SemEval-2014 Task 10: Multilingual Semantic Textual Similarity

Eneko Agirre[a], Carmen Banea[b*], Claire Cardie[c], Daniel Cer[d], Mona Diab[e*]
Aitor Gonzalez-Agirre[a], Weiwei Guo[f], Rada Mihalcea[b], German Rigau[a], Janyce Wiebe[g]

[a]University of the Basque Country
Basque Country, Spain

[b]University of Michigan
Ann Arbor, MI

[c]Cornell University
Ithaca, NY

[d]Google Inc.
Mountain View, CA

[e]George Washington University
Washington, DC

[f]Columbia University
New York, NY

[g]University of Pittsburgh
Pittsburgh, PA

Abstract

In Semantic Textual Similarity, systems rate the degree of semantic equivalence between two text snippets. This year, the participants were challenged with new data sets for English, as well as the introduction of Spanish, as a new language in which to assess semantic similarity. For the English subtask, we exposed the systems to a diversity of testing scenarios, by preparing additional OntoNotes-WordNet sense mappings and news headlines, as well as introducing new genres, including image descriptions, DEFT discussion forums, DEFT newswire, and tweet-newswire headline mappings. For Spanish, since, to our knowledge, this is the first time that official evaluations are conducted, we used well-formed text, by featuring sentences extracted from encyclopedic content and newswire. The annotations for both tasks leveraged crowdsourcing. The Spanish subtask engaged 9 teams participating with 22 system runs, and the English subtask attracted 15 teams with 38 system runs.

1 Introduction and motivation

Given two snippets of text, Semantic Textual Similarity (STS) captures the notion that some texts are more similar than others, measuring their degree of semantic equivalence. Textual similarity can range from complete unrelatedness to exact semantic equivalence, and a graded similarity intuitively captures the notion of intermediate shades of similarity, as pairs of text may differ from some minor nuanced aspects of meaning, to relatively important semantic differences, to sharing only some details, or to simply being related to the same topic (cf. Section 2).

One of the goals of the STS task is to create a unified framework for combining several semantic components that otherwise have historically tended to be evaluated independently and without characterization of impact on NLP applications. By providing such a framework, STS allows for an extrinsic evaluation of these modules. Moreover, such an STS framework itself could in turn be evaluated intrinsically and extrinsically as a grey/black box within various NLP applications such as Machine Translation (MT), Summarization, Generation, Question Answering (QA), etc.

STS is related to both Textual Entailment (TE) and Paraphrasing, but differs in a number of ways and it is more directly applicable to a number of NLP tasks. STS is different from TE inasmuch as it assumes bidirectional 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. STS also differs from both TE and Paraphrasing (in as far as both tasks have been defined to date in the literature) in that, rather than being a binary yes/no decision (e.g. *a vehicle is not a car*), we define STS to be a graded similarity notion (e.g. *a vehicle* and *a car* are more similar than *a wave* and *a car*). A quantifiable graded bidirectional notion of textual similarity is useful for a myriad of NLP tasks such as MT evaluation, information extraction, question answering, summarization, etc.

In 2012 we held the first pilot task at SemEval 2012, as part of the *SEM 2012 conference, with great success: 35 teams participated with 88 system runs (Agirre et al., 2012). In addition, we held

*carmennb@umich.edu, mtdiab@gwu.edu

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 81–91,
Dublin, Ireland, August 23-24, 2014.

year	dataset	pairs	source
2012	MSRpar	1500	newswire
2012	MSRvid	1500	videos
2012	OnWN	750	glosses
2012	SMTnews	750	MT eval.
2012	SMTeuroparl	750	MT eval.
2013	HDL	750	newswire
2013	FNWN	189	glosses
2013	OnWN	561	glosses
2013	SMT	750	MT eval.
2014	HDL	750	newswire headlines
2014	OnWN	750	glosses
2014	Deft-forum	450	forum posts
2014	Deft-news	300	news summary
2014	Images	750	image descriptions
2014	Tweet-news	750	tweet-news pairs

Table 2: English subtask: Summary of train (2012 and 2013) and test (2014) datasets.

a DARPA sponsored workshop at Columbia University.[1] In 2013, STS was selected as the official Shared Task of the *SEM 2013 conference, with two subtasks: The Core task, which is similar to the 2012 task; and a Pilot task on Typed-similarity between semi-structured records. The Core task attracted 34 participants with 89 runs, and the Typed-similarity task attracted 6 teams with 14 runs.

For STS 2014 we defined two subtasks: English and Spanish. For the English subtask we provided five test datasets: two datasets that extend already released genres (the OntoNotes-WordNet sense mappings and news headlines) and three new genres: image descriptions, DEFT discussion forum data and newswire, as well as tweet-newswire headline mappings. Participants could use all datasets released in 2012 and 2013 as training data. The Spanish subtask introduced two diverse datasets on different genres, namely encyclopedic descriptions extracted from the Spanish Wikipedia and contemporary Spanish newswire. For the Spanish subtask, the participants had access to a limited amount of labeled data, consisting of 65 sentence pairs, which they could use for training.

