DLS@CU at SemEval-2016 Task 1: Supervised Models of Sentence Similarity

Md Arafat Sultan[†] **Steven Bethard**[‡] **Tamara Sumner**[†]
[†]Institute of Cognitive Science and Department of Computer Science,
University of Colorado Boulder
[‡]Department of Computer and Information Sciences,
University of Alabama at Birmingham,
arafat.sultan@colorado.edu, bethard@uab.edu, sumner@colorado.edu

Abstract

We describe a set of systems submitted to the SemEval-2016 English Semantic Textual Similarity (STS) task. Given two English sentences, the task is to compute the degree of their semantic similarity. Each of our systems uses the SemEval 2012–2015 STS datasets to train a ridge regression model that combines different measures of similarity. Our best system demonstrates 73.6% correlation with average human annotations across five test sets.

1 Introduction

Identification of short-text semantic similarity is an important research problem with application in a multitude of NLP tasks: question answering (Yao et al., 2013; Severyn and Moschitti, 2013), short answer grading (Mohler et al., 2011; Ramachandran et al., 2015), text summarization (Dasgupta et al., 2013; Wang et al., 2013), evaluation of machine translation (Chan and Ng, 2008; Liu et al., 2011), and so on. The SemEval Semantic Textual Similarity (STS) task series (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015) is a core platform for the task: a publicly available corpus of more than 14,000 sentence pairs have been developed over a span of four years with human annotations of similarity for each pair; and about 300 system runs have been evaluated.

In this article, we describe a set of systems that participated in the SemEval-2016 English Semantic Textual Similarity (STS) task. Given two English sentences, the objective is to compute their semantic similarity in the range [0, 5], where the score increases with similarity (i.e., 0 indicates no similarity and 5 indicates identical meanings). The official evaluation metric is the Pearson product-moment correlation coefficient with human annotations. Our systems leverage different measures of sentence similarity and train ridge regression models that learn to combine predictions from these different sources using past SemEval data. The best of our three system runs achieves 73.6% with human annotations among all submitted systems on five test sets (containing a total of 1186 test pairs).

Early work in sentence similarity (Mihalcea et al., 2006; Li et al., 2006; Islam and Inkpen, 2008) established the basic procedural framework under which most modern algorithms operate: computing sentence similarity as a mean of word similarities across the two input sentences. With no human annotated STS dataset available, these algorithms are unsupervised and were evaluated extrinsically on tasks like paraphrase detection and textual entailment recognition. The SemEval STS task series has made an important contribution through the large annotated dataset, enabling intrinsic evaluation of STS systems and making supervised STS systems a reality.

At SemEval 2012–2015, most of the top-performing STS systems used a regression algorithm to combine different measures of similarity (Bär et al., 2012; Šarić et al., 2012; Wu et al., 2013; Lynum et al., 2014; Sultan et al., 2015), with the notable exception of a couple of unsupervised systems that relied primarily on alignment of related words in the two sentences (Han et al., 2013; Sultan et al., 2014b).

Our models are based on the successful linear re-

Proceedings of SemEval-2016, pages 668–673,
San Diego, California, June 16-17, 2016.

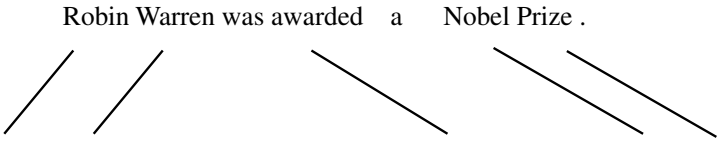

Figure 1: Words aligned by our aligner across two sentences taken from the MSR alignment corpus (Brockett, 2007). (We show only part of the second sentence.) Besides exact word/lemma matches, it identifies and aligns semantically similar word pairs using PPDB (`awarded` ↔ `received` in this example).

gression architecture of past SemEval systems in general, and the winning system of SemEval-2015 (Sultan et al., 2015) in particular. We use the features in the latter system unchanged in one of our runs and augment them with simple word and character n-gram overlap features in the other two runs.

