DLS@CU: Sentence Similarity from Word Alignment and Semantic Vector Composition

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@cis.uab.edu, sumner@colorado.edu

Abstract

We describe a set of top-performing systems at the SemEval 2015 English Semantic Textual Similarity (STS) task. Given two English sentences, each system outputs the degree of their semantic similarity. Our unsupervised system, which is based on word alignments across the two input sentences, ranked 5th among 73 submitted system runs with a mean correlation of 79.19% with human annotations. We also submitted two runs of a supervised system which uses word alignments and similarities between compositional sentence vectors as its features. Our best supervised run ranked 1st with a mean correlation of 80.15%.

1 Introduction

Identification of short text similarity is an important research problem with application in a multitude of areas: natural language processing (machine translation, text summarization), information retrieval (question answering), education (short answer scoring), 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) has become a central platform for the task: a publicly available corpus of more than 14,000 sentence pairs have been developed over the past four years with human annotations of similarity for each pair; and a total of 290 system runs have been evaluated.

In this article, we describe a set of systems that were submitted at the SemEval 2015 English STS task (Agirre et al., 2015). 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 identicality). The official evaluation metric was the Pearson correlation coefficient with human annotations. The best of our three system runs achieved the highest mean correlation (80.15%) with human annotations among all submitted systems on five test sets (containing a total of 3000 test pairs).

Early work on sentence similarity (Corley and Mihalcea, 2005; 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 data set available, these algorithms were 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 data set, enabling intrinsic evaluation of STS systems and making supervised STS systems a reality.

At SemEval 2012, domain-specific training data was provided for most of the test pairs (Agirre et al., 2012) and consequently, supervised systems were the most successful (Bär et al., 2012; Šarić et al., 2012). These systems combined different similarity measures, e.g., lexico-semantic, syntactic and string similarity, using regression models. However, at the 2013 and 2014 STS events, no such training data was provided; instead, the systems were allowed to use all past data to train their systems. Interestingly, the best systems at these two events were unsupervised (Han et al., 2013; Sultan et al., 2014b); some super-

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 148–153,
Denver, Colorado, June 4-5, 2015.

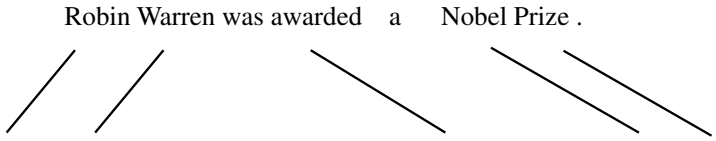

Figure 1: Words aligned by our aligner across two sentences taken from the MSR alignment corpus (Brockett, 2007). (We show only part of the second sentence.) Besides exact word/lemma matches, it identifies and aligns semantically similar word pairs using PPDB (*awarded – received* in this example).

vised systems did well, too (Wu et al., 2013; Lynum et al., 2014). The core component of a typical unsupervised system is term alignment: semantically related terms across the two sentences are aligned at first and then their semantic similarity is computed as a monotonically increasing function of the degree of alignment.

At SemEval 2015, we submitted an unsupervised system based on word alignments which is almost identical to our winning system at SemEval 2014 (Sultan et al., 2014b). We also submitted a supervised ridge regression model that uses (1) the output of our unsupervised system, and (2) the cosine similarity between the vector representations of the two sentences (derived from neural word embeddings of their content words (Baroni et al., 2014)) as its features. Our unsupervised system ranked 5th and the two supervised runs ranked 1st and 3rd. Evaluation also shows that our best run outperforms the winning systems at all past SemEval STS events.

2 System Overview

We describe our three system runs in this section in order of their complexity – new capabilities and/or features are added with each run.

2.1 Run 1: U

This is an unsupervised system that first aligns related words across the two input sentences and then outputs the proportion of aligned content words as their semantic similarity. It is similar to our last year's system (Sultan et al., 2014b) based on the word aligner described in (Sultan et al., 2014a). However, where last year's system computed a separate proportion for each sentence and then took their harmonic mean, this year's system computes a single proportion over all words in the two sentences. In other words, given sentences $S^{(1)}$ and $S^{(2)}$,

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

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. This is a conceptually simpler step and yielded better experimental results on data from past STS events.

The 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).

We also consider a levenshtein distance[1] of 1 between a misspelled word and a correctly spelled word (of length > 2) to be a match. In all runs, we truncate at the extremes to keep the score in [0, 5].

2.2 Run 2: S_1

A fundamental limitation of our unsupervised system 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

[1] 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.

'related'). In this run, we leverage neural word embeddings to overcome this limitation. We use the 400-dimensional vectors[2] developed by Baroni et al. (2014). They used the word2vec toolkit[3] to extract these vectors from a corpus of about 2.8 billion tokens. These vectors performed exceedingly well across different word similarity data sets 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 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 the centroid (i.e., the componentwise average) of its content lemma vectors.

