SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation

Daniel Cer[a], **Mona Diab**[b], **Eneko Agirre**[c], **Iñigo Lopez-Gazpio**[c], **and Lucia Specia**[d]

[a]Google Research
Mountain View, CA

[b]George Washington University
Washington, DC

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

[d]University of Sheffield
Sheffield, UK

Abstract

Semantic Textual Similarity (STS) measures the meaning similarity of sentences. Applications include machine translation (MT), summarization, generation, question answering (QA), short answer grading, semantic search, dialog and conversational systems. The STS shared task is a venue for assessing the current state-of-the-art. The 2017 task focuses on multilingual and cross-lingual pairs with one sub-track exploring MT quality estimation (MTQE) data. The task obtained strong participation from 31 teams, with 17 participating in *all language tracks*. We summarize performance and review a selection of well performing methods. Analysis highlights common errors, providing insight into the limitations of existing models. To support ongoing work on semantic representations, the *STS Benchmark* is introduced as a new shared training and evaluation set carefully selected from the corpus of English STS shared task data (2012-2017).

1 Introduction

Semantic Textual Similarity (STS) assesses the degree to which two sentences are semantically equivalent to each other. The STS task is motivated by the observation that accurately modeling the meaning similarity of sentences is a foundational language understanding problem relevant to numerous applications including: machine translation (MT), summarization, generation, question answering (QA), short answer grading, semantic search, dialog and conversational systems. STS enables the evaluation of techniques from a diverse set of domains against a shared interpretable performance criteria. Semantic inference tasks related to STS include textual entailment (Bentivogli et al., 2016; Bowman et al., 2015; Dagan et al., 2010), semantic relatedness (Bentivogli et al., 2016) and paraphrase detection (Xu et al., 2015; Ganitkevitch et al., 2013; Dolan et al., 2004). STS differs from both textual entailment and paraphrase detection in that it captures *gradations of meaning overlap* rather than making binary classifications of particular relationships. While semantic relatedness expresses a graded semantic relationship as well, it is non-specific about the nature of the relationship with contradictory material still being a candidate for a high score (e.g., "night" and "day" are highly related but not particularly similar).

To encourage and support research in this area, the STS shared task has been held annually since 2012, providing a venue for evaluation of state-of-the-art algorithms and models (Agirre et al., 2012, 2013, 2014, 2015, 2016). During this time, diverse similarity methods and data sets[1] have been explored. Early methods focused on lexical semantics, surface form matching and basic syntactic similarity (Bär et al., 2012; Šarić et al., 2012a; Jimenez et al., 2012a). During subsequent evaluations, strong new similarity signals emerged, such as Sultan et al. (2015)'s alignment based method. More recently, deep learning became competitive with top performing feature engineered systems (He et al., 2016). The best performance tends to be obtained by ensembling feature engineered and deep learning models (Rychalska et al., 2016).

Significant research effort has focused on STS over English sentence pairs.[2] English STS is a

[1]i.a., news headlines, video and image descriptions, glosses from lexical resources including WordNet (Miller, 1995; Fellbaum, 1998), FrameNet (Baker et al., 1998), OntoNotes (Hovy et al., 2006), web discussion fora, plagiarism, MT post-editing and Q&A data sets. Data sets are summarized on: http://ixa2.si.ehu.es/stswiki.

[2]The 2012 and 2013 STS tasks were English only. The 2014 and 2015 task included a Spanish track and 2016 had a

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 1–14,
Vancouver, Canada, August 3 - 4, 2017.

well-studied problem, with state-of-the-art systems often achieving 70 to 80% correlation with human judgment. To promote progress in other languages, the 2017 task emphasizes performance on Arabic and Spanish as well as cross-lingual pairings of English with material in Arabic, Spanish and Turkish. The *primary* evaluation criteria combines performance on all of the different language conditions except English-Turkish, which was run as a surprise language track. Even with this departure from prior years, the task attracted 31 teams producing 84 submissions.

STS shared task data sets have been used extensively for research on sentence level similarity and semantic representations (i.a., Arora et al. (2017); Conneau et al. (2017); Mu et al. (2017); Pagliardini et al. (2017); Wieting and Gimpel (2017); He and Lin (2016); Hill et al. (2016); Kenter et al. (2016); Lau and Baldwin (2016); Wieting et al. (2016b,a); He et al. (2015); Pham et al. (2015)). To encourage the use of a common evaluation set for assessing new methods, we present the STS Benchmark, a publicly available selection of data from English STS shared tasks (2012-2017).

2 Task Overview

STS is the assessment of pairs of sentences according to their degree of semantic similarity. The task involves producing real-valued similarity scores for sentence pairs. Performance is measured by the Pearson correlation of machine scores with human judgments. The ordinal scale in Table 1 guides human annotation, ranging from 0 for no meaning overlap to 5 for meaning equivalence. Intermediate values reflect interpretable levels of partial overlap in meaning. The annotation scale is designed to be accessible by reasonable human judges without any formal expertise in linguistics. Using reasonable human interpretations of natural language semantics was popularized by the related textual entailment task (Dagan et al., 2010). The resulting annotations reflect both pragmatic and world knowledge and are more interpretable and useful within downstream systems.

3 Evaluation Data

The Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) is the primary evaluation data source with the exception that one of the cross-lingual tracks explores data from the WMT 2014 quality estimation task (Bojar et al., 2014).[3]

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.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*
	Two boys on a couch are playing video games. Two boys are playing a video game.
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.
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.
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.
0	*The two sentences are completely dissimilar.*
	The black dog is running through the snow. A race car driver is driving his car through the mud.

Table 1: Similarity scores with explanations and English examples from Agirre et al. (2013).

Sentences pairs in SNLI derive from Flickr30k image captions (Young et al., 2014) and are labeled with the entailment relations: entailment, neutral, and contradiction. Drawing from SNLI allows STS models to be evaluated on the type of data used to assess textual entailment methods. However, since entailment strongly cues for semantic relatedness (Marelli et al., 2014), we construct our own sentence pairings to deter gold entailment labels from informing evaluation set STS scores.

Track 4b investigates the relationship between STS and MT quality estimation by providing STS labels for WMT quality estimation data. The data includes Spanish translations of English sentences from a variety of methods including RBMT, SMT, hybrid-MT and human translation. Translations are annotated with the time required for human correction by post-editing and Human-targeted Translation Error Rate (HTER) (Snover et al., 2006).[4] Participants are not allowed to use the gold quality estimation annotations to inform STS scores.

pilot track on cross-lingual Spanish-English STS. The English tracks attracted the most participation and have the largest use of the evaluation data in ongoing research.

[3] Previous years of the STS shared task include more data sources. This year the task draws from two data sources and includes a diverse set of languages and language-pairs.

[4] HTER is the minimal number of edits required for correction of a translation divided by its length after correction.

Track	Language(s)	Pairs	Source
1	Arabic (ar-ar)	250	SNLI
2	Arabic-English (ar-en)	250	SNLI
3	Spanish (es-es)	250	SNLI
4a	Spanish-English (es-en)	250	SNLI
4b	Spanish-English (es-en)	250	WMT QE
5	English (en-en)	250	SNLI
6	Turkish-English (tr-en)	250	SNLI
	Total	1750	

Table 2: STS 2017 evaluation data.

