多言語音声コーパスの人ー機械品質検査手法

水上 悦雄　榎本 成悟　テオリン アクセル エリック　加藤 宏明　河井 恒
情報通信研究機構 先進的音声翻訳研究開発推進センター
先進的音声技術研究室

{etsuo.mizukami,seigo.enomoto,axel.theorin}@.nict.go.jp
{kato.hiroaki,hisashi.kawai}@.nict.go.jp

1　はじめに

機械学習による統計的なモデル構築手法が主流である音声認識・機械翻訳・音声合成技術において大規模なコーパスの存在が重要であることは論を待たない。特に複数言語間の音声翻訳（Speech-To-Speech Translation）技術においては、多言語の大規模音声コーパス、テキストコーパスが必要となるため、一つの研究機関が自前で収録、書き起こし、テキスト作成からコーパス構築までを一貫して実施することは困難である。特に我々日本の研究機関においては日本語以外の他言語のコーパスの構築には、外注やクラウドサービスを用いて音声収録や書き起こしの業務を担ってもらうことが現実的な解となる。そこで問題となるのは、如何にその品質を保つかである。近年、クラウドソーシングを用いたデータ構築の際の品質担保のための様々な手法が検討されているが [1]、実際の大規模データを如何に検査するかは言語の壁の問題もあり、容易ではない。本稿では情報通信研究機構（以降、NICT と記述）が大規模多言語音声コーパスを構築するために用いている検査手法、構築上の課題や対策について述べる。

2　NICT 多言語音声コーパス

2014 年に総務省が掲げたグローバルコミュニケーション計画（以降、GCP と記述）に基づき、NICT では東京オリンピックを見据えた言語の壁を超えるための音声翻訳技術の研究開発を推進している。訪日インバウンド対象国、日本からのアウトバウンド対象国の公用語に鑑み、2020 年までに、日本語、英語、中国語、韓国語、タイ語、ベトナム語、インドネシア語、ミャンマー語、フランス語、スペイン語（以降、GCP10 言語と記述）の実用レベルの相互音声翻訳技術の実現を目標とし、そのために必要な音声コーパスを構築している。

2.1　コーパス設計

音声コーパスは、その目的に依存して、大きく分けて、ドメイン（ジャンル）、話者属性、発話スタイル、音響環境の四要素を如何に適応的かつバランスよく収集するかがポイントとなる。また、音声コーパスは、音響モデル学習のために用いる[1]ため、当該言語で発話され得る、多様な人による、多様な音素のパターンが、しかるべき頻度で出現する必要があるが、有限のコーパスでそれを網羅することは難しい。そのため、ある程度目的や対象を絞ってコーパスも収集することが現実的となる。

NICT が目指しているのは、訪日外国人が日本国内で旅行や生活をする際に、母語で、日本人接遇者とコミュニケーションする、あるいは、邦人が海外に赴いた際に、日本語で同様にコミュニケーションすることを支援するための音声翻訳技術である。そのため、話される内容は、観光や生活に関係することを主とし、発話のスタイルはどちらかと言えば、初対面の人同士で話される程度の、フランク過ぎず、機械へのコマンド入力でもないような発話スタイルを想定している。

2.1.1　模擬会話/独話/発話ログ

そのため、開発当初の設計では、買物場面、病院窓口、公共機関窓口、災害場面、チケット購入および公共交通機関窓口での会話を想定して、客と接客者、患者と病院関係者、申請者と公共機関窓口担当者のよう

[1]書き起こしテキストは言語モデル学習にも用いるが、それだけでは十分ではないので、Web 等から収集したより大規模なテキストコーパスをベースとするのが一般的である。

な二者間の一連の会話を模した「模擬会話」を様々な状況を想定して収録するという形をとった。実際の場面ではこの二者間では異なる言語が話されるわけであるが、インバウンドとアウトバウンドで双方の立場があり得るため、それぞれの言語でこれらを収録すれば、両方の立場の人の音声入力に対応できることになる。

