時事通信社ニュースの日英均衡対訳コーパスの構築–第 1 報

田中 英輝 † 美野 秀弥 †† 後藤 功雄 †† 山田 一郎 ††
川上 貴之 ††† 大嶋 聖一 ††† 朝賀 英裕 †††
†NHK エンジニアリングシステム ††NHK 放送技術研究所 ††† 時事通信社
tanaka.hideki@nes.or.jp
{mino.h-gq, goto.i-es, yamada.i-hy}@nhk.or.jp
{kawakami, sohshima, asaka}@jiji.co.jp

1 はじめに

深層学習を使った機械翻訳の研究プロジェクトが国立研究開発法人情報通信研究機構の委託で進められている[1]．このプロジェクトには現在 3 つの課題があり[2]，その中の 1 つがニュースを対象とした日英機械翻訳システムの研究である（以下,「ニュース翻訳プロジェクト」と呼ぶ）．本稿では同プロジェクトで開発中のニュースの日英均衡対訳コーパスの構築方針と現状を報告する．

機械翻訳の研究に大規模な対訳コーパスが必須であることは言うまでもなく，ニュース翻訳プロジェクトでもアルゴリズムの研究と並行して，時事通信社の日英ニュース記事を対象に対訳コーパスを開発していく方針である．

近年，特許 [1] や科学技術文献 [2] を対象とした大規模な日英対訳コーパスが開発され，評価型ワークショップで活発に利用されている [3, 4]．これらのコーパスは特許や科学技術文献の大規模な日英対訳文書に自動文アラインメント（以下，アラインメントと略記する[3]）を適用して構築されている．さらに最近ではニュースの日英対訳コーパスがアラインメント技術で作られ評価型ワークショップでの利用が開始されている [4]．

アラインメントは，対象の日英文書が「日英の文の順序が同一で，かつ日英の文の情報が同等である時」に最大の効果を生ずる．このような日英文書を本稿では「均衡翻訳」と呼ぶ．しかし，日英文書が均衡でない，すなわち不均衡な場合には，後述するよう，得られる対訳コーパスにノイズが混入する問題や，文書全体の対訳が得られない問題が発生する．時事通信社の日本語記事は高度な編集を経て英訳されているため日英記事は不均衡となり，上述の問題が顕著となる．

そこでニュース翻訳プロジェクトでは，アラインメントによるコーパス開発に加えて，日本語記事のすべての文を過不足なく人手で翻訳することで均衡な日英対訳コーパスを構築する計画を立てた．以下では，時事通信社の日英記事にアラインメントを適用する場合の課題と，著者らが進めている均衡な日英対訳コーパス開発の方針，および現状を報告する．

表 1: コーパス構築に利用中の記事

日本語記事	1,561,143
日英対応記事	57,154

2 時事通信社の日英記事と文アラインメント

コーパス作成の元となる時事通信社の記事の数を表 1 に示す．表 1 より日本語記事のおよそ 3.7%が英訳されていることが分かる．

2.1 時事通信社の日英記事の特徴

図 1 に 2018 年 1 月 4 日の時事通信社の日本語の記事と対応する英語記事を示す．日英記事とも先頭から 1 行づつ表示し，著者が対応すると判断した文同士を線で連結した．なお日本語記事の第 3 文と第 4 文は，

[1]多言語音声翻訳高度化のためのディープラーニング技術の研究開発（課題番号 197）
[2]1) 高度な文脈理解技術（インテリジェント翻訳技術）の研究開発，2) 新語・新トピックへの即時対応技術（ニュース対応翻訳技術）の研究開発，3) マルチモーダル翻訳技術の研究開発
[3]本稿の日英記事は対応済で記事アラインメントは不要である．

