RTM-DCU: Referential Translation Machines for Semantic Similarity

Ergun Biçici
Centre for Global Intelligent Content
School of Computing
Dublin City University, Dublin, Ireland.
ergun.bicici@computing.dcu.ie

Andy Way
Centre for Global Intelligent Content
School of Computing,
Dublin City University, Dublin, Ireland.
away@computing.dcu.ie

Abstract

We use referential translation machines (RTMs) for predicting the semantic similarity of text. RTMs are a computational model for identifying the translation acts between any two data sets with respect to interpretants selected in the same domain, which are effective when making monolingual and bilingual similarity judgments. RTMs judge the quality or the semantic similarity of text by using retrieved relevant training data as interpretants for reaching shared semantics. We derive features measuring the closeness of the test sentences to the training data via interpretants, the difficulty of translating them, and the presence of the acts of translation, which may ubiquitously be observed in communication. RTMs provide a language independent approach to all similarity tasks and achieve top performance when predicting monolingual cross-level semantic similarity (Task 3) and good results in semantic relatedness and entailment (Task 1) and multilingual semantic textual similarity (STS) (Task 10). RTMs remove the need to access any task or domain specific information or resource.

1 Semantic Similarity Judgments

We introduce a fully automated judge for semantic similarity that performs well in three semantic similarity tasks at SemEval-2014, Semantic Evaluation Exercises - International Workshop on Semantic Evaluation (Nakov and Zesch, 2014). RTMs provide a language independent solution for the semantic textual similarity (STS) task (Task 10) (Agirre et al., 2014), achieve top performance when predicting monolingual cross-level semantic similarity (Task 3) (Jurgens et al., 2014), and achieve good results in the semantic relatedness and entailment task (Task 1) (Marelli et al., 2014a).

Referential translation machine (Section 2) is a computational model for identifying the acts of translation for translating between any given two data sets with respect to a reference corpus selected in the same domain. An RTM model is based on the selection of interpretants, training data close to both the training set and the test set, which allow shared semantics by providing context for similarity judgments. In semiotics, an interpretant I interprets the signs used to refer to the real objects (Biçici, 2008). Each RTM model is a data translation and 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 present an accurate and language independent solution for making semantic similarity judgments.

We describe the tasks we participated below. Section 2 describes the RTM model and the features used. Section 3 presents the training and test results we obtain on the three tasks we competed and the last section concludes.

Task 1 Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Entailment (SRE) (Marelli et al., 2014a):

Given two sentences, produce a relatedness score indicating the extent to which the sentences express a related meaning: a number in the range $[1, 5]$.

We model the problem as a translation performance prediction task where one possible interpretation is obtained by translating S_1 (the source to translate, S) to S_2 (the target translation, T). Since linguistic processing can reveal deeper similarity relationships, we also look at the translation task at different granularities of information: plain

text (R for regular) and after lemmatization (L). We lowercase all text.

Task 3 Cross-Level Semantic Similarity (CLSS) (Jurgens et al., 2014):

Given two text from different levels, produce a semantic similarity rating: a number in the range $[0, 4]$.

CLSS task targets semantic similarity comparisons between text having different levels of granularity and we address the following level crossings: paragraph to sentence, sentence to phrase, and phrase to word. We model the problem as a translation performance prediction task among text from different levels.

Task 10 Multilingual Semantic Textual Similarity (MSTS) (Agirre et al., 2014)

Given two sentences S_1 *and* S_2 *in the same language, quantify the degree of similarity: a number in the range* $[0, 5]$.

MSTS task addresses the problem in English and Spanish (score range is $[0, 4]$). We model the problem as a translation performance prediction task between S_1 and S_2.

2 Referential Translation Machine (RTM)

Referential translation machines provide a computational model for quality and semantic similarity judgments in monolingual and bilingual settings using retrieval of relevant training data (Biçici, 2011; Biçici and Yuret, 2014) as interpretants for reaching shared semantics (Biçici, 2008). RTMs are a language independent approach and achieve top performance when predicting the quality of translations (Biçici, 2013; Biçici and Way, 2014) and when predicting monolingual cross-level semantic similarity (Jurgens et al., 2014), and good performance when evaluating the semantic relatedness of sentences and their entailment (Marelli et al., 2014a), as an automated student answer grader (Biçici and van Genabith, 2013b), and when judging the semantic similarity of sentences (Biçici and van Genabith, 2013a; Agirre et al., 2014). We improve the RTM models by:

- using a parameterized, fast implementation of FDA, FDA5, and our Parallel FDA5 instance selection model (Biçici et al., 2014),
- better modeling of the language in which similarity judgments are made with improved optimization and selection of the LM data,
- using a general domain corpus to select interpretants from,
- increased feature set for also modeling the structural properties of sentences,
- extended learning models.

