SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation

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

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

[b]University of Michigan
Ann Arbor, MI

[d]Google Inc.
Mountain View, CA

[e]George Washington University
Washington, DC

[f]University of Pittsburgh
Pittsburgh, PA

Abstract

Semantic Textual Similarity (STS) seeks to measure the degree of semantic equivalence between two snippets of text. Similarity is expressed on an ordinal scale that spans from semantic equivalence to complete unrelatedness. Intermediate values capture specifically defined levels of partial similarity. While prior evaluations constrained themselves to just monolingual snippets of text, the 2016 shared task includes a pilot subtask on computing semantic similarity on cross-lingual text snippets. This year's traditional monolingual subtask involves the evaluation of English text snippets from the following four domains: Plagiarism Detection, Post-Edited Machine Translations, Question-Answering and News Article Headlines. From the question-answering domain, we include both question-question and answer-answer pairs. The cross-lingual subtask provides paired Spanish-English text snippets drawn from the same sources as the English data as well as independently sampled news data. The English subtask attracted 43 participating teams producing 119 system submissions, while the cross-lingual Spanish-English pilot subtask attracted 10 teams resulting in 26 systems.

1 Introduction

Semantic Textual Similarity (STS) assesses the degree to which the underlying semantics of two segments of text are equivalent to each other. This assessment is performed using an ordinal scale that ranges from complete semantic equivalence to complete semantic dissimilarity. The intermediate levels capture specifically defined degrees of partial similarity, such as topicality or rough equivalence, but with differing details. The snippets being scored are approximately one sentence in length, with their assessment being performed outside of any contextualizing text. While STS has previously just involved judging text snippets that are written in the same language, this year's evaluation includes a pilot subtask on the evaluation of cross-lingual sentence pairs.

The systems and techniques explored as a part of STS have a broad range of applications including Machine Translation (MT), Summarization, Generation and Question Answering (QA). STS allows for the independent evaluation of methods for computing semantic similarity drawn from a diverse set of domains that would otherwise be only studied within a particular subfield of computational linguistics. Existing methods from a subfield that are found to perform well in a more general setting as well as novel techniques created specifically for STS may improve any natural language processing or language understanding application where knowing the similarity in meaning between two pieces of text is relevant to the behavior of the system.

Paraphrase detection and textual entailment are both highly related to STS. However, STS is more similar to paraphrase detection in that it defines a bi-directional relationship between the two snippets being assessed, rather than the non-symmetric propositional logic like relationship used in textual entailment (e.g., $P \rightarrow Q$ leaves $Q \rightarrow P$ unspecified). STS also expands the binary yes/no catego-

The authors of this paper are listed in alphabetic order.

Score	English	Cross-lingual Spanish-English
5	*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. Birdie is washing itself in the water basin.
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.	En mayo de 2010, las tropas intentaron invadir 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. "He is not a suspect anymore." John said.
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. They flew into the nest together.
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. The young lady enjoys listening to the guitar.
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. Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

Table 1: Similarity scores with explanations and examples for the English and the cross-lingual Spanish-English subtasks.

rization of both paraphrase detection and textual entailment to a finer grained similarity scale. The additional degrees of similarity introduced by STS are directly relevant to many applications where intermediate levels of similarity are significant. For example, when evaluating machine translation system output, it is desirable to give credit for partial semantic equivalence to human reference translations. Similarly, a summarization system may prefer short segments of text with a rough meaning equivalence to longer segments with perfect semantic coverage.

STS is related to research into machine translation evaluation metrics. This subfield of machine translation investigates methods for replicating human judgements regarding the degree to which a translation generated by an machine translation system corresponds to a reference translation produced by a human translator. STS systems plausibly could be used as a drop-in replacement for existing translation evaluation metrics (e.g., BLEU, MEANT, METEOR, TER).[1] The cross-lingual STS subtask that is newly introduced this year is similarly related to machine translation quality estimation.

The STS shared task has been held annually since 2012, providing a venue for the evaluation of state-of-the-art algorithms and models (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015). During this time, a diverse set of genres and data sources have been explored (i.a., news headlines, video and image descriptions, glosses from lexical resources including WordNet (Miller, 1995; Christiane Fellbaum, 1998), FrameNet (Baker et al., 1998), OntoNotes (Hovy et al., 2006), web discussion forums, and Q&A data sets). This year's

[1]Both monolingual and cross-lingual STS score what is referred to in the machine translation literature as adequacy and ignore fluency unless it obscures meaning. While popular machine translation evaluation techniques do not assess fluency independent from adequacy, it is possible that the deeper semantic assessment being performed by STS systems could benefit from being paired with a separate fluency module.

evaluation adds new data sets drawn from plagiarism detection and post-edited machine translations. We also introduce an evaluation set on Q&A forum question-question similarity and revisit news headlines and Q&A answer-answer similarity. The 2016 task includes both a traditional monolingual subtask with English data and a pilot cross-lingual subtask that pairs together Spanish and English texts.

2 Task Overview

STS presents participating systems with paired text snippets of approximately one sentence in length. The systems are then asked to return a numerical score indicating the degree of semantic similarity between the two snippets. Canonical STS scores fall on an ordinal scale with 6 specifically defined degrees of semantic similarity (see Table 1). While the underlying labels and their interpretation are ordinal, systems can provide real valued scores to indicate their semantic similarity prediction.

Participating systems are then evaluated based on the degree to which their predicted similarity scores correlate with STS human judgements. Algorithms are free to use any scale or range of values for the scores they return. They are not punished for outputting scores outside the range of the interpretable human annotated STS labels. This evaluation strategy is motivated by a desire to maximize the flexibility in the design of machine learning models and systems for STS. It reinforces the assumption that computing textual similarity is an enabling component for other natural language processing applications, rather than being an end in itself.

Table 1 illustrates the ordinal similarity scale the shared task uses. Both the English and the cross-lingual Spanish-English STS subtasks use a 6 point similarity scale. A similarity label of 0 means that two texts are completely dissimilar; this can be interpreted as two sentences with no overlap in their meanings. The next level up, a similarity label of 1, indicates that the two snippets are not equivalent but are topically related to each other. A label of 2 indicates that the two texts are still not equivalent but agree on some details of what is being said. The labels 3 and 4, both indicate that the two sentences are approximately equivalent. However, a score of 3 implies that there are some differences in important details, while a score of 4 indicates that the differing details are not important. The top score of 5, denotes that the two texts being evaluated have complete semantic equivalence.

In the context of the STS task, meaning equivalence is defined operationally as two snippets of text that mean the same thing when interpreted by a reasonable human judge. The operational approach to sentence level semantics was popularized by the recognizing textual entailment task (Dagan et al., 2010). It has the advantage that it allows the labeling of sentence pairs by human annotators without any training in formal semantics, while also being more useful and intuitive to work with for downstream systems. Beyond just sentence level semantics, the operationally defined STS labels also reflect both world knowledge and pragmatic phenomena.

