深井裕美子 (2022) 「「読解ゼミ」をオンラインで再開します」http://trans-class.blog.jp/archives/2159503.html

[3] この注記は本稿の内容に少し関係しています。

[4] 「詩を読む」は「詠む」ではなく「読む」方で、三好達治 (1952)『詩を読む人のために』至文堂（岩波文庫版は 1991 年）を頭に置いています。なお、教育のための科学研究所（https://www.s4e.jp/）のリーディングスキルテストは、言語的な単位としては文ですが、読解としては文書を想定していると考えることができます。Arai, N., Todo, N., Arai, T., Bunji, K., Sugawara, S., Inuzuka, M., Matsuzaki, T. and Ozaki, K. (2017) "Reading skill test to diagnose basic language skills in comparison to machines," *CogSci 2017*, pp. 1556-1561.も参照。

[5] 大学の「語学」授業で行われることがある訳読は、文書を対象としているという点で、高校までの語学と異なります。

[6] この例と日本語訳は、マイケル・ダメット (1998)『分析哲学の起源』（野本和幸ほか訳，勁草書房. Dummett, M. (1993) *Origins of Analytical Philosophy*. Duckworth.）の第七章「指示を欠いた意義」（岡本賢吾訳，p. 82）より。

[7] Rainer Maria Rilke (2004) *Letters to a Young Poet* (M. D. Herter Norton. trans. W. W. Norton & Company. Rainer Maria Rilke (1929) *Briefe an einen jungen Dichter*. Insel Verlag). p. 27.

[8] アリストテレス『分析論前書』。この英語は Ragged University "Logic and Reasoning" (https:// raggeduniversity.co.uk/2014/03/24/logic-reasoning/#:~:text=A%20syllogism%2C%20as%20defined%20by,necessity%20from%20their%20being%20so%27.より。

[9] ダメット，前掲，p. 81-82. 太字強調は引用者。

[10] ムーアの主張は引きませんが、以下の引用で「ムーア」が出てくるため言及しました。

[11] 作品鑑賞では、月は地球の衛星であるのにここでは太陽の周りを回っているとされていることの効果が論ぜられたりするので、訳読と同様少し「本物の発話」、文書としての扱いに踏み込むことになります。また、英語の授業や国語の授業でアディソンや樋口一葉について学ぶことがあり、それらは、他の主題科目における知識と同様に位置づけられます。

[12] 例えばこの原稿を書いている 2022 年 9 月の時点で東京都が導入を進めている、高校入試における英語スピーキング・テスト ESAT-J が、大きな問題となっていますが、試験としての深刻な問題が指摘されていることとは別に、語学が基本的にタイプに関わることを認識せず、「実際に使えない」という漠然とした観点からスピーキングについては「本物の発話」をやはり漠然と想定しながら、しかし学習と試験において「本物の発話」を実現するあるいはそれに関わる要素を入れることが読解（この場合は題材を文書として扱うことが環境の編成としては比較的容易です）以上に難しいことが認識されていないという面がありそうです。これについては大津由紀雄 (2022)「東京都「中 3 英語スピーキングテスト」深刻な問題　高校入試に活用するのはあまりに「危険」だ」東洋経済 ONLINE（https://toyokeizai. net/articles/-/590086）が、入試としてと語学学習としての両面を考慮して解説しています。

[13] ISO (2015) *ISO 17100: 2015 Translation Services – Requirements for Translation Services.*

[14] European Master's in Translation (2017) *European Master's in Translation Competence Framework*.

温故知新

新シリーズ「温故知新」

内山　将夫
AAMT

AAMT では、AAMT 創立３０周年記念事業として、過去の AAMT ジャーナルおよび JAMT ジャーナル（AAMT の前身である JAMT（日本機械翻訳協会）の会誌）を PDF 化して公開しています。

https://www.aamt.info/act/journal/

公開されているジャーナルは次の通りです。

- JAMT ジャーナル No. 01, 1991 年 7 月～No. 07, 1992 年 8 月　（今回 PDF 化）
- AAMT ジャーナル No. 01, 1992 年 11 月～No. 70, 2019 年 6 月　（今回 PDF 化）
- AAMT ジャーナル「機械翻訳」No. 71, 2019 年 12 月～（当初より PDF 版を公開）

本号から始まった「温故知新」シリーズでは、過去の AAMT ジャーナルおよび JAMT ジャーナルの記事を紹介します。

過去のジャーナル記事を読むと、当時の MT の事情が分かるとともに、現在も解決していない課題が残っていることが分かるなど、これからの MT 発展に役立つものが多いです。

本号では、今から 30 年前の 1991 年 7 月と 9 月号の JAMT ジャーナルから 2 つの記事を転載して紹介します。これらは、PDF から OCR で読み取ったテキストを修正したものです。画像等を含むオリジナルの原稿については、上記サイトをぜひご覧ください。

【1991 年 7 月　JAMT ジャーナル No.1 より】

日本機械翻訳協会設立にあたって

長尾　真（日本機械翻訳協会）

1.　ご挨拶にかえて

このたび、日本機械翻訳協会が設立され、機械翻訳システムの普及と発展、より良いシステムの作成のために、機械翻訳システムのユーザ、メーカ、および研究者がお互いに情報を交換し、一致協力をする場が出来たことはまことに喜ばしいことであります。この協会の会誌第 1 号の発刊に際し、この協会を設立するに到った目的と経緯をご説明いたしますと共に、機械翻訳システムの現状と将来、諸外国における状況と国際協力、今後の本会の進むべき方向などについて私見を述べさせていただき、できるだけ多くの方々が本協会の活動の主旨に賛同して入会され、本協会の発展を通じて社会のため、世界のために貢献していただくことを期待したいと存じます。

2.　機械翻訳の現状

機械翻訳の研究は 1950 年代にさかのぼることができます。京都大学、九州大学、電子技術総合研究所、その他で地道な研究が 30 年近くつづけられて来ました結果、ある程度のシステムが作られるようになって来ました。1970 年代の後半から 1980 年代にはいるにつれ、日本の企業活動が本格的に国際的なものとなって、国内、国外ともに日本語と英語の間の翻訳に対する需要が急激にふえ、人手による翻訳だけではまかないきれないという状況になって来ました。こういった状況から多くの計算機関係の企業は機械翻訳システムの商用化に真剣にとり組み、1980 年代の中期からいく

New series "Learning from the past"
Masao Utiyama
AAMT

つかのシステムが発表され、今日では数多くの日英、英日機械翻訳システムが市場に出ております。

3. 機械翻訳協会設立の必要性

機械翻訳の現状はこのようなものでありますので、そこに多くの誤解と混乱が生じます。人によっては機械翻訳システムは現状でも使えるという人もいれば、使えないという人もいるという状況が生じています。そしてメーカの側としてはユーザはどのような使い方、どのようなレベルの翻訳を期待し、どこに不満を持っているかが明確に把握できないというわけであります。研究者も複雑な言語現象のどこに理論上の問題点があり機械翻訳の研究をこれからどこに焦点をあてて進める必要があるのかを利用の現状を通じて認識することが必要なのであります。この困難な状況を打開するためには、機械翻訳システムのメーカ、ユーザ、研究者が一堂に会して種々の問題を自由に議論し、相互協力をする場が必要なことは明らかであります。そしてまた、長期的な視野から機械翻訳システムの発展をうながし、国際協力を押し進めてゆくためには、政府などにおけるこの分野に関する政策立案をしてゆく立場の方々の理解を得ることが必要不可欠であります。

4. これまでの努力

このような状況について通産省は大きな理解を示して下さった結果、通産省の強力な後援によって 1987 年の夏に箱根で「機械翻訳サミット」と称する国際会議を開催しました。20 ヶ国近くの国々から国の政策立案者、機械翻訳に係わる方々の参加を得て、多くの問題点について議論を行ったのでした。この会議の成果はいろいろとあったのですが、中でも機械翻訳システムは現状でもうまく使えば効果があることが明らかとなり、ユーザ、メーカ、研究者、政策立案者が一堂に会することの意義が確認されたことは非常に大きな収穫でした。これでメーカもほんとうに腰をすえてより良いシステムの開発に進むとともに、ユーザ層も急速にひろがってゆくことになりました。そして 1989 年夏にはミュンヘンで「第 2 回機械翻訳サミット」が開催され、ヨーロッパの翻訳界に大きな刺激を与え、1991 年 7 月初めにはワシントンで「第 3 回機械翻訳サミット」が開かれ、アメリカ大陸にも同様の認識が広がりつつあります。この間をぬって、翻訳に直接たずさわっている方々の考え方を聞き、またこれらの方々に機械翻訳システムの現状を正しく理解していただこうという主旨から、1989 年の春に大磯で「翻訳技術国際フォーラム」を開催しました。そこにはイギリスの翻訳者協会会長・翻訳者国際連盟の副会長をしていました V. Lawson 女史、米国翻訳者協会理事で機械翻訳に造詣の深い M. Vasconcellos 女史等を招き、種々の討論を行いました。この会議には多くの翻訳者、機械翻訳システムのユーザの方々の参加を得、機械翻訳システムに対する大きな理解をえました。特に、これからの機械翻訳の健全な発展のためには、ユーザ、メーカ、研究者が一堂に集まって意見を交換できる国際的な場を作ることの必要性が認識され、その実現について日、米、欧でそれぞれ検討をしようということになったことは、この会議の最も大きな意義の一つであったと存じます。これを受けて、同じ年の夏に開かれた第 2 回機械翻訳サミットのまとめのセッションにおいて、2 年後にワシントンで開催される第 3 回機械翻訳サミットにおいて機械翻訳国際連盟を作ることを目標とし、それまでに、日本、ヨーロッパ、アメリカ大陸のそれぞれにおいて地域的な協会の設立を考えようということが確認されたのでした。

