USAAR-SHEFFIELD: Semantic Textual Similarity with Deep Regression and Machine Translation Evaluation Metrics

Liling Tan[α], Carolina Scarton[β], Lucia Specia[β] and Josef van Genabith[γ]
[α]Universität des Saarlandes / Campus A2.2, Saarbrücken, Germany
[β]University of Sheffield / Regent Court, 211 Portobello, Sheffield, UK
[γ]Deutsches Forschungszentrum für Künstliche Intelligenz / Saarbrücken, Germany
alvations@gmail.com, c.scarton@sheffield.ac.uk,
l.specia@sheffield.ac.uk, josef.van_genabith@dfki.de

Abstract

This paper describes the USAAR-SHEFFIELD systems that participated in the Semantic Textual Similarity (STS) English task of SemEval-2015. We extend the work on using machine translation evaluation metrics in the STS task. Different from previous approaches, we regard the metrics' robustness across different text types and conflate the training data across different subcorpora. In addition, we introduce a novel deep regressor architecture and evaluated its efficiency in the STS task.

1 Introduction

Semantic Textual Similarity (STS) is the task of measuring the degree to which two text snippets have the same meaning (Agirre et al., 2014). For instance, given the two texts, "*a dog sprints across the water*" and "*a dog jumps through water*", participating systems are required to predict a real number similarity score on a scale of 0 (no relation) to 5 (semantic equivalence).

This paper presents a collaborative submission between Saarland University and University of Sheffield to the STS English shared task at SemEval-2015. We have submitted three models that use Machine Translation (MT) evaluation metrics as features to build supervised regressors that predict the similarity scores for the STS task. We introduce two variants of a novel deep regressor architecture and a classical baseline regression system that uses MT evaluation metrics as input features.

2 Related Work

Previously, research teams have applied MT evaluation metrics for the STS task with increasingly better results (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014). Rios et al. (2012) trained a Support Vector Regressor scoring a Pearson correlation mean of 0.3825 (Baseline[1]: 0.4356). Barrón-Cedeño et al. (2013) also used a Support Vector Regressor and did better than the baseline at 0.4037 mean score (Baseline: 0.3639). Huang and Chang (2014) used a linear regressor and scored 0.792 beating the baseline system (Baseline: 0.613).

Another notable mention of MT technology in the STS task is the use of referential translation machines to predict and derive features instead of using MT evaluation metrics (Biçici and van Genabith, 2013; Biçici and Way, 2014).

These previous approaches have trained a different system for each subcorpus provided by the task organizers. We have chosen to combine the different subcorpora since MT evaluation metrics are expected to be robust against text types and domains (Han et al., 2012; Padó et al., 2009).

Much of the previous work on using MT evaluation metrics is based on improving the regressors through algorithm choice, feature selection and parameters tuning. We introduce a novel architecture of hybrid supervised machine learning, *Deep Regression*, which attempts to combine different regressors and automating feature selection by means of dimensionality reduction.

[1]Refers to the token cosine baseline system (`baseline-tokencos`) from the task organizers.

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 85–89,
Denver, Colorado, June 4-5, 2015.

3 Deep Regression Architecture

Ensemble learning constructs a set of models based on different algorithms and then labels new data points by taking a (weighted) vote from the algorithms' predictions (Dietterich, 2000). A typical single layer feed-forward neural network creates a layer of perceptrons that receives inputs and predicts a series of outputs converted by means of an activation function and then the outputs will enter a final layer of a single classifier to provide a final prediction (Auer et al., 2008). We propose a deep regression architecture that is a unique way to combine a single-layer feed-forward neural net architecture with ensemble-like supervised learning.