As in prior years, 2016 shared task participants are allowed to make use of existing resources and tools (e.g., WordNet, Mikolov et al. (2013)'s word2vec). Participants are also allowed to make unsupervised use of arbitrary data sets, even if such data overlaps with the announced sources of the evaluation data.

3 English Subtask

The English subtask builds on four prior years of English STS tasks. Task participants are allowed to use all of the trial, train and evaluation sets released during prior years as training and development data. As shown in table 2, this provides 14,250 paired snippets with gold STS labels. The 2015 STS task annotated between 1,500 and 2,000 pairs per data set that were then filtered based on annotation agreement and to achieve better data set balance. The raw annotations were released after the evaluation, providing an additional 5,500 pairs with noisy STS annotations (8,500 total for 2015).[2]

The 2016 English evaluation data is partitioned into five individual evaluation sets: Headlines, Plagiarism, Postediting, Answer-Answer and Question-Question. Each evaluation set has between 209 to 254 pairs. Participants annotate a larger number of pairs for each dataset without knowledge of what pairs will be included in the final evaluation.

[2] Answers-forums: 2000; Answers-students: 1500; Belief: 2000; Headlines: 1500; Images: 1500.

year	dataset	pairs	source
2012	MSRpar	1500	newswire
2012	MSRvid	1500	videos
2012	OnWN	750	glosses
2012	SMTnews	750	WMT eval.
2012	SMTeuroparl	750	WMT 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
2015	HDL	750	newswire headlines
2015	Images	750	image descriptions
2015	Ans.-student	750	student answers
2015	Ans.-forum	375	Q&A forum answers
2015	Belief	375	committed belief
2016	HDL	249	newswire headlines
2016	Plagiarism	230	short-answer plag.
2016	Postediting	244	MT postedits
2016	Ans.-Ans.	254	Q&A forum answers
2016	Quest.-Quest.	209	Q&A forum questions

Table 2: English subtask: Train (2012, 2013, 2014, 2015) and test (2016) data sets.

3.1 Data Collection

The data for the English evaluation sets are collected from a diverse set of sources. Data sources are selected that correspond to potentially useful domains for application of the semantic similarity methods explored in STS systems. This section details the pair selection heuristics as well as the individual data sources we use for the evaluation sets.

3.1.1 Selection Heuristics

Unless otherwise noted, pairs are heuristically selected using a combination of lexical surface form and word embedding similarity between a candidate pair of text snippets. The heuristics are used to find pairs sharing some minimal level of either surface or embedding space similarity. An approximately equal number of candidate sentence pairs are produced using our lexical surface form and word embedding selection heuristics. Both heuristics make use of a Penn Treebank style tokenization of the text provided by CoreNLP (Manning et al., 2014).

year	dataset	pairs	source
2016	Trial	103	Sampled $\leq$ 2015 STS
2016	News	301	en-es news articles
2016	Multi-source	294	en news headlines, short-answer plag., MT postedits, Q&A forum answers, Q&A forum questions

Table 3: Spanish-English subtask: Trial and test data sets.

Surface Lexical Similarity Our surface form selection heuristic uses an information theoretic measure based on unigram overlap (Lin, 1998). As shown in equation (1), surface level lexical similarity between two snippets s_1 and s_2 is computed as a log probability weighted sum of the words common to both snippets divided by a log probability weighted sum of all the words in the two snippets.

$$\text{sim}_l(s_1, s_1) = \frac{2 \times \sum_{w \in s_1 \cap s_2} \log P(w)}{\sum_{w \in s_1} \log P(w) + \sum_{w \in s_2} \log P(w)} \quad (1)$$

Unigram probabilities are estimated over the evaluation set data sources and are computed without any smoothing.

Word Embedding Similarity As our second heuristic, we compute the cosine between a simple embedding space representation of the two text snippets. Equation (2) illustrates the construction of the snippet embedding space representation, $\mathbf{v}(s)$, as the sum of the embeddings for the individual words, $\mathbf{v}(w)$, in the snippet. The cosine similarity can then be computed as in equation (3).

$$\mathbf{v}(s) = \sum_{w \in s} \mathbf{v}(w) \quad (2)$$

$$\text{sim}_\text{v}(s_1, s_2) = \frac{\mathbf{v}(s_1)\mathbf{v}(s_2)}{||\mathbf{v}(s_1)||||\mathbf{v}(s_2)||} \quad (3)$$

Three hundred dimensional word embeddings are obtained by running the GloVe package (Pennington et al., 2014) with default parameters over all the data collected from the 2016 evaluation sources.[3]

[3]The evaluation source data contained only 10,352,554 tokens. This is small relative to the data sets used to train embedding space models that typically make use of $>$ 1B tokens. However, the resulting embeddings are found to be functionally

3.1.2 Newswire Headlines

The newswire headlines evaluational set is collected from the Europe Media Monitor (EMM) (Best et al., 2005) using the same extraction approach taken for STS 2015 (Agirre et al., 2015) over the date range July 28, 2014 to April 10, 2015. The EMM clusters identify related news stories. To construct the STS pairs, we extract 749 pairs of headlines that appear in same cluster and 749 pairs of headlines associated with stories that appear in different clusters. For both groups, we use a surface level string similarity metric found in the Perl package `String::Similarity` (Myers, 1986) to select an equal number of pairs with high and low surface similarity scores.

3.1.3 Plagiarism

The plagiarism evaluation set is based on Clough and Stevenson (2011)'s Corpus of Plagiarised Short Answers. This corpus provides a collection of short answers to computer science questions that exhibit varying degrees of plagiarism from related Wikipedia articles.[4] The short answers include text that was constructed by each of the following four strategies: 1) copying and pasting individual sentences from Wikipedia; 2) light revision of material copied from Wikipedia; 3) heavy revision of material from Wikipedia; 4) non-plagiarised answers produced without even looking at Wikipedia. This corpus is segmented into individual sentences using CoreNLP (Manning et al., 2014).

3.1.4 Postediting

The Specia (2011) EAMT 2011 corpus provides machine translations of French news data using the Moses machine translation system (Koehn et al., 2007) paired with postedited corrections of those translations.[5] The corrections were provided by human translators instructed to perform the minimum number of changes necessary to produce a publishable translation. STS pairs for this evaluation set are selected both using the surface form and embedding space pairing heuristics and by including the existing explicit pairs of each machine translation with its postedited correction.

3.1.5 Question-Question & Answer-Answer

The question-question and answer-answer evaluation sets are extracted from the Stack Exchange Data Dump (Stack Exchange, Inc., 2016). The data include long form Question-Answer pairs on a diverse set of topics ranging from highly technical areas such as programming, physics and mathematics to more casual topics like cooking and travel.

Pairs are constructed using questions and answers from the following less technical Stack Exchange sites: *academia*, *cooking*, *coffee*, *diy*, *english*, *fitness*, *health*, *history*, *lifehacks*, *linguistics*, *money*, *movies*, *music*, *outdoors*, *parenting*, *pets*, *politics*, *productivity*, *sports*, *travel*, *workplace* and *writers*. Since both the questions and answers are long form, often being a paragraph in length or longer, heuristics are used to select a one sentence summary of each question and answer. For questions, we use the title of the question when it ends in a question mark.[6] For answers, a one sentence summary of each question is constructed using LexRank (Erkan and Radev, 2004) as implemented by the Sumy[7] package.