2 Task Description

2.1 English Subtask

The English dataset comprises pairs of news headlines (HDL), pairs of glosses (OnWN), image descriptions (Images), DEFT-related discussion forums (Deft-forum) and news (Deft-news), and tweet comments and newswire headline mappings (Tweets).

For **HDL**, we used naturally occurring news headlines gathered by the Europe Media Monitor (EMM) engine (Best et al., 2005) from several different news sources. EMM clusters together related news. Our goal was to generate a balanced data set across the different similarity ranges, hence we built two sets of headline pairs: (i) a set where the pairs come from the same EMM cluster, (ii) and another set where the headlines come from a different EMM cluster, then we computed the string similarity between those pairs. Accordingly, we sampled 375 headline pairs of headlines that occur in the same EMM cluster, aiming for pairs equally distributed between minimal and maximal similarity using simple string similarity. We sampled other 375 pairs from the different EMM cluster in the same manner.

For **OnWN**, we used the sense definition pairs of OntoNotes (Hovy et al., 2006) and WordNet (Fellbaum, 1998). Different from previous tasks, the two definition sentences in a pair belong to different senses. We sampled 750 pairs based on a string similarity ranging from 0.5 to 1.

The **Images** data set is a subset of the PASCAL VOC-2008 data set (Rashtchian et al., 2010), which consists of 1,000 images and has been used by a number of image description systems. It was also sampled from string similarity values between 0.6 and 1.

Deft-forum and **Deft-news** are from DEFT data.[2] Deft-forum contains the forum post sentences, and Deft-news are news summaries. We selected 450 pairs for Deft-forum and 300 pairs for Deft-news. They are sampled evenly from string similarities falling in the interval 0.6 to 1.

The **Tweets** data set contains tweet-news pairs selected from the corpus released in (Guo et al., 2013), where each pair contains a sentence that pertains to the news title, while the other one represents a Twitter comment on that particular news. They are evenly sampled from string similarity values between 0.5 and 1.

Table 1 shows the explanations and values associated with each score between 5 and 0. As in prior years, we used Amazon Mechanical Turk (AMT)[3] to crowdsource the annotation of the English pairs.[4] Annotators are presented with the

[1] http://www.cs.columbia.edu/~weiwei/workshop/

[2] LDC2013E19, LDC2012E54

[3] www.mturk.com

[4] For STS 2013, we used CrowdFlower as a front-end to

Score	English	Spanish
5/4	*The two sentences are completely equivalent, as they mean the same thing.*	
	The bird is bathing in the sink. Birdie is washing itself in the water basin.	El pájaro se esta bañando en el lavabo. El pájaro se está lavando en el aguamanil.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*	
	In May 2010, the troops attempted to invade Kabul. The US army invaded Kabul on May 7th last year, 2010.	
3	*The two sentences are roughly equivalent, but some important information differs/missing.*	
	John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.	John dijo que él es considerado como testigo, y no como sospechoso. "Él ya no es un sospechoso," John dijo.
2	*The two sentences are not equivalent, but share some details.*	
	They flew out of the nest in groups. They flew into the nest together.	Ellos volaron del nido en grupos. Volaron hacia el nido juntos.
1	*The two sentences are not equivalent, but are on the same topic.*	
	The woman is playing the violin. The young lady enjoys listening to the guitar.	La mujer está tocando el violín. La joven disfruta escuchar la guitarra.
0	*The two sentences are completely dissimilar.*	
	John went horse back riding at dawn with a whole group of friends. Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.	Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. La salida del sol al amanecer es una magnífica vista que puede presenciar si usted se despierta lo suficientemente temprano para verla.

Table 1: Similarity scores with explanations and examples for the English and Spanish subtasks, where the sentences in Spanish are translations of the English ones.
A similarity score of 5 in English is mirrored by a maximum score of 4 in Spanish; the definitions pertaining to scores 3 and 4 in English were collapsed under a score of 3 in Spanish, with the definition "The two sentences are mostly equivalent, but some details differ."

detailed instructions provided in Figure 1, and are asked to label each STS sentence pair on our six point scale, selecting from a dropdown box. Five sentence pairs are presented to each annotator at once, per human intelligence task (HIT), at a payrate of $0.20; we collect five separate annotations per sentence pair. Annotators were only eligible to work on the task if they had the Mechanical Turk Master Qualification. This is a special qualification conferred by AMT (using a priority statistical model) to annotators who consistently maintain a very high level of quality across a variety of tasks from numerous requesters). Access to these skilled workers entails a 20% surcharge.

To monitor the quality of the annotations, we use the gold dataset of 105 pairs that were manually annotated by the task organizers during STS 2013. We include one of these gold pairs in each set of five sentence pairs, where the gold pairs are indistinguishable from the rest. Unlike when we ran on CrowdFlower for STS 2013, the gold pairs are not used for training purposes, nor are workers automatically banned from the task if they make too many mistakes on annotating them. Rather, the gold pairs are only used to help in identifying and removing the data associated with poorly performing annotators. With few exceptions, 90% of the answers from each individual annotator fall within +/-1 of the answers selected by the organizers for

Amazon Mechanical Turk, since it provides numerous useful tools to assist in running a successful annotation project using crowdsourcing, such as support for hidden 'golden' questions that can be used both to train annotators and to automatically stop people who repeatedly make mistakes from contributing to the task. However, in 2013, CrowdFlower dropped Amazon Mechanical Turk as an annotation source. When we tried running pairs for STS 2014 on CrowdFlower using the same templates that were successfully used for the 2013 task, we found that we obtained significantly degraded annotation quality, with an average Pearson (AMT provider vs. rest of AMT providers) of only 22.8%. In contrast, when we ran the task for 2014 on AMT, we obtained a one-vs-rest annotation of 73.6%.

