

E-010

## クラウドソーシング翻訳のコスト削減のための翻訳者割り当て手法の提案

江原 遥\*<sup>1</sup>  
Yo Ehara馬場 雪乃\*<sup>2</sup>  
Yukino Baba内山 将夫\*<sup>1</sup>  
Masao Utiyama隅田 英一郎\*<sup>1</sup>  
Eiichiro Sumita\*<sup>1</sup>情報通信研究機構  
NICT\*<sup>2</sup>京都大学  
Kyoto University

## 1. はじめに

クラウドソーシングによって、不特定多数の翻訳者に翻訳タスクを依頼する事が可能となった。しかし、クラウドソーシング翻訳者の翻訳能力は不明であるため、翻訳の質は保証されていない。そのため、許容可能な訳の入手を保証するための品質管理が必須となる。

クラウドソーシングにおける品質管理の先行研究は、大別して2種類に分けられる(図1)。翻訳に限らない分類であるが、簡単のため、翻訳を例に説明する。**後処理型**では、1つの原文を多くの翻訳者に翻訳させ**冗長性**を持たせることにより、訳の品質を保証する。得られた複数の訳に対して後処理を施すことによって、許容可能な訳を得る。この手法では、冗長性を持たせるためコストが増加する欠点があるが、翻訳者の能力について全く事前知識なく高品質な訳を得ることができるという利点がある。一方、**前選択型**では、その時点で都合のつく翻訳者の中から、与えられた原文を最もうまく翻訳する事が期待される翻訳者を選び、その翻訳者に翻訳を依頼する。この手法では、コストは削減できるものの、「翻訳者の翻訳能力についての事前知識」を得る必要が生ずる。以上の2手法は独立しており、2手法を組み合わせることも可能である。

クラウドソーシング翻訳の品質管理では、これまで、殆どの研究で**後処理**を用いている。例えば、後編集[11, 1, 7, 10]や、クラウドソーシングを再度用いて翻訳品質を評価する方法[2]などがあげられる。これは、翻訳者の翻訳能力を推定するために必要な事前知識を得る事が困難であるためと推測される。

本稿では、クラウドソーシング翻訳における前選択手法を提案する。我々の知る限り、クラウドソーシング翻訳においては、前選択手法を提案するのは本研究が初めてである。前選択手法の性質により、我々の手法はコストを削減しながら高品質の訳を得られる。

提案手法では、前選択手法が必要となる「翻訳者の翻訳能力の事前知識」として、単語テストから推定可能な、翻訳者の原言語における語彙能力を用いる。翻訳者は、翻訳前に、原文中の単語のうち知らない単語をクリックを通じてシステムに通知する。このクリック情報を安価な単語テストとして用いる。提案手法は、単語テストから推定した翻訳者の語彙能力を通じて、翻訳能力をも推定し、最も翻訳能力の高い翻訳者に仕事を割り当てる。

語彙の情報の他に、翻訳能力の推定に役立つ情報として、過去の翻訳の翻訳品質に関する教師情報があげられるが、この教師情報の作成には別途コストを要する。本稿では、教師あり(全ての原文について教師情報が利用可能)、半教師あり(一部の原文についてのみ利用可能)、教師なし(全原文で利用不可能)の設定全てで動作する手法を提案する。

本研究の貢献は以下のとおりである。

- 本研究は、クラウドソーシング翻訳におけるコスト削減手法を初めて提案した。
- 教師なし学習の設定で、翻訳コストが33.4%減少した。

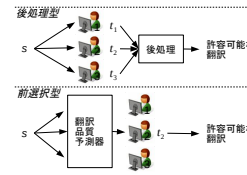


図1: 後処理型と前選択型の比較。原文  $s$  に対し、それぞれの翻訳者が翻訳  $t_1, t_2, t_3$  を生成する。**後処理型**では  $s$  を全翻訳者に翻訳させる必要があるのに対し、**前選択型**では中央の翻訳者にしか翻訳させずにすむ。

