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

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

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

[b]Google Inc.
Mountain View, CA

[c]George Washington University
Washington, DC

[d]University of Sheffield
Sheffield, UK

Abstract

Semantic Textual Similarity (STS) measures the meaning similarity of sentences. Pairwise scores are on an ordinal scale, conveying both a degree of similarity and a categorical interpretation of the semantic relationship. Applications of STS methods 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. While prior years primarily emphasized English STS, the 2017 task focuses on STS in the multilingual and cross-lingual setting. We also introduce a track that assess performance specifically for MT quality estimation. While encouraging, we find that performance lags in less well studied STS languages and language pairings (e.g., Arabic). The MT quality estimation focused track serves as a good benchmark and highlights the importance of difficult fine grained distinctions. While more challenging than prior years, the task still obtained strong participation from 31 teams, with 17 participating in *all of the language tracks* for 2017. Finally, with the aim of providing a standard benchmark for the future, we present STS Benchmark, a selection of the English datasets in previous STS tasks (2012-2017).

1 Introduction

Semantic Textual Similarity (STS) assesses the degree to which two snippets of text 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 snippets being scored are approximately one sentence in length, with their assessment being performed outside of any contextualizing text.

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 may improve any natural language processing or language understanding application where knowing the similarity in meaning between two pieces of text is relevant to the behavior of the system.

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, a diverse set of genres and data sources have been explored (i.a., news headlines, video and image descriptions, glosses from lexical resources including WordNet (Miller, 1995; Christiane Fellbaum, 1998), FrameNet (Baker et al., 1998), OntoNotes (Hovy et al., 2006), web discussion forums, plagiarism, MT postediting and Q&A data sets). The information and files regard-

The authors of this paper are listed in alphabetic order.

ing the current and preious tasks is summarized in a website[1].

With the aim of providing a standard benchmark for the future, we present STS Benchmark, a selection of the English datasets in previous STS tasks in this period (2012-2017). STS Benchmark is presented in Section 6) and is publicly available. We have gathered a selection

English STS is at this point a well-studied problem. Initially, the shared task only involved pairs of English text. The 2014 shared task introduced a complementary Spanish track and the 2016 shared task introduced a pilot track on cross-lingual Spanish-English STS. However, in terms of participation numbers, English STS has historically remained as the focus of the shared task, with state-of-the-art systems achieving over 81% correlation (see Section 6).

To promote progress in other languages, the 2017 STS shared task equally emphasizes performance on not just English pairs but also Arabic and Spanish pairs as well as cross-lingual pairings of English text with material in Arabic, Spanish and Turkish. This emphasis is reflected in that the primary ranking for the shared task combines performance on all of these different language conditions except for English-Turkish, as the latter was run as a surprise language track.

Even with this significant departure from prior years, the shared task attracted 31 teams producing 84 system submissions. Within the participants, 17 teams produced a total of 44 system submissions that processed pairs in all of the languages necessary for placement under the primary evaluation metric. Interestingly, all 44 submissions ranked under the primary evaluation metrics also provided STS scores for the Turkish-English surprise language track.

The landscape of the participant performance suggests that existing methods perform well on English and Spanish, with both languages being explored during prior STS evaluations. However, performance was markedly degraded on monolingual Arabic pairs, cross-lingual Arabic-English pairs as well as Turkish-English pairs. This suggests that further work may necessary on STS models that perform adequately on new languages or possibly equivalently when less training data is available for a particular language.

[1] http://ixa2.si.ehu.es/stswiki

2 Task Overview

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

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

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

In the context of the STS task, meaning equivalence is defined operationally as two snippets of text that mean the same thing when interpreted by a reasonable human judge. The operational approach to sentence level semantics was popularized by the recognizing textual entailment task (Dagan et al., 2010). It has the advantage that it

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.*
	In May 2010, the troops attempted to invade Kabul. The US army invaded Kabul on May 7th last year, 2010.
3	*The two sentences are roughly equivalent, but some important information differs/missing.*
	John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.
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.*
	John went horse back riding at dawn with a whole group of friends. Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

Table 1: Similarity scores with explanations and examples for the English task.

allows the labeling of sentence pairs by human annotators without any training in formal semantics, while also being more useful and intuitive to work with for downstream systems. Beyond just sentence level semantics, the operationally defined STS labels also reflect both world knowledge and pragmatic phenomena.

As in prior years, 2017 shared task participants are allowed to make use of existing resources and tools (e.g., WordNet, Mikolov et al. (2013)'s word2vec). Participants are also allowed to make unsupervised use of arbitrary data sets, even if such data overlaps with the announced sources of the evaluation data.