2 System Description

Our system employs a ridge regression model (linear regression with L_2 error and L_2 regularization) to combine a set of similarity measures. The model is trained on SemEval 2012–2015 data. Our three runs differ in the subset of features drawn from the feature pool. We describe the feature set in this section; the individual runs will be discussed in Section 4.

2.1 Features

Word Alignment Proportion. This feature operationalizes the hypothesis that highly semantically similar sentences should also have a high degree of conceptual alignment between their semantic units, i.e., words and phrases. To that end, we apply the monolingual word aligner developed by Sultan et al. (2014a) to input sentence pairs.[1]

This aligner aligns words based on their semantic similarity and the similarity between their local semantic contexts in the two sentences. It uses the paraphrase database PPDB (Ganitkevitch et al., 2013) to identify semantically similar words, and relies on dependencies and surface-form neighbors of the two words to determine their contextual similarity. Word pairs are aligned in decreasing order of a weighted sum of their semantic and contextual similarity. Figure 1 shows an example set of alignments. For more details, see (Sultan et al., 2014a).

Additionally, we also consider a Levenshtein distance[2] of 1 between a misspelled word and a correctly spelled word (of length > 2) to be a match.

Given sentences $S^{(1)}$ and $S^{(2)}$, the alignment-based similarity measure is computed as follows:

$$sim(S^{(1)}, S^{(2)}) = \frac{n_c^a(S^{(1)}) + n_c^a(S^{(2)})}{n_c(S^{(1)}) + n_c(S^{(2)})}$$

where $n_c(S^{(i)})$ and $n_c^a(S^{(i)})$ are the number of content words and the number of aligned content words in $S^{(i)}$, respectively.

Sentence Embedding. A fundamental limitation of the above feature is that it only relies on PPDB to identify semantically similar words; consequently, similar word pairs are limited to only lexical paraphrases. Hence it fails to utilize semantic similarity or relatedness between non-paraphrase word pairs (e.g., `sister` and `related`). In the current feature, we leverage neural word embeddings to overcome this limitation. We use the 400-dimensional vectors[3] developed by Baroni et al. (2014). They used the word2vec toolkit[4] to extract these vectors from a corpus of about 2.8 billion tokens. These vectors perform well across different word similarity datasets in their experiments. Details on their approach and findings can be found in (Baroni et al., 2014).

Instead of comparing word vectors across the two input sentences, we adopt a simple vector composition scheme to construct a vector representation of

[1] `https://github.com/ma-sultan/monolingual-word-aligner`

[2] The minimum number of single-character edits needed to change one word into the other, where an edit is an insertion, a deletion or a substitution.

[3] `http://clic.cimec.unitn.it/composes/semantic-vectors.html`

[4] `https://code.google.com/p/word2vec/`

Dataset	Source of Text	# of Pairs
answer-answer	Q&A forums	254
headlines	news headlines	249
plagiarism	plagiarised answers	230
postediting	post-edited MT pairs	244
question-question	Q&A forums	209

Table 1: Test sets at SemEval STS 2016.

each input sentence and then take the cosine similarity between the two sentence vectors as our second feature for this run. The vector representing a sentence is simply the sum of its content lemma vectors.

Word n-gram Overlap. This feature computes the proportion of word n-grams (lemmatized) that are in both $S^{(1)}$ and $S^{(2)}$. We employ separate instances of this feature for $n = 1, 2, 3$. The goal is to identify high local similarities in the two snippets and learn the influence that such local similarities might have on human judgment of sentence similarity.

Character n-gram Overlap. This feature computes the proportion of character n-grams that are in both $S^{(1)}$ and $S^{(2)}$ in their surface form. We employ separate instances of this feature for $n = 3, 4$. The goal is to identify and correct for spelling errors as well as incorrect lemmatizations.

