

# グローバルコミュニケーション計画のための 多言語パラレルコーパス

今村 賢治      隅田 英一郎

国立研究開発法人 情報通信研究機構

{kenji.imamura,eiichiro.sumita}@nict.go.jp

## 1 はじめに

総務省では、世界の「言語の壁」をなくすことをミッションとした「グローバルコミュニケーション計画 (Global Communication Plan; GCP)」を 2014 年度より実施している (2 節参照)。この計画中で大きなウェイトを占めているのは、多言語音声翻訳の社会実装である。病院や商業施設、観光地等のさまざまな拠点で多言語音声翻訳システムを利用可能するとともに、ユーザインタフェースや翻訳品質の改善などの利便性の向上を図っている。

多言語音声翻訳システムを開発するのに必須となるのが、多言語の対訳コーパスである。グローバルコミュニケーション計画の中にも、社会実装用の対訳コーパス開発があり、情報通信研究機構 (NICT) が中心となって開発を進めている。

本稿では、この社会実装用の多言語対訳コーパス (GCP コーパスと呼ぶ) の概要と、現在の開発状況について述べる<sup>1</sup>。このコーパスは、日本にきた外国人が、現地の日本人と会話するシチュエーションを想定し、シナリオライターが疑似対話を作文、それを翻訳する形式で作成されているものが多数を占める (一部、フレーズブック調の単独発話も含む)。疑似ではあるが、対話形式になっている。対象言語は、グローバルコミュニケーション計画のターゲットである 10 言語 (日本語を含む)、対象分野は、医療、防災、ショッピング、観光の 4 分野である。

GCP コーパスは、音声翻訳システム開発向けに作成されているが、以下の特徴を持つため、さまざまな研究への活用も期待できる。

1. アジア言語を含む 10 言語をカバーする文対応の多言語対訳コーパスである。そのため、最大 90 言語対の機械翻訳器を作成できる。また、ピボット翻訳やゼロショット翻訳 (Johnson et al., 2016) の比較研究にも利用することができる。
2. 医療、防災、ショッピング、観光の 4 分野をカバー

している。そのため、ドメイン適応の研究にも利用できる。

3. 疑似対話となっている。対話文脈を考慮した機械翻訳の研究など、文脈処理の研究に利用できる。なお、対話破綻がないので、実対話より文脈はシンプルである。

本稿では、第 1 の特徴に着目して、日本語から/への翻訳を行い、基本的翻訳性能を確認する。また、直接翻訳、ピボット翻訳、ゼロショット翻訳の比較を行う。

多言語パラレルコーパスは、現在、欧州連合議会の議事録である Europarl (Koehn, 2005) が多く使われている。GCP コーパスも、基本的には Europarl と同じ特徴を持つ。しかし、GCP コーパスはアジア言語を中心に構成されていること、外国人と現地人が会話するシチュエーションを想定し、携帯端末用音声翻訳システムが構築しやすいデータとなっている点で、Europarl とは異なる利用方法もできると考えている。

以下、2 節ではグローバルコミュニケーション計画の概要を、第 3 節では GCP コーパスの現状について述べる。第 4 節および第 5 節では、GCP コーパスを用いて機械翻訳を構成し、直接翻訳、ピボット翻訳、ゼロショット翻訳の品質を測定する。そして第 6 節でまとめる。

## 2 グローバルコミュニケーション計画

グローバルコミュニケーション計画は、総務省新藤大臣 (当時) が中心となって、2014 年に提唱された。そのミッションは、世界の「言語の壁」をなくすことである。そのために、3 つのビジョンが設定された。

1. グローバルで自由な交流の実現
2. 日本のプレゼンスの向上
3. 東京オリンピック・パラリンピックでの「おもてなし」

<sup>1</sup>対訳コーパスと並行して、音声認識・合成のための音声コーパスも開発しているが、本稿では対訳コーパスについてのみ述べる。

表 1: 医療分野における疑似対話例

大分類	中分類	小分類	発話者種別	発話
医療	病気・けが になったら	緊急時の対 応をする	外国人	このあたりに病院はありませんか？
			日本人	あのコンビニを右に曲がったところに内科の医院があります。
			日本人	具合が悪いんですか？
			外国人	めまいがします。
			日本人	日本語は得意ではないんですか？
			外国人	得意ではありません。
			日本人	あその医院だと英語は通じないかもしれません。
			日本人	隣の駅前の総合病院なら英語のできるスタッフがいます。

