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

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

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

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

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

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. 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 CPAN¹ 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

¹http://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.
- Not equivalent, but share some becaus.
 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-

²At 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 scores³. 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 0.5891⁴.

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-

³http://en.wikipedia.org/wiki/Pearson_ product-moment_correlation_coefficient# Calculating_a_weighted_correlation

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

| 00-baselind-nakeb-useline 1110 87 6752 88.5 4756 70 4334 2996 4342 586 aca08ls/task/University.Of.Sheffield-MineLearning 7241 17 8169 18 5166 8187 4859 650 acaputotask/University.Of.Sheffield-Vector Space 6054 48 7946 44 5943 1 452 7613 5126 6593 acaputotask/UNIBA-LSARI 6214 44 8077 30 5728 40 3886 7908 4671 4581 8708 6671 4533 4531 650 4531 650 4531 650 4531 650 4531 650 4531 650 4531 650 4531 650 4531 650 4531 650 4531 650 4531 650 4531 650 4531 650 4533 650 453 453 450 453 4531 650 4533 4577 11 653 5303 | $\begin{array}{r} .3908\\ .4280\\ .2089\\ .4238\\ .4238\\ .4238\\ .4238\\ .4487\\ .4672\\ .4672\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4290\\ .4164\\ .4290\\ .3235\\ .3693\\ .3480\\ .0988\\ .3255\\ .3612\\ .4753\\ .1964\\ .4057\\ .5305\\ .3505\end{array}$ |
|---|---|
| ace08ls/task-0-University.OC.Sheffield-VetCor.Space 054 48 (a) 8770 38 5716 8.187 4.859 6.360 acaputo/task-0-UNIBA-DEPR1 6.141 46 8.027 33 5.281 31 4.542 7.73 5.126 6.593 acaputo/task-0-UNIBA-LSAR1 6.221 44 6.8027 30 5728 40 3886 7.708 4.679 6.826 acaputo/task-0-UNIBA-LSAR1 6.225 41 7.751 4.855 44 6.781 4.859 4.850 4.852 7.751 4.851 6.659 bear/task-0-UNFP-run1 8.17 41 6.285 41 7.751 4 6.783 8.739 5.280 6.661 bear/task-0-UNFP-run1 8.239 1.8579 4 6.708 4.672 1.708 8.739 5.280 6.661 bear/task-0-UNFP-run1 8.239 1.8579 4 6.708 4.672 1.708 8.739 5.280 6.661 bear/task-0-UNF0-r_LREGRESSION-ALL-FEATURES 7.747 1 13 8.292 1.6316 10 5.695 8.217 5.168 6.659 croce/task-0-UNTOR-1_REGRESSION-ALL-FEATURES 7.747 1 13 8.297 11 6.323 9 5.763 3.817 5.102 6.591 croce/task-0-UNTOR-2_REGRESSION-ALL-FEATURES 7.747 51 21 8.297 11 6.323 9 5.763 3.817 5.102 6.591 croce/task-0-UNTOR-2_REGRESSION-ALL-FEATURES 7.747 51 21 8.297 11 6.323 9 5.763 3.817 5.102 6.591 croce/task-0-UNTOR-2_REGRESSION-ALL-FEATURES 7.747 51 22 9.540 2.1 4.728 6.539 4.815 6.196 danielecr/stanford_fsaT 6.354 38 7.712 70 5.407 61 4.374 8.037 4.553 4.857 6.169 danielecr/stanford_fsaT 6.354 38 7.712 7.160 7.2 5.047 61 7.374 6.528 3.1 4.718 6.758 4.618 danielecr/stanford_teT 5.580 4.812 5.587 7.807 55 5.067 4.67 4.374 8.037 4.553 4.857 6.458 davide-buscald/task-RIT-pg1 4.229 77 7.106 7.2 5.047 61 5.322 8.6125 4.952 5.387 davide-buscald/task-RIT-Pg3 4.813 68 7.690 4.898 6.4326 5.324 4.1648 darettros_glino.87a8.6-ATA-BASE 3.3454 83 6.990 81 5.20 2.58 4.171 6.728 3.3485 6.1517 demetrics_glino.87a8.6-ATA-CHNK 4.976 64 7.0 2.788 6.482 2.349 4.1648 demetrics_glino.87a8.6-ATA-CHNK 4.976 64 7.2187 6.828 2.349 6.482 7.378 6.259 2.250 devide-buscald/task-RIT-Pg3 4.413 8.808 2.774 8 4.513 6.462 7.278 4.278 | $\begin{array}{c} 2089\\ .4230\\ .4636\\ .4238\\ .4672\\ .4937\\ .0520\\ .4713\\ .4713\\ .4713\\ .4713\\ .4290\\ .4164\\ .4769\\ .3235\\ .3616\\ .3235\\ .3603\\ .3480\\ .0988\\ .1426\\ .1336\\ .2439\\ .1426\\ .1336\\ .2439\\ .1426\\ .5377\\ .1964\\ .4057\\ .5305\\ .3505\end{array}$ |
| acaoblitzakó-University Of.Sheffield-Vector.Space 6054 48 7946 49 5943 27 5460 7241 4858 6676 acaputo/taskó-UNBA-LSARI 6221 44 8079 30 5728 40 3866 7708 516 6593 acaputo/taskó-UNBA-LSARI 6221 44 8079 30 5728 40 3866 7708 512 6593 baer/taskó-UKP-run2.plus.psoptrocessing.smt.tws 8117 4 8859 4 6708 4 6821 8708 518 6641 baer/taskó-UKP-run2.plus.psoptrocessing.smt.tws 8117 4 8859 2 6773 1 6830 8739 5280 -0620 croce/taskó-UNTOR-1.REGRESSION.ALL_FEATURES 7475 12 8297 11 6323 9 5763 8217 5180 6591 croce/taskó-UNTOR-2.REGRESSION.ALL_FEATURES 7475 12 8297 11 6323 9 5763 8217 5180 6591 croce/taskó-UNTOR-3.REGRESSION.ALL_FEATURES 7475 12 8297 11 6323 9 5402 530 -4352 6593 dialecter/stanford.LFFATURES 7475 12 8297 11 6323 9 5402 51 4432 6593 4335 6196 dialecter/stanford.LFFATURES 7475 12 8297 71 70 848 66 3795 530 4377 6552 dialecter/stanford.LFFATURES 7475 12 8297 71 70 848 66 3795 530 4377 6552 dialecter/stanford.pdaAlt 4229 77 7.160 72 5044 62 4409 4698 4558 6591 dialecter/stanford.pdaAlt 4229 77 7.160 72 5044 62 4409 4698 4558 6531 3077 diavide.buscald/taskó-IRT-pg1 4280 76 7.379 65 500 63 4225 6125 43952 5387 4376 4552 4583 4856 5317 dialecter/stanford.pdaAlt 4229 77 7.160 72 5044 62 4409 4698 4558 6531 3077 diavide.buscald/taskó-IRT-pg1 4280 76 7.379 75 312 86 12312 6595 1594 4273 6592 dialecter/stanford.pdaAlt 4229 77 7.160 72 5044 62 4409 4698 4558 6531 70 diavide.buscald/taskó-IRT-pg3 4413 68 71.278 769 61 5202 84 1471 763 5128 5179 diavide.buscald/taskó-IRT-pg1 4280 76 7.379 55 412 7358 62 5007 63 422 5769 5250 desouza/taskó-FBK-run1 5633 54 7.172 76 3312 86 12312 6595 1504 4273 518 622 5269 desouza/taskó-FBK-run3 6517 32 8180 29 5888 2 5128 773 368 62 5200 63 4373 4758 542 2769 desouza/taskó-FBK-run3 6517 32 8108 29 588 82 5128 773 348 578 458 6591 diavida.busch/taskó-UNED-H34mesures 4381 75 7.578 62 526 430 737 458 578 452 5769 40317 456 5721 georginaa.dinu/taskó-SAARLAND-MUTN 3488 471 8288 11 6659 5 4018 7.118 2686 65721 georginaa.dinu/taskó-SAARLAND-AUGN 458 478 438 38 4 3618 5 6958 480 3.425 7.448 5.428 7 | $\begin{array}{r} .4280\\ .4238\\ .4238\\ .4238\\ .4238\\ .4238\\ .4238\\ .4297\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4713\\ .4290\\ .4164\\ .4769\\ .3235\\ .3614\\ .3693\\ .3480\\ .0988\\ .3693\\ .3693\\ .3693\\ .3693\\ .3693\\ .3480\\ .0988\\ .42439\\ .$ |
