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

ＮＨＫ日英機械翻訳開発プロジェクト
後藤功雄

2022/9/28　第1回AAMTセミナー　第17回AAMT長尾賞受賞記念講演

NHK

ＮＨＫ日英機械翻訳開発プロジェクト

- 目的
  - 英語ニュースの制作を支援する日英機械翻訳の開発
- メンバー
  - 放送技術研究所 機械翻訳グループ
  - 国際放送局 担当者
  - 技術局 担当者
  - 報道局 担当者

NHKの英語ニュース配信

- テレビ・ラジオ ニュース国際放送
- インターネットサービス
  - NHKワールド JAPANのNews
  - スマホアプリによる通知
  - **英語字幕**付き総合テレビ特設ニュースのライブ配信

制作支援に機械翻訳を利用

英語ニュース原稿を利用

日本語字幕を機械翻訳で英訳

- テレビニュース　二カ国語放送　英語副音声

英語ニュース制作支援での利用

制作工程

日本語ニュース

下訳などに機械翻訳を利用

↓ ライターが**英語原稿を作成**
（この工程が一番時間が必要）

英語ニュース

↓ デスクが確認

↓ ネイティブのリライターが修正

↓ デスクが確認

英語ニュース

国際放送局担当者からのコメント

2019年台風19号の際、 機械翻訳の利用で通常より1.5～2倍の出稿数

短時間で出稿できることで、刻々と変わる状況をきめ細かく出せる

泊まり、土日のライターが手薄で多くの出稿が必要になった時に助かる

特設ニュースの英語字幕ライブ配信での利用

（2022年6月開始）

- 総合テレビ特設ニュースにＡＩ翻訳の英語字幕を付けてインターネットでライブ配信
- 総合テレビの日本語字幕データをＡＩ翻訳で英訳し、英語字幕として利用
- 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
- 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

ＡＩ翻訳による英語字幕

英語ニュース原稿の一般的な特徴（一部）

[Papper, 2020; David and Peter, 2019; Strunk and White, 1999]

① 逆ピラミッド型

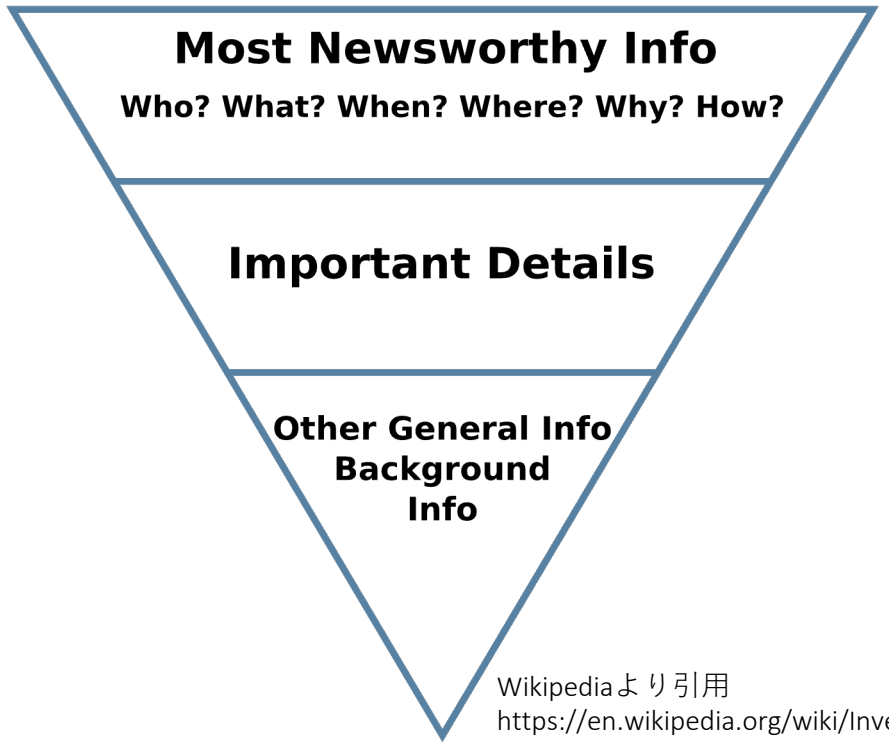
最初に、ニュースで最も重要なことを伝え、
つぎに、重要なことの詳細を伝え、
その後、その他の背景情報などを伝える