この設計の問題は、その立場に成り切って発話する必要があるため、客側の経験はあっても、接遇者側の経験がなければ、何を話してよいかもわからず、話し方もどうしても不自然になってしまう点である。そのため、ある程度はどのようなことを話すのかの台詞に近い原稿を用意する場合がある。そこで、原稿を用意することはあっても、それを読み上げるのではなく、会話の流れから、自分の言葉で発話するよう教示するようにし、書き起こしは実際の発話に基づいて作成するように指示した。

模擬会話は、二者による交互の対話であることで、より自然な会話を収録することができる一方で、二者がその場にいなければならないという制約がある。まずその二者の予定を合わせる必要がある上に、当日になってどちらかが来られないことも少なくなく、その管理コストは多大である。最初から原稿が用意されていれば、相手がいることを想定して自分の発話だけをすればよく、後にこの「独話」スタイルの会話も収録することとした。原稿として用いたのは、NICT 翻訳技術研究室が構築した多言語パラレルコーパス [3] である。多言語パラレルコーパスは、機械翻訳を目的としているため、対訳の対応関係がつきやすいように、会話と言いながらも直訳的な印象を受ける文スタイルとなっている。そのため、各言語で自然に発話できるように修正した原稿を用い、さらに現場でも各自が話しやすいように適宜変更して発話するように指示した。

ただし、実際に機械を介したコミュニケーションをする場合、必ずしも人同士の発話スタイルのようにはならず、かつ、音声認識や機械翻訳が長文を受理できない（実際には、ある程度の長さの文であるほうがよいのだが）と考える話者の発話は短いフレーズのようなものが多い。また、様々な音響環境に対応するために、学習データに雑音を重畳して耐雑音性を強化することもできるが、実環境で収録された音声を用いることができるなら、それに越したことはない。NICT では、音声翻訳技術の社会実証、社会還元を目的として、無料の音声翻訳アプリ「VoiceTra[2][4]」を公開している。VoiceTra はまさに実環境における実ユーザ発話が収録されているため、このユーザ発話記録（「発話ログ」）を書き起こして、学習データに用いることができれば、より現場音声に強い音声認識モデルを構築することができる。

NICT の音声認識は、これら、模擬会話、独話、発話ログの三種の音声コーパスを用いて、音声認識モデルを構築しており、これらはそれぞれに、音声認識の精度向上に寄与している。

2.1.2 話者バランス

母語話者と一口に言っても、様々な属性があり得る。音声言語的に最も大きく関係するのは、性別、年齢、方言であろう。NICT の音声認識は、音声アプリケーションを高頻度に使用するであろう年齢層（15〜60 歳）の男女による、各言語の標準語[3]を対象としているが、方言話者による標準語音声も認識できることを目標としている。方言話者による音声は標準音声に比して音韻に変異があるため、これら方言話者の音声もしかるべき比率で音声コーパスに含まれている必要がある。模擬会話、独話を収録する際には、これら性別、年代、方言地域のバランスを考慮している。

2.2 アノテーション仕様

音声コーパスの構築には、収集した音声から、一定の基準で書き起こしを作成する必要がある。模擬会話、独話に関しては、そもそも書き起こせないような音声は収集対象にならないが、発話ログは、音声入力の失敗、言語選択の失敗による他言語音声、高雑音下で判別不能な音声、非母語話者による音声、など、学習データとして望ましくない音声を多く含む。よって、これらを除外したうえで、書き起こしを行う。

NICT における言語共通のアノテーション基本仕様は以下である。

- 各言語の標準表記（正書法）で書き起こす。
- フィラーや感嘆詞はマークする（例：[あのー]）。
- 読みが特定できない、外来語表記や二桁以上の数字に関しては、標準表記とスペルアウトを併記する（例：[110/百十] 番、[110/百トウ] 番）。
- 一文の終わりには句点/疑問符/感嘆符を打つ（通常句点を用いないタイ語等の言語でも）。
- 不明な言葉は [<unk>] とする。
- 不用意に記号は使用せず、スペルアウトあるいは併記する。

[2]http://voicetra.nict.go.jp/

[3]標準語をどう規定するかは実は容易ではないが、NICT では「国営放送あるいはそれに準ずる公共機関のアナウンサーによる放送音声」を一つの基準としている。

- 引用符は必要に応じ当該言語で一般的に使用する記号を統一的に使用する。
- unicode に準拠した文字コード（UTF-8）を使用し、改行コードは LF とする。

