Assessing Translation Ability through Vocabulary Ability Assessment

(appeared in IJCAI'16)

江原 遥（産総研）, 馬場 雪乃（京大）
内山 将夫 (NICT), 隅田 英一郎 (NICT)

2016年8月9日 ERATO感謝祭 Season III

概要

語彙テストを用いた、簡便な翻訳者評価法を提案

従来の翻訳者評価：
高コスト

翻訳者

翻訳テスト・採点 ¥

翻訳テスト・採点 ¥

翻訳テスト・採点 ¥

評価結果

本研究による翻訳者評価：
低コスト

翻訳者

翻訳テスト・採点 ¥

語彙テスト

語彙テスト

評価結果

クラウドソーシング翻訳では品質が課題

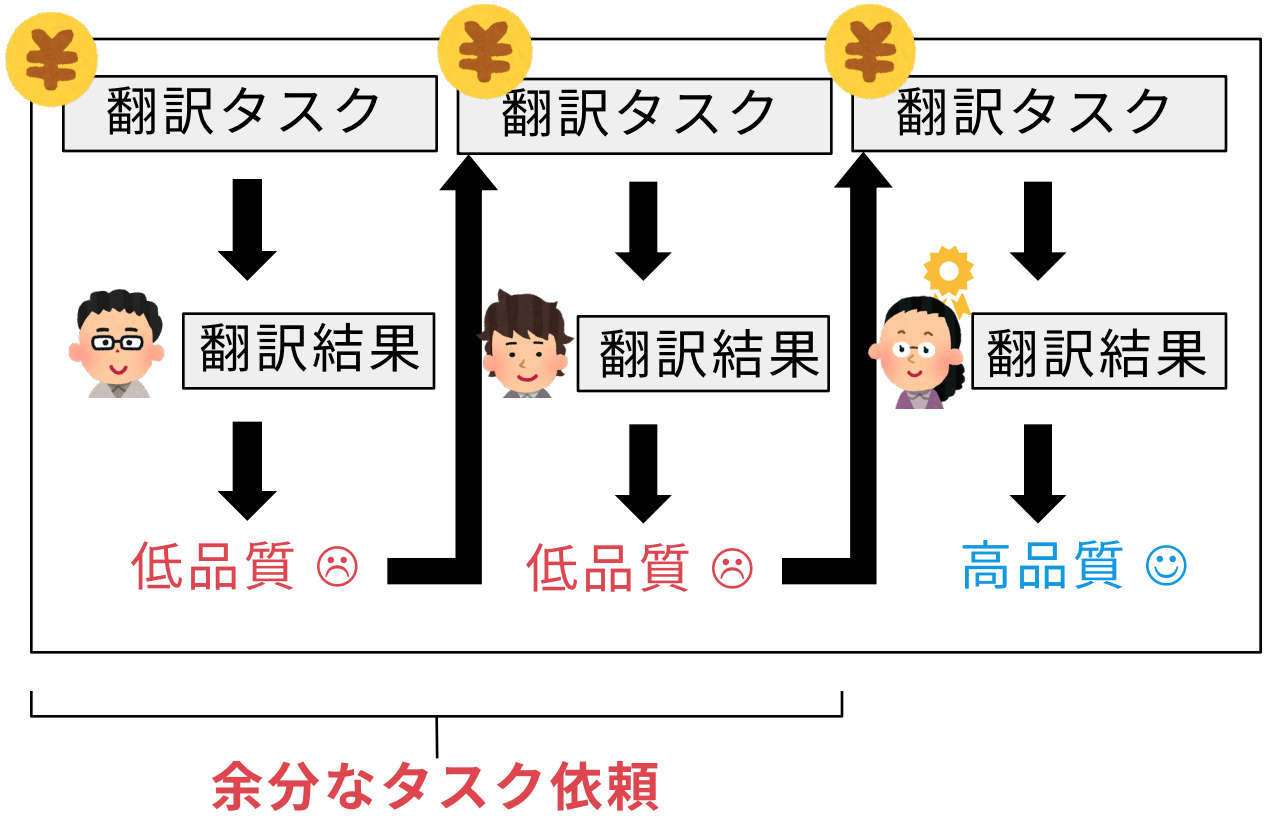
- クラウドソーシング翻訳では、非専門家も翻訳作業に参加
- 非専門家が、低品質の翻訳を返すことがしばしばある