② 短く、シンプル

同じ情報の重複がない。

③ 同じ表現の繰り返しを避ける

etc…



Wikipediaより引用
https://en.wikipedia.org/wiki/Inverted_pyramid_(journalism)

英語ニュースは日本語ニュースの単なる翻訳ではない

NHK日英ニュースの違いの例

	気象庁「東京でサクラ開花」発表 平年より12日早く	対応関係	Start of cherry blossom season declared in Tokyo
リード	[1] 気象庁は14日午後、「東京でサクラが開花した」と発表しました。		[1] The Japan Meteorological Agency has declared the start of cherry blossom season in Tokyo.
リード	[2] 平年より12日早く、去年と並んで統計を取り始めてから最も早い開花となりました。		[2] Agency officials confirmed on Sunday afternoon that at least five blossoms had opened on the benchmark tree of the Somei-yoshino variety at Yasukuni Shrine in central Tokyo.
本文	[3] 気象庁によりますと、14日は東北などで雨雲が広がりましたが、東日本や西日本を中心に晴れ、各地で平年の気温を上回る春の陽気となりました。		[3] The declaration came 12 days earlier than average. It was the same as last year, and the earliest since statistics were first kept in 1953.
本文	[4] 東京 千代田区の靖国神社には、14日午後2時に気象庁の担当者が訪れ、開花の目安となっているソメイヨシノに5輪以上の花が咲いているのを確認し、「東京でサクラが開花した」と発表しました。		[4] Temperatures on Sunday exceeded the seasonal average in some parts of the country.
本文	[5] 東京のサクラの開花の発表は、平年より12日早く、去年と並んで、昭和28年に統計を取り始めてから最も早い開花となりました。		[5] Although rain clouds spread over the northeastern Tohoku region and other areas, eastern and western Japan were mainly sunny.
本文	[6] 気象庁の担当者は「去年と同じように、このところ平年より暖かい日が続いたことから早い開花となった」と話しています。		[6] A commercial weather-information firm says cherry trees will likely start blooming earlier than usual in many western and eastern parts of the country.
本文	[7] 民間の気象会社によりますと、今後も西日本や東日本の各地で、平年よりも早くサクラが開花する見込みです。		

青：対応あり　紫：対応あり（意訳・変換）　赤：対応なし

リードの違い

東京で桜が開花した
去年同様最も早い開花

リード

[1] 気象庁は14日午後、「東京でサクラが開花した」と発表しました。

[2] 平年より12日早く、去年と並んで統計を取り始めてから最も早い開花となりました。

本文

[3] 気象庁によ…ましたが、東日本や西…回る春の陽気となりまし…

リード：全体の概要

[4] 東京 千代田…の担当者が訪れ、開花の目安となっている…の花が咲いているのを確認し、「東京でサクラが開花した」と発表しました。

[5] 東京のサクラの開花の発表は、平年より12日早く、去年と並んで、昭和28年に統計を取り始めてから最も早い開花となりました。

[6] 気象庁の担当者は「去年と同じように、このところ平年より暖かい日が続いたことから早い開花となった」と話しています。

[7] 民間の気象会社によりますと、今後も西日本や東日本の各地で、平年よりも早くサクラが開花する見込みです。

対応関係

東京で桜が開花した

Start of c…rry blossom season declared in Tokyo

[1] The Japan Meteorological Agency has declared the start of cherry blossom season in Tokyo.

[2] Agency…n that at least five blo…ee of the Somei-…ntral Tokyo.

リード：ニュースで最も重要なことだけ

[3] The de…age. It was …nce statistics were first kept in 1953.

[4] Temperatures on Sunday exceeded the seasonal average in some parts of the country.

[5] Although rain clouds spread over the northeastern Tohoku region and other areas, eastern and western Japan were mainly sunny.

[6] A commercial weather-information firm says cherry trees will likely start blooming earlier than usual in many western and eastern parts of the country.