Algorithm 1: Referential Translation Machine

Input: Training set `train`, test set `test`, corpus $\mathcal{C}$, and learning model M.
Data: Features of `train` and `test`, $\mathcal{F}_{\texttt{train}}$ and $\mathcal{F}_{\texttt{test}}$.
Output: Predictions of similarity scores on the test $\hat{q}$.

1 FDA5$(\texttt{train}, \texttt{test}, \mathcal{C}) \rightarrow \mathcal{I}$
2 MTPP$(\mathcal{I}, \texttt{train}) \rightarrow \mathcal{F}_{\texttt{train}}$
3 MTPP$(\mathcal{I}, \texttt{test}) \rightarrow \mathcal{F}_{\texttt{test}}$
4 learn$(M, \mathcal{F}_{\texttt{train}}) \rightarrow \mathcal{M}$
5 predict$(\mathcal{M}, \mathcal{F}_{\texttt{test}}) \rightarrow \hat{q}$

We use the Parallel FDA5 (Feature Decay Algorithms) instance selection model for selecting the interpretants (Biçici et al., 2014; Biçici and Yuret, 2014) this year, which allows efficient parameterization, optimization, and implementation of FDA, and build an MTPP model (Section 2.1). We view that acts of translation are ubiquitously used during communication:

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

Translation need not be between different languages and paraphrasing or communication also contain acts of translation. When creating sentences, we use our background knowledge and translate information content according to the current context.

The inputs to the RTM algorithm Algorithm 1 are a training set `train`, a test set `test`, some corpus $\mathcal{C}$, preferably in the same domain as the training and test sets, and a learning model. Step 1 selects the interpretants, $\mathcal{I}$, relevant to both the training and test data. Steps 2 and 3 use $\mathcal{I}$ to map `train` and `test` to a new space where similarities between translation acts can be derived more easily. Step 4 trains a learning model M over the training features, $\mathcal{F}_{\texttt{train}}$, and Step 5 obtains the predictions. Figure 1 depicts the RTM.

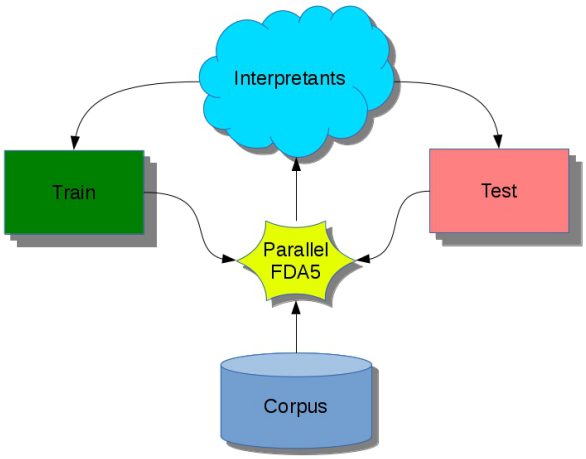


Figure 1: RTM depiction.

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 the semantic similarity of text. RTM and MTPP models are not data or language specific and their modeling power and good performance are applicable in different domains and tasks. RTM expands the applicability of MTPP by making it feasible when making monolingual quality and similarity judgments and it enhances the computational scalability by building models over smaller and more relevant set of interpretants.

2.1 The Machine Translation Performance Predictor (MTPP)

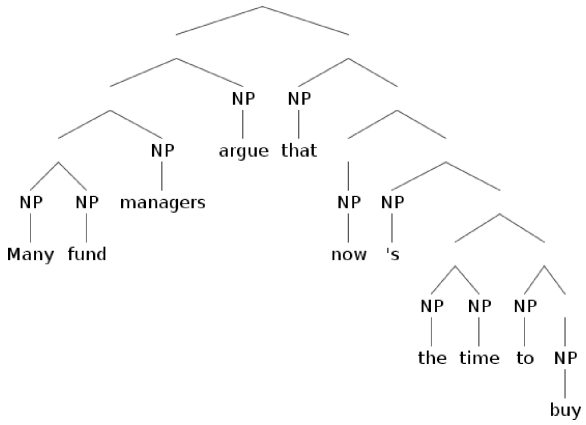
MTPP (Biçici et al., 2013) is a state-of-the-art and top performing machine translation performance predictor, which uses machine learning models over features measuring how well the test set matches the training set to predict the quality of a translation without using a reference translation. MTPP measures the coverage of individual test sentence features found in the training set and 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.

