時事通信社ニュースの日英対訳コーパスの構築–第３報

田中 英輝
NHK エンジニアリングシステム
hideki.tanaka@nes.or.jp

中澤 敏明
東京大学
nakazawa@logos.t.u-tokyo.ac.jp

美野 秀弥　伊藤 均　後藤 功雄　山田 一郎
NHK 放送技術研究所
{mino.h-gq, itou.h-ce, goto.i-es, yamada.i-hy}@nhk.or.jp

川上 貴之　大嶋 聖一　朝賀 英裕
時事通信社
{kawakami, sohsima, asaka}@jiji.co.jp

1　はじめに

著者らはニュースの日英機械翻訳の研究開発の基礎となる言語資源の構築のため，時事通信社のニュース記事を利用した日英対訳コーパスの開発を進めている [1][2]．時事通信社では，日々日本語の記事を作成し，内外の報道機関に記事を提供すると共に，その一部を英訳して同様に配信している．日々の業務で生み出される日英対応記事に文アラインメントを適用すると文単位の対訳コーパスを構築できる．しかし，日英対応記事の数は日本語記事に比べてかなり少なく，英訳時に内容が大きく編集されることから [1]，自動文アラインメントだけで現在の機械翻訳システムの学習に適した大規模コーパスを構築するのは困難である．そこで，文アラインメントに加えて，英訳を持たない日本語単独記事を翻訳者が翻訳する手法，日英対応記事の日本語記事を英語に合わせて翻訳者が修正，翻訳する手法の 2 手法を加えてコーパスを開発している．以下では，過去 3 年で構築してきた 3 種類コーパスの概要を報告すると共に，日英翻訳実験と調査によりその特徴を明らかにする．

2　コーパスの概要と諸元

時事通信社の翻訳は，内容の大きな編集を伴うことが多いが，編集と翻訳を高精度に実現する翻訳システムの実現は現時点では困難である．また，このようなシステムが実現できたとしても，人が修正する下訳用途には使いにくいだろう．むしろ内容を忠実に翻訳するシステムの方が下訳にはふさわしいと考え，著者らはその実現のため，文単位で対応するコーパスを以下の 3 手法で構築した．

文アラインメントコーパス: Align　時事通信社の日英対応記事に文アラインメントアルゴリズム [3] を適用し文単位で対応する対訳コーパスを構築した．本稿では文対応が 1 対 1 で，かつその類似度が 0.3 以上の文対を抽出して作成したコーパスを報告する[1)]．時事通信社では日本語記事の内容を大幅に編集して英語記事を作成するため，文アラインメントされた文ペアの情報には不均衡が生ずる．実際，表 2 の文長の相関係数は 0.599 と他の 2 種類に比べると低い．また，対応付けされる文は記事の一部であり，今回の抽出条件で 582K の日英文対から抽出された文対は 240K に留まった．一方，構築コストは安価である．またコーパスの話題，日英の表現スタイルは日英対応記事と一致する．本稿ではこのコーパスを “Align” あるいは “a” と略記する．

日英翻訳コーパス: Manual　日本語単独記事の一部を抽出し，翻訳者に文ごとに翻訳してもらいコーパスを構築した．得られたコーパスの日英の情報は均衡しており，実際，文長の相関係数は 0.899 と高かった（表 2）．記事の文数，キーワードリストを利用して，翻訳されにくい記事を除外しているが，日英対応記事より広い話題を含んでいると考えられる．翻訳にあたっては，翻訳スタイルガイドを事前に作成して翻訳者に提供し，時事通信社で使う用語，翻訳スタイルにできるだけ準拠するようにした．作成コストは 3 手法の中で一番高い．必要に応

1)　本コーパスは昨年報告したコーパスと同じである [2]．現在はさらに規模の大きなコーパスを構築している．

表 1 コーパス構築の 3 手法と特徴

	作業方向	日本語	英語	均衡性	文脈	話題	コスト
Align	-	時事	時事	△	×	= 時事	低
Manual	日 → 英	時事	翻訳	○	○	> 時事	高
Repair	英 → 日	修正・翻訳	時事	○	○	= 時事	中