青：対応あり　紫：対応あり（意訳・変換）　赤：対応なし

リードの次の文の違い

東京で桜が開花した
去年同様最も早い開花

各地で暖かい気温

東京で桜が開花した

靖国神社で5輪以上咲いた

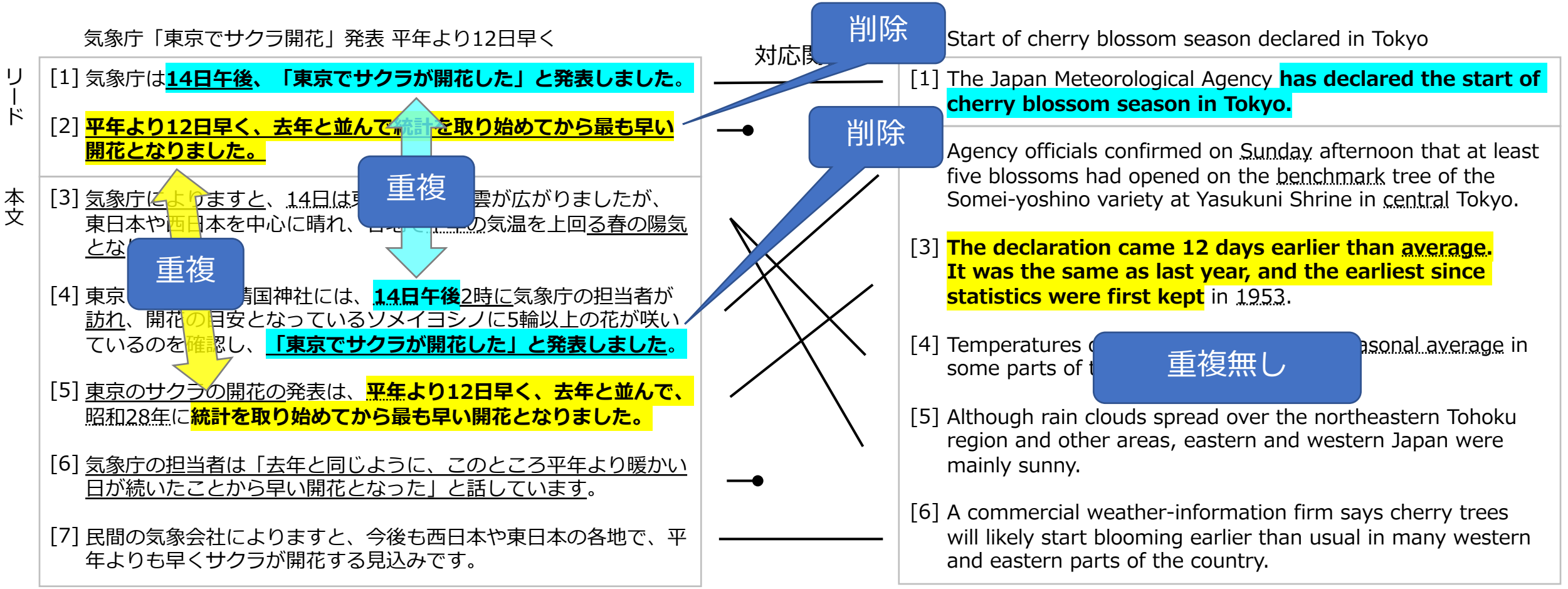
		対応関係	
リード	[1] 気…は14日午後、「東京でサクラが開花した」と発表し…		[1] The Japan Meteorological Agency has declared … of cherry blossom season in Tokyo.
	[2] 平年より12日早く、去年と並んで統計を取り始めてから最…い開花となりました。		[2] Agency officials confirmed on Sunday afternoon that at least five blossoms had opened on the benchmark tree of the Somei-yoshino variety at Yasukuni Shrine in central Tokyo.
本文	[3] 気象庁によりますと、14日は東北などで雨雲が広がりましたが、東日本や西日本を中心に晴れ、各地で平年の気温を上回る春の陽気となりました。	順番の変更	[3] The decla… It was the… were first… …tistics
	[4] 東京 千代…の担当者が訪れ、開花…の花が咲いているの…表しました。		[4] Temperat… …ge in some parts…
	[5] 東京のサク…年と並んで、昭和28年に統計を取り始めてから最も早い開花となりました。		[5] Although rain clouds spread over the northeastern Tohoku region and other areas, eastern and western Japan were mainly sunny.
	[6] 気象庁の担当者は「去年と同じように、このところ平年より暖かい日が続いたことから早い開花となった」と話しています。		[6] A commercial weather-information firm says cherry trees will likely start blooming earlier than usual in many western and eastern parts of the country.
	[7] 民間の気象会社によりますと、今後も西日本や東日本の各地で、平年よりも早くサクラが開花する見込みです。		