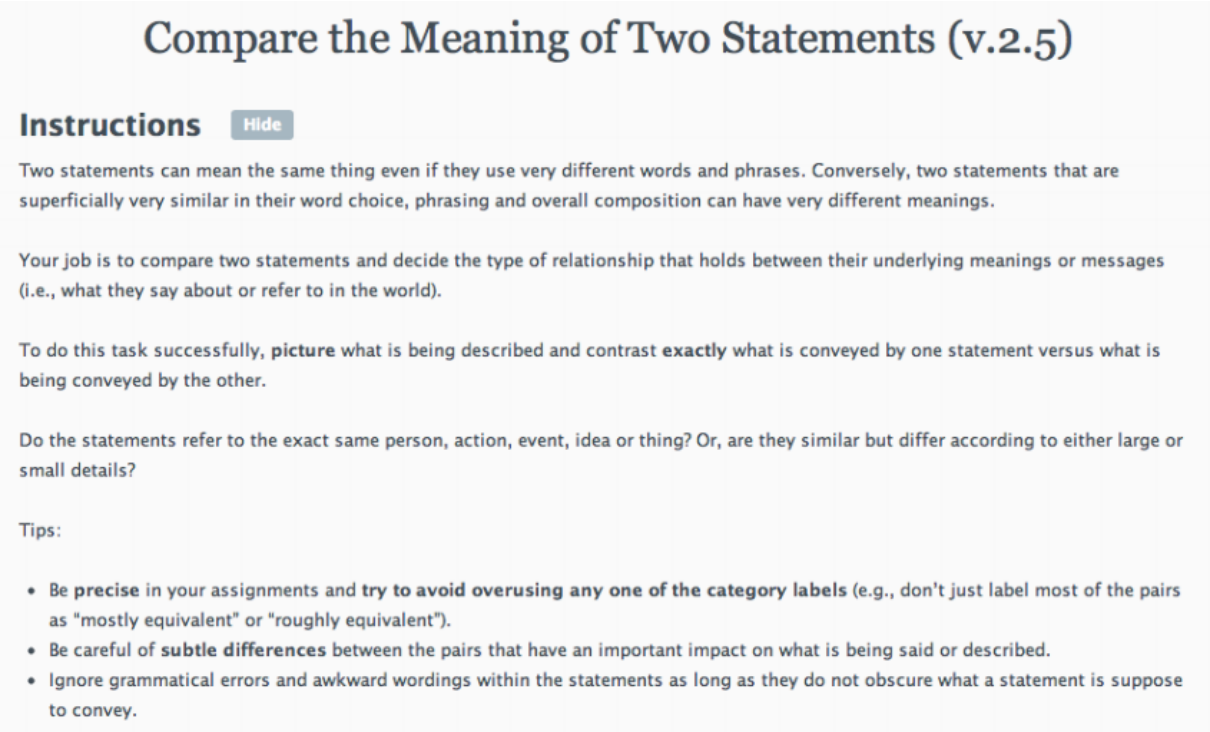

Figure 1: Annotation instructions for English subtask.

the gold dataset.

The distribution of scores obtained from the AMT providers in the Deft-forum, Deft-news, OnWN and tweet-news datasets is roughly uniform across the different grades of similarity, although the scores are slightly higher for tweet-news. Compared to the other data sets, the scores for OnWN, were more bimodal, ranging between 4.6 to 5 and 0 to 0.4, when compared to middle values (2.6-3.4).

In order to assess the annotation quality, we measure the correlation of each annotator with the average of the rest of the annotators, and then average the results. This approach to estimate the quality is identical to the method used for evaluations (see Section 3), and it can thus be considered as the upper bound of the systems. The inter-tagger correlation for each English dataset is as follows:

- HDL: 79.4%
- OnWN: 67.2%
- Deft-forum: 58.6%
- Deft-news: 70.7%
- Images: 83.6%
- Tweets-news: 74.4%

The correlation figures are generally high (over 70%), with the exception of the OnWN and Deft datasets, which score 67.2% and 58.6%, respectively. The reason for the low inter-tagger correlation on OnWN compared to the higher correlations in previous years is that we only used unmapped sense definitions, i.e., the two sentences in a pair belong to two different senses. For the Deft-forum dataset, we found that similarity values tend to be lower than in the other datasets, and more annotation disagreements happen in these low similarity values.

2.2 Spanish Subtask

The Spanish subtask follows a setup similar to the English subtask, except that the similarity scores were adapted to fit a range from 0 to 4 (see Table 1). We thought that the distinction between a score of 3 and 4 for the English task will pose more difficulty for us in conveying into Spanish, as the sole difference between the two lies in how the annotators perceive the importance of additional details or missing information with respect to the core semantic interpretation of the pair. As this aspect entails a subjective judgement, and since it is the first time that a Spanish STS evaluation is organized, we casted the annotation guidelines into straightforward and unambiguous instructions, and thus opted to use a similarity range from 0 to 4.