2.2 MTPP Features for Translation Acts

MTPP feature functions use statistics involving the training set and the test sentences to determine their closeness. Since they are language independent, MTPP allows quality estimation to be performed extrinsically. MTPP uses n-gram features defined over text or common cover link (CCL) (Seginer, 2007) structures as the basic units of information over which similarity calculations are made. Unsupervised parsing with CCL extracts links from base words to head words, representing the grammatical information instantiated in the training and test data.

We extend the MTPP model we used last year (Biçici, 2013) in its learning module and the features included. Categories for the features (S for source, T for target) used are listed below where the number of features are given in brackets for S and T, $\{$#S, #T$\}$, and the detailed descriptions for some of the features are presented in (Biçici et al., 2013). The number of features for each task differs since we perform an initial feature selection step on the tree structural features (Section 2.3). The number of features are in the range $337-437$.

- *Coverage* $\{56, 54\}$*:* Measures the degree to which the test features are found in the training set for both S ($\{56\}$) and T ($\{54\}$).
- *Perplexity* $\{45, 45\}$*:* Measures the fluency of the sentences according to language models (LM). We use both forward ($\{30\}$) and backward ($\{15\}$) LM features for S and T.
- *TreeF* $\{0, 10\text{-}110\}$*:* 10 base features and up to 100 selected features of T among parse tree structures (Section 2.3).
- *Retrieval Closeness* $\{16, 12\}$*:* Measures the degree to which sentences close to the test set are found in the selected training set, $\mathcal{I}$, using FDA (Biçici and Yuret, 2011a) and BLEU, F_1 (Biçici, 2011), *dice*, and tf-idf cosine similarity metrics.
- *IBM2 Alignment Features* $\{0, 22\}$*:* Calculates the sum of the entropy of the distribution of alignment probabilities obtained according to the IBM2 model (Brown et al., 1993) for S ($\sum_{s \in S} -p \log p$ for $p = p(t|s)$ where s and t are tokens) and T, their average for S and T, the number of entries with $p \geq 0.2$ and $p \geq 0.01$, the entropy of the word alignment between S and T and its average, and word alignment log probability and its value in terms of bits per word. We also compute word alignment percentage as in (Camargo de Souza et al., 2013) and potential BLEU, F_1, WER, PER scores for S and T.
- *IBM1 Translation Probability* $\{4, 12\}$*:* Calculates the translation probability of test sentences using the selected training set, $\mathcal{I}$ (Brown et al., 1993).



CCL

numB	depthB	avg depthB	R/L	avg R/L
24.0	9.0	0.375	2.1429	3.401

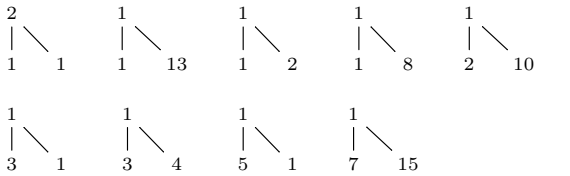


Table 1: Tree features for a parsing output by CCL (immediate non-terminals replaced with NP).

- *Feature Vector Similarity {8, 8}:* Calculates similarities between vector representations.
- *Entropy {2, 8}:* Calculates the distributional similarity of test sentences to the training set over top N retrieved sentences (Biçici et al., 2013).
- *Length {6, 3}:* Calculates the number of words and characters for S and T and their average token lengths and their ratios.
- *Diversity {3, 3}:* Measures the diversity of co-occurring features in the training set (Biçici et al., 2013).
- *Synthetic Translation Performance {3, 3}:* Calculates translation scores achievable according to the n-gram coverage.
- *Character n-grams {5}:* Calculates cosine between character n-grams (for n=2,3,4,5,6) obtained for S and T (Bär et al., 2012).
- *Minimum Bayes Retrieval Risk {0, 4}:* Calculates the translation probability for the translation having the minimum Bayes risk among the retrieved training instances.
- *Sentence Translation Performance {0, 3}:* Calculates translation scores obtained according to $q(T, R)$ using BLEU (Papineni et al., 2002), NIST (Doddington, 2002), or F_1 (Biçici and Yuret, 2011b) for q.
- *LIX {1, 1}:* Calculates the LIX readability score (Wikipedia, 2013; Björnsson, 1968) for S and T. [1]

2.3 Bracketing Tree Structural Features

We use the parse tree outputs obtained by CCL to derive features based on the bracketing structure. We derive 5 statistics based on the geometric properties of the parse trees: number of brackets used (numB), depth (depthB), average depth (avg depthB), number of brackets on the right branches over the number of brackets on the left (R/L) [2], average right to left branching over all internal tree nodes (avg R/L). The ratio of the number of right to left branches shows the degree to which the sentence is right branching or not. Additionally, we capture the different types of branching present in a given parse tree identified by the number of nodes in each of its children.

