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 Inc.
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 of the language tracks*. We summarize performance and review a selection of well performing methods. Analysis highlights common errors, providing insight into the limitations of the current state-of-the-art. To support ongoing work on semantic representations, *STS Benchmark* is introduced as a new shared training and evaluation set based on a multi-year selection of English STS pairs (2012-2017).

1 Introduction

Semantic Textual Similarity (STS) assesses the degree to which two sentences are semantically 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 assessment is performed outside of any contextualizing text.

Accurately inferring the meaning similarity between sentences is a foundational natural language understanding problem. The systems and techniques explored as a part of STS have a broad range of applications including machine translation (MT), summarization, generation, question answering (QA), short answer grading, semantic search, dialog and conversational systems. STS allows for the independent evaluation of methods for computing semantic similarity drawn from a diverse set of domains that are otherwise 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 improve natural language processing and language understanding applications where knowing the similarity in meaning between two pieces of text is relevant to the behavior of the system.

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 a graded degree of meaning overlap rather than making a binary classification of a particular relationship. Semantic relatedness also captures a graded semantic relationship between two texts. However, relatedness is non-specific about the nature of the relationship with contradictory material still being a candidate for a high score.

To encourage and support research in this area, 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, 2013, 2014, 2015, 2016). During this time, diverse similarity methods and data sets[1] have been

[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 forums, plagiarism, MT post-editing and Q&A data sets. Data sets are summarized on: http://ixa2.si.ehu.es/stswiki.

explored. Early methods focus on lexical semantics, surface form matching and basic syntactic similarity (Bär et al., 2012; Šarić et al., 2012a; Jimenez et al., 2012a). Strong new similarity signals emerged during subsequent evaluations, such as Sultan et al. (2015)'s alignment based method. Deep learning methods have recently become competitive with top performing feature engineered systems (He et al., 2016). The best performance tends to be obtained by ensembling traditional feature engineered signals with deep learning models (Rychalska et al., 2016).

Significant research effort has focused on STS over English sentence pairs.[2] English STS is a 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 is structured to emphasizes performance on Arabic and Spanish as well as cross-lingual pairings of English with material in Arabic, Spanish and Turkish. The primary ranking for the task combines performance on all of these different language conditions except English-Turkish, which was run as a surprise language track. Even with the departure from prior years, the task attracted 31 teams producing 84 system submissions. 17 teams produced a total of 44 system submissions that scored pairs in all of the languages necessary for placement under the primary ranking. Each of these 44 submissions also took part in the English-Turkish surprise language track.

STS data sets from varying years have been used extensively for research on state-of-the-art models of sentence level semantic representations. To encourage the use of a common evaluation set for assessing sentence level semantic representations, we present STS Benchmark, a publicly available selection of the English data sets from previous STS tasks during the period (2012-2017).

2 Task Overview

STS is the assessment of pairs of sentences according to their degree of semantic similarity. Performing the task involves producing real-valued similarity scores for pairs of sentences. No constraints are placed on the data or tools that can be used by STS systems, with the only exception that supervised annotations over the test data are not allowed.

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

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).

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.

Performance is measured by the Pearson correlation of the model scores with human judgments. The ordinal scale in Table 1 guides human assignment, ranging from 0 for no meaning overlap to 5 for meaning equivalence. Intermediate values reflect interpretable levels of partial overlap.

Scores are designed to be appropriate according to a reasonable human judge. Using reasonable human interpretations of natural language semantics was popularized by the related textual entailment task (Dagan et al., 2010). This results in the task being more challenging as the resulting similarity scores reflect both pragmatic and world knowledge. The resulting similarity scores are more interpretable and useful for downstream systems.

3 Evaluation Data

The Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) is the primary evalu-

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	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	dataset	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.

ation 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]

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

Track 4b Spanish-English investigates the relationship between STS and MT quality estimation. The WMT quality estimation data includes Spanish translations of English sentences from a variety of methods including RBMT, SMT, hybrid-MT

[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.

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
2017	Trial	23	Mixed STS 2016
2017	MT	1000	WMT13 Translation Task

Table 5: Spanish-English training data.

year	dataset	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.

and human translation. Translations are annotated with the time required to correct them using post-editing and Human-targeted Translation Error Rate (HTER) (Snover et al., 2006).[4] Participants are not allowed to use gold quality estimation annotations to inform semantic similarity scores.

3.1 Tracks

Table 2 summaries the evaluation data by track. There are six tracks spanning four languages: Arabic, English, Spanish and Turkish. The tracks are: *(1) Arabic*; *(2) Arabic-English*; *(3) Spanish*; *(4 a/b) Spanish-English*; *(5) English*; *(6) Turkish-English*. Track 4 has subtracks: 4a draws from SNLI; 4b pulls 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 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 annotation 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. The transla-

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

year	dataset	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	dataset	pairs	source
2017	MSRpar	1039	newswire
2017	MSRvid	749	videos
2017	SMTeuroparl	422	WMT eval.

Table 8: Arabic-English parallel data.

tors are given an English sentences and its Arabic machine translation[5] and 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 from being used to inform semantic similarity scores. Creating pairs uniformly at random would result in predominately low similarity scores. The *word embedding similarity* selection heuristic from STS 2016 (Agirre et al., 2016) is used to help find interesting pairs. First, sentence embeddings are computed as the sum of sentence word embeddings, $\mathbf{v}(s) = \sum_{w \in s} \mathbf{v}(w)$.[7] Sentence pairs likely to have some 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 semantic similarity 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 using Amazon Mechanical Turk.[8] Annotation is performed on the English sentences from SNLI. Semantic similarity 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 are the mean of the five individual annotations.

4.2 Expert Annotation

English-Spanish WMT quality estimation pairs for Track 4b are annotated for semantic similarity by a University of Sheffield graduate student who is 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 more challenging than assessing the heuristically constructed pairs in other tracks. To accurately assess semantic similarity 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.

English, Spanish and English-Spanish training data is 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

[5] Obtained from 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.

scoring 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.

Similarity scores for the heuristically paired SNLI sentences tend to be slightly lower than recent shared task years and much lower than early years due to differences in data selection, annotation and filtering. The average 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. The average similarity score on the MT quality estimation data from track 4b is 4.0.

6 System Evaluation

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

6.1 Participation

The task saw strong participation with 31 teams producing 84 submissions. Table 9 summarizes participation by track. 17 teams provided 44 systems that participated in all tracks. Traces of the focus on English STS 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.

6.2 Evaluation Metric

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

6.3 CodaLab

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

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

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	45

Table 9: Participation by shared task track.

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

Table 10 summarizes performance. ECNU is best overall (avg r: 0.7316) and obtained first place on: track 2, Arabic-English (r: 0.7493); track 3, Spanish (r: 0.8559); and track 6, Turkish-English (r: 0.7706). BIT achieved the best performance on track 1, Arabic, (r: 0.7543). CompiLIG placed first on track 4a, SNLI Spanish-English pairs (r: 0.8302). SEF@UHH performed best on the track 4b WMT quality estimation pairs (r: 0.3407). RTV did best on the track 5 English data (r: 0.8547), followed closely by DT_Team (r: 0.8536).

The most 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 semantic similarity methods for MT quality estimation may benefit using alternatives to Pearson correlation for evaluation.[14]

[12] Words obtained using ar, es and en treebank tokenizers.

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

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

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)	0.7316	**0.7440**	**0.7493•**	**0.8559•**	**0.8131**	**0.3363**	**0.8518**	**0.7706•**
ECNU (Tian et al., 2017)	0.7044	**0.738**	0.7126	**0.8456**	0.7495	**0.3311**	**0.8181**	0.7362
ECNU (Tian et al., 2017)	0.6940	**0.7271**	0.6975	0.8247	0.7649	**0.2633**	**0.8387**	0.7420
BIT (Wu et al., 2017)*	0.6789	**0.7417**	0.6965	**0.8499**	0.7828	0.1107	**0.8400**	**0.7305**
BIT (Wu et al., 2017)*	0.6703	0.7535	0.7007	0.8323	0.7813	0.0758	0.8161	0.7327
BIT (Wu et al., 2017)	0.6662	**0.7543•**	0.6953	**0.8289**	0.7761	0.0584	**0.8222**	**0.7280**
HCTI (Shao, 2017)	0.6598	**0.7130**	0.6836	**0.8263**	0.7621	0.1483	0.8113	0.6741
MITRE (Henderson et al., 2017)	0.6590	**0.7294**	0.6753	0.8202	0.7802	0.1598	0.8053	0.6430
MITRE (Henderson et al., 2017)	0.6587	**0.7304**	0.6740	0.8201	0.7799	0.1574	0.8048	0.6441
FCICU (Hassan et al., 2017)	0.6190	**0.7158**	0.6782	**0.8484**	0.6926	0.0254	**0.8272**	0.5452
neobility (Zhuang and Chang, 2017)	0.6171	0.6821	0.6459	0.7928	0.7169	0.0200	0.7927	0.6696
FCICU (Hassan et al., 2017)	0.6166	**0.7158**	0.6781	**0.8489**	0.6854	0.0214	**0.8280**	0.5390
STS-UHH (Kohail et al., 2017)	0.6058	0.6781	0.6307	0.7713	0.7201	0.0481	0.7989	0.5937
RTV	0.605	0.6713	0.5595	0.7485	0.7050	0.0761	**0.8541**	0.6204
HCTI (Shao, 2017)	0.5988	0.4373	0.6836	0.6709	0.7621	0.1483	**0.8156**	0.6741
RTV	0.5980	0.6689	0.5482	0.7424	0.6999	0.0734	**0.8541**	0.5989
MatrusriIndia	0.5960	0.6860	0.5464	0.7614	0.7118	0.0572	0.7744	0.6349
STS-UHH (Kohail et al., 2017)	0.5725	0.6104	0.5910	0.7204	0.6338	0.1205	0.7339	0.5972
SEF@UHH (Duma and Menzel, 2017)	0.5676	0.5790	0.5384	0.7423	0.5866	0.1802	0.7256	0.6211
SEF@UHH (Duma and Menzel, 2017)	0.5644	0.5588	0.4789	0.7456	0.5739	**0.3069**	0.7880	0.4990
RTV	0.5633	0.6143	0.4832	0.6863	0.6140	0.0829	**0.8547•**	0.6079
SEF@UHH (Duma and Menzel, 2017)	0.5528	0.5774	0.4813	0.6979	0.5660	**0.3407•**	0.7186	0.4878
neobility (Zhuang and Chang, 2017)	0.5195	0.1369	0.6259	0.7792	0.6930	0.0044	0.7556	0.6418
neobility (Zhuang and Chang, 2017)	0.5025	0.0369	0.6207	0.7690	0.6947	0.0147	0.7535	0.6279
MatrusriIndia	0.4975	0.5703	0.4340	0.6786	0.5563	0.0857	0.6579	0.4994
NLPProxem	0.4902	0.5193	0.5313	0.6642	0.5144	0.0996	0.6256	0.4767
UMDeep (Barrow and Peskov, 2017)	0.4792	0.4753	0.4939	0.5165	0.5615	0.1609	0.6174	0.5293
NLPProxem	0.4790	0.5506	0.4369	0.6381	0.5079	0.1414	0.6463	0.4320
UMDeep (Barrow and Peskov, 2017)	0.4773	0.4587	0.5199	0.5148	0.5232	0.1300	0.6222	0.5725
Lump (España Bonet and Barrón-Cedeño, 2017)*	0.4725	0.6052	0.1829	0.7574	0.4327	0.0116	0.7376	0.5800
Lump (España Bonet and Barrón-Cedeño, 2017)*	0.4704	0.5508	0.1357	0.7676	0.4825	0.1112	0.7269	0.5179
Lump (España Bonet and Barrón-Cedeño, 2017)*	0.4438	0.6287	0.1805	0.7380	0.4447	0.0151	0.7347	0.3652
NLPProxem	0.4070	0.5327	0.4773	0.0016	0.5506	0.1440	0.6681	0.4746
RTM (Biçici, 2017)*	0.3669	0.3365	0.1711	0.6990	0.6004	0.1455	0.5468	0.0687
UMDeep (Barrow and Peskov, 2017)	0.3521	0.3905	0.3713	0.4588	0.3482	0.0586	0.4727	0.3644
RTM (Biçici, 2017)*	0.3291	0.3365	0.0025	0.5682	0.5054	0.1368	0.6405	0.1136
RTM (Biçici, 2017)*	0.3278	0.4156	0.1332	0.4841	0.4583	0.2347	0.5632	0.0055
ResSim (Bjerva and Östling, 2017)	0.3148	0.2892	0.1045	0.6613	0.2389	0.0305	0.6906	0.1884
ResSim (Bjerva and Östling, 2017)	0.2938	0.3120	0.1288	0.6920	0.1002	0.0162	0.6877	0.1195
ResSim (Bjerva and Östling, 2017)	0.2145	0.0033	0.1098	0.5465	0.2262	0.0199	0.5057	0.0902
LIPN-IIMAS (Arroyo-Fernández and Meza Ruiz, 2017)	0.1067	0.0471	0.0769	0.1527	0.1719	0.1446	0.0738	0.0800
LIPN-IIMAS (Arroyo-Fernández and Meza Ruiz, 2017)	0.0926	0.0214	0.1292	0.0458	0.0120	0.0191	0.2038	0.2168
hjpwhu	0.0480	0.0412	0.0639	0.0617	0.0204	0.0624	0.0114	0.0753
hjpwhu	0.0294	0.0477	0.0204	0.0763	0.0046	0.0257	0.0069	0.0246
compiLIG (Ferrero et al., 2017)					**0.8302•**	0.1550		
compiLIG (Ferrero et al., 2017)					0.7684	0.1464		
compiLIG (Ferrero et al., 2017)					0.7910	0.1494		
DT_TEAM (Maharjan et al., 2017)							**0.8536**	
DT_TEAM (Maharjan et al., 2017)							**0.8360**	
DT_TEAM (Maharjan et al., 2017)							**0.8329**	
FCICU (Hassan et al., 2017)							**0.8217**	
ITNLPAiKF (Liu et al., 2017)							**0.8231**	
ITNLPAiKF (Liu et al., 2017)							**0.8231**	
ITNLPAiKF (Liu et al., 2017)							**0.8159**	
L2F/INESC-ID (Fialho et al., 2017)*				0.7616	0.0191	0.0544	0.7811	0.0293
L2F/INESC-ID (Fialho et al., 2017)							0.6952	
L2F/INESC-ID (Fialho et al., 2017)*				0.6385	0.1561	0.0524	0.6661	0.0356
LIM-LIG (Nagoudi et al., 2017)		**0.7463**						
LIM-LIG (Nagoudi et al., 2017)		**0.7309**						
LIM-LIG (Nagoudi et al., 2017)		0.5957						
MatrusriIndia		0.6860		0.7614	0.7118	0.0572	0.7744	0.6349
NRC*					0.4225	0.0023		
NRC					0.2808	0.1133		
OkadaNaoya							0.7704	
OPI-JSA (Śpiewak et al., 2017)							0.7850	
OPI-JSA (Śpiewak et al., 2017)							0.7342	
OPI-JSA (Śpiewak et al., 2017)							0.6796	
PurdueNLP (Lee et al., 2017)							0.7928	
PurdueNLP (Lee et al., 2017)							0.5535	
PurdueNLP (Lee et al., 2017)							0.5311	
QLUT (Meng et al., 2017)*							0.6433	
QLUT (Meng et al., 2017)							0.6155	
QLUT (Meng et al., 2017)*							0.4924	
SIGMA							0.8047	
SIGMA							0.8008	
SIGMA							0.7912	
SIGMA_PKU_2							0.8134	
SIGMA_PKU_2							0.8127	
SIGMA_PKU_2							0.8061	
STS-UHH (Kohail et al., 2017)							0.8093	
UCSC-NLP							0.7729	
UdL (Al-Natsheh et al., 2017)							0.8004	
UdL (Al-Natsheh et al., 2017)*							0.7901	
UdL (Al-Natsheh et al., 2017)							0.7805	
cosine baseline	0.5370	0.6045	0.5155	0.7117	0.6220	0.0320	0.7278	0.5456

* Corrected or late submission

Table 10: STS 2017 rankings ordered by average correlation across tracks 1-5. 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).

Results tend to decrease on cross-lingual tracks. On SNLI data, the baseline drops by almost 10 points for Arabic-English and Spanish-English vs. Arabic and Spanish. Many systems show smaller decreases. ECNU's top ranking entry does slightly better on track 2 than track 1, Arabic-English vs. Arabic, with only a 4 point drop from track 3 to 4b, Spanish vs. Spanish-English.

6.6 Methods

Participating teams explore techniques ranging from deep learning models to elaborate feature engineered systems. Prediction signals include surface similarity scores such as edit distance or 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] Below we highlight interesting and successful methods.

ECNU (Tian et al., 2017) The best system overall is ECNU that uses a large feature set including: n-gram overlap; edit distance; longest common prefix/suffix/substring; tree kernel similarity (Moschitti, 2006); monolingual alignment (Sultan et al., 2015); summarization and MT evaluation metrics (BLEU, GTM-3, NIST, WER, METEOR, ROUGE); and kernel similarity of vectors defined by bags-of-words, bags-of-dependency-triples and pooled word-embeddings. Models are trained with RandomForest (RF), Gradient Boosting (GB) and XGBoost (XGB). Deep learning similarity scores are computed using a variety of paraphrastic sentence embeddings methods: averaged word embeddings, projected word embeddings, a deep averaging network (DAN) and LSTM (Wieting et al., 2016). The best run ensembles all three classifier types with the deep learning similarity scores. Two other runs use a single classifier, either RF or GB, and no ensembling with deep learning similarity scores. ECNU took first place on the combined primary evaluation, Arabic-English (Track 2), Spanish (Track 3) and Turkish-English (Track 7).

BIT (Wu et al., 2017) Second place overall is achieved by BIT that focused on a WordNet based information content (IC) feature. BIT developed three systems one that exclusively made use of the IC feature. Another ensembles this feature with Sultan et al. (2015)'s alignment based similarity method, while the third system ensembles the IC feature with cosine similarity of summed word embeddings with an IDF derived weighting scheme. The IC feature in isolation is able to out perform every other system except those submitted by ECNU. Combining the IC feature with weighted word embedding similarity provides the best performance. The BIT team took 1st place on Arabic (Track 1).

HCTI (Shao, 2017) Third place overall is obtained by HCTI using a deep learning model that is similar to a convolutional Deep Structured Semantic Model (CDSSM) (Chen et al., 2015; Huang et al., 2013). The model has twin convolutional neural networks (CNNs) that generate sentence level embeddings. Sentence level embeddings are compared using cosine similarity and element wise difference with the scores being feed to another neural network to generate a similarity label. UMDeep (Barrow and Peskov, 2017) took a similar approach but using LSTMs rather than CNNs to generate the sentence embeddings.

MITRE (Henderson et al., 2017) Fourth place overall is MITRE that, like ECNU, took an ambitious feature engineering approach with some of the features based on deep learning models. Features include the cosine similarity of aligned word embeddings, the output of the TakeLab STS system (Šarić et al., 2012b), Summarization and MT evaluation features (BLEU, WER, PER, ROUGE), an RNN over similarity signals and a BiLSTM model that represents the current state-of-the-art for the SNLI entailment task (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 similarity to English pairs. Alignment based 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 of FCICU's runs average

[15] Within the highlighted submissions, the following used a monolingual English system fed by MT: ECNU, BIT, HCTI and MITRE. HCTI also submitted a run using separate ar, es and en trained models that underperformed using their en model with MT on ar and es. CompiLIG's model is cross-lingual but includes a word alignment feature that depends on MT for the cross-lingual pairs. SEF@UHH built separate ar, es, and en models with bi-directional MT used for cross-lingual pairs. LIM-LIG and DT_Team only participated in monolingual tracks.

the Babelnet based scores with soft-cardinality similarity scores (Jimenez et al., 2012b).

CompiLIG (Ferrero et al., 2017) The best Spanish-English performance on SNLI data was achieved by CompiLIG, which only participated in the two Spanish-English tracks. The system makes use of featured engineered cross-language signals including: character n-grams, cross-lingual conceptual similarity using DBNary (Serasset, 2015) and k-best word embeddings, 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.[16] Word embeddings are trained on 5.8B+ words and summed into sentence embeddings using uniform, POS and IDF weighting schemes. Sentence similarity is computed by cosine. 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 track 5 systems.

DT_Team (Maharjan et al., 2017) Second place on English (Track 5)[17] is DT_Team using feature engineering combined with the following deep learning models: DSSM (Huang et al., 2013), CDSSM (Shen et al., 2014) and skip-thoughts (Kiros et al., 2015). The feature sets include: unigram overlap, summed word alignments scores, fraction of unaligned words, difference in word counts by type (all, adj, advert, nouns, verbs), and min to max ratios of words by type. Select features have a multiplicative penalty for unaligned words. Similarity prediction uses linear SVM regression, linear regression or gradient boosted regression.

SEF@UHH (Duma and Menzel, 2017) First place on the challenging Spanish-English MT pairs (Track 4b) is SEF@UHH and uses of no supervised training data. Similarity scores compare paragraph vectors (Le and Mikolov, 2014) using cosine, negation of Bray-Curtis dissimilarity or vector correlation. Bray-Curtis performs well overall, while cosine does best on the Spanish-English MT pairs.

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

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

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

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).

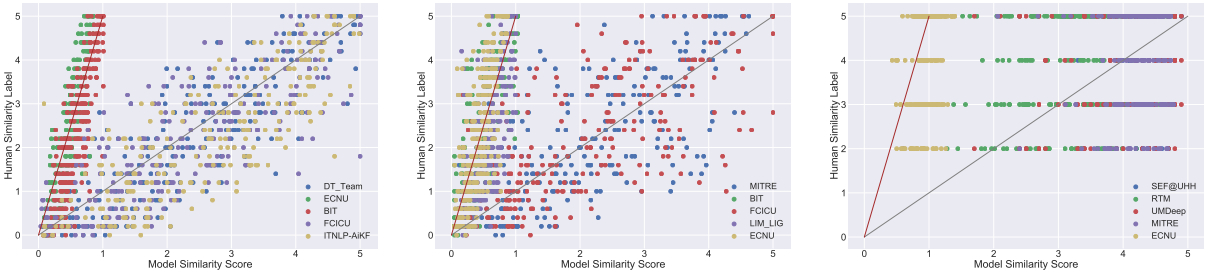
7 Analysis

Figure (1) plots model similarity scores against human semantic similarity 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,[19] we find that 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 plot shows a very weak relationship between the predicted and gold scores.

Table 12 provides examples of difficult sentence pairs for participant systems. The examples illustrate common sources of error even for well-ranking systems such as: (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" and "deserted" are minor details when contrasting "in a deserted field" with "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" appear to be instances of (v) *semantic blending*, whereby appending "with a woman" to "boy" mistakenly brings its representation closer to that of "girl".

In the multilingual and cross-lingual setting, theses issues are magnified by translation errors for systems that use MT followed by the application of a monolingual similarity model. For the track

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



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

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

Track 5: English-English	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
Track 4b: Spanish-English MT	Human	SEF@UHH	ECNU	RTM	UMDeep	MITRE
Give it a chance. Darle una oportunidad.	5.0	4.3	2.0	3.1	2.4	2.7
Later, I settled on ”Mexican-American.” Ms tarde, he asentado en ”mexicano-americana”.	4.0	3.3	0.3	2.8	3.2	4.6
The most common verbal indicators are subtle. Los indicadores ms comunes son sutiles verbal.	3.0	4.7	3.5	4.3	4.5	4.8
And in the United States, we’re considered Mexican. Y en los Estados Unidos, estamos considerando mexicanos.	2.0	4.3	2.4	2.2	2.8	3.4

Table 12: Difficult sentence pairs and scores assigned by top performing systems.[18]

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	answers-forums	15	0	375	0
forum	answer-answer	16	0	0	254

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

4b Spanish-English MT pairs, the poor predictions can in part be attributed to many systems using MT to re-translate the output of another MT system.

7.1 Contrasting Cross-lingual Semantic Similarity 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.[20] The actual distribution and spread 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 semantic similarity scores for track 4b indicate that MT systems are surprisingly good at preserving meaning. However, even for a human, interpreting the meaning 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 semantic similarity scores is 0.41, which shows that translation quality measures and semantic similarity are only moderately correlated. Differences are in part explained by translation quality scores penalizing all mismatches between the source segment and its translation, whereas semantic similarity focuses only on differences in meaning. However, the difficult interpretation work required for semantic similarity annotation may increase the risk of inconsistent and subjective labels. The annotations for MT quality estimation are produced as by-product of

[20]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.

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)	80.0	72.6
InferSent	Sentence embedding from a 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. Results for select participants and baseline models.

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

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 a training, development and test sets.[21] 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 be optimized on the development data and then locked against further changes. Using the benchmark to evaluate models will enable comparable assessments across different research efforts and a means for tracking and establishing state-of-the-art semantic similarity performance.

Table 14 shows the results of some of the best systems from Track 5 (EN-EN)[22] 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,[23] with the exception of PV-DBOW that 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 pre-processed to closely match the tokenization of the pre-trained models.[24] Default inference hyperpa-

[21] 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 sources as long as no supervised annotations of the development or test set are used.

[22] 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.

[23] **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://nlp.stanford.edu/projects/glove/, embeddings from glove.840B.300d.zip, first 10 principle components removed, $\alpha = 0.001$, dev experiments varied α, principle components removed and whether GloVe or Word2Vec word embeddings were used; **C-PHRASE**: http://clic.cimec.unitn.it/composes/cphrase-vectors.html; **PV-DBOW**: https://github.com/jhlau/doc2vec, APNEWS 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.

[24] **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 Fast-

rameters 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, even better performance is obtained by task participant systems. There is still substantial room for further improvement. To follow the current state-of-the-are, visit the leaderboard on the semantic textual similarity website.[25]

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 (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 with semantic similarity labels is available.

To provide a standard benchmark for English semantic similarity, we present 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.

Text: 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.

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

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, 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, 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.

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/.

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.

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.

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 the 10th International Workshop on Semantic Evaluation (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. 2016. Towards universal paraphrastic sentence embeddings. In *Proceedings of ICLR 2016*. http://arxiv.org/abs/1511.08198.

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.