リードの次の文：
背景情報

リードの次の文：
リードで伝えた内容を詳しく

青：対応あり　紫：対応あり（意訳・変換）　赤：対応なし

リードと本文との重複の違い



同じ表現の繰り返しの違い

気象庁「東京でサクラ開花」発表 平年より12日早く

リード

[1] 気象庁は14日午後、「東京でサクラが開花した」と発表しました。

[2] 平年より12日早く、去年と並んで統計を取り始めてから最も早い開花となりました。

本文

[3] 気象庁によりますと、14日は東北などで雨雲が広がりましたが、東日本や西日本を中心に晴れ、各地で平年の気温を上回る春の陽気となりました。

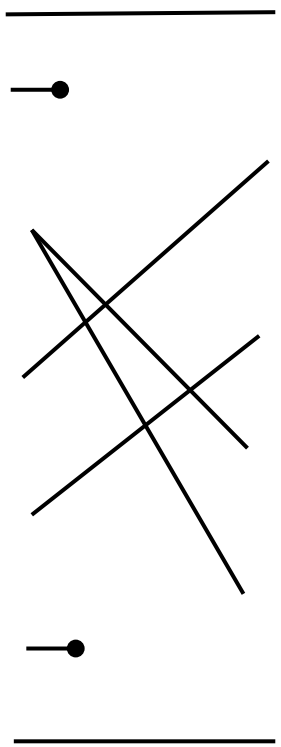
[4] 東京 千代田区の靖国神社には、14日午後2時に気象庁の担当者が訪れ、開花の目安となっているソメイヨシノに5輪以上の花が咲いているのを確認し、「東京でサクラが開花した」と発表しました。

[5] 東京のサクラの開花の発表は、平年より12日早く、去年と並んで、昭和28年に統計を取り始めてから最も早い開花となりました。

[6] 気象庁の担当者は「去年と同じように、このところ平年より暖かい日が続いたことから早い開花となった」と話しています。

[7] 民間の気象会社によりますと、今後も西日本や東日本の各地で、平年よりも早くサクラが開花する見込みです。

対応関係



Start of cherry blossom season declared in Tokyo

[1] The Japan Meteorological Agency has declared the start of cherry blossom season in Tokyo.

[2] Agency officials confirmed on Sunday afternoon that at least five blossoms had opened on the benchmark tree of the Somei-yoshino variety at Yasukuni Shrine in central Tokyo.

[3] The declaration came 12 days earlier than average.
It was the same as last year, and the earliest since statistics were first kept in 1953.

[4] Temperatures on Sunday exceeded the seasonal average in some parts of the country.

[5] Although rain clouds spread over the northeastern Tohoku region and other areas, eastern and western Japan were mainly sunny.

[6] A commercial weather-information firm says cherry trees will likely start blooming earlier than usual in many western and eastern parts of the country.

繰り返しが多い

繰り返しを避ける

英語ニュース制作支援の目標

- 現段階：内容を正確に翻訳する
- 今後　：日本語ニュースから英語ニュースを生成する

日英機械翻訳の研究開発

- 日英ニュース記事は文レベルでは正確な対訳になっていないものが多い
- 新聞・通信社の日英報道記事もある程度同様の傾向

（訳抜けを多く含む対訳データ → これで学習すると訳抜けが頻出）

ニュース分野の日英対訳で高品質な対訳文対が存在しない

高性能な機械翻訳に最も重要なものは高品質な対訳データ
→　**人手でNHKニュースの高品質な対訳データを構築**

ニュースの日英対訳文の構築

- NHK日本語ニュース文 → 人手英訳
- NHK日本語ニュース文 → 日英機械翻訳 → 人手ポストエディット
- NHK英語ニュース文　→ 英日機械翻訳 → 人手ポストエディット