3.1 Tracks

Table 2 summarizes the evaluation data by track. The six tracks span four languages: Arabic, English, Spanish and Turkish. Track 4 has subtracks with 4a drawing from SNLI and 4b pulling from WMT's quality estimation task. Track 6 is a surprise language track with no annotated training data and the identity of the language pair first announced when the evaluation data was released.

3.2 Data Preparation

This section describes the preparation of the evaluation data. For SNLI data, this includes the selection of sentence pairs, annotation of pairs with STS labels and the translation of the original English sentences. WMT quality estimation data is directly annotated with STS labels.

3.3 Arabic, Spanish and Turkish Translation

Sentences from SNLI are human translated into Arabic, Spanish and Turkish. Sentences are translated independently from their pairs. Arabic translation is provided by CMU-Qatar by native Arabic speakers with strong English skills. Translators are given an English sentence and its Arabic machine translation[5] where they perform post-editing to correct errors. Spanish translation is completed by a University of Sheffield graduate student who is a native Spanish speaker and fluent in English. Turkish translations are obtained from SDL.[6]

3.4 Embedding Space Pair Selection

We construct our own pairings of the SNLI sentences to deter gold entailment labels being used to inform STS scores. The *word embedding similarity* selection heuristic from STS 2016 (Agirre et al., 2016) is used to find interesting pairs. Sentence embeddings are computed as the sum of individual word embeddings, $\mathbf{v}(s) = \sum_{w \in s} \mathbf{v}(w)$.[7] Sentences with likely meaning overlap are identified using cosine similarity, Eq. (1).

$$\text{sim}_{\text{v}}(s_1, s_2) = \frac{\mathbf{v}(s_1)\mathbf{v}(s_2)}{\|\mathbf{v}(s_1)\|_2 \|\mathbf{v}(s_2)\|_2} \quad (1)$$

4 Annotation

Annotation of pairs with STS labels is performed using Crowdsourcing, with the exception of Track 4b that uses a single expert annotator.

4.1 Crowdsourced Annotations

Crowdsourced annotation is performed on Amazon Mechanical Turk.[8] Annotators examine the STS pairings of English SNLI sentences. STS labels are then transferred to the translated pairs for cross-lingual and non-English tracks. The annotation instructions and template are identical to Agirre et al. (2016). Labels are collected in batches of 20 pairs with annotators paid $1 USD per batch. Five annotations are collected per pair. The MTurk *master*[9] qualification is required to perform the task. Gold scores average the five individual annotations.

4.2 Expert Annotation

English-Spanish WMT quality estimation pairs for Track 4b are annotated for STS by a University of Sheffield graduate student who is a native speaker of Spanish and fluent in English. This track differs significantly in label distribution and the complexity of the annotation task. Sentences in a pair are translations of each other and tend to be more semantically similar. Interpreting the potentially subtle meaning differences introduced by MT errors is challenging. To accurately assess STS performance on MT quality estimation data, no attempt is made to balance the data by similarity scores.

5 Training Data

The following summarizes the training data: Table 3 English; Table 4 Spanish;[10] Table 5 Spanish-English; Table 6 Arabic; and Table 7 Arabic-English. Arabic-English parallel data is supplied by translating English training data, Table 8.

[5] Produced by the Google Translate API.

[6] http://www.sdl.com/languagecloud/managed-translation/

[7] We use 50-dimensional GloVe word embeddings (Pennington et al., 2014) trained on a combination of Gigaword 5 (Parker et al., 2011) and English Wikipedia available at http://nlp.stanford.edu/projects/glove/.

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

[9] A designation that statistically identifies workers who perform high quality work across a diverse set of tasks.

[10] Spanish data from 2015 and 2014 uses a 5 point scale that collapses STS labels 4 and 3, removing the distinction between unimportant and important details.

Year	Data set	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	post-editing	244	MT postedits
2016	Ans.-Ans.	254	Q&A forum answers
2016	Quest.-Quest.	209	Q&A forum questions
2017	Trial	23	Mixed STS 2016

Table 3: English training data.

Year	Data set	Pairs	Source
2014	Trial	56	
2014	Wiki	324	Spanish Wikipedia
2014	News	480	Newswire
2015	Wiki	251	Spanish Wikipedia
2015	News	500	Sewswire
2017	Trial	23	Mixed STS 2016

Table 4: Spanish training data.

English, Spanish and English-Spanish training data pulls from prior STS evaluations. Arabic and Arabic-English training data is produced by translating a subset of the English training data and transferring the similarity scores. For the MT quality estimation data in track 4b, Spanish sentences are translations of their English counterparts, differing substantially from existing Spanish-English STS data. We release one thousand new Spanish-English STS pairs sourced from the 2013 WMT translation task and produced by a phrase-based Moses SMT system (Bojar et al., 2013). The data is expert annotated and has a similar label distribution to the track 4b test data with 17% of the pairs scoring an STS score of less than 3, 23% scoring 3, 7% achieving a score of 4 and 53% scoring 5.

5.1 Training vs. Evaluation Data Analysis

Evaluation data from SNLI tend to have sentences that are slightly shorter than those from prior years of the STS shared task, while the track 4b MT quality estimation data has sentences that are much longer. The track 5 English data has an average sentence length of 8.7 words, while the English sentences from track 4b have an average length of 19.4. The English training data has the following average lengths: 2012 10.8 words; 2013 8.8 words (excludes restricted SMT data); 2014 9.1 words; 2015 11.5 words; 2016 13.8 words.

Year	Data set	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
2017	Trial	23	Mixed STS 2016
2017	MT	1000	WMT13 Translation Task

Table 5: Spanish-English training data.

Year	Data set	Pairs	Source
2017	Trial	23	Mixed STS 2016
2017	MSRpar	510	newswire
2017	MSRvid	368	videos
2017	SMTeuroparl	203	WMT eval.

Table 6: Arabic training data.

Similarity scores for our pairings of the SNLI sentences are slightly lower than recent shared task years and much lower than early years. The change is attributed to differences in data selection and filtering. The average 2017 similarity score is 2.2 overall and 2.3 on the track 7 English data. Prior English data has the following average similarity scores: 2016 2.4; 2015 2.4; 2014 2.8; 2013 3.0; 2012 3.5. Translation quality estimation data from track 4b has an average similarity score of 4.0.

6 System Evaluation

This section reports participant evaluation results for the SemEval-2017 STS shared task.

6.1 Participation

The task saw strong participation with 31 teams producing 84 submissions. 17 teams provided 44 systems that participated in all tracks. Table 9 summarizes participation by track. Traces of the focus on English are seen in 12 teams participating just in track 5, English. Two teams participated exclusively in tracks 4a and 4b, English-Spanish. One team took part solely in track 1, Arabic.

Year	Data set	Pairs	Source
2017	Trial	23	Mixed STS 2016
2017	MSRpar	1020	newswire
2017	MSRvid	736	videos
2017	SMTeuroparl	406	WMT eval.

Table 7: Arabic-English training data.

Year	Data set	Pairs	Source
2017	MSRpar	1039	newswire
2017	MSRvid	749	videos
2017	SMTeuroparl	422	WMT eval.

Table 8: Arabic-English parallel data.

6.2 Evaluation Metric

Systems are evaluated on each track by their Pearson correlation with gold labels. The overall ranking averages the correlations across tracks 1-5 with tracks 4a and 4b individually contributing.