表 2: 2017 年末現在の GCP コーパスのサイズ

言語	トータル	文数				
		医療	防災	ショッピング	観光	その他
日本語	2,029,111 (25.2 文字/文)	420,270	249,495	355,429	527,056	476,861
英語	2,029,111 (11.2 単語/文)	420,270	249,495	355,429	527,056	476,861
中国語	2,026,608	420,270	249,495	355,429	527,056	474,358
韓国語	2,026,608	420,270	249,495	355,429	527,056	474,358
タイ語	1,150,070	145,054	117,636	180,843	232,179	474,358
ベトナム語	1,150,070	145,054	117,636	180,843	232,179	474,358
インドネシア語	1,150,070	145,054	117,636	180,843	232,179	474,358
ミャンマー語	1,150,070	145,054	117,636	180,843	232,179	474,358
スペイン語	337,654	145,054	117,636	9,512	18,944	46,508
フランス語	340,499	145,054	117,636	9,867	19,593	48,349

この計画に基づき、グローバルコミュニケーション開発推進協議会<sup>2</sup>が組織され、1) 多言語音声翻訳の対応領域、対応言語を拡大するための研究開発、2) 産学官連携による病院、商業施設、観光地等での実証実験を行っている。

グローバルコミュニケーション計画では、4つの分野が設定された。

- 医療: 病院等における患者と医療スタッフ（医師、看護師、事務など）の対話を含む。
- 防災: 外国人への災害情報の提供や、災害時の自治体等の対応を想定
- ショッピング: 外国訪問客の買い物時に、店員等が対応する対話を想定
- 観光: 訪問客の宿泊、移動や、観光地での案内などを想定

言語は、日本語、英語、中国語（簡体字）、韓国語、タイ語、インドネシア語、ベトナム語、ミャンマー語、スペイン語、フランス語の10言語を対象としている。これは、訪日外国人数などを考慮して決定された。

### 3 GCP コーパス

GCP コーパスは、グローバルコミュニケーション計画に基づいて開発された、音声翻訳を実現するためのコーパスである。大きくは音声コーパスと対訳コーパスに分けられるが、本稿では、対訳コーパスについて説明する。

<sup>2</sup><http://gcp.nict.go.jp/about/index.html>

コーパスの対象分野は、GCP で設定された4分野である。本コーパスの主な用途が音声翻訳システムの実現であるため、日本を訪れた外国人が、現地の日本人と会話を行うというシチュエーションが想定された。そのため、コーパスの大部分は対話形式となっている。ただし、実際の対話を録音・書き起こしたのではなく、シナリオライターが状況を想定して作文した疑似対話となっている。医療分野における疑似対話の例を表1に示す。また、挨拶や旅行会話の常套句のような発話に関しては、単独発話として整備している。各発話には、付加情報として、大分類、中分類、小分類に分類されたカテゴリ、および発話者種別（日本人、外国人別）が付与されている。

このように、GCP コーパスは、日本を訪問した外国人との対話を想定しているため、まず日本語で疑似対話を作成した。そして、日本語原文を多言語に翻訳する形で整備した。対象言語は、GCP の10言語（日本語を除くと9言語）である。2017 年末時点での、言語ごとの GCP コーパスのサイズを表2に示す。日本語、英語、中国語、韓国語を優先的に整備しているため、この4言語は他の言語に比べて大きなサイズとなっている。

### 4 GCP コーパスの翻訳器の品質

GCP コーパスは、音声翻訳を実現するための対訳コーパスであるので、本節では、まずニューラル機械翻訳器を学習させ、ベースラインとなる翻訳品質（本稿では BLEU (Papineni et al., 2002) を用いる) を測定する。

表 3: 日本語から/への翻訳品質

言語	訓練文数	BLEU スコア	
		日本語から	日本語へ
英語	1,954,477	27.20	31.01
中国語	1,952,475	35.82	42.34
韓国語	1,952,475	52.87	58.13
タイ語	1,110,232	25.64	27.64
ベトナム語	1,110,232	30.32	30.64
インドネシア語	1,110,232	22.16	25.94
ミャンマー語	1,110,232	23.90	30.82
スペイン語	326,433	22.28	24.82
フランス語	329,160	22.05	23.39