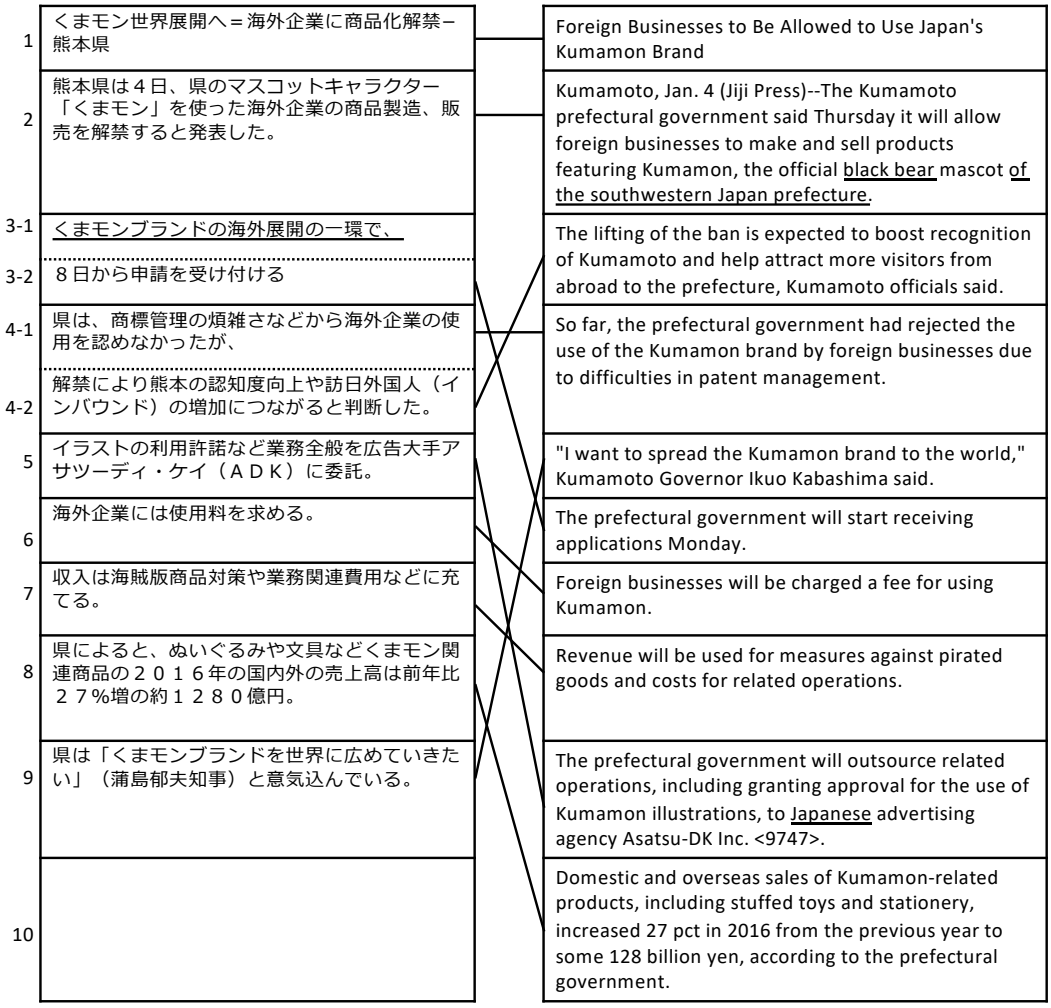

図 1: 日英記事の例と表現対応

節が英語に対応するため文を 2 つの節に分割した．また，日本語，英語の一方のみにある情報に下線を付加した．この図から，以下の全体的な特徴を読み取ることができる．

- 翻訳単位
  1 文対 1 文の翻訳ではない．日本語の第 3 文，第 4 文はそれぞれ 2 つの節に分割され翻訳されている．
- 情報の加除
  日本語の第 3-1 文は訳出されていない．また，英語の第 2 文には「くまモン」の説明が付け加えられ，さらに第 9 文には “Asatsu-Dk” が「日本」の会社であることが付け加えられている．
- 文の順序の変更
  対応の線の交差は英訳時に文の順序が変更されたことを示す．図の例では日本語の第 3-2, 4-2, 5, 9 文の位置が変更されている．

以上の特徴は日英記事が不均衡な対訳であることを示す．その背景には日英ニュースの執筆スタイルの違いの反映や，英語読者への配慮のため高度な編集を加えて英訳している事実がある．同様の特徴は NHK のニュースでも観察されており [5]，ニュースの日英翻訳に広く見られる特徴と考える．

さらに英語記事には以下の特徴が見られる．

- タイトル固有の表記と文体の利用
  表記に大文字小文字が混用される．未来を表す「be to」構文の be が省略されるなどタイトル固有の文体が使われる．
- 日付と曜日の変換

同じ週の日付は曜日に変換される．日本語の第 3-2 文の「8 日」は英語の第 6 文では “Monday” と翻訳されている．

- 表記や用語の統一
  用語集によって使う語や表記が統制されている．英語の第 10 文ではパーセントの訳に “pct” が使われている．

2.2 文アラインメントの課題

現在よく用いられているアラインメント手法では，文などの言語単位に分割された日本語と英語の文書を入力とし，日英の言語単位間が交差しないように結びつけた結果が出力される [1]．言語単位には機械翻訳システムの学習，利用の利便性の面から文が採用されることが多い．以上に基づくアラインメントを不均衡な時事通信社の日英記事に適用すると以下の問題を生ずる．

- ノイズの混入
  図 1 では，日本語の 3-1 や 3-2 のように，文より小さな節が英語の 1 文に翻訳される例を示した．このような翻訳を文単位で対応付けると英語にない日本語の節，すなわちノイズが対訳に混入する．
- 文脈の欠如
  図 1 の線を交差させないためには，一部の線の消去が必要となる．すなわち記事の一部しか対応付けられず，得られる対訳コーパスには文脈情報が完全には反映されないことになる．

3 日英均衡対訳コーパスの構築

上述の問題はあるが，大規模な対訳コーパスを安価に構築できるアラインメントは高い有用性を持つ．そこで著者らは，表 1 の記事のうち，日英対訳記事からはアラインメントで対訳コーパスを作り，また，単独の日本語記事からは人手の翻訳により均衡コーパスを作成することにした．以下，構築方針を説明する．

3.1 記事選択

アラインメントに比べて人手翻訳のコストは高い．このため表 1 の日本語単独記事をすべて翻訳することは困難であり，以下の方針で記事を選択した．

- 記事の期間
  本プロジェクトは 2020 年度まで実施される予定であり，できるだけ新しい 2016 年から 2020 年の 4 年の日本語記事を翻訳対象とした．
- 記事の形式
  時事通信社の記事には金利などの価格を表形式で伝えるもの，人事異動のように，人名が羅列されているものもある．これらは通常の文とは形式が異なるため翻訳対象から除外した．また，英訳される日本語記事の 80%が 5 文から 15 文の範囲にあることから，文数がこの区間に入る記事を選択した．
- 記事の内容
  時事通信社で英訳されている日本語記事は全体の一部である．そこで，日本語記事に与えられているキーワードリストを使い，英訳されている日本語記事のキーワードリストと，日本語記事のキーワードリストの分布の類似性をコサインで表し，分布が極端に異なる日本語記事を排除することにした．

3.2 翻訳の方針

均衡コーパスを作成するため以下の方針で翻訳した．また方針に従って翻訳した例を図 2 に示す．

- 翻訳単位
  日本語の 1 文を 1 文あるいは複数の英文で翻訳する[4]．日本語の複数文をまとめた翻訳はしない．また，日英の文の順序は同一とする．
- 文脈の利用
  翻訳者は記事全体を参照して，文脈を反映して翻訳する．著者らの予備調査によると人手の翻訳も文脈の有無により変化することを確認している．文脈により，省略された主語の補完，主題の捉え方による態の変換（主語の変換）などが発生する．本プロジェクトではこのような現象の機械翻訳も視野に入れ，文脈を反映した翻訳を実施する．図 2 の第 3 文では日本語にない節 “The product can be exported” が補完されている．
- スタイルガイド
  2.1 節で示したようにタイトル，用語は時事通信

