RTM at SemEval-2016 Task 1: Predicting Semantic Similarity with Referential Translation Machines and Related Statistics

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

Abstract

We use referential translation machines (RTMs) for predicting the semantic similarity of text in both STS Core and Cross-lingual STS. RTMs pioneer a language independent approach to all similarity tasks and remove the need to access any task or domain specific information or resource. RTMs become 14th out of 26 submissions in Cross-lingual STS. We also present rankings of various prediction tasks using the performance of RTM in terms of MRAER, a normalized relative absolute error metric.

1 Semantic Agreement

We participated in Semantic Textual Similarity task at SemEval-2016 (Bethard et al., 2016) with RTMs. RTMs identify translation acts between any two data sets with respect to interpretants, data close to the task instances, effectively judging monolingual and bilingual similarity. We use RTMs for predicting the semantic similarity of text. Interpretants are used to derive features measuring the closeness of the test sentences to the training data, the difficulty of translating them, and the presence of the acts of translation, which may ubiquitously be observed in communication.

Semantic Web's dream is to allow machines to share, exploit, and understand knowledge on the web. As more and more shared conceptualizations of domains emerge, we get closer to this goal. Semantic textual similarity (STS) task (Agirre et al., 2016) at SemEval-2016 (Bethard et al., 2016) is about quantifying the degree of similarity between two given sentences S_1 and S_2 in the same language (English) in STS Core (STS English) or in different languages (English or Spanish) in Cross-lingual STS (STS Spanish), with a real number in $[0, 5]$. S_1 and S_2 may be constructed using different models and with different conceptualizations of the world or different ontologies and different vocabulary. Even if two instances are categorized as same, they may have different implications for commonsense reasoning (both albatros and penguin are a bird) (Biçici, 2002).

The existence of a single ontology that can cover all the required conceptual information for reaching semantic understanding is questionable because it would presume an agreement among all ontology experts. Yet, semantic agreement using heterogeneous ontologies may not be possible as well since in the most extreme case, they would not use the same tokens. Therefore, semantic textual similarity is harder than the Chinese room thought experiment (Internet Encyclopedia of Philosophy, 2016) since we are not given any instructions about how to answer queries. Our goal is to quantify the level of semantic agreement between S_1 and S_2 and RTMs use interpretants, data close to the task instances for building prediction models for semantic similarity.

2 Referential Translation Machine

Each 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

Proceedings of SemEval-2016, pages 1342–1348,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

