RTM at SemEval-2017 Task 1: Referential Translation Machines for Predicting Semantic Similarity

Ergun Biçici
orcid.org/0000-0002-2293-2031
bicici.github.com

Abstract

We use referential translation machines for predicting the semantic similarity of text in all STS tasks which contain Arabic, English, Spanish, and Turkish this year. RTMs pioneer a language independent approach to semantic similarity and remove the need to access any task or domain specific information or resource. RTMs become 6th out of 52 submissions in Spanish to English STS. We average prediction scores using weights based on the training performance to improve the overall performance.

1 Referential Translation Machines (RTMs)

Semantic textual similarity (STS) task (Cer et al., 2017) at SemEval-2017 (Bethard et al., 2017) is about quantifying the degree of similarity between two given sentences S_1 and S_2 in the same language or in different languages. RTMs use interpretants, data close to the task instances, to derive features measuring the closeness of the test sentences to the training data, the difficulty of translating them, and to identify translation acts between any two data sets for building prediction models. RTMs are applicable in different domains and tasks and in both monolingual and bilingual settings. Figure 1 depicts RTMs and explains the model building process.

RTMs use ParFDA (Biçici, 2016a) for instance selection and machine translation performance prediction system (MTPPS) (Biçici and Way, 2015) for generating features for the training and the test set mapping both to the same space where the total number of features in each task becomes 368. The new features we include are about punctuation: number of tokens about punctuation (Kozlova et al., 2016) and the cosine between the punctuation vectors.

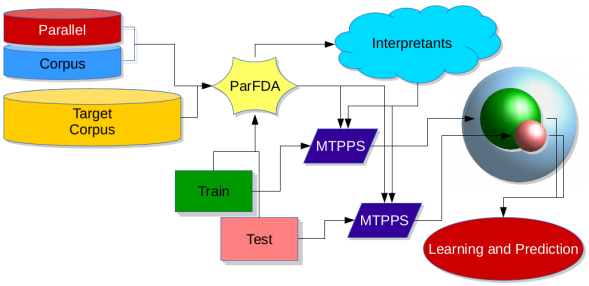

Figure 1: RTM depiction: ParFDA selects interpretants close to the training and test data using parallel corpus in bilingual settings and monolingual corpus in the target language or just the monolingual target corpus in monolingual settings; an MTPPS use interpretants and training data to generate training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

RTMs are providing a language independent text processing and machine learning model able to use predictions from different predictors. We use ridge regression (RR), k-nearest neighors (KNN), support vector regression (SVR), AdaBoost (Freund and Schapire, 1997), and extremely randomized trees (TREE) (Geurts et al., 2006) as learning models in combination with feature selection (FS) (Guyon et al., 2002) and partial least squares (PLS) (Wold et al., 1984). For most of the models, we use `scikit-learn`. [1] We optimize the models using a subset of the training data for the following parameters: λ for RR, k for KNN, γ, C, and ϵ for SVR, minimum number of samples for leaf nodes and for splitting an in-

[1] `http://scikit-learn.org/`. For RR, contains different solvers, support for sparse matrices, and checks for size and errors.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 203–207,
Vancouver, Canada, August 3 - 4, 2017.

	all	Tr.1	Tr.2	Tr.3	Tr.4a	Tr.4b	Tr.5	Tr.6
		ar-ar	ar-en	es-es	es-en	en-es	en-en	en-tr
test		250	250	250	250	250	250	500
ranks	34	'34'	34	'27'	26	6	61	'39'
out of	44	48	44	47	52	52	78	47
r	0.37	0.34	0.17	0.7	0.6	0.15	0.55	0.07

Table 1: RTM ranks and the number of instances in the STS test sets with abbreviations: Arabic (ar), English (en), Spanish (es), Turkish (tr). Only 250 instances are evaluated in en-tr. Results within single quotes used mismatching corpora and therefore we reran our experiments (Section 3).