- 提案モデルの分析により、語彙能力と翻訳能力の間に有意な相関があることを明らかにした。

## 2. 準備

原文の集合を  $S$  とする。各原文  $s \in S$  を単語列とみなす。 $s$  は単語集合ではなく単語列なので同種の単語の重複を許す。簡単のため、 $s$  中の各単語を  $w \in s$  と書く。単語  $w$  の素性ベクトルを  $w \in \mathbb{R}^N$ 、原文  $s$  の素性ベクトルを  $s \in \mathbb{R}^M$  とする。ただし、 $N, M$  はそれぞれ単語と原文の素性ベクトルの次元数とする。翻訳者の集合を  $K$  とする。翻訳者  $k$  が訳した原文の集合を  $S_k \subset S$  とする。翻訳者  $k \in K$  が原文  $s \in S$  を翻訳した場合の翻訳品質(訳質)を表す訳質ラベルを  $z_{ks} \in \{1, 0\}$  とする。 $z_{ks} = 1$  の時、許容可能(acceptable)、 $z_{ks} = 0$  の時、許容不可能(unacceptable)とする。翻訳者  $k$  の単語  $w$  に対する語彙知識ラベルを  $y_{kw} \in \{0, 1\}$  で表す。翻訳者  $k$  が単語  $w$  を知らない場合、 $y_{kw} = 0$  とし、そうでない場合、 $y_{kw} = 1$  とする。本研究の目的は、翻訳者  $k$ 、新しい文  $s' \notin S$  に対して、翻訳品質を予測するモデル  $\Pr(z_{ks'} | s', \{w | w \in s'\}, k, \{y_{kw} | w \in s'\})$  を構築することである。ただし、 $s'$  は  $s'$  の素性ベクトルである。また、 $s'$  中の各単語  $w \in s'$  に対し、 $w$  はその素性ベクトル、 $y_{kw}$  は翻訳者  $k$  の  $w$  に対する語彙知識ラベルを表す。

$y_{kw}$  は常に与えられているものとする。教師あり学習の設定では、訓練データ中の  $z_{ks}$  は与えられているものとする。教師なし学習の設定では、 $z_{ks}$  は与えられていないものとし、推定する必要がある。半教師あり学習の設定では、訓練データ中の一部のデータのみに対して  $z_{ks}$  が与えられているものとし、与えられていない場合は教師なし学習の設定と同様に推定する。

## 3. 提案モデル

提案モデル(Translation Quality Predictor, TQP モデル)は、尤度と事前分布のモデルからなる。事前分布を変えることで、後述する複数のモデルを提案する。一方、尤度は、**翻訳品質モデル**(図2左)と、**語彙知識モデル**(図2右)の2つの部分モデルからなり、全ての提案モデルで共通である。どちらも既存の Rasch モデル[9]を用いている。Rasch モデルは、項目反応理論[3]と呼ばれる人間の能力や項目の難易度を予測・分析するためのモデルの1つであり、言語テストの他、教育や心理学の分野で広く使われている。しかし、Rasch モデルの

連絡先: 江原 遥, 情報通信研究機構 ユニバーサルコミュニケーション研究所 多言語翻訳研究室, 619-0289 京都府相楽郡精華町光台 3-5, 0774-98-6832, ehara@nict.go.jp.

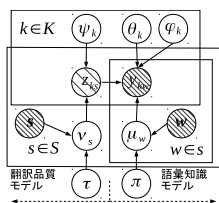


図2: 提案するTQPモデルのグラフィカルモデル表現。斜線の入った変数は、観測されたことを表す。訳質ラベル $z_{ks}$ のみ特殊であり、教師あり学習時には観測値になるが、教師なし学習時には未観測値になる。簡単のため事前分布は省略した。

考え方をもとに、前選択タスク用のモデルを提案するのは我々が初めてである。

全体として、図2のモデルが表す確率は次の通りである。ただし、 $\Pr[y_{kw} = 1 | z_{ks}] \equiv \sigma(\theta_k - \mu_w) z_{ks} \sigma(\phi_k - \mu_w)^{1-z_{ks}}$ とする。

$$\prod_k \prod_{s \in S_k} \prod_{w \in S} \Pr[y_{kw} | z_{ks}] \Pr[z_{ks}] \quad (1)$$

過学習を防ぐため、また、パラメタ間の関係の事前知識をモデルに入れるため、事前分布を定める。過学習を防ぐために、 $\pi, \tau, \theta_k, \phi_k$ については、通常通り、原点から離れた値に罰を与えるような事前分布を導入する。ただし、 $\mathcal{N}$ は、多次元正規分布の確率密度関数とする。