「日本語」「英語」列の「時事」は日英対応記事の文がそのまま使われていることを示す．
「話題」列は日英対応記事の話題との比較で = は同一，> はより広いことを示す．

表 2 コーパスの諸元

	Align	Manual	Repair
収集開始	2011/10	2016/01	2017/01
収集終了	2018/06	2020/09	2020/09
記事数	-	41,972	20,400
文数	239,718	372,877	223,486
日本語文平均長	54.6	49.3	49.3
英語文平均長	26.0	25.5	24.9
文長相関係数	0.599	0.899	0.867
日英の文対応	1-1 のみ	1-1 から 1-5	1-1 のみ

じて日本語 1 文を複数の英文で翻訳することを許容した結果，表 2 に示すように 1-1 から 1-5 の日英文対応が含まれ，この対応の 95%が 1-1 であった．本稿ではこのコーパスを “Manual” あるいは “m” と略記する．

日本語修正コーパス: Repair Align の情報の均衡性を高めるため，日英対応記事の日本語記事を英語記事に合わせて翻訳者が修正，あるいは翻訳する手法でコーパスを作成した．得られた文長相関係数は Manual と同様 0.867 と高かった（表 2）．Align 同様，コーパスの話題は日英対応記事と一致する．英語 1 文に合うように日本語文を修正，翻訳する際，複数文にすることを許容したが，結果的に全て 1 対 1 の文対応となった（表 2）．日本語記事の修正，翻訳時に固有名詞の調査が不要，文や表現の再利用が可能などの利点があり，作成コストは Align と Manual の間となった．本稿ではこのコーパスを “Repair” あるいは “r” と略記する．以上のコーパスの特徴を表 1 に，数値的な諸元を表 2 に示す．

3 日英翻訳実験と調査

2 節で報告した 3 種類のコーパスの効果を確認するため Sockeye toolkit[4] による実装の Transformer[5] を使った日英翻訳実験を行った．翻訳モデル学習の設定は以下の通りである．日英記事をそれぞれトークナイズした後[2]，語彙サイズ 30K の条件で日英同時 byte-pair encoding 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 回チェックポイントでのパープレキシティ改善がなければ終了するようにした．その他はデフォールトの設定に従った．

2) 日本語は KyTea[6]，英語は Moses toolkit を使った．

表 3 実験データの文数

	Align	Manual	Repair
train	236,380	199,507	311,777
dev	449	756	657
test	1,731	1,484	2,081

自動評価には Bleu 値を使い，同一学習データから 5 つの初期値でモデルを作成し，5 種類の翻訳結果の Bleu 値を平均を採用した．

実験に使ったコーパスの大きさを表 3 に示す．実験用コーパスは見出しを含まない本文だけであり，Manual コーパスは，Align と Repair コーパスに合わせるため，日英 1 対 1 の本文を抽出して作成した．Align の開発，評価データは日英の対応が良い文対を翻訳者が選択して作成している．

3.1 学習データ追加効果

3 種類のコーパスの学習データを 10K 文ずつ最大サイズまで増加してモデルを学習し，評価データの翻訳結果の Bleu 値を測定した（図 1）．評価データと学習データの種類は同一である．3 種類のコーパスとも学習データの増加と共に Bleu 値は上昇した．どの Bleu もまだ飽和には達していないが Manual の Bleu の上昇幅は小さくなりつつある．

Align，Repair，Manual の Blue 値にはかなりの開きがある．同じ日英ニュースのコーパスではあるが，性質に何らかの違いがある可能性がある．そこで，その原因を探るため以下の実験を行った．

3.2 学習・評価データ種類の交差

Align，Manual，Repair の各 173K 文（図 1 の赤線に相当）を使って学習した 3 種類の翻訳モデルで前記 3 種類の評価データを翻訳し Bleu を測定した（図 2）．グラフを観察すると，まず，Align と Repair の評価データに対する Manual のモデルの Bleu は