Soft Cardinality. Soft Cardinality (Jimenez et al., 2012) is a measure of set cardinality where similar items in a set contribute less to its cardinality than dissimilar items. Jimenez et al. (2012) propose a parameterized measure of semantic similarity based on soft cardinality that computes sentence similarity from word similarity and the latter from character n-gram similarity. This measure was highly successful at SemEval-2012 (Agirre et al., 2012). We employ this measure with untuned parameter values as a feature for our model: $p = 1$, $bias = 0$, $\alpha = 0.5$, $bias_{sim} = 0$, $\alpha_{sim} = 0.5$, $q_1 = 2$, and $q_2 = 4$. (Please see the original article for a detailed description of these parameters as well as the similarity measure.)

3 Data

The 1186 test sentence pairs at SemEval-2016 are divided into five sets, each consisting of pairs from a particular domain. Each pair is assigned a similarity score in the range [0, 5] by human annotators (0: no similarity, 5: identicality). Test sets are discussed briefly in Table 1.

We train our supervised systems using data from the past four years of SemEval STS (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015). The selections vary by test set, which we discuss in the next section.

4 Runs

We submit three runs at SemEval-2016. Each run employs a ridge regression model; we use Scikit-learn (Pedregosa et al., 2011) for model implementation. Different training data from SemEval 2012–2015 are used for different test sets. For *headlines*, we train the model on *headlines* (2013, 2014, 2015), *deft-news* (2014), *tweet-news* (2014), and *smtnews* (2012) pairs. For *postediting*, the model is trained on *smt* (2013), *smteuroparl* (2012) and *smtnews* (2012) pairs. These selections are based on the similarity between the source and the target domains—news data for *headlines*, machine translation data for *postediting*. For the other three test sets, all past annotations (except those for *fnwn* (2013)) are used, as we did not find any close matches for these test sets in the SemEval 2012–2015 data.

4.1 Run 1

Run 1 is a ridge regression model based only on the first two features—alignment and embeddings. The regularization strength parameter α is set using cross-validation on training data.

4.2 Run 2

Run 2 employs a model similar to the run 1 model, but uses the entire feature set described in Section 2.1. The same training sets are used for each test set and the model parameter α is again set using cross-validation on training data.

4.3 Run 3

Run 3 is identical to run 2, except that it assigns a lower value to the regularization parameter α (100 as opposed to 500 in run 2).

5 Evaluation

Table 2 shows the performances of the three runs (measured by Pearson's r, the official evaluation metric at SemEval STS) alongside the score for the

Dataset	Runs			Best
	1	2	3	Score
answer-answer	.5523	.5599	.5453	.6924
headlines	.8008	.8033	.8033	.8275
plagiarism	.8229	.8123	.8195	.8414
postediting	.8426	.8442	.8442	.8669
question-question	.6599	.6423	.6666	.7471
Weighted Mean	.7356	.7330	.7355	.7781

Table 2: Performance on STS 2016 data. Each number in rows 1–5 is the correlation (Pearson's r) between system output and human annotations for the corresponding test set. The rightmost column shows the best score by any system. The last row shows the value of the final evaluation metric for each run as well as the top-performing system.

best performing system for each test set. The last row shows the official evaluation metric that computes a weighted sum of correlations over all test sets, where the weight of a test set is proportional to the number of sentence pairs it contains.

Runs 1 and 3 have very similar overall performances, slightly better than that of run 2. Among the different test sets, the models perform well on news headlines, plagiarism and machine translation data, but poorly on the Q&A forums data.

5.1 Ablation Study

From the overall performances in Table 2, it is clear that the three new features added to the Sultan et al. (2015) model do not improve performance. Therefore, we run a feature ablation study only on the run 1 model. Table 3 shows the results. Similar to the findings reported in (Sultan et al., 2015), the alignment-based feature performs better across test sets. However, the addition of the embedding feature improves performance on almost all test sets.

5.2 Relation between the Runs

We compute pairwise correlations between the predictions of our three runs to see how different they are. As Table 4 shows, the predictions are highly correlated, which is expected given the results in Table 2.