Finally, we combine the two features – output of our unsupervised run ($\boldsymbol{U}$) and the sentence vectors' cosine similarity – using a ridge regression model (implemented in scikit-learn (Pedregosa et al., 2011), with $\alpha = 1.0$ and the 'auto' solver that automatically selects a feature weight learning algorithm from a pool depending on the type of the data). The model is trained using annotations from SemEval 2012–2014 (details in Section 3).

2.3 Run 3: S_2

The aligner used in our previous two runs aligns content words even if there are no similarities between their contexts in the two sentences. In this run, we use an alignment-based feature (in addition to our two features in $\boldsymbol{S_1}$) where content words are aligned only if they have some contextual similarity – a common word either in their dependencies or in a neighborhood of 3 words to the left and 3 words to the right (considering only content words for the latter).

3 Data

The 3000 test sentence pairs at SemEval 2015 were divided into five sets, each consisting of pairs from a different domain. Each pair was assigned similarity scores in the range [0, 5] by multiple human annotators (0: no similarity, 5: identicality) and the average of the annotations was taken as their final similarity score. We describe each data set briefly in Table 1.

Data Set	Source of Text	# of Pairs
answers-forums	forum answers	375
answers-students	student short answers	750
belief	belief annotations	375
headlines	news headlines	750
images	image descriptions	750

Table 1: Test sets at SemEval STS 2015.

We trained our supervised systems using data from the past three years of SemEval STS (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014). For *answers-forums*, *answers-students* and *belief*, we used all past annotations. For *headlines*, we used all *headlines* (2013), *headlines* (2014), *deft-news* (2014) and *smtnews* (2012) pairs. For *images*, we used all *msrpar* (2012; train and test), *msrvid* (2012; train and test) and *images* (2014) pairs. The specific training corpus selections for the two latter data sets were based on our experiments with past *headlines* and *images* data, where these subsets yielded better results than an all-inclusive training set (seemingly due to the fact that they were drawn from similar domains and were still large-enough to provide the model with effective supervision).

4 Evaluation

In addition to the official evaluation at SemEval 2015, we report evaluation results on past STS (2012–2014) test data. For all these evaluations, the performance metric is the Pearson correlation coefficient between system output and average human annotations. Correlation is computed for each individual test set, and a weighted sum of all correlations (i.e. over all test sets) is used as the final evaluation metric. The weight of a test set is proportional to the number of sentence pairs it contains.

Before presenting the results, we describe a preprocessing step for one of the 2015 test sets. Identifying the right stop words (some of which can be domain-specific) proved key in our past investigation of STS (Sultan et al., 2014b); therefore we consider it very important to manually examine individual domains to ensure proper categorization of words. An inspection of the trial data for the *answers-students* set indicated that the expressions in the

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

[3] https://code.google.com/p/word2vec/

Data Set	Runs			Best Score
	$\boldsymbol{U}$	$\boldsymbol{S_1}$	$\boldsymbol{S_2}$	
answers-forums	0.6821	***0.7390***	0.7241	**0.7390**
answers-students	***0.7879***	0.7725	0.7569	**0.7879**
belief	0.7325	*0.7491*	0.7223	**0.7717**
headlines	0.8238	*0.8250*	*0.8250*	**0.8417**
images	0.8485	*0.8644*	0.8631	**0.8713**
Weighted Mean	0.7919	***0.8015***	0.7921	-
Rank	5	***1***	3	-

Table 2: Performance on STS 2015 data. Each number in rows 1–5 is the correlation between system output and human annotations for the corresponding data set. The rightmost column shows the best score by any system. The last two rows show the value of the final evaluation metric and the system rank, respectively, for each run.

following pairs are semantically equivalent for the given domain: {'battery terminal', 'terminal'} and {'electrical state', 'state'}. Therefore, we treated the two words 'battery' and 'electrical' as special stop words during occurrences of these pairs across the input sentences.

4.1 STS 2015 Results

Performances of our three runs on each of the STS 2015 test sets are shown in Table 2. Each bold number represents the best score by any system on the corresponding test set and each italic number shows the best score among our runs. The weighted mean of correlations and rank for each run is also shown.

Our best run ($\boldsymbol{S_1}$) did not perform the best on all test sets (in fact it does so on only one test set), but it maintained the best balance across all test sets. The second best overall system run (ExBThemis-themisexp) had a mean correlation of 79.42%. We found the difference of 0.73% between this system and $\boldsymbol{S_1}$ to be statistically significant at $p < 0.0001$ in a two-sample one-tailed z-test[4] (unlike last year's 0.05% (Agirre et al., 2014)).

The third feature in S_2 did not prove useful as S_2 performed worse than S_1 on almost all test sets. This result falls in line with our observation reported in (Sultan et al., 2014a): "more often than not content words are inherently sufficiently meaningful to be aligned even in the absence of contextual evidence when there are no competing pairs."

[4]Standard deviation was computed from the frequency distribution of correlations across the five test sets.

Year	$\boldsymbol{S_1}$	Winning System
2014	**0.779**	0.761
2013	**0.6542**	0.6181
2012	**0.6803**	0.6773

Table 3: Performance of our top system ($\boldsymbol{S_1}$) on past STS test sets (mean correlation with human annotations). The score of the winning system at each event is shown on column 3. $\boldsymbol{S_1}$ outperforms all past winning systems.

Contrary to our findings from past years' data, the special stop words for the *answers-students* test set (discussed in the previous section) did not improve performance – considering these words as content words, we observed a slightly higher correlation of 0.7895 for our unsupervised system $\boldsymbol{U}$.

4.2 Results on Past Test Sets

Table 3 shows the performance of our best system $\boldsymbol{S_1}$ on test data from SemEval 2012–2014. To ensure fair comparison with other systems, for years 2013 and 2014, we used only past data to train our model. For year 2012, we used the designated training data for test sets *msrpar*, *msrvid* and *smteuroparl*, and all 2012 training pairs for the other two test sets.

$\boldsymbol{S_1}$ outperformed all winning systems from 2012 through 2014. Without any domain-specific training data, the top systems at SemEval 2013 and 2014 were unsupervised. $\boldsymbol{S_1}$ achieved the best performance on both despite its supervised nature.

4.3 Ablation Study

We performed a feature ablation study for $\boldsymbol{S_1}$ on STS 2015 data to determine the relative importances of its two features. Table 4 shows the results. Columns 2 and 4 show performances of our $\boldsymbol{U}$ and $\boldsymbol{S_1}$ systems. (Remember that the former is used as a feature by the latter.) Column 3 shows the performance of the second feature of $\boldsymbol{S_1}$ (i.e. cosine similarity between the sentence vectors) as a measure of STS.

On four of the five test sets, $\boldsymbol{U}$ outperformed sentence vector similarity. However, combining the two features improved system performance on four out of five test sets, and overall. These results indicate that each feature captures aspects of STS that the other does not and consequently the two complement each other when used together.

Data Set	U	Vector Sim	S_1
answers-forums	0.6821	0.7330	**0.7390**
answers-students	**0.7879**	0.6899	0.7725
belief	0.7325	0.6981	**0.7491**
headlines	0.8238	0.7511	**0.8250**
images	0.8485	0.8411	**0.8644**
Weighted Mean	0.7919	0.7494	**0.8015**

Table 4: Performance of each individual feature of our best run (S_1) on STS 2015 test sets. Combining the two features improves performance on all but one test set.

5 Conclusions and Future Work

At SemEval 2014, we reported a top-performing unsupervised STS system (Sultan et al., 2014b) that relied only on word alignment. This year, we present a supervised system that is statistically significantly better than our last year's system. Combining a vector similarity feature derived from word embeddings with alignment-based similarity, it outperforms all past and current STS systems. Since it makes use of only off-the-shelf software[5] and data, it is easily replicable as well.

The primary limitation of our system is the 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. Finally, rather than studying STS as a standalone problem, the time has come to develop algorithms that can adapt to requirements posed by different data domains and applications.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant Numbers EHR/0835393 and EHR/0835381. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[5]Our aligner is also available at: `https://github.com/ma-sultan/monolingual-word-aligner`

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 385-393, Montreal, Canada.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 Shared Task: Semantic Textual Similarity. In *Proceedings of the 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, Iñigo 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, Denver, Colorado, USA.

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 6th International Workshop on Semantic Evaluation, held in conjunction with the 1st Joint Conference on Lexical and Computational Semantics*, SemEval '12, pages 435-440, Montreal, Canada.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski 2014. Dont 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, USA.

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

Courtney Corley and Rada Mihalcea. 2005. Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13-18, Ann Arbor, Michigan, USA.

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.* Association for Computational Linguistics, 758-764.

Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the 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 (TKDD)*, 2(2):10:1-10:25.

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

André Lynum, Partha Pakray, Björn Gambäck 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*, pages 775-780, Boston, Massachusetts, 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*, vol. 12, pages 2825-2830.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. TakeLab: Systems for Measuring Semantic Text Similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation, held in conjunction with the 1st Joint Conference on Lexical and Computational Semantics*, SemEval '12, pages 441-448, Montreal, Canada.

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 (May), pages 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.

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.