| acaputo/ask6-UNBA-LSARI 6421 44 6827 38 5891 31 4542 7673 5126 6593 acaputo/ask6-UNBA-LSARI 6221 44 7879 30 5728 40 3886 7708 4679 6826 acaputo/ask6-UNPLA-LSARI 6228 44 7751 45 4128 7712 4531 6506 baertask6-UKP-run3 plus-anadom 7790 8 8299 1 8579 4 6708 4 6821 771 830 8719 5280 6641 baertask6-UKP-run3 plus-anadom 7790 8 8297 11 6320 71 6830 8739 5280 .0620 croce/task6-UNTIOR-1.REGRESSION.ALL-FEATURES 7474 13 8292 1 6316 10 5695 8217 5168 6591 croce/task6-UNTIOR-2.REGRESSION.ALL-FEATURES 7474 13 8292 11 6320 9 5763 8217 5108 6591 croce/task6-UNTIOR-2.REGRESSION.ALL-FEATURES 7474 13 8292 11 6320 9 5763 8217 5108 6591 croce/task6-UNTIOR-3.REGRESSION.ALL-FEATURES 7474 13 8292 11 6320 9 5763 8317 5102 croce/task6-UNTIOR-2.REGRESSION.ALL-FEATURES 7474 13 8292 12 5399 28 4686 8027 4574 6591 danielcer/stanford_LSaT 6320 7716 72 5044 66 7375 5350 4377 6052 danielcer/stanford_LSaT 6320 7716 72 5047 67 4374 8037 3533 3077 danielcer/stanford_LFBT 6358 55 7807 55 500 43 4295 6125 4952 5387 danielcer/stanford_LFBT 6358 55 7807 55 4674 67 4374 8037 3533 3077 danielcer/stanford_LFBT 6358 55 7807 55 500 4486 65 4326 525 4352 danielcer/stanford_LFBT 634 4813 68 7569 61 520 58 4212 5387 danielcer/stanford_LFBT 64 538 55 7807 55 500 56 4027 533 4856 5317 danielcer/stanford_LFBT 64 548 55 7807 56 500 9 488 65 4326 525 4352 danielcer/stanford_LFBT 792 4481 68 7569 61 8202 58 8171 6728 5317 demetrios_glinos/task6-ATA-CINK 4451 79 7129 75 6312 85 1387 6452 1273 520 desouza/task6-FBK-run1 5633 54 7170 73 4212 85 1818 6425 1273 540 demetrios_glinos/task6-ATA-CINK 4451 79 7129 76 5316 62 3120 773 448 1419 2038 dvilarinooyalarask6-BLAP-RUN-1 4907 63 753 808 02 5588 82 5177 346 4526 5244 149 2038 dvilarinooyalarask6-BLAP-RUN-1 4907 63 763 45 1109 0357 0348 1788 dvilarinooyalarask6-BLAP-RUN-1 4907 63 763 45 1109 0357 0348 1788 dvilarinooyalarask6-BLAP-RUN-1 4907 63 763 45 1109 0357 0348 1788 dvilarinooyalarask6-BLAP-RUN-1 4907 63 763 82 8050 54 0353 7573 86 000 dvilarinooyalarask6-BLAP-RUN-1 4907 63 7751 80 506 40 | $\begin{array}{c} .4636\\ .4238\\ .4887\\ .4672\\ .4937\\ .4713\\ .4713\\ .4713\\ .4713\\ .4290\\ .4164\\ .4769\\ .3235\\ .3614\\ .3693\\ .3480\\ .0988\\ .0988\\ .0988\\ .0988\\ .1426\\ .1336\\ .1426\\ .1336\\ .2439\\ .5474\\ .6085\\ .5305\\ .4537\\ .1964\\ .4465\\ .5305\\ .3505\\ \end{array}$ |
| acaputo/task6-UKP-run1 6285 41 7951 43 5551 45 4128 7.012 4531 6306 baer/task6-UKP-run2.plus.postprocessing.smLtwsi 8239 1 8.870 4 6708 4 6820 87739 5280 -66641 baer/task6-UKP-run2.plus.postprocessing.smLtwsi 7290 8 8166 19 4320 11 6330 8739 5280 -6621 erroce/task6-UNTOR-1.REGRESSION_ALL_FEATURES 7474 13 8292 12 6331 6376 5217 516 6591 croce/task6-UNTOR-2.REGRESSION_ALL_FEATURES 7474 13 8292 51 4758 5616 438 5717 510 6379 55350 4377 6512 6337 5535 6477 63 4808 3077 5448 63 3797 5330 4377 6512 4587 451 453 5817 641 479 451 452 5837 456 5317 547 67 | $\begin{array}{r} .4887\\ .4672\\ .4937\\ .4937\\ .4937\\ .4937\\ .4713\\ .4713\\ .4713\\ .4713\\ .4290\\ .4164\\ .4769\\ .3235\\ .3614\\ .3235\\ .3693\\ .32480\\ .0988\\ .1426\\ .1336\\ .2439\\ .5474\\ .6085\\ .1336\\ .2439\\ .5474\\ .4537\\ .1964\\ .4537\\ .1964\\ .4055\\ .5305\\ .3505\end{array}$ |
| bacrtask6-UKP-run1 8117 4 8579 2 6773 1 6.680 8739 5280 6649 bacrtask6-UKP-run2, plus, sopprocessing, ant.uwi 7790 8 8.166 19 4720 11 6830 8739 5280 .6620 croce/task6-UNTOR-1.REGRESSION.ALL_FEATURES 7474 13 8.292 12 6310 5.595 8217 5108 6591 croce/task6-UNTOR-2.REGRESSION.ALL_FEATURES 7474 13 742 9 592 24 4853 6196 danielcer/stanford_fsa1 6354 38 7212 70 4484 66 3795 5350 4373 5333 3077 davide_buscald/task6-IRT-pg1 4280 76 13202 58 4111 6728 5179 5526 davide_buscald/task6-IRT-pg3 4813 68 7560 61 5202 583 4865 5317 davide_buscald/task6-IRT-pg3 4814 83 6900 81 2772 77 | $\begin{array}{r} .4672\\ .4937\\0520\\ .4713\\ .4713\\ .4713\\ .4290\\ .4164\\ .4769\\ .3235\\ .3614\\ .3693\\ .3480\\ .0988\\ .1426\\ .1336\\ .1426\\ .1336\\ .2439\\ .5474\\ .6085\\ .5305\\ .4537\\ .1964\\ .4055\\ .5305\\ .3505\end{array}$ |
| baedraskó-UKP-nuz-plus-postprocessing-smL.vsi R239 1 R576 2 6773 1 6830 8739 5280 -0620 croce/taskó-UNTOR-1. REGRESSION_ALL_FEATURES 7474 13 8297 11 6323 9 5680 8739 5280 -0620 croce/taskó-UNTOR-2. REGRESSION_ALL_FEATURES 7474 13 8297 11 6323 9 5763 8217 5102 6591 croce/taskó-UNTOR-3. REGRESSION_ALL_FEATURES 7471 6324 8 7212 70 8795 5350 617 4717 6535 7807 53 4717 6052 danielcer/stanford_pda/li† 4229 77 7160 72 5044 62 4295 633 377 3333 3077 davide-buscaldi/task6-IRT-pg1 4280 76 7370 65 5009 63 4295 5125 4952 5381 davide-buscaldi/task6-IRT-wa 4064 81 7278 76 8218 5012 453 | $\begin{array}{c} .4937\\ -0520\\ .4713\\ .4713\\ .4713\\ .4290\\ .4164\\ .4769\\ .3235\\ .3614\\ .3235\\ .3614\\ .3693\\ .3480\\ .0988\\ .3693\\ .3480\\ .1426\\ .1336\\ .2439\\ .5474\\ .1964\\ .4537\\ .1964\\ .4057\\ .5305\\ .3505\end{array}$ |
| baer/task6-UNTOR-1, EGRESSION_BEST FFATURES 7474 13 8.292 12 6.316 10 6.830 8.739 5.280 -0.620 croce/task6-UNTOR-1, REGRESSION_ALL_FEATURES 7474 12 8.292 12 6.316 10 6.838 8.217 5108 6.591 croce/task6-UNTOR-3.REGRESSION_ALL_FEATURES 6.28 31 7.642 59 5.492 51 4.788 6.593 4.835 6.196 daniclecer/stanford_taf1 4.220 77 7160 72 74484 66 .3795 .5350 4.377 6023 daniclecer/stanford_tet ⁺ .5589 55 7.807 55 .4674 67 .3737 65 .4674 61 .4925 .6125 .4925 .5127 .581 .4326 .5833 .3077 davide_buscaldi/task6-IRT-pg1 .4280 76 .7379 65 .4674 .6189 .5102 584 .517 .526 .5474 .1684 .6266 .2245 .5127 <td< td=""><td>$\begin{array}{c} .4713\\ .4713\\ .4713\\ .4290\\ .4164\\ .4769\\ .3235\\ .3614\\ .3693\\ .3480\\ .0988\\ .0988\\ .0988\\ .1426\\ .1336\\ .2439\\ .5474\\ .4055\\ .5305\\ .4753\\ .3505 \end{array}$</td></td<> | $\begin{array}{c} .4713\\ .4713\\ .4713\\ .4290\\ .4164\\ .4769\\ .3235\\ .3614\\ .3693\\ .3480\\ .0988\\ .0988\\ .0988\\ .1426\\ .1336\\ .2439\\ .5474\\ .4055\\ .5305\\ .4753\\ .3505 \end{array}$ |