Table 1 depicts the parsing output obtained by CCL for the following sentence from WSJ23 [3]:

Many fund managers argue that now 's the time to buy .

We use Tregex (Levy and Andrew, 2006) for visualizing the output parse trees presented on the left. The bracketing structure statistics and features are given on the right hand side. The root node of each tree structural feature represents the number of times that feature is present in the parsing output of a document.

3 SemEval-14 Results

We develop individual RTM models for each task and subtask that we participate at SemEval-2014 with the RTM-DCU team name. The interpretants are selected from the LM corpora distributed by the translation task of WMT14 (Bojar et al., 2014) and the LM corpora provided by LDC for English (Parker et al., 2011) and Spanish (Ângelo Mendonça et al., 2011) [4]. We use the Stanford

[1] LIX=$\frac{A}{B} + C\frac{100}{A}$, where A is the number of words, C is words longer than 6 characters, B is words that start or end with any of ".", ":", "!", "?" similar to (Hagström, 2012).

[2] For nodes with uneven number of children, the nodes in the odd child contribute to the right branches.

[3] Wall Street Journal (WSJ) corpus section 23, distributed with Penn Treebank version 3 (Marcus et al., 1993).

[4] English Gigaword 5th, Spanish Gigaword 3rd edition.

POS tagger (Toutanova et al., 2003) to obtain the lemmatized corpora for the SRE task. For each RTM model, we extract the features both on the training set and the test set. The number of instances we select for the interpretants in each task is given in Table 2.

Task	Setting	Train	LM
Task 1, SRE	English	770	10770
Task 3, CLSS	Par2S	302	2802
Task 3, CLSS	S2Phrase	202	2702
Task 3, CLSS	Phrase2W	102	2602
Task 10, MSTS	English	504	8002
Task 10, MSTS	English OnWN	504	8004
Task 10, MSTS	Spanish	502	8002

Table 2: Number of sentences in $\mathcal{I}$ (in thousands) selected for each task.

We use ridge regression (RR), support vector regression (SVR) with RBF (radial basis functions) kernel (Smola and Schölkopf, 2004), and extremely randomized trees (TREE) (Geurts et al., 2006) as the learning models. TREE is an ensemble learning method over randomized decision trees. These models learn a regression function using the features to estimate a numerical target value. We also use these learning models after a feature subset selection 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), both of which are described in (Biçici et al., 2013). We optimize the learning parameters, the number of features to select, the number of dimensions used for PLS, and the parameters for parallel FDA5. More detailed descriptions of the optimization processes are given in (Biçici et al., 2013; Biçici et al., 2014). We optimize the learning parameters by selecting ε close to the standard deviation of the noise in the training set (Biçici, 2013) since the optimal value for ε is shown to have linear dependence to the noise level for different noise models (Smola et al., 1998). At testing time, the predictions are bounded to obtain scores in the corresponding ranges. We obtain the confidence scores using support vector classification (SVC).

3.1 Task 1: Semantic Relatedness and Entailment

SRE contains sentence pairs from the SICK (Sentences Involving Compositional Knowledge) data set (Marelli et al., 2014b), which contain sentence pairs that contain rich lexical, syntactic and semantic phenomena. Official evaluation metric in SRE is the Pearson's correlation score, which is used to select the top systems on the training set. SRE task allows the submission of 5 entries. We present the performance of the top 5 individual RTM models on the training set in Table 3. ACC is entailment accuracy, r_P is Pearson's correlation, r_S is Spearman's correlation, MSE is mean squared error, MAE is mean absolute error, and RAE is relative absolute error. R uses the regular lowercased corpora and L uses the lemmatized corpora. R+L correspond to the perspective using the features from both R and L, which doubles the number of features. We compute the entailment by SVC.

