Performance Prediction

Anonymous ACL submission

Abstract

TODO

1 Introduction

Motivation: Training and testing translaton model can be expensive- LRLs don't have the privilege- nice if we could tell directly if a model works for a language directly without actually running it.

State of the art: TOADD literature that mentioned factors that we also consider- explain why they're important- currently best/ most popular model is XXX

Flaws in state of the art: Point out factors that they are missing, not much on LRLs

Our proposition: can't immediately come out with a perfect model? Let's try a bunch: first consider single factor then multifactor- rigorous model selection that considered over/under fit, trade off between factors, and worst case

Evaluation: Compare with baselines from State of the art- even if worse, at least we tried on LRLs

2 Problem Formulation

- Want a good math model that can predict the performance of a Neural Machine Translation (NMT) model with respect to quantifiable factors.
- Performance of NMT in this paper: sp-BLEU scores.

Definition 1. The *Pareto-Efficiency Principle* suggests that a model A is better than a model B if A *Pareto-dominates* B, i.e., A performs better than B in at least one situation without being worse in any other situation.

Definition 2. A model A is said to be *Pareto-optimal*, if there is no other model that performs better than A in **every** situation, i.e., no other model *Pareto-dominates* it.

- Choosing a Pareto-optimal model ensures that the model represent the best trade-off in terms of minimizing cost across all situation simultaneously, without any other trial function outperforming them in every situation.

Definition 3. A model A is said to be a *Rawlsian choice* among all possible models if, under the principles of Rawlsian Fairness, it ensures that even its worst performance is better than the worst performance of any other model.

- Rawlsian Fairness is taken into consideration to prioritize the performance of the model in worst-case scenarios. - We seek for a Rawlsian choice math model that is also Pareto-optimal.

3 Methodology

3.1 Data

We collected experimental records from prior training and testing of the mBart model across different datasets and target languages. Each experimental record consists of a sp-BLEU score along with a ID of corresponding descriptive features, including training dataset in stage 1, ϕ_{t_1} and its size, ϕ_{s_1}, training dataset in stage 2, ϕ_{t_2} an its size, ϕ_{s_2}, testing dataset ϕ_τ, source language (always English (en)), and target language, ϕ_l. In the process of modeling, we often slice the experimental records by grouping records that share similar feature(s), as described by slice ID. These groups of records with similar features are referred as slices of experimental records.

3.2 Parameters

We considered three potential factors that affect the performance of the translation model listed below. Each factor is parametrized by different sets of variables.

Size of training datasets: We took ϕ_{s_1} and ϕ_{s_2} of the experimental records directly as the variables under this factor, denoted by s_1 and s_2 respectively.

Domain relatedness: We calculated Jensen-Shannon divergence (JSD) between each training

dataset and testing dataset, denoted by j_1 and j_2 respectively. JSD between a training dataset t and a testing dataset τ is calculated as follows: Suppose the unigram distribution of a word i, w_i is $\mathbb{P}_t(w_i)$ in the source language text of t and is $\mathbb{P}_\tau(w_i)$ in the unigram distribution τ, then the JSD between the training and testing datasets is given by

$$j_t = \frac{1}{2}KL(t, M) + \frac{1}{2}KL(\tau, M)$$

where M is the merged distribution of t and τ such that

$$\mathbb{P}_M(w_i) = \begin{cases} \frac{1}{2}\mathbb{P}_t(w_i) + \frac{1}{2}\mathbb{P}_\tau(w_i) & \text{if } w_i \in t \cap \tau \\ \mathbb{P}_t(w_i) & \text{if } w_i \in t \setminus \tau \\ \mathbb{P}_\tau(w_i) & \text{if } w_i \in \tau \setminus t \end{cases}$$

and $KL(\mathcal{D}, M)$ is the Kullback–Leibler (KL) divergence between the original unigram distribution $\mathcal{D}$ and the merged distribution M such that

$$KL(\mathcal{D}, M) = \sum_{\forall w_i \in M} (\mathbb{P}_M(w_i) - \mathbb{P}_\mathcal{D}(w_i)) \log\left(\frac{\mathbb{P}_M(w_i)}{\mathbb{P}_\mathcal{D}(w_i)}\right)$$

Language relatedness: We considered following two categories of variables under this factor:

(a) **Dataset independent variables:** We utilized six distance features from URIEL Typological Database to measure the level of relatedness between the source and target language, namely, syntatic distance, d_{syc}, phonological distance, d_{pho}, inventory distance, d_{inv}, featural distance, d_{fea}, geographical distance, d_{geo}, and genetic distance, d_{gen}.