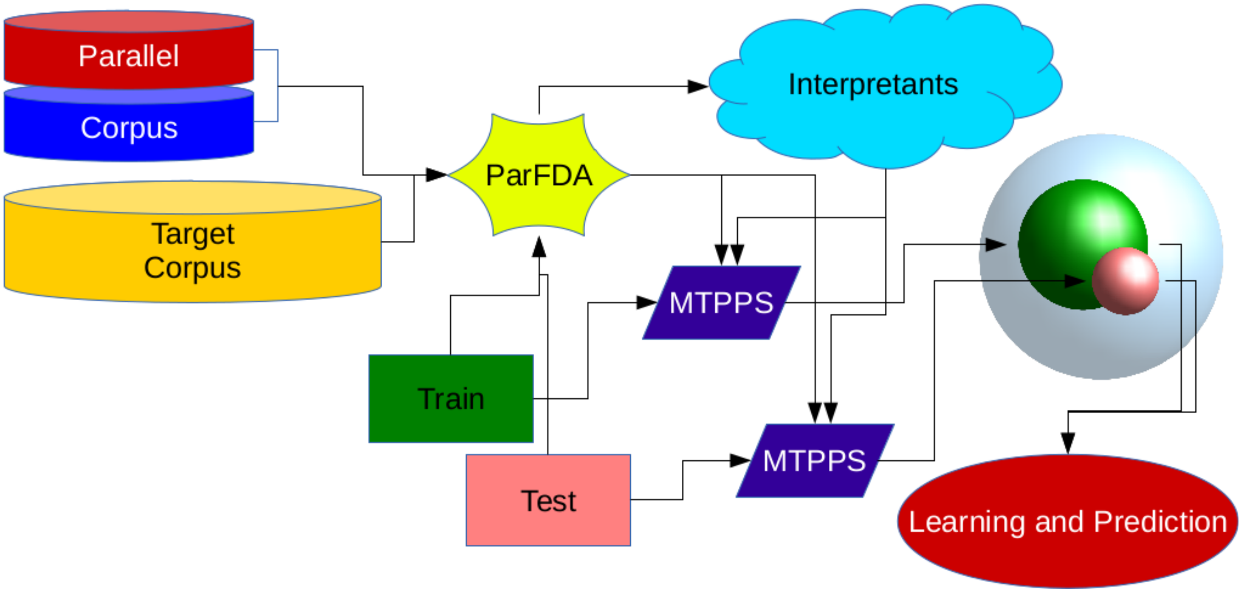


Figure 1: RTM depiction: ParFDA selects interpretants close to the training and test data using parallel corpus in bilingual settings and monolingual corpus in the target language or just the monolingual target corpus in monolingual settings; an MTPPS use interpretants and training data to generate training 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.

		ans.-ans.	headlines	plagiarism	postediting	que.-que.
STS	base	1572	1498	1271	3287	1555
English	eval.	254	249	230	244	209
		multisource	newswire			
STS	base	2973	301			
Spanish	eval.	294	301			

Table 1: Number of instances in the STS test set. Only some of the instances are actually evaluated (eval. row).

monolingual (Biçici and Way, 2015) and bilingual settings (Biçici et al., 2015b). Our encouraging results in the semantic similarity tasks increase our understanding of the acts of translation we ubiquitously use when communicating and how they can be used to predict semantic similarity.

Figure 1 depicts RTMs and explains the model building process. Given a training set $\mathtt{train}$, a test set $\mathtt{test}$, and some corpus $\mathcal{C}$, preferably in the same domain, the RTM steps are:

1. select$(\mathtt{train}, \mathtt{test}, \mathcal{C}) \rightarrow \mathcal{I}$
2. MTPP$(\mathcal{I}, \mathtt{train}) \rightarrow \mathcal{F}_{\mathtt{train}}$
3. MTPP$(\mathcal{I}, \mathtt{test}) \rightarrow \mathcal{F}_{\mathtt{test}}$
4. learn$(M, \mathcal{F}_{\mathtt{train}}) \rightarrow \mathcal{M}$
5. predict$(\mathcal{M}, \mathcal{F}_{\mathtt{test}}) \rightarrow \hat{y}$

RTMs use ParFDA (Biçici et al., 2015a) for instance selection and machine translation performance prediction system (MTPPS) (Biçici and Way, 2015) for generating features.

We use support vector regression (SVR) for building the predictor in combination with feature selection (FS) and partial least squares (PLS). Assuming that $\hat{\mathbf{y}}, \mathbf{y} \in \mathbb{R}^n$ are the prediction and the target respectively, evaluation metrics we use are defined in Equation (1) where metrics are Pearson's correlation (r), mean absolute error (MAE), relative absolute error (RAE), relative Pearson's correlation (r_R), MAER (mean absolute error relative), and MRAER (mean relative absolute error relative).

We use MAER and MRAER for easier replication and comparability. MAER is the mean absolute error relative to the magnitude of the target and MRAER is the mean absolute error relative to the absolute error of a predictor always predicting the target mean assuming that target mean is known (Biçici and Way, 2015). $\lfloor\ .\ \rfloor_\epsilon$ caps its argument from below to ϵ where $\epsilon =$ MAE$(\hat{\mathbf{y}}, \mathbf{y})/2$, which represents half of the score step with which a decision about a change in measurement's value can be made.

	Domain r											
Model	ans-ans.	headlines	plagiarism	postediting	que.-que.	Weighted r	r	r_R	MAE	RAE	MAER	MRAER
SVR	0.4486	0.6634	0.8038	0.8133	0.6237	0.6685	0.6506	0.7563	1.015	0.679	0.5819	0.726
PLS-SVR	0.344	0.6605	0.8064	0.8231	0.6454	0.6518	0.6386	0.7786	1.0228	0.684	0.5779	0.739
FS+PLS-SVR	0.3533	0.6529	0.8049	0.823	0.648	0.6524	0.6369	0.7733	1.0243	0.685	0.5766	0.742

Table 2: STS English test results for each domain.

