Referential Translation Machines for Predicting Translation Quality and Related Statistics

Ergun Biçici
ADAPT Research Center
School of Computing
Dublin City University, Ireland
ergun.bicici@computing.dcu.ie

Qun Liu
ADAPT Research Center
School of Computing
Dublin City University, Ireland
qliu@computing.dcu.ie

Andy Way
ADAPT Research Center
School of Computing
Dublin City University, Ireland
away@computing.dcu.ie

Abstract

We use referential translation machines (RTMs) for predicting translation performance. RTMs pioneer a language independent approach to all similarity tasks and remove the need to access any task or domain specific information or resource. We improve our RTM models with the ParFDA instance selection model (Biçici et al., 2015), with additional features for predicting the translation performance, and with improved learning models. We develop RTM models for each WMT15 QET (QET15) subtask and obtain improvements over QET14 results. RTMs achieve top performance in QET15 ranking 1st in document- and sentence-level prediction tasks and 2nd in word-level prediction task.

1 Referential Translation Machine (RTM)

Referential translation machines are a computational model effectively judging monolingual and bilingual similarity while identifying translation acts between any two data sets with respect to interpretants. RTMs achieve top performance in automatic, accurate, and language independent prediction of machine translation performance and reduce our dependence on any task dependent resource. Prediction of translation performance can help in estimating the effort required for correcting the translations during post-editing by human translators. We improve our RTM models (Biçici and Way, 2014):

- by using improved ParFDA instance selection model (Biçici et al., 2015) allowing better language models (LM) in which similarity judgments are made to be built with improved optimization and selection of the LM data,
- by selecting TreeF features over source and translation data jointly instead of taking their intersection,
- with extended learning models including bayesian ridge regression (Tan et al., 2015), which did not obtain better performance than support vector regression in training results (Section 2.2).

We present top results with Referential Translation Machines (Biçici, 2015; Biçici and Way, 2014) at quality estimation task (QET15) in WMT15 (Bojar et al., 2015). RTMs pioneer a computational model for quality and semantic similarity judgments in monolingual and bilingual settings using retrieval of relevant training data (Biçici and Yuret, 2015) as interpretants for reaching shared semantics. RTMs use Machine Translation Performance Prediction (MTPP) System (Biçici et al., 2013; Biçici, 2015), which is a state-of-the-art performance predictor of translation even without using the translation by using only the source. We use ParFDA for selecting the interpretants (Biçici et al., 2015; Biçici and Yuret, 2015) and build an MTPP model. MTPP derives indicators of the closeness of test sentences to the available training data, the difficulty of translating the sentence, and the presence of acts of translation for data transformation. We view that acts of translation are ubiquitously used during communication:

Every act of communication is an act of translation (Bliss, 2012).

Figure 1 depicts RTM. Our encouraging results in QET provides a greater understanding of the acts of translation we ubiquitously use and how they can be used to predict the performance of translation. RTMs are powerful enough to be applicable in different domains and tasks while achieving top performance.

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 304–308,
Lisboa, Portugal, 17-18 September 2015.

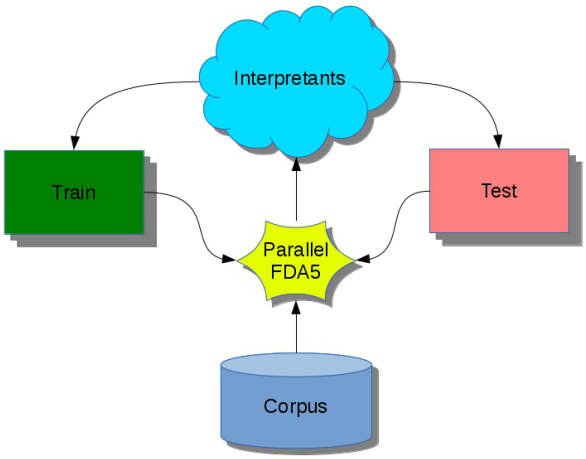

Figure 1: RTM depiction.

