# 時事通信社ニュースの日英対訳コーパスの構築–第２報

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 ††††

†NHK エンジニアリングシステム †† 東京大学 †††NHK 放送技術研究所 †††† 時事通信社

tanaka.hideki@nes.or.jp, nakazawa@logos.t.u-tokyo.ac.jp
{mino.h-gq, itou.h-ce, goto.i-es, yamada.i-hy}@nhk.or.jp
{kawakami, sohshima, asaka}@jiji.co.jp

## 1 はじめに

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．これに対して著者らは言語資源の開発が遅れている日英ニュースの機械翻訳の研究開発を加速するため，時事通信社のニュースを利用した日英ニュースコーパスの開発を進めている [3].

時事通信社は日本語の記事を作成し，内外の報道機関や一般に提供すると共に，日本語記事の一部を英訳して同様のサービスを行なっている．すなわちコーパス構築の資源としては，英訳を持たない「日本語単独記事」および，日本語記事とその英訳の「日英対応記事」を利用できる．日英対応記事には，日英の内容が完全に等価な文ペアもあるが，多くの文ペアは日英の片側にしかない情報を含む．また，文そのものが片側にしかない場合も多い．この日英間の情報の不均衡のため，日英対応記事をそのまま機械翻訳システムの学習に使うのは難しい．

前回の報告 [3] では，日英対応記事に文アラインメントを適用してコーパスを作成する際，日英の情報不均衡の問題が発生することを述べた．また情報不均衡の問題を軽減するため，文アラインメントコーパスに加えて，別に日本語単独記事の内容を人手で忠実に英訳してコーパスを構築する計画を説明した．

本報告では，文アラインメントによって作成したコーパス，人手の日英翻訳で作成したコーパス，さらに，本稿の新たな試みとして，日英対応記事の日本語記事を英語記事に合わせて修正して作成するコーパスについて説明する．また，翻訳実験によりそれぞれのコーパスの効果を報告する．

## 2 日英ニュースコーパス

本節では時事通信社の「日英対応記事」と「日本語単独記事」を利用して作成している 3 種類の日英ニュースコーパスを説明する．各コーパスの諸元を表 1 にまとめた．

### 2.1 文アラインメントコーパス

2011 年から 2018 年 6 月までの 57,154 本の日英対応記事に文アラインメントアルゴリズム [1] を適用して日英ニュースコーパスを作成した．本稿ではこのコーパスを Align と略記する．著者らは文アラインメント結果から，1 対 1 対応で文類似度が 0.3 以上の文ペアを抽出して Align コーパスを作成した．Align コーパスには以下の特徴がある．

- 情報の均衡性
  日英片方にしかない情報が混入する可能性があり，完全な均衡性は得られない．
- 完全性
  文単位での抽出となり，記事全体の文が抽出されるとは限らず，部分的なコーパスとなる．
- 表現スタイル
  コーパスの日英の文とも実際に出現した原文であり，完全に時事通信社のスタイルに一致する．
- 経済性
  安価で短時間にコーパスを構築できる．

表 1 の Align の諸元は抽出の類似度閾値により変わる．まず，獲得できる文ペアの数が変化する．1 対 1 対応の文ペアの総数は 582K だが，今回の類似度閾値 0.3

表 1: 各コーパスの諸元

| | Align | Manual | Repair |
|---|---|---|---|
| 抽出時期 | 2011-2018 | 2016-2018 | 2018-2019 |
| 文数 | 239,718 | 199,170 | 28,452 |
| 日本語平均長 | 54.5 | 48.9 | 48.3 |
| 英語平均長 | 26.0 | 25.2 | 24.7 |
| 文長相関係数 | 0.599 | 0.896 | 0.868 |

日本語基準文長:47.7 文字, 英語基準文長:25.0 単語

で文数は 240K に縮小した．次に，抽出される文の平均文長が閾値によって変化する．抽出元の 57,154 本の日英記事の平均文長「日本語 47.7 文字，英語 25.0 単語」は翻訳される日英の文の平均長と考えられるが，類似度 0.3 以上に限ることで日本語の平均長は上昇し，54.5 文字となった．さらに，翻訳の均衡性が変化する．本稿では翻訳の均衡性の指標として日英の文長の相関係数を利用する[1]．1 対 1 対応の文ペアの相関係数を調べたところ，文類似度が大きくなるほど相関係数が高くなる傾向が見られた．今回の条件，類似度 0.3 以上で抽出したコーパスの相関係数は 0.599 であり，人手で作成した他の 2 つのコーパスより小さくなっている．

### 2.2 日英翻訳コーパス

文アラインメントによるコーパス作成では日英情報の均衡性の確保が困難なため，日本語記事の内容を人手で忠実に英訳してコーパスを作成している．本稿ではこのコーパスを Manual と略記する．作成方針の詳細は [3] に記載の通りである．Manual コーパスには以下の特徴がある．