3 Evaluation Tracks

The evaluation includes five different tracks and includes data spanning four languages: Arabic, English, Spanish and Turkish. The majority of the data is sourced from the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) with additional data related to machine translation quality estimation sourced from the WMT 2014 quality estimation shared task (Bojar et al., 2014).

The evaluation is organized into six tracks with each track corresponding to the following languages and language pairs:

- Track 1: Arabic
- Track 2: Arabic-English
- Track 3: Spanish
- Track 4(a/b): Spanish-English
- Track 5: English
- Track 6: Turkish-English

Tracks 2, 4 and 6 require systems to assess cross-lingual pairs as the individual members within each pair of text snippets are written in different languages. Track 4 is split into two subtracts: 4a draws from data sourced from SNLI; 4b is built using data sourced from WMT's quality estimation task. Track 6 is a surprise language track, with participants only being informed of the language pair at the beginning of the evaluation period and with no STS annotated training data being provided for this pair. The overall performance is the average of the seven scores from the individual distinct evaluation sets, with tracks 4a and 4b contributing as individual scores.

The amount of training data available for each track is given in: Table (2) English; Table (3) Spanish; Table (4) Spanish-English; Table (5) Arabic; and Table (6) Arabic-English. No training data is provided for the Turkish-English surprise language track. The 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. For the tracks involving Arabic, participants are also supplied with

year	dataset	pairs	source
2012	MSRpar	1500	newswire
2012	MSRvid	1500	videos
2012	OnWN	750	glosses
2012	SMTnews	750	WMT eval.
2012	SMTeuroparl	750	WMT eval.
2013	HDL	750	newswire
2013	FNWN	189	glosses
2013	OnWN	561	glosses
2013	SMT	750	MT eval.
2014	HDL	750	newswire headlines
2014	OnWN	750	glosses
2014	Deft-forum	450	forum posts
2014	Deft-news	300	news summary
2014	Images	750	image descriptions
2014	Tweet-news	750	tweet-news pairs
2015	HDL	750	newswire headlines
2015	Images	750	image descriptions
2015	Ans.-student	750	student answers
2015	Ans.-forum	375	Q&A forum answers
2015	Belief	375	committed belief
2016	HDL	249	newswire headlines
2016	Plagiarism	230	short-answer plag.
2016	Postediting	244	MT postedits
2016	Ans.-Ans.	254	Q&A forum answers
2016	Quest.-Quest.	209	Q&A forum questions
2017	Trial	23	Mixed STS 2016

Table 2: English training data (2012, 2013, 2014, 2015. 2016).

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 3: Spanish training data.

a small amount of parallel data drawn from prior STS evaluations sets, Table (7).

3.1 Data Collection

Pairs are constructed from English sentences samples from the SNLI data set for tracks 1, 2, 3, 4a, 5 and 6. As described in more detail below, SNLI pairs are heuristically constructed using bag-of-words based embeddings to guide the process. For the test data, we did not directly use the pairings present in SNLI data. Since the pairings present in the SNLI corpus are annotated with entailment labels, directly including theses pairs could have made it too easy for teams to use this information to inform their labels across most of the 2017 evaluation sets.

In the case of track 4b, the source sentences come from the test set of the WMT13 translation shared task (Bojar et al., 2013), while the machine

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 4: 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 5: Arabic training data.

translations were produced by a phrase-based statistical MT system (referred to as the 'uedin' baseline at WMT). This data was selected such that we could compare STS labels with translation quality labels that were already available for the same sentences.

3.2 Translation into Arabic, Spanish and Turkish

English sentences from SNLI are human translated to appropriate track specific languages. Arabic translation is provided by the Qatar Computing Research Institute (QCRI). Spanish sentences are produced by a native Spanish speaker. Turkish sentences are obtained through SDL.

3.3 Embedding Space Pair Selection

Constructing pairings of sentences at random from a corpus would be dominated by low STS scores. In order to achieve better data set balance, the SNLI pairs are selected at random but constrained to have a minimal degree of similarity according to a bag-of-words based embedding representation of the two sentences. To assess the degree of similarity, we compute the cosine between an embedding space representation of the two text snippets.

Equation (1) illustrates the construction of the snippet embedding space representation, $\mathbf{v}(s)$, as the sum of the embeddings for the individual words, $\mathbf{v}(w)$, in the snippet. The cosine similarity can then be computed as in equation (2).

$$\mathbf{v}(s) = \sum_{w \in s} \mathbf{v}(w) \quad (1)$$

year	dataset	pairs	source
2017	Trial	23	Mixed STS 2016
2017	MSRpar	1020	newswire
2017	MSRvid	736	videos
2017	SMTeuroparl	406	WMT eval.

Table 6: Arabic-English training data.

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

Table 7: Arabic-English parallel data.

$$sim_v(s_1, s_2) = \frac{\mathbf{v}(s_1)\mathbf{v}(s_2)}{||\mathbf{v}(s_1)||||\mathbf{v}(s_2)||} \quad (2)$$

We make use of pre-trained 50-dimensional embeddings available from the GloVe package (Pennington et al., 2014).[2] The embeddings were trained using GloVe over a combination of Gigaword 5 (Robert Parker and Maeda, 2011) and English Wikipedia.

4 Annotation

Annotation of pairs with STS labels made use of a mixture of Crowdsourcing using mechanical turk and expert annotation. Crowdsourcing was used for tracks 3 (Spanish), 4a (SNLI English-Spanish), track 6 (Turkish-English), and track 5 (English). Track 4b (WMT English-Spanish) was annotated by a single expert annotator. Tracks 1 (Arabic) and Tracks 2 (Arabic-English) were annotated by a team of annotators from the Qatar foundation.

4.1 Crowdsourced Annotations

Crowdsourced annotation of STS pairs is performed using Amazon Mechanical Turk.[3] This section describes the templates and annotation parameters and how the gold standard annotations are computed from multiple annotations from crowd workers. Crowdsourcing is used for the SNLI pairs of English-Spanish, Spanish-Spanish, Turkish-English and English-English sentences. The crowd workers annotated English pairs and the annotations were then transferred to the translated pairs.

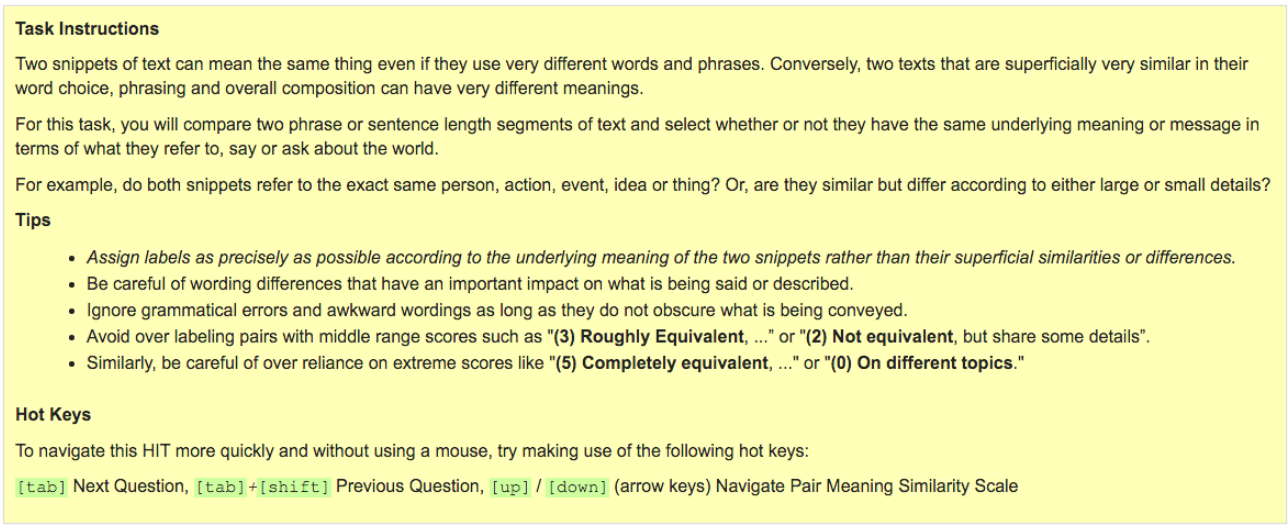
The annotation instructions and template are based on those from last year's English STS subtask (Agirre et al., 2016). Figure 1 provides the instructions. The STS pairs are annotated in batches of 20 pairs. For each batch, annotators are paid $1 USD. Five annotations are collected per pair. Only workers with the MTurk *master* qualification are allowed to perform the annotation, a designation by the MTurk platform that statistical identifies workers who perform high quality work across a diverse set of tasks. Gold annotations are selected as the mean value of the crowdsourced annotations.