Data	Model	ACC	r_P	r_S	RMSE	MAE	RAE
L	SVR	67.52	.7372	.6918	.6946	.5511	.6856
L	PLS-SVR	67.04	.7539	.6927	.6763	.5369	.668
R+L	PLS-SVR	66.76	.75	.6879	.6815	.539	.6705
R+L	SVR	66.66	.7295	.6814	.7027	.5591	.6956
L	PLS-RR	66.56	.7247	.6765	.7054	.5687	.7075

Table 3: SRE training results of the top 5 RTM systems selected.

SRE challenge results on the test set are given in Table 4. The setting R using PLS-SVR learning becomes the 8th out of 17 submissions when predicting the semantic relatedness and 17th out of 18 submissions when predicting the entailment.

Data	Model	ACC	r_P	r_S	RMSE	MAE	RAE
L	PLS-SVR	67.20	.7639	.6877	.655	.5246	.6645
R+L	PLS-SVR	67.65	.7688	.6918	.6492	.5194	.658
L	SVR	67.65	.7559	.6887	.664	.531	.6726
R+L	SVR	67.44	.7625	.6899	.6555	.5251	.6651
R	PLS-SVR	66.61	.7570	.6683	.6637	.5324	.6744

Table 4: RTM-DCU test results on the SRE task.

	Model	r_P	RMSE	MAE	RAE
Par2S	TREE	0.8013	0.8345	0.6277	0.5083
Par2S	PLS-TREE	0.7737	0.8824	0.673	0.5449
Par2S	SVR	0.7718	0.8863	0.6791	0.5499
S2Phrase	TREE	0.6756	0.9887	0.7746	0.6665
S2Phrase	PLS-TREE	0.6119	1.0616	0.8582	0.7384
S2Phrase	SVR	0.6059	1.0662	0.8668	0.7458
Phrase2W	TREE	0.201	1.3275	1.1353	0.9706
Phrase2W	RR	0.1255	1.3463	1.1594	0.9912
Phrase2W	SVR	0.0847	1.3548	1.1663	0.9972

Table 5: CLSS training results of the top 3 RTM systems for each subtask. Levels correspond to paragraph to sentence (Par2S), sentence to phrase (S2Phrase), and phrase to word (Phrase2W).

3.2 Task 3: Cross-Level Semantic Similarity

CLSS contains sentence pairs from different genres including text from newswire, travel, reviews, metaphoric text, community question answering sites, idiomatic text, descriptions, lexicographic text, and search. Official evaluation metric in CLSS is the sum of the Pearson's correlation scores for different levels [5]. CLSS task allows the submission of 3 entries per subtask. We present the performance of the top 3 individual RTM models on the training set in Table 5. RMSE is the root mean squared error. As the compared text size decrease, the performance decrease since it can become harder and more ambiguous to find the similarity using less context. RTM-DCU results on the CLSS challenge test set are provided in Table 6.

	Model	r_P	RMSE	MAE	RAE
Par2S	TREE	.8445	.7417	.5622	.4579
Par2S	PLS-TREE	.7847	.853	.6456	.5258
Par2S	SVR	.7858	.8428	.6539	.5325
S2Phrase	TREE	.75	.8827	.7053	.6255
S2Phrase	PLS-TREE	.6979	.9491	.7781	.69
S2Phrase	SVR	.6631	.9835	.7992	.7088
Phrase2W	TREE	.3053	1.3351	1.14	.9488
Phrase2W	RR	.2207	1.3644	1.1574	.9633
Phrase2W	SVR	.1712	1.3792	1.1792	.9815

Table 6: RTM-DCU test results on CLSS for the top 3 RTM systems for each subtask.

Table 7 lists the results along with their ranks for r_P and r_S, Spearman's correlation, out of 38 submissions. The baseline in Table 7 is normalized longest common substring (LCS) scaled in the range $[0, 4]$. Top individual rank row lists the ranks in each subtask. We present the results for both our official and late (about 1 day) submissions including word to sense (W2S) results [6]. RTM-DCU is able to obtain the top result in Par2S in the CLSS task.

3.3 Task 10: Multilingual Semantic Textual Similarity

MSTS contains sentence pairs from different domains: sense definitions from semantic lexical resources such as OnWN (from OntoNotes (Pradhan et al., 2007) and WordNet (Miller, 1995)) and FNWN (from FrameNet (Baker et al., 1998) and WordNet), news headlines, image descriptions, news title tweet comments, deft forum and news,

[5]Giving advantage to participants submitting to all levels.

