Referential Translation Machines for Predicting Translation Performance

Ergun Biçici
ergunbicici@yahoo.com
bicici.github.com

Abstract

Referential translation machines (RTMs) pioneer a language independent approach for predicting translation performance and to all similarity tasks with top performance in both bilingual and monolingual settings and remove the need to access any task or domain specific information or resource. RTMs achieve to become 1st in document-level, 4th system at sentence-level according to mean absolute error, and 4th in phrase-level prediction of translation quality in quality estimation task.

1 Referential Translation Machines

Prediction of translation performance can help in estimating the effort required for correcting the translations during post-editing by human translators if needed. Referential translation machines achieve top performance in automatic and accurate prediction of machine translation performance independent of the language or domain of the prediction task. Each referential translation machine (RTM) model is a data translation prediction model between the instances in the training set and the test set and translation acts are indicators of the data transformation and translation. RTMs are powerful enough to be applicable in different domains and tasks while achieving top performance in both monolingual (Biçici and Way, 2015) and bilingual settings (Biçici et al., 2015b).

Figure 1 depicts RTMs and explains the model building process (Biçici, 2016). RTMs use ParFDA (Biçici et al., 2015a) for selecting instances and interpretants, data close to the task instances for building prediction models and machine translation performance prediction system (MTPPS) (Biçici and Way, 2015) for generating features. We improve our RTM models (Biçici et al., 2015b) with numeric expression identification using regular expressions and replace them with a label (Biçici, 2016).

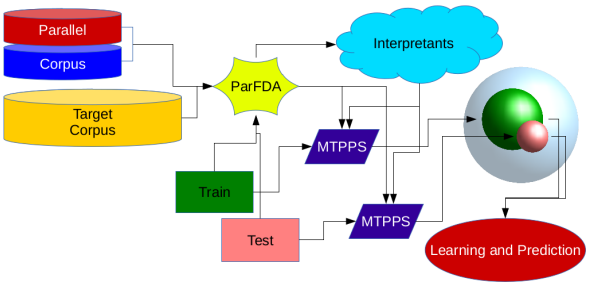

Figure 1: RTM depiction: ParFDA selects interpretants close to the training and test data using parallel corpus in bilingual settings and monolingual corpus in the target language or just the monolingual target corpus in monolingual settings; an MTPPS uses interpretants and training data to generate training features and another uses interpretants and test data to generate test features in the same feature space; learning and prediction takes place taking these features as input.

2 RTM in the Quality Estimation Task

We develop RTM models for all of the four subtasks of the quality estimation task (QET) in WMT16 (Bojar et al., 2016) (QET16), which include English to Spanish (en-es), English to German (en-de), and German to English (de-en) translation directions. The subtasks are: sentence-level prediction (Task 1), word-level prediction (Task 2), phrase-level prediction (Task 2p), and document-level prediction (Task 3). Task 1 is about predicting HTER (human-targeted translation edit rate) (Snover et al., 2006) scores of sentence translations, Task 2 is about binary classification of word-level quality, Task 2p is about binary classification of phrase-level quality, and

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 777–781,
Berlin, Germany, August 11-12, 2016.

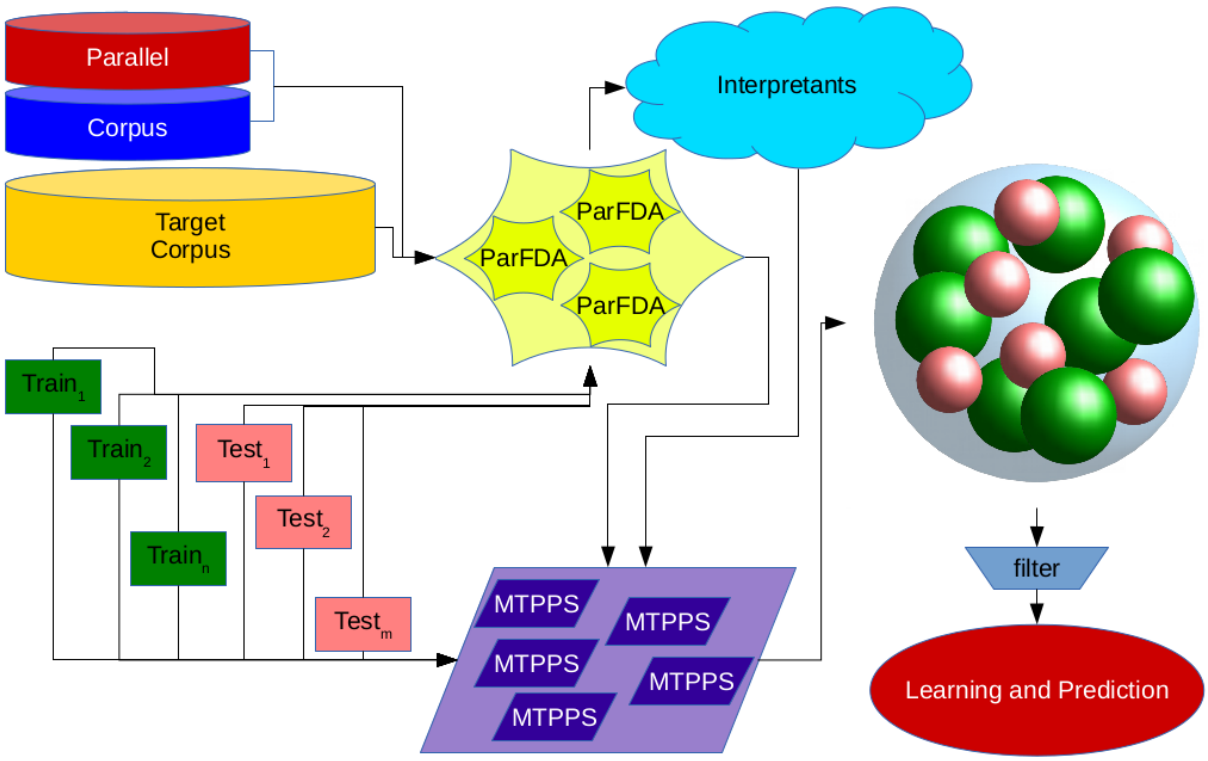

Figure 2: RTM depiction for Task 3 where document-level translation performance is predicted. Separate MTPPS instances are run for each train and test document to obtain corresponding feature representations, which are filtered and processed before learning and prediction.

			RTM Interpretants	
Task	Train	Test	Training	LM
Task 1 (en-de)	13000	3000	400K	10M
Task 2 (en-de)	13000	2000	500K	10M
Task 3 (en-es)	146	62	1M	10M

Table 1: Number of instances in different tasks and the number of sentences used as interpretants by the RTM models.

Task 3 is about predicting weighted HTER scores of document translations.

Language model (LM) are built using KENLM (Heafield et al., 2013). We tokenize and truecase all of the corpora using code released with Moses (Koehn et al., 2007) [1]. Table 1 lists the number of sentences in the training and test sets for each task. We also list the size of the interpretants used by the corresponding RTM models (K for thousand, M for million). We use the same number of interpretants for training as last year in Task 1. We increase the number of instances used for the LM to 10M. This year, we did not include features from backward LM in MTPPS and we used numeric expression identification in Task 1 and Task 3.

[1] https://github.com/moses-smt/mosesdecoder/tree/master/scripts

2.1 RTM Prediction Models

We present results using support vector regression (SVR) with RBF (radial basis functions) kernel (Smola and Schölkopf, 2004) and extremely randomized trees (TREE) (Geurts et al., 2006) for sentence and document translation prediction tasks. We also use them after a feature subset selection (FS) with recursive feature elimination (RFE) (Guyon et al., 2002) or a dimensionality reduction and mapping step using partial least squares (PLS) (Specia et al., 2009), or PLS after FS (FS+PLS). We use Global Linear Models (GLM) (Collins, 2002) with dynamic learning (GLMd) (Biçici et al., 2015b) for word-level translation performance prediction. GLMd uses weights in a range $[a, b]$ to update the learning rate dynamically according to the error rate as shown in Figure 3.

Figure 2 depicts how RTMs are used to build predictors for Task 3, where we run a separate MTPPS instance for each train or test document

Task	Translation	Model	r	MAE	RAE	MAER	MRAER
Task 1	en-de	SVR	0.39	0.1449	0.874	0.7653	0.824
	en-de	FS SVR	0.4	0.1453	0.877	0.7704	0.826
Task 3	en-es	FS+PLS TREE	0.55	0.3058	0.823	0.4394	0.815
	en-es	FS SVR	0.33	0.3383	0.91	0.4308	0.8

Table 2: Training performance of the top 2 individual RTM models prepared for different tasks.

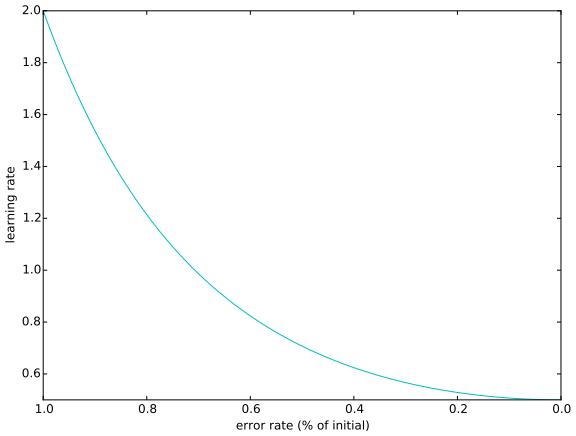

Figure 3: Learning rate curve.

Model	# splits	% error	weight range
GLMd	4	0.0688	$[0.5, 2]$
GLMd	5	0.0757	$[0.5, 2]$

Table 3: RTM Task 2 training results where GLMd parallelized over 4 splits is referred as GLMd s4 and GLMd with 5 splits as GLMd s5.

and obtain corresponding features (depicted with a green or salmon colored sphere). We obtain an RTM representation vector instance from each of these by using only the document-level features from MTPPS and the min, max, and average of the sentence-level features.

2.2 Training Results

We use mean absolute error (MAE), relative absolute error (RAE), root mean squared error (RMSE), Pearson's correlation (r_P), and Spearman's correlation (r_S) as well as relative MAE (MAER) and relative RAE (MRAER) to evaluate (Biçici and Way, 2015). MAER and MRAER consider both the predictor's error and the fluctuations of the target scores at the instance level. RTM test performance on various tasks sorted according to MRAER can help identify which tasks and subtasks may require more work. DeltaAvg (Callison-Burch et al., 2012) calculates the average quality difference between the top $n-1$ quartiles and the overall quality for the test set. Table 2 presents the training results for Task 1 and Task 3. Table 3 presents Task 2 training results obtained after the challenge.

2.3 Test Results

The results on the test set are listed in Table 4 [2] and Table 5. Ranks are out of 9, 8, 6, and 5 system submissions in Task 1, Task 2, Task 2p, and Task 3 respectively. RTMs with FS SVR is able to achieve the 6th rank in Task 1 according to r_P and 4th according to MAE. The top MAE is 12.3 where RTM obtains 9% more MAE. RTMs with FS+PLS TREE is able to achieve the 1st rank in Task 3.

2.4 Target Optimized Results

Table 6 lists the RTM results optimizing the target evaluation metric, r, obtained after the challenge. The results show that numerical expression identification did not improve the test results for QET Task 1 but we have observed improvements in semantic textual similarity in English (Biçici, 2016).

2.5 Comparison with Previous Results

We compare the difficulty of tasks according to MRAER levels achieved. In Table 7, we list the RTM test results for tasks and subtasks that predict HTER or METEOR from QET16, QET15 (Biçici et al., 2015b), QET14 (Biçici and Way, 2014), and QET13 (Biçici, 2013). Compared with QET15 Task 1 performance, MAER improved in QET16 and obtained the top MAER performance in sentence-level prediction. Compared with QET15 Task 2 performance, both F_1 OK and F_1 BAD improved even though the training error tripled. wF_1 calculation in QET16 is different than the calculation used in QET15.

[2] We calculate r_S using `scipy.stats`.

Task	Model	DeltaAvg	r_P	r_S	RMSE	MAE	RAE	MAER	MRAER	Rank
Task 1	en-de SVR	6.38	0.3581	0.3841	18.06	13.59	0.8992	0.7509	0.8567	7
	en-de FS SVR	6.66	0.3764	0.4003	17.81	13.46	0.8905	0.7537	0.8388	6
Task 3	en-es FS+PLS TREE	0.12	0.3562	0.46	0.3437	0.2533	0.8996	0.3285	0.8505	1
	en-es FS SVR	0.12	0.2929	0.3546	0.3529	0.2676	0.9505	0.333	0.9018	2

Table 4: Test performance of the top 2 individual RTM models prepared for different tasks.

	Model	wF_1	F_1 OK	F_1 BAD	Rank
Word	GLMd s4	0.2725	**0.8884**	0.3068	9
	GLMd s5	0.3081	0.8820	0.3494	$\sim$8
Phrase	GLMd s4	0.3070	**0.8145**	0.3770	5
	GLMd s5	0.3274	**0.8016**	0.4084	4

Table 5: RTM Task 2 results on the test set. wF_1 is the average weighted F_1 score. **bold** results obtain top performance.

3 Conclusion

Referential translation machines achieve top performance in automatic, accurate, and language independent prediction of translation performance. RTMs pioneer a language independent approach for predicting translation performance and to all similarity tasks and remove the need to access any task or domain specific information or resource.

Acknowledgments

We thank the reviewers for providing constructive comments.

References

Ergun Biçici and Andy Way. 2014. Referential translation machines for predicting translation quality. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, 6. Association for Computational Linguistics.

Ergun Biçici and Andy Way. 2015. Referential translation machines for predicting semantic similarity. *Language Resources and Evaluation*, pages 1–27.

Ergun Biçici, Qun Liu, and Andy Way. 2015a. ParFDA for fast deployment of accurate statistical machine translation systems, benchmarks, and statistics. In *Proceedings of the EMNLP 2015 Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, 9. Association for Computational Linguistics.

Ergun Biçici, Qun Liu, and Andy Way. 2015b. Referential translation machines for predicting translation quality and related statistics. In *Proceedings of the EMNLP 2015 Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, 9. Association for Computational Linguistics.

Ergun Biçici. 2013. Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, 8. Association for Computational Linguistics.

Ergun Biçici. 2016. RTM at SemEval-2016 task 1: Predicting semantic similarity with referential translation machines and related statistics. In *SemEval-2016: Semantic Evaluation Exercises - International Workshop on Semantic Evaluation*, San Diego, USA, 6.

Ondrej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Pavel Pacina, Martin Poppel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi. 2016. Proc. of the 2016 conference on statistical machine translation. In *Proc. of the First Conference on Statistical Machine Translation (WMT16)*, Berlin, Germany, August.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proc. of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June.

Michael Collins. 2002. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proc. of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Associa-*

Model	r	MAE	RAE	MAER	MRAER
FS-SVR	0.37	0.135	0.893	0.7471	0.846
+numeric PLS-SVR	0.37	0.1358	0.898	0.7572	0.865

Table 6: RTM top predictor testing results for Task 1 optimized for r.

Task	Translation	Model	r	MAE	RAE	MAER	MRAER
QET16 Task 1 HTER	en-de	FS SVR	0.3764	13.4589	0.8905	0.7537	0.8388
QET16 Task 3 HTER	en-es	FS+PLS TREE	0.3562	0.2533	0.8996	0.3285	0.8505
QET15 Task 1 HTER	en-es	FS+PLS SVR	0.349	0.1335	0.903	0.8284	0.8353
QET15 Task 3 METEOR	en-de	FS SVR	0.6668	0.0728	0.7279	0.3249	0.6467
	de-en	FS+PLS SVR	0.6373	0.0494	0.7482	0.2996	0.68
QET14 Task 1.2 HTER	en-es	SVR	0.5499	0.134	0.8532	0.7727	0.8758
QET13 Task 1.1 HTER	en-es	PLS-SVR	0.5596	0.1326	0.8849	2.3738	1.6428

Table 7: Test performance of the top RTM results when predicting HTER or METEOR.

tion for Computational Linguistics, pages 690–696, Sofia, Bulgaria, August.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of Association for Machine Translation in the Americas,*.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Proc. of the 13th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 28–35, Barcelona, Spain, May.