合計100万文対を構築

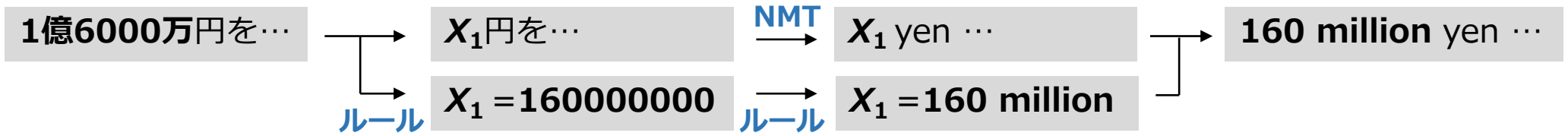
NHKニュースの翻訳品質が向上

ニュース翻訳の課題は対訳文とNMTだけで解決できるわけではない

数字の翻訳

- ニュースで数字は重要
- 数の単位の変換が必要：＜例＞1億6000万 → 160 million
  - NMTで単語列として翻訳すると間違える時がある

- ルールベースとの融合（大きな数字・小数点を含む数字）



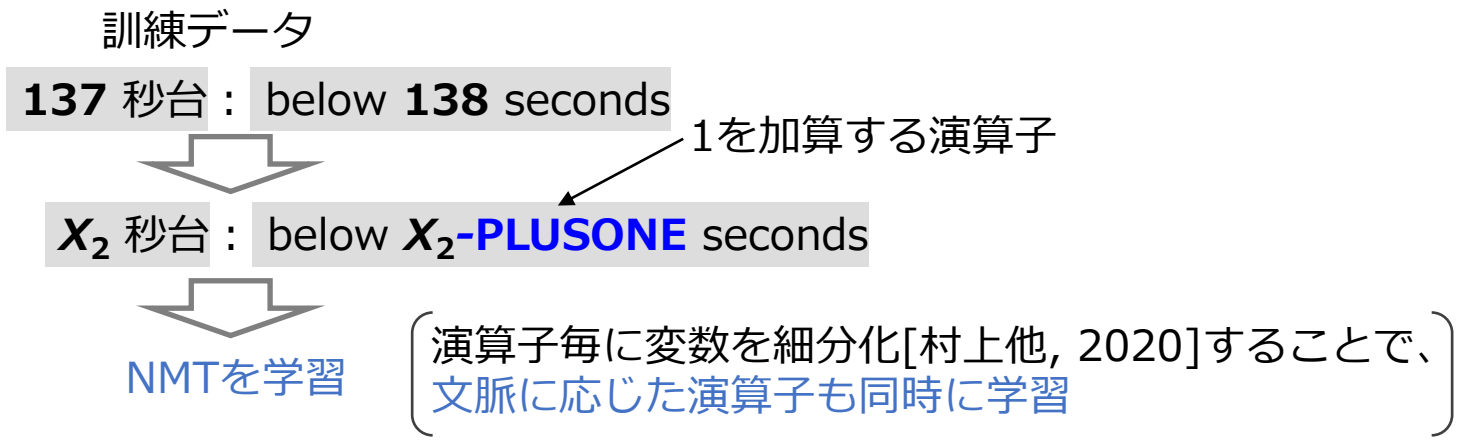
- ニュース特有の表現への対応
  - 風速：メートル／秒 → キロメートル／時に変換

＜例＞風速30メートル → Winds of 108 kilometers per hour

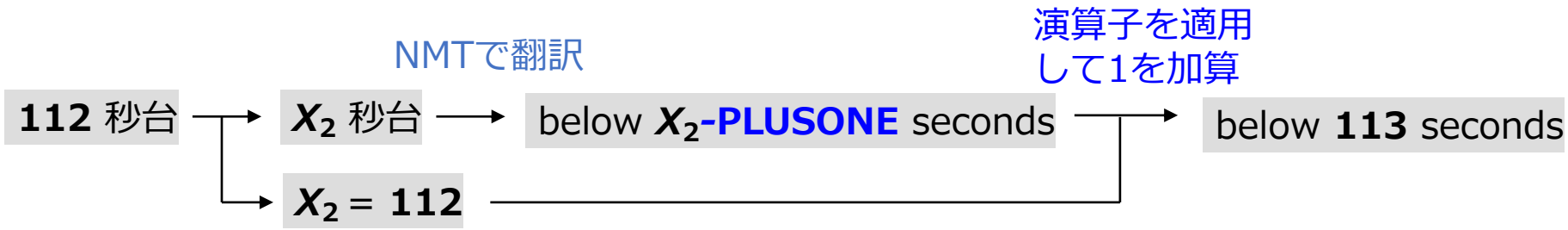
数字の翻訳（文脈依存）

- 目的言語文脈に依存する値の変化への対応
  <例> **112** 秒台 → below **113** seconds

学習時



翻訳時



日付表現の翻訳 [Kinugawa et al., 2022]

前処理で曜日に置換
もしくは月を挿入

曜日への置換や月を挿入→誤訳

日本語表現の日付・期間の曖昧性を自動識別

日英対訳文を用いて、
期間と日付で表層が異なる英語表現から
日本語日付表現の日付・期間の分類を特定

その分類結果を学習して識別器を構築

ユーザー辞書機能の導入

- 辞書機能として、相補関係にある2つの手法を組み合わせて利用
  - 入力に訳を挿入することで出力を制御[Dinu et al., 2019]