クラウドソーシング翻訳の結果例

原文	The four brothers died of smallpox in 737 and there was a rumor that it was a punishment for driving the prince to suicide.
良い翻訳結果	4人の兄弟は737年に天然痘で死亡し、それは皇子を自殺に追い込んだことへの罰だというううわさが広まった。
悪い翻訳結果	4人の兄弟は737年スモールポックスで死んだ。そして自殺するためにプリンスを運転して消えたというううわさがたった。

低品質な翻訳は依頼回数（＝費用）を増加させる

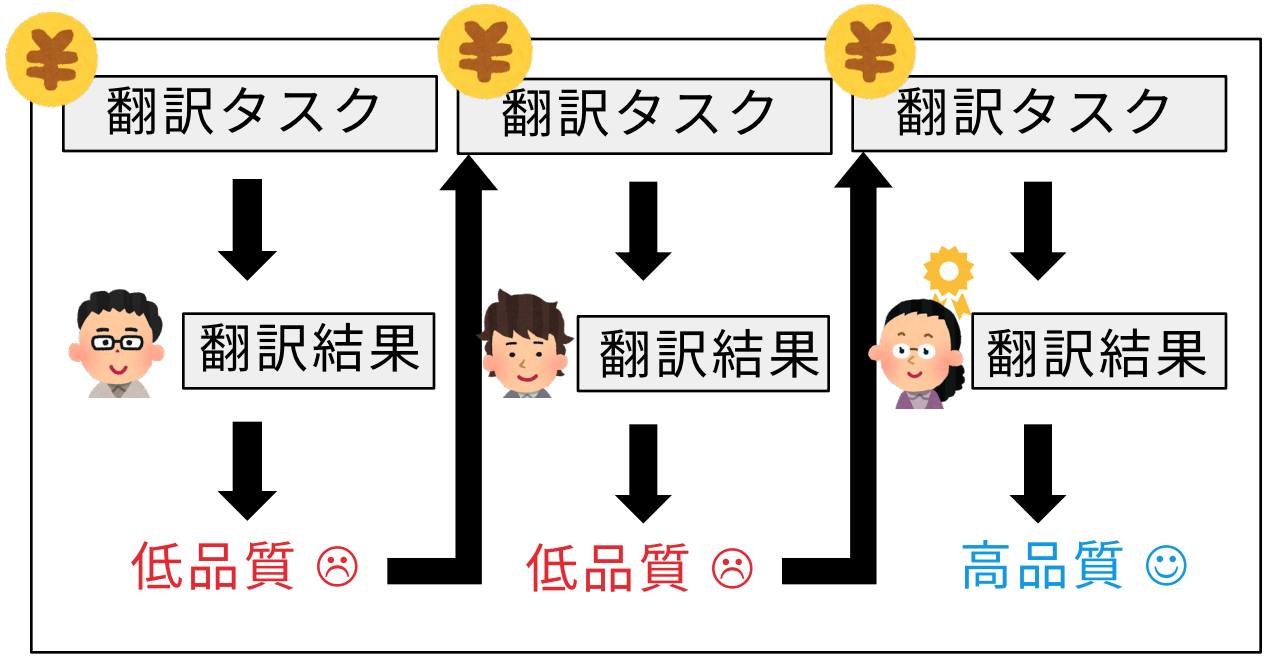
- 十分な品質の翻訳が得られるまでタスクを投げ続ける場合、低品質な翻訳は余分なタスク依頼を生じさせる



背景：クラウドソーシング翻訳 (3)