4 Cross-lingual Subtask

The pilot cross-lingual subtask explores the expansion of STS to paired snippets of text in different languages. The 2016 shared task pairs snippets in Spanish and English, with each pair containing exactly one Spanish and one English member. A trial set of 103 pairs was released prior to the official evaluation window containing pairs of sentences randomly selected from prior English STS evaluations, but with one of the snippets being translated into Spanish by human translators.[8] The similarity scores associated with this set are taken from the manual STS annotations within the original English data.

useful for finding semantically similar text snippets that differ in surface form.

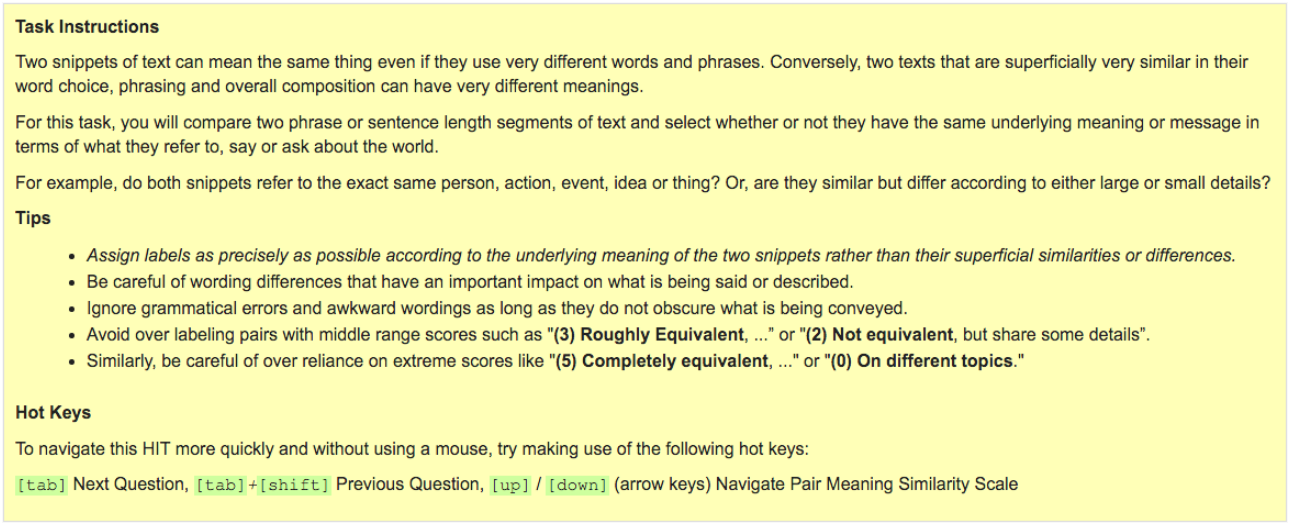
[4]Questions: A. What is inheritance in object orientated programming?, B. Explain the PageRank algorithm that is used by the Google search engine, C. Explain the Vector Space Model that is used for Information Retrieval., D. Explain Bayes Theorem from probability theory, E. What is dynamic programming?

[5]The corpus also includes English news data machine translated into Spanish and the postedited corrections of these translations. We use the English-Spanish data in the cross-lingual task.

[6]Questions with titles not ending in a "?" are discarded.

[7]`https://pypi.python.org/pypi/sumy`

[8]SDL was used to translate the trial data set, `http://sdl.com`.



Task Instructions

Two snippets of text can mean the same thing even if they use very different words and phrases. Conversely, two texts that are superficially very similar in their word choice, phrasing and overall composition can have very different meanings.

For this task, you will compare two phrase or sentence length segments of text and select whether or not they have the same underlying meaning or message in terms of what they refer to, say or ask about the world.

For example, do both snippets refer to the exact same person, action, event, idea or thing? Or, are they similar but differ according to either large or small details?

Tips

- *Assign labels as precisely as possible according to the underlying meaning of the two snippets rather than their superficial similarities or differences.*
- Be careful of wording differences that have an important impact on what is being said or described.
- Ignore grammatical errors and awkward wordings as long as they do not obscure what is being conveyed.
- Avoid over labeling pairs with middle range scores such as "**(3) Roughly Equivalent**, ..." or "**(2) Not equivalent**, but share some details".
- Similarly, be careful of over reliance on extreme scores like "**(5) Completely equivalent**, ..." or "**(0) On different topics**."

Hot Keys

To navigate this HIT more quickly and without using a mouse, try making use of the following hot keys:

`[tab]` Next Question, `[tab]`+`[shift]` Previous Question, `[up]` / `[down]` (arrow keys) Navigate Pair Meaning Similarity Scale

Figure 1: Annotation instructions for the STS English subtask.

Participants are allowed to use the labeled STS pairs from any of the prior STS evaluations. This includes STS pairs from all four prior years of the English STS subtasks as well as data from the 2014 and 2015 Spanish STS subtasks.

4.1 Data Collection

The cross-lingual evaluation data is partitioned into two evaluation sets: news and multi-source. The news data set is manually harvested from multilingual news sources, while the multi-source dataset is sampled from the same sources as the 2016 English data, with one of the snippets being translated into Spanish by human translators.[9] As shown in Table 3, the news set has 301 pairs, while the multi-source set has 294 pairs. For the news evaluation set, participants are provided with exactly the 301 pairs that will be used for the final evaluation. For the multi-source dataset, we take the same approach as the English subtask and release 2,973 pairs for annotation by participant systems, without providing information on what pairs will be included in the final evaluation.

4.1.1 Cross-lingual News

The cross-lingual news dataset is manually culled from less mainstream news sources such as Russia Today[10], in order to pose a more natural challenge in terms of machine translation accuracy. Articles on the same or differing topics are collected, with particular effort being spent to find articles on the same or somewhat similar story (many times written by the same author in English and Spanish), that exhibit a natural writing pattern in each language by itself and do not amount to an exact translation. When compared across the two languages, such articles exhibit different sentence structure and length. Additional paragraphs are also included by the writer that would cater to the readers' interests. For example, in the case of articles written about the Mexican drug lord Joaquin "El Chapo" Guzman, who was recently captured, the English articles typically have less extraneous details, focusing more on facts, while the articles written in Spanish provide additional background information with more narrative. Such articles allow for the manual extraction of high quality pairs that enable a wider variety of testing scenarios: from exact translations, to paraphrases exhibiting a different sentence structure, to somewhat similar sentences, to sentences sharing common vocabulary but no topic similarity, and ultimately to completely unrelated sentences. This ensures that semantic similarity systems that rely heavily on lexical features (which have been also typically used in STS tasks to derive test and train datasets) are at

[9]The evaluation set was translated into Spanish by Gengo: `http://gengo.com/`

[10]English: `https://www.rt.com/`, Spanish: `https://actualidad.rt.com/`

a disadvantage, and rather systems that actually explore semantic information receive due credit.