Track	Language(s)	Participants
1	Arabic	49
2	Arabic-English	45
3	Spanish	48
4a	Spanish-English	53
4b	Spanish-English MT	53
5	English	77
6	Turkish-English	48
Primary	All except Turkish	44

Table 9: Participation by shared task track.

6.3 CodaLab

As directed by the SemEval workshop organizers, the CodaLab research platform hosts the task.[11]

6.4 Baseline

The baseline is the cosine of binary sentence vectors with each dimension representing whether an individual word appears in a sentence.[12] For cross-lingual pairs, non-English sentences are translated into English using state-of-the-art machine translation.[13] The baseline achieves an average correlation of 53.7 with human judgment on tracks 1-5 and would rank 23rd overall out the 44 system submissions that participated in all tracks.

6.5 Rankings

Participant performance is provided in Table 10. ECNU is best overall (avg r: 0.7316) and achieves the highest participant evaluation score on: track 2, Arabic-English (r: 0.7493); track 3, Spanish (r: 0.8559); and track 6, Turkish-English (r: 0.7706). BIT attains the best performance on track 1, Arabic (r: 0.7543). CompiLIG places first on track 4a, SNLI Spanish-English (r: 0.8302). SEF@UHH exhibits the best correlation on the difficult track 4b WMT quality estimation pairs (r: 0.3407). RTV has the best system for the track 5 English data (r: 0.8547), followed closely by DT_Team (r: 0.8536).

Especially challenging tracks with SNLI data are: track 1, Arabic; track 2, Arabic-English; and track 6, English-Turkish. Spanish-English performance is much higher on track 4a's SNLI data than track 4b's MT quality estimation data. This highlights the difficulty and importance of making fine grained distinctions for certain downstream applications. Assessing STS methods for quality estimation may benefit from using alternatives to Pearson correlation for evaluation.[14]

Results tend to decrease on cross-lingual tracks. The baseline drops $> 10\%$ relative on Arabic-English and Spanish-English (SNLI) vs. monolingual Arabic and Spanish. Many participant systems show smaller decreases. ECNU's top ranking entry performs slightly better on Arabic-English than Arabic, with a slight drop from Spanish to Spanish-English (SNLI).

6.6 Methods

Participating teams explore techniques ranging from state-of-the-art deep learning models to elaborate feature engineered systems. Prediction signals include surface similarity scores such as edit distance and matching n-grams, scores derived from word alignments across pairs, assessment by MT evaluation metrics, estimates of conceptual similarity as well as the similarity between word and sentence level embeddings. For cross-lingual and non-English tracks, MT was widely used to convert the two sentences being compared into the same language.[15] Select methods are highlighted below.

[11] https://competitions.codalab.org/competitions/16051

[12] Words obtained using Arabic (ar), Spanish (es) and English (en) Treebank tokenizers.

[13] http://translate.google.com

[14] e.g., Reimers et al. (2016) report success using STS labels with alternative metrics such as normalized Cumulative Gain (nCG), normalized Discounted Cumulative Gain (nDCG) and F1 to more accurately predict performance on the downstream tasks: text reuse detection, binary classification of document relatedness and document relatedness within a corpus.

[15] Within the highlighted submissions, the following use a monolingual English system fed by MT: ECNU, BIT, HCTI