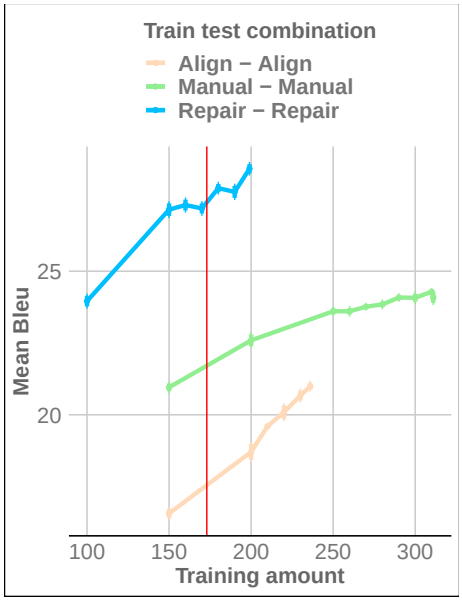

図 1　3 種類のコーパスの学習データの追加効果

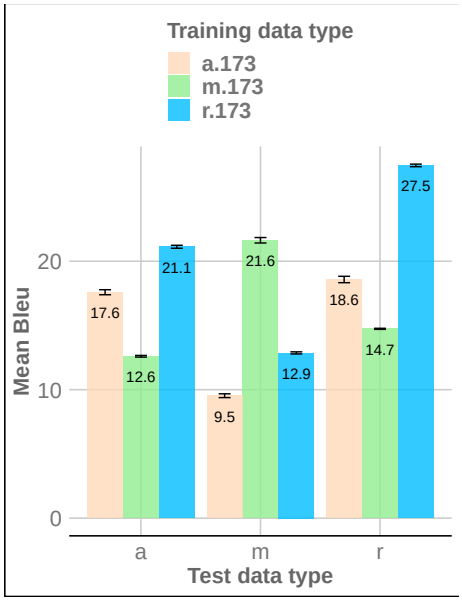

図 2　3 種類のコーパスの交差評価

表 4　異なり語の Dice 係数

	コーパス種別の組合せ	Dice 係数
日本語	Align-Manual	0.590
	Align-Repair	**0.636**
	Manual-Repair	0.591
英語	Align-Manual	0.446
	Align-Repair	**0.556**
	Manual-Repair	0.455

表 5　語の平均出現数

	コーパス種別	語の平均出現数
日本語	Align	143.0
	Manual	**97.2**
	Repair	136.8
英語	Align	81.9
	Manual	**53.0**
	Repair	83.9

低い．一方，Manual の評価データに対する Align と Repair のモデルの Bleu は低い．以上の事実は Align と Repair コーパスの性質と Manual コーパスには何らかの違いが存在することを示唆している．

次に，Align と Repair の評価データに対しては Repair のモデルが Align のモデルよりかなり大きな Bleu を達成したことが分かる．両者の出典は日英対応記事だが，Repair コーパスは人手をかけて情報の均衡性を高めており，高い Bleu 値はその効果の現れだと推察できる．

3.3　語彙の類似性と語の平均出現数の調査

Manual，Align，Repair の学習データに含まれる語彙の類似性を調査した．それぞれのトークナイズされた学習データから（Repair のほぼ全数である）195K 文を取り出し，異なり語の Dice 係数を計算した（表 4）．Dice 係数を見ると日英とも Align-Manual と Repair-Manual はほぼ同じだが，Align-Repair の Dice 係数はこれらより大きい．つまり Manual コーパスと Align，Repair コーパスには語の分布の差，すなわち話題の差があることが分かった．

次に語の平均出現数（token-type ratio）を計算した（表 5）．Manual の語の平均出現数は日英とも Align，Repair よりかなり小さく，データがより希薄であることが分かる．これは Manual コーパスは Align や Repair に比べて広い話題を扱っていることの証左とも言える．図 1 で Repair コーパスに比べて Manual コーパスの Bleu が全領域で低く，また Manual コーパスではデータ追加の効果が現れにくいことが観察されたが，その理由の 1 つは Manual コーパスの語彙の希薄性にあると思われる．

3.4　Manual と Repair の主観評価

図 2 に示した Manual と Repair で作成したモデルを互いに交差評価した部分（図の右から 1,2,4,5 番目の棒グラフの領域，以後，評価領域と呼ぶ）の Bleu にはかなりの差が見られる．そこで，この領域について特許情報機構の 5 段階の内容伝達レベルによる主観評価を実施した [7] [3]．Repair の評価データからは 93 文，Manual の評価データからはは 94 文の評