4.2 Track 4b English-Spanish Annotation

The annotation of the WMT English-Spanish pairs for Track 4b is performed by a graduate student who is native speaker of Spanish and fluent speaker of English. The data is annotated with STS scores by a single annotator, who has received training by the organizers. This dataset differs significantly from the others in terms of the data distribution and the complexity of the annotation task. Only a subset of the STS labels apply to it: all pairs are at least topically equivalent, since they are translations of each other. The distribution and spread of STS scores is such that only 16%/13% of the training and test instances score below 3, 23% of the instances score 3, and 53% score 5. In other words, most sentences are considered completely equivalent. This data, which had been primarily created for quality estimation, was particularly hard to judge for STS because most translations contained MT errors that rendered them inaccurate and/or disfluent. This also made it hard for STS systems to make accurate predictions. The original annotations for quality estimation were produced as by-product of a post-editing task, where humans were requested to fix the MT output and the edit distance between this output and its post-edited version was used as a quality score. This is a less subjective task which humans can perform consistently. For comparison, the Pearson correlation between the original (gold) quality scores and the gold STS scores is 0.41, which shows that the two annotations (translation quality and semantic similarity) are only moderately correlated. This makes sense, since for translation quality all mismatches between the source segment and its MT are penalised, whereas STS focuses on differences in meaning.

[2] http://nlp.stanford.edu/projects/glove/

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



Task Instructions

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

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

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

Tips

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

Hot Keys

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

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

Figure 1: STS annotation instructions for the SNLI source data.

4.3 Track 1 Arabic and Track 2 Arabic-English Annotation

The data for track 1 and track 2 were annotated by a team of expert annotators from the Qatar foundation. These experts directly annotated the Arabic-Arabic and Arabic-English pairs.

5 System Evaluation

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

5.1 Participation

Even though the 2017 shared task was substantially more challenging than prior years that focused on the well studied problem of monolingual English STS, the shared task attracted strong participation with 31 participating teams submitting 84 systems. A total of 17 teams produced 44 systems that participated in all of the STS 2017 languages and language pairs. Traces of the traditional participation focus on monolingual English STS can still be seen in 12 of the participating teams producing systems exclusively for track 5's English STS pairs. Two teams participate exclusively in tracks 4a and 4b, Spanish STS, which can be seen as a continuation of the Spanish STS subtasks run as a part of STS 2014 and 2015. One team participates exclusively in track 1, Arabic STS, a new language for STS introduced this year.

5.2 Evaluation Metric

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

5.3 CodaLab

As per direction of the SemEval workshop organizers, this year's shared task uses CodaLab,[4] a project out of Microsoft Research for reproducible research.

5.4 Baseline

Similar to prior years, we include a baseline built using a very simple vector space representation. To use this baseline with the cross-lingual pairs, the non-English sentence in each pair is translated into English using a state-of-the-art machine translation system.[5] For all tracks, a language appropriate treebank style tokenizer is then used to split the snippets in each pair into the individual words that will be used for the cosine similarity. The snippets are then projected to a one-hot vector representation such that each dimension corresponds to a word observed in the snippets. If a word appears in a snippet one or more times, the corresponding dimension in the vector is set to one and is otherwise set to zero. The textual similarity score is then computed as the cosine between these vector representations of the two snippets.