[6]W2S results for the late submission is obtained from the LCS baseline to calculate the ranks.

r_P	Par2S	S2Phrase	Phrase2W	W2S	Rank
LCS	0.527	0.562	0.165	0.109	25
Official	0.780	0.677	0.208		14
	0.747	0.588	0.164		19
	0.786	0.666	0.171		18
Late	0.845	0.750	0.305	0.109	6
	0.785	0.698	0.221	0.109	13
	0.786	0.663	0.171	0.109	17
Top Rank	1	5	3		
r_S	Par2S	S2Phrase	Phrase2W	W2S	Rank
LCS	0.527	0.562	0.165	0.13	23
Official	0.780	0.677	0.208		17
	0.747	0.588	0.164		22
	0.786	0.666	0.171		18
Late	0.829	0.734	0.295	0.13	8
	0.778	0.687	0.219	0.13	15
	0.778	0.667	0.166	0.13	16
Top Rank	1	5	5		

Table 7: RTM-DCU test results on CLSS.

paraphrases. Official evaluation metric in MSTS is the Pearson's correlation score.

MSTS task provides 7622 training instances and 3750 test instances. For the OnWN domain, 1316 training instances are available and therefore, we build a separate RTM model for this domain. Separate modeling of the OnWN dataset results with higher confidence scores on the test instances than we would obtain using the overall model to predict. MSTS task allows the submission of 3 entries per subtask. We present the performance of the top 3 individual RTM models on the training set in Table 8.

Lang	Model	r_P	RMSE	MAE	RAE
English	TREE	0.6931	1.0627	0.8058	0.6649
	PLS-TREE	0.6875	1.0753	0.8038	0.6632
	PLS-SVR	0.6884	1.0698	0.8157	0.6730
OnWN	TREE	0.8094	0.9295	0.694	0.5245
	PLS-TREE	0.7953	0.9604	0.7203	0.5444
	PLS-SVR	0.7888	0.9779	0.7234	0.5468
Spanish	TREE	0.6513	0.7341	0.5904	0.7508
	PLS-TREE	0.4157	0.9007	0.7108	0.9039
	PLS-SVR	0.4239	1.1427	0.8293	1.0545

Table 8: MSTS training results on the English, English OnWN, and Spanish tasks.

RTM results on the MSTS challenge test set are provided in Table 9 along with the RTM results in STS 2013 (Biçici and van Genabith, 2013a). Table 10 and Table 11 lists the official results on English and Spanish tasks with rankings calculated according to weighted r_P, which weights according to the number of instances in each domain. RTM-DCU is able to become 10th in the OnWN domain and 19th overall out of 38 submissions in MSTS English and 18th out of 22 submissions in

		Model	r_P	RMSE	MAE	RAE
English	deft-forum	TREE	.4341	1.4306	1.1609	1.0908
		PLS-TREE	.3965	1.4115	1.1472	1.078
		PLS-SVR	.3078	1.6277	1.3482	1.2669
	deft-news	TREE	.6974	1.1469	.9032	.8716
		PLS-TREE	.6811	1.1229	.8769	.8462
		PLS-SVR	.5562	1.2803	.9835	.9491
	headlines	TREE	.6199	1.1495	.9254	.7845
		PLS-TREE	.6125	1.1552	.9314	.7896
		PLS-SVR	.6301	1.1041	.8807	.7467
	images	TREE	.6995	1.2034	.9499	.7395
		PLS-TREE	.6656	1.2298	.9692	.7545
		PLS-SVR	.6474	1.4406	1.1057	.8607
	OnWN	TREE	.8058	1.3122	1.0028	.5585
		PLS-TREE	.7992	1.2997	.9815	.5467
		PLS-SVR	.8004	1.2913	.9449	.5263
	tweet-news	TREE	.6882	.9869	.831	.8093
		PLS-TREE	.6691	1.0101	.8433	.8213
		PLS-SVR	.5531	1.0633	.8653	.8427
Spanish	News	TREE	.7	1.5185	1.351	1.4141
		PLS-TREE	.6253	1.6523	1.4464	1.514
		PLS-SVR	.6411	1.554	1.3196	1.3813
	Wikipedia	TREE	.4216	1.5433	1.298	1.3579
		PLS-TREE	.3689	1.6655	1.4015	1.4662
		PLS-SVR	.4242	1.5998	1.3141	1.3748
STS 2013 English	headlines	L+S SVR	.6552	1.5649	1.2763	1.0231
		L+P+S SVR	.651	1.4845	1.1984	.9607
		L+P+S SVR TL	.6385	1.4878	1.2008	.9626
	OnWN	L+S SVR	.6943	1.7065	1.3545	.8255
		L+P+S SVR	.6971	1.6737	1.333	.8124
		L+P+S SVR TL	.6755	1.7124	1.3598	.8287
	SMT	L+S SVR	.3005	.8833	.6886	1.6132
		L+P+S SVR	.2861	.8810	.6821	1.598
		L+P+S SVR TL	.3098	.8635	.6547	1.5339
	FNWN	L+S SVR	.2016	1.2957	1.0604	1.2633
		L+P+S SVR	.118	1.4369	1.1866	1.4136
		L+P+S SVR TL	.1823	1.3245	1.0962	1.3059