#### 4.1 実験設定

言語対 GCP コーパスは 10 言語の平行コーパスであるので、最大  $10 \times 9 = 90$  種類の翻訳器を構成することができる。このうち、本稿では日本語に対する翻訳 (日  $\leftrightarrow$  X の 18 種類) の翻訳品質を測定する。データセット 本稿では、表 2 のコーパスを、訓練、開発、テストに分割して使用する。まず、コーパス全体から、開発用、テスト用として各 2 千から 2 万程度の文を取り出し (全体のサイズが言語によって異なるため、ヘルドアウト量も異なる)、残った文を訓練セットとした。ヘルドアウトセットから均等に各 2,000 文を取り出して、開発セット、テストセットとした。翻訳システム 上記コーパスは、まず、内部開発の単語分かち書き器を用いて単語分割した。次に、各言語ごとにバイトペア符号化 (Sennrich et al., 2016) を行い、約 1.6 万のサブワードに分割した。

本稿では OpenNMT<sup>3</sup> (Klein et al., 2017) を翻訳システムとして使用した。設定は以下のとおり。

- エンコーダーは、2 層 Bi-LSTM(500+500 次元)。単語分散表現は 500 次元。
- デコーダーは、2 層 LSTM(1,000 次元)。単語分散表現は 500 次元。
- 学習には確率的勾配降下法 (SGD) を使用。学習率 1.0 で 14 エポック、その後半減させながら 6 エポック実行した。ミニバッチサイズは 64 とした。
- ビーム幅 10 で翻訳し、長さランキング (今村他, 2018) で翻訳文を選択した。

#### 4.2 翻訳品質

表 3 は、BLEU で測定した翻訳品質である。まず、BLEU スコアは、日本語からの翻訳で 22.05–52.87、日本語への翻訳で 23.39–58.13 と、言語によってかなりの異なりがある。しかし、全体としては、訓練文数が多い方が BLEU スコアが高い傾向がある。

<sup>3</sup><http://opennmt.net/>

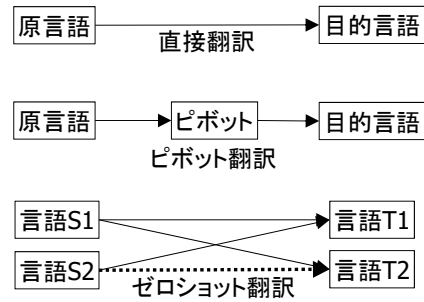


図 1: 直接翻訳, ピボット翻訳, ゼロショット翻訳

次に日本語からの翻訳と日本語への翻訳を比較すると、全言語について、日本語への翻訳の方がスコアが高い。日本語は主語が省略される場合など、対応する言語では必須の要素が省略されることがあるため、日本語からの翻訳が難しくなったと考えられる。

また、韓国語との翻訳のみ、BLEU スコアが 50 を超えるという結果になった。韓国語は、日本語と文法が似通っている (どちらも SOV 形式) ため、従来の統計翻訳でも翻訳品質が高いことが知られている。今回用いたニューラル機械翻訳でも同様な結果が得られた。

### 5 ピボット翻訳, ゼロショット翻訳

多言語対訳コーパスからは、直接全言語対の翻訳器を作成することができる。しかし、そのような対訳コーパスが存在しない場合、通常はピボット翻訳を行うことになる。これは、リソースが豊富な言語をピボットにし、それを介して翻訳を行うものである (たとえば (Utiyama and Isahara, 2007))。

一方、ニューラル機械翻訳では、原言語のエンコーダーと、目的言語のデコーダーを訓練すると、直接の対訳データがなくても翻訳器が構成できる可能性が指摘されている。これをゼロショット翻訳と呼ぶ (Johnson et al., 2016) (図 1)。

本稿の多言語対訳コーパスを用いると、直接翻訳の場合の精度を上限とし、それに対してピボット翻訳、ゼロショット翻訳がどの程度近づけたかを評価することができる。また、同一の語彙セットが利用できるなど、直接翻訳、ピボット翻訳、ゼロショット翻訳で実験条件を共有できるため、三者の比較を行うことができるという利点がある。

本節では、日英中韓の 4 国語を例にとり、直接翻訳、ピボット翻訳、ゼロショット翻訳の比較を行う。

#### 5.1 実験設定

まず、4 言語で共通する約 200 万文から、4.1 節と同じ方法で開発セット、テストセット (各 2,000 文/言語) を抽出し、単語分割した。次に、4 言語をまとめ

表 4: 直接・ピボット・ゼロショット翻訳の比較  
(a) 直接翻訳

		目的言語			
		日	英	中	韓
原言語	日	–	27.31	35.60	52.81
	英	30.84	–	22.21	26.27
	中	42.33	24.14	–	34.85
	韓	57.92	24.85	30.66	–