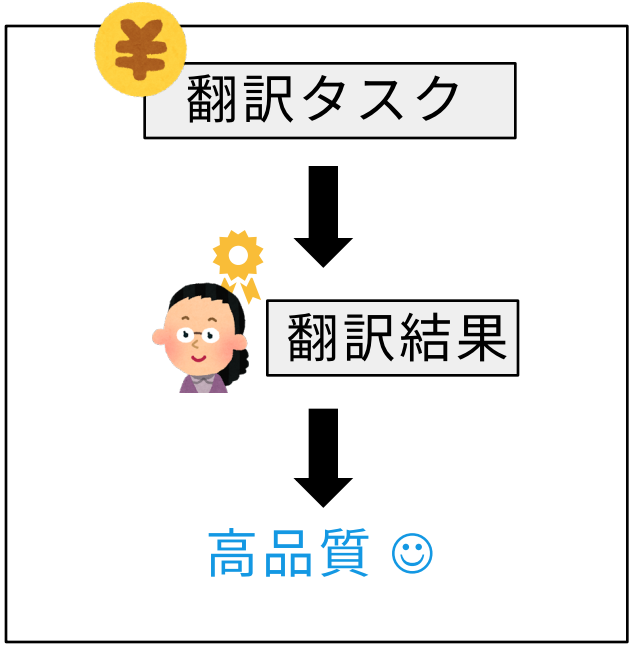
良い翻訳者が事前にわかれば依頼を効率化できる

「良い翻訳者」が
事前にわからない場合



3回のタスク依頼が必要

「良い翻訳者」が
事前にわかる場合



1回のタスク依頼で済む

翻訳テストによる翻訳者評価は正確だが高コスト

- 翻訳者評価の従来法：翻訳テスト
- 翻訳テストは正確だが採点コストが掛かりスケールしない
- 簡便な翻訳者評価方法が求められている

翻訳テストの例

Translate the following sentences:
Q1: “The accident triggered the …”
その事件がきっかけとなり…

Q2: “His style was handed down …”
彼の手法は江戸時代の終わりから続く…

専門家が採点

自己申告語彙テストの利用で採点コストを減らす

- 自己申告語彙テスト
  - 各単語の意味を知っているかどうかを答える
  - 自己申告なので、専門家による採点は不要
- 翻訳能力と語彙テストの回答を関連づける確率モデルを構築