4.1.2 Multi-source

The raw multi-source data sets annotated by participating systems are constructed by first sampling 250 pairs from each of the following four data sets from the English task: Answer-Answer, Plagiarism, Question-Question and Headlines. One sentence from each sampled pair is selected at random for translation into Spanish by human translators.[11] An additional 1973 pairs are drawn from the English-Spanish section of EAMT 2011. We include all pairings of English source sentences with their human post-edited Spanish translations, resulting in 1000 pairs. We also include pairings of English source sentences with their Spanish machine translations. This only produced an additional 973 pairs, since 27 of the pairs are already generated by human post-edited translations that exactly match their corresponding machine translations. The gold standard data are selected by randomly drawing 60 pairs belonging to each data set within the raw multi-source data, except for EMM where only 54 pairs were drawn.[12]

5 Annotation

Annotation of pairs with STS scores is performed using crowdsourcing on Amazon Mechanical Turk.[13] This section describes the templates and annotation parameters we use for the English and cross-lingual Spanish-English pairs, as well as how the gold standard annotations are computed from multiple annotations from crowd workers.

5.1 English Subtask

The annotation instructions for the English subtask are modified from prior years in order to accommodate the annotation of question-question pairs. Figure 1 illustrates the new instructions. References to *statements* are replaced with *snippets*. The new instructions remove the wording suggesting that annotators "picture what is being described" and provide tips for navigating the annotation form quickly. The annotation form itself is also modified from prior years to make use of radio boxes to annotate the similarity scale rather than drop-down lists.

The English STS pairs are annotated in batches of 20 pairs. For each batch, annotators are paid $1 USD. Five annotations are collected per pair. Only workers with the MTurk *master* qualification are allowed to perform the annotation, a designation by the MTurk platform that statistical identifies workers who perform high quality work across a diverse set of MTurk tasks. Gold annotations are selected as the median value of the crowdsourced annotations after filtering out low quality annotators. We remove annotators with correlation scores < 0.80 using a simulated gold annotation computed by leaving out the annotations from the worker being evaluated. We also exclude all annotators with a kappa score < 0.20 against the same simulated gold standard.

The official task evaluation data are selected from pairs having at least three remaining labels after excluding the low quality annotators. For each evaluation set, we attempt to select up to 42 pairs for each STS label. Preference was given to pairs with a higher number of STS labels matching the median label. After the final pairs are selected, they are spot checked with some of the pairs having their STS score corrected.

5.2 Cross-lingual Spanish-English Subtask

The Spanish-English pairs are annotated using a slightly modified template from the 2014 and 2015 Spanish STS subtask. Given the multilingual nature of the subtask, the guidelines consist of alternating instructions in either English or Spanish, in order to dissuade monolingual annotators from participating (see Figure 2). The template is also modified to use the same six point scale used by the English subtask, rather than the five point scale used in the Spanish subtasks in the past (which did not attempt to distinguish between differences in unimportant details). Judges are also presented with the cross-lingual example pairs and explanations listed on Table 1.

The cross-lingual pairs are annotated in batches of 7 pairs. Annotators are paid $0.30 USD per batch and each batch receives annotations from 5 workers. The annotations are restricted to workers who

[11] Inspection of the data suggests the translation service provider may have used a postediting based process.

[12] Our annotators work on batches of 7 pairs. Drawing 54 pairs from the EMM data results in a total number pairs that is cleanly divisible by 7.

[13] https://www.mturk.com/

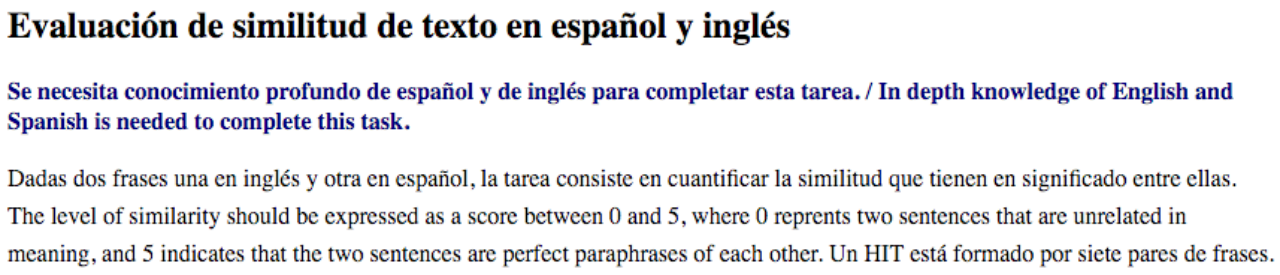


Figure 2: Annotation instructions for the STS Spanish-English subtask.

have completed 500 HITs on the MTurk platform and have less than 10% of their lifetime annotations rejected. The gold standard is computed by averaging over the 5 annotations collected for each pair.

6 System Evaluation

This section reports the evaluation results for the 2016 STS English and cross-lingual Spanish-English subtasks.

6.1 Participation

Participating teams are allowed to submit up to three systems.[14] For the English subtask, there were 119 systems from 43 participating teams. The cross-lingual Spanish-English subtask saw 26 submissions from 10 teams. For the English subtask, this is a 45% increase in participating teams from 2015. The Spanish-English STS pilot subtask attracted approximately 53% more participants than the monolingual Spanish subtask organized in 2015.

6.2 Evaluation Metric

On each test set, systems are evaluated based on their Pearson correlation with the gold standard STS labels. The overall score for each system is computed as the average of the correlation values on the individual evaluation sets, weighted by the number of data points in each evaluation set.

6.3 Baseline

Similar to prior years, we include a baseline built using a very simple vector space representation. For this baseline, both text snippets in a pair are first tokenized by white-space. The snippets are then projected to a one-hot vector representation such that each dimension corresponds to a word observed in one of the snippets. If a word appears in a snippet one or more times, the corresponding dimension in the vector is set to one and is otherwise set to zero. The textual similarity score is then computed as the cosine between these vector representations of the two snippets.

6.4 English Subtask

The rankings for the English STS subtask are given in Tables 4 and 5. The baseline system ranked 100th. Table 6 provides the best and median scores for each of the individual evaluation sets as well as overall.[15] The table also provides the difference between the best and median scores to highlight the extent to which top scoring systems outperformed the typical level of performance achieved on each data set.

The best overall performance is obtained by Samsung Poland NLP Team's EN1 system, which achieves an overall correlation of 0.778 (Rychalska et al., 2016). This system also performs best on three out of the five individual evaluation sets: answer-answer, headlines, plagiarism. The EN1 system achieves competitive performance on the postediting data with a correlation score of 0.83516. The best system on the postediting data, RICOH's Run-n (Itoh, 2016), obtains a score of 0.867. Like all systems, EN1 struggles on the question-question data, achieving a correlation of 0.687. Another system submitted by the Samsung Poland NLP Team named

[14]For the English subtask, the SimiHawk team was granted permission to make 4 submissions as the team has 3 submissions using distinct machine learning models (feature engineered, LSTM, and Tree LSTM) and they asked to separately submit an ensemble of the three methods.

