DLS@CU: Sentence Similarity from Word Alignment

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 present an algorithm for computing the semantic similarity between two sentences. It adopts the hypothesis that semantic similarity is a monotonically increasing function of the degree to which (1) the two sentences contain similar semantic units, and (2) such units occur in similar semantic contexts. With a simplistic operationalization of the notion of semantic units with individual words, we experimentally show that this hypothesis can lead to state-of-the-art results for sentence-level semantic similarity. At the SemEval 2014 STS task (task 10), our system demonstrated the best performance (measured by correlation with human annotations) among 38 system runs.

1 Introduction

Semantic textual similarity (STS), in the context of short text fragments, has drawn considerable attention in recent times. Its application spans a multitude of areas, including natural language processing, information retrieval and digital learning. Examples of tasks that benefit from STS include text summarization, machine translation, question answering, short answer scoring, and so on.

The annual series of SemEval STS tasks (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014) is an important platform where STS systems are evaluated on common data and evaluation criteria. In this article, we describe an STS system which participated and outperformed all other systems at SemEval 2014.

The algorithm is a straightforward application of the monolingual word aligner presented in (Sultan et al., 2014). This aligner aligns related words in two sentences based on the following properties of the words:

1. They are semantically similar.
2. They occur in similar semantic contexts in the respective sentences.

The output of the word aligner for a sentence pair can be used to predict the pair's semantic similarity by taking the proportion of their aligned content words. Intuitively, the more semantic components in the sentences we can meaningfully align, the higher their semantic similarity should be. In experiments on STS 2013 data reported by Sultan et al. (2014), this approach was found highly effective. We also adopt this hypothesis of semantic compositionality for STS 2014.

We implement an STS algorithm that is only slightly different from the algorithm in (Sultan et al., 2014). The approach remains equally successful on STS 2014 data.

2 Background

We focus on two relevant topics in this section: the state of the art of STS research, and the word aligner presented in (Sultan et al., 2014).

2.1 Semantic Textual Similarity

Since the inception of textual similarity research for short text, perhaps with the studies reported by Mihalcea et al. (2006) and Li et al. (2006), the topic has spawned significant research interest. The majority of systems have been reported as part of the SemEval 2012 and *SEM 2013 STS tasks (Agirre et al., 2012; Agirre et al., 2013). Here we confine our discussion to systems that participated in these tasks.

With designated training data for several test sets, supervised systems were the most successful in STS 2012 (Bär et al., 2012; Šarić et al., 2012;

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 241–246,
Dublin, Ireland, August 23-24, 2014.

Jimenez et al., 2012). Such systems typically apply a regression algorithm on a large number of STS features (e.g., string similarity, syntactic similarity and word or phrase-level semantic similarity) to generate a final similarity score. This approach continued to do well in 2013 (Han et al., 2013; Wu et al., 2013; Shareghi and Bergler, 2013) even without domain-specific training data, but the best results were demonstrated by an unsupervised system (Han et al., 2013). This has important implications for STS since extraction of each feature adds to the latency of a supervised system. STS systems are typically important in the context of a larger system rather than on their own, so high latency is an obvious drawback for such systems.

We present an STS system that has simplicity, high accuracy and speed as its design goals, can be deployed without any supervision, operates in a linguistically principled manner with purely semantic sentence properties, and was the top system at SemEval STS 2014.

2.2 The Sultan et al. (2014) Aligner

The word aligner presented in (Sultan et al., 2014) has been used unchanged in this work and plays a central role in our STS algorithm. We give only an overview here; for the full details, see the original article.

We will denote the sentences being aligned (and are subsequently input to the STS algorithm) as $S^{(1)}$ and $S^{(2)}$. We describe only content word alignment here; stop words are not used in our STS computation.

The aligner first identifies word pairs $w_i^{(1)} \in S^{(1)}$ and $w_j^{(2)} \in S^{(2)}$ such that:

1. $w_i^{(1)}$ and $w_j^{(2)}$ have non-zero semantic similarity, sim_{Wij}. The calculation of sim_{Wij} is described in Section 2.2.1.
2. The semantic contexts of $w_i^{(1)}$ and $w_j^{(2)}$ have some similarity, sim_{Cij}. We define the semantic context of a word w in a sentence S as a set of words in S, and the semantic context of the word pair $(w_i^{(1)}, w_j^{(2)})$, denoted by $context_{ij}$, as the Cartesian product of the context of $w_i^{(1)}$ in $S^{(1)}$ and the context of $w_j^{(2)}$ in $S^{(2)}$. We define several types of context (i.e., several selections of words) and describe the corresponding calculations of sim_{Cij} in Section 2.2.2.
3. There are no competing pairs scoring higher than $(w_i^{(1)}, w_j^{(2)})$ under $f(sim_W, sim_C) = 0.9 \times sim_W + 0.1 \times sim_C$. That is, there are no pairs $(w_k^{(1)}, w_j^{(2)})$ such that $f(sim_{Wkj}, sim_{Ckj}) > f(sim_{Wij}, sim_{Cij})$, and there are no pairs $(w_i^{(1)}, w_l^{(2)})$ such that $f(sim_{Wil}, sim_{Cil}) > f(sim_{Wij}, sim_{Cij})$. The weights 0.9 and 0.1 were derived empirically via a grid search in the range $[0, 1]$ (with a step size of 0.1) to maximize alignment performance on the training set of the (Brockett, 2007) alignment corpus. This set contains 800 human-aligned sentence pairs collected from a textual entailment corpus (Bar-Haim et al., 2006).

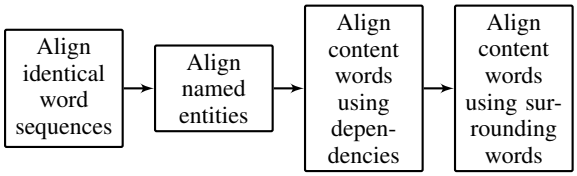

Figure 1: The alignment pipeline.

The aligner then performs one-to-one word alignments in decreasing order of the f value.

This alignment process is applied in four steps as shown in Figure 1; each step applies the above process to a particular type of context: identical words, dependencies and surrounding content words. Additionally, named entities are aligned in a separate step (details in Section 2.2.2).

Words that are aligned by an earlier module of the pipeline are not allowed to be re-aligned by downstream modules.

2.2.1 Word Similarity

Word similarity (sim_W) is computed as follows:

1. If the two words or their lemmas are identical, then $sim_W = 1$.
2. If the two words are present as a pair in the lexical XXXL corpus of the Paraphrase Database[1] (PPDB) (Ganitkevitch et al., 2013), then $sim_W = 0.9$.[2] For this step, PPDB was augmented with lemmatized forms of the already existing word pairs.[3]

[1] PPDB is a large database of lexical, phrasal and syntactic paraphrases.

[2] Again, the value 0.9 was derived empirically via a grid search in [0, 1] (step size = 0.1) to maximize alignment performance on the (Brockett, 2007) training data.

[3] The Python NLTK WordNetLemmatizer was used to lemmatize the original PPDB words.

3. For any other word pair, $sim_W = 0$.

2.2.2 Contextual Similarity

Contextual similarity (sim_C) for a word pair $(w_i^{(1)}, w_j^{(2)})$ is computed as the sum of the word similarities for each pair of words in the context of $(w_i^{(1)}, w_j^{(2)})$. That is:

$$sim_{Cij} = \sum_{(w_k^{(1)}, w_l^{(2)}) \in context_{ij}} sim_{Wkl}$$

Each of the stages in Figure 1 employs a specific type of context.

Identical Word Sequences. Contextual similarity for identical word sequences (a word sequence W which is present in both $S^{(1)}$ and $S^{(2)}$ and contains at least one content word) defines the context by pairing up each word in the instance of W in $S^{(1)}$ with its occurrence in the instance of W in $S^{(2)}$. All such sequences with length ≥ 2 are aligned; longer sequences are aligned before shorter ones. This simple step was found to be of very high precision in (Sultan et al., 2014) and reduces the overall computational cost of alignment.

Named Entities. Named entities are a special case in the alignment pipeline. Even though the context for a named entity is defined in the same way as it is defined for any other content word (as described below), named entities are aligned in a separate step before other content words because they have special properties such as coreferring mentions of different lengths (e.g. Smith and John Smith, BBC and British Broadcasting Corporation). The head word of the named entity is used in dependency calculations.

Dependencies. Dependency-based contextual similarity defines the context for the pair $(w_i^{(1)}, w_j^{(2)})$ using the syntactic dependencies of $w_i^{(1)}$ and $w_j^{(2)}$. The context is the set of all word pairs $(w_k^{(1)}, w_l^{(2)})$ such that:

- $w_k^{(1)}$ is a dependency of $w_i^{(1)}$,
- $w_l^{(2)}$ is a dependency of $w_j^{(2)}$,
- $w_i^{(1)}$ and $w_j^{(2)}$ have the same lexical category,
- $w_k^{(1)}$ and $w_l^{(2)}$ have the same lexical category, and,
- The two dependencies are either identical or semantically "equivalent" according to the equivalence table provided by Sultan et al. (2014). We explain semantic equivalence of dependencies using an example below.

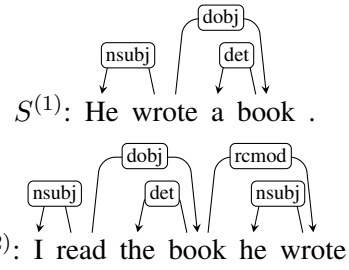

Figure 2: Example of dependency equivalence.

Equivalence of Dependency Structures. Consider $S^{(1)}$ and $S^{(2)}$ in Figure 2. Note that $w_2^{(1)} = w_6^{(2)}$ = 'wrote' and $w_4^{(1)} = w_4^{(2)}$ = 'book' in this pair. Now, each of the two following typed dependencies: *dobj*$(w_2^{(1)}, w_4^{(1)})$ in $S^{(1)}$ and *rcmod*$(w_4^{(2)}, w_6^{(2)})$ in $S^{(2)}$, represents the relation "thing that was written" between the verb 'wrote' and its argument 'book'. Thus, to summarize, an instance of contextual evidence for a possible alignment between the pair $(w_2^{(1)}, w_6^{(2)})$ ('wrote') lies in the pair $(w_4^{(1)}, w_4^{(2)})$ ('book') and the equivalence of the two dependency types *dobj* and *rcmod*.

The equivalence table of Sultan et al. (2014) is a list of all such possible equivalences among different dependency types (given that $w_i^{(1)}$ has the same lexical category as $w_j^{(2)}$ and $w_k^{(1)}$ has the same lexical category as $w_l^{(2)}$).

If there are no word pairs with identical or equivalent dependencies as defined above, i.e. if $sim_{Cij} = 0$, then $w_i^{(1)}$ and $w_j^{(2)}$ will not be aligned by this module.

Surrounding Content Words. Surrounding-word-based contextual similarity defines the context of a word in a sentence as a fixed window of 3 words to its left and 3 words to its right. Only content words in the window are considered. (As explained in the beginning of this section, the context of the pair $(w_i^{(1)}, w_j^{(2)})$ is then the Cartesian product of the context of $w_i^{(1)}$ in $S^{(1)}$ and $w_j^{(2)}$ in $S^{(2)}$.) Note that $w_i^{(1)}$ and $w_j^{(2)}$ can be of different lexical categories here.

A content word can often be surrounded by stop words which provide almost no information about its semantic context. The chosen window size is assumed, on average, to effectively make

Data Set	Source of Text	# of Pairs
deft-forum	discussion forums	450
deft-news	news articles	300
headlines	news headlines	750
images	image descriptions	750
OnWN	word sense definitions	750
tweet-news	news articles and tweets	750

Table 1: Test sets for SemEval STS 2014.

sufficient contextual information available while avoiding the inclusion of contextually unrelated words. But further experiments are necessary to determine the best span in the context of alignment.

Unlike dependency-based alignment, even if there are no similar words in the context, i.e. if $sim_{Cij} = 0$, $w_i^{(1)}$ may still be aligned to $w_j^{(2)}$ if $sim_{Wij} > 0$ and no alignments for $w_i^{(1)}$ or $w_j^{(2)}$ have been found by earlier stages of the pipeline.

2.2.3 The Alignment Sequence

The rationale behind the specific sequence of alignment steps (Figure 1) was explained in (Sultan et al., 2014): (1) Identical word sequence alignment was found to be the step with the highest precision in experiments on the (Brockett, 2007) training data, (2) It is convenient to align named entities before other content words to enable alignment of entity mentions of different lengths, (3) Dependency-based evidence was observed to be more reliable (i.e. of higher precision) than textual evidence on the (Brockett, 2007) training data.

3 Method

Our STS score is a function of the proportions of aligned content words in the two input sentences.

The proportion of content words in $S^{(1)}$ that are aligned to some word in $S^{(2)}$ is:

$$prop_{Al}^{(1)} = \frac{|\{i : [\exists j : (i,j) \in Al] \text{ and } w_i^{(1)} \in C\}|}{|\{i : w_i^{(1)} \in C\}|}$$

where C is the set of all content words in English and Al are the predicted word alignments. A word alignment is a pair of indices (i, j) indicating that word $w_i^{(1)}$ is aligned to $w_j^{(2)}$. The proportion of aligned content words in $S^{(2)}$, $prop_{Al}^{(2)}$, can be computed in a similar way.

We posit that a simple yet sensible way to obtain an STS estimate for $S^{(1)}$ and $S^{(2)}$ is to take a mean of $prop_{Al}^{(1)}$ and $prop_{Al}^{(2)}$. Our two submitted runs use the harmonic mean:

$$sim(S^{(1)}, S^{(2)}) = \frac{2 \times prop_{Al}^{(1)} \times prop_{Al}^{(2)}}{prop_{Al}^{(1)} + prop_{Al}^{(2)}}$$

It is a more conservative estimate than the arithmetic mean, and penalizes sentence pairs with a large disparity between the values of $prop_{Al}^{(1)}$ and $prop_{Al}^{(2)}$. Experiments on STS 2012 and 2013 data revealed the harmonic mean of the two proportions to be a better STS estimate than the arithmetic mean.

Data Set	Run 1	Run 2
deft-forum	0.4828	0.4828
deft-news	0.7657	0.7657
headlines	0.7646	0.7646
images	0.8214	0.8214
OnWN	0.7227	0.8589
tweet-news	0.7639	0.7639
Weighted Mean	0.7337	0.7610

Table 2: Results of evaluation on SemEval STS 2014 data. Each value on columns 2 and 3 is the correlation between system output and human annotations for the corresponding data set. The last row shows the value of the final evaluation metric.

4 Data

STS systems at SemEval 2014 were evaluated on six data sets. Each test set consists of a number of sentence pairs; each pair has a human-assigned similarity score in the range [0, 5] which increases with similarity. Every score is the mean of five scores crowdsourced using the Amazon Mechanical Turk. The sentences were collected from a variety of sources. In Table 1, we provide a brief description of each test set.

5 Evaluation

We submitted the results of two system runs at SemEval 2014 based on the idea presented in Section 3. The two runs were identical, except for the fact that for the *OnWN* test set, we specified the following words as additional stop words during run 2 (but not during run 1): *something*, *someone*, *somebody*, *act*, *activity*, *some*, *state*.[4] For both

[4] *OnWN* has many sentence pairs where each sentence is of the form "the act/activity/state of verb+ing something/somebody". The selected words act merely as fillers in such pairs and consequently do not typically contribute to the similarity scores.

Data Set	Run 1	Run 2
FNWN	0.4686	0.4686
headlines	0.7797	0.7797
OnWN	0.6083	0.8197
SMT	0.3837	0.3837
Weighted Mean	0.5788	0.6315

Table 3: Results of evaluation on *SEM STS 2013 data.

Data Set	Run 1	Run 2
MSRpar	0.6413	0.6413
MSRvid	0.8200	0.8200
OnWN	0.7227	0.7004
SMTeuroparl	0.4267	0.4267
SMTnews	0.4486	0.4486
Weighted Mean	0.6476	0.6423

Table 4: Results of evaluation on SemEval STS 2012 data.

runs, the *tweet-news* sentences were preprocessed by separating the hashtag from the word for each hashtagged word.

Table 2 shows the performance of each run. Rows 1 through 6 show the Pearson correlation coefficients between the system scores and human annotations for all test sets. The last row shows the value of the final evaluation metric, which is a weighted sum of all correlations in rows 1–6. The weight assigned to a data set is proportional to its number of pairs. Our run 1 ranked 7th and run 2 ranked 1st among 38 submitted system runs.

An important implication of these results is the fact that knowledge of domain-specific stop words can be beneficial for an STS system. Even though we imparted this knowledge to our system during run 2 via a manually constructed set of additional stop words, simple measures like TF-IDF can be used to automate the process.

5.1 Performance on STS 2012 and 2013 Data

We applied our algorithm on the 2012 and 2013 STS test sets to examine its general utility. Note that the STS 2013 setup was similar to STS 2014 with no domain-dependent training data, whereas several of the 2012 test sets had designated training data.

Over all the 2013 test sets, our two runs demonstrated weighted correlations of 0.5788 (rank: 4) and 0.6315 (rank: 1), respectively. Table 3 shows performances on individual test sets. (Descriptions of the test sets can be found in (Agirre et al., 2013).) Again, run 2 outperformed run 1 on *OnWN* by a large margin.

On the 2012 test sets, however, the performance was worse (relative to other systems), with respective weighted correlations of 0.6476 (rank: 8) and 0.6423 (rank: 9). Table 4 shows performances on individual test sets. (Descriptions of the test sets can be found in (Agirre et al., 2012).)

This performance drop seems to be an obvious consequence of the fact that our algorithm was not trained on domain-specific data: on STS 2013 data, the top two STS 2012 systems, with respective weighted correlations of 0.5652 and 0.5221 (Agirre et al., 2013), were outperformed by our system by a large margin.

In contrast to the other two years, our run 1 outperformed run 2 on the 2012 *OnWN* test set by a very small margin. A closer inspection revealed that the previously mentioned sentence structure "the act/activity/state of verb+ing something/somebody" is much less common in this set, and as a result, our additional stop words tend to play more salient semantic roles in this set than in the other two *OnWN* sets (i.e. they act relatively more as content words than stop words). The drop in correlation with human annotations is a consequence of this role reversal. This result again shows the importance of a proper selection of stop words for STS and also points to the challenges associated with making such a selection.

6 Conclusions and Future Work

We show that alignment of related words in two sentences, if carried out in a principled and accurate manner, can yield state-of-the-art results for sentence-level semantic similarity. Our system has the desired quality of being both accurate and fast. Evaluation on test data from different STS years demonstrates its general applicability as well.

The idea of STS from alignment is worth investigating with larger semantic units (i.e. phrases) in the two sentences. Another possible research direction is to investigate whether the alignment proportions observed for the two sentences can be used as features to improve performance in a supervised setup (even though this scenario is arguably less common in practice because of unavailability of domain or situation-specific training data).

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.

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, Dublin, Ireland.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

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.

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

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*, NAACL-HLT '13, pages 758-764, Atlanta, Georgia, USA.

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.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012. Soft Cardinality: a parameterized similarity function for text comparison. 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 449-453, Montreal, 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, vol.18, no.8.* 1138-1150.

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, Boston, Massachusetts, 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 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.

Ehsan Shareghi and Sabine Bergler. 2013. CLaC-CORE: Exhaustive Feature Combination for Measuring Textual Similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, *SEM '13, pages 202-206, Atlanta, Georgia, USA.

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

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.