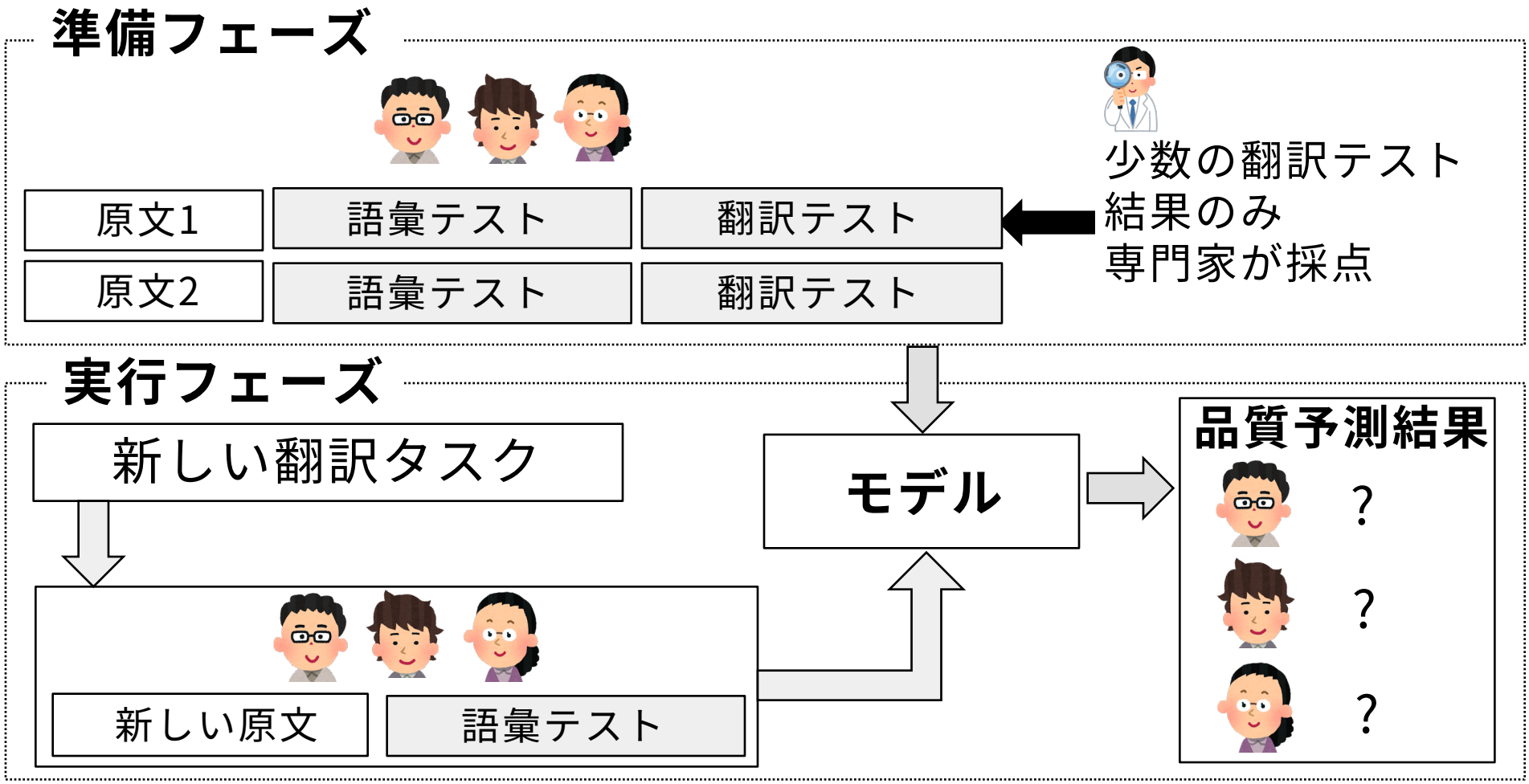
自己申告語彙テストの例

意味を知っている単語をチェックしてください

Q1. “**The accident triggered the establishment of the Law for the …**”