3）　翻訳に反映されている原文の重要情報の量を 5 段階に分類した評価基準で 1 が 20%以下，5 が 100%となっている．

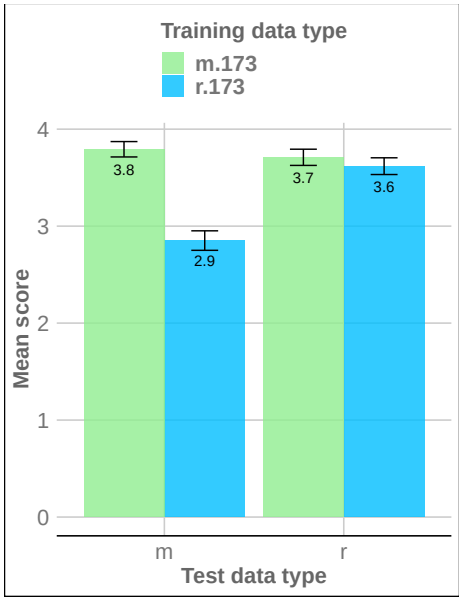

図 3　Manual と Repair コーパス翻訳結果の主観評価

価データを抽出し，全てを同一の 2 名の翻訳者で評価した．評価に当たっては誤りの一部を許容するようにした．例えば，時事通信社では日本語記事の日付を英語記事では曜日で翻訳する．現在の日英コーパスは基本的にこのような形になっているが，日付の曜日への正確な変換は今回の翻訳システムには望めない．また，日本語記事には出現しない企業名の証券コードが英語記事には出現する，日本語に出現しない記事の出典情報が英語記事の冒頭に “Tokyo, Sept. 18 (Jiji Press)” のように付与されることがあるといった問題も正しく対応できない[4]．以上の現象は別処理，あるいはコーパスのクリーニングで対応すべき現象であり，今回の評価対象から外した．一方，Bleu の評価ではこのような差異も検出される．

評価の平均値を図 3 に示す．これを図 2 の評価領域と比較する．まず，Manual の評価データに対する性能は Bleu 値，主観評価値とも Manual のモデルが高い．次に，Repair の評価データに着目すると，Bleu 値は Repair のモデルが Manual のモデルより圧倒的に高いが，主観評価スコアは両モデルでほぼ同等となった．以上の傾向は 2 名の評価者とも同じであった．Repair の評価結果を観察したところ，主観評価で許容した誤りが Bleu で問題になったケースが見られた．具体例は割愛するが，本質的に扱えない現象を主観評価で許容したことが Bleu 値と主観評価の差につながった可能性が高い．おそらく，実感的な性能は現在の Bleu 値ではなく，主観評価値に近いと考える．

4) Repair の英語記事には証券コードと出典が完全に記載されているが，Manual の英語記事には一部を除いてない．

4　おわりに

本稿では 3 種類のニュースの日英対訳コーパス構築の概要，および，それぞれの性質を翻訳実験，語彙調査，翻訳結果の主観評価で調査した結果を報告した．ニュースの日英対訳コーパス構築は 2021 年度で一旦区切りとなる．今後は，コーパスのクリーニングや改修を加えてより詳細な評価を実施する．さらに，本コーパスを利用した [8] などの新たな研究へ軸足を移していく．

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] 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.

[2] 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対訳コーパスの構築–第 2 報. 言語処理学会第 26 回年次大会（NLP2020）発表論文集, 2020.

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

[4] 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.

[5] 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.

[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] 特許庁情報技術統括室. 特許文献機械翻訳の品質評価手順 Ver.1.0, 2014.

[8] 美野秀弥, 衣川和尭, 伊藤均, 後藤功雄, 山田一郎, 田中英輝, 川上貴之, 大嶋聖一, 朝賀英裕. 単言語データを用いた逆翻訳と順翻訳によるデータ拡張の効果の比較. 言語処理学会第 27 回年次大会（NLP2021）発表論文集, 2021.