	Domain r									
Model	Multisource r	News r	Weighted r	Rank	r	r_R	MAE	RAE	MAER	MRAER
FS+PLS-SVR	0.5204	0.5915	0.5564	14	0.5244	0.5291	1.241	0.809	0.8812	0.856
SVR	0.5294	0.4985	0.5137	16	0.4455	0.4075	1.3473	0.878	0.9933	0.924
FS-SVR	0.5284	0.536	0.5322	15	0.4691	0.444	1.3094	0.853	0.9441	0.891

Table 3: STS Spanish test results.

$$
\begin{aligned}
r &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
\text{MAE}(\hat{\mathbf{y}}, \mathbf{y}) &= \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n} \\
\text{RAE}(\hat{\mathbf{y}}, \mathbf{y}) &= \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{\sum_{i=1}^{n}|\bar{y} - y_i|} \\
r_R &= \frac{\sum_{i=1}^{n}\left(\frac{\hat{y}_i - \bar{\hat{y}}}{\lfloor |y_i| \rfloor_\epsilon}\right)\left(\frac{y_i - \bar{y}}{\lfloor |y_i| \rfloor_\epsilon}\right)}{\sqrt{\sum_{i=1}^{n}\left(\frac{\hat{y}_i - \bar{\hat{y}}}{\lfloor |y_i| \rfloor_\epsilon}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{\lfloor |y_i| \rfloor_\epsilon}\right)^2}} \\
\text{MAER}(\hat{\mathbf{y}}, \mathbf{y}) &= \frac{\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{\lfloor |y_i| \rfloor_\epsilon}}{n} \\
\text{MRAER}(\hat{\mathbf{y}}, \mathbf{y}) &= \frac{\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{\lfloor |\bar{y} - y_i| \rfloor_\epsilon}}{n}
\end{aligned}
\tag{1}
$$

We compare different tasks in Table 9 with evaluation results that are calculated relative to the magnitude of each target score instance. r multiplies distance of $\hat{y}_i$ and y_i to their own means (Equation (1)). We obtain normalized correlation, r_R, using $\epsilon = \sigma(\mathbf{y})/2$.

3 SemEval-16 STS Results

SemEval-2016 STS contains sentence pairs from different domains: answer-answer, headlines, plagiarism, postediting, question-question for English and multisource and newswire for Spanish. Official evaluation metric in STS is the Pearson's correlation score. Table 1 lists the number of instances in the test set where only some of the instances are actually evaluated.

We build individual RTM models for each subtask. Our team name is RTM. Interpretants are selected from the corpora distributed by the translation task of WMT16 (Bojar et al., 2016) and they consist of monolingual sentences used to build the LM and parallel sentence pair instances used by MTPPS to derive features and for word alignment features. We use monolingual corpora in English for STS English to select interpretants and also for STS Spanish shuffled dataset, which is the official format that was made available to the participants.

We used English-Spanish parallel corpus and English and Spanish monolingual corpora for our STS Spanish experiments in Section 4 after the challenge using the language identified version. We built RTM models using 200 thousand sentences for training data and 5 million sentences for the language model, which corresponds to the fixed training set size setting in (Biçici and Way, 2015). We identified numeric expressions using regular expressions as a preprocessing step, which replaces them with a label. For training RTM models for STS Spanish, we use STS English training data and STS Spanish data from SemEval-2015 after scaling the scores to range

Task	Setting		r	MAE	RAE	MAER	MRAER
train		SVR	0.74	0.8028	0.612	0.4745	0.69
	+numerics	SVR	0.73	0.8108	0.618	0.4758	0.698
test		SVR	0.65	1.0224	0.684	0.6074	0.719
	+numerics	SVR	**0.66**	**1.0052**	**0.673**	**0.5954**	0.719

Table 4: RTM top predictor results on STS English show that performance improve after identification of numerics on the test set.