☑ ☑ ☐ ☑

翻訳・語彙テストを同時実施、一部の翻訳だけ採点



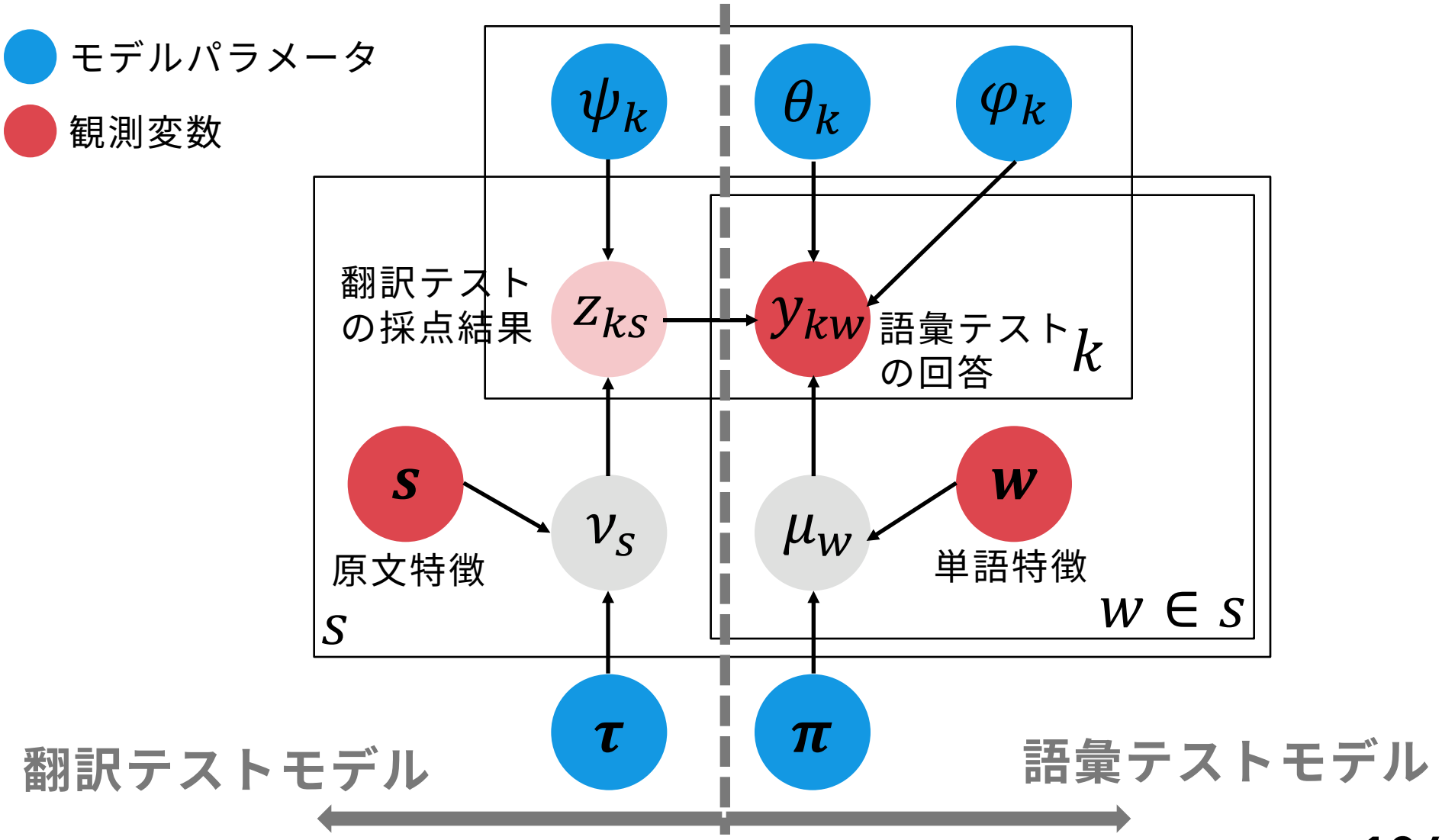
翻訳・語彙テストの結果から品質予測器を学習

- 入力

$\{k\}_k$	翻訳者集合
$\{s\}_s$	翻訳・語彙テストの原文集合, 原文は単語の集合で表現
$\{z_{ks}\}_{k,s}$	翻訳テストの採点結果の集合, $z_{ks} \in \{0, 1, \text{“?”}\}$
$\{y_{kw}\}_{k,w \in s}$	語彙テストの回答結果の集合, $y_{kw} \in \{0, 1\}$
$\{\boldsymbol{s}\}_s$	原文特徴の集合
$\{\boldsymbol{w}\}_{w \in s}$	単語特徴の集合

- 出力：品質予測器 $\Pr\left[z_{ks^*} = 1 \mid \{y_{kw}\}_{w \in s^*}\right]$
  （新しい原文s^*に対して翻訳者kが、
  高品質の翻訳を返す確率を予測）

翻訳・語彙テストの結果を関連づける確率モデル



翻訳能力と原文難易度に従い翻訳テスト結果が決まる

- 翻訳者kが原文sに対して高品質の翻訳を返す確率を
  ラッシュモデルで表現

$$\Pr\left[z_{ks}=1\right]=\frac{1}{1+\exp\left(-\left(\psi_k-\nu_s\right)\right)}$$