- 情報の均衡性
  人手で忠実に翻訳するため，日英の情報には高い均衡性が確保される．表 1 に示したように相関係数は 0.896 と非常に高い．
- 完全性
  記事全体を翻訳するため，記事全体からなる完全なコーパスが得られる．
- 表現スタイル
  日本語は原文で英語は翻訳である．時事通信社の英語スタイルに近づけるため用語集，スタイルガイドを提供して翻訳を依頼している．
- 経済性
  人手に頼った作業のため，コーパスの構築には多大のコストがかかる．

表 1 の Manual の諸元は 2018 年度作成部分であり，2019 年度末には 70K 文程度が追加される予定である．

### 2.3 日本語修正コーパス

日英対応記事の英語の記事の内容と等価になるよう日本語記事を翻訳者が修正する．この修正した日本語記事と元の英語記事からなるコーパスを作成している．このコーパスを Repair と略記する．Repair コーパスには以下の特徴がある．

- 情報の均衡性
  英語記事に合わせて日本語記事を修正するため，Manual と同等の高い均衡性が確保される．文長の相関係数は 0.868 と高い．
- 完全性
  記事全体を対象に修正するため，Manual と同等の記事全体のコーパスが得られる．
- 表現スタイル
  日本語は修正結果で英語は原文である．日本語記事をゼロから作成するのではなく，利用できる表現はそのまま使う．このため日本語記事は時事通信社の日本語ニュースのスタイルにかなり近いことが期待できる．ただし，英語に合わせて主語が補完されるなど英語ニュースの特徴が反映されている．
- 経済性
  日本語の固有名詞や専門用語の調査を省ける場合があるため，完全な人手の Manual コーパス作成より安価である．

2019 年 12 月現在，28K 文程度の量であるが，2019 年度末には 100K 文程度まで拡張される予定である．

## 3 翻訳実験

コーパスの効果を測定するため，Transformer[4] による翻訳実験を行なった．利用したシステムは Sockeye toolkit[5] による実装である．

[1] コーパスの日英の文の情報が常に均衡していれば，日英の文長の変化は一致し，高い相関係数を示すと考えられる．

### 3.1 設定

日英記事をトークナイズしたあと[2]，語彙サイズ 30K の条件で日英同時 byte-pair encoding[7] を実施した．学習はバッチサイズ 5,000 トークン，最大文長 90 トークン，最大エポック数 30，チェックポイント間隔 5,000 に設定した．また連続して 3 回チェックポイトでパープレキシティが改善しない場合は終了するようにした．その他はデフォルトの設定である．

Manual コーパスと Align コーパスの学習には最大で 195,304 文（195K と略記）を利用した．Repair コーパスは 28,452 文（28K と略記）を利用した．

評価データは Manual コーパスの 200 記事（1,564 文），開発データは 100 記事（791 文）を使った．評価データを Manual コーパスから選択した理由は，日本語が実際に出現する日本語記事であること，日本語記事の内容が過不足なく英語記事に反映されており，下訳として時事通信社で使うための条件に近いと考えたからである．

### 3.2 結果

まず学習に使う Manual および Align コーパスの量を変化させ，翻訳性能の変化を観察した．具体的には Manual コーパス，Align コーパスの学習量を 100K，150K，160K，…，190K，195K と変化させ，それぞれのポイントで 5 回モデルを学習し，評価データを翻訳した．そして各ポイントで BLEU スコアの平均を算出した（図 1）．Align コーパスの BLEU スコアは 100K で 7.86，195K で 9.86 である．一方，Manual コーパスの BLEU スコアは 100K で 17.9，195K で 21.9 である．どの時点でも Manual コーパスの BLEU スコアは Align コーパスより高い．また，Align コーパスのデータ追加による BLEU スコアの改善は小さいが Manual コーパスの改善はこれより大きい．

次に，Repair コーパスの効果を確認することにした．ただしコーパスのサイズが小さいため単独の学習は行わず Manual コーパス全体の 195K 文に Repair コーパス全体の 28K 文を加えた時の BLEU スコアの変化を調査した（図 2）．195K での BLEU スコアは 21.9 で 28K を加えると 22.0 とわずかに上昇したがその差は有意ではなかった．