[15]The median scores reported here do not include late or corrected systems. The median scores for the on-time systems without corrections are: ALL 0.68923; plagiarism 0.78949; answer-answer 0.48018; postediting 0.81241; headlines 0.76439; question-question 0.57140.

Team	Run	ALL	Ans.-Ans.	HDL	Plagiarism	Postediting	Ques.-Ques.	Run Rank	Team Rank
Samsung Poland NLP Team	EN1	**0.77807**	**0.69235**	**0.82749**	**0.84138**	0.83516	0.68705	1	1
UWB	sup-general	0.75731	0.62148	0.81886	0.82355	0.82085	0.70199	2	2
MayoNLPTeam	Run3	0.75607	0.61426	0.77263	0.80500	0.84840	**0.74705**	3	3
Samsung Poland NLP Team	EN2	0.75468	**0.69235**	**0.82749**	0.81288	0.83516	0.58567	4	
NaCTeM	micro+macro	0.74865	0.60237	0.80460	0.81478	0.82858	0.69367	5	4
ECNU	S1-All	0.75079	0.56979	0.81214	0.82503	0.82342	0.73116	5*	4*
UMD-TTIC-UW	Run1	0.74201	0.66074	0.79457	0.81541	0.80939	0.61872	6	5
SimiHawk	Ensemble	0.73774	0.59237	0.81419	0.80566	0.82179	0.65048	7	6
MayoNLPTeam	Run2	0.73569	0.57739	0.75061	0.80068	0.82857	0.73035	8	
Samsung Poland NLP Team	AE	0.73566	0.65769	0.81801	0.81288	0.78849	0.58567	9	
DLS@CU	Run1	0.73563	0.55230	0.80079	0.82293	0.84258	0.65986	10	7
DLS@CU	Run3	0.73550	0.54528	0.80334	0.81949	0.84418	0.66657	11	
DTSim	Run1	0.73493	0.57805	0.81527	0.83757	0.82286	0.61428	12	8
NaCTeM	macro	0.73391	0.58484	0.79756	0.78949	0.82614	0.67039	13	
DLS@CU	Run2	0.73297	0.55992	0.80334	0.81227	0.84418	0.64234	14	
Stasis	xgboost	0.73050	0.50628	0.77824	0.82501	0.84861	0.70424	15	9
IHS-RD-Belarus	Run1	0.72966	0.55322	0.82419	0.82634	0.83761	0.59904	16	10
USFD	COMB-Features	0.72869	0.50850	0.82024	0.83828	0.79496	0.68926	17	11
USFD	CNN	0.72705	0.51096	0.81899	0.83427	0.79268	0.68551	18	
saarsheff	MT-Metrics-xgboost	0.72693	0.47716	0.78848	0.83212	0.84960	0.69815	19	12
MayoNLPTeam	Run1	0.72646	0.58873	0.73458	0.76887	0.85020	0.69306	20	
UWB	unsup	0.72622	0.64442	0.79352	0.82742	0.81209	0.53383	21	
UMD-TTIC-UW	Run2	0.72619	0.64427	0.78708	0.79894	0.79338	0.59468	22	
SERGIOJIMENEZ	Run2	0.72617	0.55257	0.78304	0.81505	0.81634	0.66630	23	13
IHS-RD-Belarus	Run2	0.72465	0.53722	0.82539	0.82558	0.83654	0.59072	24	
DTSim	Run3	0.72414	0.56189	0.81237	0.83239	0.81498	0.59103	25	
ECNU	U-SEVEN	0.72427	0.47748	0.76681	0.83013	0.84239	0.71914	25*	
SERGIOJIMENEZ	Run1	0.72411	0.50182	0.78646	0.83654	0.83638	0.66519	26	
NaCTeM	Micro	0.72361	0.55214	0.79143	0.83134	0.82660	0.61241	27	
SERGIOJIMENEZ	Run3	0.72215	0.49068	0.77725	0.82926	0.84807	0.67291	28	
DTSim	Run2	0.72016	0.55042	0.79499	0.82815	0.81508	0.60766	29	
DCU-SEManiacs	Fusion	0.71701	0.58328	0.76392	0.81386	0.84662	0.56576	30	14
DCU-SEManiacs	Synthetic	0.71334	0.68762	0.72227	0.81935	0.80900	0.50560	31	
RICOH	Run-b	0.71165	0.50871	0.78691	0.82661	0.86554	0.56245	32	15
ECNU	S2	0.71175	0.57158	0.79036	0.77338	0.74968	0.67635	32*	
HHU	Overlap	0.71134	0.50435	0.77406	0.83049	0.83846	0.60867	33	16
UMD-TTIC-UW	Run3	0.71112	0.64316	0.77801	0.78158	0.77786	0.55855	34	
University of Birmingham	CombineFeatures	0.70940	0.52460	0.81894	0.82066	0.81272	0.56040	35	17
University of Birmingham	MethodsFeatures	0.70911	0.52028	0.81894	0.81958	0.81333	0.56451	36	
SimiHawk	F	0.70647	0.44003	0.77109	0.81105	0.81600	0.71035	37	
UWB	sup-try	0.70542	0.53333	0.77846	0.74673	0.78507	0.68909	38	
Stasis	boostedtrees	0.70496	0.40791	0.77276	0.82903	0.84635	0.68359	39	
RICOH	Run-n	0.70467	0.50746	0.77409	0.82248	**0.86690**	0.54261	40	
Stasis	linear	0.70461	0.36929	0.76660	0.82730	0.83917	0.74615	41	
RICOH	Run-s	0.70420	0.51293	0.78000	0.82991	0.86252	0.52319	42	
University of Birmingham	CombineNoFeatures	0.70168	0.55217	0.82352	0.82406	0.80835	0.47904	43	
MathLingBudapest	Run1	0.70025	0.40540	0.81187	0.80752	0.83767	0.64712	44	18
ISCAS_NLP	S1	0.69996	0.49378	0.79763	0.81933	0.81185	0.57218	45	19
ISCAS_NLP	S3	0.69996	0.49378	0.79763	0.81933	0.81185	0.57218	46	
UNBNLP	Regression	0.69940	0.55254	0.71353	0.79769	0.81291	0.62037	47	20
DCU-SEManiacs	task-internal	0.69924	0.62702	0.71949	0.80783	0.80854	0.51580	48	
MathLingBudapest	Run2	0.69853	0.40540	0.80367	0.80752	0.83767	0.64712	49	
MathLingBudapest	Run3	0.69853	0.40540	0.80366	0.80752	0.83767	0.64712	50	
ISCAS_NLP	S2	0.69756	0.49651	0.79041	0.81214	0.81181	0.57181	51	
UNBNLP	Average	0.69635	0.58520	0.69006	0.78923	0.82540	0.58605	52	
NUIG-UNLP	m5all3	0.69528	0.40165	0.75400	0.80332	0.81606	0.72228	53	21
wolvesaar	xgboost	0.69471	0.49947	0.72410	0.79076	0.84093	0.62055	54	22
wolvesaar	lotsa-embeddings	0.69453	0.49415	0.71439	0.79655	0.83758	0.63509	55	
Meiji-WSL-A	Run1	0.69435	0.58260	0.74394	0.79234	0.85962	0.47030	56	23
saarsheff	MT-Metrics-boostedtrees	0.69259	0.37717	0.77183	0.81529	0.84528	0.66825	57	
wolvesaar	DLS-replica	0.69244	0.48799	0.71043	0.80605	0.84601	0.61515	58	
saarsheff	MT-Metrics-linear	0.68923	0.31539	0.76551	0.82063	0.83329	0.73987	59	
EECS	Run2	0.68430	0.48013	0.77891	0.76676	0.82965	0.55926	60	24
NUIG-UNLP	m5dom1	0.68368	0.41211	0.76778	0.75539	0.80086	0.69782	61	
EECS	Run3	0.67906	0.47818	0.77719	0.77266	0.83744	0.51840	62	
PKU	Run1	0.67852	0.47469	0.77881	0.77479	0.81472	0.54180	63	25
EECS	Run1	0.67711	0.48110	0.77739	0.76747	0.83270	0.51479	64	
PKU	Run2	0.67503	0.47444	0.77703	0.78119	0.83051	0.49892	65	
HHU	SameWordsNeuralNet	0.67502	0.42673	0.75536	0.79964	0.84514	0.54533	66	
PKU	Run3	0.67209	0.47271	0.77367	0.77580	0.81185	0.51611	67	
NSH	Run1	0.66181	0.39962	0.74549	0.80176	0.79540	0.57080	68	26
RTM	SVR	0.66847	0.44865	0.66338	0.80376	0.81327	0.62374	68*	26*
Meiji-WSL-A	Run2	0.65871	0.51675	0.58561	0.78700	0.81873	0.59035	69	
BIT	Align	0.65318	0.54530	0.78140	0.80473	0.79456	0.29972	70	27
UNBNLP	tf-idf	0.65271	0.45928	0.66593	0.75778	0.77204	0.61710	71	
UTA_MLNLP	100-1	0.64965	0.46391	0.74499	0.74003	0.71947	0.58083	72	28
RTM	FS+PLS-SVR	0.65237	0.35333	0.65294	0.80488	0.82304	0.64803	72*	
RTM	PLS-SVR	0.65182	0.34401	0.66051	0.80641	0.82314	0.64544	72*	
SimiHawk	LSTM	0.64840	0.44177	0.75703	0.71737	0.72317	0.60691	73	
BIT	VecSim	0.64661	0.48863	0.62804	0.80106	0.79544	0.51702	74	
NUIG-UNLP	m5dom2	0.64520	0.38303	0.76485	0.74351	0.76549	0.57263	75	