[4] 英文が長すぎて不自然な場合には複数文で翻訳した．すなわち文対応は完全に均衡しているわけではない．

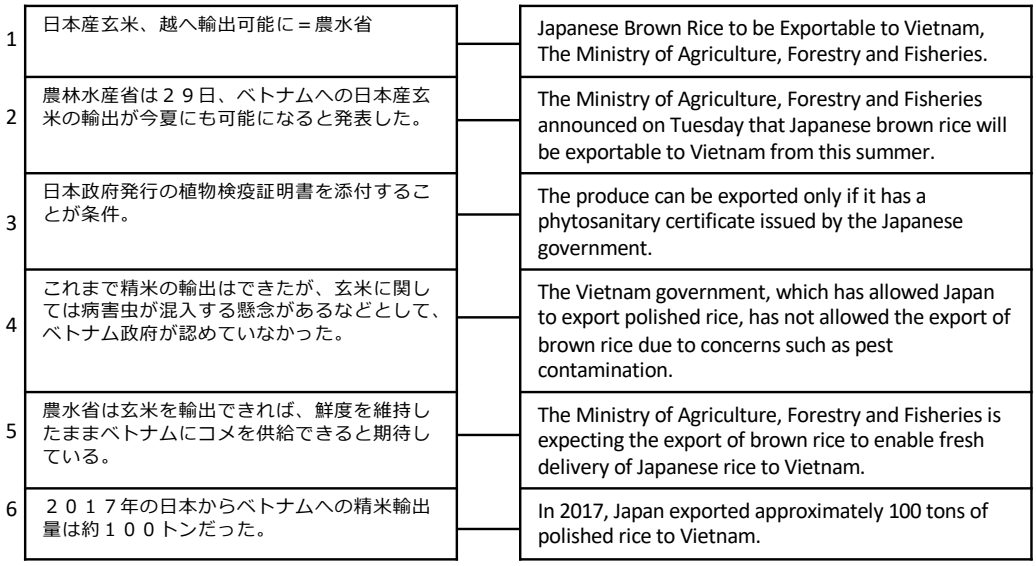

図 2: 均衡翻訳の例

社固有の基準に従っている．アラインメントで作成したコーパスとの併用を考えると，できるだけその基準に従いたいが，時事通信社の翻訳者以外に準拠させるのは難しい．そこで時事通信社のスタイルガイド，さらに記事の観察を加えて作成したスタイルガイドを翻訳会社に提供し，これに準拠するよう依頼した．

以上の方針に従い現在複数の翻訳会社で翻訳を進めている．中間検査をすでに 4 回実施しており，スタイルや翻訳方針の誤解等はほぼなくなったと考える．2018 年度は 20 万文対のコーパスの構築，約 1,000 万文字の翻訳を目標に翻訳を進めており，2018 年度末には目標をほぼ達成できる見込みである．

4　おわりに

本稿ではニュースの均衡した日英対訳コーパス構築の方針と現状について述べた．ニュースの日英均衡対訳コーパスは現時点で存在せず，プロジェクト終了時までに大規模なコーパスが構築できれば機械翻訳の研究を大きく加速できると考える．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究により得られたものです．

参考文献

[1] Masao Utiyama and Hitoshi Isahara. A Japanese-English Patent Parallel Corpus. In *MT Summit XI*, pp. 474–482, 2007.

[2] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian Scientific Paper Excerpt Corpus. In *LREC*, 2016.

[3] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the Patent Translation task at NTCIR-7 Workshop. In *NTCIR-7 Workshop Meeting*, pp. 389–400, 2008.

[4] Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. Overview of the 5th Workshop on Asian Translation, 2018.

[5] 荒牧英治, 黒橋禎夫, 柏岡秀紀, 田中英輝. 用例ベース翻訳のための日英アラインメント確信度語類似度を用いた訳語選択. 自然言語処理, Vol. 11, No. 1, pp. 107–123, 2004.