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

Research and Development of Japanese–English Machine Translation System for News

日本放送協会　放送技術研究所スマートプロダクション研究部主任研究員
後藤　功雄

2014 年京都大学大学院情報学研究科知能情報学専攻博士課程修了。博士（情報学）。ATR、情報通信研究機構への出向を経て、現在日本放送協会放送技術研究所スマートプロダクション研究部主任研究員。自然言語処理の研究に従事。

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。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。[1]

2　NHK での機械翻訳システムの利用

NHK 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[2]、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。

1　本稿は AAMT Journal「機械翻訳」No.77 に掲載の「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（CC BY-SA 4.0）[23] を基に一部改訂したものである。

2　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。

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）

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。日英で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と図 1 の例での該当箇所を示す。

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）[3] 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。

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。

3　ニュース制作者により明示されることが多い。

②　短く、シンプルに書く

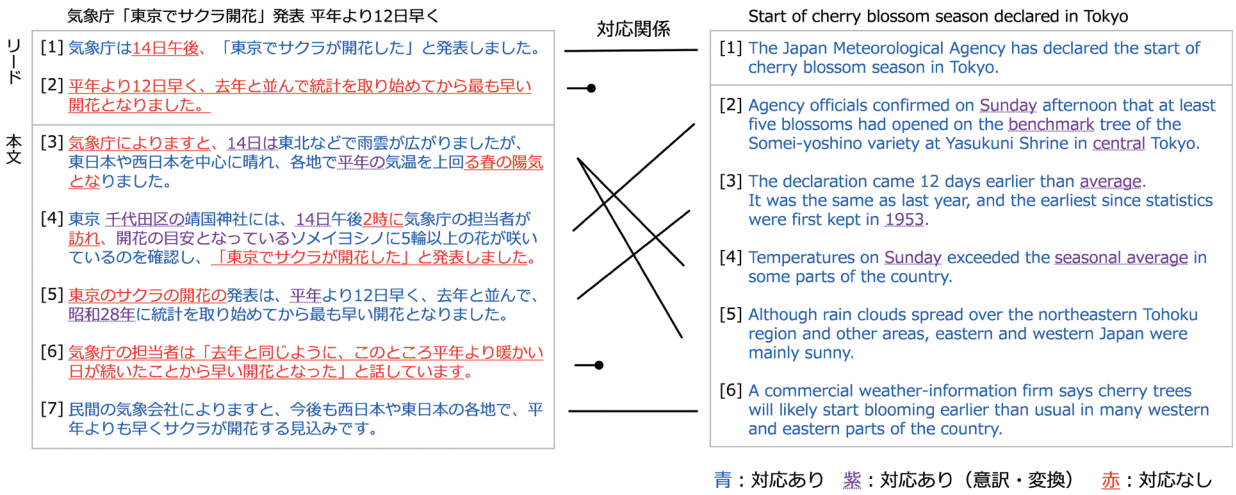
同じ情報の重複がない（冗長でない）。詳細は重要度により省略することもある。

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) の例 1 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。

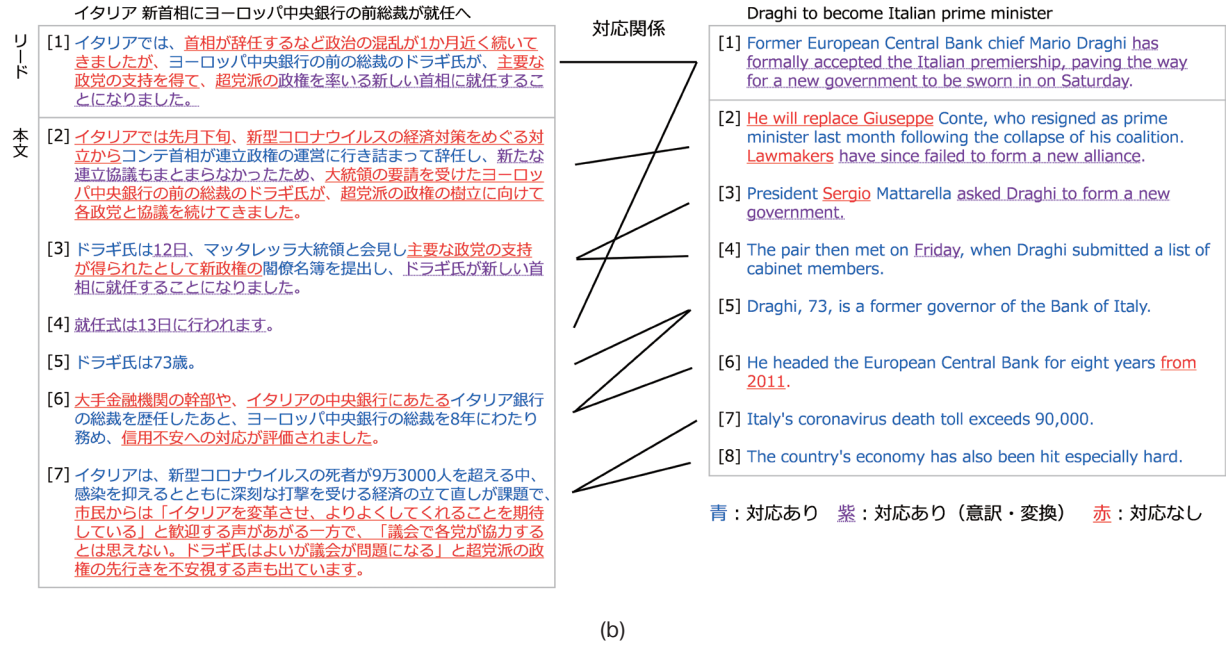
③　同じ表現の繰り返しを避ける

日本語ニュースでは同じ表現の繰り返しがしばしば多く出現する。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。

これら以外の違いとして、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。また、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。



(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

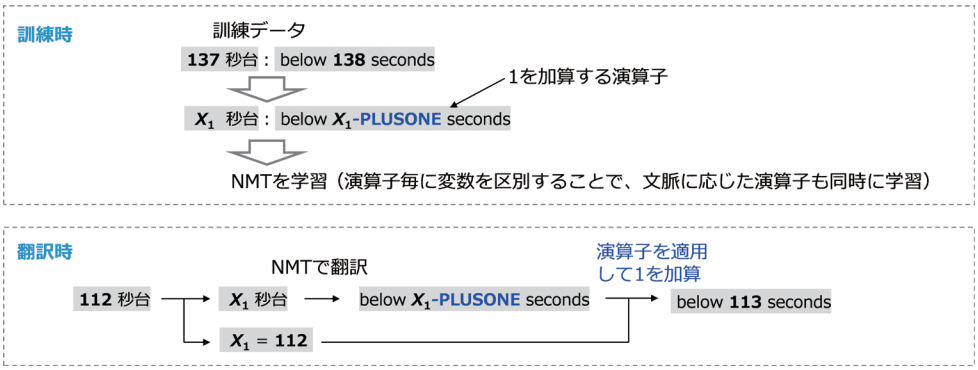


図２　目的言語文脈に依存する値の変化への対応

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]。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、NHK では 2017 年から NMT を活用して、ニュースを対象とした日英機械翻訳の研究開発と検証を進めている。

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、英語ニュース制作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.

[23] 後藤 功雄 , 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，AAMT Journal「機械翻訳」No.77，2022，p. 4-10.