(b) **Dataset dependent variables:** We considered the following variables to measure the language similarities between the source language text $D_t^{(S)}$ and target language text $D_t^{(T)}$ in each training dataset t:

(i) Ratio of dataset size

$$\rho_t = \frac{|D_t^{(S)}|}{|D_t^{(T)}|}$$

where $|D_t^{(\cdot)}|$ is the number of tokens in dataset $D_t^{(\cdot)}$.

(ii) Distance of type-token ratio

$$d_{ttr,t} = \left(1 - \frac{TTR_{D_t^{(S)}}}{TTR_{D_t^{(T)}}}\right)^2$$

where $TTR_{D_t^{(\cdot)}}$ = type-token ratio of dataset $D_t^{(\cdot)}$.

(iii) Word overlap

$$o_{w,t} = \frac{|W_{D_t^{(S)}} \cap W_{D_t^{(T)}}|}{|W_{D_t^{(S)}}| + |W_{D_t^{(T)}}|}$$

where $W_{D_t^{(\cdot)}}$ = set of types in dataset $D_t^{(\cdot)}$.

(iv) Subword overlap

$$o_{sw,t} = \frac{|S_{D_t^{(S)}} \cap S_{D_t^{(T)}}|}{|S_{D_t^{(S)}}| + |S_{D_t^{(T)}}|}$$

where $s_{D_t^{(\cdot)}}$ = set of subwords, Obtained by unsupervised word segmentation, in dataset $D_t^{(\cdot)}$.

(v) Total distance of word alignment done when running AWESOME (Aligning Word Embedding Spaces of Multilingual Encoders) on the training dataset, denoted by $d_{a,t}$.

3.3 Modeling

Each model is defined by a trial function that is used to model the performance score with respect to some selected variables. In the first phase of modeling, trial functions focused on variables within a single factor, while in the second phase, multifactor trial functions were developed using insights from the single-factor models. The trial functions were used to plot the line of best fit for each slice, and the resulting fit coefficients were tabulated. The root mean square error (RMSE) was then calculated to evaluate the fit of each trial function to its corresponding slice. Relevant overfitting and underfitting were also measured. Through these two analysis, the trial functions were refined by adjusting the range of the coefficients or by constructing an improved trial function.

3.4 Evaluation

To ensure the performance of each trial function is well-measured, we conducted the following procedure:

1. Partition the set of all slices into k partitions such that slices within a partition share common features as described by the partition ID.
2. A partition is chosen at random to be used for evaluation, denoted by π_k.
3. Determine the *most representative fits* (MRF) within each $k-1$ partition.

4. Using the MRFs from $k-1$ partition, determine the fits estimator as values of coefficients in the trial function to curve fit on all slices in π_k.

5. Calculate the RMSE for each curve fitting. Record the average RMSE in the cost vectors of the trial function.

Most Representative Fits (MRF): Consider a slice partition π_i with n slices $\psi_1^{(i)}, ..., \psi_n^{(i)}$. Suppose the trial function f has m coefficients, slice j in the partition would have fits vector $\mathbf{b}^{(i,j)} = (\beta_1^{(i,j)}, ..., \beta_m^{(i,j)})$ that correspond to each coefficient value in the equation of best fit line. The *most representative fits* (MRF) for this slice is

$$\bar{\mathbf{b}}^{(i)} = (\bar{\beta}_1^{(i)}, ..., \bar{\beta}_m^{(i)})$$

where $\bar{\mathbf{b}}^{(i)}$ is determined using one of the following approaches:

I. Simple Average: The set of average fits across all slices in the fold, i.e.,

$$\bar{\mathbf{b}}^{(i)} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{b}^{(i,j)}$$

II. Best set of fits: The set of fits from a slice in the partition that yields the lowest average RMSE when used to fit other slices in the fold, i.e.,

$$\bar{\mathbf{b}}^{(i)} = \arg\min_{\mathbf{b}^{(i,b)}} \frac{1}{n-1}\sum_{j\neq b} R(f, B^{(i,b)}, \psi_j^{(i)})$$

where R calculates the RMSE when fitting the trial function f onto $\psi_j^{(i)}$ using $\mathbf{b}^{(i,b)}$ as coefficient values.