特に最後の仕様は計算機で扱う上で重要であるが、後述するミャンマー語においては単純でない問題がある。

3 NICT のコーパス検査手法

上述までの収録設計、音声書き起こしのアノテーション仕様に基づいて、話者募集、収録、音声の書き起こしまでのほとんどを NICT では外注によって実施している。その場合、作成された音声、書き起こしの品質をどのように検査するのか、という問題が生じる[4]。そこで NICT では、機械的な自動検査手法と人手の検査手法の二通りの検査によって、作成物の品質をチェック、フィードバックすることで、最終的な音声と書き起こしの品質を保っている。

3.1 自動検査手法

数百時間、数十万文に及ぶ大規模な音声収録・書き越しには、多数の作業者が関わる。特に書き起こし作業において、仕様通りに書き起こしが完璧になされる（100%の書き起こし精度）ことは不可能と言ってよく、言語によっては全体の 5 割を超える文に何らかの誤りを含むことは珍しくない。これを全件チェックすることは非現実的である。そのため、機械による自動検査は必須である。

3.1.1 テキスト解析による仕様合致検査

NICT では、テキストレベルの検査は、主に以下のような仕様の処理ツールを作成して検査を実施している。

1. 正規表現で検出可能な、フィラーやスペルアウトの記号の整合性（[括弧] の閉じ忘れなど）、句点の不備、不要な記号の検出
2. ゼロ幅スペースや不正な文字コードの検出

完全にエラーと特定できない場合は、アラートとして人手チェックを促すようにしている。例えば、ミャンマー語ではミャンマー数字を文字として代用してしまうことがあり注意が必要である。また、韓国語でも、二桁以上の数字表記を [括弧] で囲んだ場合、続く数詞をつなげるか、スペースを入れるかという分かち書きも規則化が難しく、人手でのチェックが必要となる。また、必要に応じてスペルチェックも実施する[5]。

3.1.2 VAD による分量・品質検査

前後の無音区間長を厳密に指定することは難しいので（厳密に指定すると、コストに跳ね返る）、一定の無音が発声の前後に入ることは許容されるが[6]、仕様上要求している総時間数のうち、半分が無音という状況は望ましくない。そこで、NICT では VAD（Voice Activity Ditection）ツールを提供して[7]正味の音声区間長で仕様要求量を指定、検査するようにしている。また、音声区間が特定できれば、簡易的な S/N も計算ができ、逆に音声区間が特定できないものは、そもそも音声として問題がある場合があるため、音声自体の検査にもなる。

3.1.3 ASR による品質検査

既に NICT では、GCP10 言語の音声認識モデルを構築しており、音声と書き起こしと認識（ASR）結果の比較をすることで、これらの乖離があるものは、書き起こしが誤っている可能性があるものとして検査に役立てている。ただし、ASR の性能は 100%でないため、認識結果と一致しないことが書き起こし誤りと言えるわけでもないし、そもそも、認識性能を向上させるために、音響的にも言語的にも既存の学習データにない音声を求めているわけであるから、むしろ、認識できない音声があってしかるべきである。後述する人手チェックと合わせて、より頑健な検査となるように、補助的な位置づけとして利用している[8]。

3.2 人手検査手法

機械的手法による自動検査には限界がある。なぜなら、スペルチェックや正規表現によるフォーマット違反チェック、ASR によるチェックは完全ではないし、結局「音声を聞いてみないとわからない」というのが現実である。そのため、NICT では、GCP10 言語の母語話者をチェッカーとして雇用し、抜き取り検査を行っ

[4] 日本語（あるいは英語）であれば、分担して一定の分量を検査することはできなくはないが、その他の言語はそうもいかない。

[5] 現状、スペルチェッカーはフリーのツール（例えば Hunspell, https://github.com/hunspell/hunspell）を使用している。