Setting	Model	ans.-ans.	headlines	plagiarism	postediting	que.-que.	Weighted r	r	r_R	MAE	RAE	MAER	MRAER
		Domain % numerics											
		1.4	3.2	0.33	1.2	0.3	1.01 (% of total)						
		Domain r											
	SVR	0.4458	0.6813	0.8	0.7881	0.6218	0.6654	0.6549	0.7441	1.0224	0.684	0.6074	0.719
+numerics	SVR	**0.4978**	0.6767	0.7956	**0.7983**	0.6096	**0.6746**	**0.6632**	**0.7612**	**1.0052**	**0.673**	**0.5954**	**0.719**

Table 5: STS English test results for each domain from new experiments. Domain % numerics lists the percentage of tokens classified as numerics in each domain.

$[0, 5]$.

Table 2 and Table 3 list the results on the test set. Ranks are out of 26 submissions in STS Spanish. We also observe that r over all of the test set, which does not compute the weighted average of r according to the number of instances in each domain can differ from the weighted r scores.

4 Experiments After the Challenge

In this section, we detail the training performance of our model based on major modeling differences with our previous RTM models on SemEval tasks (Biçici and Way, 2015). This year, we identified numeric expressions using regular expressions as a preprocessing step, which mainly identifies integers and real numbers that can have exponents. After sending the test results, we further worked on the numeric expression identification to expand the types of identified expressions. We also experimented with language identification for STS Spanish. Language identification is done using the manually corrected results starting from the output of automatic language identification tool mguesser. [1] After language identification, corpora were split into English and Spanish rather than the shuffled format that was made available to the participants. We compare the performance after identification of numeric expressions and identification of the language using SVR. Both STS Spanish models use previous years' training data from both STS English and STS Spanish for training, which total to 13823 instances.

Table 4 and Table 5 presents the results before and after identification of numerics on STS English. We observe that identification of numerics improve the performance on the test set (**bolded** results).

Table 6 presents the results on STS Spanish with the default shuffled setting and with the setting where we model the prediction as machine translation performance prediction from English to Spanish after identifying the language of each sentence in the training set. STS Spanish training dataset contains English sentences in majority and shuffled +numerics setting only use English corpora even though Spanish sentences are shuffled in the test set and eventually, shuffled +numerics setting obtains better results than language identified +numerics setting on the training set. Even so, we observe that identification of the language improve the performance on the test set (**bolded** results). Training results on setting language identified +numerics is lower, which may be due to the RTM model using language identified test corpus and the same training corpus as the shuffled +numerics setting.

Table 8 plots the performance on the test set where instances are sorted according to the magnitude of the target scores. For STS English, we observe decreasing AER and a valley of absolute errors, which may be due to SVR preferring predictions close to the mean of train score distribution.

5 RTMs Across Tasks and Years

We compare the difficulty of various prediction tasks where RTMs participated (Biçici and Way, 2015) ac-

[1] `http://www.mnogosearch.org`

Task	Setting		r	MAE	RAE	MAER	MRAER
train	shuffled +numerics	SVR	0.72	0.8224	0.639	0.4864	0.718
	language identified +numerics	SVR	0.69	0.8567	0.666	0.5261	0.74
test	shuffled +numerics	SVR	0.3687	1.4589	0.951	1.0949	1.04
	language identified +numerics	SVR	**0.6739**	**1.0529**	**0.686**	**0.7087**	**0.729**

Table 6: RTM SVR results on STS Spanish show that performance improve after language identification on the test set.

		Domain r									
Setting	Model	Multisource r	News r	Weighted r	Rank	r	r_R	MAE	RAE	MAER	MRAER
shuffled +numerics	SVR	0.5375	0.4498	0.4931	17	0.3687	0.3722	1.4589	0.951	1.0949	1.04
language identified +numerics	SVR	**0.6066**	**0.7225**	**0.6652**	13	**0.6739**	**0.6604**	**1.0529**	**0.686**	**0.7087**	**0.729**

Table 7: STS Spanish test results from new experiments.