6 Conclusions and Future Work

We present three supervised models of sentence similarity based on the winning system at SemEval-2015 (Sultan et al., 2015). Our additional features

Data Set	Run 1	Alignment	Embedding
answer-answer	.5523	.5196	.4972
headlines	.8008	.8013	.7240
plagiarism	.8229	.8149	.7813
postediting	.8426	.8226	.8214
question-question	.6599	.5767	.6833
Weighted Mean	.7356	.7084	.6994

Table 3: Performance of each individual feature of our best run (run 1) on STS 2016 test sets. Combining the two features improves performance on most test sets.

Runs	Pearson's r
1, 2	.9834
1, 3	.9952
2, 3	.9920

Table 4: Pairwise correlations between the three runs.

do not improve performance and results show similar influences of alignment and embedding features as in SemEval-2015.

Besides high performance, the run 1 model has the key advantage of simplicity and high replicability. All the major design components are also available for free download (links provided in Section 2).

A key limitation of the system is its inability to model semantics of units larger than words (phrasal verbs, idioms, and so on). This is an important future direction not only for our system but also for STS and text comparison tasks in general. Incorporation of stop word semantics is key to identifying similarities and differences in subtle aspects of sentential semantics like polarity and modality. Domain-specific learning of the word vectors can also improve results.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, 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 '13, pages 32–43, Atlanta, Georgia, USA.

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 *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 81–91, Dublin, Ireland.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 252–263, Denver, Colorado.

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. In *Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 435–440, Montréal, Canada.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 238–247, Baltimore, Maryland.

Chris Brockett. 2007. Aligning the RTE 2006 Corpus. Technical Report MSR-TR-2007-77, Microsoft Research.

Yee Seng Chan and Hwee Tou Ng. 2008. MAXSIM: A Maximum Similarity Metric for Machine Translation Evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62, Columbus, Ohio.

Anirban Dasgupta, Ravi Kumar, and Sujith Ravi. 2013. Summarization Through Submodularity and Dispersion. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 1014–1022, Sofia, Bulgaria.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, NAACL '13, pages 758–764, Atlanta, Georgia, USA.

Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Second Joint Conference on Lexical and Computational Semantics*, *SEM '13, pages 44–52, Atlanta, Georgia, USA.

Aminul Islam and Diana Inkpen. 2008. Semantic Text Similarity Using Corpus-based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2):10:1–10:25.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012. Soft Cardinality: A Parameterized Similarity Function for Text Comparison. In *Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 449–453, Montréal, Canada.

Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. 2006. Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2011. Better Evaluation Metrics Lead to Better Machine Translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 375–384, Edinburgh, Scotland, UK.

André Lynum, Partha Pakray, Björn Gambäck, and Sergio Jimenez. 2014. NTNU: Measuring Semantic Similarity with Sublexical Feature Representations and Soft Cardinality. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 448–453, Dublin, Ireland.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence*, AAAI '06, pages 775–780.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to Grade Short Answer Questions Using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACL '11, pages 752–762, Portland, Oregon, USA.

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.

Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. 2015. Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, NAACL-BEA '15, pages 97–106.

Aliaksei Severyn and Alessandro Moschitti. 2013. Automatic Feature Engineering for Answer Selection and Extraction. In *Proceedings of the 2013 Conference*

on Empirical Methods in Natural Language Processing, EMNLP '13, pages 458–467, Seattle, Washington, USA.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014a. Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014b. DLS@CU: Sentence Similarity from Word Alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 241–246, Dublin, Ireland.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 148–153, Denver, Colorado, USA.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. TakeLab: Systems for Measuring Semantic Text Similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 441–448, Montréal, Canada.

Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie. 2013. A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 1384–1394, Sofia, Bulgaria.

Stephen Wu, Dongqing Zhu, Ben Carterette, and Hongfang Liu. 2013. MayoClinicNLP-CORE: Semantic Representations for Textual Similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, *SEM '13, pages 148–154, Atlanta, Georgia, USA.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Answer Extraction as Sequence Tagging with Tree Edit Distance. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, NAACL '13, pages 858–867, Atlanta, Georgia, USA.