Team	Primary	Track 1 AR-AR	Track 2 AR-EN	Track 3 SP-SP	Track 4a SP-EN	Track 4b SP-EN-WMT	Track 5 EN-EN	Track 6 EN-TR
ECNU (Tian et al., 2017)	73.16	**74.40**	**74.93•**	**85.59•**	**81.31**	**33.63**	**85.18**	**77.06•**
ECNU (Tian et al., 2017)	70.44	**73.80**	71.26	**84.56**	74.95	**33.11**	**81.81**	73.62
ECNU (Tian et al., 2017)	69.40	**72.71**	69.75	82.47	76.49	**26.33**	**83.87**	74.20
BIT (Wu et al., 2017)*	67.89	**74.17**	69.65	**84.99**	78.28	11.07	**84.00**	**73.05**
BIT (Wu et al., 2017)*	67.03	75.35	70.07	83.23	78.13	7.58	81.61	73.27
BIT (Wu et al., 2017)	66.62	**75.43•**	69.53	**82.89**	77.61	5.84	**82.22**	**72.80**
HCTI (Shao, 2017)	65.98	**71.30**	68.36	**82.63**	76.21	14.83	81.13	67.41
MITRE (Henderson et al., 2017)	65.90	**72.94**	67.53	82.02	78.02	15.98	80.53	64.30
MITRE (Henderson et al., 2017)	65.87	**73.04**	67.40	82.01	77.99	15.74	80.48	64.41
FCICU (Hassan et al., 2017)	61.90	**71.58**	67.82	**84.84**	69.26	2.54	**82.72**	54.52
neobility (Zhuang and Chang, 2017)	61.71	68.21	64.59	79.28	71.69	2.00	79.27	66.96
FCICU (Hassan et al., 2017)	61.66	**71.58**	67.81	**84.89**	68.54	2.14	**82.80**	53.90
STS-UHH (Kohail et al., 2017)	60.58	67.81	63.07	77.13	72.01	4.81	79.89	59.37
RTV	60.50	67.13	55.95	74.85	70.50	7.61	**85.41**	62.04
HCTI (Shao, 2017)	59.88	43.73	68.36	67.09	76.21	14.83	**81.56**	67.41
RTV	59.80	66.89	54.82	74.24	69.99	7.34	**85.41**	59.89
MatrusriIndia	59.60	68.60	54.64	76.14	71.18	5.72	77.44	63.49
STS-UHH (Kohail et al., 2017)	57.25	61.04	59.10	72.04	63.38	12.05	73.39	59.72
SEF@UHH (Duma and Menzel, 2017)	56.76	57.90	53.84	74.23	58.66	18.02	72.56	62.11
SEF@UHH (Duma and Menzel, 2017)	56.44	55.88	47.89	74.56	57.39	**30.69**	78.80	49.90
RTV	56.33	61.43	48.32	68.63	61.40	8.29	**85.47•**	60.79
SEF@UHH (Duma and Menzel, 2017)	55.28	57.74	48.13	69.79	56.60	**34.07•**	71.86	48.78
neobility (Zhuang and Chang, 2017)	51.95	13.69	62.59	77.92	69.30	0.44	75.56	64.18
neobility (Zhuang and Chang, 2017)	50.25	3.69	62.07	76.90	69.47	1.47	75.35	62.79
MatrusriIndia	49.75	57.03	43.40	67.86	55.63	8.57	65.79	49.94
NLPProxem	49.02	51.93	53.13	66.42	51.44	9.96	62.56	47.67
UMDeep (Barrow and Peskov, 2017)	47.92	47.53	49.39	51.65	56.15	16.09	61.74	52.93
NLPProxem	47.90	55.06	43.69	63.81	50.79	14.14	64.63	43.20
UMDeep (Barrow and Peskov, 2017)	47.73	45.87	51.99	51.48	52.32	13.00	62.22	57.25
Lump (España Bonet and Barrón-Cedeño, 2017)*	47.25	60.52	18.29	75.74	43.27	1.16	73.76	58.00
Lump (España Bonet and Barrón-Cedeño, 2017)*	47.04	55.08	13.57	76.76	48.25	11.12	72.69	51.79
Lump (España Bonet and Barrón-Cedeño, 2017)*	44.38	62.87	18.05	73.80	44.47	1.51	73.47	36.52
NLPProxem	40.70	53.27	47.73	0.16	55.06	14.40	66.81	47.46
RTM (Biçici, 2017)*	36.69	33.65	17.11	69.90	60.04	14.55	54.68	6.87
UMDeep (Barrow and Peskov, 2017)	35.21	39.05	37.13	45.88	34.82	5.86	47.27	36.44
RTM (Biçici, 2017)*	32.91	33.65	0.25	56.82	50.54	13.68	64.05	11.36
RTM (Biçici, 2017)*	32.78	41.56	13.32	48.41	45.83	23.47	56.32	0.55
ResSim (Bjerva and Östling, 2017)	31.48	28.92	10.45	66.13	23.89	3.05	69.06	18.84
ResSim (Bjerva and Östling, 2017)	29.38	31.20	12.88	69.20	10.02	1.62	68.77	11.95
ResSim (Bjerva and Östling, 2017)	21.45	0.33	10.98	54.65	22.62	1.99	50.57	9.02
LIPN-IIMAS (Arroyo-Fernández and Meza Ruiz, 2017)	10.67	4.71	7.69	15.27	17.19	14.46	7.38	8.00
LIPN-IIMAS (Arroyo-Fernández and Meza Ruiz, 2017)	9.26	2.14	12.92	4.58	1.20	1.91	20.38	21.68
hjpwhu	4.80	4.12	6.39	6.17	2.04	6.24	1.14	7.53
hjpwhu	2.94	4.77	2.04	7.63	0.46	2.57	0.69	2.46
compiLIG (Ferrero et al., 2017)					**83.02•**	15.50		
compiLIG (Ferrero et al., 2017)					76.84	14.64		
compiLIG (Ferrero et al., 2017)					79.10	14.94		
DT_TEAM (Maharjan et al., 2017)							**85.36**	
DT_TEAM (Maharjan et al., 2017)							**83.60**	
DT_TEAM (Maharjan et al., 2017)							**83.29**	
FCICU (Hassan et al., 2017)							**82.17**	
ITNLPAiKF (Liu et al., 2017)							**82.31**	
ITNLPAiKF (Liu et al., 2017)							**82.31**	
ITNLPAiKF (Liu et al., 2017)							**81.59**	
L2F/INESC-ID (Fialho et al., 2017)*				76.16	1.91	5.44	78.11	2.93
L2F/INESC-ID (Fialho et al., 2017)							69.52	
L2F/INESC-ID (Fialho et al., 2017)*				63.85	15.61	5.24	66.61	3.56
LIM-LIG (Nagoudi et al., 2017)		**74.63**						
LIM-LIG (Nagoudi et al., 2017)		**73.09**						
LIM-LIG (Nagoudi et al., 2017)		59.57						
MatrusriIndia		68.60		76.14	71.18	5.72	77.44	63.49
NRC*					42.25	0.23		
NRC					28.08	11.33		
OkadaNaoya							77.04	
OPI-JSA (Śpiewak et al., 2017)							78.50	
OPI-JSA (Śpiewak et al., 2017)							73.42	
OPI-JSA (Śpiewak et al., 2017)							67.96	
PurdueNLP (Lee et al., 2017)							79.28	
PurdueNLP (Lee et al., 2017)							55.35	
PurdueNLP (Lee et al., 2017)							53.11	
QLUT (Meng et al., 2017)*							64.33	
QLUT (Meng et al., 2017)							61.55	
QLUT (Meng et al., 2017)*							49.24	
SIGMA							80.47	
SIGMA							80.08	
SIGMA							79.12	
SIGMA_PKU_2							81.34	
SIGMA_PKU_2							81.27	
SIGMA_PKU_2							80.61	
STS-UHH (Kohail et al., 2017)							80.93	
UCSC-NLP							77.29	
UdL (Al-Natsheh et al., 2017)							80.04	
UdL (Al-Natsheh et al., 2017)*							79.01	
UdL (Al-Natsheh et al., 2017)							78.05	
cosine baseline	53.70	60.45	51.55	71.17	62.20	3.20	72.78	54.56

* Corrected or late submission

Table 10: STS 2017 rankings ordered by average correlation across tracks 1-5. Performance is reported by convention as Pearson's $r \times 100$. For tracks 1-6, the top ranking result is marked with a • symbol and results in bold have no statistically significant difference with the best result on a track, $p > 0.05$ Williams' t-test (Diedenhofen and Musch, 2015).

ECNU (Tian et al., 2017) The best overall system is from ENCU and ensembles well performing a feature engineered models with deep learning methods. Three feature engineered models use Random Forest (RF), Gradient Boosting (GB) and XGBoost (XGB) regression methods with features based on: n-gram overlap; edit distance; longest common prefix/suffix/substring; tree kernels (Moschitti, 2006); word alignments (Sultan et al., 2015); summarization and MT evaluation metrics (BLEU, GTM-3, NIST, WER, METEOR, ROUGE); and kernel similarity of bags-of-words, bags-of-dependencies and pooled word-embeddings. ECNU's deep learning models are differentiated by their approach to sentence embeddings using either: averaged word embeddings, projected word embeddings, a deep averaging network (DAN) (Iyyer et al., 2015) or LSTM (Hochreiter and Schmidhuber, 1997). Each network feeds the element-wise multiplication, subtraction and concatenation of paired sentence embeddings to additional layers to predict similarity scores. The ensemble averages scores from the four deep learning and three feature engineered models.[16]

BIT (Wu et al., 2017) Second place overall is achieved by BIT primarily using sentence information content (IC) informed by WordNet and BNC word frequencies. One submission uses sentence IC exclusively. Another ensembles IC with Sultan et al. (2015)'s alignment method, while a third ensembles IC with cosine similarity of summed word embeddings with an IDF weighting scheme. Sentence IC in isolation outperforms all systems except those from ECNU. Combining sentence IC with word embedding similarity performs best.

HCTI (Shao, 2017) Third place overall is obtained by HCTI with a model similar to a convolutional Deep Structured Semantic Model (CDSSM) (Chen et al., 2015; Huang et al., 2013). Sentence embeddings are generated with twin convolutional neural networks (CNNs). The embeddings are then compared using cosine similarity and element wise difference with the resulting values fed to additional layers to predict similarity labels. The architecture is abstractly similar to ECNU's deep learning models. UMDeep (Barrow and Peskov, 2017) took a similar approach using LSTMs rather than CNNs for the sentence embeddings.

MITRE (Henderson et al., 2017) Fourth place overall is MITRE that, like ECNU, takes an ambitious feature engineering approach complemented by deep learning. Ensembled components include: alignment similarity; TakeLab STS (Šarić et al., 2012b); string similarity measures such as matching n-grams, summarization and MT metrics (BLEU, WER, PER, ROUGE); a RNN and recurrent convolutional neural networks (RCNN) over word alignments; and a BiLSTM that is state-of-the-art for textual entailment (Chen et al., 2016).

FCICU (Hassan et al., 2017) Fifth place overall is FCICU that computes a sense-base alignment using BabelNet (Navigli and Ponzetto, 2010). BabelNet synsets are multilingual allowing non-English and cross-lingual pairs to be processed similarly to English pairs. Alignment similarity scores are used with two runs: one that combines the scores within a string kernel and another that uses them with a weighted variant of Sultan et al. (2015)'s method. Both runs average the Babelnet based scores with soft-cardinality (Jimenez et al., 2012b).