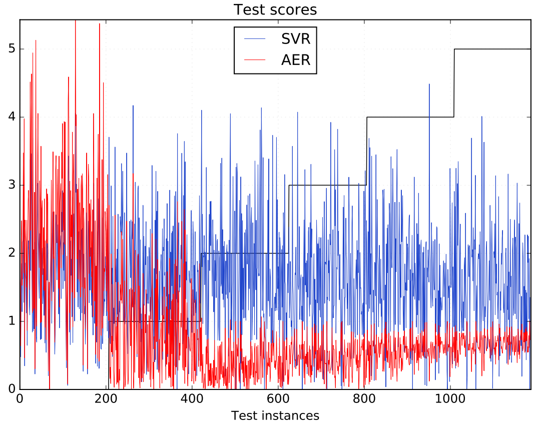
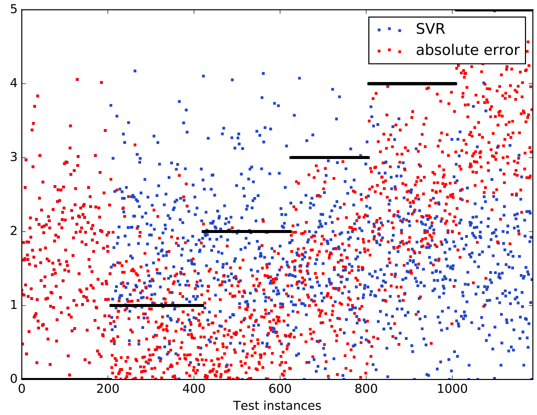
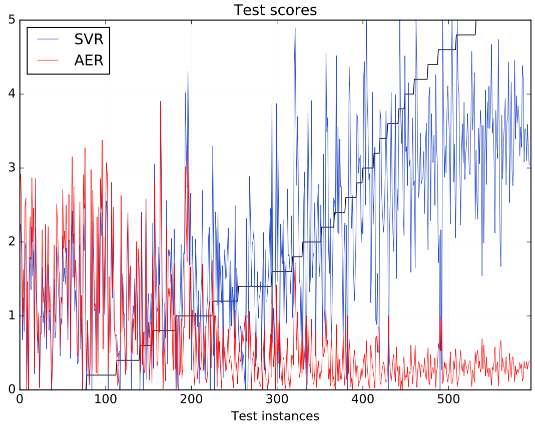
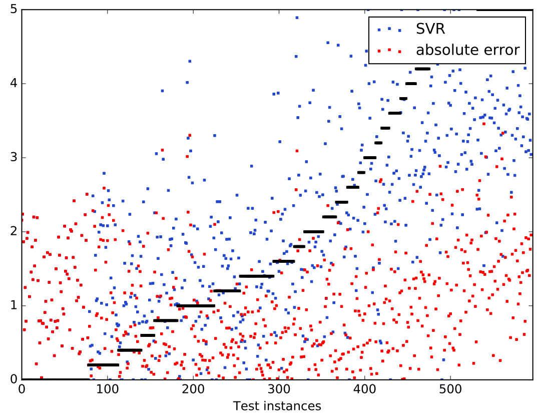
Task	absolute error relative	absolute error
STS English		
STS Spanish		

Table 8: RTM SVR performance on the test set in STS 2016. Left figure in each row is the absolute error relative to the magnitude of the target (AER) and the right figure is the absolute error.

cording to MRAER in Table 9. MAER and MRAER considers both the predictor's error and the fluctuations of the target scores at the instance level, which is at the sentence level in STS 2016. The best results are obtained for the CLSS 2014 paragraph-to-sentence subtask, which may be due to the larger contextual information that paragraphs can provide for the RTM models. We observe that the performance in STS improved in 2016 compared to STS in previous years. Table 9 can be used to evaluate the difficulty of various tasks and domains based on RTM. We separated the results having MRAER greater than 1 as in these tasks and subtasks RTM does not perform significantly better than the mean predictor, and fluctuations render these as tasks that may require more work. Our findings are negative towards re-use of those datasets and results obtained without further work. STS Spanish is able to achieve MRAER less than 1 in 2016. We also note that RTMs achieve the top result in both CLSS 2014 (Jurgens et al., 2014) and in all QET tasks in Table 9, including the QET 2015 German-English METEOR task (Bojar et al., 2015).

