Siamese LSTMs for Translation Post-Edit Ranking

Manasi Patwardhan, Kanika Kalra, Shirish Karande
TCS Research, Pune, India
manasi.patwardhan@tcs.com, kalra.kanika@tcs.com

Abstract

We consider the scenario of partially ranked crowd translations, where workers collaboratively post and peer edit prior translations. The iterative nature of the contributions, leads to interdependencies in the quality. We pose the problem as pair-wise comparisons of translations considering their history and model it using Siamese LSTM architecture. The LSTMs model translation dependencies and Siamese network model the preference function. We consider a supervised setting and predict the pair-wise comparisons for non-ranked translations. A sorting algorithm is used to get the complete set of rankings. The sequential nature of the problem is modeled well by LSTMs yielding 88.83% accuracy and 0.95 rank correlation, against a non-sequential Siamese DNN network providing an accuracy of 74.35% and 0.85 rank correlation, thus establishing the efficacy of the proposed approach.

Introduction

Post editing is an attractive alternative to completely manual translation offering time and cost efficiency, provided the quality of the translations is ensured (Zbib et al. 2013; Läubli et al. 2013; Callison-Burch 2009; Goto, Lin, and Ishida 2014; Potepa et al. 2011). Crowd ranking can serve the purpose (Callison-Burch 2009; Goto, Lin, and Ishida 2014; Bentivogli et al. 2011; Paul et al. 2012); but the quality of ranks still remain questionable. (Kilian et al. 2012; Zaidan and Callison-Burch 2011) use expert's rank as one of the features to train classifiers for quality prediction, where all contributions are independent. Such independence cannot be guaranteed when the workers post-edit machine or peer-edit prior human contributions, as the quality of a translation is not only a function of its quality parameters; but also depends on the quality of prior refereed translations.

We consider a workflow (Figure 1), where a crowd worker inspects all the machine and prior crowd translations for a sentence; rates the prior contributions on a scale of 1 (worst) to 5 (best); selects one and submits a post-edited version, forming a trail of translation history. We enable an expert to rank all the available translations for a subset of source sentences. Lower rank denotes higher quality. All ranked contributions of a source sentence are compared with each other to generate pairs. If the rank of the first translation in a pair is less than the second; the pair is labeled as '1', otherwise as '0'. We compare the paired translations along with their edit-history and train a Siamese recurrent neural architecture to identify better quality translation in a pair. Such networks are earlier used by (Mueller and Thyagarajan 2016) for identifying semantic similarity between paired sentences. They are designed to implement the symmetries naturally present in a preference function. We use this trained network to compare the remaining non-ranked translation sequences of a sentence to get complete set of paired comparisons and further solve the rank completion problem.

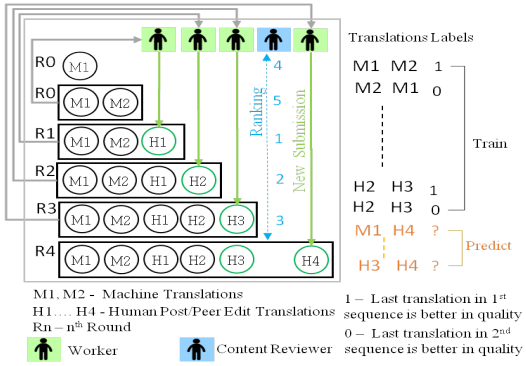

Figure 1: Workflow for Post and Peer-Editing Translations

Prior approaches (Rigutini et al. 2011; Fürnkranz and Hüllermeier 2010) learn preference function using shared weight approach to compare an paired objects, by providing their feature representations. However, to the best of our knowledge, ours is the first attempt to apply the shared weight architectures for translation ranking problem. Moreover, considering the sequential nature of the post-edit problem, we use Siamese adaption of LSTMs, for pair-wise translation comparisons along with their post-edit history. We benchmark against Siamese DNN which compares translations without consideration of the history.