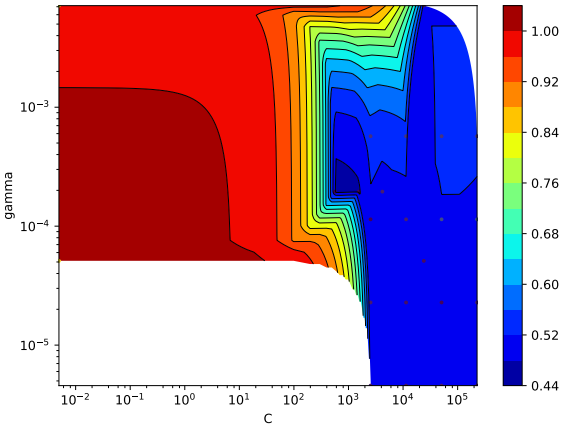

Figure 2: Sample SVR optimization plot (en-en).

ternal node for TREE, the number of features for FS, and the number of dimensions for PLS. For AdaBoost, we do not optimize but use exponential loss and 500 estimators like we use also with the TREE model. We use grid search for SVR. Figure 2 plots sample search contours.

Evaluation metrics we use are Pearson's correlation (r), mean absolute error (MAE), relative absolute error (RAE), MAER (mean absolute error relative), and MRAER (mean relative absolute error relative) (Biçici and Way, 2015). Official evaluation metric is r.

This year, we experiment with averaging scores from different models. The predictions, $\hat{y}$, are sorted according to their performance on the training set and the mean of the top k predictions (equally weighted averaging) or their weighted average according to their performance are used:

$$\hat{\boldsymbol{y}}_{\mu_k} = \frac{1}{k}\sum_{i=1}^{k}\hat{\boldsymbol{y}}_i \quad (1)$$

$$\hat{\boldsymbol{y}}_{w_k} = \frac{1}{\sum_{i=1}^{k}\frac{1}{w_i}}\sum_{i=1}^{k}\frac{1}{w_i}\hat{\boldsymbol{y}}_i \quad (2)$$

The weights are inverted since we are trying to decrease MAER and normalize by the sum. We use MAER for sorting and selecting predictions.

2 SemEval-17 STS Results

SemEval-2017 STS contains STS sentence pairs from the languages listed in Table 1 where the top r from among our officially submitted results are listed, which contain a mean averaged, a weight averaged, and a top prediction corresponding to weight 3, mean 3, and SVR model predictions. These results do not contain AdaBoost results and they are optimized less. We build individual RTM models for each subtask with RTM team name. Interpretants are selected from the corpora distributed by the translation task of WMT17 (Bojar et al., 2017) and they consist of monolingual sentences used to build the LM and parallel sentence pair instances used by MTPPS to derive the features. For monolingual STS, we use the corresponding monolingual corpora. We built RTM models using:

- 275 thousand sentences for en-en, 200 thousand sentences for en-tr, and 250 thousand sentences for others for training data
- 7 million sentences for the language model

which are close to the fixed training set size setting in (Biçici and Way, 2015).

We identified numeric expressions using regular expressions as a pre-processing step, which replaces them with a label. Identification of numerics improve the performance on the test set (Biçici, 2016b). For en-es or es-en, we did not use any language identification tool and separated sentences based on left/right difference rather than using the mixed format that was made available to the participants even though identification of the language increase r on the test set from 0.5375 to 0.6066 while decreasing error (Biçici, 2016b). For en-tr, we were not provided any training data; therefore, we used the training data from other subtasks.

3 Experiments After the Challenge

Table 2 compares the top averaging result with the top result without averaging on the test set. The

Task	r	MAE	RAE	MAER	MRAER	model
ar-ar	0.5302	1.4072	1.122	1.3068	1.331	weight 7
ar-ar	0.5286	1.3909	1.109	1.2941	1.304	TREE
ar-en	0.2144	1.5793	1.276	1.4937	1.456	mean 2
ar-en	0.2235	1.565	1.264	1.4556	1.432	FS-SVR
es-es	0.7398	0.9689	0.708	0.7756	0.746	weight 4
es-es	0.7409	0.9673	0.7072	0.7739	0.7467	FS-TREE
es-en	0.5481	1.4072	1.137	1.3229	1.362	mean 3
es-en	0.5197	1.4176	1.146	1.3483	1.328	FS-TREE
en-es	0.1101	1.3122	1.305	0.3306	1.377	weight 2
en-es	0.0847	1.3263	1.319	0.3351	1.388	TREE
en-en	0.7103	1.0261	0.852	0.8678	1.042	weight 11
en-en	0.6528	1.0644	0.883	0.9126	1.052	FS+PLS-SVR
en-tr	-0.0204	1.6094	1.2849	1.4614	1.3533	weight 8
en-tr	-0.0527	1.7121	1.3669	1.4955	1.4569	FS+PLS SVR
all	0.4105	averaging				
all	0.4011	others				