翻訳能力　　原文難易度

語彙能力と単語難易度に従い語彙テスト回答が決まる

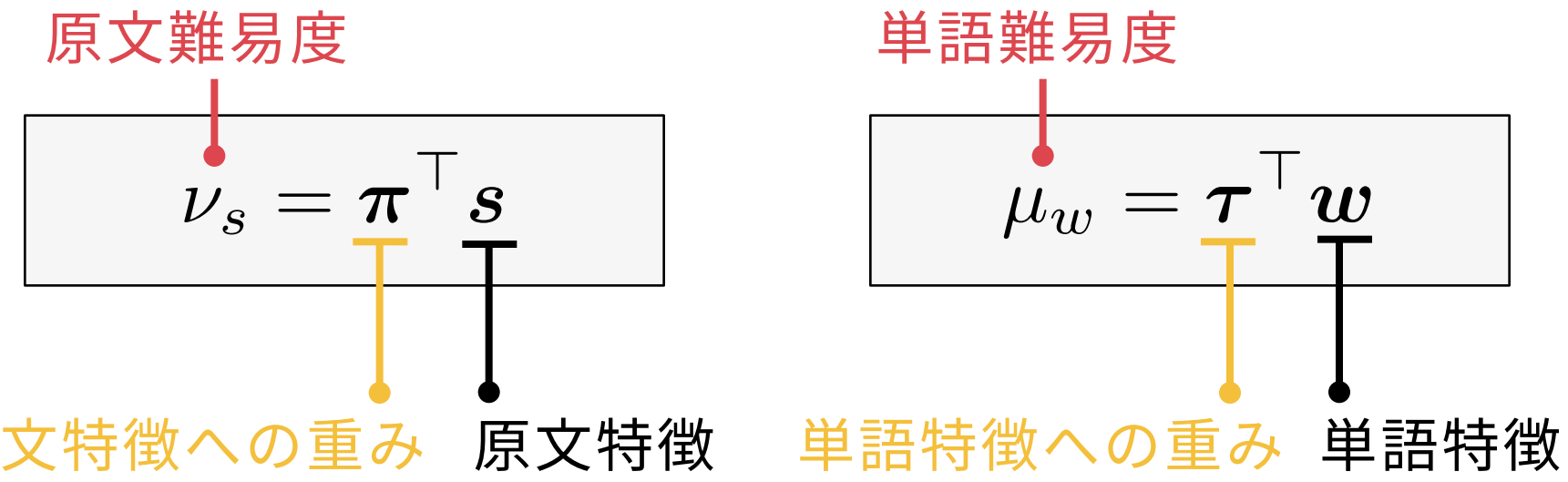
- 翻訳者kが単語wを「知っている」と答える確率を
  ラッシュモデルで表現

$$\Pr\left[y_{kw}=1 \mid z_{ks}=1\right]=\frac{1}{1+\exp\left(-\left(\theta_k-\mu_w\right)\right)}$$

$$\Pr\left[y_{kw}=1 \mid z_{ks}=0\right]=\frac{1}{1+\exp\left(-\left(\phi_k-\mu_w\right)\right)}$$

原文・単語の難易度は各特徴に従って決まる

- 原文・単語難易度のモデル：



- 最終的なモデルパラメータ：
  各翻訳者kの翻訳能力 (ψ_k), 語彙能力 (θ_k, ϕ_k),
  原文特徴への重み $(\boldsymbol{\pi})$, 単語特徴への重み $(\boldsymbol{\tau})$

翻訳能力と語彙能力を近づける事前分布を導入

- モデルの尤度：

S_k: 翻訳者kが翻訳した原文集合
λ and ξ: ハイパーパラメータ

$$L\left(\{\psi_k\},\{\theta_k\},\{\phi_k\},\boldsymbol{\pi},\boldsymbol{\tau}\right)$$
$$= \prod_k \mathcal{N}\left(\psi_k \mid \theta_k, \lambda^{-1}\right) \mathcal{N}\left(\theta_k \mid 0, \lambda^{-1}\right) \mathcal{N}\left(\phi_k \mid 0, \lambda^{-1}\right)$$
$$\times \mathcal{N}\left(\boldsymbol{\pi} \mid \xi^{-1}\boldsymbol{I}\right) \times \mathcal{N}\left(\boldsymbol{\tau} \mid \xi^{-1}\boldsymbol{I}\right)$$
$$\times \prod_k \prod_{s\in\mathcal{S}_k} \prod_{w\in s} \Pr\left[y_{kw} \mid z_{ks}\right] \Pr\left[z_{ks}\right]$$

翻訳能力 ψ_k が
語彙能力 θ_k に近づく
ような事前分布を導入

- EMアルゴリズムでモデルパラメータを推定
  - 翻訳テストの採点結果 z_{ks}が既知の場合、
    $\mathrm{E}\left[z_{ks}\right] = \Pr\left[z_{ks}=1 \mid \{y_{kw}\}_{w\in s}\right]$ の代わりにz_{ks}を直接使う