5. 日本機械翻訳協会設立までの経緯

国内では、日本翻訳連盟、日本翻訳協会、その他翻訳に関する会が従来から活動しておられます。また日本電子工業振興協会の中には 10 年あまり前から機械翻訳システム調査委員会が種々の調査・研究を行っております。そして学会としては情報処理学会、電子情

報通信学会にそれぞれ自然言語処理の研究会があり、活発に研究活動を行っております。こういったところに、上記のような経緯をご説明し、ご理解を得ながら、昨年暮れあたりから、これらのグループの中心的な方々に何度かお集まりいただき、日本としてどのような形の団体を作るのがよいかを検討してまいりました。その間、通産省等にもご説明をし、ご理解を得るよう努力をしてまいりました。その結果、翻訳会社、ユーザ企業、メーカ企業を中心とし、また個人も十分に活動のできる形で日本機械翻訳協会を作ることが合意され、平成 3 年 4 月 17 日に設立の会合が行われ、会則の決定、理事、監事、役員の決定等を行い、ここに正式に協会が発足したわけであります。その具体的内容は本誌に掲載されております。

6. 国際的連携の重要性

そもそも翻訳は外国語を相手とするものでありますから、常に外国のことを知る努力が必要であります。特に機械翻訳においては、言語の良質な文法・辞書についてはその言語を母国語とする国と協力することが最も大切なことであります。また、多くの国で同じ言語対の機械翻訳システムの研究開発が行われておりますので、国際的な情報の交換が非常に大切となります。さらに、たとえば日英機械翻訳システムは日本国内だけでなく、アメリカ、ヨーロッパの各国の政府、大学、研究所、企業等で関心がもたれており、いわば世界中がマーケットであるということもできます。したがって、機械翻訳に関するあらゆる種類の情報交換を行い、意見を交換する場としての機械翻訳に関する国際的なグループ活動が必要なことはいうまでもありません。これについては既に上記 4 で述べました通りの経過をへて、現在アメリカに Association for Machine Translation in the Americas （AMTA）が設立され、7 月初めにワシントンで開催される第 3 回機械翻訳サミットの際に第 1 回総会が行われます。ヨーロッパでも同様の協会を作るべく現在準備がなされており、これも European Association for Machine Translation （EAMT）として、同じく第 3 回機械翻訳サミットの機会に設立総会をする予定となっています。このように日本を始めとして、アメリカ大陸、ヨーロッパで機械翻訳に関する協会が作られますので、これら 3 者の連合として機械翻訳国際連盟を作ろうということになり、これも第 3 回機械翻訳サミットで設立されることになるでしょう(本誌が皆様のお手元にとどく時点では、未来形でなく完了形の表現となっているにちがいありません)。そうして機械翻訳に関する世界中の情報が刻々とニュースレターによって皆様の手元にとどくことになるわけです。また、我々としても日本から世界に向かって積極的に発言してゆく道が開かれることになります。

7. 機械翻訳システムの将来

機械翻訳の研究が行われ始めてから約 40 年、本格的な開発と商品化の努力が行われ始めてから約 10 年になります。現在の機械翻訳システムは、まだまだ不完全とはいえ、うまく使うと人手だけにたよる翻訳よりも早く、また安くできるようになって来ました。効果をあげるためには、どのような文章が機械翻訳に向いているかをよく検討し、またテキストの分野に対応する辞書を用意すること、機械翻訳の結果を人手によってうまく後修正を行うことが必要になります。こういったことの詳細は今後発行されるこの会誌において説明され、皆様のお役に立つことになると期待しております。いずれにしても、数年前は大型計算機でしか利用できなかった機械翻訳システムが、最近はほとんどがワークステーションで使えるようになり、持ち運びも可能となって来ましたし、電話等の通信線を通じた機械翻訳のパソコンネットワークサービスも行われ始めております。ノート型パソコンに入れられて使われるという日もそのうち来るでしょう。そして学習機能をうまく利用することによって、機械翻訳システムを自分流に育てあげ、自分の分身として使うという

日も来ることは確実だと思います。

8. おわりに

以上述べて来ましたことを実現するためには、この分野に係わるすべての方々の協力が必要です。ベテランの翻訳者に近いようなシステムを作りあげるには、まだまだ研究・開発を行う必要があり、また利用技術などにおいても研究すべきことは山積しております。この日本機械翻訳協会、さらには国際機械翻訳連盟を通じた皆様方の積極的なご協力により、より良いシステムを作り上げ、広く普及することによって、21 世紀には言語の障壁をかなり低くすることができ、国際的な理解と協力、協調に機械翻訳が大きな貢献をするということを期待するものであります。

【1991 年 9 月　JAMT ジャーナル No.2 より】

MT SUMMIT-III 開催される

さる、7 月 1 日から 4 日米国ワシントン D.C のメイフラワー・ホテルで開催された。参加者約 400 名の内、日本からも 60 名位の出席があった。

まず、本稿では、MT SUMMIT-III における論文発表(Contributed Papers)およびパネル討論について報告する。2 日半に亙った会議において 8 件の論文発表と、7 件のパネル討論が行なわれた。これらを網羅的に紹介することも可能ではあるが、それよりは焦点を絞って、サミットの雰囲気を感じて貰えるようなものとしたい。

したがって、筆者が個人的に興味を持った論文発表 1 件とサミットの中心であるパネル討論 2 件（特に 1 件はサミットの予稿集に載せられていないもの）について報告する。なお、論文発表およびパネル討論の全題目を以下に示す。

なお、今回のサミットの発表において特徴的なものとしては、大規模コーパスの統計的利用と、大規模辞書作成支援が、あげられよう。前者は、論文発表 2、5 で、また後者は、論文発表 2、3、8 で触れられている。また、パネル討論においても、関連するコメントがなされていた。

論文発表

1. An Architecture Sketch of Eurotra-II
2. Advances in Machine Translation Research in IBM
3. Ultra: A Multi-lingual Machine Translator
4. Capturing Language-Specific Semantic Distinctions in Interlingua-based MT
5. ArchTran: A Corpus-based Statistics-oriented English-Chinese MT System

6. The METAL System

7. Applying an experimental MT system to a realistic problem

8. Automation of Bilingual Lexicon Compilation

パネル討論

1. The MT User Experience

2. Building the Customer Base

3. International Perspectives on MT

4. Where do Translators Fit into Machine Translation?

5. Evaluation of MT Systems

6. At the Fore front of MT Research

7. Applications of MT Technology

[内容紹介]

【論文発表報告】

Capturing Language-Specific Semantic Distinctions in Interlingua-based MT

中間言語方式による機械翻訳システムについての発表。このシステムは、Translation divergences （T.D）や Translation mismatches （T.M）が存在した場合にも、正しく、自然な翻訳ができることと、システムに新しい言語を追加する場合のコストを少なくすることを目的としている。

ここで、T.D とは、ソース言語の文とターゲット言語の文とが、同じ情報を持っているが、その構造が異なっていることをいい、T.M とは、両者の持つ情報そのものが異なっている場合をいう。たとえば、英語では「数」の情報が必要だが、日本語では必要でない。

したがって、日英翻訳の場合には、システムが「数」の情報を補ってやらなくてはならないし、英日翻訳の場合には、その情報を無視しなくては自然な翻訳にはならない。このような例は、単語に関しても起こりうる。例えば、スペイン語では、英語の fish という単語に対応するものとして、魚が自然の状態にあるか、捕らえられ食べられる状態にあるかによって、別の単語を用いる。このような場合には、翻訳システムは文脈あるいは領域知識を用いて情報を得、適切に訳し分けなくてはならない。このシステムでは、意味に注目した中間表現を用いることにより、これを可能にしている。

T.D や T.M を考慮して文生成を行なう場合には次の 2 つの課題を処理することが必要である。

・言うべきことを表現した、自然なテキストの生成

・ソーステキストの構造を利用することにより対応するターゲットテキストの構造を生成

トランスファ方式では、このような処理をトランスファ規則を用いて行なうことになるが、T.M の処理に必要な情報をそのような親則を用いて得ることには、困難が伴う場合がある。したがって、本システムでは中間言語方式が採用された。

T.D や T.M に関わる問題は、実は翻訳における問題なのではなくて、生成における問題なのである。ただ、単一言語の生成システムの場合とは違って、機械翻訳においては、「そのような問題を起こさないような意味表現」を採用することができないために、よりシステマティックな解決法が要求されるのである。

たとえば、ある意味表現を（生成に適した）別の構造の意味表現に変換することが要求される。

本システムでは翻訳は次のように行なわれる。

・ソース言語の文をソース DRS （Discourse Representation Structures）に変換する。

・ソース DRS を中間言語に変換する。(現在の所、この段階では、何も処理を行なわない。)

・中間言語をターゲット DRS に変換する。(何を表現するかを決定)

・ターゲット DRS からターゲット言語の文を生成（どう表現するかを決定）

後半の 2 つは互いにフィードバックを繰り返しながら、文を生成していく。

【パネル討論報告】

1. Evaluation of MT Systems（機械翻訳システムの

評価はどのようにするべきか)

University of Geneva の Margaret King 氏をチェアパーソンとし、New Mexico State University の Yorick Wilks 氏、University of Gothenburg の Sture Allen 氏;、University of Stuttgart の Ulrich Heid 氏、Union Bank of Switzerland の Doris Albisser 氏らによって行なわれた。