Task	Train	Test
Task 1 (en-es)	12271	1817
Task 2 (en-es)	12271	1817
Task 3 (en-de)	800	415
Task 3 (de-en)	800	415

Table 1: Number of sentences in different tasks.

2 RTM in the Quality Estimation Task

We participate in all of the three subtasks of the quality estimation task (QET) (Bojar et al., 2015), which include English to Spanish (en-es), English to German (en-de), and German to English (de-en) translation directions. There are three subtasks: sentence-level prediction (Task 1), word-level prediction (Task 2), 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, and Task 3 is about predicting METEOR (Lavie and Agarwal, 2007) scores of document translations.

Instance selection for the training set and the language model (LM) corpus is handled by ParFDA (Biçici et al., 2015), whose parameters are optimized for each translation task. LM are trained using SRILM (Stolcke, 2002). 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.

[1] mosesdecoder/scripts/

2.1 RTM Prediction Models and Optimization

We present results using support vector regression (SVR) with RBF (radial basis functions) kernel (Smola and Schölkopf, 2004) for sentence and document translation prediction tasks and Global Linear Models (GLM) (Collins, 2002) with dynamic learning (GLMd) (Biçici, 2013; Biçici and Way, 2014) for word-level translation performance prediction. We also use these learning models 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).

GLM relies on Viterbi decoding, perceptron learning, and flexible feature definitions. GLMd extends the GLM framework by parallel perceptron training (McDonald et al., 2010) and dynamic learning with adaptive weight updates in the perceptron learning algorithm:

$$\mathbf{w} = \mathbf{w} + \alpha \left(\Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \hat{\mathbf{y}}) \right), \quad (1)$$

where Φ returns a global representation for instance i and the weights are updated by α, which dynamically decays the amount of the change during weight updates at later stages and prevents large fluctuations with updates.

The learning rate updates the weight values with weights in the range $[a, b]$ using the following function taking error rate as the input:

$$f(x) = (\log_a b - 1)x^2 + 1 \quad (2)$$

Learning rate curve for $a = 0.5$ and $b = 1.0$ is provided in Figure 2:

2.2 Training Results

We use mean absolute error (MAE), relative absolute error (RAE), root mean squared error (RMSE), and correlation (r) as well as relative MAE (MAER) and relative RAE (MRAER) to evaluate (Biçici, 2015; Biçici, 2013). MAER is mean absolute error relative to the magnitude of the target and MRAER is mean absolute error relative to the absolute error of a predictor always predicting the target mean assuming that target mean is known (Biçici, 2015). 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.,

Task	Translation	Model	r	MAE	RAE	MAER	MRAER
Task1	en-es	FS SVR	0.355	0.1387	0.895	0.782	0.821
	en-es	FS+PLS SVR	0.362	0.1389	0.896	0.784	0.824
Task3	en-de	FS SVR	0.517	0.0737	0.734	0.289	0.678
	en-de	SVR	0.503	0.0765	0.761	0.307	0.737
	de-en	FS SVR	0.479	0.0473	0.738	0.267	0.665
	de-en	FS+PLS SVR	0.391	0.0515	0.804	0.288	0.81

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

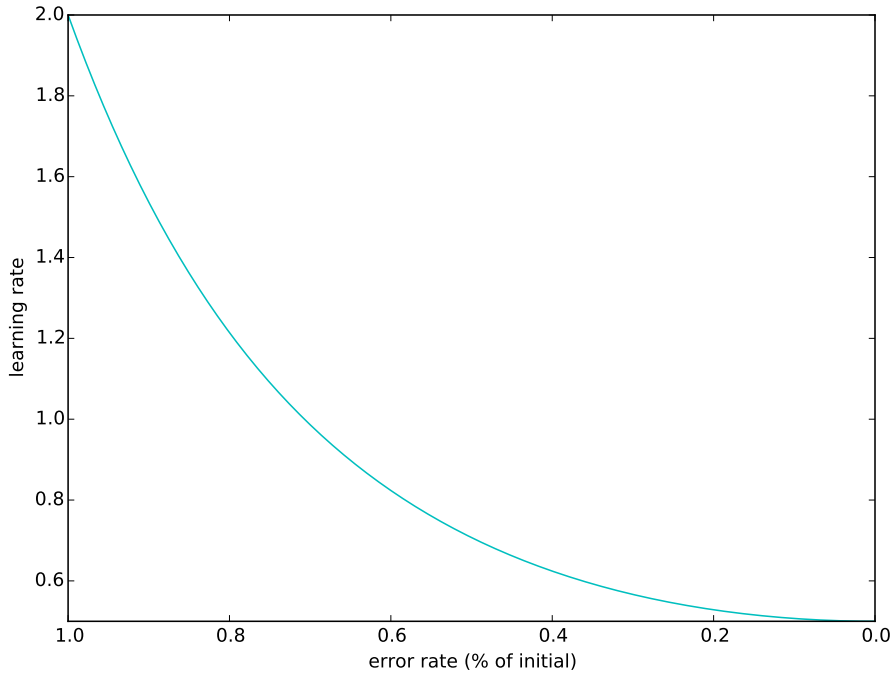

Figure 2: Learning rate curve.

Model	# splits	% error	weight range
GLMd	4	0.0227	$[0.5, 2]$
GLMd	5	0.0234	$[0.5, 2]$

Table 3: RTM-DCU Task 2 training results.

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. We refer to GLMd parallelized over 4 splits as GLMd s4 and GLMd with 5 splits as GLMd s5.

2.3 Test Results

Task 1: Predicting the HTER for Sentence Translations The results on the test set are given in Table 4. Rank lists the overall ranking in the task out of about 9 submissions. We obtain the rankings by sorting according to the predicted scores and randomly assigning ranks in case of ties. RTMs with FS followed by PLS and learning with SVR is able to achieve the top rank in this task.

Task 2: Prediction of Word-level Translation Quality Task 2 is about binary classification of word-level quality. We develop individual RTM models for each subtask and use GLMd model (Biçici, 2013; Biçici and Way, 2014), for predicting the quality at the word-level. The results on the test set are in Table 5 where the ranks are out of about 17 submissions. RTMs with GLMd becomes the second best system this task.

Task 3: Predicting METEOR of Document Translations Task 3 is about predicting METEOR (Lavie and Agarwal, 2007) and their ranking. The results on the test set are given in Table 4 where the ranks are out of about 6 submissions using wF_1. RTMs achieve top rankings in this task.

Task	Translation	Model	DeltaAvg	r	MAE	RAE	MAER	MRAER	Rank
Task1	en-es	FS SVR	0.61	0.3665	0.1325	0.8963	0.8344	0.8488	3
	en-es	FS+PLS SVR	0.63	0.349	0.1335	0.903	0.8284	0.8353	1
Task3	en-de	FS SVR	0.65	0.6668	0.0728	0.7279	0.3249	0.6467	2
	en-de	SVR	0.76	0.6247	0.075	0.7499	0.3623	0.7245	1
	de-en	FS SVR	0.49	0.5521	0.0578	0.8763	0.395	0.9159	1
	de-en	FS+PLS SVR	0.42	0.6373	0.0494	0.7482	0.2996	0.68	2

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

Model	wF_1	Rank	F_1 GOOD	F_1 BAD
GLMd s5	0.76	3	0.2391	0.8812
GLMd s4	0.7588	4	0.2269	0.8826

Table 5: RTM-DCU Task 2 results on the test set. wF_1 is the average weighted F_1 score.

2.4 RTMs Across Tasks and Years

We compare the difficulty of tasks according to MRAER levels achieved. In Table 6, we list the RTM test results for tasks and subtasks that predict HTER or METEOR from QET15, QET14 (Biçici and Way, 2014), and QET13 (Biçici, 2013). The best results when predicting HTER are obtained this year.

3 Conclusion

Referential translation machines achieve top performance in automatic, accurate, and language independent prediction of document-, sentence-, and word-level statistical machine translation (SMT) performance. RTMs remove the need to access any SMT system specific information or prior knowledge of the training data or models used when generating the translations. RTMs achieve top performance when predicting translation performance.

Acknowledgments

This work is supported in part by SFI as part of the ADAPT research center (www.adaptcentre.ie, 07/CE/I1142) at Dublin City University and in part by SFI for the project "Monolingual and Bilingual Text Quality Judgments with Translation Performance Prediction" (computing.dcu.ie/~ebicici/Projects/TIDA_RTM.html, 13/TIDA/I2740). We also thank the SFI/HEA Irish Centre for High-End Computing (ICHEC, www.ichec.ie) for the provision of computational facilities and support.

References

Ergun Biçici and Andy Way. 2014. Referential translation machines for predicting translation quality. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 313–321, Baltimore, Maryland, USA, June.

Ergun Biçici and Deniz Yuret. 2015. Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions On Audio, Speech, and Language Processing (TASLP)*, 23:339–350.

Ergun Biçici, Declan Groves, and Josef van Genabith. 2013. Predicting sentence translation quality using extrinsic and language independent features. *Machine Translation*, 27:171–192, December.

Ergun Biçici, Qun Liu, and Andy Way. 2015. Parallel FDA5 for fast deployment of accurate statistical machine translation systems, benchmarks, and statistics. In *Proc. of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September. Association for Computational Linguistics.

Ergun Biçici. 2013. Referential translation machines for quality estimation. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 343–351, Sofia, Bulgaria, August.

Ergun Biçici. 2015. RTM-DCU: Predicting semantic similarity with referential translation machines. In *SemEval-2015: Semantic Evaluation Exercises - International Workshop on Semantic Evaluation*, Denver, Colorado, USA, 4-5 June.

Chris Bliss. 2012. Comedy is translation, February. http://www.ted.com/talks/chris_bliss_comedy_is_translation.html.

Ondrej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proc. of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012.

Task	Translation	Model	r	MAE	RAE	MAER	MRAER
QET15 Task1 HTER	en-es	FS SVR	0.3665	0.1325	0.8963	0.8344	0.8488
	en-es	FS+PLS SVR	0.349	0.1335	0.903	0.8284	0.8353
QET15 Task3 METEOR	en-de	FS SVR	0.6668	0.0728	0.7279	0.3249	0.6467
	en-de	SVR	0.6247	0.075	0.7499	0.3623	0.7245
	de-en	FS SVR	0.5521	0.0578	0.8763	0.395	0.9159
	de-en	FS+PLS SVR	0.6373	0.0494	0.7482	0.2996	0.68
QET14 Task1.2 HTER	en-es	SVR	0.5499	0.134	0.8532	0.7727	0.8758
QET13 Task1.1 HTER	en-es	PLS-SVR	0.5596	0.1326	0.8849	2.3738	1.6428

Table 6: Test performance of the top individual RTM results when predicting HTER or METEOR also including results from QET14 (Biçici and Way, 2014) and QET13 (Biçici, 2013).

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.

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.

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 *ACL*, pages 177–180, Prague, Czech Republic, June.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June.

Ryan McDonald, Keith Hall, and Gideon Mann. 2010. Distributed training strategies for the structured perceptron. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 456–464, Los Angeles, California, June.

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.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, pages 901–904.

Liling Tan, Carolina Scarton, Lucia Specia, and Josef van Genabith. 2015. Usaar-sheffield: Semantic textual similarity with deep regression and machine translation evaluation metrics. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 85–89. Association for Computational Linguistics.