CompiLIG (Ferrero et al., 2017) The best Spanish-English performance on SNLI sentences was achieved by CompiLIG using features including: cross-lingual conceptual similarity using DBNary (Serasset, 2015), cross-language MultiVec word embeddings (Berard et al., 2016), and Brychcin and Svoboda (2016)'s improvements to Sultan et al. (2015)'s method.

LIM-LIG (Nagoudi et al., 2017) Using only weighted word embeddings, LIM-LIG took second place on Arabic.[17] Arabic word embeddings are summed into sentence embeddings using uniform, POS and IDF weighting schemes. Sentence similarity is computed by cosine similarity. POS and IDF outperform uniform weighting. Combining the IDF and POS weights by multiplication is reported by LIM-LIG to achieve r 0.7667, higher than all submitted Arabic (track 1) systems.

DT_Team (Maharjan et al., 2017) Second place on English (track 5)[18] is DT_Team using feature en-

and MITRE. HCTI submitted a separate run using ar, es and en trained models that underperformed using their en model with MT for ar and es. CompiLIG's model is cross-lingual but includes a word alignment feature that depends on MT. SEF@UHH built ar, es, and en models and use bi-directional MT for cross-lingual pairs. LIM-LIG and DT_Team only participate in monolingual tracks.

[16]The two remaining ECNU runs only use either RF or GB and exclude the deep learning models.

[17]The approach is similar to SIF (Arora et al., 2017) but without removal of the common principle component

[18]RTV took first place on track 5, English, but submitted no system description paper.

Genre	Train	Dev	Test	Total
news	3299	500	500	4299
caption	2000	625	525	3250
forum	450	375	254	1079
total	5749	1500	1379	8628

Table 11: STS Benchmark annotated examples by genres (rows) and by train, dev. test splits (columns).

gineering combined with the following deep learning models: DSSM (Huang et al., 2013), CDSSM (Shen et al., 2014) and skip-thoughts (Kiros et al., 2015). Engineered features include: unigram overlap, summed word alignments scores, fraction of unaligned words, difference in word counts by type (all, adj, adverbs, nouns, verbs), and min to max ratios of words by type. Select features have a multiplicative penalty for unaligned words.

SEF@UHH (Duma and Menzel, 2017) First place on the challenging Spanish-English MT pairs (Track 4b) is SEF@UHH. Unsupervised similarity scores are computed from paragraph vectors (Le and Mikolov, 2014) using cosine, negation of Bray-Curtis dissimilarity and vector correlation. MT converts cross-lingual pairs, L_1-L_2, into two monolingual pairs, L_1-L_1 and L_2-L_2, with averaging used to combine the monolingual similarity scores. Bray-Curtis performs well overall, while cosine does best on the Spanish-English MT pairs.

7 Analysis

Figure 1 plots model similarity scores against human STS labels for the top 5 systems from tracks 5 (English), 1 (Arabic) and 4b (English-Spanish MT). While many systems return scores on the same scale as the gold labels, 0-5, others return scores from approximately 0 and 1. Lines on the graphs illustrate perfect performance for both a 0-5 and a 0-1 scale. Mapping the 0 to 1 scores to range from 0-5,[20] approximately 80% of the scores from top performing English systems are within 1.0 pt of the gold label. Errors for Arabic are more broadly distributed, particularly for model scores between 1 and 4. The English-Spanish MT plots the weak relationship between the predicted and gold scores.

Table 12 provides examples of difficult sentence pairs for participant systems and illustrates common sources of error for even well-ranking systems including: (i) *word sense disambiguation* "making" and "preparing" are very similar in the context of "food", while "picture" and "movie" are not similar when picture is followed by "day"; (ii) *attribute importance* "outside" vs. "deserted" are smaller details when contrasting "The man is in a deserted field" with "The man is outside in the field"; (iii) *compositional meaning* "A man is carrying a canoe with a dog" has the same content words as "A dog is carrying a man in a canoe" but carries a different meaning; (iv) *negation* systems score "...with goggles and a swimming cap" as nearly equivalent to "...without goggles or a swimming cap". Inflated similarity scores for examples like "There is a young girl" vs. "There is a young boy with the woman" demonstrate (v) *semantic blending*, whereby appending "with a woman" to "boy" brings its representation closer to that of "girl".

For multilingual and cross-lingual pairs, these issues are magnified by translation errors for systems that use MT followed by the application of a monolingual similarity model. For track 4b Spanish-English MT pairs, some of the poor performance can in part be attributed to many systems using MT to re-translate the output of another MT system, obscuring errors in the original translation.

7.1 Contrasting Cross-lingual STS with MT Quality Estimation

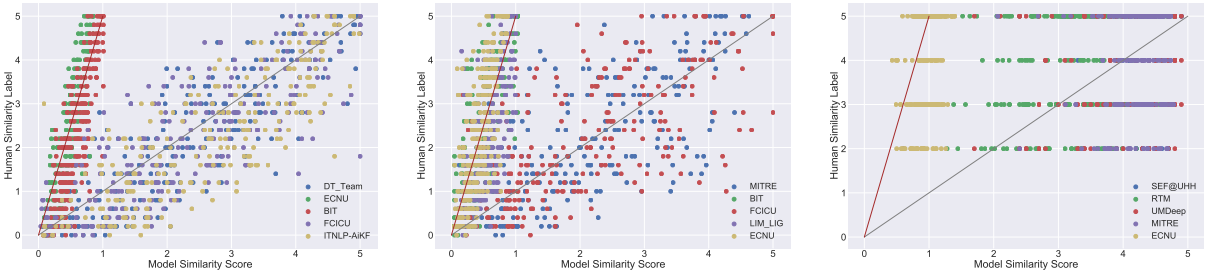
Since MT quality estimation pairs are translations of the same sentence, they are expected to be minimally on the same topic and have an STS score ≥ 1.[21] The actual distribution of STS scores is such that only 13% of the test instances score below 3, 22% of the instances score 3, 12% score 4 and 53% score 5. The high STS scores indicate that MT systems are surprisingly good at preserving meaning. However, even for a human, interpreting changes caused by translations errors can be difficult due both to disfluencies and subtle errors with important changes in meaning.

The Pearson correlation between the gold MT quality scores and the gold STS scores is 0.41, which shows that translation quality measures and STS are only moderately correlated. Differences are in part explained by translation quality scores penalizing all mismatches between the source segment and its translation, whereas STS focuses on differences in meaning. However, the difficult interpretation work required for STS annotation may

[19] ECNU, BIT and LIM-LIG are scaled to the range 0-5.

[20] $s_{new} = 5 \times \frac{s - min(s)}{max(s) - min(s)}$ is used to rescale scores.

[21] The evaluation data for track 4b does in fact have STS scores that are ≥ 1 for all pairs. In the 1,000 sentence training set for this track, one sentence that received a score of zero.



(a) Track 5: English (b) Track 1: Arabic (c) Track 4b: English-Spanish MT

Figure 1: Model vs. human similarity scores for top systems.