Prior to the evaluation window, we released 65 Spanish sentence pairs for trial / training. In order to evaluate system performance under differ-

ent scenarios, we developed two test datasets, one extracted from the Spanish Wikipedia[5] (December 2013 dump) and one from contemporary news articles collected from media in Spanish (February 2014).

2.2.1 Spanish Wikipedia

The Wikipedia dump was processed using the Parse::MediaWikiDump Perl library. We removed all titles, html tags, wiki tags and hyperlinks (keeping only the surface forms). Each article was split into paragraphs, where the first paragraph was considered to be the article's abstract, while the remaining ones were deemed to be its content. Each of these were split into sentences using the Perl library Uplug::PreProcess::SentDetect, and only the sentences longer than eight words were used. We iteratively computed the lexical similarity[6] between every sentence in the abstract and every sentence in the content, and retained those pairs whose sentence length ratio was higher than 0.5, and their similarity scored over 0.35.

The final set of sentence pairs was split into five bins, and their scores normalized to range from 0 to 1. The more interesting and difficult pairs were found, perhaps not surprisingly, in bins 0 and 1, where synonyms/short paraphrases where more frequent. An example extracted from those bins, where the text in italics highlights the differences between the two sentences:

- "America*" es el* segund*o continente* más grande del planeta, *después* de Asia.
  "America" is the second largest continent in the world, following Asia.
- America *corresponde a la* segund*a masa de tierra* más grande del planeta, *luego* de Asia.
  America is the second largest land mass on the planet, after Asia.

The Spanish verb "*Es*" maps to (En:[7] is), "*corresponde a*" (En: corresponds to), the phrase "*el segundo continente*" (En: the second continent) is equivalent to "*la segunda masa de tierra*" (the second land mass), and "*despues*" (En: following) to "*luego*" (En: after). Despite the difference in vocabulary choice, the two sentences are paraphrases of each other.

From the candidate pairs, we manually selected 324 sentence pairs, in order to ensure a diverse and challenging set. This set was annotated in two ways, first by two graduate students in Computer Science who are native speakers of Spanish, and second by using AMT.

The AMT framework was set up to contain seven sentence pairs per HIT, where six of them were part of the test dataset, while one was used for control. AMT providers were eligible to complete a task if they had more than 500 accepted HITs, with 90%+ acceptance rate.[8] We paid $0.30 per HIT, and each HIT was annotated by five AMT providers. We sought to ensure that only Spanish speaking annotators would complete the HITs by providing all the information related to the task (its title, abstract, description, guidelines and examples), as well as the control pair in Spanish only. The participants were instructed to label the pairs on a scale from 0 to 4 (see Table 1). Each sentence pair was followed by a comment text box, which the AMT providers used to provide the topic of the sentences, corrections, etc.

The two students achieved a Pearson correlation of 0.6974 on the Wikipedia dataset. To see how their judgement compares to the crowd wisdom, we averaged the AMT scores for each pair, and computed their correlation with our annotators, obtaining 0.824 and 0.742, respectively. Surprisingly enough, both these correlation values are higher than the correlation among the annotators themselves. When averaging the annotator scores and comparing them with the AMT providers' average score per pair, the correlation becomes 0.8546, indicating that the task is well defined, and that the annotations contributed by the AMT providers are of satisfactory quality. Given these scores, the gold standard was annotated using the average AMT provider judgement per pair.

2.2.2 Spanish News

The second Spanish dataset was extracted from news articles published in Spanish language media from around the world in February 2014. The hyperlinks to the articles were obtained by parsing the "International" page of Spanish Google News,[9] which aggregates or clusters in real time articles describing a particular event from a diverse pool of news sites, where each grouping

[5] `es.wikipedia.org`

[6] Algorithm based on the Linux *diff* command (Algorithm::Diff Perl module).

[7] "En" stands for English.

[8] Initially, Amazon had automatically upgraded our annotation task to require Master level providers (as those participating in the English annotations), yet after approximately 4 days, no HIT had been completed.

[9] `news.google.es`

is labeled with the title of one of the predominant articles. By leveraging these clusters of links pointing to the sites where the articles were originally published, we are able to gather raw text that has a high probability of containing semantically similar sentences. We encountered several difficulties while mining the articles, ranging from each article having its own formatting depending on the source site, to advertisements, cookie requirements, to encoding for Spanish diacritics. We used the *lynx text-based browser*,[10] which was able to standardize the raw articles to a degree. The output of the browser was processed using a rule based approach taking into account continuous text span length, ratio of symbols and numbers to the text, etc., in order to determine when a paragraph is part of the article content. After that, a second pass over the predictions corrected mislabeled paragraphs if they were preceded and followed by paragraphs identified as content. All the content pertaining to articles on the same event was joined, sentence split, and *diff* pairwise similarities were computed. The set of candidate sentences followed the same requirements as for the Wikipedia dataset, namely length ratio higher than 0.5 and similarity score over 0.35. From these, we manually extracted 480 sentence pairs which were deemed to pose a challenge to an automated system.