[6] むしろ前後の無音は重要な音声要素である。

[7] https://github.com/ASTL-NICT/VAD から公開中。

[8] 実際の検査時には、純粋な抜き取りによる人手チェックをしている。

ている[9]。NICT では、現在、JIS 規格（JIS Z 9015-1）に準拠した、OC 曲線に基づく抜き取り検査を採用し、第一種、第二種の過誤を防ぐよう努めている。

4　課題について

以上の方法を用いても課題は残る。以下、これまでに直面している課題例を述べる。

4.1　文字コード

前述のように、文字コードは unicode に統一されている必要があるが、単純ではない。例えば、ミャンマー語は、その歴史的背景から、その統一が難しい言語である。ミャンマー国内で最も使われているのは、Zawgyi-One というフォント/文字コード[10]であるが、このフォントは unicode に準拠していない。ゆえに、unicode に準拠した [2] 文字コードで書き起こしを作成するように指示しても、作業者にその経験がないために、書き起こし作業自体が容易でないため、一旦 Zawgyi 系のフォントで作成して、それを unicode に変換するような場合もあるが、この変換が十分でないことがある[11]。さらに、Windows や Mac OS などの一般的な OS で unicode に準拠したタイプオーダーで入力して正確にミャンマー語を表示することができ、かつ unicode に準拠したタイプオーダーでないと正しく表示されないエディタが存在しないことが大きな壁となっている。そもそも、ミャンマー語の手書き順序と unicode 指定のタイプオーダーには矛盾する点もあり、一般的なエディタではこれを補完するので、間違っていることに気づけないことが多い。NICT では、先に示したテキストの自動検査で、最低限、あり得ないコードオーダーを検出するようにしている。

4.2　音声と表記の整合性

実際の音声は言い怠けて発話されることも多い。また、正しい発話と、一般的な発話が必ずしも一致しない場合がある[12]。日本語同様、韓国語やタイ語、ミャンマー語など、音をそのまま表記できてしまう言語において、標準表記と実音声のどちらを採用すべきかは悩ましいが、NICT では、書き起こしを正書法/標準表記で記載し、正規化するように指示している[13]。音声認識という点に関して言えば、発声のまま認識結果が出力されるほうがよいという見方もあるが、音声翻訳という観点から言えば、正規化されているほうが望ましい。将来的に end-to-end の音声翻訳技術開発が進めば、この点は問題とならない可能性はある。

5　おわりに

大規模音声コーパスを構築するに際して、100%の精度を求めることは難しい。では、何割程度であれば、誤りが許容されるかは、学習の結果、その誤りがどの程度、音声認識精度に影響するかに依存するだろう[14]。精度がよいに越したことはないが、それはそのままコストに跳ね返るので、費用対効果を考慮した精度指定と、検査手法が今後も求められるだろう。

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

[1] 馬場雪乃. ヒューマンコンピュテーションとクラウドソーシング. 言語処理学会第 23 回年次大会チュートリアル, 2017.

[2] Martin Hosken and Maung Tuntunlwin. *Representing Myanmar in Unicode — Details and Examples.* http://unicode.org/notes/tn11/, 2012.

[3] 今村賢治, 隅田英一郎. 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. 言語処理学会第 24 回年次大会, 3 月 2018.

[4] 松田繁樹, 林輝昭, 葦苅豊, 志賀芳則, 柏岡秀紀, 安田圭志, 大熊英男, 内山将夫, 隅田英一郎, 河井恒, 中村哲. Development of the “VoiceTra” multi-lingual speech translation system. *IEICE TRANSACTIONS on Information and Systems*, Vol. E100-D, No. 4, pp. 621–632, April 2017.

[9]検査を外部の専門業者に外注することもできるが、結局、その専門業者の作業結果を検査する必要があり、堂々巡りとなる。

[10]ミャンマー語では、フォントと文字コードは非分離である。

[11]現在ではほぼ 100%変換できるツールもある。

[12]例えば、十個を「ジッコ」と発話する人は多くはない。

[13]読み方が一般的とは言い難い場合は、実発声をスペルアウトして併記するように指示している。

[14]現状、十分な量（数百時間レベル）のコーパスであれば、一文のうち、一単語の書き起こしが間違っている文が全体の 10%程度あっても、大きな問題はないと考えている。