III. Cross average fitting: The set of average fits from $n-1$ slices in the partition that yields the lowest RMSE when used to fit the remaining slice, i.e.,

$$\bar{\mathbf{b}}^{(i)} = \arg\min_{\tilde{\mathbf{b}}^{(i,l)}} R(f, \tilde{\mathbf{b}}^{(i,l)}, \psi_l^{(i)})$$

where

$$\tilde{\mathbf{b}}^{(i,l)} = \frac{1}{n-1}\sum_{j\neq l} \mathbf{b}^{(i,j)}$$

Fits Estimator: The average of most representative fits from $k-1$ partitions,

$$\hat{\mathbf{b}} = \frac{1}{k-1}\sum_{i=1}^{k-1} \bar{\mathbf{b}}^{(i)}$$

is used to estimate the fits for the remaining partition. The following average RMSE is calculated as the cost of this evaluation:

$$C_{\mathcal{M}}(f, \Phi_{com}) = \frac{1}{m-1}\sum_{j=1}^{n} R(f, \hat{\mathbf{b}}, \psi_j^{(k)})$$

where $\mathcal{M}$ is the chosen MRF approach while Φ_{com} is the set of common features within a partition.

Cost vectors: Suppose there are p ways of partitioning, a trial function has three cost vectors, correspond to $\mathcal{M} = \{\text{I, II, III}\}$, that records the cost of each evaluation for each way of partitioning.

$$\mathbf{c}_{\mathcal{M}}(f) = \big(C_{\mathcal{M}}(f, \Phi_{com_1}), ..., C_{\mathcal{M}}(f, \Phi_{com_p})\big)$$

Note: The left out fold, π_k should be kept constant for all cost vectors.

Choosing the best trial function: The best trial function was determined based on its cost vectors. For each MRF approach, we first identified the set of *Pareto-optimal* trial functions, ensuring that no other function had a lower cost than the Pareto-optimal functions for every entry in the cost vector.

$$\mathcal{P} = \{f \in F : \nexists g \in F \text{ s.t. } \mathbf{c}_{\mathcal{M}}(g) < \mathbf{c}_{\mathcal{M}}(f)\}$$

where $\mathbf{c}_{\mathcal{M}}(g) < \mathbf{c}_{\mathcal{M}}(f)$ means all entries in $\mathbf{c}_{\mathcal{M}}(g)$ are strictly less than its corresponding entry in $\mathbf{c}_{\mathcal{M}}(f)$ for each coordinate.

Rawlsian Fairness analysis was then conducted among all Pareto-optimal trial functions by selecting the trial function with the lowest maximum RMSE among all entries in its corresponding cost vector.

$$f^* = \arg\min_{f\in\mathcal{P}}\{\max C_{\mathcal{M}}(f, \Phi_{com_\iota})\} \text{ for } \iota = \{1, ..., p\}$$

The selected trial function, f^* is the overall best trial function.

3.5 Baselines

- Simple linear regression over factors Neubig considered

- Simple linear regression over ALL variables WE considered

- Stepwise regression over ALL variables WE considered

- Stepwise regression over lang-idp variables we considered, then combine with other variables for linear regression

- What about using MRF-Simple Average as baselines for Best set of fits and Cross average fitting?

4 Results

4.1 Phase 1 Modeling

Something like this? Should have one for each MRF approach- see if the best trial functions agree.

Factor	f^*	$\mathbf{c}_{\mathcal{M}}(f^*)$	$\bar{\mathbf{c}}(F)$
Size			
Domain			
Lang-idp			
Lang-dp			

where f^* is the best trial function, $\mathbf{c}_{\mathcal{M}}(f^*)$ is its performance matrix, $\bar{\mathbf{c}}(F)$ is the entry-wise average of all cost vectors.

Put some graphs of the best trial functions here.

Put cost vectors of all trial functions in appendix?

4.2 Phase 2 Modeling

Should have only one best trial function. Compare with baselines here.

Put graph of best trial function here.

5 Discussion

- Emphasis significance of our factors and why we did 2 phases

- LRLs specific arguments?

- Instead of doing complicated MRF stuff- could have use cost matrix e.g. lang vs test set

- Not comprehensive because of outliers

- Also too many dimensions- hard to see when considering >3 factors/ variables

Last updated: 6/26 12:00am