Due to the high correlations obtained between the AMT providers' scores and the annotators' scores on Wikipedia, the news dataset was only annotated using AMT, following exactly the same task setup as for Wikipedia.

3 Evaluation

Evaluation of STS is still an open issue. STS experiments have traditionally used Pearson product-moment correlation between the system scores and the GS scores, or, alternatively, Spearman rank order correlation. In addition, we also need a method to aggregate the results from each dataset into an overall score. The analysis performed in (Agirre and Amigó, In prep) shows that Pearson and averaging across datasets are the best suited combination in general. In particular, Pearson is more informative than Spearman, in that Spearman only takes the rank differences into account, while Pearson does account for value differences as well. The study also showed that other alternatives need to be considered, depending on the requirements of the target application.

We leave application-dependent evaluations for future work, and focus on average Pearson correlation. When averaging, we weight each individual correlation by the size of the dataset. In order to compute statistical significance among system results, we use a one-tailed parametric test based on Fisher's z-transformation (Press et al., 2002, equation 14.5.10). In addition, English subtask participants could provide an optional confidence measure between 0 and 100 for each of their predictions. Team RTM-DCU is the only one who has provided these, and the evaluation of their runs using weighted Pearson (Pozzi et al., 2012) is listed at the end of Table 3.

Participants[11] could take part in the shared task with a maximum of 3 system runs per subtask.

3.1 English Subtask

In order to provide a simple word overlap baseline (Baseline-tokencos), we tokenize the input sentences splitting on white spaces, and then represent each sentence as a vector in the multidimensional token space. Each dimension has 1 if the token is present in the sentence, 0 otherwise. Vector similarity is computed using the cosine similarity metric.

We also run the freely available system, TakeLab (Šarić et al., 2012), which yielded state of the art performance in STS 2012 and strong results out-of-the-box in 2013.[12]

15 teams participated in the English subtask, submitting 38 system runs. One team submitted the results past the deadline, as explicitly marked in Table 3. After the submission deadline expired, the organizers published the gold standard and participant submissions on the task website, in order to ensure a transparent evaluation process.

Table 3 shows the results of the English subtask, with runs listed in alphabetical order. The correlation in each dataset is given, followed

[10] `lynx.browser.org`

[11] **Participating teams**: Bielefeld_SC (McCrae et al., 2013), BUAP (Vilariño et al., 2014), DLS@CU (Sultan et al., 2014b), FBK-TR (Vo et al., 2014), IBM_EG (no information), LIPN (Buscaldi et al., 2014), Meerkat_Mafia (Kashyap et al., 2014), NTNU (Lynum et al., 2014), RTM-DCU (Biçici and Way, 2014), SemantiKLUE (Proisi et al., 2014), StanfordNLP (Socher et al., 2014), TeamZ (Gupta, 2014), UMCC_DLSI_SemSim (Chavez et al., 2014), UNAL-NLP (Jimenez et al., 2014), UNED (Martinez-Romo et al., 2011), UoW (Rios, 2014).

