4Q-02

集合知を利用した対訳知識のカバレッジ向上

牛久 敦§　河原 大輔†　黒橋 禎夫†　颯々野 学‡
§京都大学工学部　†京都大学大学院情報学研究科　‡ヤフー株式会社

1　はじめに

高精度な機械翻訳を実現する上で、対訳知識のカバレッジの向上・充実は必要不可欠である。日常的に用いられる口語表現・フレーズに関しては、数十万項目を持つ対訳辞書でもカバレッジが不足しているのが現状である。そのため、特に日常会話等の翻訳において大きな悪影響を与えている。

本研究では、日英対訳辞書においてカバーできていない日常的表現をスマートフォン用アプリケーションにより収集し、クラウドソーシングによって訳語を獲得することでカバレッジの向上を目的とするフレームワーク (図 1) を提案する。このフレームワークの背景は次の通りである。

- 現在、日本においてスマートフォンのユーザーは 50 %を超えており [1]、日常的な疑問をスマートフォンを通してインターネットで解決する環境が構築されつつあると考えられる。
- 近年、インターネットを介して、多数の人間に安価に仕事を依頼できる仕組みであるクラウドソーシングが注目されている。クラウドソーシングにより、専門家の翻訳と同品質程度の翻訳が得られることも示されている [1]。

また、実際にそのフレームワークに基づく実験を行った結果について報告する。

2　カバレッジ向上フレームワーク

提案するスマートフォンとクラウドソーシングを用いたカバレッジ向上のフレームワークについて説明する。

日常的表現は、専門性の高くない単語・句・短文であり、口語的な表現も含まれる。このような日常的表現を音声エージェントアプリによって収集する。音声エージェントアプリとはユーザーからの発話に反応して、スマートフォンの機能を動作させるものである。例えば、「明日の天気は？」という発話に反応して、「明日の天気は晴れです」のように反応する。この機能の一部と

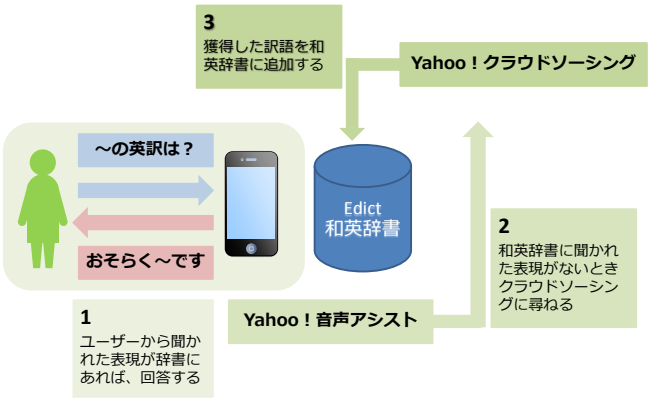

図 1: 集合知を用いた対訳知識カバレッジ向上フレームワーク

して簡易翻訳を考える。これは、ユーザーからの「アガるの翻訳は？」に反応して、「get nervous です」のように応答する。翻訳は対訳辞書を引き、エントリーに対応する訳を返す簡易的なものである。そのため、単語や複合語は翻訳可能であるが、定形表現ではない短文等の翻訳は行えない。「アガる」のようにユーザーが翻訳しようとしている表現を以下では「ターゲット表現」と呼ぶ。翻訳に失敗したターゲット表現をクラウドソーシングを用いて翻訳する。

ターゲット表現は、日常的表現の可能性が高いと考えられる。既に述べたように、スマートフォンを日常的に使う環境は構築されつつあり、専門的でない話し言葉が使われると考えられるからである。

獲得したターゲット表現及びその対訳を辞書に追加することにより、辞書のカバレッジを向上させることができる。これが、提案するフレームワークであり、これを繰り返すことでカバレッジの向上が見込まれる。

3　実験

3.1　ターゲット表現の収集

上記のフレームワークに基づく実験を行った。音声エージェントアプリは Yahoo！音声アシストを使用した。Yahoo!音声アシストは、Android 及び iPhone 上で利用できる音声エージェントアプリである。翻訳機能

Wisdom of Crowds Improves Coverage of Bilingual Knowledge
Ushiku Atsushi, Kawahara Daisuke, Kurohashi Sadao, Sassano Manabu
§†Kyoto University
‡Yahoo Japan Corporation
[1]http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2014/h25mediariyou_1sokuhou.pdf