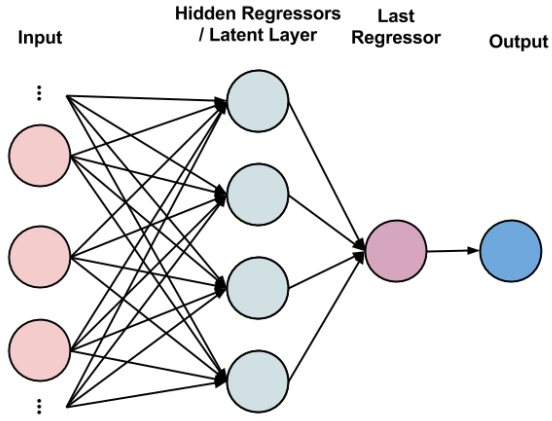

Figure 1: Deep Regression Architecture.

Figure 1 presents the Deep Regression architecture where the inputs are fed into the different hidden regressors and unlike traditional neural network, each regressor produces a discrete output with a different cost function unlike the consistent activation function in neural nets. Different from ensemble learning, the voting/selection determinant has been replaced by a last layer of a single regressor that takes latent layer as input to produce the final output STS score.

By designing the architecture in this way, the feature space from the input is reduced to the number of hidden regressors and the input for the last layer regressors is a latent layer in the higher dimensional space. Within a standard neural net, every node in the latent layer is influenced by all the perceptrons in the previous layer. In contrast, each latent dimension is only dependent on one regressor; in this respect it resembles ensemble learning where the regressors/classifiers are trained independently.

4 Feature Matrix

Machine Translation evaluation metrics consider varying degrees of information at the lexical, syntactic and semantic levels. Each metric comprises several features that compute the translation quality by comparing every translation against one or several reference translations. We consider three sets of features: n-gram overlaps, Shallow Parsing metrics and METEOR. These metrics correspond to the lexical, syntactic and semantic levels respectively.

4.1 N-gram Overlaps

Gonzàlez et al. (2014) reintroduces the notion of language independent metrics relying on n-gram overlaps. This is similar to the BLEU metric that calculates the geometric mean of n-gram precision by comparing the translation against its reference(s) (Papineni et al., 2002) without the brevity penalty.

Different from BLEU, the n-gram overlaps are computed as similarity coefficients instead of taking the crude proportion of overlap n-gram.

$$n\text{-gram}_{overlap} = sim\left(n\text{-gram}_{trans} \cap n\text{-gram}_{ref}\right)$$

We use 16 features of n-gram overlap by considering both the cosine similarity and Jaccard Index in calculating the n-gram overlaps for character and token n-gram from the order of bigrams to 5-grams. In addition, we use the ratio of n-gram lengths and the Jaccard similarity of pseudo-cognates (Simard et al., 1992) as the 17th and 18th n-gram overlap features.

4.2 Shallow Parsing

The Shallow Parsing (SP) metric measures the syntactic similarities by computing the overlaps between the translation and the reference translation at the Parts-Of-Speech (POS), word lemmas and base phrase chunks level. The purpose of the SP metric is to capture the proportion of lexical items correctly translated according to their shallow syntactic realization.

The base phrase chunks are tagged using the BIOS toolkit (Surdeanu et al., 2005) and POS tag-

ging and lemmatization are achieved using SVM-Tool (Giménez and Màrquez, 2004). For instance, given a pair of sentences in the format (word/POS/lemma/chunk):

- ***NP***(*a*/DT/*a*/B-NP *dog*/NN/*dog*/I-NP) *sprints*/VBZ/*sprint*/B-VP *across*/IN/*across*/O ***NP***(*the*/DET/*the*/B-NP *water*/NN/*water*/I-NP)

- ***NP***(*a*/DT/*a*/B-NP *dog*/NN/*dog*/I-NP) *jumps*/VBZ/*jump*/B-VP *through*/IN/*through*/O *water*/NN/*water*/B-NP

We consider the overlap proportions for the POS features, lemma, IOB features, shallow chunks. The Inside, Outside, Begin (IOB) features refer to the shallow parsing tags at the lexical level, e.g. `B-NP` represents the beginning of a noun phrase (Sang et al., 2000). The IOB features are measured lexically by considering each IOB tag while the shallow chunk features only consider the number of bracketed chunks.