(b) ピボット翻訳 (カッコは直接翻訳の再掲)

		目的言語			
		日	英	中	韓
原言語	日	–	(27.31)	(35.60)	(52.81)
	英	(30.84)	–	20.82	25.49
	中	(42.33)	23.85	–	33.87
	韓	(57.92)	24.55	30.82	–

(c) ゼロショット翻訳 (英中韓)

		目的言語			
		日	英	中	韓
原言語	日	–	26.32	34.17	51.73
	英	29.49	–	2.90	5.11
	中	40.71	8.87	–	10.00
	韓	56.66	10.63	6.44	–

てバイトペア符号化を行い、サブワード 5 万タイプに分割した。この実験では、すべて同一の語彙セットを用いた。

直接翻訳は、4 節と同じシステムを用いて、12 種類の言語方向のモデルを学習し、翻訳品質を測定する。

ピボット翻訳では、日本語をピボット言語として、英中韓の 6 種類の翻訳品質を測定する。すなわち、直接翻訳用に学習した日 ↔ (英 | 中 | 韓) モデルを用いて、(英 | 中 | 韓) → 日 → (英 | 中 | 韓) の翻訳品質を測定する。なお、ピボット言語への翻訳は、1 ベストを使用する。

ゼロショット翻訳では、ピボット翻訳と同様に、日 ↔ (英 | 中 | 韓) コーパスを用いて、(英 | 中 | 韓) → (英 | 中 | 韓) 翻訳器を構成する。まず、日 ↔ (英 | 中 | 韓) コーパスの原言語に、目的言語名を表すタグを埋め込んだ訓練データを作成する。そして、全言語対の訓練データを結合して、一つのモデルを訓練する。本実験では、約 1,200 万文の訓練データとなる。翻訳時は、原言語テストデータの先頭に目的言語タグを付加し、翻訳を行うと、学習に用いていない言語対についても、タグに従った目的言語が生成される。

## 5.2 実験結果

表 4 は、(a) 直接翻訳、(b) ピボット翻訳、(c) ゼロショット翻訳の結果である。まず、(b) ピボット翻訳を見ると、ピボット翻訳は、直接翻訳に比べ、翻訳品質が低下する言語がほとんどであるが、韓 → 中のように、若干向上する場合もあり、ピボット翻訳でも直接翻訳に近い精度が出せることがわかる。

(c) ゼロショット翻訳に関しては、非常に低い BLEU スコアしか出ない。ただし、学習に用いた言語対 (日 ↔ X) に関しては、直接翻訳よりは低くなるものの、非学習言語対に比べると高い数値を示しており、多言語モデル自体の学習はある程度成功していることがわかる。ゼロショット翻訳は、まだ研究途上の方式であるので、今後さらなる研究が必要と考える。

## 6 まとめ

本稿では、グローバルコミュニケーション計画に則って作成している GCP コーパスについて紹介した。本コーパスは、音声翻訳器を作成するために構築している文対応の多言語コーパスである。本稿では、多言語対訳であるという特徴について、翻訳品質評価を通じて基本特性を確認した。また、直接翻訳、ピボット翻訳、ゼロショット翻訳の比較が可能であることを示した。

今後は、東京オリンピックに向けて、さらなる拡充を行うとともに、本コーパスを音声翻訳器などに組み込み、幅広く提供してゆく予定である<sup>4</sup>。

## 謝辞

本研究は総務省の情報通信技術の研究開発「グローバルコミュニケーション計画の推進-多言語音声翻訳技術の研究開発及び社会実証- I. 多言語音声翻訳技術の研究開発」の一環として行われました。

## 参考文献

- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google’s multilingual neural machine translation system: Enabling zero-shot translation. ArXiv e-prints. 1611.04558.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. of ACL 2017, System Demonstrations*, pages 67–72.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT Summit X*, pages 79–86.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL-2002*, pages 311–318.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL-2016 (Volume 1: Long Papers)*, pages 1715–1725.
- Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proc. of HLT-NAACL 2007*, pages 484–491.
- 今村賢治, 藤田篤, 隅田英一郎. 2018. 単言語コーパスと逆翻訳を用いたエンコーダー・デコーダーの訓練法. 言語処理学会第 24 回年次大会.

<sup>4</sup>VoiceTra として提供中. <http://voicetra.nict.go.jp/>