Pairs	Human	DT_Team	ECNU	BIT	FCICU	ITNLP-AiKF
There is a cook preparing food. A cook is making food.	5.0	4.1	4.1	3.7	3.9	4.5
The man is in a deserted field. The man is outside in the field.	4.0	3.0	3.1	3.6	3.1	2.8
A girl in water without goggles or a swimming cap. A girl in water, with goggles and swimming cap.	3.0	4.8	4.6	4.0	4.7	0.1
A man is carrying a canoe with a dog. A dog is carrying a man in a canoe.	1.8	3.2	4.7	4.9	5.0	4.6
There is a young girl. There is a young boy with the woman.	1.0	2.6	3.3	3.9	1.9	3.1
The kids are at the theater watching a movie. it is picture day for the boys	0.2	1.0	2.3	2.0	0.8	1.7

Table 12: Difficult English sentence pairs (Track 5) and scores assigned by top performing systems.[19]

Genre	File	Yr.	Train	Dev	Test
news	MSRpar	12	1000	250	250
news	headlines	13/6	1999	250	250
news	deft-news	14	300	0	0
captions	MSRvid	12	1000	250	250
captions	images	14/5	1000	250	250
captions	track5.en-en	17	0	125	125
forum	deft-forum	14	450	0	0
forum	ans-forums	15	0	375	0
forum	ans-ans	16	0	0	254

Table 13: STS Benchmark detailed break-down by files and years.

increase the risk of inconsistent and subjective labels. The annotations for MT quality estimation are produced as by-product of post-editing. Humans fix MT output and the edit distance between the output and its post-edited correction provides the quality score. This post-editing based procedure is known to produce relatively consistent estimates across annotators.

8 STS Benchmark

The STS Benchmark is a careful selection of the English data sets used in SemEval and *SEM STS shared tasks between 2012 and 2017. Tables 11 and 13 provide details on the composition of the benchmark. The data is partitioned into training, development and test sets.[22] The development set can be used to design new models and tune hyperparameters. The test set should be used sparingly and only after a model design and hyperparameters have been locked against further changes. Using the STS Benchmark enables comparable assessments across different research efforts and improved tracking of the state-of-the-art.

Table 14 shows the STS Benchmark results for some of the best systems from Track 5 (EN-EN)[23] and compares their performance to competitive baselines from the literature. All baselines were run by the organizers using canonical pre-trained models made available by the originator of each method,[24] with the exception of PV-DBOW that

[22] Similar to the STS shared task, while the training set is provided as a convenience, researchers are encourage to incorporate other supervised and unsupervised data as long as no supervised annotations of the test partitions are used.

[23] Each participant submitted the run which did best in the development set of the STS Benchmark, which happened to be the same as their best run in Track 5 in all cases.

[24] **sent2vec**: https://github.com/epfml/sent2vec, trained model sent2vec_twitter_unigrams; **SIF**: https://github.com/epfml/sent2vec Wikipedia trained word frequencies enwiki_vocab_min200.txt, https://github.com/alexandres/lexvec embeddings from lexvec.commoncrawl.300d.W+C.pos.vectors, first 15 principle components removed, $\alpha = 0.001$, dev

STS 2017 Participants on STS Benchmark			
Name	Description	Dev	Test
ECNU	Ensemble (Tian et al., 2017)	84.7	81.0
BIT	WordNet+Embeddings (Wu et al., 2017)	82.9	80.9
DT_TEAM	Ensemble (Maharjan et al., 2017)	83.0	79.2
HCTI	CNN (Shao, 2017)	83.4	78.4
SEF@UHH	Doc2Vec (Duma and Menzel, 2017)	61.6	59.2
Sentence Level Baselines			
sent2vec	Sentence spanning CBOW with words & bigrams (Pagliardini et al., 2017)	78.7	75.5
SIF	Word embedding weighting & principle component removal (Arora et al., 2017)	80.1	72.0
InferSent	Sentence embedding from bi-directional LSTM trained on SNLI (Conneau et al., 2017)	80.1	75.8
C-PHRASE	Prediction of syntactic constituent context words (Pham et al., 2015)	74.3	63.9
PV-DBOW	Paragraph vectors, Doc2Vec DBOW (Le and Mikolov, 2014; Lau and Baldwin, 2016)	72.2	64.9
Averaged Word Embedding Baselines			
LexVec	Weighted matrix factorization of PPMI (Salle et al., 2016a,b)	68.9	55.8
FastText	Skip-gram with sub-word character n-grams (Joulin et al., 2016)	65.3	53.6
Paragram	Paraphrase Database (PPDB) fit word embeddings (Wieting et al., 2015)	63.0	50.1
GloVe	Word co-occurrence count fit embeddings (Pennington et al., 2014)	52.4	40.6
Word2vec	Skip-gram prediction of words in a context window (Mikolov et al., 2013a,b)	70.0	56.5

Table 14: STS Benchmark. Pearson's $r \times 100$ results for select participants and baseline models.

uses the model from Lau and Baldwin (2016) and InferSent which was reported independently. When multiple pre-trained models are available for a method, we report results for the one with the best dev set performance. For each method, input sentences are preprocessed to closely match the tokenization of the pre-trained models.[25] Default inference hyperparameters are used unless noted otherwise. The *averaged word embedding baselines* compute a sentence embedding by averaging word embeddings and then using cosine to compute pairwise sentence similarity scores.

While state-of-the-art baselines for obtaining sentence embeddings perform reasonably well on the benchmark data, improved performance is obtained by top 2017 STS shared task systems. There is still substantial room for further improvement. To follow the current state-of-the-art, visit the leaderboard on the STS wiki.[26]

9 Conclusion

We have presented the results of the 2017 STS shared task. This year's shared task differed substantially from previous iterations of STS in that the primary emphasis of the task shifted from English to multilingual and cross-lingual STS involving four different languages: Arabic, Spanish, English and Turkish. Even with this substantial change relative to prior evaluations, the shared task obtained strong participation. 31 teams produced 84 system submissions with 17 teams producing a total of 44 system submissions that processed pairs in all of the STS 2017 languages. For languages that were part of prior STS evaluations

experiments varied α, principle components removed and whether GloVe, LexVec, or Word2Vec word embeddings were used; **C-PHRASE**: http://clic.cimec.unitn.it/composes/cphrase-vectors.html; **PV-DBOW**: https://github.com/jhlau/doc2vec, AP-NEWS trained apnews_dbow.tgz; **LexVec**: https://github.com/alexandres/lexvec, embedddings lexvec.commoncrawl.300d.W.pos.vectors.gz; **FastText**: https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md, Wikipedia trained embeddings from wiki.en.vec; **Paragram**: http://ttic.uchicago.edu/~wieting/, embeddings trained on PPDB and tuned to WS353 from Paragram-WS353; **GloVe**: https://nlp.stanford.edu/projects/glove/, Wikipedia and Gigaword trained 300 dim. embeddings from glove.6B.zip; **Word2vec**: https://code.google.com/archive/p/word2vec/, Google News trained embeddings from GoogleNews-vectors-negative300.bin.gz.