For instance, the POS tag DT occurs twice in first sentence one and once in second sentence, thus we extract the feature `SP-POS`(DT) = 1/2 = 0.5.

- `SP-POS` (DT,NN,VBZ,IN) = [0.5,1,1,1]
- `SP-LEMMA` (*a,dog,jump,through,water*) = [1,1,0,0,1]
- `SP-IOB` (B-NP,I-NP,B-VP,O) = [1,1,-0.5,1,1]
- `SP-CHUNK` (NP) = [0.5]

For `SP-POS`, `SP-LEMMA` and `SP-IOB`, we use the NIST-like measure where we not only consider the individual POS, LEMMA or IOB tags but an accumulated score over a sequence of 1-5 n-grams, e.g. `SP-POS` (DT+NN,DT+NN+VBZ, ...) or `SP-LEMMA` (*a+dog,a+dog+jump, ...*) .

5 METEOR

METEOR aligns the translation to a reference translation first then it uses unigram mapping to match words at their surface forms, word stems, synonym matches and paraphrase matches (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010).

Different from the n-gram and shallow parsing features, METEOR makes a distinction between content words and function words and the precision and recall is measured by weighing them differently. It also accounts for word order differences by penalizing chunks from the translation that do not appear in the translation.

We use the METEOR 1.5 system with tuned weights and penalty using the WMT12 data. For the STS experiment, we use all four variants of METEOR: exact matches, stem matches, synonym matches and paraphrase matches.

6 Experiments and Results

6.1 Training Data

We conflated all training and test data of various text types from previous SemEval STS shared tasks into a single training set with 10597 paragraph/sentence/caption pairs. The MT metrics for each text pair were computed with the Asiya toolkit (Giménez and Màrquez, 2010). Tokenization and preprocessing operations, such as lemmatization, POS tagging, parsing and n-gram extraction, are performed by the Asiya toolkit.

6.2 Models

We submitted three models to the SemEval-2015 STS English Task:

- `ModelX`: Deep Regression framework with the full feature set from n-gram overlaps, Shallow Parsing and METEOR.

- `ModelY`: Bayesian Ridge Regressor with the full feature set

- `ModelZ`: Deep Regression framework with only METEOR features

For the hidden regressors layer of the deep regression models, we have used the multivariate linear, logistic, Bayesian ridge, elastic net, random sample consensus and support vector (radial basis function kernel) regressors.[2] The final layer regressor is a Bayesian ridge regressor. These supervised regressors are implemented in `scikit-learn` (Pedregosa et al., 2011).

[2]No comprehensive parameter tuning was attempted on the models and the default parameters for each regressor can be found on our code repository, https://github.com/alvations/USAAR-SemEval-2015.

	Ans-Forums	Ans-Student	Belief	Headlines	Images	Mean	Rank
ModelX	0.3706	0.3609	0.4767	0.5183	0.5436	0.4616	68
ModelY	0.6264	0.7386	0.705	0.7927	0.8162	**0.7275**	**21**
ModelZ	0.4237	0.6757	0.6994	0.5239	0.6833	0.6111	58

Table 1: Spearman's Results for STS English Task @ SemEval-2015.

6.3 Results

Table 1 presents the official results for the English STS task where our baseline model (ModelY) strikingly outperforms the deep regressor models (ModelX and ModelZ).

Our baseline model achieved modest results ranking 24 out of 73 submissions, however our deep regressors have failed to function on par with a simple baseline regressor. We note that the deep regressor with the full feature set (ModelX) scored lower than the deep regressor with only the METEOR features (ModelZ). This reiterates the effectiveness of semantically motivated METEOR features in determining similarity as previously indicated by Huang and Chang (2014).