| croce/task6-UNITOR-2.REGRESSION_ALL_FEATURES 7475 12 827 11 6323 9 5763 8217 5102 6591 croce/task6-UNITOR-3.REGRESSION_ALL_FEATURES_ALL_DOMAINS 6282 31 7642 59 5492 51 4728 6593 4335 6105 danielcer/stanford_Isa† .6354 38 .7217 71 60 21 4747 4804 66 3795 5.535 6468 danielcer/stanford_rte† .5589 55 760 61 5202 58 4.747 4.803 .3333 .3077 davide_buscaldi/task6-RIT-pg1 .4280 76 7.7379 65 5009 61 .5202 58 .4171 .6728 .5179 .5326 davide_buscaldi/task6-RIT-pg1 .4284 81 .6909 81 .2727 7 .1644 .6256 .2214 .1648 demetrics_glinos/task6-ATA-CINK .4976 .746 7 .212 .518 .1847 .212 .6438 .51 .212 .518 .817 .2124 .1648 .212 | $\begin{array}{c} 4713\\ 4713\\ 4290\\ 4164\\ 4769\\ 3235\\ 3614\\ 3693\\ 3480\\ 0988\\ 1426\\ 1336\\ 2439\\ 5474\\ 4057\\ 4453\\ 4537\\ 1964\\ 4057\\ 4465\\ 5305\\ 4753\\ 3505\\ \end{array}$ |
| croce/task6-UNTOR-3-REGRESSION_ALL_FEATURES_ALL_DOMAINS 6289 40 150 21 593 28 4666 80.27 4574 6591 danielcer/stanford_Isaf 6354 38 7212 70 4848 66 3795 5350 4377 6052 danielcer/stanford_all† 4229 77 7160 72 5044 62 4409 4698 4538 6464 danielcer/stanford_all† 4229 77 7160 72 5044 62 4407 4538 6464 davide_buscaldi/task6-IRT-pg3 4818 68 7569 61 5202 58 4326 5333 4856 5317 demetrios_glinos/task6-ATA-BASE 3454 83 6990 81 2772 76 648 2718 769 2950 desotza/task6-FBK-run1 5635 54 712 76 3212 85 1877 6322 4212 65027 344 4521 6507 20 564 4244 4101 1499 4212 demetrios_glinos/task6-ATA-CHNK 4907 </td <td>$\begin{array}{c} 4713\\ 4290\\ 4164\\ 4769\\ 3235\\ 3693\\ 3480\\ 0988\\ 1426\\ 1336\\ 2439\\ 5474\\ 6085\\ 4537\\ 1964\\ 4057\\ 4465\\ 5305\\ 4753\\ 3505\\ \end{array}$</td> | $\begin{array}{c} 4713\\ 4290\\ 4164\\ 4769\\ 3235\\ 3693\\ 3480\\ 0988\\ 1426\\ 1336\\ 2439\\ 5474\\ 6085\\ 4537\\ 1964\\ 4057\\ 4465\\ 5305\\ 4753\\ 3505\\ \end{array}$ |
| csjxu/task6-PolyUCOMP-RUN1 652 31 7.742 59 5492 51 4728 6533 4835 6196 danielcer/stanford_fsa† 6354 31 7.742 70 4848 66 .7375 .5350 .4377 .6052 danielcer/stanford_tre† .5580 56 .7807 55 .4674 67 .4374 .8037 .3533 .3533 .3573 davide-buscald/task6-IRIT-pg1 .4280 76 7.769 61 .5202 58 .4171 .6728 .517 .5526 davide-buscald/task6-IRIT-pg1 .4280 76 7.716 61 .5202 .58 .4171 .6728 .533 .5850 .5350 .4565 .5244 .1648 demetrios_glinos/task6-ATA-CHNK .4976 64 .7120 76 .3212 .58 .188 .6482 .2769 .2550 desouza/task6-FBK-run1 .6633 54 .7120 .712 .76 .322 .8 .1887 .6492 .500 desouza/task6-FBK-run2 .6631 .577 .25 | $\begin{array}{c} .4290\\ .4164\\ .4769\\ .3235\\ .3614\\ .3693\\ .3480\\ .0988\\ .1426\\ .1336\\ .2439\\ .5474\\ .6085\\ .4537\\ .1964\\ .4057\\ .4465\\ .5305\\ .4753\\ .3505 \end{array}$ |
| danielcer/stanford_pdaÅli† 4229 77 7160 72 5044 62 4409 4698 4588 6468 danielcer/stanford_tref 558 55 7807 55 4674 67 4374 8037 5333 3077 davide_buscaldi/task6-IRIT-pg1 4280 76 7779 65 5009 61 5202 58 4171 6728 5179 55 6614 5202 583 3017 davide_buscaldi/task6-IRIT-pg1 4813 68 7569 61 5202 583 3171 6 4806 5335 5317 demetrios_glinos/task6-ATA-BASE 3454 83 6900 81 2728 77 1684 6256 .2244 1648 desouza/task6-FBK-run2 6433 55 7807 26 5888 32 5128 .7807 6328 482 .7127 76 3628 2128 .7807 .6228 .2494 .0171 .4929 .4250 .4251 .6050 .0173 .4419 .6228 .6481 .7127 .7632 | $\begin{array}{c} 4769\\ 3235\\ 3613\\ 3693\\ 3480\\ 0988\\ 1426\\ 1336\\ 2439\\ 5474\\ 6085\\ 4537\\ 1964\\ 4057\\ 4465\\ 5305\\ 4753\\ 3505\\ \end{array}$ |
| danielcer/stanford/ret 558 7807 55 4674 67 4374 8037 3533 3077 davide_buscaldi/task6-IRIT-pg1 4280 76 5009 63 4295 6125 5526 davide_buscaldi/task6-IRIT-wu 44064 81 7877 69 4898 65 4326 5833 4856 5317 demetrios_glinos/task6-ATA-CHNK 4976 64 7160 73 3212 85 1648 6250 2244 1648 demetrios.glinos/task6-ATA-CHNK 4976 64 7160 73 3212 85 1887 6482 2769 2950 desouza/task6-FBK-run1 .5633 54 7127 76 3312 850 7379 622 82 2494 6117 1495 4212 desouza/task6-FBK-run3 .6517 32 8106 25 .5077 20 5169 .7737 419 .5232 .518 .632 .4521 .6050 dvilarinoayala/task6-BUAP-RUN-3 .6610 25 .7744 64 .517 .258 | $\begin{array}{c} 3235\\ 3614\\ 3693\\ 3480\\ 0988\\ 1426\\ 1336\\ 2439\\ 5474\\ 6085\\ 4537\\ 1964\\ 4057\\ 4465\\ 5305\\ 3505\\ \end{array}$ |
| davide.buscaldi/task6-IRIT-pg1 4280 76 7.379 65 5009 63 4295 6.125 4952 5387 davide.buscaldi/task6-IRIT-pg3 4813 68 7.569 61 5.202 58 4171 6728 5516 davide.buscaldi/task6-IRIT-wu 4064 81 7.287 69 4898 65 3.326 5333 4.885 5317 demetrios.glinos/task6-ATA-BASE 3454 83 6900 81 2.772 87 1.684 6.255 1.504 2.735 demetrios.glinos/task6-ATA-STAT 4165 79 7.129 75 3.312 85 1.887 6.482 2.769 2.950 desouza/task6-FBK-run1 .5633 54 7.127 76 3.628 82 2.494 6.117 .495 4.212 desouza/task6-BUAP-RUN-1 .4906 35 .8008 29 .588 32 .518 7.378 .6512 .4521 .6050 dvilarinoayala/task6-BUAP-RUN-1 .4906 351 35 .610 5.99 .4114 .6378 <t< td=""><td>$\begin{array}{c} .3614\\ .3693\\ .3480\\ .0988\\ .1426\\ .1336\\ .2439\\ .5474\\ .6085\\ .4537\\ .1964\\ .4057\\ .4465\\ .5305\\ .5305\\ .3505 \end{array}$</td></t<> | $\begin{array}{c} .3614\\ .3693\\ .3480\\ .0988\\ .1426\\ .1336\\ .2439\\ .5474\\ .6085\\ .4537\\ .1964\\ .4057\\ .4465\\ .5305\\ .5305\\ .3505 \end{array}$ |
| davide buscaldi/ask6-IRT-wu 4064 81 .7287 69 488 65 .4326 .5833 .4856 .5131 demetrios_glinos/task6-ATA-BASE .3454 83 .6990 81 .2772 87 .1684 .6256 .2244 .1648 demetrios_glinos/task6-ATA-STAT .4165 .79 .7129 75 .3312 85 .1887 .6482 .2769 .2950 desouza/task6-FBK-run1 .5633 54 .7127 .76 .6328 82 .2148 .6177 .4149 .6228 desouza/task6-FBK-run2 .6438 .55 .8080 29 .588 32 .5128 .7807 .3796 .6228 dvilarinoayala/task6-BUAP-RUN-1 .4997 63 .5777 20 .516 .917 .0418 .6378 .4785 .6691 enrique/task6-UNED-H34measures .2711 .88 .6694 87 .4286 .278 .4785 .6692 enrique/task6-UNED-H34measures .2711 .88 .6694 87 .4286 .2312 .6303 .625 | .3480 .0988 .1426 .1336 .2439 .5474 .6085 .4537 .1964 .4057 .4465 .5305 .4753 .3505 |
| demetrios.glinos/task6-ATA-BASE .3454 83 .6990 81 .2772 87 .1684 .6256 .2244 .1648 demetrios.glinos/task6-ATA-STAT .4165 79 .7129 75 .3121 85 .1887 .6482 .2769 .2950 desouza/task6-FBK-run1 .5633 54 .7127 76 .3628 82 .2494 .6117 .1495 .4212 desouza/task6-FBK-run2 .6438 53 .800 29 .588 .22 .1548 .7376 .6228 dvilarinoayala/task6-BUAP-RUN-1 .4997 63 .7568 62 .5007 20 .5189 .7773 .4149 .6298 dvilarinoayala/task6-BUAP-RUN-1 .4997 63 .7568 62 .500 .4037 .6532 .4521 .6050 dvilarinoayala/task6-BUAP-RUN-2 .0260 89 .5933 89 .1016 89 .1109 .0577 .334 .458 .6062 enrique/task6-UNED-Hall/measures .7318 .6694 87 .4286 .72 .3610 .312 < | .0988 .1426 .1336 .2439 .5474 .6085 .4537 .1964 .4057 .4465 .5305 .4753 .3505 |
| demetrios_glinos/task6-ATA-CHNK 4976 64 7160 73 3212 85 1504 2735 demetrios_glinos/task6-ATA-STAT 4165 79 7129 75 3312 85 1887 6482 .2769 .2950 desouza/task6-FBK-run1 .5633 54 .7127 76 .3628 82 .2494 .6117 .1495 .4212 desouza/task6-FBK-run3 .6517 32 .8106 25 .6077 20 .5169 .7737 .4419 .6228 dvilarinoayala/task6-BUAP-RUN-1 .4997 63 .7576 62 .5260 56 .4037 .6532 .4521 .6008 dvilarinoayala/task6-BUAP-RUN-2 .0260 .9533 89 .1016 89 .109 .0057 .0348 .1788 dvilarinoayala/task6-UNED-H34measures .2791 88 .6694 87 .4286 72 .3861 .2570 .4086 .5035 .47 .5166 .515 .516 .516 .516 .516 .516 .516 .5166 .516 .516 .5166 </td <td>.1426 .1336 .2439 .5474 .6085 .4537 .1964 .4057 .4465 .5305 .4753 .3505</td> | .1426 .1336 .2439 .5474 .6085 .4537 .1964 .4057 .4465 .5305 .4753 .3505 |
| demetrios.glinos/task6-ATA-STAT 4165 79 7129 75 3312 85 1.887 6.482 2769 .2950 desouza/task6-FBK-run1 .5633 54 .7127 76 .3628 82 .2494 .6117 .1495 .4212 desouza/task6-FBK-run2 .6438 35 .8080 29 .588 32 .5128 .7773 .4419 .6298 dvilarinoayala/task6-BUAP-RUN-1 .4997 63 .7568 62 .5077 20 .518 .4787 .6512 .6501 dvilarinoayala/task6-BUAP-RUN-2 .0260 89 .5933 89 .1016 89 .1109 .0057 .0348 .1788 dvilarinoayala/task6-BUAP-RUN-3 .6630 25 .7474 64 .5105 59 .4018 .6378 .4785 .6691 enrique/task6-UNED-HallMeasures .2791 88 .6694 87 .4286 72 .3861 .2570 .4086 .5006 georgiana_dinu/task6-SAARLAND-ALIGN_VSSIM .4525 .7871 50 .5065 .60 .4043 | .2439 .5474 .6085 .4537 .1964 .4057 .4465 .5305 .4753 .3505 |
| desouza/task6-FBK-run2 6438 35 8080 29 5888 32 5128 7.807 3.796 6228 desouza/task6-FBK-run3 .6517 32 8.106 25 6.077 20 .5169 .773 .4419 .6528 dvilarinoayala/task6-BUAP-RUN-1 .4997 63 .5768 62 .5260 56 .4037 .6532 .4521 .6008 dvilarinoayala/task6-BUAP-RUN-2 .0260 89 .5933 89 .1016 89 .103 .6318 .4758 .6691 enrique/task6-UNED-H34measures .2791 88 .6694 87 .4286 72 .3861 .2570 .4086 .6006 enrique/task6-UNED-H34measures .2791 88 .6694 87 .4286 72 .3861 .2570 .4086 .6006 enrique/task6-UNED-SUNST .4680 69 .7625 .60 .4043 .718 .2686 .5721 georgiana_dinu/task6-SAARLAND-ALIGN_VSSIM .4524 .7818 .6502 .3 .6601 .5753 .6310 .8312 < | .5474 .6085 .4537 .1964 .4057 .4465 .5305 .4753 .3505 |
| desouza/task6-FBK-run3 6517 32 8106 25 607 20 516 7773 4419 6298 dvilarinoayala/task6-BUAP-RUN-1 4997 63 7568 62 5260 56 4037 6532 4521 6050 dvilarinoayala/task6-BUAP-RUN-2 -0260 89 5933 89 1016 89 1109 0057 .0348 .1788 dvilarinoayala/task6-UNED-H3measures .6630 25 .7474 64 .5105 59 .4018 .6378 .4785 .6691 enrique/task6-UNED-H3measures .2791 88 .6694 87 .4286 72 .3861 .2570 .4086 .6006 enrique/task6-UNED-SP-INIST .460 69 .7625 60 .515 47 .5166 .6303 .4625 .6442 georgiana.dinu/task6-SAARLAND-MIXT/VSSIM .4584 71 .8258 13 .5602 .3 .610 .3312 .1391 .5966 jan.snajder/task6-takelab-syntax .8138 2 .8569 3 .6601 5 .6985 <td>.6085 .4537 .1964 .4057 .4465 .5305 .4753 .3505</td> | .6085 .4537 .1964 .4057 .4465 .5305 .4753 .3505 |
| dvilarinoayala/task6-BUAP-RUN-1 .4997 63 .7568 62 .5260 56 .4037 .5532 .4521 .6050 dvilarinoayala/task6-BUAP-RUN-2 .0260 89 .5933 89 .1016 89 .109 .057 .0348 .1788 dvilarinoayala/task6-BUAP-RUN-3 .6630 25 .7474 64 .5105 59 .4018 .5378 .4758 .6692 enrique/task6-UNED-H3dmeasures .2791 88 .6694 87 .4286 72 .3861 .2700 .4006 .6001 georgiana.dinu/task6-SAARLAND-ALIGN-VSSIM .4952 65 .7871 50 .5065 60 .4043 .7171 .2686 .5721 georgiana.dinu/task6-takelab-simple .8133 3 .8625 1 .6753 2 .7343 .8803 .4771 .6797 jan_snajder/task6-takelab-syntax .8138 2 .8569 3 .6601 5 .6985 .8620 .3131 .6040 janardhar/task6-penn-EReg .6212 27 .8048 .34 .5654 <td< td=""><td>.4537 .1964 .4057 .4465 .5305 .4753 .3505</td></td<> | .4537 .1964 .4057 .4465 .5305 .4753 .3505 |
| dvilarino-yala/task6-BUAP-RUN-3 6630 25 7474 64 5105 59 4018 6378 4758 5691 enrique/task6-UNED-H34measures .4381 75 .7518 63 .577 48 .5328 .5788 .4785 .6692 enrique/task6-UNED-SP_INIST .4680 69 .7625 60 .5115 47 .5166 .6303 .4625 .6442 georgiana_dinu/task6-SAARLAND-ALIGN_VSSIM .492 65 .7871 50 .5065 60 .4043 .718 .2686 .5721 georgiana_dinu/task6-SAARLAND-MIXT_VSSIM .4548 71 .8258 13 .5662 43 .6310 .8122 .1391 .5966 janarsnajder/task6-takelab-syntax .8138 2 .8569 3 .6610 .5614 .5504 .3755 .2888 jhaaneha/task6-Penn-ELReg .6422 .27 .8048 34 .562 44 .5480 .7844 .513 .936 .6214 .7049 jhaaneha/task6-Penn-ELReg .6422 .27 .8048 34 .5659< | .4057 .4465 .5305 .4753 .3505 |
| enrique/task6-UNED-H34measures 4381 75 7518 63 577 48 5328 5788 4785 6692 enrique/task6-UNED-HallMeasures .2791 88 6694 87 4286 72 3861 .270 4086 6006 enrique/task6-UNED-SP_INIST .4680 69 .7625 60 .6043 .7718 .2686 .5721 georgiana.dinu/task6-SAARLAND-ALIGN-VSSIM .4952 65 .7871 50 .5662 43 .6310 .812 .1391 .5966 jan.snajder/task6-takelab-simple .8133 3 .8653 1 .6753 2 .7343 .8803 .4771 .6797 jan.snajder/task6-takelab-syntax .8138 2 .8569 3 .6601 .6985 .8620 .3612 .7049 janardhar/task6-penn-ELReg .6622 27 .8048 .481 83 .1936 .5504 .3753 .6604 .5679 .71818 .3547 .560 .4377 .560 .418 .5413 .560 .611 .5613 .562 .5619 | .4465 .5305 .4753 .3505 |
| enrique/task6-UNED-HallMeasures 2.791 88 6694 87 4.286 72 3.861 2.570 4.086 6006 enrique/task6-UNED-SP_INIST 4.660 69 7.625 60 5.615 47 5.166 6.303 .4625 .6421 georgiana_dinu/task6-SAARLAND-ALIGN_VSSIM .4952 65 .7871 50 .5065 60 .4043 .7118 .2066 .5721 georgiana_dinu/task6-SAARLAND-MIXT_VSSIM .4513 71 .8258 13 .5662 43 .6310 .8312 .1391 .5966 jan_snajder/task6-takelab-syntax .8138 2 .8559 3 .6601 5 .6985 .8620 .3612 .7049 janardhan/task6-frane-Elkeg .6622 27 .8043 36 .5594 .3547 .3588 .214 .540 .7857 .3568 .6124 .5460 .7857 .3568 .614 .5469 .5460 .7857 .3568 .6144 .5133 .5406 .5 | .5305 .4753 .3505 |
| enrique/task6-UNED-SP_INIST 4680 69 7.625 60 5.615 47 5.166 6.303 .4625 6.442 georgiana_dinu/task6-SAARLAND-ALIGN_VSSIM .495 65 .7871 50 5.015 60 .4043 .7718 .2686 .5721 georgiana_dinu/task6-SAARLAND-MIXT_VSSIM .4548 71 .8258 13 .5662 43 .6310 .8312 .1391 .5966 janardipet/task6-takelab-simple .8138 2 .8569 3 .6610 5 .6820 .6102 .7049 janardipet/task6-takelab-syntax .8138 2 .8569 3 .6610 .5614 .5504 .3755 .2888 jhaneha/task6-Penn-ELReg .6622 27 .8048 34 .564 44 .5480 .7844 .5516 .7871 .5610 .7857 .3568 .6014 jhasneha/task6-Penn-ELReg .6572 .8083 28 .5757 .571 .5610 .7857 .3568 .6214 jhasneha/task6-SAGAN-RUN1 .5522 .577 .904 47 .50 | .4753 .3505 |
| georgiana_dinu/task6-SAARLAND-MIXT-VSSIM 4548 71 .8258 13 .5662 43 .6310 .8312 .1391 .5966 jan_snajder/task6-takelab-simple .8133 3 .8663 1 .6753 2 .7343 .8803 .4771 .6797 jan_snajder/task6-takelab-syntax .8138 2 .8569 3 .6601 5 .6985 .8620 .3612 .7049 janardhan/task6-takelab-syntax .8138 2 .8659 .8669 .6601 5 .6985 .8620 .3612 .7049 jhasneha/task6-Penn-ELReg .6622 .27 .8048 .84 .5544 .5480 .7844 .3513 .6040 jhasneha/task6-Penn-ELReg .6497 .33 .8043 .36 .5699 .11 .7857 .3568 .6212 jotacastillo/task6-SAGAN-RUN1 .5522 .575 .71 .5400 .7818 .5514 .5519 .7113 .4739 .6542 jotacastillo/task6-SAGAN-RUN2 | |
| jan_smajder/task6-takelab-simple 8133 3 . 8635 1 . 6753 2 .7343 .8803 .4771 .6797 jan_smajder/task6-takelab-syntax 8138 2 . 8569 3 .6601 5 .6985 .8620 .3612 .7049 janardhan/task6-janardhan-UNL_matching .3431 84 .6878 84 .3481 83 .1936 .5504 .3755 .2888 jhasneha/task6-Penn-ELReg .6622 27 .8048 34 .5654 44 .5480 .7844 .3513 .6040 jhasneha/task6-Penn-EReg .6573 28 .8083 28 .5755 37 .5610 .7857 .3568 .6214 jhasneha/task6-Penn-LReg .6497 33 .8043 36 .5699 41 .5460 .7818 .3547 .5969 jotacastillo/task6-SAGAN-RUN1 .5522 57 .7904 47 .5906 29 .5659 .7113 .4739 .6542 jotacastillo/task6-SAGAN-RUN2 .6272 42 .8032 37 .5838 34 .5538 .7706 .4480 .6135 jotacastillo/task6-SAGAN-RUN3 .6311 39 .7943 45 .5649 46 .5394 .7560 .4181 .5904 Konstantin_Z/task6-ABBYY-General .5636 53 .8052 33 .5759 36 .4797 .7821 .4576 .6488 M.Rios/task6-UOW-LEX_PARA .6397 36 .7187 71 .3825 80 .3628 .6426 .3074 .2806 M.Rios/task6-UOW-LEX_PARA .5981 49 .6955 82 .3473 84 .3529 .5724 .3066 .2643 M.Rios/task6-UOW-LEX_PARA .5981 49 .6257 88 .2567 88 .2995 .2910 .1611 .2571 mheilman/task6-ETS-PERP | 2006 |
| jan.snajder/task6-takelab-syntax 8138 2 8569 3 6001 5 6985 8.620 3.612 7.049 janardhan/task6-janardhan-UNL_matching 3.431 84 6.878 84 3.481 83 1.936 5.504 3.755 2.888 jhasneha/task6-Penn-EReg 6.622 27 8.048 34 5.554 4 5.540 3.755 2.888 jhasneha/task6-Penn-EReg 6.673 28 8.043 26 5.559 37 5.610 7.857 3.568 6.214 jhasneha/task6-SAGAN-RUN1 5.522 57 7.904 47 5.906 29 5.659 7.113 4.739 6.542 jotacastillo/task6-SAGAN-RUN2 6.272 42 8.023 37 5.838 34 5.538 7.706 4.480 6.135 jotacastillo/task6-SAGAN-RUN2 6.272 42 8.032 37 5.838 34 5.554 46 5.394 7.560 4.418 5.904 Konstantin_Z/task6-ABBYY-General 5.636 53 8.052 33 5.759 36 4.797 7.821 4.576 6.488 M.Rios/task6-UOW-LEX_PARA_SEM 5.981 49 6.955 82 3.473 84 3.529 5.724 3.066 2.643 M.Rios/task6-UOW-LEX_PARA_SEM 5.981 49 6.955 82 2.3473 84 3.529 5.724 3.066 2.643 M.Rios/task6-UOW-LEX_PARA_SEM 5.961 79 6.955 88 2.967 88 2.957 88 2.951 2.910 1.611 2.571 mheilman/task6-ETS-PERP | .3806 .3989 |
| jhasneha/task6-Penn-ELReg 6622 27 8048 34 5654 44 5480 7.844 .3513 6040 jhasneha/task6-Penn-EReg .6573 28 8083 28 5755 37 .5610 .7857 .3568 .6214 jhasneha/task6-Penn-LReg .6497 33 .8043 36 .5699 41 5460 .7857 .3568 .6214 jotacastillo/task6-SAGAN-RUN1 .5522 57 .7904 47 .5906 29 .5659 .7113 .4739 .6542 jotacastillo/task6-SAGAN-RUN2 .6272 42 .8032 37 .5838 34 .5538 .7706 .4480 .6135 jotacastillo/task6-SAGAN-RUN3 .6311 39 .7943 45 .5649 46 .5394 .7560 .4181 .5904 Konstantin.Z/task6-ABBYY-General .5636 53 .8052 33 .5759 36 .4797 .7821 .4576 .6488 M.Rios/task6-UOW-LEX_PARA .6397 36 .7187 71 .3825 80 .3628 .6426 .3074 .2806 M.Rios/task6-UOW-LEX_PARA .5981 49 .6955 82 .3473 84 .3529 .5724 .3066 .2643 M.Rios/task6-UOW-SEM .5361 59 .6287 88 .2567 88 .2995 .2910 .1611 .2571 mheilman/task6-ETS-PERP | .4683 |
| jhasneha/task6-Penn-ERRe [*] 6573 28 8083 28 5755 37 5610 7.857 .3568 6214 jhasneha/task6-Penn-LReg 6.6497 33 8043 36 5699 41 5400 7.818 .3547 5969 jotacastillo/task6-SAGAN-RUN1 5522 57 7.904 47 5906 29 5659 7.113 .4739 6542 jotacastillo/task6-SAGAN-RUN2 6272 42 8032 37 5838 34 5538 7.706 .4480 .6135 jotacastillo/task6-SAGAN-RUN2 6311 39 7.943 45 5549 46 5394 7.560 .4480 .6135 Konstantin_Z/task6-ABBYY-General 5636 53 .8052 33 .5759 36 .4797 .7821 .4576 .6488 M.Rios/task6-UOW-LEX_PARA 6397 36 .7187 71 .3825 80 .3628 .6426 .3074 .2806 M.Rios/task6-UOW-LEX_PARA.5EM 5981 49 .6955 82 .3473 84 .3529 .5724 .3066 .2643 M.Rios/task6-UOW-EEX_PARA.5EM 7361 59 .6287 88 .2957 88 .2959 .2910 .1611 .2571 mheilman/task6-ETS-PERP 780 7 .808 7 .8089 27 .6399 7 .6397 .720 .4470 .7124 | .3387 |
| jhasneha/task6-Penn-LReg .6497 33 .8043 36 .5699 41 .5400 .7818 .3547 .5969 jotacastillo/task6-SAGAN-RUN1 .5522 .57 .7904 47 .5906 29 .5659 .7113 .4739 .6542 jotacastillo/task6-SAGAN-RUN2 .6272 42 .8032 37 .5838 34 .5538 .7706 .4480 .6135 jotacastillo/task6-SAGAN-RUN3 .6311 39 .7943 45 .5649 46 .5394 .7560 .4181 .5904 Konstantin_Z/task6-ABBYY-General .5636 .5805 .38 .5759 .6477 .7821 .4576 .6488 M.Rios/task6-UOW-LEX_PARA .6397 .6377 .787 .718 .3825 80 .3628 .6262 .3044 .2806 M.Rios/task6-UOW-LEX_PARA_SEM .5981 49 .6955 82 .3473 84 .3529 .5724 .3066 .2643 M.Rios/task6-UOW-SEM .5361 59 .6287 88 .2567 .88 .2995 .2910 | .3607 .3732 |
| jotacastillo/task6-SAGAN-RUN1 5522 57 7904 47 5906 29 5659 7.113 4739 6542 jotacastillo/task6-SAGAN-RUN2 627 42 8032 37 588 34 5538 7706 4480 6135 jotacastillo/task6-SAGAN-RUN3 6311 39 7943 45 584 46 5394 7560 44181 5904 Konstantin_Ztask6-ABBYY-General 5636 53 8052 33 5759 36 4797 7821 4576 6488 M_Rios/task6-UOW-LEX_PARA 6397 36 718 71 3825 80 3628 6426 3074 2806 M_Rios/task6-UOW-LEX_PARA,SEM 5981 49 6955 82 3473 84 3529 5724 3066 2643 M_Rios/task6-UOW-SEM 5361 59 6287 88 2957 88 2959 2910 .1611 2571 mheilman/task6-ETS-PERP 786 788 7 8064 32 6305 11 6211 7210 4722 7080 mheilman/task6-ETS-PERPptnases 7834 6 8089 27 6399 7 6397 7200 4850 7.124 | .4137 |
| jotacastillo/task6-SAGAN-RUN3 .6311 39 .7943 45 .5649 46 .5394 .7560 .4181 .5904 Konstantin_Z/task6-ABBYY-General .5636 53 .8052 33 .5759 36 .4797 .7821 .4576 .6488 M.Rios/task6-UOW-LEX_PARA .6397 36 .7187 71 .3825 80 .3628 .6426 .3074 .2806 M.Rios/task6-UOW-LEX_PARA_SEM .5981 49 .6955 82 .3473 84 .3529 .5724 .3066 .2643 M.Rios/task6-UOW-SEM .5361 59 .6287 88 .2957 88 .2995 .2910 .1611 .2571 mheilman/task6-ETS-PERP .7808 7 .8064 32 .6305 11 .6211 .7210 .4722 .7080 | .4253 |
| Konstantin Z/task6-ABBYY-General 5636 53 8052 33 5759 36 479 7821 4576 6488 M.Rios/task6-UOW-LEX_PARA .6397 36 .7187 71 .3828 80 .3628 .6426 .3074 .2806 M.Rios/task6-UOW-LEX_PARA_SEM .5981 49 .6955 82 .3473 84 .3529 .5724 .3066 .2643 M.Rios/task6-UOW-SEM .5361 59 .6287 88 .2567 88 .2910 .1611 .2571 mheilman/task6-ETS-PERP .7804 .8064 32 .6305 11 .6211 .7210 .4722 .7080 mheilman/task6-ETS-PERPptrases .7834 6 .8089 27 .6397 .7200 .4820 .7120 .7120 | .3894 .3746 |
| M_Rios/task6-UOW-LEX_PARA .6397 36 .7187 71 .3825 80 .3628 .6426 .3074 .2806 M_Rios/task6-UOW-LEX_PARA_SEM .5981 49 .6955 82 .3473 84 .3529 .5724 .3066 .2643 M_Rios/task6-UOW-SEM .5361 59 .6287 88 .2567 88 .2910 .1611 .2571 mheilman/task6-ETS-PERP .7804 32 .6305 1 .6211 .7210 .4722 .7080 mheilman/task6-ETS-PERPphrases .7834 6 .8089 27 .6399 7 .6397 .7200 .4850 .7124 | .3682 |
| M_Rios/task6-UOW-SEM .5361 59 .6287 88 .2567 88 .2995 .2910 .1611 .2571 mheilman/task6-ETS-PERP .7808 7 .8064 32 .6305 11 .6211 .7210 .4722 .7080 mheilman/task6-ETS-PERP .7834 6 .8089 27 .6399 7 .6397 .7200 .4850 .7124 | .2082 |
| mheilman/task6-ETS-PERP .7808 7 .8064 32 .6305 11 .6211 .7210 .4722 .7080 mheilman/task6-ETS-PERPphrases .7834 6 .8089 27 .6399 7 .6397 .7200 .4850 .7124 | .1164 |
| mheilman/task6-ETS-PERPphrases .7834 6 .8089 27 .6399 7 .6397 .7200 .4850 .7124 | .2212 .5149 |
| $mhailman hash \in ETC TED_n$ $1.4477 = 72 + 7001 = 60 + 5052 = 57 + 5040 = 5017 + 4740 = 6160$ | .5312 |
| mheilman/task6-ETS-TERp | .4566 |
| nitish_aggarwal/task6-aggarwal-run1 * .5777 52 .8158 20 .5466 52 .3675 .8427 .3534 .6030 nitish_aggarwal/task6-aggarwal-run2 * .5833 51 .8183 17 .5683 42 .3720 .8330 .4238 .6513 | .4430 .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 .4918 66 .7646 58 .5061 61 .4989 .7092 .4437 .4879 | .3446 .2441 |
| parthapatray/task6-JU_CSE_NLP-Sematic_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 rada/task6-UNT-IndividualRegression .7846 5 .8440 6 .6162 13 .5353 .8750 .4203 .6715 | .2256 .4033 |
| rada/task6-UNT-IndividualRegression .7846 5 .8440 6 .6162 13 .5353 .8750 .4203 .6715 sbdlrhmn/task6-sbdlrhmn-Run1 .6663 23 .7842 53 .5376 54 .5440 .7335 .3830 .5860 | .4035 |
| sbdlrhmn/task6-sbdlrhmn-Run2 .4169 78 .7104 77 .4986 64 .4617 .4489 .4719 .6353 | .4353 |
| sgjimenezv/task6-SOFT-CARDINALITY 7.7331 15 8526 5 6708 3 6405 8562 5152 7.109 sgjimenezv/task6-SOFT-CARDINALITY-ONE-FUNCTION 7.107 19 8397 8 6486 6 6316 8237 4320 7.109 | .4833 |
| sgjimenezv/task6-SOFT-CARDINALITY-ONE-FUNCTION siva/task6-DSS-alignheuristic | .4833 .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 sokolov/task6-LIMSI-cosprod .6392 37 .7344 67 .3940 78 .3948 .6597 .0143 .4157 | .3465 .2889 |
| sokolovitaskó-LIMSI-gradtree 6.782 27 7.7 66 4.118 75 4.848 6.636 0.934 4.370 | .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 spirin2/task6-UIUC-MLNLP-CCM .7269 16 .8217 16 .6104 17 .5769 .8203 .4667 .5835 | .4603 .4945 |
| spirinZ/ask6-010C-ML/NLP-CCWi | .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 tiantianzhu7/task6-tiantianzhu7-1 .4533 72 .7134 74 .4192 73 .4184 .5630 .2083 .4822 | .3968 .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*† | .4383 |
| yehrasko-SRUBC-SYSTEM2† (7562 10 8111 24 5858 33 6050 7939 4.294 5871 | |
| yeh/task6-SRIUBC-SYSTEM3† .6876 21 .7812 54 .4668 68 .4791 .7901 .2159 .3843 | .3994 .3366 |
| yguiterrez/task6-UMCC_DLSI-MultiLex 6630 26 7922 46 5560 49 6022 7709 4435 4327 | .3366 .2801 |
| ygutierrez/task6-UMCC.DLSI-MultiSem 6529 29 8115 23 6116 16 5269 7756 4688 6539 ygutierrez/task6-UMCC.DLSI-MultiSemLex 7213 18 8239 14 6158 15 6205 8104 4325 6256 | .3366 .2801 .4264 |
| yrkakde/task6-yrkakde-DiceWordnet .5977 50 .7902 48 .5742 39 .5294 .7470 .5531 .5698 | .3366 .2801 |
| yrkakde/task6-yrkakde-JaccNERPenalty .6067 47 .8078 31 .5955 24 .5757 .7765 .4989 .6257 | .3366 .2801 .4264 .5470 |

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, \dagger team involving one of the organizers.

| 80. | | | | | monum | own cui | Sivi i-cui w | OII-WIN | $OII-WIN_W$ | SWIT-news | SMT-news w |
|---------------|--|--|---|---|--|--|---|--|--|--|--|
| ου · | .4946 | .4295 | .4082 | .6125 | .6593 | .4952 | .5273 | .5387 | .5574 | .3614 | .4674 |
| 13. | .5503 | .4171 | .4033 | .6728 | .7048 | .5179 | .5529 | .5526 | .5950 | .3693 | .4648 |
| 64. | .4682 | .4326 | .4035 | .5833 | .6253 | .4856 | .5138 | .5317 | .5189 | .3480 | .4482 |
| 81. | .2615 | .5328 | .4494 | .5788 | .4913 | .4785 | .4660 | .6692 | .6440 | .4465 | .3632 |
| 91. | .2002 | .3861 | .3802 | .2570 | .2343 | .4086 | .4212 | .6006 | .5947 | .5305 | .4858 |
| 80. | .3754 | .5166 | .5082 | .6303 | .5588 | .4625 | .4801 | .6442 | .5761 | .4753 | .4143 |
| 80. | .3636 | .3427 | .3498 | .3549 | .3353 | .4271 | .3989 | .5298 | .4619 | .4034 | .3228 |
| 33. | .5442 | .4184 | .4241 | .5630 | .5630 | .2083 | .4220 | .4822 | .5031 | .2745 | .3536 |
| 57. | .5249 | .4260 | .4340 | .5628 | .5758 | .1546 | .4776 | .4552 | .4926 | .1923 | .3362 |
| 46. | .5229 | .3411 | .3611 | .5946 | .5899 | .1868 | .4769 | .4029 | .4365 | .1823 | .4014 |
| 0 2 9 2 2 3 3 | 54 81 91 80 80 83 57 | 54 .4682 81 .2615 91 .2002 80 .3754 80 .3636 33 .5442 | 54 .4682 .4326 81 .2615 .5328 91 .2002 .3861 80 .3754 .5166 80 .3636 .3427 33 .5442 .4184 57 .529 .4260 | 54 .4682 .4326 .4035 11 .2615 .5328 .4494 11 .2002 .3861 .3802 30 .3754 .5166 .5082 30 .3636 .3427 .3498 33 .5442 .4184 .4241 37 .5249 .4260 .4340 | 54 .4682 .4326 .4035 .5833 11 .2615 .5328 .4494 .5788 11 .2002 .3861 .3802 .2570 30 .3754 .5166 .5082 .6303 30 .3363 .3427 .3498 .3549 33 .5442 .4184 .4241 .5630 77 .5249 .4260 .4340 .5628 | 54 .4682 .4326 .4035 .5833 .6253 \$1 .2615 .5328 .4494 .5788 .4913 \$1 .2015 .5328 .4494 .5788 .4913 \$1 .2015 .5328 .4494 .5788 .4913 \$1 .2015 .5328 .4494 .5788 .913 \$0 .3754 .5166 .5082 .6303 .5588 \$0 .3636 .3427 .3498 .3549 .3353 \$3 .5442 .4184 .4241 .5630 .5630 \$7 .5249 .4260 .4340 .5628 .5788 | 54 .4682 .4326 .4035 .5833 .6253 .4856 \$1 .2615 .5328 .4494 .5788 .4913 .4785 \$1 .2015 .5328 .4494 .5788 .4913 .4785 \$1 .2012 .3861 .3802 .2570 .2343 .4086 \$0 .3754 .5166 .5082 .6303 .5588 .4625 \$0 .3636 .3427 .3498 .3549 .3353 .4271 \$3 .5442 .4184 .4241 .5630 .2083 .7084 \$7 .5249 .4260 .4340 .5628 .5758 .1546 | 54 .4682 .4326 .4035 .5833 .6253 .4856 .5138 11 .2615 .5328 .4494 .5788 .4913 .4785 .4660 11 .2012 .3861 .3802 .2570 .2343 .4086 .4212 30 .3754 .5166 .5082 .6303 .5588 .4025 .4801 30 .3354 .3407 .3498 .3549 .3353 .4271 .3989 30 .35.42 .4184 .4241 .5630 .5630 .2083 .4220 31 .5442 .4260 .4344 .5638 .5788 .1546 .4776 | 54 .4682 .4326 .4035 .5833 .6253 .4856 .5138 .5317 11 .2615 .5328 .4494 .5788 .4913 .4785 .4660 .6692 10 .2002 .3861 .3802 .2570 .2343 .4086 .4212 .6006 30 .3754 .5166 .5082 .6303 .5588 .4625 .4801 .6422 30 .3636 .3427 .3498 .3549 .3353 .4271 .3989 .5298 33 .5442 .4184 .4241 .5630 .5630 .2083 .4220 .4822 71 .5249 .4260 .4340 .5528 .5758 .1546 .4776 .4552 | 54 .4682 .4326 .4035 .5833 .6253 .4856 .5138 .5317 .5189 11 .2615 .5328 .4494 .5788 .4913 .4785 .4660 .6692 .6440 11 .2015 .5328 .4494 .5788 .4913 .4785 .4660 .6692 .6440 11 .2002 .3861 .3802 .2570 .2343 .4086 .4212 .6006 .5947 30 .3754 .5166 .5082 .6303 .5588 .4625 .4801 .6442 .5761 30 .3354 .3427 .3498 .3353 .4271 .3989 .5298 .4619 33 .5442 .4184 .4241 .5630 .5638 .2271 .3989 .5298 .4619 33 .5442 .4184 .4260 .4340 .5628 .5788 .1546 .4776 .4552 .4926 70 .2420 .4260 | 54 .4682 .4326 .4035 .5833 .6253 .4856 .5138 .5317 .5189 .3480 11 .2615 .5328 .4494 .5788 .4913 .4785 .4660 .6692 .6440 .4465 11 .2015 .5328 .4494 .5788 .4913 .4785 .4660 .6692 .6440 .4465 10 .2002 .3861 .3802 .2570 .2343 .4086 .4212 .6006 .5947 .5305 30 .3754 .5166 .5082 .6303 .5588 .4625 .4801 .6442 .5761 .4753 30 .3363 .3427 .3498 .33549 .3353 .4271 .3989 .5298 .4619 .4034 33 .5442 .4184 .4241 .5630 .2083 .4220 .4812 .5031 .2745 37 .5249 .4260 .4340 .5628 .5738 .1546 .477 |

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

⁵http://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 | | | | MT evaluation | MWE Named Entity recognition | POS tagger | Semantic Role Labeling | | | Syntax | Textual entailment | Word Sense Disambiguation | - |
|--|-------------|--------------|--------------------------|---------------------|----------------------|------------|-----------------------|-----------|---------|-----------|---------------------------|---------------|--------|--------|--------|---------------|---------------------------------|------------|------------------------|-----------|--------------|--------------|--------------------|---------------------------|-----------|
| aca08ls/task6-University_Of_Sheffield-Hybrid | | | | | | | | | х | | | х | х | | х | | | _ | _ | | x | _ | _ | X | |
| aca08ls/task6-University_Of_Sheffield-Machine_Learning aca08ls/task6-University_Of_Sheffield-Vector_Space | | + | - | | | _ | _ | | x x | \square | | X X | X X | _ | х | + | + | _ | ┢ | \vdash | x | + | + | X X | |
| baer/task6-UKP-run1 | ╢ | x | x | х | х | - | | х | X | \square | х | | X | - | | + | + | x | + | x | x | + | x | +^ | x |
| baer/task6-UKP-run2_plus_postprocessing_smt_twsi | | x | х | х | х | | | х | х | | х | х | х | | | 1 | T | x | t | | x | T | х | + | x |
| baer/task6-UKP-run3_plus_random | | X | x | | х | | | X | х | | х | х | х | | | | | x | | x | x | _ | x | \Box | x |
| croce/task6-UNITOR-1_REGRESSION_BEST_FEATURES croce/task6-UNITOR-2_REGRESSION_ALL_FEATURES | | + | - | x | | _ | | | | \square | x | | X X | | x | + | + | X | - | \vdash | _ | x | + | + | \vdash |
| croce/task6-UNITOR-3_REGRESSION_ALL_FEATURES_ALL_DOMAIN | s | + | - | X X | | - | | | | | X X | | X | _ | x x | + | + | X | | ++ | _ | x x | + | + | + |
| csjxu/task6-PolyUCOMP-RUN | | + | | | | | | x | х | | ~ | х | ~ | | ~ | + | + | x | _ | ++ | + | + | + | + | + |
| danielcer/stanford_fsa | | T | | | | | х | | х | | | | х | х | | Х | | x | T | | | x | | \pm | \square |
| danielcer/stanford_pdaAll | | | | | | | x | | х | | | | х | x | | X | | x | | \square | | x | _ | \perp | \square |
| danielcer/stanford_rte davide_buscaldi/task6-IRIT-pg1 | | + | | x | | _ | _ | | x x | x | | x | x | _ | х | 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 | | | | | | | | | х | | х | х | | | | | | x | | | x | | | | x |
| demetrios_glinos/task6-ATA-CHNK | | | | | | | | | х | | х | х | | _ | | | | x | | | x | \downarrow | _ | ╞ | x |
| demetrios_glinos/task6-ATA-STAT desouza/task6-FBK-run1 | , | - | - | x | | _ | x | x | x x | | X X | X X | х | x | | + | , | x x | | ++ | x | x | + | + | X X |
| desouza/task6-FBK-run2 | | + | | X | | - | X | X | X | | х | X | X | | | + | ť | X | | ++ | ÷ | + | + | + | + |
| desouza/task6-FBK-run3 | | $^{+}$ | | x | | | | x | x | | х | | | | | + | $^{+}$ | x | - | Ħ | + | + | - | + | \square |
| dvilarinoayala/task6-BUAP-RUN-1 | | X | | | | | | | | | | | х | | | | | | | | | | | | |
| dvilarinoayala/task6-BUAP-RUN-2 | | | | | | | | | х | | | | _ | | | | \perp | | | \square | \downarrow | _ | _ | \perp | \square |
| dvilarinoayala/task6-BUAP-RUN-3 jan_snajder/task6-takelab-simple | | - | ~ | v | | × | _ | v | v | | v | х | X X | v | v | + | + | - | ╞ | ++ | + | + | _ | + | X X |
| jan_snajder/task6-takelab-syntax | | x | x | X X | | X | - | x | x x | | X X | X | x | X X | х | + | , | X X | _ | ++ | + | x | + | + | |
| janardhan/task6-janardhan-UNL_matching | ╢ | + | | - | | - | | | x | \square | - | ~ | x | ~ | | + | , | | | ++ | | x | + | x | + |
| jotacastillo/task6-SAGAN-RUN1 |) | : | | х | | | | | x | | | | х | х | | | T | T | T | Ħ | | | x x | ĸх | \square |
| jotacastillo/task6-SAGAN-RUN2 | > | | | x | | | | | x | | | | х | х | | | | | \perp | \square | \perp | | X X | | |
| jotacastillo/task6-SAGAN-RUN3 Konstantin_Z/task6-ABBYY-General | , | - | - | x | | _ | _ | | x | \square | | | х | x | | + | + | _ | ╞ | \vdash | + | + | x x | x x | \vdash |
| M_Rios/task6-UOW-LEX_PARA | | + | x | | | _ | - | | | | х | | x | _ | | x | , | x | x | ++ | + | x | + | +- | + |
| M_Rios/task6-UOW-LEX_PARA_SEM | | + | x | | | | | | | | x | | х | | | x | , | | | + + | _ | x | - | + | + |
| M_Rios/task6-UOW-SEM | | | х | | | | | | | | х | | х | | | | , | ι x | Х | | | x | | | |
| mheilman/task6-ETS-PERP | | | | х | | _ | | | х | | | х | х | х | | x | | _ | \vdash | | x | + | _ | + | \square |
| mheilman/task6-ETS-PERPphrases mheilman/task6-ETS-TERp | | + | - | X X | X X | _ | _ | | x x | \square | | X X | X X | x x | - | x | + | _ | ┢ | \vdash | x | + | + | + | X X |
| parthapakray/task6-JU_CSE_NLP-Semantic_Syntactic_Approach | - | | | • | ^ | х | - | | х | \vdash | _ | ^ | х | ^ | | × | | x x | + | ++ | + | x | x | + | X |
| rada/task6-UNT-CombinedRegression | | T | | | | | | х | х | х | х | х | х | | х | | T | + | t | Ħ | + | Ŧ | - | x | x |
| rada/task6-UNT-IndividualDecTree | | | | | | | | | х | х | х | х | х | | х | | | | | | \square | | | X | |
| rada/task6-UNT-IndividualRegression sgjimenezv/task6-SOFT-CARDINALITY | | + | - | | | | | x | x | х | x | X | X X | _ | х | + | + | _ | ╞ | \vdash | + | + | + | x | x |
| sgjimenezv/task6-SOFT-CARDINALITY-ONE-FUNCTION | | + | - | | | X X | - | | | | | X X | x | - | | + | + | | + | ++ | + | + | + | + | + |
| skamler_/task6-EHU-RUN1v2 | | + | | | | ~ | x | | | х | _ | ~ | x | | | + | + | x | + | x | + | + | + | + | + |
| sokolov/task6-LIMSI-cosprod | | T | | х | | | | | | | х | | х | | х | | T | | T | | | T | | | \square |
| sokolov/task6-LIMSI-gradtree | | | | x | | | | | | | x | | х | | х | | | | \perp | Ш | \perp | | _ | \perp | \square |
| sokolov/task6-LIMSI-sumdiff spirin2/task6-UIUC-MLNLP-Blend | ╢. | + | - | X X | | _ | x | x | | | х | | х | _ | х | - | | x | - | \vdash | + | x | x | + | x |
| spirin2/task6-UIUC-MLNLP-CCM | > | _ | | X | | _ | x | X | | | | | _ | _ | | X | () () | | - | | _ | _ | x | +- | X |
| spirin2/task6-UIUC-MLNLP-Puzzle | , | | | x | | | x | x | | | | | | | | X | | | | | | | x | + | x |
| sranjans/task6-sranjans-1 | | | | х | | | | х | х | | х | х | | | | | | x | T | | 1 | 1 | | x | \square |
| sranjans/task6-sranjans-2 | | | | х | | | | x | х | | х | _ | х | х | | | > | _ | | \square | \perp | _ | | x x | \square |
| sranjans/task6-sranjans-3 tiantianzhu7/task6-tiantianzhu7-1 | | + | - | x | | _ | _ | x | X X | | х | X X | x | x | | + | , | x x | - | \vdash | + | x | x | x | \vdash |
| tiantianzhu7/task6-tiantianzhu7-1 | ╢ | + | + | - | | - | - | | X | \square | _ | Х | - | - | | + | + | X | _ | ++ | + | + | + | +^ | + |
| tiantianzhu7/task6-tiantianzhu7-3 | \parallel | + | 1 | 1 | | х | | | x | Η | | x | | + | | + | + | x | - | + | + | + | + | + | \square |
| weiwei/task6-weiwei-run1 | | x | - | х | | | | | х | | х | | х | | | | | x | | \square | | | | T | \square |
| yeh/task6-SRIUBC-SYSTEM1 | \parallel | 1 | X | | \square | | | X | X | Ц | x | X | X | _ | - | | 1 | X | _ | \square | \downarrow | \downarrow | \perp | + | \square |
| yeh/task6-SRIUBC-SYSTEM2 yeh/task6-SRIUBC-SYSTEM3 | \parallel | + | X X | - | | _ | _ | X X | X X | \square | X X | X X | X X | - | + | + | + | X | - | ++ | + | + | + | + | + |
| ygutierrez/task6-UMCC_DLSI-MultiLex | + | + | ^ | x | | | | ^ | X | х | A | | X | + | + | + | + | X | - | ++ | + | + | + | + | x |
| ygutierrez/task6-UMCC_DLSI-MultiSem | | + | | x | | | | | x | | | | х | | | + | t | X | _ | \square | + | + | + | x | <u>+</u> |
| ygutierrez/task6-UMCC_DLSI-MultiSemLex | | | | х | | | | | х | х | | х | х | | | | | x | | \square | | | | х | X |
| yrkakde/task6-yrkakde-DiceWordnet | | | | | | | | | x | | x | | x | | | | | | | | | - I | | 1 | 1 |

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 se-

mantic 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.