6 Contributions

Referential translation machines pioneer a clean and intuitive computational model for automatically measuring semantic similarity by measuring the acts of translation involved. We show that identification of numeric expressions in STS English and identification of the language in STS Spanish improve the performance on the test set. RTM test performance on various tasks sorted according to MRAER can identify which tasks and subtasks and the datasets provided are mature enough for further results.

Acknowledgments

We thank the reviewers for providing constructive comments.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, USA, June. Association for Computational Linguistics.

Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch. 2016. Proc. of the 10th international workshop on semantic evaluation (semeval 2016). In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. 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 *Proc. of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September. Association for Computational Linguistics.

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

Ergun Biçici. 2002. Prolegomenon to commonsense reasoning in user interfaces. *ACM Crossroads*, 9(1).

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.

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 workshop on statistical machine translation. In *Proc. of the Eleventh Workshop on Statistical Machine Translation*, Berlin, Germany, August.

Internet Encyclopedia of Philosophy. 2016. Chinese room thought experiment.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. SemEval-2014 Task 3: Cross-level semantic similarity. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland, August.

Task	Subtask	Domain	RAE	MAER	MRAER
CLSS 2014	Paragraph to Sentence	Mixed	0.458	0.5112	0.504
STS 2014	English	OnWN	0.558	0.7975	0.546
QET 2014	English-Spanish PEE	Europarl	1.079	0.304	0.614
STS 2015	English	Images	0.588	0.5424	0.623
STS 2015	English	Headlines	0.589	0.4844	0.638
CLSS 2014	Sentence to Phrase	Mixed	0.626	0.6857	0.644
QET 2015	English-German METEOR	Europarl	0.7279	0.3249	0.647
QET 2014	German-English PEE	Europarl	0.82	0.3575	0.679
QET 2014	English-German PEE	Europarl	0.86	0.3692	0.698
STS 2016	English	ALL	0.673	0.5954	0.719
STS 2014	English	Images	0.74	0.8338	0.725
STS 2016	Spanish	ALL	0.686	0.7087	0.729
QET 2014	Spanish-English PEE	Europarl	0.9	0.3798	0.749
STS 2014	English	ALL	0.745	0.7274	0.757
STS 2013	English	ALL	0.779	0.8494	0.77
QET 2014	English-Spanish PET	Europarl	0.722	0.4651	0.779
STS 2014	English	Headlines	0.784	0.6711	0.785
STS 2015	English	ALL	0.722	0.7379	0.788
STS 2014	English	Tweet-news	0.714	0.4225	0.797
SRE 2014	English	SICK	0.664	0.1827	0.818
STS 2015	English	Answers-students	0.782	0.5542	0.84
CLSS 2014	Phrase to Word	Mixed	0.949	1.1454	0.848
QET 2015	English-Spanish HTER	Europarl	0.896	0.8344	0.849
STS 2013	English	OnWN	0.826	1.2875	0.86
ParSS 2015	English	Tweets	0.788	0.6788	0.862
QET 2014	English-Spanish HTER	Europarl	0.853	0.7727	0.876
STS 2014	English	Deft-news	0.872	0.6271	0.881
QET 2015	German-English METEOR	Europarl	0.876	0.395	0.916
STS 2015	Spanish	News	0.898	0.3757	1.089
STS 2015	Spanish	ALL	0.889	0.3883	1.094
STS 2015	English	Answers-forums	1.06	1.3883	1.107
STS 2015	Spanish	Wikipedia	0.868	0.413	1.121
STS 2013	English	Headlines	1.023	1.0456	1.144
STS 2014	English	Deft-forum	1.091	0.7724	1.216
STS 2015	English	Belief	1.153	1.5882	1.224
STS 2013	English	FNWN	1.263	1.5087	1.405
STS 2014	Spanish	News	1.157	0.4773	1.492
QET 2013	English-Spanish HTER	Europarl	0.885	2.3738	1.643
STS 2014	Spanish	ALL	1.251	0.5345	1.657
STS 2014	Spanish	Wikipedia	1.358	0.65	1.661
STS 2013	English	SMT	1.613	0.1669	2.072

Table 9: Best RTM test results for different tasks and subtasks sorted according to MRAER.