Table 2: RTM top averaged result compared with the top non averaged result on the test set. Averaging improve the performance on the test set.

	all	Tr.1	Tr.2	Tr.3	Tr.4a	Tr.4b	Tr.5	Tr.6
		ar-ar	ar-en	es-es	es-en	en-es	en-en	en-tr
ranks	33	33	34	25	33	6	53	45
out of	44	48	44	47	52	52	78	47

Table 3: RTM ranks in the STS test sets with results from Table 2.

results warn us that ar-ar, ar-en, en-es, and es-en obtain MRAER larger than 1 suggesting more work towards these tasks. en-en has slightly more than 1 in MRAER and this is worse than the 0.719 MRAER obtained by RTMs in STS in 2016. For es-es, we obtain slightly lower results compared with 0.729 MRAER of RTMs in STS in 2016 where we used language identification. The test set domain is different this year; Stanford Natural Language Inference corpus (Bowman et al., 2015) is focusing on inference and entailment tasks and entailment assumes direction and in contrast the goal in STS is the bidirectional grading of equivalence (Agirre et al., 2015). Table 3 list the ranks we can obtain with RTMs these new results. Figure 3 plots the performance on the test set where instances are sorted according to the magnitude of the target scores.

Also in this section, we present results about transfer of learning. Transfer learning attempt to re-use and transfer knowledge from models developed in different domains or for different tasks such as using models developed for handwritten digit recognition for handwritten character recognition (Guyon et al., 2012). We cross use RTM SVR models developed for different tasks as a cross-task TL [2] and present the results in Table 4 with #train listing the size of the training set used for each task. Cross use of RTM es-es model increase r for en-en from 0.71 to 0.75 and for en-ar from 0.19 to 0.50 while making all tasks except 4b en-es below the 1 MRAER threshold we seek for showing improvements in prediction performance relatively better than a predictor knowing and using the mean of the target scores on the test set.

4 Conclusion

Referential translation machines pioneer a clean and intuitive computational model for automatic prediction of semantic similarity by measuring the acts of translation involved. Averaging predictions improve the correlation on the test set.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada

[2] www.youtube.com/watch?v=9ChVn3xVNDI; we have the same domain of STS but we use the models for different tasks.

			test						
	r	#train	ar-ar	en-ar	es-es	es-en	en-es	en-en	en-tr
train	ar-ar	1105	0.4391	0.1053	0.0885	-0.0153	-0.0554	0.5535	-0.1235
	en-ar	2186	-0.0773	0.1938	0.0596	-0.1587	-0.0138	-0.0861	0.0036
	es-es	1644	0.5235	0.4953	**0.7342**	0.4051	0.0238	**0.7503**	0.3888
	es-en	1722	**0.5947**	0.3572	0.6886	0.4017	0.1591	0.6798	0.4781
	en-es	1722	0.5643	**0.5616**	0.666	**0.6052**	**0.2141**	0.6794	**0.4998**
	en-en	15672	0.57	0.2963	0.6841	0.2213	-0.0933	0.7109	0.0817
	en-tr	22329	0.4242	0.2222	0.3914	-0.0671	-0.0638	0.4075	-0.0074
	MAER	# train	ar-ar	en-ar	es-es	es-en	en-es	en-en	en-tr
train	ar-ar	1105	1.2202	1.5205	1.4414	1.5653	0.3624	1.0899	1.676
	en-ar	2186	1.6913	1.5928	1.6145	1.819	0.4261	1.6371	1.8628
	es-es	1644	**0.8667**	**0.9702**	**0.7136**	**0.9997**	0.3966	**0.6175**	**1.0874**
	es-en	1722	1.332	1.4329	1.3814	1.444	**0.3051**	1.2728	1.4783
	en-es	1722	1.012	1.1449	0.978	1.099	0.3246	0.8882	1.2638
	en-en	15672	0.9224	1.3786	0.9324	1.3048	0.4329	0.8031	1.4932
	en-tr	22329	1.044	1.2837	1.1252	1.3961	0.4787	1.0383	1.4959
	MRAER	# train	ar-ar	en-ar	es-es	es-en	en-es	en-en	en-tr
train	ar-ar	1105	1.24	1.357	1.251	1.459	1.549	1.16	1.415
	en-ar	2186	1.775	1.6	1.546	1.731	1.663	1.648	1.644
	es-es	1644	**0.943**	**0.935**	**0.759**	**0.962**	1.896	**0.735**	**0.934**
	es-en	1722	1.168	1.203	1.126	1.21	1.408	1.08	1.173
	en-es	1722	1.2	1.249	1.038	1.25	**1.385**	1.104	1.193
	en-en	15672	1.127	1.434	0.982	1.297	1.785	1.081	1.446
	en-tr	22329	1.271	1.415	1.146	1.416	1.97	1.169	1.492

