SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity

Eneko Agirre
University of the Basque Country
Donostia, 20018, Basque Country
e.agirre@ehu.es

Daniel Cer
Stanford University
Stanford, CA 94305, USA
danielcer@stanford.edu

Mona Diab
Center for Computational Learning Systems
Columbia University
mdiab@ccls.columbia.edu

Aitor Gonzalez-Agirre
University of the Basque Country
Donostia, 20018, Basque Country
agonzalez278@ikasle.ehu.es

Abstract

Semantic Textual Similarity (STS) measures the degree of semantic equivalence between two texts. This paper presents the results of the STS pilot task in Semeval. The training data contained 2000 sentence pairs from previously existing paraphrase datasets and machine translation evaluation resources. The test data also comprised 2000 sentences pairs for those datasets, plus two surprise datasets with 400 pairs from a different machine translation evaluation corpus and 750 pairs from a lexical resource mapping exercise. The similarity of pairs of sentences was rated on a 0-5 scale (low to high similarity) by human judges using Amazon Mechanical Turk, with high Pearson correlation scores, around 90%. 35 teams participated in the task, submitting 88 runs. The best results scored a Pearson correlation >80%, well above a simple lexical baseline that only scored a 31% correlation. This pilot task opens an exciting way ahead, although there are still open issues, specially the evaluation metric.

1 Introduction

Semantic Textual Similarity (STS) measures the degree of semantic equivalence between two sentences. STS is related to both Textual Entailment (TE) and Paraphrase (PARA). STS is more directly applicable in a number of NLP tasks than TE and PARA such as Machine Translation and evaluation, Summarization, Machine Reading, Deep Question Answering, etc. STS differs from TE in as much as it assumes symmetric graded equivalence between the pair of textual snippets. In the case of TE the equivalence is directional, e.g. a car is a vehicle, but a vehicle is not necessarily a car. Additionally, STS differs from both TE and PARA in that, rather than being a binary yes/no decision (e.g. a vehicle is not a car), STS incorporates the notion of graded semantic similarity (e.g. a vehicle and a car are more similar than a wave and a car).

STS provides a unified framework that allows for an extrinsic evaluation of multiple semantic components that otherwise have tended to be evaluated independently and without broad characterization of their impact on NLP applications. Such components include word sense disambiguation and induction, lexical substitution, semantic role labeling, multiword expression detection and handling, anaphora and coreference resolution, time and date resolution, named-entity handling, underspecification, hedging, semantic scoping and discourse analysis. Though not in the scope of the current pilot task, we plan to explore building an open source toolkit for integrating and applying diverse linguistic analysis modules to the STS task.

While the characterization of STS is still preliminary, we observed that there was no comparable existing dataset extensively annotated for pairwise semantic sentence similarity. We approached the construction of the first STS dataset with the following goals: (1) To set a definition of STS as a graded notion which can be easily communicated to non-expert annotators beyond the likert-scale; (2) To gather a substantial amount of sentence pairs from diverse datasets, and to annotate them with high quality; (3) To explore evaluation measures for STS; (4) To explore the relation of STS to PARA and Machine Translation Evaluation exercises.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 385–393,
Montréal, Canada, June 7-8, 2012.

In the next section we present the various sources of the STS data and the annotation procedure used. Section 4 investigates the evaluation of STS systems. Section 5 summarizes the resources and tools used by participant systems. Finally, Section 6 draws the conclusions.

2 Source Datasets

Datasets for STS are scarce. Existing datasets include (Li et al., 2006) and (Lee et al., 2005). The first dataset includes 65 sentence pairs which correspond to the dictionary definitions for the 65 word pairs in Similarity(Rubenstein and Goodenough, 1965). The authors asked human informants to assess the meaning of the sentence pairs on a scale from 0.0 (minimum similarity) to 4.0 (maximum similarity). While the dataset is very relevant to STS, it is too small to train, develop and test typical machine learning based systems. The second dataset comprises 50 documents on news, ranging from 51 to 126 words. Subjects were asked to judge the similarity of document pairs on a five-point scale (with 1.0 indicating "highly unrelated" and 5.0 indicating "highly related"). This second dataset comprises a larger number of document pairs, but it goes beyond sentence similarity into textual similarity.

When constructing our datasets, gathering naturally occurring pairs of sentences with different degrees of semantic equivalence was a challenge in itself. If we took pairs of sentences at random, the vast majority of them would be totally unrelated, and only a very small fragment would show some sort of semantic equivalence. Accordingly, we investigated reusing a collection of existing datasets from tasks that are related to STS.

We first studied the pairs of text from the Recognizing TE challenge. The first editions of the challenge included pairs of sentences as the following:

> T: The Christian Science Monitor named a US journalist kidnapped in Iraq as freelancer Jill Carroll.
> H: Jill Carroll was abducted in Iraq.

The first sentence is the text, and the second is the hypothesis. The organizers of the challenge annotated several pairs with a binary tag, indicating whether the hypothesis could be entailed from the text. Although these pairs of text are interesting we decided to discard them from this pilot because the length of the hypothesis was typically much shorter than the text, and we did not want to bias the STS task in this respect. We may, however, explore using TE pairs for STS in the future.