クラウドソーシングで1,498件の英日翻訳結果を収集

- 「Wikipedia日英京都関連文書対訳コーパス」から英文を選択
- ランサーズで英日翻訳タスクを依頼（1件10円）
  - 翻訳時に語彙テストも受験させる
- 各翻訳結果に品質ラベル（「高品質」or「低品質」）を付与

データセット概要

原文数	104
翻訳結果数	1,498
翻訳者数	55

言語非依存の単純な特徴を利用

原文特徴

文章中の語数
文章中のカンマ数
文のパープレキシティ
文中の未知語数

単語特徴

Google n-gramコーパスにおける、1-gram確率の負の対数尤度
現代アメリカ英語コーパスCOCAにおける、1-gram確率の負の対数尤度
Brownコーパスそれぞれにおける、1-gram確率の負の対数尤度
12段階の単語難易度指標（SVL 12000)

高品質の翻訳を得るまでの翻訳費用で評価

- 提案法により「高品質の翻訳を返す確率が高い」順に翻訳者を並べることができる
- 出力された順に翻訳を依頼した場合の翻訳費用で評価
  - 翻訳費用：
    「高品質」ラベル付きの翻訳に到達するまでに必要な翻訳者数

翻訳費用の例

翻訳タスク ⇒  ⇒  ⇒  この場合、翻訳費用＝３

低品質☹　低品質☹　高品質☺

ロジスティック回帰とSVMをベースラインに使用

ベースラインと提案法のバリエーション

	各手法で用いる情報		
	原文特徴	単語特徴	語彙テスト結果
ランダム			
SVM-S	✓		
SVM-SW	✓	✓	✓
LR-S	✓		
LR-SW	✓	✓	✓
OURS-in	✓	✓	✓
OURS	✓	✓	✓

OURS-in:
提案法のバリエーション。翻訳能力と語彙能力を関連づけない

結果 (1)

提案法は、翻訳テストの採点率が小さい場合（採点コストを掛けない場合）に翻訳費用の削減効果大

翻訳費用の対ランダム比

	翻訳テスト結果の採点率			
	0%	1%	6%	12%
ランダム	1.000	1.000	1.000	1.000
SVM-S	-	0.825	0.703	0.672
SVM-SW	-	0.756	0.679	**0.659**
LR-S	-	0.820	0.706	0.669
LR-SW	-	0.731	0.678	0.660
OURS-in	0.908	0.736	0.699	0.661
OURS	**0.673**	**0.660**	**0.675**	0.664

翻訳能力と語彙能力の関連付けは予測精度向上に効果アリ

翻訳費用の対ランダム比

	翻訳テスト結果の採点率			
	0%	1%	6%	12%
ランダム	1.000	1.000	1.000	1.000
SVM-S	-	0.825	0.703	0.672
SVM-SW	-	0.756	0.679	**0.659**
LR-S	-	0.820	0.706	0.669
LR-SW	-	0.731	0.678	0.660
OURS-in	0.908	0.736	0.699	0.661
OURS	**0.673**	**0.660**	**0.675**	0.664

語彙テスト結果は他手法の予測精度向上にも有効

翻訳費用の対ランダム比

	翻訳テスト結果の採点率			
	0%	**1%**	**6%**	**12%**
ランダム	1.000	1.000	1.000	1.000
SVM-S	-	0.825	0.703	0.672
SVM-SW	-	0.756	0.679	**0.659**
LR-S	-	0.820	0.706	0.669
LR-SW	-	0.731	0.678	0.660
OURS-in	0.908	0.736	0.699	0.661
OURS	**0.673**	**0.660**	**0.675**	0.664

まとめ

語彙テストを用いた、簡便な翻訳者評価法を提案

- 自己申告語彙テストを用い、
  採点コストを掛けずに翻訳者能力ひいては翻訳品質を予測
- 翻訳・語彙テストの結果を関連づける確率モデルを提案
- 結果
  - 提案法はクラウドソーシング翻訳費用の削減に貢献
  - 翻訳能力と語彙能力の関連付けが有効