[12] Code is available at `http://ixa2.si.ehu.es/stswiki`

Run Name	deft forum	deft news	Headl	images	OnWN	tweet news	Weighted mean	Rank
Baseline-tokencos	0.353	0.596	0.510	0.513	0.406	0.654	0.507	-
TakeLab	0.333	0.716	0.720	0.742	0.793	0.650	0.678	-
Bielefeld_SC-run1	0.211	0.432	0.321	0.368	0.367	0.415	0.354	32
Bielefeld_SC-run2	0.211	0.431	0.311	0.356	0.361	0.409	0.347	33
BUAP-EN-run1	0.456	0.686	0.689	0.697	0.654	0.771	0.671	19
DLS@CU-run1	0.483	0.766	0.765	0.821	0.723	0.764	0.734	7
DLS@CU-run2	0.483	0.766	0.765	0.821	0.859	0.764	0.761	1
FBK-TR-run1	0.322	0.523	0.547	0.601	0.661	0.462	0.535	25
FBK-TR-run2	0.167	0.421	0.485	0.521	0.572	0.359	0.441	28
FBK-TR-run3	0.305	0.405	0.471	0.489	0.551	0.438	0.459	27
IBM_EG-run1	0.474	0.743	0.737	0.801	0.760	0.730	0.722	8
IBM_EG-run2	0.464	0.641	0.710	0.747	0.732	0.696	0.684	15
LIPN-run1	0.454	0.640	0.653	0.809	-	0.551	0.508	26
LIPN-run2	0.084	-	-	-	-	-	0.010	35
Meerkat_Mafia-Hulk	0.449	0.785	0.757	0.790	0.787	0.757	0.735	6
Meerkat_Mafia-pairingWords	0.471	0.763	0.760	0.801	0.875	0.779	0.761	2
Meerkat_Mafia-SuperSaiyan	0.492	0.771	0.767	0.768	0.802	0.765	0.741	5
NTNU-run1	0.437	0.714	0.722	0.800	0.835	0.411	0.663	20
NTNU-run2	0.508	0.766	0.753	0.813	0.777	0.792	0.749	4
NTNU-run3	0.531	0.781	0.784	0.834	0.850	0.675	0.755	3
SemantiKLUE-run1	0.337	0.608	0.728	0.783	0.848	0.632	0.687	14
SemantiKLUE-run2	0.349	0.643	0.733	0.773	0.855	0.640	0.694	13
StanfordNLP-run1	0.319	0.635	0.636	0.758	0.627	0.669	0.627	22
StanfordNLP-run2	0.304	0.679	0.621	0.715	0.625	0.636	0.610	24
StanfordNLP-run3	0.342	0.650	0.602	0.754	0.609	0.638	0.614	23
UMCC_DLSI_SemSim-run1	0.475	0.662	0.632	0.742	0.813	0.675	0.682	16
UMCC_DLSI_SemSim-run2	0.469	0.662	0.625	0.739	0.814	0.654	0.676	18
UMCC_DLSI_SemSim-run3	0.283	0.385	0.267	0.436	0.603	0.278	0.381	30
UNAL-NLP-run1	0.504	0.721	0.762	0.807	0.782	0.614	0.711	12
UNAL-NLP-run2	0.383	0.730	0.765	0.771	0.827	0.403	0.657	21
UNAL-NLP-run3	0.461	0.722	0.761	0.778	0.843	0.658	0.721	9
UNED-run22_p_np	0.104	0.315	0.037	0.324	0.509	0.490	0.310	34
UNED-runS5K_10_np	0.118	0.506	0.057	0.498	0.488	0.579	0.379	31
UNED-runS5K_3_np	0.094	0.564	0.018	0.607	0.577	0.670	0.431	29
UoW-run1	0.342	0.751	0.754	0.776	0.799	0.737	0.714	11
UoW-run2	0.342	0.587	0.754	0.788	0.799	0.628	0.682	17
UoW-run3	0.342	0.763	0.754	0.788	0.799	0.753	0.721	10
†RTM-DCU-run1	0.434	0.697	0.620	0.699	0.806	0.688	0.671	
†RTM-DCU-run2	0.397	0.681	0.613	0.666	0.799	0.669	0.651	
†RTM-DCU-run3	0.308	0.556	0.630	0.647	0.800	0.553	0.608	
†RTM-DCU-run1	0.418	0.685	0.622	0.698	0.833	0.687	0.673	
†RTM-DCU-run2	0.383	0.674	0.609	0.663	0.826	0.669	0.653	
†RTM-DCU-run3	0.273	0.553	0.633	0.644	0.825	0.568	0.611	

Table 3: English evaluation results. Results at the top correspond to out-of-the-box systems. Results at the bottom correspond to results using the confidence score.
Notes: "-" for not submitted, "†" for post-deadline submission.

by the mean correlation (the official measure), and the rank of the run. The highest correlations are for OnWN (87.5%, by Meerkat_Mafia) and images (83.4%, by NTNU), followed by Tweets (79.2%, by NTNU), HEADL (78.4%, by NTNU) and deft news and forums (78.1% and 53.1%, respectively, by NTNU). Compared to the inter-annotator agreement correlation, the ranking among datasets is very similar, with the exception of OnWN, as it gets the best score but has very low agreement. One possible reason is that the participants used previously available data. The results of the best 4 top system runs are significantly different (p-value < 0.05) from the 5th top scoring system run and below. The top 4 systems did not show statistical significant variation among them.

Only three runs (cf. lower rows in Table 3) included non-uniform confidence scores, barely affecting their ranking.

Interestingly, the two top performing systems on the English STS sub-task are both unsupervised. DLS@CU (Sultan et al., 2014b) presents an unsupervised algorithm which predicts the STS score based on the proportion of word alignments in the two sentences. Two related words are aligned depending on how similar the two words are, and also on how similar the contexts of the words are in the respective sentences (Sultan et al., 2014a). Meerkat_Mafia_pairingWords (Kashyap et al., 2014) also follows a fully unsupervised approach. The authors train LSA on an English corpus of three billion words using a sliding window approach, resulting in a vocabulary size of 29,000 words associated with 300 dimensions. They account for named entities and out-of-vocabulary words by leveraging external resources such as DBpedia[13] and Wordnik.[14] In Spanish, the system equivalent to this run ranked second following a cross-lingual approach, by applying the English system to the translated version of the dataset (see 3.2).

The Table also shows the results of TakeLab, which was trained with all datasets from previous years. TakeLab would rank 18th, ten absolute points below the best system, a smaller difference than in 2013.

[13] dbpedia.org

[14] wordnick.com

3.2 Spanish Subtask

The Spanish subtask attracted 9 teams with 22 participating systems, out of which 16 were supervised and 6 unsupervised. The participants were from both Spanish (Colombia, Cuba, Mexico, Spain), and non-Spanish speaking countries (two teams from France, Germany, Ireland, UK, US). The evaluation results appear in Table 4.