Microsoft Research (MSR) has pioneered the acquisition of paraphrases with two manually annotated datasets. The first, called MSR Paraphrase (MSRpar for short) has been widely used to evaluate text similarity algorithms. It contains 5801 pairs of sentences gleaned over a period of 18 months from thousands of news sources on the web (Dolan et al., 2004). 67% of the pairs were tagged as paraphrases. The inter annotator agreement is between 82% and 84%. Complete meaning equivalence is not required, and the annotation guidelines allowed for some relaxation. The pairs which were annotated as not being paraphrases ranged from completely unrelated semantically, to partially overlapping, to those that were almost-but-not-quite semantically equivalent. In this sense our graded annotations enrich the dataset with more nuanced tags, as we will see in the following section. We followed the original split of 70% for training and 30% for testing. A sample pair from the dataset follows:

> The Senate Select Committee on Intelligence is preparing a blistering report on prewar intelligence on Iraq.
>
> American intelligence leading up to the war on Iraq will be criticized by a powerful US Congressional committee due to report soon, officials said today.

In order to construct a dataset which would reflect a uniform distribution of similarity ranges, we sampled the MSRpar dataset at certain ranks of string similarity. We used the implementation readily accessible at CPAN[1] of a well-known metric (Ukkonen, 1985). We sampled equal numbers of pairs from five bands of similarity in the [0.4 .. 0.8] range separately from the paraphrase and non-paraphrase pairs. We sampled 1500 pairs overall, which we split 50% for training and 50% for testing.

The second dataset from MSR is the MSR Video Paraphrase Corpus (MSRvid for short). The authors showed brief video segments to Annotators from Amazon Mechanical Turk (AMT) and were asked

[1] `http://search.cpan.org/~mlehmann/String-Similarity-1.04/Similarity.pm`

- A person is slicing a cucumber into pieces.
- A chef is slicing a vegetable.
- A person is slicing a cucumber.
- A woman is slicing vegetables.
- A woman is slicing a cucumber.
- A person is slicing cucumber with a knife.
- A person cuts up a piece of cucumber.
- A man is slicing cucumber.
- A man cutting zucchini.

Figure 1: Video and corresponding descriptions from MSRvid

Compare Two Similar Sentences

Score how similar two sentences are to each other according to the following scale.

The sentences are:

(5) Completely equivalent, as they *mean the same thing*.
(4) Mostly equivalent, but some *unimportant details differ*.
(3) Roughly equivalent, but some *important information differs/missing*.
(2) Not equivalent, but *share some details*.
(1) Not equivalent, but are *on the same topic*.
(0) On different topics.

Select a similarity rating for each sentence pair below:

Figure 2: Definition and instructions for annotation

to provide a one-sentence description of the main action or event in the video (Chen and Dolan, 2011). Nearly 120 thousand sentences were collected for 2000 videos. The sentences can be taken to be roughly parallel descriptions, and they included sentences for many languages. Figure 1 shows a video and corresponding descriptions.

The sampling procedure from this dataset is similar to that for MSRpar. We construct two bags of data to draw samples. The first includes all possible pairs for the same video, and the second includes pairs taken from different videos. Note that not all sentences from the same video were equivalent, as some descriptions were contradictory or unrelated. Conversely, not all sentences coming from different videos were necessarily unrelated, as many videos were on similar topics. We took an equal number of samples from each of these two sets, in an attempt to provide a balanced dataset between equivalent and non-equivalent pairs. The sampling was also done according to string similarity, but in four bands in the [0.5 .. 0.8] range, as sentences from the same video had a usually higher string similarity than those in the MSRpar dataset. We sampled 1500 pairs overall, which we split 50% for training and 50% for testing.

Given the strong connection between STS systems and Machine Translation evaluation metrics, we also sampled pairs of segments that had been part of human evaluation exercises. Those pairs included a reference translation and a automatic Machine Translation system submission, as follows:

> The only instance in which no tax is levied is when the supplier is in a non-EU country and the recipient is in a Member State of the EU.
>
> The only case for which no tax is still perceived "is an example of supply in the European Community from a third country.

We selected pairs from the translation shared task of the 2007 and 2008 ACL Workshops on Statistical Machine Translation (WMT) (Callison-Burch et al., 2007; Callison-Burch et al., 2008). For consistency, we only used French to English system submissions. The training data includes all of the Europarl human ranked fr-en system submissions from WMT 2007, with each machine translation being paired with the correct reference translation. This resulted in 729 unique training pairs. The test data is comprised of all Europarl human evaluated fr-en pairs from WMT 2008 that contain 16 white space delimited tokens or less.

In addition, we selected two other datasets that were used as out-of-domain testing. One of them comprised of all the human ranked fr-en system submissions from the WMT 2007 news conversation test set, resulting in 351 unique system reference pairs.[2] The second set is radically different as it comprised 750 pairs of glosses from OntoNotes 4.0 (Hovy et al., 2006) and WordNet 3.1 (Fellbaum, 1998) senses. The mapping of the senses of both resources comprised 110K sense pairs. The similarity between the sense pairs was generated using simple word overlap. 50% of the pairs were sampled from senses which were deemed as equivalent senses, the rest from senses which did not map to one another.

3 Annotation

In this first dataset we defined a straightforward likert scale ranging from 5 to 0, but we decided to provide definitions for each value in the scale (cf. Figure 2). We first did pilot annotations of 200 pairs se-

[2] At the time of the shared task, this data set contained duplicates resulting in 399 sentence pairs.

lected at random from the three main datasets in the training set. We did the annotation, and the pairwise Pearson ranged from 84% to 87% among ourselves. The agreement of each annotator with the average scores of the other was between 87% and 89%.

In the future, we would like to explore whether the definitions improve the consistency of the tagging with respect to a likert scale without definitions. Note also that in the assessment of the quality and evaluation of the systems performances, we just took the resulting SS scores and their averages. Using the qualitative descriptions for each score in analysis and evaluation is left for future work.