の原言語は日本語、目標言語は英語である。翻訳機能に用いた辞書は Edict[2] を編集したものを用いた。元々の Edict の日本語エントリー数は 15 万程度であったが、編集により 5 万程度になった。編集は「有る」「ある」といった同一の日本語エントリーのうち、ひらがなのエントリーを削除、また、常用漢字以外の漢字を使うエントリーを削除した。これは、アプリケーション側の容量削減のために行われており、なるべく影響のでないような項目の削除を行ったものである[3]。ユーザーから聞かれたターゲット表現が辞書に含まれているときは、訳を返し、含まれていないときは、「『ターゲット表現』 英語」を検索した結果を表示する。

1ヶ月間の期間で、ユーザーが翻訳を意図する発話は 294 件存在した。その内訳は、翻訳を返すことができたものが 113 件、できなかったものが 181 件であった。翻訳に失敗した主な原因は、ターゲット表現が短文であることや、単語ではあるが Edict 編集時に削除してしまったこと、元々Edict がカバーしていないことである。

3.2 クラウドソーシングによる翻訳

これら翻訳に失敗したターゲット表現のうちから、重複するターゲット表現や Edict 編集時に削除したものを除いた、89 項目をクラウドソーシングで翻訳した。内訳は、単語・句が 47 個、短文が 42 個であった。ターゲット表現には翻訳が不可能と思われるものも含まれていた。例としては「Google」や、「おでこはこんにちは」のように元々日本語ではない、あるいは意味が理解できないものなどである。後者は、音声認識の失敗によるものと考えられる。

クラウドソーシングサービスとして Yahoo!クラウドソーシングを利用した。言語能力は特に問わなかったが、Yahoo！クラウドソーシングが日本語のサービスであるため、主に日本語を母国語とする者が翻訳したと考えられる。クラウドソーシングで 1 つのターゲット表現に対し 10 人のワーカーに回答を求めた。10 の訳のうちで最も数が多い訳を、ターゲット表現に対する対訳とした。得られた 89 項目のターゲットとその対訳を和英辞書に追加した。

翻訳の正確性について、英語のネイティブスピーカーに評価をしてもらった。評価結果について、表 1 に示す。短文で翻訳結果で採用されたものに関して、Web 上の翻訳サービスの機械翻訳の結果と一致するものが多数存在していた。このことが、対訳の評価で誤訳が多い理由である可能性があり、ターゲット表現を画像

翻訳評価	単語・句	短文
正しく翻訳できている	40	20
間違いではないが、この表現は使わない	0	2
誤訳である	7	20

表 1: 翻訳の正確性

にして提示することで、コピーアンドペーストを防ぐ等の対策が必要であると考えられる。

ターゲット表現には翻訳サービスでは翻訳できないような単語も存在した。たとえば「ちょい先」は Google 翻訳を用いると「Choi ahead」の訳になってしまう。「ちょい先」のクラウドソーシングでの訳語は「A little earlier」であり、これは Google 翻訳で「ちょっと先」を翻訳した「A little earlier」と一致する。これはワーカーが「ちょっと先」のように翻訳できる形に言い換えて翻訳サービスを利用したと推測される。

また、ターゲット表現の単語・句において、インターネット上の大規模辞書である Weblio 和英辞書[4] に載っていない項目は 47 項目中 18 項目存在した。具体的には、「復帰登板」や、「レスキュー車」等である。このように大規模辞書でもカバーしていない日常的表現が収集できた。

4 まとめ

本論文では、スマートフォンとクラウドソーシングを用いたカバレッジ向上フレームワークを提案し、実験によってカバレッジの向上を確認した。また、大規模辞書でもカバーしていない表現を獲得できた。

ログから、日常的表現の翻訳を聞かれる頻度が分かり、表現の重要性も調べる事ができる。今後はこのようなことについても分析していく予定である。一方で、翻訳にインターネット上の翻訳サービスを用いているワーカーが多いためか、句・短文の翻訳の信頼性には問題が残った。句・短文の翻訳精度の評価・向上を行い、カバレッジのさらなる向上を目指したい。

参考文献

[1] Omar F. Zaidan and Chris Callison-Burch. Crowdsourcing Translation: Professional Quality from Non-Professionals, In *Proceedings of HLT'11*, pp.1220-1229, 2011.

[2] Matt Post and Chris Callison-Burch and Miles Osborne. Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing,In *Proceedings of WMT'12*, pp.401-409, 2012.

[2]http://www.edrdg.org/jmdict/edict.html
[3]2015 年 1 月の段階で、容量の問題が改善され、同一エントリーを削除した 10 万項目程度の編集した Edict を用いている。
[4]http://ejje.weblio.jp/