The top ranking system is the 2nd run of UMCC_DLSI_SemSim (Chavez et al., 2014), which achieves a weighted correlation of 0.807. It entails a cross-lingual approach, as it leverages a SVM-based English framework, by mapping the Spanish words to their English equivalent using the most common sense in WordNet 3.0. The classifier uses a combination of features, such as those derived from traditional knowledge-based ((Leacock and Chodorow, 1998; Wu and Palmer, 1994; Lin, 1998), and others) and corpus-based metrics (LSA (Landauer et al., 1997)), paired with lexical features (such as Dice-Similarity, Euclidean-distance, etc.). It is trained on a cumulative English STS dataset comprising train and test data released as part of tasks in SemEval2012 (Agirre et al., 2012) and *Sem 2013 (Agirre et al., 2013), as well as training data available from tasks 1 and 10 in SemEval 2014. Interestingly enough, run 2 of the system performs better than run 1, despite the fact that it uses half the features, and focuses on string based similarity measures only. This difference between runs is noticed on the Wikipedia dataset only, and it amounts to 4% Pearson correlation. While the system had a robust performance on the Spanish subtask, for English, its overall rank was 16, 18, and 33, respectively.

Coming in close at only 0.3% difference, is Meerkat-Mafia PairingAvg (run 2) (Kashyap et al., 2014), which also follows a cross-lingual approach, by applying the system the team developed for the English subtask to the translated version of the datasets (see 3.1). The interesting aspect of their work is that in their first submission (run 1), they only consider the similarity resulting from the sentence pair translation through the Google Translate service.[15] In the second run, they expand each sentence to 20 possible combinations by accounting for the multiple translation meanings of a given word, and considering the average similarity of all resulting pairs. While the first run achieves a weighted correlation of 73.8%,

[15] translate.google.com

Run Name	System type	Wikipedia	News	Weighted mean	Rank
Bielefeld-SC-run1	unsupervised*	0.263	0.554	0.437	22
Bielefeld-SC-run2	unsupervised*	0.265	0.555	0.438	21
BUAP-run1	supervised	0.550	0.679	0.627	17
BUAP-run2	unsupervised	0.640	0.764	0.714	14
RTM-DCU-run1	supervised	0.422	0.700	0.588	18
RTM-DCU-run2	supervised	0.369	0.625	0.522	20
RTM-DCU-run3	supervised	0.424	0.641	0.554	19
LIPN-run1	supervised	0.652	0.826	0.756	11
LIPN-run2	supervised	0.716	0.832	0.785	6
LIPN-run3	supervised	0.716	0.809	0.771	10
Meerkat-Mafia-run1	unsupervised	0.668	0.785	0.738	13
Meerkat-Mafia-run2	unsupervised	0.743	**0.845**	0.804	2
Meerkat-Mafia-run3	supervised	0.738	0.822	0.788	5
TeamZ-run1	supervised	0.610	0.717	0.674	15
TeamZ-run2	supervised	0.604	0.710	0.667	16
UMCC-DLSI-run1	supervised	0.741	0.825	0.791	4
UMCC-DLSI-run2	supervised	0.7802	0.825	**0.807**	1
UNAL-NLP-run1	weakly supervised	**0.7803**	0.815	0.801	3
UNAL-NLP-run2	supervised	0.757	0.783	0.772	9
UNAL-NLP-run3	supervised	0.689	0.796	0.753	12
UoW-run1	supervised	0.748	0.800	0.779	7
UoW-run2	supervised	0.748	0.800	0.779	8

Table 4: Spanish evaluation results in terms of Pearson correlation.

the second one performs significantly better at 80.4%, indicating that the additional context may also include multiple instances of accurate translations, hence significantly impacting the overall similarity score. In English, the system equivalent to run 2 in Spanish, namely Meerkat MafiapairingWords, achieves a competitive ranked performance across all six datasets, ranking second, at an order of 10^{-4} distance from the top system. This supports the claim that, despite its unsupervised nature, the system is quite versatile and highly competitive with the top performing supervised frameworks, and that it may achieve an even higher performance in Spanish if accurate sentence translations were provided.

Overall, most systems were cross-lingual, relying on different translation approaches, such as 1) translating the test data into English (as the two systems above), and then exporting the score obtained for the English sentences back to Spanish, or 2) performing automatic translation of the English training data, and learning a classifier directly in Spanish. (Buscaldi et al., 2014) supplemented their training dataset with human annotations conducted in Spanish, using definition pairs extracted from a Spanish dictionary. A different angle was explored by (Rios, 2014), who proposed a multilingual framework using transfer learning across English and Spanish by training on traditional lexical, knowledge-based and corpus-based features. The semantic similarity task was approached from a monolingual perspective as well (Gupta, 2014), by focusing on Spanish resources, such as the trial data we released as part of the subtask, and the Spanish WordNet;[16] these were leveraged using meta-learning over variations of overlap-based metrics. Following the same line, (Biçici and Way, 2014) pursued language independent methods, who avoided relying on task or domain specific information through the usage of referential translation machines. This approach models textual semantic similarity as a decision in terms of translation quality between two datasets (in our case Spanish STS trial and test data) given relevant examples from an in-language reference corpus.

In comparison to the correlations obtained in the English subtask, where the highest weighted mean was 76.1%, for Spanish, we obtained 80.7%, probably due to the more formal nature of the datasets, since Wikipedia and news articles employ mostly well formed and grammatically correct sentences, and we selected all snippets to be longer than 8 words. The overall correlation scores obtained for English were hurt by the deft-forum data, which scored significantly lower (at a maximum correlation of 50.8%), when compared to all the other datasets whose correlation was higher than 70%. The OnWN data was most similar to our test sets, and it attained a maximum of 85.9%.