翻訳能力 $\psi_k$ については、語彙能力 $\theta_k, \phi_k$ との関係によって、2種類の事前分布を用意した。TQP1では、 $\psi_k$ についても、他の能力パラメタと同様、過学習を防ぐための事前分布を導入する。式で表すと、 $\Pr[\psi_k] \equiv \mathcal{N}(0, \xi^{-1})$ となる。一方、TQP2では、翻訳能力 $\psi_k$ と、語彙能力 $\theta_k, \phi_k$ を近づけるような事前分布を導入した。 $\theta_k$ と $\phi_k$ のどちらが大きき $\psi_k$ に影響するかについては事前知識が何もないので、 $\psi_k$ は $\frac{1}{2}(\theta_k + \phi_k)$ に近づけた。式で表すと、 $\Pr[\psi_k] \equiv \mathcal{N}(\frac{1}{2}(\theta_k + \phi_k), \xi^{-1})$ となる。

## 4. 評価

### 4.1 データセット

評価では、英日翻訳を対象にした、Wikipedia日英京都関連文書対訳コーパス\*1中の、対応する日本語訳が10文字以上の文から、104文をランダムに選んだ。

クラウドソーシングサービスには、ランサーズ\*2を利用した。ランサーズの作業者の大半は日本語母語話者と想定される。55人の翻訳者から、1,498件の翻訳を得た。翻訳コストは、翻訳者が1文を訳すのに付き、10円であった。原文1文に対して、平均14.4個の翻訳が得られた。1人の翻訳者は、平均27.2文を訳した。翻訳者が翻訳する前に、原文中の知らない単語をクリックさせた。その後、翻訳者に翻訳を開始させた。辞書の使用は翻訳時のみ解禁した。翻訳品質は、2名の英語に堪能な日本語母語話者であるアノテータによって、各翻訳に付与された。翻訳品質は5段階で付けられており、5のみが許容可能で、1から4が許容不可能な誤りを含むとした。許容可能・許容不可能の判断のカップ値は、0.619であり、“significant agreement”とみなせる[8]。

実験を通して、提案モデルでは、文のパープレキシティ、文中の未知語(Out-of-vocabulary)数、文中の語数、文中のコマ数、パープレキシティ/語数を文素性に用いた。単語素性は、3つのコーパスと1の単語難易度指標を用いた。具体的には、Google n-gramコーパス、現代アメリカ英語コーパスCOCA、Brownコーパスにおける1-gram確率の負の対数尤度と、12,000語・12段階の人手による難易度指標SVL 12000である。これらの素性は、日本人英語学習者の語彙能力を正

表1: 実験で比較した手法の一覧。“TQP”が名前に入っているものが提案手法、他のものがベースライン。

SVM	原文素性の名を用いたSVM
Rasch-based	Rasch-basedモデル[6]。これは、ユーザIDと原文素性のみを素性を用いたロジスティック回帰をのみ定める。
SVM+CatWF	原文素性に加え、原文中の単語を識別者が知らない単語と知っている単語に分け、それぞれの単語の素性ベクトルを単純に繋げた長い素性ベクトルとし、SVMをかけたもの。
LR+CatWF	単語ごとに素性で、ロジスティック回帰を分類に用いたもの。
SVM+AvgWF	原文素性に前文、原文中の単語を識別者が知らない単語と知っている単語に分け、それぞれ単語の素性ベクトルの平均ベクトルを素性ベクトルとし、SVMをかけたもの。
LR+AvgWF	単語ごとに素性で、ロジスティック回帰を分類に用いたもの。
TQP1	TQPモデル1
TQP2	TQPモデル2
SEMI-TQP1	教師なし・半教師あり学習設定でのTQPモデル1
SEMI-TQP2	教師なし・半教師あり学習設定でのTQPモデル2

表2: 訓練データ量に対するコスト削減率。それぞれの列で、太字が最良値、下線を引いた値が次に良い値を示す。アスタリスクは、太字が下線に対して有意である事を示す。(\*:  $p < .01$ , \*\* $p < .10$ , ウィルコクソンの符号順位検定を用いた)。

訓練データ量	0	10	55	112	896
Rasch-based	-	0.149	0.297	0.338	0.349
SVM+CatWF	-	0.128	0.199	0.268	0.267
LR+CatWF	-	0.233	0.321	0.341	0.341
SVM+AvgWF	-	0.269	0.322	0.340	0.335**
TQP1	-	0.256	0.298	0.337	0.337
TQP2	-	0.341	0.382	0.338	0.337
SEMI-TQP1	0.092	0.261	0.301	0.339	0.334
SEMI-TQP2	0.334**	0.344**	0.342*	0.341*	0.312

確に測定するのに有効である事が示されている[6]。実験は、5-fold nested cross validationの設定で行った。

### 4.2 コスト削減率による評価

実際に許容可能な訳を得るまでの注文数の減少率(コスト削減率)によって提案手法の性能を評価した(表2)。例えば、3人の翻訳者のうち、1人だけが許容可能な訳を返す「良い翻訳者」とする。ランダムに翻訳者を選び良い翻訳者に当たるまでには平均 $1 \times \frac{1}{3} + 2 \times \frac{2}{3} \times \frac{1}{2} + 3 \times \frac{2}{3} \times \frac{1}{2} = 2$ 回注文しなければならぬが、提案手法で良い翻訳者を最初から当てられたとすれば、注文数は1回になり、コスト削減率は $1/2 = 0.5$ となる。コスト削減率による評価に用いた実験設定は、精度実験の時と同じである。

訓練データ数 $TS = 0$ の時、SEMI-TQP2は、33.4%のコストを削減している表2。一方、SVM+AvgWFでは、このコスト削減率を得るのに、全翻訳者数と同程度の55個の評価が必要となる。すなわち、既存手法では、翻訳者数と同じ数だけ翻訳品質の教師データを作成しなければならず、品質評価のコストが高い事がわかる。

また、翻訳能力と語彙能力の相関関係を仮定するSEMI-TQP2, TQP2が、仮定しないSEMI-TQP1, TQP1を、それぞれ常に上回っていることから、この仮定が翻訳能力の推定に有効である事が示唆される。また、提案モデルのうち、“SEMI”のついているものの方が、ついでにないものよりも性能が良いことから、半教師あり学習が有効に働いていることもわかる。

## 参考文献

- AMBATI, V., VOGEL, S., AND CARBONELL, J. Collaborative workflow for crowdsourcing translation. In *Proc. of CSCW* (2012), pp. 1191–1194.
- BABA, Y., AND KASHIMA, H. Statistical quality estimation for general crowdsourcing tasks. In *Proc. of KDD* (New York, NY, USA, 2013), ACM, pp. 554–562.
- BAKER, F. B., AND KIM, S.-H. *Item Response Theory: Parameter Estimation Techniques*, second ed. Marcel Dekker, New York, 2004.
- BEGLAR, D. A rasch-based validation of the vocabulary size test. *Language Testing* 27, 1 (2010), 101–118.
- DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. Ser. B* 39, 1 (1977), 1–38.
- EHARA, Y., SHIMIZU, N., NINOMIYA, T., AND NAKAGAWA, H. Personalized reading support for second-language web documents. *ACM Trans. Intell. Syst. Technol.* 4, 2 (2013).
- GREEN, S., HEER, J., AND MANNING, C. D. The efficacy of human post-editing for language translation. In *Proc. of CHI* (2013), pp. 439–448.
- LANDIS, J. R., AND KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics* 33 (1977), 159–174.
- RASCH, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen, 1960.
- YAN, R., GAO, M., PAVLIC, E., AND CALLISON-BURCH, C. Are two heads better than one? crowdsourced translation via a two-step collaboration of non-professional translators and editors. In *Proc. of ACL* (2014), pp. 1134–1144.
- Z Aidan, O. F., AND CALLISON-BURCH, C. Crowdsourcing translation: Professional quality from non-professionals. In *Proc. of ACL-HLT* (2011), pp. 1220–1229.

\*1 [http://alaginrc.nict.go.jp/WikiCorpus/index\\_E.html](http://alaginrc.nict.go.jp/WikiCorpus/index_E.html)

\*2 <http://www.lancers.jp>