Table 9: RTM-DCU test results on MSTS for the top 3 RTM systems for each subtask as well as RTM results in STS 2013 (Biçici and van Genabith, 2013a).

MSTS Spanish. The performance difference between MSTS English and MSTS Spanish may be due to the fewer training data available for the MSTS Spanish task, which may be decreasing the performance of our supervised learning approach.

3.4 RTMs Across Tasks and Years

We compare the difficulty of tasks according to the RAE levels achieved. RAE measures the error relative to the error when predicting the actual mean. A high RAE is an indicator that the task is hard. In Table 12, we list the RAE obtained for different tasks and subtasks, also listing RTM results in STS 2013 (Biçici and van Genabith, 2013a) and RTM results (Biçici and Way, 2014) on the quality estimation task (QET) (Bojar et al., 2014) where post-editing effort (PEE), human-targeted translation edit rate (HTER), or post-editing time (PET) of translations are predicted.

Model	Wikipedia	News	Weighted r_P	Rank
TREE	0.4216	0.7000	0.5878	18
PLS-TREE	0.3689	0.6253	0.5219	20
PLS-SVR	0.4242	0.6411	0.5537	19

Table 11: RTM-DCU test results on MSTS Spanish task. Rankings are calculated according to the weighted Pearson's correlation.

The best results are obtained for the CLSS Par2S subtask, which may be due to the larger contextual information that paragraphs can provide for the RTM models. For the SRE task, we can only reduce the error with respect to knowing and predicting the mean by about 35%. Prediction of bilingual similarity as in quality estimation of translation can be expected to be harder and RTMs achieve state-of-the-art performance in this task as well (Biçici and Way, 2014).

4 Conclusion

Referential translation machines provide a clean and intuitive computational model for automatically measuring semantic similarity by measuring the acts of translation involved and achieve to be the top on some semantic similarity tasks at SemEval-2014. RTMs make quality and semantic similarity judgments possible based on the retrieval of relevant training data as interpretants for reaching shared semantics.

Acknowledgments

This work is supported in part by SFI (07/CE/I1142) as part of the CNGL Centre for Global Intelligent Content (`www.cngl.org`) at Dublin City University, in part by SFI (13/TIDA/I2740) for the project "Monolingual and Bilingual Text Quality Judgments with Translation Performance Prediction" (`www.computing.dcu.ie/~ebicici/Projects/TIDA_RTM.html`), and in part by the European Commission through the QTLaunchPad FP7 project (No: 296347). We also thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

	Model	deft-forum	deft-news	headlines	images	OnWN	tweet-news	Weighted r_P	Rank
	TREE	.4341	.6974	.6199	.6995	.8058	.6882	.6706	20
	PLS-TREE	.3965	.6811	.6125	.6656	.7992	.6691	.6513	23
	PLS-SVR	.3078	.5562	.6301	.6475	.8004	.5531	.6076	27
	Top Rank	17	16	25	26	16	13		
With Conf.	TREE	.4181	.6846	.6216	.6981	.8331	.6870	.6729	19
	PLS-TREE	.3831	.6739	.6094	.6629	.8260	.6691	.6534	23
	PLS-SVR	.2731	.5526	.6330	.6441	.8246	.5683	.6110	26
	Top Rank	18	18	23	27	10	14		

Table 10: RTM-DCU test results with ranks on MSTS English task.