Margaret King 氏は、各パネリストが、系統だったコメントを行なえるように、機械翻訳システムの評価をいくつかの側面から切り出してみた。まず、評価者として、研究者・研究スポンサ・開発企業・システム開発者・商用システム購入者を考える。何を評価しているか（質・速度・取り扱える言語現象・コスト等）また、何が評価されているのか（今、開発段階のシステムなのか、あるいは最終の商品であるのか等）を認識することが必要である。システムの評価は、システムの内部を知って行なうのと、入出力のみを見て行なうのとでは、違ってこよう。出力をどのような基準で評価するのだろうか？実際に評価を行なうためには、システムの出力を検討し、値をつけることが必要となる。このような場合に、入力文として、言語現象を網羅するように選択された文のセット(test suites)を用意することは難しい。

一方、既存の大量の文に対して出力を検討することは、時間と費用がかかる。総合的な評価をどう行なうかが難しい課題である。たとえば、誤りの解析を考えると、1) エラーそのものが個別のシステムに依存し、2）しかも、(a）何がエラーか？（b）どんなエラーがあるか？(c) エラーの原因は？というようにすべてシステムに依存している。

Wilks 氏は研究者の立場から、研究者も翻訳の対象を予め規定して研究を行なっていること、test suites や旧来の言語理論から離れて現実の文章を対象とするべきであることなどを示した。また、評価が出力の質あるいはコストで行なわれるとして、翻訳支援システムは高い評価を得ることができるかも知れないが、研究者にとっては満足できるものではないとも述べた。評価基準としては 1）Linguistic Method 2）Technical Method 3）Organizational Method 等が示された。

Allen 氏は、研究スポンサの立場から、EUROTORA の経験を踏まえて、総合的なコメントを行なった。

Heid 氏は開発者の立場から、システム作成者と使用者（例えば、研究室とソフトウェアハウス）の協力と、システムの見通しの良さ（transparency）の重要性を述べた。また、評価において注目すべき点として、システムの拡張性（アーキテクチャ・文法・辞書等）、ハードウェア依存性、リインプリメンテーションなどを示した。

Albisser 氏は、スイスユニオン銀行での実際のシステム使用経験からの機械翻訳システムの定量的評価について述べた。これは、単に出力の質のみではなく、システムの総合的な評価を言語的側面・環境的側面・組織的側面・供給者との関係の側面の 4 つで行なうものである。言語的側面では、文の複雑さを 4 段階に、また、誤りのタイプを 2 段階に評価し、得点化する。環境的側面では、移植性・入出力インタフェース・外部辞書との関係・操作性等が評価される。組織的側面では、利用者の内部事情（辞書作成者が必要か？現在いる翻訳者を訓練することができるか？翻訳量は？システムにかけることのできるコストは？等)に沿って評価が行なわれる。実際の評価は、Logos、ALPS、METAL に対して行なわれた。

2. At the Forefront of MT Research（機械翻訳研究の最前線）

CMU の Jaime Carbonell 氏をチェアパーソンに、IBM の Peter Brown 氏、ATR の Akira Kurematsu 氏、CICC を代表して東京工業大学の Hozumi Tanaka 氏、慶応大学の Masaru Tomita 氏によって行なわれた。

まず、Carbonell 氏が、機械翻訳の最近の新しいパラダイムとして、超並列（massively paralell）、ニューラルネット、Example based MT 等を紹介した。また、

ユーザフレンドリーなOCRが組み込まれたシステム、DTP 機能を備えたシステムが普通に出回るようになっていることにも触れた。今後の機械翻訳研究の大きな流れとしては機械翻訳そのものの進化と、機械翻訳の拡張（例えば、speech-to-speech）という二つの流れがあるとの意見であった。

次に Brown 氏が、IBM で行なわれている確率的（統計的）手法を用いた文脈的アプローチの機械翻訳システムについて述べた。これは、カナダ議会の英語およびフランス語の文書を対象としており、まず、英語からフランス語への翻訳時の確率を例文等を用いて求めておき、その情報を用いて逆方向の翻訳を行なう。100個の短い文でテストしたところ、翻訳の精度は約 50%であった。最も性能の良い商用システムでも、60%の精度であるので、ある程度の精度は保たれる。しかし、より高精度の翻訳を実現するには、世界知識や文脈が必要であるという。

Kurematsu 氏は ATR のプロジェクト、SL-TRANS（音声言語翻訳実験システム）についての説明を行なった。ビデオによるデモが準備されていたが、残念ながらトラブルがあり、実現しなかった。

会議の受け付けにおける会話文に分野を限定（specific domain）した音声による日英機械翻訳システムであり、現状のシステムは語彙数が 400 語程度で、制限付の会話文（Controlled Spoken Language）を対象としている。システムは音声認識、ユニフィケーション・パーザーを利用した言語処理、音声合成などで構成されている。今後の計画としては、今後 5 年で語彙数を 3,000 語に、今後 10 年で 10,000 語程度にし、限られたドメイン（現在の Specific Domain よりも広い Limited Domain）を扱うようにする計画である。

今後の課題としては、ill-formed な入力でも解析できるロバストなシステム、不特定話者への対応、音声認識の精度向上、リアルタイム処理、多量の翻訳対の収集語彙の増大等があげられていた。

Tanaka 氏は CICC プロジェクトについて発表した。このプロジェクトでは政府開発援助のもとに、中間言語方式で日本語、中国語、タイ語、マレーシア語、インドネシア語の間の多言語機械翻訳システムを開発中である。現在、中間言語の概念の集合及び概念間の関係の集合を作っている。今後の 5 年で大規模なコーパスを用いて評価を行い標準化を進めて行く。今後 10 年では、EDR の大規模辞書を利用していく予定である。また、基礎研究あるいは意味処理・文脈処理の研究も行なう。

Tomita 氏は慶応大学の計算機環境を中心にユーザの教育の必要性に触れた。慶応大学の藤沢キャンパスでは、400 台以上の UNIX マシンがあり、24 時間体制で電子メイルやニュースが利用できる環境である。このようにキャンパス内にネットワーク環境を整備し、そのもとで機械翻訳システムを利用できるようにする。機械翻訳システムあるいは、ユーザインタフェースの改良も大切だが、ユーザの"改良"も大切である。機械翻訳システムとしては、SUN/Sparc station 上に沖電気の PENSEE/EJ を置き、それを学生に使ってもらうというプランを持っている。学生に日本語でエッセイを書いてもらい、そのエッセイを機械翻訳を用いて英語に翻訳する。その際、英語の後編集は行なわずに、良い英語が出て来るまで、何度も日本語を書き直し、機械翻訳にかける。その過程で自然に機械翻訳の使い方が分かるようになるのではないだろうか。将来は英日について同じことをする計画である。

質疑応答では、Tanaka 氏から Brown 氏に確率的手法ではこれまでに、構築されてきた、言語理論等を無視することになるのではないかとのコメントがなされた。また、MT に理論はあるか？という質問があったが、部分的な理論はあるが統合理論がないとの考えが、同じく Tanaka 氏から示された。

（電子技術総合研究所　知能情報部自然言語研究室
主任研究官　井佐原 均）

【ライブデモ報告】

展示と併行して 7 月 2 日～3 日に同ホテル・チヤイニーズルームで、各社毎にライブデモ（30 分）が実施

された。20～30 名程度の聴衆が参加していたが、質問は比較的少なかった。LOGOS と SYSTRAN の説明時に「日本語の場合、一文中に動詞が 3 個以上ある重文や複文、重複文の解析はかなり困難であり、前編集による解析結果の修正が必要になるが、英語や仏語の場合はどうか？」という質問をしたところ、けげんな顔をされ「英語や仏語ではそんなことはない。どのような場合でも 70 点位の翻訳品質が得られる。」という返答があった。「日本の研究者だけなぜこんなに苦労しているのだろうか。」という思いが一瞬、頭の中を駆け抜けた。

比較的よく似たヨーロッパ語族問の翻訳を考えてきた欧米の研究者と、全く異質な言語間の翻訳を考えてきた日本の研究者とは、翻訳に対する考え方が違っているのではないかという感じがした。また、日本で開発された技術を、外国の翻訳システムが積極的に導入すれば、さらに翻訳品質が向上するのではないだろうか。

【ポスター報告】

7 月 3 日 10 時～15 時キャビネットルームで開催された。

ポスターに採録された論文は、日本と米国が各 4 件、欧州の 2 件計 10 件であったが、論文の著者が常時待機し説明を行ったものは、このうち半数にとどまり人気は薄かった。

その中でも実用性の面から、利用者に使い易いカスタマイズ機能を整備した東芝の発表と、研究面からプリエディットせずに高品質の翻訳が得られるかを示した NTT の発表は好対照であった。以下、日・米から発表されたポスターセッション論文の概要と所感等を報告する。

・EJ/JE Machine Translation System ASTRASAC-Extension toward Personalization

<<Hideki Hirakawa 他、東芝、日本>>

・東芝が開発した日英/英日機械翻訳システムの概要と、MT SUMMIT 以降に新たに追加された利用者に対するカスタマイズ（Personalization）機能が発表された。

・主な Personalization 機能は、以下の通りである。

＊英語ドキュメントに対する OCR 入力のサポート（印刷文書の認識率 99.7%）

＊DTP 文書とのリンク

＊翻訳方法のカスタマイズ（各種の linguistic parameter が用意されている）

[例]日→英 主語のない文の扱い

一受身/人称代名詞/it/命令形/

利用者定義ストリング

[例]英→日 分詞構文の訳出

一して/するので/しながら

＊辞書のカスタマイズ

＊利用者の文書を使ったカスタマイズ

翻訳対象文書中の他の情報を使用して曖昧性を除去する。

［例］output primitives の output は、4 つの categorial ambiguities （noun、verb （infinitive、past form、present participle form））を持つが文書の他の部分に of output primitives という表現が出現すれば、output は noun と推定できる。