    <入力例>　自由民主党は・・・

    → <t> 自由民主党 <d/> Liberal Democratic Party of Japan </t> は・・・

    語彙の情報を翻訳で利用できる。
    頻度ゼロの語と未学習の語には対応できない。低頻度でも出力されない場合あり

    ⇩

    学習データで高頻度語を対象

  - 該当表現を変数に置換して翻訳し、出力中の変数を訳語に置換[Long et al., 2016]

    <入力例>　ザポリージャ原発で・・・　→　<TERM$_1$>に・・・

    <出力例>　・・・at <TERM$_1$>　→　・・・at Zaporizhzhia Nuclear Power Plant

    低頻度語、登録直後の表現を対象

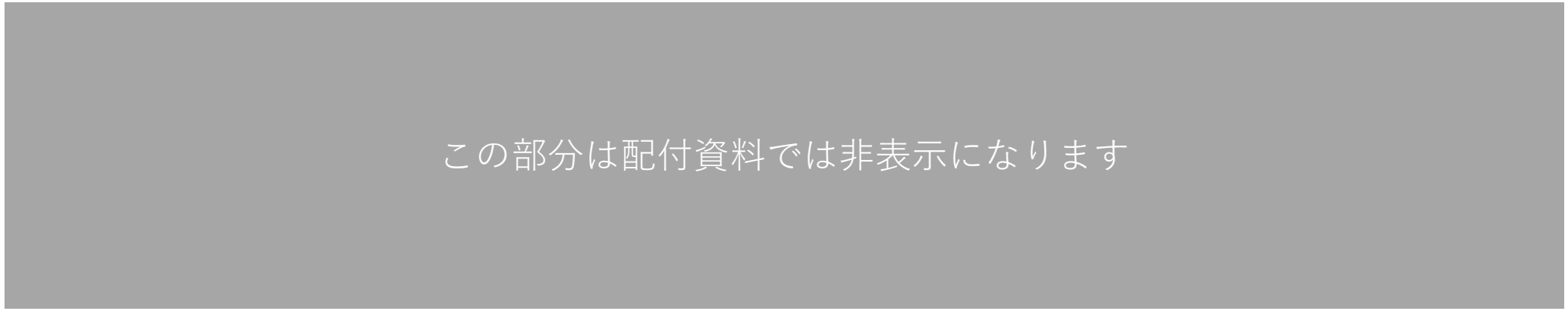
文脈の利用（主語補完）(1/2) [Goto et al., 2020]

- 既存の文脈利用手法（2-to-1） [Tiedemann and Scherrer, 2017]
  - 学習時と翻訳時に、直前の文を区切り記号(<delimiter>)と共に翻訳元文に追加

機械翻訳への入力

文脈

翻訳元文



- 機械翻訳の学習の枠組み（正解訳からの誤差逆伝播）だけでは、多くの文脈の情報から必要な情報を推測して、そこを活用するように学習することは困難
- 2文以上前の文脈を利用できない

文脈の利用（主語補完）(2/2) [Goto et al., 2020]

文脈

この部分は配付資料では非表示になります

文脈を述語・項構造解析し、主語・主題を抽出

配付資料では
非表示になります

翻訳元文に追加（直前の文に主語・主題がない場合はそれ以前の文から抽出して追加）

機械翻訳
への入力

この部分は配付資料では非表示になります

2-to-1では学習が難しい部分を構造解析に基づいて実施

文脈の利用（目的言語文脈）[Mino et al., 2020]

- 目的言語側文脈
  - 英語側の一貫性や同一表現の繰り返し抑制の制御に必要
- 利用方法
  - 入力文に直前の英語文を追加
- 従来手法[Kim et al., 2019]の課題
  - 文脈として利用する直前の目的言語文の違いによる悪影響
    - [学習時]：訓練データ
    - [翻訳時]：機械翻訳出力（翻訳誤りやtranslationeseあり）
- 対策
  - 学習時に用いる直前の目的言語文：最初は訓練データ、次第に機械翻訳出力を増加
    → 学習時と翻訳時の文脈の違いを低減

日英ニュースの翻訳知識の活用(1/2) [後藤他, 2020]

学習データでの訳抜けは、翻訳時に訳抜けの原因となる

日英ニュースの翻訳知識の活用(2/2) [後藤他, 2020]

ラベルを自動推定　**D**:訳出なし　**T**:訳出あり　**訳出されていない内容語をラベルで区別して学習**

学習

ラベル：　**D　T　D　T　T　T　T　T　T　T　T　T**

訓練データ　参拝　を　終えた　人　たち　は　縁起物　を　買い　まし　た　。

Visitors　bought　mascots　.