[25] **sent2vec**: results shown here tokenized by tweetTokenize.py constrasting dev experiments used wikiTokenize.py, both distributed with sent2vec. **LexVec**: numbers were converted into words, all punctuation was removed, and text is lowercased; **FastText**: Since, to our knowledge, the tokenizer and preprocessing used for the pre-trained FastText embeddings is not publicly described. We use the following heuristics to preprocess and tokenize sentences for FastText: numbers are converted into words, text is lowercased, and finally prefixed, suffixed and infixed punctuation is recursively removed from each token that does not match an entry in the model's lexicon; **Paragram**: Joshua (Matt Post, 2015) pipeline to pre-process and tokenized English text; **C-PHRASE**, **GloVe**, **PV-DBOW** & **SIF**: PTB tokenization provided by Stanford CoreNLP (Manning et al., 2014) with post-processing based on dev OOVs; **Word2vec**: Similar to FastText, to our knownledge, the preprocessing for the pre-trained Word2vec embeddings is not publicly described. We use the following heuristics for the Word2vec experiment: All numbers longer than a single digit are converted into a '#' (e.g., 24 → ##) then prefixed, suffixed and infixed punctuation is recursively removed from each token that does not match an entry in the model's lexicon.

[26] http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

(e.g., English and Spanish), state-of-the-art systems are able to achieve strong correlations with human judgment. However, we obtain weaker correlations from participating systems for Arabic, Arabic-English and Turkish-English. This suggests further research is necessary in order to develop robust models that can both be readily applied to new languages and perform well even when less supervised training data is available. To provide a standard benchmark for English STS, we present the STS Benchmark, a careful selection of the English data sets from previous STS tasks (2012-2017). To assist in interpreting the results from new models, a number of competitive baselines and select participant systems are evaluated on the benchmark data. Ongoing improvements to the current state-of-the-art is available from an online leaderboard.

Acknowledgments

We thank Alexis Conneau for the evaluation of InferSent on the STS Benchmark. This material is based in part upon work supported by QNRF-NPRP 6 - 1020-1-199 OPTDIAC that funded Arabic translation, and by a grant from the Spanish MINECO (projects TUNER TIN2015-65308-C5-1-R and MUSTER PCIN-2015-226 cofunded by EU FEDER) that funded STS label annotation and by the QT21 EU project (H2020 No. 645452) that funded STS labels and data preparation for machine translation pairs. Iñigo Lopez-Gazpio is supported by the Spanish MECD. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of QNRF-NPRP, Spanish MINECO, QT21 EU, or the Spanish MECD.

References

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 SemEval 2015*. http://www.aclweb.org/anthology/S15-2045.

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 SemEval 2014*. http://www.aclweb.org/anthology/S14-2010.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the SemEval-2016*. http://www.aclweb.org/anthology/S16-1081.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *Proceedings of *SEM 2012/SemEval 2012*. http://www.aclweb.org/anthology/S12-1051.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Proceedings of *SEM 2013*. http://www.aclweb.org/anthology/S13-1004.

Hussein T. Al-Natsheh, Lucie Martinet, Fabrice Muhlenbach, and Djamel Abdelkader ZIGHED. 2017. UdL at SemEval-2017 Task 1: Semantic textual similarity estimation of english sentence pairs using regression model over pairwise features. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2013.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR 2017*. https://openreview.net/pdf?id=SyK00v5xx.

Ignacio Arroyo-Fernández and Ivan Vladimir Meza Ruiz. 2017. LIPN-IIMAS at SemEval-2017 Task 1: Subword embeddings, attention recurrent neural networks and cross word alignment for semantic textual similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2031.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING '98*. http://aclweb.org/anthology/P/P98/P98-1013.pdf.

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of *SEM 2012/SemEval 2012*. http://www.aclweb.org/anthology/S12-1059.

Joe Barrow and Denis Peskov. 2017. UMDeep at SemEval-2017 Task 1: End-to-end shared weight LSTM model for semantic textual similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2026.

Luisa Bentivogli, Raffaella Bernardi, Marco Marelli, Stefano Menini, Marco Baroni, and Roberto Zamparelli. 2016. SICK through the SemEval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Lang Resour Eval* 50(1):95–124. https://doi.org/10.1007/s10579-015-9332-5.

Alexandre Berard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. MultiVec: a multilingual and multilevel representation learning toolkit for NLP. In *Proceedings of LREC 2016*. http://www.lrec-conf.org/proceedings/lrec2016/pdf/666_Paper.pdf.

Ergun Biçici. 2017. RTM at SemEval-2017 Task 1: Referential translation machines for predicting semantic similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2030.

Johannes Bjerva and Robert Östling. 2017. ResSim at SemEval-2017 Task 1: Multilingual word representations for semantic textual similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2021.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of WMT 2014*. http://www.aclweb.org/anthology/W/W14/W14-3302.pdf.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of WMT 2013*. http://www.aclweb.org/anthology/W13-2201.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP 2015*. http://aclweb.org/anthology/D/D15/D15-1075.pdf.

Tomas Brychcin and Lukas Svoboda. 2016. UWB at SemEval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic information. In *Proceedings of SemEval 2016*. https://www.aclweb.org/anthology/S/S16/S16-1089.pdf.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. Enhancing and combining sequential and tree LSTM for natural language inference. *CoRR* abs/1609.06038. http://arxiv.org/abs/1609.06038.

Yun-Nung Chen, Dilek Hakkani-Tür, and Xiaodong He. 2015. Learning bidirectional intent embeddings by convolutional deep structured semantic models for spoken language understanding. In *Proceedings of NIPS-SLU, 2015*. https://www.microsoft.com/en-us/research/publication/learning-bidirectional-intent-embeddings-by-convolutional-deep-structured-semantic-models-for-spoken-language-understanding/.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *CoRR* abs/1705.02364. http://arxiv.org/abs/1705.02364.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches. *J. Nat. Language Eng.* 16:105–105. https://doi.org/10.1017/S1351324909990234.

Birk Diedenhofen and Jochen Musch. 2015. cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE* 10(4). http://dx.doi.org/10.1371/journal.pone.0121945.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING 04*. http://aclweb.org/anthology/C/C04/C04-1051.pdf.

Mirela-Stefania Duma and Wolfgang Menzel. 2017. SEF@UHH at SemEval-2017 Task 1: Unsupervised knowledge-free semantic textual similarity via paragraph vector. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2024.

Cristina España Bonet and Alberto Barrón-Cedeño. 2017. Lump at SemEval-2017 Task 1: Towards an interlingua semantic similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2019.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press. https://books.google.com/books?id=Rehu8OOzMIMC.

Jérémy Ferrero, Laurent Besacier, Didier Schwab, and Frédéric Agnès. 2017. CompiLIG at SemEval-2017 Task 1: Cross-language plagiarism detection methods for semantic textual similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2012.

Pedro Fialho, Hugo Patinho Rodrigues, Luísa Coheur, and Paulo Quaresma. 2017. L2f/inesc-id at semeval-2017 tasks 1 and 2: Lexical and semantic features in word and textual similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2032.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL/HLT 2013*. http://cs.jhu.edu/ ccb/publications/ppdb.pdf.

Basma Hassan, Samir AbdelRahman, Reem Bahgat, and Ibrahim Farag. 2017. FCICU at SemEval-2017 Task 1: Sense-based language independent semantic textual similarity approach. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2015.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of EMNLP*. pages 1576–1586. http://aclweb.org/anthology/D15-1181.

Hua He and Jimmy Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of NAACL/HLT*. http://www.aclweb.org/anthology/N16-1108.

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 SemEval 2016*. http://www.anthology.aclweb.org/S/S16/S16-1170.pdf.