Methodology

Let t_{ij} be a post-edit translation submitted for the i^{th} sentence in j^{th} post-edit round. The index of round, j, is utilized as the time index for the LSTMs. x_{ij} be the input vector for the j^{th} post-edit translation of the i^{th} sentence. y_{ijk}

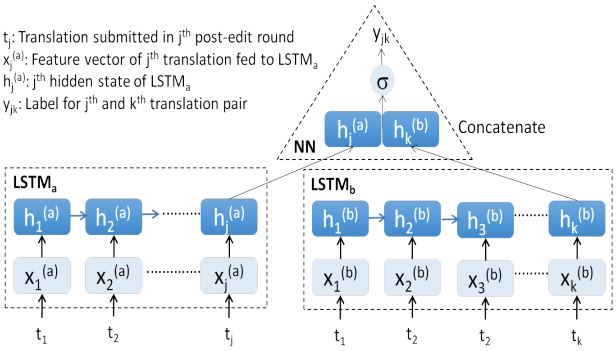

Figure 2: Siamese LSTM Network

is the label for ordered translation pair t_{ij} and t_{ik}, where $j \neq k$. $y_{ijk} \in \{0, 1\}$, 1 indicating the 1^{st} translation t_{ij} is better than the 2^{nd} translation t_{ik} and 0 otherwise. The problem we are trying to solve is: Given an ordered pair of translation along with their post-edit sequences for a source sentence, determine the better quality translation.

There are two networks $LSTM_a$ and $LSTM_b$ (Figure 2). Each process one of the translation post-edit sequence in a given pair. We model siamese architectures with tied weights such that $LSTM_a$ = $LSTM_b$. A LSTM learns the mapping from the space of variable length post-edit sequences of n-dimensional translation feature vectors $x_{i1}, x_{i2}, \ldots, x_{ij-1}, x_{ij}$ into a an embedding h_{ij}, provided by the last hidden state. Following equations demonstrate the mapping defined by the LSTM model. The equations are valid for all the sentences i and both $LSTM_a$ and $LSTM_b$.

$$I_j = \tanh(W_{xI}x_j + W_{hI}h_{j-1} + b_I) \quad (1)$$

$$\tilde{c}_j = \sigma(W_{x\tilde{c}}x_j + W_{h\tilde{c}}h_{j-1} + b_{\tilde{c}}) \quad (2)$$

$$F_j = \sigma(W_{xF}x_j + W_{hF}h_{j-1} + b_F) \quad (3)$$

$$O_j = \tanh(W_{xO}x_j + W_{hO}h_{j-1} + b_O) \quad (4)$$

$$c_j = c_{j-1} \odot F_j + I_j\dot{\tilde{c}}_j \quad (5)$$

$$h_j = O_j \odot \tanh c_j \quad (6)$$

The last hidden states h_{ij} and h_{ik} of the $LSTM_a$ and $LSTM_b$ respectively are concatenated and provided as an input to one dense sigmoid layer to map it to the output y_{ijk}.

$$y_{jk} = \sigma(W_{yh}[h_j, h_k])) \quad (7)$$

The siamese LSTM network solves the problem by computing the conditional probability $p(y_{ijk}|[x_{i1}, x_{i2}, \ldots, x_{ij-1}, x_{ij}], [x_{i1}, x_{i2}, \ldots, x_{ik-1}, x_{ik}]) = p(y_{ijk}|h_{ij}, h_{ik})p(h_{ij}|x_{i1}, x_{i2}, \ldots, x_{ij-1}, x_{ij})p(h_{ik}|x_{i1}, x_{i2}, \ldots, x_{ik-1}, x_{ik})$, where [.] denotes a post-edit sequence. Translation input feature vector x_{ij} consists of: i) Edit distance of the translation with the parent translation selected for post-edit. (ii) Word length of the source sentence (iii) Time taken in seconds by a worker to perform the task (iv) Average peer rating of a translation (v) Worker one-hot vector encoding.

For validating the results we split our ranked translation data (i) Sentence-wise (ii) Translation-wise and (iii) Pair-wise. For (ii) and (iii) the predicted labels along with the training labels provide us with a complete set of comparisons for a sentence. We rank all the translations of a sentence by sorting them based on the number of wins. In case of a tie, a translation submitted in the later round is considered to be better quality. Spearman's coefficient is used to determine rank correlation between actual and predicted rankings.