[2]日本語は KyTea[6] を使用し，英語は Moses toolkit を使用した．

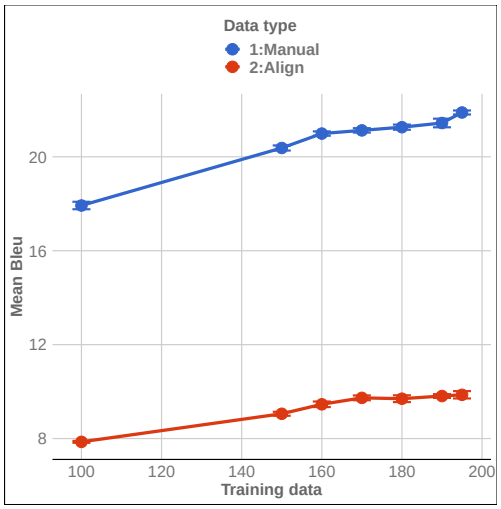


図 1: Manual と Align コーパスの比較

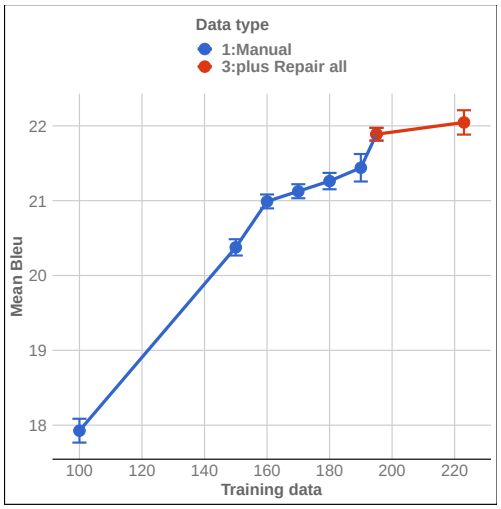


図 2: Manual への Repair の追加

効果を個別に確認するため，Manual コーパス 195K で学習したシステム，および Manual コーパス 195K に Repair コーパス 28K を加えて学習したシステムで評価セットを翻訳し，文単位の BLEU スコアを計算した．この結果，評価事例 1,546 の内，Manual+Repair コーパスの BLEU スコアの方が高かった事例が 475，逆に Manual コーパスの BLEU スコアの方が高かった事例が 441 であった．前者の例を表 2 に示す．全体の効果は見られなかったが，個別には改善している可能性があり，今後調査を続けたい．

今回の実験より，Manual コーパスがニュースの翻訳システム構築の上で有用であることが確認できた．また，Manual コーパスの学習データ量を増やすことによ

表 2: Manual+Repair コーパスの BLEU スコアが Manual コーパスより高かった事例

| | |
|---|---|
| Source | 米軍岩国基地（山口県岩国市）には今月、最新鋭Ｆ３５Ｂステルス戦闘機が米国外で初めて配備され、トランプ氏は就任演説で「古くからの同盟を強化する」と明言した。 |
| Reference* | A state-of-the-art F35B stealth fighter was deployed for the first time outside the U.S. at the Iwakuni base of the U.S. Forces (Iwakuni City, Yamaguchi Prefecture) this month. Trump clearly stated that he would reinforce old alliances in his inaugural address. |
| Manual corpus | In the U.S. Iwakuni Base (Iwakuni City, Yamaguchi Prefecture), the latest F-35B stealth fighter jets was deployed outside the U.S. this month, and Trump clearly stated that he will "strengthen the alliance from old." |
| Manual +Repair corpus | This month, the state-of-the-art F35B stealth fighter jets were deployed outside the U.S. for the first time in Iwakuni, Yamaguchi Prefecture, and Trump declared in his inaugural address, "We will strengthen our alliance from old." |

*Source は 1 文だが Reference は 2 文である．翻訳システムは 1 文を出力している．

り BLEU スコアが上昇することを確認できた．Repair コーパスの学習データ追加の効果は今後さらに確認したい．現在，Manual，Repair 両コーパスの構築を続けており，また Align コーパスの追加も行う予定である．これらの追加の効果も今後確認していく．

なお，本稿では各コーパスの性能測定が目的のため，各コーパス単独，あるいは単純な合併により機械翻訳システムを学習した．実際に本稿のような性質の異なるコーパスを混合する場合には Mino らの報告 [8] のようにタグを使った適応化によって，より高い効果が得られることを付記しておく．

## 4 おわりに

本稿では開発を進めている 3 種類の日英ニュースコーパスの概要を説明し，翻訳実験によってそれらの効果を報告した．コーパス開発は 2020 年度まで継続する予定である．今回の実験の結果を含め，さまざまな実験を行い，より効果的な開発手法を見出したい．

## 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] 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. 時事通信社ニュースの日英均衡コーパスの構築–第 1 報. 言語処理学会第 25 回年次大会（NLP2019）発表論文集, 2019.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. 2017.

[5] Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. The Sockeye Neural Machine Translation Toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pp. 200–207, 2018.

[6] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 529–533, 2011.

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, 2016.

[8] Hideya Mino, Hitoshi Ito, Isao Goto, Ichiro Yamada, Hideki Tanaka, and Takenobu Tokunaga. Neural Machine Translation System using a Content-equivalently Translated Parallel Corpus for the Newswire Translation Tasks at WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pp. 106–111, 2019.