Table 4: STS 2016: English Rankings. Late or corrected systems are marked with a * symbol.

Team	Run	ALL	Ans.-Ans.	HDL	Plagiarism	Postediting	Ques.-Ques.	Run Rank	Team Rank
UTA_MLNLP	150-1	0.64500	0.43042	0.72133	0.71620	0.74471	0.62006	76	
SimiHawk	TreeLSTM	0.64140	0.52277	0.74083	0.67628	0.70655	0.55265	77	
NORMAS	SV-2	0.64078	0.36583	0.68864	0.74647	0.80234	0.61300	78	29
UTA_MLNLP	150-3	0.63698	0.41871	0.72485	0.70296	0.69652	0.65543	79	
LIPN-IIMAS	SOPA	0.63087	0.44901	0.62411	0.69109	0.79864	0.59779	80	30
NORMAS	ECV-3	0.63072	0.27637	0.72245	0.72496	0.79797	0.65312	81	
JUNITMZ	Backpropagation-1	0.62708	0.48023	0.70749	0.72075	0.77196	0.43751	82	31
NSH	Run2	0.62941	0.34172	0.74977	0.75858	0.82471	0.46548	82*	
LIPN-IIMAS	SOPA1000	0.62466	0.44893	0.59721	0.75936	0.76157	0.56285	83	
USFD	Word2Vec	0.62254	0.27675	0.64217	0.78755	0.75057	0.68833	84	
HHU	DeepLDA	0.62078	0.47211	0.58821	0.62503	0.84743	0.57099	85	
ASOBEK	T11	0.61782	0.52277	0.63741	0.78521	0.84245	0.26352	86	32
ASOBEK	M11	0.61430	0.47916	0.68652	0.77779	0.84089	0.24804	87	
LIPN-IIMAS	SOPA100	0.61321	0.43216	0.58499	0.74727	0.75560	0.55310	88	
Telkom University	WA	0.60912	0.28859	0.69988	0.69090	0.74654	0.64009	89	33
BIT	WeightedVecSim	0.59560	0.37565	0.55925	0.75594	0.77835	0.51643	90	
3CFEE	grumlp	0.59603	0.36521	0.72092	0.74210	0.76327	0.37179	90*	34*
ASOBEK	F1	0.59556	0.42692	0.67898	0.75717	0.81950	0.26181	91	
JUNITMZ	Recurrent-1	0.59493	0.44218	0.66120	0.73708	0.69279	0.43092	92	
VRep	withLeven	0.58292	0.30617	0.68745	0.69762	0.73033	0.49639	93	34
Meiji_WSL_teamC	Run1	0.58169	0.53250	0.64567	0.74233	0.54783	0.42797	94	35
JUNITMZ	FeedForward-1	0.58109	0.40859	0.66524	0.76752	0.66522	0.38711	95	
VRep	withStopRem	0.57805	0.29487	0.68185	0.69730	0.72966	0.49029	96	
VRep	noStopRem	0.55894	0.34684	0.67856	0.69768	0.74088	0.30908	97	
UNCC	Run-3	0.55789	0.31111	0.59592	0.64672	0.75544	0.48409	98	36
Telkom University	CS	0.51602	0.06623	0.72668	0.50534	0.73999	0.56194	99	
NORMAS	RF-1	0.50895	0.16095	0.58800	0.62134	0.72016	0.46743	100	
STS Organizers	*baseline*	0.51334	0.41133	0.54073	0.69601	0.82615	0.03844	100†	37†
meijiuniversity_teamb	4features_^LCS_pos	0.45748	0.26182	0.60223	0.55931	0.66989	0.16276	101	37
meijiuniversity_teamb	5features	0.45656	0.26498	0.60000	0.55231	0.66894	0.16518	102	
meijiuniversity_teamb	4features_^LCS	0.45621	0.26588	0.59868	0.55687	0.66690	0.16103	103	
Amrita_CEN	SEWE-2	0.40951	0.30309	0.43164	0.63336	0.66465	-0.03174	104	38
VENSESEVAL	Run1	0.44258	0.41932	0.62738	0.64538	0.27274	0.22581	104*	38*
UNCC	Run-1	0.40359	0.19537	0.49344	0.38881	0.59235	0.34548	105	
UNCC	Run-2	0.37956	0.18100	0.50924	0.26190	0.56176	0.38317	106	
WHU_NLP	CNN20	0.11100	-0.02488	0.16026	0.16854	0.24153	0.00176	107	39
3CFEE	bowkst	0.33174	0.20076	0.42288	0.50150	0.19111	0.35970	107*	
WHU_NLP	CNN10	0.09814	0.05402	0.19650	0.08949	0.16152	-0.02989	108	
DalGTM	Run1	0.05354	-0.00557	0.29896	-0.07016	-0.04933	0.08924	109	40
DalGTM	Run2	0.05136	-0.00810	0.28962	-0.07032	-0.04818	0.08987	110	
DalGTM	Run3	0.05027	-0.00780	0.28501	-0.07134	-0.05014	0.09223	111	
WHU_NLP	CNN5	0.04713	0.00041	0.13780	0.03264	0.12669	-0.08105	112	
IHS-RD-Belarus	Run3					0.83761		113	

Table 5: STS 2016: English Rankings (continued). As before, late or corrected systems are marked with a * symbol. The baseline run by the organizers is marked with a † symbol (at rank 100).