・Toward an MT System without Pre-Editing: Effects of a New Method in ALT-J/E

<<Satoru lkehara 他、NTT、日本>>

・前編集なしで高品質な日英機械翻訳をするために必要な方式の概要を発表したものである。

・直訳困難な日本語も意訳できるように、以下の点に着目した方式を採用している。

＊言語過程節(時枝文法)をベースにしている。

＊多段翻訳方式（Idiomatic Expression Transfer、Semantic Valents Transfer 、 General Pattern Transfer）を採用している。

＊内部で、翻訳し易い日本文へ自動書換えしている。

＊文間の意味関係を解析し、文脈処理（省略要素の補

完）を実現している。

＊言語知識が体系的に集められている。

・意味属性体系:3000 カテゴリ

・構文辞書:15,000 パターン

・日本語単語辞書（見出し語数）:40 万語

・前編集せずに翻訳するためには、どのような機械/どの程度のルール等を用意すればよいかを、実際の新聞記事の翻訳に焦点をあてて検証している。言語知識も着実に集大成されている。

CMU から 3 件、知識ベースを導入した INTER-LINGUA 方式の KANT（Knowledge-based Accurate Natural-language Translation)の設計思想、構成概要を紹介したもので、人間の介在なく高品質な訳結果を得ることが狙いである。電子工学系の技術マニュアルを対象に、英/日、英/仏、英/独に翻訳するシステムを開発中である。

英語辞書の規模は一般用語 14,000 語、専門用語が約 4～500 語である。

2 件目は、並列処理プロセッサ SNAP（Semantic Network Array Processor）上に開発された Memory-Based MT システムの発表である。ホストマシーンとして Sun3/280 が使用され、実験は比較的単純な文を例にしていたが、今後の成果が期待される。

3 件目は、JANU System 連続音声発生の英語を、日本語と独語の音声に翻訳するシステムである。国際会議予約の対話を例に LR Parser（富田氏）と Connectionist Parser（Jain 氏）を使用した場合の翻訳結果を比較している。

(NTT 情報通信処理研究所メッセージシステム研究部
主幹研究員 坂間 保雄)

【展示報告】7 月 1 日 19 時 30 分～7 月 4 日 13 時まで、4 日間セッションが開催された同じホテルの STATE ROOM で行われた。MT システムデモンストレーションに参加したメーカ及び研究開発機関は、22 件、このうち、半数の 11 件が日本からの参加であった。特に日立、富士通は 2 こまブースを使い、説明員 2 名（英日、日英）とアメリカ人のナレータを準備し 4 名体制で説明に当たっていたし、特に日立は企画会社へ依頼して日立専用ブースを作る力の入れようだ。

（通信派）富士通はワープロの OASYS ポケットから日本文を入力の後、NIFTY と ATLAS を接続して、数分後に英文ができあがってくる。同じく SYSTRAN はパソコン（IBM PC）端末からオンラインでカリフォルニアの IBM 本体にアクセスする方法でシステムを見せてくれた。SYSTRAN のもう一つの特徴は、DTP 機能が付加されている点が良い。

（ラップトップ派）東芝、カテナは SPARCLT を使い、翻訳システムを動かす。NEC、三菱はワークステーションにラップトップパソコン（PC ノート、AX ノート）を 2～3 台接続するシステムであった。

NEC は利用ユーザである IBS 桜井社長（翻訳会社）との協力体制で参加。

ユーザとメーカの協同で、一般顧客は PCVAN を通して PIVOT を使い、翻訳家の Pre-Post Edit が必要であれば IBS が対応してくれるという事で、デモでは具体的な経験に基づく事例で説得力があった。カテナリソースは、スピードが速いというセールスポイントを全面にアピール。

（実績派）日本で最も販売しているシャープ（約 600 台）と沖（数百台）は、英日翻訳システムのバージョンアップした DUET E/J II と PENSEE II が実績をもとに説明をしていた。湾岸戦争で話題のブッシュ大統領決断のスピーチを訳し BUSH を(ブッシュ大統領と潅木）BAKER を（ベーカー国務長官とパン屋）と言うように文章の内容を選択し訳語の抽出を変えて見せていた。この説明は、日本語を充分理解できない人でも、意味の違いを正確に機械が訳し分ける事を説明していた。また翻訳結果失敗の場合に、◇印(失敗マーク)を出すことで感心した人も多いようだ。

次に各ブースの出展者に聞いた。【来場者の質問ベスト 10】をまとめてみた。

1. システムの価格はいくらか？
2. 現状展示機能と今後の他言語間との開発計画？

3. ソフトウェアーのバージョンアップをどうするか？

4. 開発言語は何か（C 言語..）？

5. 本体 CPU は何か（MC68030..）？

6. 入力方法は（OCR は接続可能か）？

7. Post-Edit の方法を操作面から見て使い易いかどうか？

8. 開発メンバーの人数及び開発期間？

9. 開発開始時期と発売予定時期？また、発売している場合は、アメリカで販売しているのか？

10. 今後の問い合わせ先は？

CNN-TV 取材もあり、各ブースでは熱心な質問や意見交換があり盛況であった。

（シャープ（株）情報システム研究所 信田 恵壱）

【MT SYSTEM DEMONSTRATIONS】

	メーカ名称及び開発研究機関名称	国名	システム名	翻訳対象言語
1.	Catena-Resource Laboratories Inc.	日本	「STAR」	E⇒J 「The Translator」 E⇒J
2.	Center of the International Cooperation for Computerization	日本	「MTS for ASIAN LANGUAGES」	J/Chi/タイ/インドネシア/マレー(相互間)
3.	CONCERN DATA	スウェーデン	「STYLUS」	E⇒Ru
4.	ECS, Inc.	アメリカ	「Linguistic Toolkits」	E/F/Sp/J/Ko (相互間)
5.	EUROTRA-D & EUROTRA-NL	ドイツ	「CAT2」	E/F/G 「MiMo2」 E/Du/Sp
6.	Fujitsu Limited	日本	「ATLAS」	E⇔J NIFTY-Serve
7.	Hitachi, Ltd.	日本	「HICATS」	E⇔J
8.	Intergraph Corporation	アメリカ	「DP／Translator」	E⇔G E⇔F E⇒Sp E⇒It
9.	Japan Electronic Dictionary Research Institute, Ltd.	日本	「電子化辞書」	
10.	Linguistic Products	アメリカ	「PC-Translator」	E⇔Sp E⇔F E⇔ﾃﾞﾝﾏｰｸ E⇔Sw
11.	Logos Corporation	アメリカ	「LOGOS」	E⇒Sp E⇒F E⇔G E⇒It
12.	Mitsubishi Electric Corporation	日本	「MELTRAN-J/E」	J⇒E
13.	NEC	日本	「PIVOT」	E⇔J
14.	Computing Research Laboratory -New Mexico State University	アメリカ	「ULTRA」	E/G/J/Sp/Chi (相互間)
15.	NTT Communications and Information Processing Laboratories	日本	「ALT-J/E」	J⇒E
16.	Oki Electric Industry Co., Ltd.	日本	「PENSEE」	E⇔J
17.	SHARP CORPORATION	日本	「DUET-E/J II」	E⇒J
18.	Siemens Nixdorf Information Systems, Inc.	ドイツ	「METAL」	G⇔E
19.	Socatra XLT Inc.	カナダ	「XLT」	E⇔F
20.	SYSTRAN	アメリカ	「SYSTRAN」	E⇔F E⇔G E⇒Sp E⇒It E⇒Du E⇒Po E⇒ｱﾗﾋﾞｱ Ru⇒E
21.	Toshiba Corporation	日本	「ASTRANSAC」	E⇔J
22.	Translation Technologies International, Inc.	アメリカ	「TOVNA」	F⇒E Sp⇒E G⇒E ｱﾗﾋﾞｱ⇒E Chi⇒E J⇒E Ko⇒E

E;英語 J;日本語 Chi;中国 F;フランス G;ドイツ Sp;スペイン Sw;スウェーデン lt;イタリア Du;オランダ Po;ポルトガル Ko;韓国 Ru;ロシア語⇔;双方向翻訳⇒;一方向翻訳 /;相互間翻訳

イベント報告

AMTA2022 参加報告

田中　英輝
情報通信研究機構

1. AMTA2022 の概要

本稿では、2022 年 9 月 12 日から 9 月 16 日までアメリカ、フロリダ州、オーランドの Sheraton Orlando Lake Buena Vista Resort で開催された AMTA2022 について報告する。AMTA は The Association for Machine Translation in the Americas、アメリカ機械翻訳協会の略称で、隔年で年次大会を開催しており、AMTA2022 は 15 回目の年次大会である。

AMTA は、私たちアジア太平洋機械翻訳協会（AAMT）とヨーロッパ機械翻訳協会（EAMT）と姉妹関係にある団体で、機械翻訳のユーザー、研究者、開発者などさまざまなステークホルダーが会員となっている。このため年次大会は、機械翻訳に関わる幅広い人が参加し、多岐にわたる内容が報告されるところに特徴がある。

今回は、2021 年にアメリカで開催された MT Summit と同様、リサーチ、ユーザー・プロバイダー、政府関係者の 3 つのトラックで発表が行われた。発表数の内訳は表 1 の通りである。また採択率は全体で 60%だったとのことである。

表 1　発表の内訳

リサーチ	25
ユーザー・プロバイダー	26
政府	11
合計	62

AMTA2022 はハイブリッド形式を採用し、現地、遠隔のどちらでも発表、聴講が可能であった。特徴的だったのは、遠隔発表者と現地発表者を混在させて同じセッションで報告させていたことである。会場の発表は遠隔に中継され、遠隔発表は現地会場のスクリーンに投影された。主催者側にとって技術上、またプログラム編成上の難しさは大きいと想像するが、遠隔発表者は現地参加者とほぼ同じ体験を享受できたと思われる。

またすべての発表は現地で録画され、以後 90 日間は AMTA2022 のサイトで参加者に、それ以後は AMTA のホームページで AMTA 会員に公開されるとのことである。

今回の参加者の内訳を表 2 に示す。遠隔参加者は現地参加者のほぼ 2 倍だったことがわかる。

表 2　参加者の内訳

現地	148
遠隔	284
合計	432

本会議のほか、2 件のワークショップ、6 件のチュートリアル、1 件のラウンドテーブルが開催された。また本会議中には 3 件の基調講演が行われ、それぞれの直後には関連した話題のパネルディスカッションが設けられていた。

2. AMTA2022 の特徴と傾向

チュートリアルのタイトル一覧を表 3 に示す。

表 3 チュートリアルのタイトル一覧

1	Introduction to MT
2	AutoML for Neural Machine Translation
3	Machine Translation User Guide 2022
4	Creating Text-to-Speech Software for Everyday Language
5	Recent Advances in Translation Quality Estimation
6	The Post-editor Toolkit

これらのタイトル自身が、機械翻訳の関係者の関心

Report on Participation in AMTA2022
Hideki Tanaka
National Institute of Information and Communications Technology

の高い分野の反映であり、AMTA2022 での流行、傾向を示している。

以下、多少内容を補足する。

チュートリアル 1 では、これから MT でホットになる話題として大規模言語モデル（GPT3 など）の応用、適応的・予測的機械翻訳、文脈利用翻訳、マルチモーダル翻訳などを挙げていた。適応的・予測的機械翻訳とは翻訳システムを使う中でユーザーが誤りを修正するとすぐさまそれが反映されるような仕組みを指している。大規模データからの学習を基本とする機械翻訳にとって実現は難しいが現場のニーズの高い技術である。

チュートリアル 2 は、機械翻訳システムのハイパーパラメタを自動的に学習しようという研究分野の解説であった。機械翻訳システムの構築には、大規模データからパラメタと呼ばれる膨大な数の数値を学習によって決定するプロセスが含まれる。これに対して、あらかじめ与えられた数値に固定され、学習されない変数をハイパーパラメタと呼ぶ。現在、ハイパーパラメタは経験や直感で決められることが多い。しかし、ハイパーパラメタを変えると翻訳性能も大きく変わることが実験で確かめられており、その最適な決定法が求められている。

チュートリアル 4 は音声合成システムを作るためのツールキットの紹介である。機械翻訳同様、音声合成でも深層学習を使うことで飛躍的に音声の質が向上している。次節で述べるが、今回のユーザー・プロバイダートラックでは同時通訳システムの発表が 4 件あった。音声合成システムは同時通訳システムに必須の技術であり、大きな関心があったものと思われる。

チュートリアル 5 は、性能向上が著しい翻訳品質の自動評価に関する解説である。ユーザー・プロバイダートラックでも評価をワークフローに入れる発表や、評価そのものに関わる発表が数多くあったことと合わせて、現在大きな関心を集めている分野だと思われる。

本節を締めくくるにあたり、本会議の各トラックで最優秀発表賞に選ばれた発表を挙げておく。これらは、現地、遠隔両方の参加者の投票によって決められた。

- リサーチ
  - Domain-Specific Text Generation for Machine Translation
  - Measuring the Effects of Human and Machine Translation on Website Engagement
- ユーザー・プロバイダートラック
  - Business Critical Errors: A Framework for Adaptive Quality Feedback
- 政府トラック
  - Dragonfly: Automated Sign Language Recognition and Machine Translation（本稿執筆時に ACL Anthology への登録なし）

ユーザー・プロバイダートラックの受賞発表は、新しく Business Critical Errors という概念を提唱した Unbabel 社の発表である。MQM に代表される人手の翻訳品質評価では、あらかじめ設定した文法カテゴリの誤りをアノテーションしてスコアを計算する手法が主流である。これに対して Unbabel 社では、文法誤りを気にするクライアントは少なく、むしろそれ以外の点を指摘するクライアントが多いことを幾度となく経験した。そこで、同社では、ユーザーが気にする文法以外の誤りを打合せで獲得し、これを MQM のように点数化する仕組みを作るワークフローを整備している。文法誤りにこだわり過ぎることへの警鐘を鳴らす発表であった。

3. ユーザー・プロバイダートラックの特徴

本トラックでは言語サービスプロバイダー、システムインテグレーター、研究者など幅広い分野からの発表が行われ、内容も多岐に渡った。中でも、翻訳品質評価、システム評価、ワークフロー提案、MT システム開発（特に学習データ収集法）などの話題が目についた。また、同時通訳（音声翻訳）システムに関連した報告が 4 件あったことは印象的であった。まだビジネスレベルに到達した同時通訳システムの事例は少ないと思われるが、今後さらに注目を集める分野になりそうである。

以下では報告者が特に興味を持った発表 3 件について報告する。

Automatic post-editing of MT output using large language models

チュートリアル 1 で指摘のあった大規模言語モデルの応用の 1 つとして興味深い報告であった。

製品名、会社名、専門用語、略語などの用語統一は翻訳時の重要なポイントの一つである。このため ULG 社では翻訳結果中の置換対象となる用語に XML タグを付与し、これを辞書で強制置換しているが、人手評価によれば性数の不一致、語順の乱れなどの問題が発生することがわかっている。そこで、Open AI の巨大言語モデル GPT3 を使ってこの誤りを修正する手法を検討している。GPT3 はプロンプトと呼ばれる命令と共に文を入力すると、命令通りに文を変換する言語モデルである。今回は英語からスペイン語への翻訳を対象に GPT3 を使った以下の実験を行なっている。

1. 対象文：用語置換を行い、文法誤りが混入したスペイン語文、プロンプト：「文法誤りを訂正すること」
2. 対象文：同上、プロンプト：「語順を訂正すること」
3. 対象文：用語置換タグのない英語文、プロンプト：「スペイン語に翻訳すること」
4. 対象文：用語置換タグのある英語文、プロンプト：「次の用語置換を行なってスペイン語に翻訳すること（置換前用語／置換語用語）」

手法 1 と 2 は GPT3 をスペイン語のポストエディットシステムとして、手法 3 と 4 は英語–スペイン語翻訳システムとして使っている。

さらに GPT3 を使った上記 4 手法に加えて ULG の翻訳システムを使った次の 2 手法を加え、合計 6 つの実験を行なっている。

5. 用語翻訳を行わない ULG 社の翻訳
6. 用語翻訳を行う ULG 社の翻訳（上記の手法 1 の入力となる）

結果を BLEU、TER で評価すると共に、用語統一率で評価している。

手法 6 は 99%を超える用語統一率を達成したが、最初に述べたような文法誤りを含んでいる。これに対して手法 1 では用語統一率は 98.11%と高いまま、性数の一致、冠詞、前置詞の変更などが適切に行なわれており、BLEU も変わっていない。これらから著者らは手法 1 を有望な手法だと結論している。一方、手法 2 では不要な用語の変更が行われていた。手法 3、4、すなわち翻訳システムとしの GPT3 の性能は不十分で、特に意訳が過ぎること、用語統一の効果が得られない問題があったと報告している。

今回の GPT3 はありものをそのまま使い、それで文法修正のポストエディットが機能したという結論を得ているが、GPT3 をチューニングした場合に性能がどう変わるか興味深い。

Data analytics meets machine translation solution

VMWare 社では、毎月 700 万セグメントにおよぶ膨大な数の機械翻訳のポストエディット（PE）を行なっている。同社で PE 率を計測して、主観評価した結果、PE 率=0%は完璧、0%＜PE 率<20%は「良好」、20%<PE 率は「不良」という関係があることがわかったという。そこで PE 率を翻訳品質とする予測システムを開発している。また、 すべての PE 率をモニタすることで、MT システムの弱点、翻訳の異常発見に役立てている。膨大な数の PE 率を把握するため Elastic Search によるデータベース管理、Kibana によるデータ可視化を行ないさまざまなグラフをダッシュボード表示できるようにし、MT 開発者、翻訳品質管理担当者などだれでもアクセスできるようにしている。たとえば、個別セグメントの PE 率の時間順のグラフを見ることで、異常な PE、すなわち MT システムの弱点を検出できる。また、期間ごとの PE 率によるカテゴリ「完璧、良好、不良」の分布の違いをみることで、翻訳品質の安定性を観察できる。大量の翻訳結果と PE 結果から得られる指標を自動的にグラフ表示することで、いろいろな気付きにつなげるデータサイエンスの手法は今

後の有望な方向の一つと思われる。

Comparison between ATA grading framework scores and auto scores

翻訳の自動評価は機械翻訳システムの出力の良し悪しを判定するために開発されている。

これに対して本報告は、人の翻訳結果に対する自動評価と人手評価の関連（相関）を調べる小規模な実験を行なっている。実験設定は以下の通りである。

- 翻訳の課題
  - 250 語程度の 2 つの英語パッセージを中国語へ翻訳
- 翻訳者
  - 翻訳の学生 2 名と経験の異なる翻訳者 6 名の合計 8 名
- 人手評価
  - ATA（アメリカ翻訳協会）の定める基準に従って 2 名で評価。これは MQM に似た基準で、事前に決めた観点の誤りがあればスコアを加算していくため、誤りが多いほどスコアは大きくなる
- 自動評価
  - 参照訳を利用する BLEU、TER と COMET、および参照訳を使わない COMET（複数）
- 参照訳
  - 2 種類：一つは、直訳。他方は意訳（発表者者は fancy な翻訳と呼んでいた）

6 名の翻訳結果に対する ATA スコアと自動評価スコアの相関係数を計算し、その大きさで関係を分析している。この際、以下の状況を含めた上で分析している：2 種類の参照訳（直訳、意訳）による違い、6 名の翻訳者に加えて、自動評価の参照訳に使わなかった他方の参照訳を評価対象に加える、自動翻訳の結果 3 種を評価に加える。主な結論は、1）参照訳を使う自動評価手法は、参照訳に大きな影響を受ける、2）参照訳を使わない COMET スコアは ATA スコアと強い相関を持つ、3）高い相関は、スコアの中間部分に限られ、両端部分では直線関係からずれる。 結論 3）は、自動翻訳結果や参照訳の一方を加えた分析から得ている。

翻訳者の能力評価に参照訳を使わない COMET が使える可能性を示唆した研究であるが、意訳、レベルの低い MT の翻訳などではうまく評価できないようである。原因を報告者に尋ねたところ、現在追及中とのことであった。

4. おわりに

コロナ禍の影響で AMTA2020、MT Summit2021 と 2 回続けて AMTA 主催の会議は遠隔開催となっていた。AMTA2022 は 2 年ぶりの現地開催に加え、遠隔参加も加えたハイブリッド会議として開催された。先に述べたように遠隔発表者は現地発表者と同じ体験ができるよう工夫されており、全員が満足できる会議になっていた。また、人とのつながりを大事にしたいという実行委員会の配慮が随所に現れており、コヒーブレークに加えて、朝食会場も会議場横のホワイエに設けられ、朝からネットワーキングができるように工夫されていた。おかげで報告者もいろいろな人と知り合う機会を得た。AMTA2022 実行委員会のみなさんに感謝したい。

2023 年には AAMT 主催で MT Summit 2023 がマカオで開催されるため AMTA の年次大会は 2024 年に開催される予定である。情報は随時 AMTA のホームページに掲載されるので、ご興味のある方は是非参照されたい。

イベント報告

AMTA 2022 Conference 参加報告

平岡　裕資
立教大学異文化コミュニケーション研究科博士後期課程

1. AMTA 2022 Conference の概要

今回で第15回目となるAMTA 2022 Conferenceは、9月 12 日～16 日のおよそ一週間、アメリカ・フロリダ州のオーランドで開催された。本会議は、機械翻訳に携わる様々な人々（ユーザー、研究者、開発者など）の知識共有・意見交換を目的とした場である。発表は合計で 50、基調講演が 3 件、ワークショップが 3 件、チュートリアルが 6 件行われ、形式はオンサイトとオンラインのハイブリット型で進行された。本稿では、報告者が参加したワークショップについて主に報告する。

2. ワークショップ：Workshop on Empirical Translation Process Research (WeTPR)

Empirical Translation Process Research とは、人間翻訳およびポストエディットのプロセスを実験に基づく手法で解明する研究のことを指す。本ワークショップでは、TPR の方法論・問題意識・最新の研究結果を機械翻訳コミュニティ内で発信することで、TPR の重要性が認知されることを目指している。発表の数は 8 件、その内大枠の理論に関するものが 1 件、翻訳者の負荷に関するものが 2 件、翻訳エラーの要因解明に関するものが 4 件、翻訳の受容者（読者）の負荷に関するものが 1 件あった。

その中でも特に興味深かったのは、Ogawa（2022）の“Predicting the Number of Errors in Human Translation using Source Text and Translator Characteristics”であった。この研究では原文及び翻訳者の特徴と翻訳エラー発生率の関係性を解明している。特に面白いと感じたのは、原文の特徴の中でも統語的複雑さ（syntactic complexity）と長い単語の割合（proportion of long words）が翻訳エラーの発生率に影響する、という点である。なぜ興味深いかというと、これは機械翻訳向け制御言語・プリエディットに関する研究成果と類似しているからだ。例えば、原文の統語的複雑さを解消することはニューラル機械翻訳の品質の向上につながることがわかっている[1]。また、機械翻訳が翻訳しやすい「産業日本語」の書き方をまとめた特許ライティングマニュアル[2]などでは、複合名詞（つまり、長い単語）は分解すべきとされている。人間翻訳者が起こすエラーを誘発する日本語と機械翻訳が起こすエラーを誘発する日本語のそれぞれの特徴が類似していることは、言い換えれば、機械翻訳向けに制御された制御言語は（すべてではないにしても）人間翻訳者にとっても有効であり翻訳しやすいものである、と言えるかもしれない。Ogawa (2022)による研究成果は、翻訳言語方向や分野など特定の状況下に制限されてはいるものの、報告者がテーマに掲げる「日英翻訳の機械翻訳および人間翻訳者のための制御言語・プリエディット」にとって非常に意義のあるものであった。

3. おわりに

以上、報告者が参加したワークショップとその中で特に興味深かった研究について報告した。今後同様の取り組みが継続され、TPR で扱われる問題と手法論が機械翻訳コミュニティだけでなく実務翻訳者や翻訳研究者にも広く共有されることは、さらなる翻訳の解明とそれに基づく機械翻訳の発展につながることは間違

Report on AMTA 2022 Conference
Yusuke Hiraoka
Rikkyo University Graduate School of Intercultural Communication PhD student

いないだろう。

参考文献

[1] Miyata, R., & Fujita, A. (2021). Understanding Pre-Editing for Black-Box Neural Machine Translation. ArXiv, abs/2102.02955.

[2] 一般財団法人日本語特許情報機構特許情報研究所 (2018). 特許ライティングマニュアル（第 2 版）. [Online] https://tech-jpn.jp/tokkyo-writing-manual/

イベント報告

第 9 回アジア翻訳ワークショップ（WAT2022）開催報告

中澤　敏明
東京大学

1. はじめに

本稿では WAT2022[1]の開催報告を行う。アジア翻訳ワークショップ（Workshop on Asian Translation, WAT）はアジア言語を中心とした評価型機械翻訳ワークショップであり、2014 年に第 1 回（WAT2014）を開催して以降、毎年開催している。本稿の著者は初回からオーガナイザーの一人としてワークショップの運営を行っている。2016 年の第 3 回（WAT2016）以降は自然言語処理の国際会議との併設ワークショップとして開催しており、2022 年の第 9 回（WAT2022）は韓国の慶州（キョンジュ）で開催された COLING 2022 の併設ワークショップとして、2022 年 10 月 17 日にオンラインと現地でのハイブリッド形式で開催された。参加者は現地参加とオンラインと合わせて約 30 名程度で、現地参加が 2/3 ほどであった。

ワークショップは様々な機械翻訳の評価タスクの実施に加えて機械翻訳に関する研究論文の募集も行っており、WAT2022 では 4 件の研究論文を採択した。採択した研究論文のタイトルと簡単な概要を以下に示す。

- Does partial pre-translation can improve low resourced-languages pairs?

本論文は日本語・フランス語間の翻訳において、数値表現、固有名詞、動詞などいくつかのレベルで入力文の一部を目的言語に事前翻訳しておき、これを NMT に入力することで翻訳精度の向上を目指している。結果としては事前翻訳による翻訳精度向上にはつながらなかったが、より深く分析することで今後良い結果が得られる可能性もあるのではないかと思われる。

- Multimodal Neural Machine Translation with Search Engine Based Image Retrieval

本論文はマルチモーダル機械翻訳において、入力文から抽出したキーワードを用いて画像検索を行い、得られた画像も併せて用いることで翻訳精度の向上を目指したものである。提案手法は画像情報のない入力においてもマルチモーダル機械翻訳が可能となる。実験により、テキストのみを用いた場合よりも良い精度であることが示され、またランダムの画像や空の画像情報を入力した場合よりも良い精度であることが示された。

- Comparing BERT-based Reward Functions for Deep Reinforcement Learning in Machine Translation

本論文は強化学習により NMT の最適化の目的関数を様々に変えた実験をおこなっている。BLEU のような単語の表層に基づく目的関数を用いるよりも、BERT に基づく指標を用いる方が翻訳精度が向上することを示している。

- Improving Jejueo-Korean Translation with Cross-Lingual Pretraining Using Japanese and Korean

本論文はチェジュ語と韓国語間の翻訳において、日本語と韓国語のモノリンガルコーパスを利用して mBART の事前学習を行うことで、翻訳精度を向上させられることを報告するものである。チェジュ語は話者が数千人しかいない低資源言語だが、言語構造的に近い日本語のデータを用いることで翻訳精度が向上することを示している点で面白い。

招待講演はニューヨーク大学の Duygu Ataman 氏より “Machine translation of Turkic languages: Current approaches and Open challenges” というタイトルで行われた。トルコ語はラテン文字、キリル文

Report of the 9th Workshop on Asian Translation (WAT2022)
Toshiaki Nakazawa
The University of Tokyo

字、ペルシャ文字などいろいろな文字が使われ、多様な正書法を持ち、比較的語順が自由で、数多くの形態素を持つ言語であるため、言語処理の研究対象としてチャレンジングである。Ataman 氏はトルコ語の言語処理に対して、特に 1. コーパスの収集とベンチマーク、2. モデルの構築、3. 評価の 3 つの課題を挙げ、それぞれにおけるこれまでの研究成果を紹介した。Ataman 氏の講演で用いられたスライドは WAT2022 のホームページ [1]でダウンロードすることができるので、興味のある方は是非ご覧いただきたい。

WAT2022 では以下の Shared Task を設定した。なお各タスクの後ろに括弧付きで示されている数字は、各タスクの参加チーム数である。数字の示されていないタスクは､参加チームがなかったことを表している。

・文書単位翻訳タスク:

- ASPEC+ParaNatCom: 英 → 日　科学技術論文
- BSD Corpus: 英 ⇔ 日　ビジネスシーン対話
- JIJI Corpus: 英 ⇔ 日　ニュース
- NICT-SAP: ヒンディー/タイ/インドネシア/マレー(1)/ベトナム ⇔ 英　非構造的文書
- 日中韓 ⇔ 英　構造的文書 (1)

・マルチモーダル翻訳タスク:

- Hindi Visual Genome: 英語 → ヒンディー語 (3)
- Malayalam Visual Genome: 英語 → マラヤーラム語 (2)
- Bengali Visual Genome: 英語 → ベンガル語 (3)
- Ambiguous MS COCO: 英 ⇔ 日
- Video Guided Ambiguous Subtitling: 日 → 英
- Khmer Speech Translation: クメール → 英/仏

・インド諸語翻訳タスク (2):

- アッサム/ベンガル/グジャラート/ヒンディー/カンナダ/マラヤーラム/マラティ/ネパール/オリヤー/パンジャーブ/シンド/シンハラ/タミル/テルグ/ウルドゥ ⇔ 英

・特許翻訳タスク:

- JPC3: 英中韓 ⇔ 日

・制限翻訳タスク:

- ASPEC: 英 (3)/中 ⇔ 日

・対訳コーパスフィルタリングタスク:

- JParaCrawl: 日 ⇔ 英 (1)

昨年度は 26 チームの参加があったのだが、残念ながら今年度は 8 チームの参加にとどまった。表 1 に参加者のリストを示す。

Team ID	Organization	Country
TMU	Tokyo Metropolitan University	Japan
NICT-5	NICT	Japan
Sakura	Rakuten Institute of Technology Singapore, Rakuten Asia.	Singapore
CNLP-NITS-PP	NIT Silchar	India
NITR	NIT Rourkela	India
HwTscSU	Huawei Translation Services Center, 2012 Lab, Huawei co. LTD; School of Computer Science and Technology, Soochow University	China
SILO_NLP	Silo AI	Finland
nlp_novices	SCTR's Pune Institute of Computer Technology	India

表 1：WAT2022 Shared Task 参加者リスト

次章では制限翻訳タスク及び対訳コーパスフィルタリングタスクについて、タスクの簡単な説明と評価結果、および得られた知見について述べる。

2. 評価結果と得られた知見

2.1 制限翻訳タスク

制限翻訳タスクは WAT2021 で新たに追加されたタスクであり、今年度は新たに日中の言語対が追加され

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2022/index.html#invited-talk.html

た。しかしながら日中のタスクへの参加者は今回はなかった。制限翻訳タスクでは入力文とともに出力文で必ず使わなければならない訳語のリストが与えられるので、システムはこれらの訳語を必ず含むように翻訳を生成しなければならない。なお与えられるのは訳語のリストのみであり、入力文のどの語に対応する訳語なのかは与えられない。

日英・日中共に ASPEC のデータ(dev/devtest/test)を対象として、人手で専門用語(制限用語)の抽出を行なっている。表 2 に文単位の制限用語の平均数を示す。

	英日	日英	中日	日中
dev	2.8	2.8	1.2	1.2
devtest	3.2	3.2	1.5	1.5
test	3.3	3.2	1.4	1.4

表 2：文単位の制限用語の平均数

自動評価は BLEU スコアと一貫性スコアにより行い、最終的なシステムのランキングはこれらを組み合わせたスコアにより行なった(表 3 中の final)。一貫性スコアは全ての制限用語が出力できた文の割合であり、最終ランキングは全ての制限用語が出力できた文のみで計算した BLEU スコアにより決定した。人手評価は特許庁が公開している「特許文献機械翻訳の品質評価手順」の中の「内容の伝達レベルの評価」[2](Accuracy と呼ぶ)と source-based Direct Assessment[2]により行なった。

日英・英日のタスクに参加したチームは、Lexical Constraint Aware NMT (LeCA[3])と Multi-Source Levenshtein Transformer (MSLevT[4])を組み合わせた手法を用いていた。LeCA のみでは全ての制限用語を必ず用いるという制約を満たさない出力が得られる場合があるため、MSLevT による Post-edit を行うことで 100%の精度で制限用語を用いることに成功している。表 3 に自動評価と人手評価の結果を示す。なお表中のアスタリスクは、Human Reference との見分けがつかないことを示す。

En-Ja Outputs	final	HE (Accuracy)	HE (src-based DA)
LeCA+LevT (*ensemble*)	**52.7**	4.18	**76.4***
LeCA+LevT	50.5	4.19	**76.6***
LeCA only	37.6	**4.24**	74.9
Human Reference	–	–	76.6
Ja-En Outputs	**final**	**HE (Accuracy)**	**HE (src-based DA)**
LeCA+LevT (*ensemble*)	**40.8**	**4.31**	**74.1***
LeCA+LevT	38.1	4.22	72.0
LeCA only	23.0	4.14	73.3
Human Reference	–	–	74.7

表 3：制限翻訳タスクの評価結果

結果を見ると LeCA+LevT のアンサンブルモデルが最も良い結果となっていることがわかる。しかし Accuracy を詳細に分析すると、このモデルは評価が最も良い出力の割合が最も高いのと同時に、最も悪い出力の割合も最も高いことがわかった。評価が最も悪い出力を見ると、制限用語以外の部分が抜け落ちてしまっていることがほとんどであった。final のスコアではこの訳抜け問題を適切に捉えられておらず、今後改良が必要であることがわかった。また、翻訳の制限を満たしつつ、高品質な翻訳を生成することにおいてはまだ課題が残されていることも明確となった。

2.2 対訳コーパスフィルタリングタスク

対訳コーパスフィルタリングタスクは今年度から新たに追加されたタスクである。本タスクは、オーガナイザーから提供されるノイズを含む大規模対訳コーパスから、機械翻訳の訓練に有害となりそうな対訳文を除外することが目標である。なお文ペアを除外するのみで、文を改変したり追加したりすることはできない。参加者はノイズ除去済みの対訳コーパスを提出し、全チーム全く同じ設定で NMT の訓練を行い、テストデータでの翻訳精度を競う。世界最大の機械翻訳ワークショップである WMT では 2018 年より同様のタスクを実施しているが、日英の言語対で行われるのは今回が初である。今回は NTT が一般公開している JParaCrawl v3.0[5] （2570 万文対）を対象としてタスクが実施され、科学技術論文コーパスである ASPEC[6]をテストデータとして評価が行われた。

[2] https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html

本タスクには 1 チームから 2 件の提出があった。このチームは feature decay algorithms (FDA)によるフィルタリングと、NMT モデルが出力する翻訳確率に基づくフィルタリングの 2 種類の結果を提出した。どちらも元の 2570 万文対から 500 万文対まで対訳コーパスの量を減らしていた。表 4 に自動評価と人手評価の結果を示す。ベースラインは元の対訳文を全て使ってモデルの訓練を行なっている。なお自動評価において NMT の翻訳確率に基づくフィルタリングの結果がベースラインよりも低かったため、人手評価は FDA によるフィルタリングに対してのみおこなった。

En-Ja Team	BLEU	Human Eval.
sakura (FDA)	**28.8**	**4.31**
baseline	27.4	4.18
sakura (NMT Prob.)	26.7	—
Ja-En **Team**	**final**	**Human Eval.**
sakura (FDA)	**21.8**	**4.49**
sakura (NMT Prob.)	19.9	—
baseline	19.4	4.35

表 4：パラレルコーパスフィルタリングタスクの評価結果

結果から、対訳コーパスを元のサイズの 20%にまで削減したにもかかわらず、どちらの言語対においてもベースラインよりも良い翻訳精度であることがわかる。ここから、闇雲に対訳コーパスを増やすのではなく、ノイズを除去し、翻訳したい文のドメインに合わせて適切な対訳文を用いてモデルの訓練を行うことが重要であることがわかる。

3. まとめ

本稿では WAT2022 全体の概要と、制限翻訳タスクおよびパラレルコーパスフィルタリングタスクの結果を報告した。アジアの翻訳研究の活性化、データ整備等を目的として 2014 年に始めた WAT は、ドメイン数や言語数の増加、参加者数の増加など一定程度の成果を得ており、WAT を通じてアジア地域の機械翻訳研究コミュニティーの連携等が行えると良いと考えている。

NMT の研究成果は毎年山のように報告されており、翻訳精度は年々向上しているが、実利用において課題となる訳抜け、過剰訳、訳語の一貫性、長文の対応など未解決の問題はまだまだ多く残されている。WAT では今後も機械翻訳の実利用を考えた上での課題解決につながるような取り組みを継続していきたいと考えている。

WAT は今後も継続して開催予定であるが、来年度の開催予定は未定である。また WAT では翻訳評価にかかる費用等のためのスポンサーを募集しているため、興味のある方はご連絡いただければ幸いである。

WAT2022 では残念ながら過去の WAT と比較して参加チームが少なかった。参加が少なかった要因を分析し、来年度はより多くの参加者が集まるよう、工夫していきたいと思う。

参考文献

[1] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondˇrej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In Proceedings of the 9th Workshop on Asian Translation (WAT2022), Gyeongju, Republic of Korea. Association for Computational Linguistics.

[2] Christian Federmann. 2018. Appraise evaluation framework for machine translation. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

[3] Guanhua Chen, Yun Chen, Yong Wang, and Victor O.K. Li. 2020. Lexical-constraint-aware

neural machine translation via data augmentation. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pages 3587–3593. International Joint Conferences on Artificial Intelligence Organization. Main track.

[4] Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with Levenshtein transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3536–3543, Online. Association for Computational Linguistics.

[5] Morishita, Makoto, Suzuki, Jun, Nagata, Masaaki, 2020. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus, in Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association, pp. 3603-3609.

[6] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, Hitoshi Isahara, 2016. ASPEC: Asian Scientific Paper Excerpt Corpus, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC2016). European Language Resources Association (ELRA), Portorož, Slovenia, pp. 2204-2208.

法人会員 PR

Web データ自動対訳作成サービス「CorpusNow」のご紹介

株式会社川村インターナショナル

1. 言語資産の英知と人の創造力をつなぐ

株式会社川村インターナショナルは創業以来、35 年以上にわたり、高品質な翻訳サービスの提供にこだわりを持ち続けています。翻訳そのものの品質はもちろんのこと、翻訳サービスの QCD をトータルで高品質化するため、翻訳に付随する作業の効率化や新たな翻訳サービスの提供にも積極的に挑戦しています。

弊社が目指すのは、翻訳を必要とするすべての方に変革（DX）をもたらすことです。そのため、社内に開発部門を擁し、長年培った翻訳のノウハウを最新の IT 技術と組み合わせることで、機械翻訳や AI の有効活用、自動化・可視化の推進、言語資産の活用といった、人とデータ・情報基盤を結び付ける複合的なソリューションを提供し続けています。

2. Web データ自動対訳作成サービス

CorpusNow は、インターネット上の多言語 Web サイトから、独自の手法で効率的に対訳データを作成するサービスです。Web サイトからテキスト情報をスクレイピング（※）し、原文と訳文を対のデータとして整形します。

従来の方法では、Web サイトの情報から対訳データを作成するには、サイトの選定・データの抽出・照合・整形など、複数の工程を個別に、かつ多くの場合は手動で実行する必要がありました。CorpusNow では、これらの処理を自動化し、独自に効率化することで、より早く、安く対訳データをお客様にご提供できるようになりました。

※スクレイピング：特定のデータから余分な情報を除去し、必要な情報のみを抽出すること。

CorpusNow は、下記のような課題をお持ちの方にお勧めのサービスです。

・翻訳メモリを作成して翻訳のコストを削減したい

・自社サイトを活用して自社専用の機械翻訳エンジンを作りたい

・多言語データを解析してグローバル市場の動向を把握したい

詳細ページ：https://www.k-intl.co.jp/corpusnow

CorpusNow ロゴ

3. CorpusNow の 3 つの特長

1 対象 URL を指定するだけ

データを取得したい原文と訳文のそれぞれの Web サイトの URL を指定するだけで、対訳データが作成されます。

2 作成データの品質を自動評価

独自の品質評価ロジックにより、一定の品質基準を上回る対訳データのみが保存されます。作成した対訳データの正誤を自動で評価できるため、人が品質を評価する場合と比較して、大幅に低い費用で精度の高い対訳データを作成できます。

3 追加のデータ加工も可能

対訳データは、将来の翻訳費用の低減を可能にする翻訳メモリや、機械翻訳エンジンをトレーニングするための教師データに加工できます。また、自然言語処

Introduction to CorpusNow
Kawamura International Co., Ltd.

理などの解析処理（形態素解析等）をご希望の場合など、お客様のご要望に応じた形式でデータを納品します。

4. ご利用されたお客様の声

「特許関連文書は一文が長く専門的なため、元々の文章も大変難解です。その文章を機械翻訳で訳出すると、少し不自然な翻訳になることもあり、訳文の修正にそれなりの時間がかかってしまっていました。

エンジンカスタマイズにより、弊社出願の技術分野に特化した自然な訳出結果が出れば…という気持ちで川村インターナショナルさんのオンラインセミナーに参加しました。

セミナーを伺ったところ、社内にある既存出願の特許公報でエンジンカスタマイズができそうだとわかりましたので、早速トライアルを試してみることにしました。

カスタマイズには日本語と英語 1 文が対になっているペアが 10 万ペアあると望ましいと伺いました。ファイル単位の公報ペアはあっても、1 文ずつを対にするというのは人手でやると途方もない時間がかかります。川村インターナショナルさんはその作業を高精度かつ自動で行う「CorpusNow」というサービスがあり、そちらを使って教師データ用の言語資産作成をサポートいただきました。翻訳の精度が上がったと実感できるまでサポートいただけましたので、スムーズに導入の運びとなりました。」

（三井・ケマーズ フロロプロダクツ株式会社様　導入事例ページより抜粋：https://ldxlab.io/case_MT03）

5. 言語資産の活用分野

川村インターナショナルの AI・アノテーションのサービス（用語集/翻訳メモリ/データ作成）を活用すれば、Word/Excel/PowerPoint/PDF をはじめ、紙文書や画像、動画など様々な形式のデータも、有益な言語資産に変換できます。言語資産となったデータは、翻訳の QCD の最適化に役立ちます。さらには、弊社の機械翻訳ソリューション「XMAT®（トランスマット）」の追加機能「LAC（Language Asset Creator）」において、自社に特化した機械翻訳エンジンのカスタムモデルの作成に活用することも可能です。

今後もこのような潜在的なご要望に応えるソリューションを開発・提供してまいります。皆様どうぞお気軽にお問い合わせください。

お問い合わせはこちら：https://ldxlab.io/contact

編集後記

編集後記

新田 順也
AAMT 編集委員会

去る 9 月に AAMT の委員会が、MT リテラシーを解説した MT ユーザーガイドを公開しました。10 月に開催された第 31 回 JTF 翻訳祭 2022 では、山田優様による MT ユーザーガイドに関するセッションがありました。これを機に、MT の適切な運用に向けて関係者の議論の場が増えていくことを願います。

巻頭言では、出内将夫様から「MT が浸透した世の中へ」という題で、MT の技術開発の歴史を踏まえ MT の浸透について考察をいただきました。

第 17 回 AAMT 長尾賞受賞者の後藤功雄様からは「ニュースを対象とした日英機械翻訳システムの研究開発」という題で、英語ニュース特有の課題への解決手法を紹介いただきました。

第 9 回 AAMT 長尾賞学生奨励賞受賞者の森下睦様と美野秀弥様からも寄稿いただきました。森下睦様からは、「JParaCrawl v3.0: 大規模日英対訳コーパス」という題で、対訳コーパスの精度向上および量の増加の方法を紹介いただきました。

同賞受賞者の美野秀弥様からは、「学習時と推論時における入力データの分布の異なりを考慮したニューラル機械翻訳モデルの学習手法」という題で、複数の特徴を区別して学習できるドメインタグを用いた学習手法と、文脈情報に前文の機械翻訳結果を用いる文脈考慮型ニューラル機械翻訳の学習手法を紹介いただきました。

名和大輔様からは、「特許庁における機械翻訳に関する取組」という題で、日英機械翻訳と中日機械翻訳における対訳コーパスの拡充による翻訳品質向上の取り組みを紹介いただきました。

影浦峡様からは、「人間の翻訳と機械の翻訳（6）：翻訳中核プロセスの一つの性質」という題で、翻訳プロセスにおける理解に関する考察をいただきました。

本号から始まった「温故知新」では、今から 30 年前の 1991 年 7 月と 9 月号の JAMT ジャーナル（AAMT の前身である JAMT（日本機械翻訳協会）の会誌）から 2 つの記事を紹介しています。当時の JAMT 会長である長尾真様の「日本機械翻訳協会設立にあたって」では、機械翻訳に対する誤解や混乱を解決するために、立場の異なる関係者が現状を認識する必要性が提唱されています。井佐原均様・坂間保雄様・信田恵壱様の「MT SUMMIT-III 開催される」では、カンファレンスでの論文発表やパネル討論の内容（機械翻訳の改良手法や評価手法）に加え、ポスターや展示の内容も紹介されております。

田中英輝様からは、「AMTA2022 参加報告」という題で、米国フロリダ州で開催されたカンファレンスの概要と、チュートリアル内容、3 件の発表内容を報告いただきました。

平岡裕資様からも、「AMTA 2022 Conference 参加報告」という題で、同カンファレンスでの、人間翻訳およびポストエディットの翻訳プロセスに関する研究のワークショップの報告をいただきました。

中澤敏明様からは、「第 9 回アジア翻訳ワークショップ（WAT2022）開催報告」という題で、WAT2022 の概要として 4 件の研究論文、招待講演、2 件の Shared Task の紹介をしていただきました。

法人会員の PR 記事として、株式会社川村インターナショナル様から「Web データ自動対訳作成サービス『CorpusNow』のご紹介」という題で寄稿をいただきました。

12 月 1 日には、「AAMT 2022, Tokyo」が現地とオンラインとのハイブリッドで開催されます。現地開催は 3 年ぶりです。みなさまに直接お目にかかり情報交換をできることを楽しみにしています。

Editor's note
Junya Nitta
AAMT Editorial Board

AAMTジャーナル「機械翻訳」No. 77

【 発 行 日 】2022年12月15日
【 発 行 】アジア太平洋機械翻訳協会（AAMT）
ホームページ：https://aamt.info/
【 住 所 】〒619-0289
京都府相楽郡精華町光台3-5
国立研究開発法人情報通信研究機構　先進的翻訳技術研究室内
【編集委員会】内山将夫　後藤功雄　中澤敏明　新田順也　園尾聡　森口功造　隅田英一郎
山田優　石川弘美　早川威士　出内将夫
【表紙デザイン】泉谷東十郎
【 題 字 】長尾真
【 事 務 局 】石川弘美
【 印 刷 所 】株式会社 プリントパック

Asia-Pacific Association for Machine Translation (AAMT)
c/o National Institute of Information and Communications Technology
3-5 Hikaridai Seika-cho Soraku-gun Kyoto 619-0289, Japan


AAMT
Asia-Pacific Association for Machine Translation
アジア太平洋機械翻訳協会