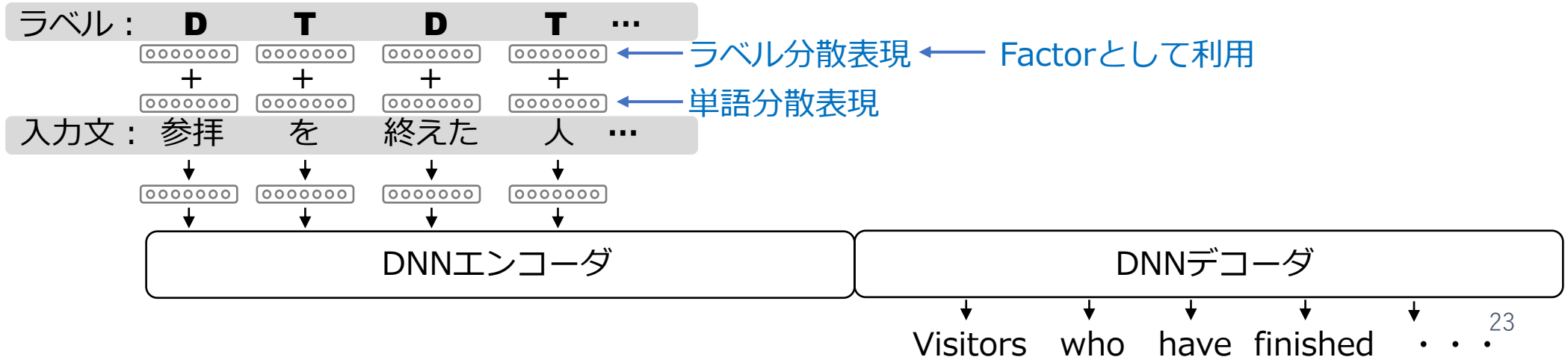
訓練データで訳出されている部分の学習結果を利用して翻訳　➡　**訳出される**

翻訳

ラベル：　**T　T　T　T　T　T　T　T　T　T　T　T**

入力文：　参拝　を　終えた　人　たち　は　お守り　を　買い　まし　た　。

ラベルの利用方法



制作支援向け訳抜け検出 [後藤他, 2018]

入力文

参拝を終えた人たちは、おみくじをひいたり、お守りや魔よけの矢などの縁起物を買い求めたりしていました。

出力文

Visitors can buy arrow amulets, good-luck charms and fortune-telling slips.

訳抜け部分が一目で分かると便利

訳抜け箇所の検出

参拝を終えた人たちは、おみくじをひいたり、お守りや魔よけの矢などの縁起物を買い求めたりしていました。

強制的に逆翻訳

Visitors can buy arrow amulets, good-luck charms and fortune-telling slips.

英語文にない日本語部分の生成確率は低くなる → 訳抜け箇所と推定

まとめと今後の予定

■ まとめ

- NHKでの機械翻訳の利用、英語ニュースの特徴、機械翻訳の研究開発の取り組みを紹介

■ 今後の予定

- 翻訳機能・制作支援
  - 新しい翻訳知識の自動獲得
  - 翻訳誤りの推定

- 日本語ニュースから英語ニュースの制作自動化
  （翻訳の範囲を超えたタスク）

選考委員の方々をはじめ関係者の皆様、セミナー聴講者の皆様、ありがとうございました。

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

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

[David and Peter, 2019] David Ingram and Peter Henshall, The News Manual, https://www.thenewsmanual.net

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

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

[村上他, 2020] 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.

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

[Dinu et al., 2019] 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, pages 3063–3068, 2019.

[Long et al., 2016] 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 (WAT2016), pages 47–57, 2016.

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

[Goto et al., 2020] Isao Goto, Hideya Mino, Hitoshi Ito, Kazutaka Kinugawa, Ichiro Yamada, and Hideki Tanaka. 2020. 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, pages 72–79.

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

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

[後藤他，2020] 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.

[後藤他，2018] 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.