Task	Subtask	Domain	Model	RAE
SRE	English	SICK	R PLS-SVR	.6645
			R+L PLS-SVR	.6580
			L SVR	.6726
			R+L SVR	.6651
			R PLS-SVR	.6744
CLSS	Par2S	Mixed	TREE	.4579
	S2Phrase		TREE	.6255
	Phrase2W		TREE	.9488
MSTS	English	deft-forum	PLS-TREE	1.078
		deft-news	PLS-TREE	.8462
		headlines	PLS-SVR	.7467
		images	TREE	.7395
		OnWN	PLS-SVR	.5263
		tweet-news	TREE	.8093
	Spanish	News	PLS-SVR	1.3813
		Wikipedia	TREE	1.3579
STS 2013	English	headlines	L+P+S SVR	.9607
		OnWN	L+P+S SVR	.8124
		SMT	L+P+S SVR TL	1.5339
		FNWN	L+S SVR	1.2633
QET PEE	Spanish-English	Europarl	FS-RR	.9000
	Spanish-English	Europarl	PLS-RR	.9409
	English-German	Europarl	PLS-TREE	.8883
	English-German	Europarl	TREE	.8602
	English-Spanish	Europarl	TREE	1.0983
	English-Spanish	Europarl	PLS-TREE	1.0794
	German-English	Europarl	RR	.8204
	German-English	Eurupparl	PLS-RR	.8437
QET HTER	English-Spanish	Europarl	SVR	.8532
	English-Spanish	Europarl	TREE	.8931
QET PET	English-Spanish	Europarl	SVR	.7223
	English-Spanish	Europarl	RR	.7536

Table 12: Best RTM-DCU RAE test results for different tasks and subtasks as well as STS 2013 results (Biçici and van Genabith, 2013a) and results from quality estimation task of translation (Bojar et al., 2014; Biçici and Way, 2014).

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual semantic textual similarity. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland, August.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90.

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proc. of the main conference and the shared task, and Volume 2: Proc. of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 435–440, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Ergun Biçici and Josef van Genabith. 2013a. CNGL-

CORE: Referential translation machines for measuring semantic similarity. In **SEM 2013: The Second Joint Conference on Lexical and Computational Semantics*, pages 234–240, Atlanta, Georgia, USA, 13-14 June. Association for Computational Linguistics.

Ergun Biçici and Josef van Genabith. 2013b. CNGL: Grading student answers by acts of translation. In **SEM 2013: The Second Joint Conference on Lexical and Computational Semantics and Proc. of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 585–591, Atlanta, Georgia, USA, 14-15 June. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Ergun Biçici and Deniz Yuret. 2011a. Instance selection for machine translation using feature decay algorithms. In *Proc. of the Sixth Workshop on Statistical Machine Translation*, pages 272–283, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ergun Biçici and Deniz Yuret. 2011b. RegMT system for machine translation, system combination, and evaluation. In *Proc. of the Sixth Workshop on Statistical Machine Translation*, pages 323–329, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ergun Biçici and Deniz Yuret. 2014. Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions On Audio, Speech, and Language Processing (TASLP)*.

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. 2014. Parallel FDA5 for fast deployment of accurate statistical machine translation systems. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 59–65, Baltimore, USA, June. Association for Computational Linguistics.

Ergun Biçici. 2011. *The Regression Model of Machine Translation*. Ph.D. thesis, Koç University. Supervisor: Deniz Yuret.

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. Association for Computational Linguistics.

Ergun Biçici. 2008. Consensus ontologies in socially interacting multiagent systems. *Journal of Multiagent and Grid Systems*, 4(3):297–314.

Carl Hugo Björnsson. 1968. *Läsbarhet*. Liber.

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

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

José Guilherme Camargo de Souza, Christian Buck, Marco Turchi, and Matteo Negri. 2013. FBK-UEdin participation to the WMT13 quality estimation shared task. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 352–358, Sofia, Bulgaria, August. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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.

Kenth Hagström. 2012. Swedish readability calculator. https://github.com/keha76/Swedish-Readability-Calculator.

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.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proc. of the fifth international conference on Language Resources and Evaluation*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014a. SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland, August.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014b. A sick cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, November.

Preslav Nakov and Torsten Zesch, editors. 2014. *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland, August.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword fifth edition, Linguistic Data Consortium.

Sameer S. Pradhan, Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2007. Ontonotes: a unified relational semantic representation. *Int. J. Semantic Computing*, 1(4):405–419.

Yoav Seginer. 2007. *Learning Syntactic Structure*. Ph.D. thesis, Universiteit van Amsterdam.

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

A. J. Smola, N. Murata, B. Schölkopf, and K.-R. Müller. 1998. Asymptotically optimal choice of ε-loss for support vector machines. In L. Niklasson, M. Boden, and T. Ziemke, editors, *Proc. of the International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, pages 105–110, Berlin. Springer.

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.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wikipedia. 2013. LIX. http://en.wikipedia.org/wiki/LIX.

Ângelo Mendonça, Daniel Jaquette, David Graff, and Denise DiPersio. 2011. Spanish Gigaword third edition, Linguistic Data Consortium.