[16] grial.uab.es/descarregues.php

4 Conclusion

This year's STS task comprised a multilingual flair, by introducing Spanish datasets alongside the English ones. In English, the datasets sought to expose the participating teams to more diverse scenarios compared to the previous years, by introducing image descriptions, forum and newswire genre, and tweet-newswire headline mappings. For Spanish, two datasets were developed consisting of encyclopedic and newswire text acquired from Spanish sources. Overall, the English subtask attracted 15 teams (with 38 system variations), while the Spanish subtask had 9 teams (with 22 system runs). Most teams from the Spanish subtask have also submitted runs for the English evaluations.

Acknowledgments

The authors are grateful to Verónica Pérez-Rosas and Vanessa Loza for their help with the annotations for the Spanish subtask. This material is based in part upon work supported by National Science Foundation CAREER award #1361274 and IIS award #1018613, by DARPA-BAA-12-47 DEFT grant #12475008, and by MINECO CHIST-ERA READERS and SKATER projects (PCIN-2013-002-C02-01, TIN2012-38584-C06-02). Aitor Gonzalez Agirre is supported by a doctoral grant from MINECO. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Defense Advanced Research Projects Agency.

References

Eneko Agirre and Enrique Amigó. In prep. Exploring evaluation measures for semantic textual similarity. In *Unpublished manuscript*.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 Shared Task: Semantic textual similarity, including a pilot on typed-similarity. In *The Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, pages 32–43.

Clive Best, Erik van der Goot, Ken Blackler, Tefilo Garcia, and David Horby. 2005. Europe media monitor - system description. In *EUR Report 22173-En*, Ispra, Italy.

Ergun Biçici and Andy Way. 2014. RTM-DCU: Referential translation machines for semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Davide Buscaldi, Jorge J. Garcia Flores, Joseph Le Roux, Nadi Tomeh, and Belem Priego Sanchez. 2014. LIPN: Introducing a new geographical context similarity measure and a statistical similarity measure based on the Bhattacharyya coefficient. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Alexander Chavez, Hector Davila, Yoan Gutierrez, Antonio Fernandez-Orquin, Andrés Montoyo, and Rafael Munoz. 2014. UMCC_DLSI_SemSim: Multilingual system for measuring semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Christiane Fellbaum. 1998. *WordNet - An electronic lexical database*. MIT Press.

Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking tweets to news: A framework to enrich online short text data in social media. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*, pages 239–249.

Anubhav Gupta. 2014. TeamZ: Measuring semantic textual similarity for Spanish using an overlap-based approach. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

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*, pages 57–60.

Sergio Jimenez, George Dueñas, Julia Baquero, and Alexander Gelbukh. 2014. UNAL-NLP: Combining soft cardinality features for semantic textual similarity, relatedness and entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Abhay Kashyap, Lushan Han, Roberto Yus, Jennifer Sleeman, Taneeya Satyapanich, Sunil Gandhi, and Tim Finin. 2014. Meerkat Mafia: Multilingual and cross-level semantic textual similarity systems. In *Proceedings of the 8th International Workshop on*

Semantic Evaluation (SemEval-2014), Dublin, Ireland.

Thomas K. Landauer, Darrell Laham, Bob Rehder, and M. E. Schreiner. 1997. How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. *Cognitive Science*.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In *WordNet: An Electronic Lexical Database*, pages 305–332.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, Madison, Wisconsin.

André Lynum, Partha Pakray, Björn Gambäck, and Sergio Jimenez. 2014. NTNU: Measuring semantic similarity with sublexical feature representations and soft cardinality. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Juan Martinez-Romo, Lourdes Araujo, Javier Borge-Holthoefer, Alex Arenas, José A. Capitán, and José A. Cuesta. 2011. Disentangling categorical relationships through a graph of co-occurrences. *Phys. Rev. E*, 84:046108, Oct.

John P. McCrae, Philipp Cimiano, and Roman Klinger. 2013. Orthonormal explicit topic analysis for cross-lingual document matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1732–1740, Seattle, Washington, USA.

Francesco Pozzi, Tiziana Di Matteo, and Tomaso Aste. 2012. Exponential smoothing weighted correlations. *The European Physical Journal B*, 85(6).

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2002. *Numerical recipes: The art of scientific computing V 2.10 with Linux or single-screen license*. Cambridge University Press.

Thomas Proisi, Stefan Evert, Paul Greiner, and Besim Kabashi. 2014. SemantiKLUE: Robust semantic similarity at multiple levels using maximum weight matching. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA.

Miguel Rios. 2014. UoW: Multi-task learning Gaussian process for semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, pages 207–218.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014a. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014b. DLS@CU: Sentence similarity from word aligment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Darnes Vilariño, David Pinto, Saúl León, Mireya Tovar, and Beatriz Beltrán. 2014. BUAP: Evaluating features for multilingual and cross-level semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Ngoc Phuoc An Vo, Tommaso Caselli, and Octavian Popescu. 2014. FBK-TR: Applying SVM with multiple linguistic features for cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Takelab: Systems for measuring semantic text similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 441–448, Montréal, Canada, 7-8 June.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico.