3J4-2in

単語テストを利用した翻訳品質事前予測モデル

江原 遥 *1 Yo Ehara　馬場 雪乃 *2 Yukino Baba　内山 将夫 *1 Masao Utiyama　隅田 英一郎 *1 Eiichiro Sumita

*1*3*4情報通信研究機構 NICT　*2京都大学 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，前選択手法で必要となる「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．提案手法の背後には，原言語における語彙能力と翻訳能力は相関するという仮定がある．これは，翻訳者の大部分が翻訳先言語の母語話者で，翻訳元言語（原言語）を外国語として学習する者であることと，語彙能力と全般的な外国語の能力との相関を示す既存研究 [4] による．

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.

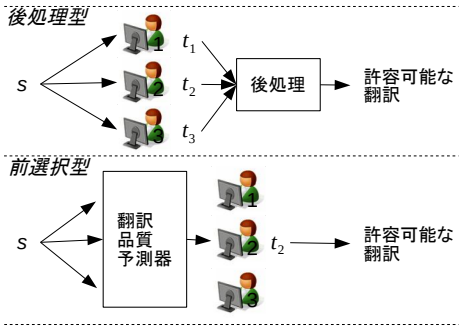

図 1: 後処理型と前選択型の比較．原文 s に対し，それぞれの翻訳者が翻訳 t_1, t_2, t_3 を生成する．**後処理型**では s を全翻訳者に翻訳させる必要があるのに対し，**前選択型**では中央の翻訳者にしか翻訳させずにすむ．

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．
- 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 $\mathcal{S}$ とする．各原文 $s \in \mathcal{S}$ 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 $\boldsymbol{w} \in \mathbb{R}^N$，原文 s の素性ベクトルを $\boldsymbol{s} \in \mathbb{R}^M$ 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 $\mathcal{K}$ する．翻訳者 k が訳した原文の集合を $\mathcal{S}_k \subseteq \mathcal{S}$ とする．翻訳者 $k \in \mathcal{K}$ が原文 $s \in \mathcal{S}$ 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

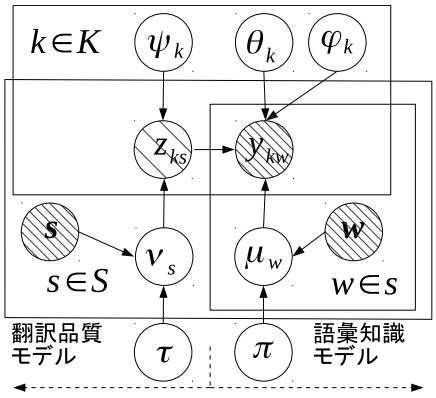

図 2: 提案する TQP モデルのグラフィカルモデル表現．斜線の入った変数は，観測されたことを表す．訳質ラベル z_{ks} のみ特殊であり，教師あり学習時には観測値になるが，教師なし学習時には未観測値になる．簡単のため事前分布は省略した．

ないと答えた場合，$y_{kw} = 0$ 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 $s' \notin \mathcal{S}$ に対して，翻訳品質を予測するモデル $\Pr(z_{ks'} | \boldsymbol{s}', \{\boldsymbol{w} | w \in s'\}, k, \{y_{kw} | w \in s'\})$ を構築する ことである．ただし，$\boldsymbol{s}'$ 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，$\boldsymbol{w}$ 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．

3.1 尤度

翻訳品質モデルでは，翻訳者 k が，各文 s を許容可能（$z_{ks} = 1$）に訳す確率を次式で定義する．[*1]

$$\Pr[z_{ks} = 1] = \sigma(\psi_k - \nu_s) \tag{1}$$

ここで，σ は，ロジスティックシグモイド関数であり，実数 a に対して $\sigma(a) \equiv \frac{1}{\exp(-a)+1}$ と定義される．推定するパラメタは 2 種あり，翻訳能力 ψ_k，文 s の翻訳難易度 ν_s である．σ の

*1 本稿ではパラメタはギリシャ文字で表し，パラメタによる条件付けは省く．例えば $\Pr[z_{ks} = 1 | \psi_k, \nu_s]$ は単に $\Pr[z_{ks} = 1]$ と書く．

表 1: 提案モデルの事前分布．ψ の事前分布を変えることによって，2 種類のモデルを構築する．

$\boldsymbol{\pi}$		$\Pr[\boldsymbol{\pi}] \equiv \mathcal{N}(\mathbf{0}, \lambda^{-1}\boldsymbol{I})$
$\boldsymbol{\tau}$		$\Pr[\boldsymbol{\tau}] \equiv \mathcal{N}(\mathbf{0}, \xi^{-1}\boldsymbol{I})$
θ_k		$\Pr[\theta_k] \equiv \mathcal{N}(0, \lambda^{-1})$
ϕ_k		$\Pr[\phi_k] \equiv \mathcal{N}(0, \lambda^{-1})$
ψ	TQP1	$\Pr[\psi_k] \equiv \mathcal{N}(0, \xi^{-1})$
	TQP2	$\Pr[\psi_k] \equiv \mathcal{N}(\frac{1}{2}(\theta_k + \phi_k), \xi^{-1})$

性質により，(1) は，翻訳能力 ψ_k が翻訳難易度 ν_s を超えた時，すなわち $\psi_k > \nu_s$ であるときのみ $\Pr[z_{ks} = 1] > \Pr[z_{ks} = 0]$ となり，予測器では $z_{ks} = 1$ と判別される．さらに，[6] をもとに難易度 ν_s を，設問 s に対する素性ベクトル $\boldsymbol{s}$ を入れられるように拡張する．この場合，重みベクトル $\boldsymbol{\tau}$ を導入して，難易度を $\nu_s = \boldsymbol{\tau}^\top \boldsymbol{s}$ と定義する．k が s を許容可能に訳した時，成功（$z_{ks} = 1$）とし，そうでない時，失敗とする．

語彙知識モデルでは，翻訳品質モデルと同様，翻訳力と単語の難易度を定義する．単語 w の難易度は，単語素性 $\boldsymbol{w}$ と，対応する重みベクトル $\boldsymbol{\pi}$ を用いて，$\mu_w \equiv \boldsymbol{\pi}^\top \boldsymbol{w}$ と定義する．語彙能力については，翻訳者 k が許容可能な翻訳 ($z_{ks} = 1$) を書いた時の語彙能力 θ_k と，許容不可能な翻訳 ($z_{ks} = 0$ を書いた時の語彙能力 ϕ_k の 2 種類に分ける．

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，$\Pr[y_{kw} = 1 \mid z_{ks}] \equiv \sigma(\theta_k - \mu_w)^{z_{ks}} \sigma(\phi_k - \mu_w)^{1 - z_{ks}}$ とする．

$$\prod_k \prod_{s \in \mathcal{S}_k} \prod_{w \in s} \Pr[y_{kw} \mid z_{ks}] \Pr[z_{ks}] \tag{2}$$

3.2 事前分布

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（表 1）．過学習を防ぐために，$\boldsymbol{\pi}$, $\boldsymbol{\tau}$, θ_k, ϕ_k については，原点から離れすぎ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 ψ_k については，語彙能力 θ_k, ϕ_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，ψ_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．一方，**TQP2** では，翻訳能力 ψ_k と，語彙能力 θ_k，ϕ_k を近づけるような事前分布を導入した．θ_k と ϕ_k のどちらが大きく ψ_k に影響するかについては事前知識が何もないので，ψ_k は $\frac{1}{2}(\theta_k + \phi_k)$ に近づけた．

3.3 教師ありの場合のパラメタ推定

パラメタ推定には，Maximum A Posteriori (MAP) 推定を用いた．下記の事後確率を最大化する．ここで，“Priors” は，前節で定義した事前分布である．

$$\begin{aligned} &L(\{\theta_k\}_K, \{\phi_k\}_K, \{\psi_k\}_K, \boldsymbol{\pi}, \boldsymbol{\tau}) \\ &= \text{Priors}[\{\theta_k\}_K, \{\phi_k\}_K, \{\psi_k\}_K, \boldsymbol{\pi}, \boldsymbol{\tau}] \\ &\times \prod_k \prod_{s \in \mathcal{S}_k} \prod_{w \in s} \Pr[y_{kw} \mid z_{ks}] \Pr[z_{ks}] \end{aligned} \tag{3}$$

事後確率 (3) の最大化は，負の対数事後確率を最大化することと等価である．z_{ks} が固定されている教師あり学習では，(3) の負の対数事後確率はパラメタに関して凸関数であり，大

域的最適解が求まる．このことは，L2 正則化ロジスティック回帰の負の対数事後確率が凸関数である事を用いて容易に証明できる．z_{ks} が固定されている時，我々は，L-BFGS 法を用いて最大化を行った．

3.4 教師なし・半教師ありの場合のパラメタ推定

教師なしの設定の場合は，z_{ks} が不明であるので，これを EM アルゴリズム [5] を用いながら推定する．半教師ありの設定の場合は，教師なしと同様に E ステップで z_{ks} を推定した後，教師データが与えられている z_{ks} のみ，所与の値と単純に置き換えて M ステップに進める．

記法を簡単にするため，下記の表記を導入する．

α_{kw} 文 s が許容可能な時，k が単語 w を知っている確率．すなわち，$\alpha_{kw} \equiv \Pr[y_{kw}=1 \mid z_{ks}=1] = \sigma(\theta_k - \mu_w)$.

β_{kw} 文 s が許容不可能な時，k が単語 w を知っている確率．すなわち，$\beta_{kw} \equiv \Pr[y_{kw}=1 \mid z_{ks}=0] = \sigma(\phi_k - \mu_w)$.

γ_{ks} k が s を訳した時，許容可能な訳が入手できる確率．すなわち，$\gamma_{ks} \equiv \Pr[z_{ks}=1] = \sigma(\psi_k - \nu_s)$.

E ステップでは，次の式で z_{ks} を更新する．

$$\Pr\left[z_{ks}=1 \mid \{y_{kw}\}_{w\in s}\right] \propto \gamma_{ks}\Big(\prod_{w\in s} \alpha_{kw}^{y_{kw}} (1-\alpha_{kw})^{(1-y_{kw})}\Big)$$

$$\Pr\left[z_{ks}=0 \mid \{y_{kw}\}_{w\in s}\right] \propto (1-\gamma_{ks})\Big(\prod_{w\in s} \beta_{kw}^{y_{kw}} (1-\beta_{kw})^{(1-y_{kw})}\Big)$$

E ステップの後，z_{ks} を新しい z_{ks} を用いて更新する．この新しい z_{ks} を $z'_{ks} = \Pr\left[z_{ks}=1 \mid \{y_{kw}\}_{w\in s}\right]$ と定義する．

M ステップでは，Q 関数を最大化する．Q 関数の形は，z_{ks} が z'_{ks} に置き換えられる事を除いて，(3) で示した関数と同じである．Q 関数の最大化においては z_{ks} は固定されているので，Q 関数の対数の負値は凸関数となり，L-BFGS 法を用いて大域的最適解を求めることができる．

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 [*2] 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 [*3] 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パ値は，0.619 であり，“significant agreement” とみなせる [8]．

*2 http://alaginrc.nict.go.jp/WikiCorpus/index_E.html
*3 http://www.lancers.jp

表 2: 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	原文素性のみを用いた SVM.
Rasch-based	(1) で示される，Rasch-based モデル．これは，ユーザ ID と原文素性のみを素性に用いたロジスティック回帰とみなせる．
SVM+CatWF	原文素性に加え，原文中の単語を翻訳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て長い素性ベクトルとし，SVM をかけたもの．
LR+CatWF	上記と同じ素性で，ロジスティック回帰を分類に用いたもの．
SVM+AvgWF	原文素性に加え，原文中の単語を翻訳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	上記と同じ素性で，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

4.2 素性

実験を通して，TQP モデルでは，表 3 に示す同じ素性を用いた．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（表 4）．これらの素性は，日本人英語学習者の語彙能力を正確に測定するのに有効である事が示されている [6]．

4.3 精度実験

まず，翻訳品質の予測精度に関する実験を行った．比較した手法は，下記の通りである．ここで，“TQP” が名前に入っているものが，提案手法である．

実験は，5-fold nested cross validation の設定で行った．全てのモデルにはハイパーパラメタがあるため，このハイパーパラメタを開発セット（validation set）を用いてチューニングした．全てのハイパーパラメタは，10^{-3}, $10^{-2.4}$, $10^{-1.8}$, $10^{-1.2}$, $10^{-0.6}$, 1, $10^{0.6}$, $10^{1.2}$, $10^{1.8}$, $10^{2.4}$, $10^{3.0}$ の中から，グリッドサーチで選んだ．ハイパーパラメタを選んだ後，テストセットで性能を評価した．訓練セット，開発セット，テストセットは，全て disjoint である．

表 5 に実験結果を示す．2 値分類であるため，頻度の多い方のラベルの割合 0.5527 が，チャンスレートであり，1 つのベースラインとなる．訓練データ数 0（$TS=0$）の時，提案手法 **SEMI-TQP2** のみがチャンスレートを有意に上回っていた．これは，**SEMI-TQP2** が仮定する翻訳能力と語彙能力の間の相関関係が，実際に予測に有効である事を示唆している．

また，**LR+AvgWF** は，右端の列でのみ，提案手法 **SEMI-TQP2** と **TQP2** をわずかに上回っている．これは，翻訳能力と語彙能力の相関関係を仮定しないと，提案手法を打ち負かすのに 896 件もの翻訳品質に関する訓練データが必要になる事を示唆している．

表 3: 文素性 $\boldsymbol{s}$.

文のパープレキシティ．パープレキシティの計算には，標準的な Kneser-Ney スムージング付き 3-gram 言語モデルを用いた [*4]．
文中の未知語 (Out-of-vocabulary) 数．上記の言語モデルに現れない語を未知語とみなした．
文中の語数．
文中のコンマ数．
パープレキシティ/語数，未知語数/語数，コンマ数/語数も素性に加えた．

表 4: 単語素性 $\boldsymbol{w}$．.

Google n-gram コーパス [*5] における 1-gram 確率の負の対数尤度
現代アメリカ英語コーパス COCA[*6] における 1-gram 確率の負の対数尤度
Brown コーパス [*7] における 1-gram 確率の負の対数尤度
12 段階の難易度指標 SVL 12000[*8]．計 12,000 語について難易度が人手で与えられている．日本人英語学習者を考慮して作成された．

表 5: 訓練データ量に対する精度．ベースラインとなる最頻ラベルの割合は 0.553 である． 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
SVM	-	0.541	0.574	0.577	0.661
Rasch-based	-	0.550	0.597	0.635	0.694
SVM+CatWF	-	0.523	0.543	0.534	0.547
LR+CatWF	-	0.528	0.559	0.559	0.566
SVM+AvgWF	-	0.545	0.603	0.634	0.693
LR+AvgWF	-	0.569	0.619	0.648	**0.704****
TQP1	-	0.571	0.604	0.637	0.695
TQP2	-	0.684	0.685	0.689	0.696
SEMI-TQP1	0.501	0.569	0.609	0.639	0.696
SEMI-TQP2	**0.681****	**0.686****	**0.687****	**0.691****	0.699

4.4 コスト削減率による評価

次に，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 6）．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 $1 * \frac{1}{3} + 2 * \frac{2}{3} * \frac{1}{2} + 3 * \frac{2}{3} * \frac{1}{2} = 2$ 回注文しなければならないが，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 6．一方，**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，ついてい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．

4.5 能力パラメタの散布図

翻訳能力と語彙能力の相関関係を仮定する TQP2 では，相関が確認できる（図 3 右側）．重要な知見として，この仮定を置かない TQP1 でも，訓練データ量を増やすと，この相関関係が有意に確認できる（同図左下）．

表 6: 訓練データ量に対するコスト削減率．太字，下線，アスタリスクの意味は表 5 と同様である．

訓練データ量	0	10	55	112	896
SVM	-	0.175	0.297	0.328	0.326
Rasch-based	-	0.180	0.294	0.331	0.329
SVM+CatWF	-	0.123	0.199	0.258	0.288
LR+CatWF	-	0.126	0.203	0.264	0.291
SVM+AvgWF	-	0.244	0.321	**0.341**	0.347
LR+AvgWF	-	0.269	0.322	0.340	**0.355****
TQP1	-	0.256	0.298	0.337	0.333
TQP2	-	0.340	0.332	0.338	0.337
SEMI-TQP1	0.092	0.264	0.301	0.339	0.334
SEMI-TQP2	**0.334****	**0.344****	**0.342***	**0.341***	0.342

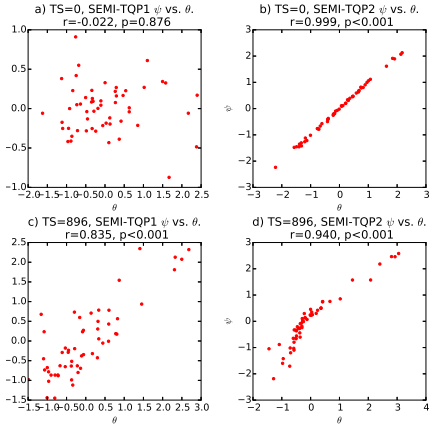

図 3: 能力パラメタの散布図．各点は 55 人の各翻訳者，r は相関係数，p は帰無仮説 $r = 0$ に対する p 値を表す．

謝辞

本研究は JSPS 科研費 26730115 の助成を受けた．

参考文献

[1] Ambati, V., Vogel, S., and Carbonell, J. Collaborative workflow for crowdsourcing translation. In *Proc. of CSCW* (2012), pp. 1191–1194.

[2] 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.

[3] Baker, F. B., and Kim, S.-H. *Item Response Theory: Parameter Estimation Techniques*, second ed. Marcel Dekker, New York, 2004.

[4] Beglar, D. A rasch-based validation of the vocabulary size test. *Language Testing 27*, 1 (2010), 101–118.

[5] 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.

[6] 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).

[7] 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.

[8] Landis, J. R., and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics 33* (1977), 159–174.

[9] Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Danish Institute for Educational Research, Copenhagen, 1960.

[10] Yan, R., Gao, M., Pavlick, 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.

[11] Zaidan, O. F., and Callison-Burch, C. Crowdsourcing translation: Professional quality from non-professionals. In *Proc. of ACL-HLT* (2011), pp. 1220–1229.

∗4 http://www.statmt.org/moses/?n=moses.baseline
∗5 https://catalog.ldc.upenn.edu/LDC2006T13
∗6 http://corpus.byu.edu/coca/
∗7 http://www.nltk.org/nltk_data/
∗8 http://www.alc.co.jp/vocgram/article/svl/