Given the good results of the pilot we decided to deploy the task in Amazon Mechanical Turk (AMT) in order to crowd source the annotation task. The turkers were required to have achieved a 95% of approval rating in their previous HITs, and had to pass a qualification task which included 6 example pairs. Each HIT included 5 pairs of sentences, and was paid at 0.20$ each. We collected 5 annotations per HIT. In the latest data collection, each HIT required 114.9 second for completion.

In order to ensure the quality, we also performed post-hoc validation. Each HIT contained one pair from our pilot. After the tagging was completed we checked the correlation of each individual turker with our scores, and removed annotations of turkers which had low correlations (below 50%). Given the high quality of the annotations among the turkers, we could alternatively use the correlation between the turkers itself to detect poor quality annotators.

4 Systems Evaluation

Given two sentences, s1 and s2, an STS system would need to return a similarity score. Participants can also provide a confidence score indicating their confidence level for the result returned for each pair, but this confidence is not used for the main results. The output of the systems performance is evaluated using the Pearson product-moment correlation coefficient between the system scores and the human scores, as customary in text similarity (Rubenstein and Goodenough, 1965). We calculated Pearson for each evaluation dataset separately.

In order to have a single Pearson measure for each system we concatenated the gold standard (and system outputs) for all 5 datasets into a single gold standard file (and single system output). The first version of the results were published using this method, but the overall score did not correspond well to the individual scores in the datasets, and participants proposed two additional evaluation metrics, both of them based on Pearson correlation. The organizers of the task decided that it was more informative, and on the benefit of the community, to also adopt those evaluation metrics, and the idea of having a single main evaluation metric was dropped. This decision was not without controversy, but the organizers gave more priority to openness and inclusiveness and to the involvement of participants. The final result table thus included three evaluation metrics. For the future we plan to analyze the evaluation metrics, including non-parametric metrics like Spearman.

4.1 Evaluation metrics

The first evaluation metric is the Pearson correlation for the concatenation of all five datasets, as described above. We will use *overall Pearson* or simply *ALL* to refer to this measure.

The second evaluation metric normalizes the output for each dataset separately, using the linear least squares method. We concatenated the system results for five datasets and then computed a single Pearson correlation. Given $Y = \{y_i\}$ and $X = \{x_i\}$ (the gold standard scores and the system scores, respectively), we transform the system scores into $X' = \{x'_i\}$ in order to minimize the squared error $\sum_i (y_i - x'_i)^2$. The linear transformation is given by $x'_i = x_i * \beta_1 + \beta_2$, where β_1 and β_2 are found analytically. We refer to this measure as *Normalized Pearson* or simply *ALLnorm*. This metric was suggested by one of the participants, Sergio Jimenez.

The third evaluation metric is the weighted mean of the Pearson correlations on individual datasets. The Pearson returned for each dataset is weighted according to the number of sentence pairs in that dataset. Given r_i the five Pearson scores for each dataset, and n_i the number of pairs in each dataset, the weighted mean is given as $\sum_{i=1..5}(r_i * n_i)/\sum_{i=1..5} n_i$ We refer to this measure as *weighted mean of Pearson* or *Mean* for short.

4.2 Using confidence scores

Participants were allowed to include a confidence score between 1 and 100 for each of their scores. We used weighted Pearson to use those confidence scores[3]. Table 2 includes the list of systems which provided a non-uniform confidence. The results show that some systems were able to improve their correlation, showing promise for the usefulness of confidence in applications.

4.3 The Baseline System

We produced scores using a simple word overlap baseline system. We tokenized the input sentences splitting at white spaces, and then represented each sentence as a vector in the multidimensional token space. Each dimension had 1 if the token was present in the sentence, 0 otherwise. Similarity of vectors was computed using cosine similarity.

We also run a random baseline several times, yielding close to 0 correlations in all datasets, as expected. We will refer to the random baseline again in Section 4.5.

4.4 Participation

Participants could send a maximum of three system runs. After downloading the test datasets, they had a maximum of 120 hours to upload the results. 35 teams participated, submitting 88 system runs (cf. first column of Table 1). Due to lack of space we can't detail the full names of authors and institutions that participated. The interested reader can use the name of the runs to find the relevant paper in these proceedings.

There were several issues in the submissions. The submission software did not ensure that the naming conventions were appropriately used, and this caused some submissions to be missed, and in two cases the results were wrongly assigned. Some participants returned Not-a-Number as a score, and the organizers had to request whether those where to be taken as a 0 or as a 5.

Finally, one team submitted past the 120 hour deadline and some teams sent missing files after the deadline. All those are explicitly marked in Table 1. The teams that included one of the organizers are also explicitly marked. We want to stress that in these teams the organizers did not allow the developers of the system to access any data or information which was not available for the rest of participants. One exception is *weiwei*, as they generated the 110K OntoNotes-WordNet dataset from which the other organizers sampled the surprise data set.

After the submission deadline expired, the organizers published the gold standard in the task website, in order to ensure a transparent evaluation process.

4.5 Results

Table 1 shows the results for each run in alphabetic order. Each result is followed by the rank of the system according to the given evaluation measure. To the right, the Pearson score for each dataset is given. In boldface, the three best results in each column.

First of all we want to stress that the large majority of the systems are well above the simple baseline, although the baseline would rank 70 on the Mean measure, improving over 19 runs.

The correlation for the non-MT datasets were really high: the highest correlation was obtained was for MSRvid (0.88 r), followed by MSRpar (0.73 r) and On-WN (0.73 r). The results for the MT evaluation data are lower, (0.57 r) for SMT-eur and (0.61 r) for SMT-News. The simple token overlap baseline, on the contrary, obtained the highest results for On-WN (0.59 r), with (0.43 r) on MSRpar and (0.40 r) on MSRvid. The results for MT evaluation data are also reversed, with (0.40 r) for SMT-eur and (0.45 r) for SMT-News.

The ALLnorm measure yields the highest correlations. This comes at no surprise, as it involves a normalization which transforms the system outputs using the gold standard. In fact, a random baseline which gets Pearson correlations close to 0 in all datasets would attain Pearson of 0.5891[4].

Although not included in the results table for lack of space, we also performed an analysis of confidence intervals. For instance, the best run according to ALL (r = .8239) has a 95% confidence interval of [.8123,.8349] and the second a confidence interval of [.8016,.8254], meaning that the differences are not statistically different.

5 Tools and resources used

The organizers asked participants to submit a description file, special emphasis on the tools and resources that they used. Table 3 shows in a simpli-

[3] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient#Calculating_a_weighted_correlation

[4] We run the random baseline 10 times. The mean is reported here. The standard deviation is 0.0005

Run	ALL	Rank	ALLnrm	Rank	Mean	Rank	MSRpar	MSRvid	SMT-eur	On-WN	SMT-news
00-baseline/task6-baseline	.3110	87	.6732	85	.4356	70	.4334	.2996	.4542	.5864	.3908
aca08ls/task6-University_Of_Sheffield-Hybrid	.6485	34	.8238	15	.6100	18	.5166	.8187	.4859	.6676	.4280
aca08ls/task6-University_Of_Sheffield-Machine_Learning	.7241	17	.8169	18	.5750	38	.5166	.8187	.4859	.6390	.2089
aca08ls/task6-University_Of_Sheffield-Vector_Space	.6054	48	.7946	44	.5943	27	.5460	.7241	.4858	.6676	.4280
acaputo/task6-UNIBA-DEPRI	.6141	46	.8027	38	.5891	31	.4542	.7673	.5126	.6593	.4636
acaputo/task6-UNIBA-LSARI	.6221	44	.8079	30	.5728	40	.3886	.7908	.4679	.6826	.4238
acaputo/task6-UNIBA-RI	.6285	41	.7951	43	.5651	45	.4128	.7612	.4531	.6306	.4887
baer/task6-UKP-run1	.8117	4	.8559	4	.6708	4	.6821	.8708	.5118	.6649	.4672
baer/task6-UKP-run2_plus_postprocessing_smt_twsi	**.8239**	1	**.8579**	2	**.6773**	1	.6830	.8739	.5280	.6641	.4937
baer/task6-UKP-run3_plus_random	.7790	8	.8166	19	.4320	71	.6830	.8739	.5280	-.0620	-.0520
croce/task6-UNITOR-1_REGRESSION_BEST_FEATURES	.7474	13	.8292	12	.6316	10	.5695	.8217	.5168	.6591	.4713
croce/task6-UNITOR-2_REGRESSION_ALL_FEATURES	.7475	12	.8297	11	.6323	9	.5763	.8217	.5102	.6591	.4713
croce/task6-UNITOR-3_REGRESSION_ALL_FEATURES_ALL_DOMAINS	.6289	40	.8150	21	.5939	28	.4686	.8027	.4574	.6591	.4713
csjxu/task6-PolyUCOMP-RUN1	.6528	31	.7642	59	.5492	51	.4728	.6593	.4835	.6196	.4290
danielcer/stanford_fsa†	.6354	38	.7212	70	.4848	66	.3795	.5350	.4377	.6052	.4164
danielcer/stanford_pdaAll†	.4229	77	.7160	72	.5044	62	.4409	.4698	.4558	.6468	.4769
danielcer/stanford_rte†	.5589	55	.7807	55	.4674	67	.4374	.8037	.3533	.3077	.3235
davide_buscaldi/task6-IRIT-pg1	.4280	76	.7379	65	.5009	63	.4295	.6125	.4952	.5387	.3614
davide_buscaldi/task6-IRIT-pg3	.4813	68	.7569	61	.5202	58	.4171	.6728	.5179	.5526	.3693
davide_buscaldi/task6-IRIT-wu	.4064	81	.7287	69	.4898	65	.4326	.5833	.4856	.5317	.3480
demetrios_glinos/task6-ATA-BASE	.3454	83	.6990	81	.2772	87	.1684	.6256	.2244	.1648	.0988
demetrios_glinos/task6-ATA-CHNK	.4976	64	.7160	73	.3215	86	.2312	.6595	.1504	.2735	.1426
demetrios_glinos/task6-ATA-STAT	.4165	79	.7129	75	.3312	85	.1887	.6482	.2769	.2950	.1336
desouza/task6-FBK-run1	.5633	54	.7127	76	.3628	82	.2494	.6117	.1495	.4212	.2439
desouza/task6-FBK-run2	.6438	35	.8080	29	.5888	32	.5128	.7807	.3796	.6228	.5474
desouza/task6-FBK-run3	.6517	32	.8106	25	.6077	20	.5169	.7773	.4419	.6298	.6085
dvilarinoayala/task6-BUAP-RUN-1	.4997	63	.7568	62	.5260	56	.4037	.6532	.4521	.6050	.4537
dvilarinoayala/task6-BUAP-RUN-2	-.0260	89	.5933	89	.1016	89	.1109	.0057	.0348	.1788	.1964
dvilarinoayala/task6-BUAP-RUN-3	.6630	25	.7474	64	.5105	59	.4018	.6378	.4758	.5691	.4057
enrique/task6-UNED-H34measures	.4381	75	.7518	63	.5577	48	.5328	.5788	.4785	.6692	.4465
enrique/task6-UNED-HallMeasures	.2791	88	.6694	87	.4286	72	.3861	.2570	.4086	.6006	.5305
enrique/task6-UNED-SP_INIST	.4680	69	.7625	60	.5615	47	.5166	.6303	.4625	.6442	.4753
georgiana_dinu/task6-SAARLAND-ALIGN_VSSIM	.4952	65	.7871	50	.5065	60	.4043	.7718	.2686	.5721	.3505
georgiana_dinu/task6-SAARLAND-MIXT_VSSIM	.4548	71	.8258	13	.5662	43	.6310	.8312	.1391	.5966	.3806
jan_snajder/task6-takelab-simple	**.8133**	3	**.8635**	1	**.6753**	2	.7343	.8803	.4771	.6797	.3989
jan_snajder/task6-takelab-syntax	**.8138**	2	**.8569**	3	.6601	5	.6985	.8620	.3612	.7049	.4683
janardhan/task6-janardhan-UNL_matching	.3431	84	.6878	84	.3481	83	.1936	.5504	.3755	.2888	.3387
jhasneha/task6-Penn-ELReg	.6622	27	.8048	34	.5654	44	.5480	.7844	.3513	.6040	.3607
jhasneha/task6-Penn-ERReg	.6573	28	.8083	28	.5755	37	.5610	.7857	.3568	.6214	.3732
jhasneha/task6-Penn-LReg	.6497	33	.8043	36	.5699	41	.5460	.7818	.3547	.5969	.4137
jotacastillo/task6-SAGAN-RUN1	.5522	57	.7904	47	.5906	29	.5659	.7113	.4739	.6542	.4253
jotacastillo/task6-SAGAN-RUN2	.6272	42	.8032	37	.5838	34	.5538	.7706	.4480	.6135	.3894
jotacastillo/task6-SAGAN-RUN3	.6311	39	.7943	45	.5649	46	.5394	.7560	.4181	.5904	.3746
Konstantin_Z/task6-ABBYY-General	.5636	53	.8052	33	.5759	36	.4797	.7821	.4576	.6488	.3682
M_Rios/task6-UOW-LEX_PARA	.6397	36	.7187	71	.3825	80	.3628	.6426	.3074	.2806	.2082
M_Rios/task6-UOW-LEX_PARA_SEM	.5981	49	.6955	82	.3473	84	.3529	.5724	.3066	.2643	.1164
M_Rios/task6-UOW-SEM	.5361	59	.6287	88	.2567	88	.2995	.2910	.1611	.2571	.2212
mheilman/task6-ETS-PERP	.7808	7	.8064	32	.6305	11	.6211	.7210	.4722	.7080	.5149
mheilman/task6-ETS-PERPphrases	.7834	6	.8089	27	.6399	7	.6397	.7200	.4850	.7124	.5312
mheilman/task6-ETS-TERp	.4477	73	.7291	68	.5253	57	.5049	.5217	.4748	.6169	.4566
nitish_aggarwal/task6-aggarwal-run1⋆	.5777	52	.8158	20	.5466	52	.3675	.8427	.3534	.6030	.4430
nitish_aggarwal/task6-aggarwal-run2⋆	.5833	51	.8183	17	.5683	42	.3720	.8330	.4238	.6513	.4499
nitish_aggarwal/task6-aggarwal-run3	.4911	67	.7696	57	.5377	53	.5320	.6874	.4514	.5827	.2818
nmalandrakis/task6-DeepPurple-DeepPurple_hierarchical	.6228	43	.8100	26	.5979	23	.5984	.7717	.4292	.6480	.3702
nmalandrakis/task6-DeepPurple-DeepPurple_sigmoid	.5540	56	.7997	41	.5558	50	.5960	.7616	.2628	.6016	.3446
nmalandrakis/task6-DeepPurple-DeepPurple_single	.4918	66	.7646	58	.5061	61	.4989	.7092	.4437	.4879	.2441
parthapakray/task6-JU_CSE_NLP-Semantic_Syntactic_Approach∗	.3880	82	.6706	86	.4111	76	.3427	.3549	.4271	.5298	.4034
rada/task6-UNT-CombinedRegression	.7418	14	.8406	7	.6159	14	.5032	.8695	.4797	.6715	.4033
rada/task6-UNT-IndividualDecTree	.7677	9	.8389	9	.5947	25	.5693	.8688	.4203	.6491	.2256
rada/task6-UNT-IndividualRegression	.7846	5	.8440	6	.6162	13	.5353	.8750	.4203	.6715	.4033
sbdlrhmn/task6-sbdlrhmn-Run1	.6663	23	.7842	53	.5376	54	.5440	.7335	.3830	.5860	.2445
sbdlrhmn/task6-sbdlrhmn-Run2	.4169	78	.7104	77	.4986	64	.4617	.4489	.4719	.6353	.4353
sgjimenezv/task6-SOFT-CARDINALITY	.7331	15	.8526	5	**.6708**	3	.6405	.8562	.5152	.7109	.4833
sgjimenezv/task6-SOFT-CARDINALITY-ONE-FUNCTION	.7107	19	.8397	8	.6486	6	.6316	.8237	.4320	.7109	.4833
siva/task6-DSS-alignheuristic	.5253	60	.7962	42	.6030	21	.5735	.7123	.4781	.6984	.4177
siva/task6-DSS-average	.5490	58	.8047	35	.5943	26	.5020	.7645	.4875	.6677	.4324
siva/task6-DSS-wordsim	.5130	61	.7895	49	.5287	55	.3765	.7761	.4161	.5728	.3964
skamler_/task6-EHU-RUN1v2⋆†	.3129	86	.6935	83	.3889	79	.3605	.5187	.2259	.4098	.3465
sokolov/task6-LIMSI-cosprod	.6392	37	.7344	67	.3940	78	.3948	.6597	.0143	.4157	.2889
sokolov/task6-LIMSI-gradtree	.6789	22	.7377	66	.4118	75	.4848	.6636	.0934	.3706	.2455
sokolov/task6-LIMSI-sumdiff	.6196	45	.7101	78	.4131	74	.4295	.5724	.2842	.3989	.2575
spirin2/task6-UIUC-MLNLP-Blend	.4592	70	.7800	56	.5782	35	.6523	.6691	.3566	.6117	.4603
spirin2/task6-UIUC-MLNLP-CCM	.7269	16	.8217	16	.6104	17	.5769	.8203	.4667	.5835	.4945
spirin2/task6-UIUC-MLNLP-Puzzle	.3216	85	.7857	51	.4376	69	.5635	.8056	.0630	.2774	.2409
sranjans/task6-sranjans-1	.6529	30	.8018	39	.6249	12	.6124	.7240	.5581	.6703	.4533
sranjans/task6-sranjans-2	.6651	24	.8128	22	.6366	8	.6254	.7538	.5328	.6649	.5036
sranjans/task6-sranjans-3	.5045	62	.7846	52	.5905	30	.6167	.7061	.5666	.5664	.3968
tiantianzhu7/task6-tiantianzhu7-1	.4533	72	.7134	74	.4192	73	.4184	.5630	.2083	.4822	.2745
tiantianzhu7/task6-tiantianzhu7-2	.4157	80	.7099	79	.3960	77	.4260	.5628	.1546	.4552	.1923
tiantianzhu7/task6-tiantianzhu7-3	.4446	74	.7097	80	.3740	81	.3411	.5946	.1868	.4029	.1823
weiwei/task6-weiwei-run1⋆†	.6946	20	.8303	10	.6081	19	.4106	.8351	.5128	.7273	.4383
yeh/task6-SRIUBC-SYSTEM1†	.7513	11	.8017	40	.5997	22	.6084	.7458	.4688	.6315	.3994
yeh/task6-SRIUBC-SYSTEM2†	.7562	10	.8111	24	.5858	33	.6050	.7939	.4294	.5871	.3366
yeh/task6-SRIUBC-SYSTEM3†	.6876	21	.7812	54	.4668	68	.4791	.7901	.2159	.3843	.2801
ygutierrez/task6-UMCC_DLSI-MultiLex	.6630	26	.7922	46	.5560	49	.6022	.7709	.4435	.4327	.4264
ygutierrez/task6-UMCC_DLSI-MultiSem	.6529	29	.8115	23	.6116	16	.5269	.7756	.4688	.6539	.5470
ygutierrez/task6-UMCC_DLSI-MultiSemLex	.7213	18	.8239	14	.6158	15	.6205	.8104	.4325	.6256	.4340
yrkakde/task6-yrkakde-DiceWordnet	.5977	50	.7902	48	.5742	39	.5294	.7470	.5531	.5698	.3659
yrkakde/task6-yrkakde-JaccNERPenalty	.6067	47	.8078	31	.5955	24	.5757	.7765	.4989	.6257	.3468

Table 1: The first row corresponds to the baseline. **ALL** for overall Pearson, **ALLnorm** for Pearson after normalization, and **Mean** for mean of Pearsons. We also show the ranks for each measure. Rightmost columns show Pearson for each individual dataset. Note: ∗ system submitted past the 120 hour window, ⋆ post-deadline fixes, † team involving one of the organizers.

Run	ALL	ALL_w	MSRpar	$MSRpar_w$	MSRvid	$MSRvid_w$	SMT-eur	$SMT\text{-}eur_w$	On-WN	$On\text{-}WN_w$	SMT-news	$SMT\text{-}news_w$
davide_buscaldi/task6-IRIT-pg1	.4280	**.4946**	**.4295**	.4082	.6125	**.6593**	.4952	**.5273**	.5387	**.5574**	.3614	**.4674**
davide_buscaldi/task6-IRIT-pg3	.4813	**.5503**	**.4171**	.4033	.6728	**.7048**	.5179	**.5529**	.5526	**.5950**	.3693	**.4648**
davide_buscaldi/task6-IRIT-wu	.4064	**.4682**	**.4326**	.4035	.5833	**.6253**	.4856	**.5138**	**.5317**	.5189	.3480	**.4482**
enrique/task6-UNED-H34measures	**.4381**	.2615	**.5328**	.4494	**.5788**	.4913	**.4785**	.4660	**.6692**	.6440	**.4465**	.3632
enrique/task6-UNED-HallMeasures	**.2791**	.2002	**.3861**	.3802	**.2570**	.2343	.4086	**.4212**	**.6006**	.5947	**.5305**	.4858
enrique/task6-UNED-SP_INIST	**.4680**	.3754	**.5166**	.5082	**.6303**	.5588	.4625	**.4801**	**.6442**	.5761	**.4753**	.4143
parthapakray/task6-JU_CSE_NLP-Semantic_Syntactic_Approach	**.3880**	.3636	.3427	**.3498**	**.3549**	.3353	**.4271**	.3989	**.5298**	.4619	**.4034**	.3228
tiantianzhu7/task6-tiantianzhu7-1	.4533	**.5442**	.4184	**.4241**	**.5630**	**.5630**	.2083	**.4220**	.4822	**.5031**	.2745	**.3536**
tiantianzhu7/task6-tiantianzhu7-2	.4157	**.5249**	.4260	**.4340**	.5628	**.5758**	.1546	**.4776**	.4552	**.4926**	.1923	**.3362**
tiantianzhu7/task6-tiantianzhu7-3	.4446	**.5229**	.3411	**.3611**	**.5946**	.5899	.1868	**.4769**	.4029	**.4365**	.1823	**.4014**

Table 2: Results according to weighted correlation for the systems that provided non-uniform confidence alongside their scores.

fied way the tools and resources used by those participants that did submit a valid description file. In the last row, the totals show that WordNet was the most used resource, followed by monolingual corpora and Wikipedia. Acronyms, dictionaries, multilingual corpora, stopword lists and tables of paraphrases were also used.

Generic NLP tools like lemmatization and PoS tagging were widely used, and to a lesser extent, parsing, word sense disambiguation, semantic role labeling and time and date resolution (in this order). Knowledge-based and distributional methods got used nearly equally, and to a lesser extent, alignment and/or statistical machine translation software, lexical substitution, string similarity, textual entailment and machine translation evaluation software. Machine learning was widely used to combine and tune components. Several less used tools were also listed but were used by three or less systems.

The top scoring systems tended to use most of the resources and tools listed (*UKP*, *Takelab*), with some notable exceptions like *Sgjimenez* which was based on string similarity. For a more detailed analysis, the reader is directed to the papers of the participants in this volume.

6 Conclusions and Future Work

This paper presents the SemEval 2012 pilot evaluation exercise on Semantic Textual Similarity. A simple definition of STS beyond the likert-scale was set up, and a wealth of annotated data was produced. The similarity of pairs of sentences was rated on a 0-5 scale (low to high similarity) by human judges using Amazon Mechanical Turk. The dataset includes 1500 sentence pairs from MSRpar and MSRvid (each), ca. 1500 pairs from WMT, and 750 sentence pairs from a mapping between OntoNotes and WordNet senses. The correlation between non-expert annotators and annotations from the authors is very high, showing the high quality of the dataset. The dataset was split 50% as train and test, with the exception of the surprise test datasets: a subset of WMT from a different domain and the OntoNotes-WordNet mapping. All datasets are publicly available.[5]

The exercise was very successful in participation and results. 35 teams participated, submitting 88 runs. The best results scored a Pearson correlation over 80%, well beyond a simple lexical baseline with 31% of correlation. The metric for evaluation was not completely satisfactory, and three evaluation metrics were finally published. We discuss the shortcomings of those measures.

There are several tasks ahead in order to make STS a mature field. The first is to find a satisfactory evaluation metric. The second is to analyze the definition of the task itself, with a thorough analysis of the definitions in the likert scale.

We would also like to analyze the relation between the STS scores and the paraphrase judgements in MSR, as well as the human evaluations in WMT. Finally, we would also like to set up an open framework where NLP components and similarity algorithms can be combined by the community. All in all, we would like this dataset to be the focus of the community working on algorithmic approaches for semantic processing and inference at large.

Acknowledgements

We would like to thank all participants, specially (in alphabetic order) Yoan Gutierrez, Michael Heilman, Sergio Jimenez, Nitin Madnami, Diana McCarthy and Shrutiranjan Satpathy for their contributions on evaluation metrics. Eneko Agirre was partially funded by the

[5] `http://www.cs.york.ac.uk/semeval-2012/task6/`

	Acronyms	Dictionaries	Distributional thesaurus	Monolingual corpora	Multilingual corpora	Stop words	Tables of paraphrases	Wikipedia	WordNet	Alignment	Distributional similarity	KB Similarity	Lemmatizer	Lexical Substitution	Machine Learning	MT evaluation	MWE	Named Entity recognition	POS tagger	Semantic Role Labeling	SMT	String similarity	Syntax	Textual entailment	Time and date resolution	Word Sense Disambiguation	Other
aca08ls/task6-University_Of_Sheffield-Hybrid									x			x	x		x							x				x	x
aca08ls/task6-University_Of_Sheffield-Machine_Learning									x			x	x		x							x				x	x
aca08ls/task6-University_Of_Sheffield-Vector_Space									x			x	x													x	x
baer/task6-UKP-run1		x	x	x	x			x	x		x	x	x						x		x	x		x			x
baer/task6-UKP-run2_plus_postprocessing_smt_twsi		x	x	x	x			x	x		x	x	x						x		x	x		x			x
baer/task6-UKP-run3_plus_random		x	x	x	x			x	x		x	x	x						x		x	x		x			x
croce/task6-UNITOR-1_REGRESSION_BEST_FEATURES				x							x		x		x				x				x				
croce/task6-UNITOR-2_REGRESSION_ALL_FEATURES				x							x		x		x				x				x				
croce/task6-UNITOR-3_REGRESSION_ALL_FEATURES_ALL_DOMAINS				x							x		x		x				x				x				
csjxu/task6-PolyUCOMP-RUN								x	x			x							x								
danielcer/stanford_fsa							x		x				x	x			x		x				x				
danielcer/stanford_pdaAll							x		x				x	x			x		x				x				
danielcer/stanford_rte									x	x			x		x		x		x				x				x
davide_buscaldi/task6-IRIT-pg1				x					x			x							x			x					
davide_buscaldi/task6-IRIT-pg3				x					x			x							x			x					
davide_buscaldi/task6-IRIT-wu				x					x			x							x			x					
demetrios_glinos/task6-ATA-BASE									x		x	x							x	x		x					x
demetrios_glinos/task6-ATA-CHNK									x		x	x							x	x		x					x
demetrios_glinos/task6-ATA-STAT									x		x	x							x	x		x					x
desouza/task6-FBK-run1	x			x			x	x	x		x	x	x	x				x	x				x				x
desouza/task6-FBK-run2				x			x	x	x		x	x	x						x								
desouza/task6-FBK-run3				x			x	x	x		x								x								
dvilarinoayala/task6-BUAP-RUN-1		x											x														
dvilarinoayala/task6-BUAP-RUN-2									x																		
dvilarinoayala/task6-BUAP-RUN-3													x														x
jan_snajder/task6-takelab-simple		x	x	x		x		x	x		x	x	x	x	x				x								x
jan_snajder/task6-takelab-syntax				x					x		x	x	x	x				x	x				x				
janardhan/task6-janardhan-UNL_matching									x				x					x	x				x			x	
jotacastillo/task6-SAGAN-RUN1	x			x					x				x	x										x	x	x	
jotacastillo/task6-SAGAN-RUN2	x			x					x				x	x										x	x	x	
jotacastillo/task6-SAGAN-RUN3	x			x					x				x	x										x	x	x	
Konstantin_Z/task6-ABBYY-General																											
M_Rios/task6-UOW-LEX_PARA			x								x		x			x		x	x	x			x				
M_Rios/task6-UOW-LEX_PARA_SEM			x								x		x			x		x	x	x			x				
M_Rios/task6-UOW-SEM			x								x		x					x	x	x			x				
mheilman/task6-ETS-PERP				x					x			x	x	x		x						x					
mheilman/task6-ETS-PERPphrases				x	x				x			x	x	x		x						x					x
mheilman/task6-ETS-TERp				x	x				x			x	x	x													x
parthapakray/task6-JU_CSE_NLP-Semantic_Syntactic_Approach	x					x			x				x				x	x	x				x	x			x
rada/task6-UNT-CombinedRegression								x	x	x	x	x	x		x											x	x
rada/task6-UNT-IndividualDecTree								x	x	x	x	x	x		x											x	x
rada/task6-UNT-IndividualRegression								x	x	x	x	x	x		x											x	x
sgjimenezv/task6-SOFT-CARDINALITY						x						x	x														
sgjimenezv/task6-SOFT-CARDINALITY-ONE-FUNCTION						x						x	x														
skamler_/task6-EHU-RUN1v2							x			x			x						x		x						
sokolov/task6-LIMSI-cosprod				x							x		x		x												
sokolov/task6-LIMSI-gradtree				x							x		x		x												
sokolov/task6-LIMSI-sumdiff				x							x		x		x												
spirin2/task6-UIUC-MLNLP-Blend	x			x			x	x									x	x	x	x			x	x			x
spirin2/task6-UIUC-MLNLP-CCM	x			x			x	x									x	x	x	x			x	x			x
spirin2/task6-UIUC-MLNLP-Puzzle	x			x			x	x									x	x	x	x			x	x			x
sranjans/task6-sranjans-1				x				x	x		x	x	x						x							x	
sranjans/task6-sranjans-2				x				x	x		x	x	x	x				x	x						x	x	
sranjans/task6-sranjans-3				x				x	x		x	x	x	x				x	x				x		x		
tiantianzhu7/task6-tiantianzhu7-1									x			x							x							x	
tiantianzhu7/task6-tiantianzhu7-2									x			x							x								
tiantianzhu7/task6-tiantianzhu7-3						x			x			x							x								
weiwei/task6-weiwei-run1		x		x					x		x		x						x								
yeh/task6-SRIUBC-SYSTEM1			x					x	x		x	x	x						x								
yeh/task6-SRIUBC-SYSTEM2			x					x	x		x	x	x						x								
yeh/task6-SRIUBC-SYSTEM3			x					x	x		x	x	x						x								
ygutierrez/task6-UMCC_DLSI-MultiLex				x					x	x		x	x						x								x
ygutierrez/task6-UMCC_DLSI-MultiSem				x					x			x	x						x							x	x
ygutierrez/task6-UMCC_DLSI-MultiSemLex				x					x	x		x	x						x							x	x
yrkakde/task6-yrkakde-DiceWordnet									x		x		x														
Total	8	6	10	33	5	5	9	20	47	7	31	37	49	13	13	4	7	12	43	9	4	13	17	10	5	15	25

Table 3: Resources and tools used by the systems that submitted a description file. Leftmost columns correspond to the resources, and rightmost to tools, in alphabetic order.

European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270082 (PATHS project) and the Ministry of Economy under grant TIN2009-14715-C04-01 (KNOW2 project). Daniel Cer gratefully acknowledges the support of the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181 and the support of the DARPA Broad Operational Language Translation (BOLT) program through IBM. The STS annotations were funded by an extension to DARPA GALE subcontract to IBM # W0853748 4911021461.0 to Mona Diab. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

References

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 136–158.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 70–106.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meetings of the Association for Computational Linguistics (ACL)*.

B. Dolan, C. Quirk, and C. Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 04: Proceedings of the 20th international conference on Computational Linguistics*, page 350.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*.

Michael D. Lee, Brandon Pincombe, and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259, Mahwah, NJ.

Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150, August.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October.

E. Ukkonen. 1985. Algorithms for approximate string matching. *Information and Contro*, 64:110–118.