Evaluation Set	Best	Median	Δ
ALL	0.77807	0.69347	0.08460
Ans.-Ans.	0.69235	0.48067	0.21169
HDL	0.82749	0.76439	0.06311
Plagiarism	0.84138	0.79445	0.04694
Postediting	0.86690	0.81272	0.05418
Ques.-Ques.	0.74705	0.57673	0.17032

Table 6: Best and median scores by evaluation set for the English subtask as well as the difference between the two, Δ.

EN2 achieves the best correlation on the question-question data at 0.747.

The most difficult data sets this year are the two Q&A evaluation sets: answer-answer and question-question. Difficulty on the question-question data was expected as this is the first year that question-question pairs are formally included in the evaluation of STS systems.[16] Interestingly, the baseline system has particular problems on this data, achieving a correlation of only 0.038. This suggests that surface overlap features might be less informative on this data, possibly leading to prediction errors by the systems that include them. An answer-answer evaluation set was previously included in the 2015 STS task. Answer-answer data is included again in this year's evaluation specifically because of the poor performance observed on this type of data in 2015.

In Table 6, it can be seen that the difficult Q&A data sets are also the data sets that exhibit the biggest difference between the top performing system for that evaluation set and typical system performance, as capture by the median on the same data. For the easier data sets, news headlines, plagiarism, and postediting, there is only a relatively modest gap between the best system for that data set and typical system performance ranging from about 0.05 to 0.06. However, on the difficult Q&A data set the difference between the best system and typical systems jumps to 0.212 for answer-answer and 0.170 for question-question. This suggests these harder data sets may be better at discriminating between different approaches with most systems now being fairly competent on assessing easier pairs.[17]

6.4.1 Methods

Participating systems vary greatly in the approaches they take to solving STS. The overall winner, Samsung Poland NLP Team, proposes a textual similarity model that is a novel hybrid of recursive auto-encoders from deep learning with penalty and reward signals extracted from WordNet (Rychalska et al., 2016). To obtain even better performance, this model is combined in an ensemble with a number of other similarity models including a version of Sultan et al. (2015)'s very successful STS model enhanced with additional features found to work well in the literature.

The team in second place overall, UWB, combines a large number of diverse similarity models and features (Brychcin and Svoboda, 2016). Similar to Samsung, UWB includes both manually engineered NLP features (e.g., character n-gram overlap) with sophisticated models from deep learning (e.g., Tree LSTMs). The third place team, MayoNLPTeam, also achieves their best results using a combination of a more traditionally engineered NLP pipeline with a deep learning based model (Afzal et al., 2016). Specifically, MayoNLPTeam combines a pipeline that makes use of linguistic resources such as WordNet and well understood concepts such as the information content of a word (Resnik, 1995) with a deep learning method known as Deep Structured Semantic Model (DSSM) (Huang et al., 2013).

The next two teams in overall performance, ECNU and NaCTeM, make use of large feature sets, including features based on word embeddings. However, they did not incorporate the more sophisticated deep learning based models explored by Samsung, UWB and MayoNLPTeam (Tian and Lan, 2016; Przybyła et al., 2016).

The next team in the rankings, UMD-TTIC-UW, *only* makes use of a single deep learning model (He et al., 2016). The team extends a multi-perspective convolutional neural network (MPCNN) (He et al., 2015) with a simple word level attentional mecha-

[16]The pair selection criteria from prior years did not explicitly exclude the presence of questions or question-question pairs. Of the 19,189 raw pairs from prior English STS evaluations as trial, train or evaluation data, 831 (4.33%) of the pairs include a question mark within at least one member of the pair, while 319 of the pairs (1.66%) include a question mark within both members.

[17]To see how much of this is related to the Q&A domain in particular, we will investigate including difficult non-Q&A evaluation data in future STS competitions.

nism based on the aggregate cosine similarity of a word in one text with all of the words in a paired text. The submission is notable for how well it performs without any manual feature engineering.

Finally, the best performing system on the post-editing data, RICOH's Run-n, introduces a novel IR-based approach for textual similarity that incorporates word alignment information (Itoh, 2016).

6.5 Cross-lingual Spanish-English Subtask

The rankings for the cross-lingual Spanish-English STS subtask are provided in Table 7. Recall that the multi-source data is drawn from the same sources as the monolingual English STS pairs. It is interesting to note that the performance of cross-lingual systems on this evaluation set does not appear to be significantly worse than the monolingual submissions even though the systems are being asked to perform the more challenging problem of evaluating cross-lingual sentence pairs. While the correlations are not directly comparable, they do seem to motivate a more direct comparison between cross-lingual and monolingual STS systems.

In terms of performance on the manually culled news data set, the highest overall rank is achieved by an unsupervised system submitted by team UWB (Brychcin and Svoboda, 2016). The unsupervised UWB system builds on the word alignment based STS method proposed by Sultan et al. (2015). However, when calculating the final similarity score, it weights both the aligned and unaligned words by their inverse document frequency. This system is able to attain a 0.912 correlation on the news data, while ranking second on the multi-source data set. For the multi-source test set, the highest scoring submission is a supervised system from the UWB team that combines multiple signals originating from lexical, syntactic and semantic similarity approaches in a regression-based model, achieving a 0.819 correlation. This is modestly better than the second place unsupervised approach that achieves 0.808.

Approximately half of the submissions are able to achieve a correlation above 0.8 on the news data. On the multi-source data, the overall correlation trend is lower, but with half the systems still obtaining a score greater than 0.6. Due to the diversity of the material embedded in the multi-source data, it seems to amount to a more difficult testing scenario. Nonetheless, there are cases of: 1) systems performing much worse on the news data set: the FBK HLT-MT systems experience an approximately 0.25 drop in correlation on the news data as compare to the multi-source setting; 2) systems performing evenly on both data sets.

6.5.1 Methods

In terms of approaches, most runs rely on a monolingual framework. They automatically translate the Spanish member of a sentence pair into English and then compute monolingual semantic similarity using a system developed for English. In contrast, the CNRC team (Lo et al., 2016) provides a true cross-lingual system that makes use of embedding space phrase similarity, the score from XMEANT, a cross-lingual machine translation evaluation metric (Lo et al., 2014), and precision and recall features for material filling aligned cross-lingual semantic roles (e.g., action, agent, patient). The FBK_HLT team (Ataman et al., 2016) proposes a model combining cross-lingual word embeddings with features from QuEst (Specia et al., 2013), a tool for machine translation quality estimation. The RTM system (Biçici, 2016) also builds on methods developed for machine translation quality estimation and is applicable to both cross-lingual and monolingual similarity. The GWU_NLP team (Aldarmaki and Diab, 2016) uses a shared cross-lingual vector space to directly assess sentences originating in different languages.[18]

7 Conclusion

We have presented the results of the 2016 STS shared task. This year saw a significant increase in participation. There are 119 submissions from 43 participating teams for the English STS subtask. This is a 45% increase in participating teams over 2015. The pilot cross-lingual Spanish-English STS subtask has 26 submissions from 10 teams, which is impressive given that this is the first year such a challenging subtask was attempted. Interestingly, the cross-lingual STS systems appear to perform competitively to monolingual systems on pairs drawn from the same sources. This suggests that it would be interesting to perform a more direct comparison between cross-lingual and monolingual systems.

[18]The GWU_NLP team includes one of the STS organizers.

Team	Run	News	Multi-Src	Mean	Run Rank	Team Rank
UWB	sup	0.90621	**0.81899**	**0.86311**	1	1
UWB	unsup	**0.91237**	0.80818	0.86089	2	
SERGIOJIMENEZ	run1	0.88724	0.81836	0.85321	3	2
SERGIOJIMENEZ	run3	0.89651	0.80737	0.85246	4	
DCU-SEManiacs	run2	0.89739	0.79262	0.84562	5	3
DCU-SEManiacs	run1	0.89408	0.76927	0.83241	6	
SERGIOJIMENEZ	run2	0.82911	0.81266	0.82098	7	
GWU_NLP	run2_xWMF	0.86768	0.73189	0.80058	8	4
GWU_NLP	run1_xWMF	0.86626	0.69850	0.78337	9	
GWU_NLP	run3_bWMF	0.83474	0.72427	0.78015	10	
CNRC	MT1	0.87562	0.64583	0.76208	11	5
CNRC	MT2	0.87754	0.63141	0.75592	12	
CNRC	EMAP	0.71943	0.41054	0.56680	13	
RTM	FS+PLS-SVR	0.59154	0.52044	0.55641	14	6
RTM	FS-SVR	0.53602	0.52839	0.53225	15	
RTM	SVR	0.49849	0.52935	0.51374	16	
FBK_HLT	MT-run3	0.25507	0.53892	0.39533	17	7
FBK_HLT	MT-run1	0.24318	0.53465	0.38720	18	
FBK_HLT	MT-run2	0.24372	0.51420	0.37737	19	
LIPM-IIMAS	sopa	0.08648	0.15931	0.12247	20	8
WHU_NLP	CNN10	0.03337	0.04083	0.03706	21	9
WHU_NLP	CNN5	0	0.05512	0.02724	22	
WHU_NLP	CNN20	0.02428	-0.06355	-0.01912	23	
JUNITMZ	backpropagation1	-0.05676	-0.48389	-0.26781	24	10
JUNITMZ	backpropagation2	-0.34725	-0.39867	-0.37266	25	
JUNITMZ	backpropagation3	-0.52951	-0.39891	-0.46498	26	

Table 7: STS 2016: Cross-lingual Spanish-English Rankings

Acknowledgments

This material is based in part upon work supported by DARPA-BAA-12-47 DEFT grant to George Washington University, by DEFT grant #12475008 to the University of Michigan, and by a MINECO grant to the University of the Basque Country (TUNER project TIN2015-65308-C5-1-R). Aitor Gonzalez Agirre is supported by a doctoral grant from MINECO (PhD grant FPU12/06243). 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 Defense Advanced Research Projects Agency.

References

Naveed Afzal, Yanshan Wang, and Hongfang Liu. 2016. MayoNLP at SemEval-2016 Task 1: Semantic textual similarity based on lexical semantic net and deep learning semantic model. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.

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.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, CO.

Hanan Aldarmaki and Mona Diab. 2016. GWU NLP at SemEval-2016 Shared Task 1: Matrix factorization for crosslingual STS. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.

Duygu Ataman, Jose G. C. de Souza, and Marco Turchi1 Matteo Negri. 2016. FBK HLT-MT at SemEval-2016 Task 1: Cross-lingual semantic similarity measurement using quality estimation features and compositional bilingual word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *COLING*

'98 Proceedings of the 17th international conference on Computational linguistics - Volume 1.

Clive Best, Erik van der Goot, Ken Blackler, Teófilo Garcia, and David Horby. 2005. Europe Media Monitor - System description. In *EUR Report 22173-En*, Ispra, Italy.

Ergun Biçici. 2016. RTM at SemEval-2016 Task 1: Predicting semantic similarity with referential translation machines and related statistics. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.

Tomas Brychcin and Lukas Svoboda. 2016. UWB at SemEval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic information. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.

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

Paul Clough and Mark Stevenson. 2011. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45(1):5–24.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 16:105–105.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1576–1586, Lisbon, Portugal.

Hua He, John Wieting, Kevin Gimpel, Jinfeng Rao, and Jimmy Lin. 2016. UMD-TTIC-UW at SemEval-2016 Task 1: Attention-based multi-perspective convolutional neural networks for textual similarity measurement. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.

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*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. ACM International Conference on Information and Knowledge Management (CIKM).

Hideo Itoh. 2016. RICOH at SemEval-2016 Task 1: Ir-based semantic textual similarity estimation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. XMEANT: Better semantic MT evaluation without reference translations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Chi-kiu Lo, Cyril Goutte, and Michel Simard. 2016. CNRC at SemEval-2016 Task 1: Experiments in crosslingual semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Scottsdale, AZ, USA.

George A. Miller. 1995. WordNet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Eugene W. Myers. 1986. Ano(nd) difference algorithm and its variations. *Algorithmica*, 1(1):251–266.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Piotr Przybyła, Nhung T. H. Nguyen, Matthew Shardlow, Georgios Kontonatsios, and Sophia Ananiadou. 2016. NaCTeM at SemEval-2016 Task 1: Inferring sentence-level semantic similarity from an ensemble of complementary lexical and sentence-level features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453.

Barbara Rychalska, Katarzyna Pakulska, Krystyna Chodorowska, Wojciech Walczak, and Piotr Andruszkiewicz. 2016. Samsung Poland NLP Team at SemEval-2016 Task 1: Necessity for diversity; combining recursive autoencoders, wordnet and ensemble methods to measure semantic similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *15th Conference of the European Association for Machine Translation*, EAMT, pages 73–80, Leuven, Belgium.

Stack Exchange, Inc. 2016. Stack exchange data dump. `https://archive.org/details/stackexchange`.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, CO, USA.

Junfeng Tian and Man Lan. 2016. ECNU at SemEval-2016 Task 1: Leveraging word embedding from macro and micro views to boost performance for semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, CA, USA.