Table 4: RTM SVR model (rows) r, MAER, and MRAER results on the test sets (columns).

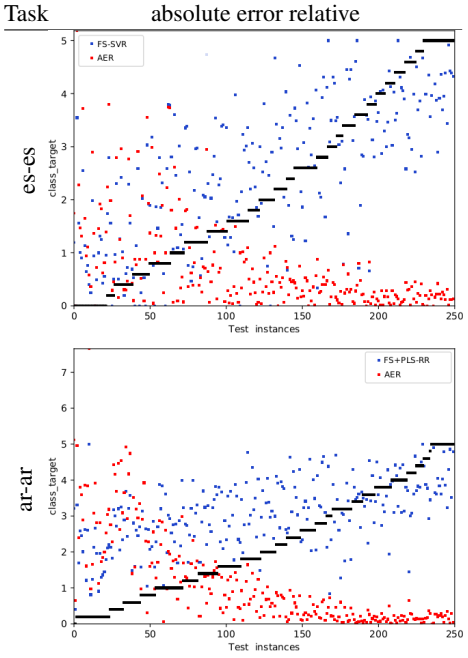

Figure 3: RTM's top predictor's absolute errors relative to the magnitude of the target.

Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, pages 252–263. www.aclweb.org/anthology/S15-2045.

Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens. 2017. Proc. of the 11th international workshop on semantic evaluation (semeval-2017). In *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics.

Ergun Biçici. 2016a. ParFDA for instance selection for statistical machine translation. In *Proc. of the First Conference on Statistical Machine Translation (WMT16)*. Association for Computational Linguistics, Berlin, Germany. http://aclanthology.info/papers/parfda-for-instance-selection-for-statistical-machine-translation.

Ergun Biçici. 2016b. 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, CA, USA. http://aclanthology.info/papers/rtm-at-semeval-2016-task-1-predicting-semantic-similarity-with-referential-translation-machines-and-related-statistics.

Ergun Biçici and Andy Way. 2015. Referential translation machines for predicting semantic similarity. *Language Resources and Evaluation* pages 1–27. https://doi.org/10.1007/s10579-015-9322-7.

Ondrej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Jimeno Antonio Yepes, Julia Kreutzer, Varvara Logacheva, Aurelie Neveol, Mariana Neves, Philipp Koehn, Christof Monz, Matteo

Negri, Matt Post, Stefan Riezler, Artem Sokolov, Lucia Specia, Karin Verspoor, and Marco Turchi. 2017. Proc. of the second conference on Machine Translation. In *Proc. of the Second Conference on Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 1–14. http://www.aclweb.org/anthology/S17-2001.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1):119–139. https://doi.org/10.1006/jcss.1997.1504.

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

Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham W. Taylor, and Daniel L. Silver, editors. 2012. *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, volume 27 of *JMLR Proceedings*. JMLR.org. http://clopinet.com/ul.

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. https://doi.org/10.1023/A:1012487302797.

Anna Kozlova, Mariya Shmatova, and Anton Frolov. 2016. Ysda participation in the wmt'16 quality estimation shared task. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 793–799. http://www.aclweb.org/anthology/W/W16/W16-2385.

S. Wold, A. Ruhe, H. Wold, and W. J. III Dunn. 1984. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing* 5:735–743.