John Henderson, Elizabeth Merkhofer, Laura Strickhart, and Guido Zarrella. 2017. MITRE at SemEval-2017 Task 1: Simple semantic similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2027.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of NAACL/HLT*. http://www.aclweb.org/anthology/N16-1162.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of NAACL/HLT 2006*. http://aclweb.org/anthology/N/N06/N06-2015.pdf.

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. In *Proceedings of CIKM*. https://www.microsoft.com/en-us/research/publication/learning-deep-structured-semantic-models-for-web-search-using-clickthrough-data/.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of ACL/IJCNLP*. http://www.aclweb.org/anthology/P15-1162.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012a. Soft cardinality: A parameterized similarity function for text comparison. In *Proceedings of *SEM 2012/SemEval 2012*. http://www.aclweb.org/anthology/S12-1061.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012b. Soft Cardinality: A parameterized similarity function for text comparison. In *Proceedings of *SEM 2012/SemEval 2012*. http://aclweb.org/anthology/S/S12/S12-1061.pdf.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *CoRR* abs/1607.01759. http://arxiv.org/abs/1607.01759.

Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. In *Proceedings of ACL*. http://www.aclweb.org/anthology/P16-1089.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. *CoRR* abs/1506.06726. http://arxiv.org/abs/1506.06726.

Sarah Kohail, Amr Rekaby Salama, and Chris Biemann. 2017. STS-UHH at SemEval-2017 Task 1: Scoring semantic textual similarity using supervised and unsupervised ensemble. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2025.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of ACL Workshop on Representation Learning for NLP*. http://www.aclweb.org/anthology/W/W16/W16-1609.pdf.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *CoRR* abs/1405.4053. http://arxiv.org/abs/1405.4053.

I-Ta Lee, Mahak Goindani, Chang Li, Di Jin, Kristen Marie Johnson, Xiao Zhang, Maria Leonor Pacheco, and Dan Goldwasser. 2017. PurdueNLP at SemEval-2017 Task 1: Predicting semantic textual similarity with paraphrase and event embeddings. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2029.

Wenjie Liu, Chengjie Sun, Lei Lin, and Bingquan Liu. 2017. ITNLP-AiKF at SemEval-2017 Task 1: Rich features based svr for semantic textual similarity computing. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2022.

Nabin Maharjan, Rajendra Banjade, Dipesh Gautam, Lasang J. Tamang, and Vasile Rus. 2017. Dt_team at semeval-2017 task 1: Semantic similarity using alignments, sentence-level embeddings and gaussian mixture model output. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2014.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL 2014 Demonstrations*. http://www.aclweb.org/anthology/P/P14/P14-5010.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 14*. http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

Yuan Cao Gaurav Kumar Matt Post. 2015. Joshua 6: A phrase-based and hierarchical statistical machine translation. *The Prague Bulletin of Mathematical Linguistics* 104:516. https://ufal.mff.cuni.cz/pbml/104/art-post-cao-kumar.pdf.

Fanqing Meng, Wenpeng Lu, Yuteng Zhang, Jinyong Cheng, Yuehan Du, and Shuwang Han. 2017. QLUT at SemEval-2017 Task 1: Semantic textual similarity based on word embeddings. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2020.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

George A. Miller. 1995. WordNet: A lexical database for english. *Commun. ACM* 38(11):39–41. https://doi.org/10.1145/219717.219748.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *Proceedings of ECML'06*. http://dx.doi.org/10.1007/11871842_32.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. Representing sentences as low-rank subspaces. *CoRR* abs/1704.05358. http://arxiv.org/abs/1704.05358.

El Moatez Billah Nagoudi, Jérémy Ferrero, and Didier Schwab. 2017. LIM-LIG at SemEval-2017 Task1: Enhancing the semantic similarity for arabic sentences with vectors weighting. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2017.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of ACL 2010*. http://aclweb.org/anthology/P/P10/P10-1023.pdf.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. *arXiv* https://arxiv.org/pdf/1703.02507.pdf.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. *Gigaword Fifth Edition LDC2011T07*. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/ldc2011t07.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP 2014*. http://www.aclweb.org/anthology/D14-1162.

Nghia The Pham, Germán Kruszewski, Angeliki Lazaridou, and Marco Baroni. 2015. Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model. In *Proceedings of ACL/IJCNLP*. http://www.aclweb.org/anthology/P15-1094.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016*. http://aclweb.org/anthology/C16-1009.

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 SemEval-2016*. http://www.aclweb.org/anthology/S16-1091.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. 2016a. Enhancing the lexvec distributed word representation model using positional contexts and external memory. *CoRR* abs/1606.01283. http://arxiv.org/abs/1606.01283.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. 2016b. Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of ACL*. http://aclweb.org/anthology/P16-2068.

Gilles Serasset. 2015. DBnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web Journal (special issue on Multilingual Linked Open Data)* 6:355–361. https://doi.org/10.3233/SW-140147.

Yang Shao. 2017. HCTI at SemEval-2017 Task 1: Use convolutional neural network to evaluate semantic textual similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2016.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Gregoire Mesnil. 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of CIKM '14*. https://www.microsoft.com/en-us/research/publication/a-latent-semantic-model-with-convolutional-pooling-structure-for-information-retrieval/.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA 2006*. http://mt-archive.info/AMTA-2006-Snover.pdf.

Martyna Śpiewak, Piotr Sobecki, and Daniel Karaś. 2017. OPI-JSA at SemEval-2017 Task 1: Application of ensemble learning for computing semantic textual similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2018.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of SemEval 2015*. http://aclweb.org/anthology/S/S15/S15-2027.pdf.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 Task 1: Leverage kernel-based traditional nlp features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2028.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012a. Takelab: Systems for measuring semantic text similarity. In *Proceedings of *SEM 2012/SemEval 2012*. http://www.aclweb.org/anthology/S12-1060.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012b. TakeLab: Systems for measuring semantic text similarity. In *Proceedings of SemEval 2012*. http://www.aclweb.org/anthology/S12-1060.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL (TACL)* 3:345–358. http://aclweb.org/anthology/Q/Q15/Q15-1025.pdf.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016a. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of EMNLP*. https://aclweb.org/anthology/D16-1157.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016b. Towards universal paraphrastic sentence embeddings. In *Proceedings of ICLR 2016*. http://arxiv.org/abs/1511.08198.

John Wieting and Kevin Gimpel. 2017. Revisiting recurrent networks for paraphrastic sentence embeddings. *CoRR* abs/1705.00364. http://arxiv.org/abs/1705.00364.

Hao Wu, Heyan Huang, Ping Jian, Yuhang Guo, and Chao Su. 2017. BIT at SemEval-2017 Task 1: Using semantic information space to evaluate semantic textual similarity. In *Proceedings of SemEval-2017*. http://www.aclweb.org/anthology/S17-2007.

Wei Xu, Chris Callison-Burch, and William B. Dolan. 2015. SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of SemEval 2015*. http://www.aclweb.org/anthology/S15-2001.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* 2:67–78. http://aclweb.org/anthology/Q14-1006.

WenLi Zhuang and Ernie Chang. 2017. Neobility at SemEval-2017 Task 1: An attention-based sentence similarity model. In *Proceedings SemEval-2017*. http://www.aclweb.org/anthology/S17-2023.