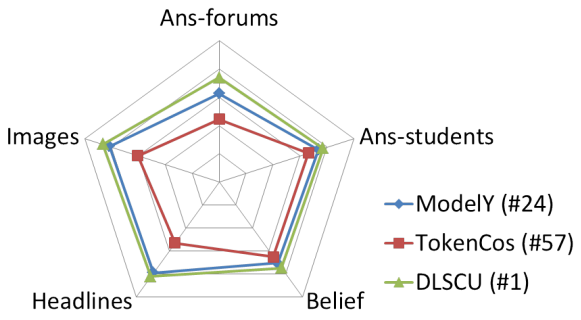

Figure 2: Comparison of Results with Best and Baseline Systems

Interestingly, the conflation of datasets has no obvious detrimental effects on the performance for any specific domains. Figure 2 presents a comparison of results between ModelY, the top system from DLSU and the organizers' baseline system (TokenCos). It shows that the distribution of Spearman's correlation for our model is as well-balanced as the best system.

7 Conclusion

In this paper, we have described our submissions to the STS English task for SemEval-2015. We have introduced a novel deep regression infrastructure with MT evaluation metrics to measure semantic similarity. Although our deep regressors performed poorly, our baseline system have achieved promising results amongst the participating systems and we showed that conflating datasets of different genres has negligible effects on a semantic similarity system based on MT evaluation metrics.

The results also confirm the good performance of METEOR, a traditional MT evaluation metric, for the STS task.

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM): Proceedings of the Main Conference and the Shared Task*, pages 385–393, Montréal, Canada.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 32–43, Atlanta, Georgia.

Eneko Agirre, Carmen Banea, Claire Cardic, 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 *Proceedings of the 8th Inter-*

national Workshop on Semantic Evaluation (SemEval 2014), pages 81–91, Dublin, Ireland.

Peter Auer, Harald Burgsteiner, and Wolfgang Maass. 2008. A learning rule for very simple universal approximators consisting of a single layer of perceptrons. *Neural Networks*, 21(5):786–795.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Alberto Barrón-Cedeño, Lluís Màrquez, Maria Fuentes, Horacio Rodríguez, and Jordi Turmo. 2013. UPC-CORE: What Can Machine Translation Evaluation Metrics and Wikipedia Do for Estimating Semantic Textual Similarity? In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 143–147, Atlanta, Georgia.

Ergun Biçici and Josef van Genabith. 2013. CNGL-CORE: Referential Translation Machines for Measuring Semantic Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 234–240, Atlanta, Georgia.

Ergun Biçici and Andy Way. 2014. RTM-DCU: Referential Translation Machines for Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 487–496, Dublin, Ireland.

Michael Denkowski and Alon Lavie. 2010. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. In *Proceedings of the HLT: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 250–253, Los Angeles, California.

Thomas G. Dietterich. 2000. Ensemble Methods in Machine Learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pages 1–15.

Jesús Giménez and Lluís Màrquez. 2004. Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. *Recent Advances in Natural Language Processing III*, pages 153–162.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Meritxell Gonzàlez, , Alberto Barrón-Cedeo, and Llus Màrquez. 2014. Ipa and stout: Leveraging linguistic and source-based features for machine translation evaluation. In *Ninth Workshop on Statistical Machine Translation*, page 8.

Aaron L.F. Han, Derek F. Wong, and Lidia S. Chao. 2012. Lepor: A robust evaluation metric for machine translation with augmented factors. In *24th International Conference on Computational Linguistics*, page 441. Citeseer.

Pingping Huang and Baobao Chang. 2014. SSMT:A Machine Translation Evaluation View To Paragraph-to-Sentence Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 585–589, Dublin, Ireland.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Chris Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 297–305.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Miguel Rios, Wilker Aziz, and Lucia Specia. 2012. UOW: Semantically Informed Text Similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM): Proceedings of the Main Conference and the Shared Task*, pages 673–678, Montréal, Canada.

Tjong Kim Sang, Erik F., and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, ConLL '00, pages 127–132, Stroudsburg, PA, USA.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Forth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, Canada.

Mihai Surdeanu, Jordi Turmo, and Eli Comelles. 2005. Named Entity Recognition from Spontaneous Open-Domain Speech. In *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech)*, Lisbon, Portugal.