% Split Siamese	Validation Accuracy DNN	Validation Accuracy LSTM	Rank Correlation DNN	Rank Correlation LSTM
Sentence-Wise Random Split				
90-10	73.67	74.32	0.5685	0.5866
Translation-Wise Split				
(n-1)-1*	68.88	71.05	0.8005	0.8168
Pair-Wise Random Splits				
90-10	74.05	92.67	0.9775	0.9947
80-20	74.04	90.84	0.9416	0.9824
70-30	74.09	89.94	0.9005	0.9656
60-40	74.35	88.83	0.8458	0.9543
50-50	73.39	85.64	0.7854	0.9062
40-60	73.00	82.95	0.7250	0.8615
30-70	72.79	70.75	0.6759	0.8086

*1^{st} n-1 sentence translations for training, last for testing

Table 1: Results

Data and Experimental Results

60 non-professional crowd workers provided us 6016 Hindi post-edit translations for 1758 English sentences from a book about an autobiography. The maximum rounds of post-edits for any sentence were limited to 4. Thus, along with 3338 machine translations, we have total of 9354 translations. Translations of 474 sentences (2693 translations) were ranked between 1 to 6 by an expert worker (in-house linguist). We formed total of 12,824 ordered pairs of post-edit translation sequences and the corresponding labels by following the process discussed in the introduction section.

We modeled siamese LSTM, where each LSTMs have 10 dimensional hidden layer, time step of 6 and with 67 dimensional feature vector. Table 1 represents the validation accuracies and rank correlations for distinct splits. Siamese LSTMs results are better as compared to the results of non-recurrent Siamese DNNs and approximately 8-10% better than purely deterministic rank completion algorithms.

Conclusion

Siamese LSTMs are able to effectively model the pairwise comparisons for translations along with capturing post-editing interdependencies. In future, we plan to apply the technique for other crowd annotation NLP tasks with free-flowing labels, such as text simplification (Coster and Kauchak 2011), content moderation (Bernstein et al. 2015), image or speech transcription (Callison-Burch and Dredze 2010), question-answering (Bian et al. 2008), etc.

References

Bentivogli, L.; Federico, M.; Moretti, G.; and Paul, M. 2011. Getting expert quality from the crowd for machine translation evaluation. *Proc. MT Summmit* 13:521–528.

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2015. Soylent: a word processor with a crowd inside. *Communications of the ACM* 58(8):85–94.

Bian, J.; Liu, Y.; Agichtein, E.; and Zha, H. 2008. Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of the 17th international conference on World Wide Web*, 467–476. ACM.

Callison-Burch, C., and Dredze, M. 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 1–12. Association for Computational Linguistics.

Callison-Burch, C. 2009. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 286–295. Association for Computational Linguistics.

Coster, W., and Kauchak, D. 2011. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 665–669. Association for Computational Linguistics.

Fürnkranz, J., and Hüllermeier, E. 2010. Preference learning and ranking by pairwise comparison. In *Preference learning*. Springer. 65–82.

Goto, S.; Lin, D.; and Ishida, T. 2014. Crowdsourcing for evaluating machine translation quality. In *LREC*, 3456–3463.

Kilian, N.; Krause, M.; Runge, N.; and Smeddinck, J. 2012. Predicting crowd-based translation quality with language-independent feature vectors. In *HComp12 Proceedings of the AAAI Workshop on Human Computation. AAAI Press, Toronto, ON, Canada*, 114–115.

Läubli, S.; Fishel, M.; Massey, G.; Ehrensberger-Dow, M.; and Volk, M. 2013. Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, 83–91.

Mueller, J., and Thyagarajan, A. 2016. Siamese recurrent architectures for learning sentence similarity. In *AAAI*, 2786–2792.

Paul, M.; Sumita, E.; Bentivogli, L.; and Federico, M. 2012. Crowd-based mt evaluation for non-english target languages. *Proceedings of the European Association for Machine Translation (EAMT12)* 229–236.

Potepa, A.; Plonka, P.; Pytel, M.; and Radziszowski, D. 2011. Iterative translation by monolinguists implementation and tests of the new approach. In *ACIIDS (1)*, 445–454.

Rigutini, L.; Papini, T.; Maggini, M.; and Scarselli, F. 2011. Sortnet: Learning to rank by a neural preference function. *IEEE transactions on neural networks* 22(9):1368–1380.

Zaidan, O. F., and Callison-Burch, C. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 1220–1229. Association for Computational Linguistics.

Zbib, R.; Markiewicz, G.; Matsoukas, S.; Schwartz, R. M.; and Makhoul, J. 2013. Systematic comparison of professional and crowdsourced reference translations for machine translation. In *HLT-NAACL*, 612–616.