[4] https://www.microsoft.com/en-us/research/project/codalab/

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

		Track 1	Track 2	Track 3	Track 4a	Track 4b	Track 5	Track 6
teamname	Primary	AR-AR	AR-EN	SP-SP	SP-EN	SP-EN-WMT	EN-EN	EN-TR
ECNU	0.7316	0.744	0.7493	0.8559	0.8131	0.3363	0.8518	0.7706
ECNU	0.7044	0.738	0.7126	0.8456	0.7495	0.3311	0.8181	0.7362
ECNU	0.694	0.7271	0.6975	0.8247	0.7649	0.2633	0.8387	0.742
BIT	0.6789	0.7417	0.6965	0.8499	0.7828	0.1107	0.84	0.7305
BIT	0.6703	0.7535	0.7007	0.8323	0.7813	0.0758	0.8161	0.7327
BIT	0.6662	0.7543	0.6953	0.8289	0.7761	0.0584	0.8222	0.728
HCTI	0.6598	0.713	0.6836	0.8263	0.7621	0.1483	0.8113	0.6741
MITRE	0.659	0.7294	0.6753	0.8202	0.7802	0.1598	0.8053	0.643
MITRE	0.6587	0.7304	0.674	0.8201	0.7799	0.1574	0.8048	0.6441
FCICU	0.619	0.7158	0.6782	0.8484	0.6926	0.0254	0.8272	0.5452
neobility	0.6171	0.6821	0.6459	0.7928	0.7169	0.02	0.7927	0.6696
FCICU	0.6166	0.7158	0.6781	0.8489	0.6854	0.0214	0.828	0.539
STS-UHH	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	0.705	0.0761	0.8541	0.6204
HCTI	0.5988	0.4373	0.6836	0.6709	0.7621	0.1483	0.8156	0.6741
RTV	0.598	0.6689	0.5482	0.7424	0.6999	0.0734	0.8541	0.5989
MatrusriIndia	0.596	0.686	0.5464	0.7614	0.7118	0.0572	0.7744	0.6349
STS-UHH	0.5725	0.6104	0.591	0.7204	0.6338	0.1205	0.7339	0.5972
SEF@UHH	0.5676	0.579	0.5384	0.7423	0.5866	0.1802	0.7256	0.6211
SEF@UHH	0.5644	0.5588	0.4789	0.7456	0.5739	0.3069	0.788	0.499
RTV	0.5633	0.6143	0.4832	0.6863	0.614	0.0829	0.8547	0.6079
SEF@UHH	0.5528	0.5774	0.4813	0.6979	0.566	0.3407	0.7186	0.4878
neobility	0.5195	0.1369	0.6259	0.7792	0.693	0.0044	0.7556	0.6418
neobility	0.5025	0.0369	0.6207	0.769	0.6947	0.0147	0.7535	0.6279
MatrusriIndia	0.4975	0.5703	0.434	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	0.4792	0.4753	0.4939	0.5165	0.5615	0.1609	0.6174	0.5293
NLPProxem	0.479	0.5506	0.4369	0.6381	0.5079	0.1414	0.6463	0.432
UMDeep	0.4773	0.4587	0.5199	0.5148	0.5232	0.13	0.6222	0.5725
Lump	0.4725	0.6052	0.1829	0.7574	0.4327	0.0116	0.7376	0.58
Lump	0.4704	0.5508	0.1357	0.7676	0.4825	0.1112	0.7269	0.5179
Lump	0.4438	0.6287	0.1805	0.738	0.4447	0.0151	0.7347	0.3652
NLPProxem	0.407	0.5327	0.4773	0.0016	0.5506	0.144	0.6681	0.4746
RTM	0.3669	0.3365	0.1711	0.699	0.6004	0.1455	0.5468	0.0687
UMDeep	0.3521	0.3905	0.3713	0.4588	0.3482	0.0586	0.4727	0.3644
RTM	0.3291	0.3365	0.0025	0.5682	0.5054	0.1368	0.6405	0.1136
RTM	0.3278	0.4156	0.1332	0.4841	0.4583	0.2347	0.5632	0.0055
ResSim	0.3148	0.2892	0.1045	0.6613	0.2389	0.0305	0.6906	0.1884
ResSim	0.2938	0.312	0.1288	0.692	0.1002	0.0162	0.6877	0.1195
ResSim	0.2145	0.0033	0.1098	0.5465	0.2262	0.0199	0.5057	0.0902
LIPN-IIMAS	0.1067	0.0471	0.0769	0.1527	0.1719	0.1446	0.0738	0.08
LIPN-IIMAS	0.0926	0.0214	0.1292	0.0458	0.012	0.0191	0.2038	0.2168
hjpwhu	0.048	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
DT_TEAM							0.8536	
DT_TEAM							0.836	
DT_TEAM							0.8329	
ITNLPAiKF							0.8231	
ITNLPAiKF							0.8231	
FCICU							0.8217	
ITNLPAiKF							0.8159	
SIGMA_PKU_2							0.8134	
SIGMA_PKU_2							0.8127	
STS-UHH							0.8093	
SIGMA_PKU_2							0.8061	
SIGMA							0.8047	
SIGMA							0.8008	
UdL							0.8004	
PurdueNLP							0.7928	
SIGMA							0.7912	
UdL							0.7901	
OPI-JSA							0.785	
L2F/INESC-ID				0.7616	0.0191	0.0544	0.7811	0.0293
UdL							0.7805	
MatrusriIndia		0.686		0.7614	0.7118	0.0572	0.7744	0.6349
UCSC-NLP							0.7729	
OkadaNaoya							0.7704	
OPI-JSA							0.7342	
L2F/INESC-ID							0.6952	
OPI-JSA							0.6796	
L2F/INESC-ID				0.6385	0.1561	0.0524	0.6661	0.0356
QLUT							0.6433	
QLUT							0.6155	
PurdueNLP							0.5535	
PurdueNLP							0.5311	
QLUT							0.4924	
LIM-LIG		0.7463						
LIM-LIG		0.7309						
LIM-LIG		0.5957						
NRC					0.4225	0.0023		
NRC					0.2808	0.1133		
compiLIG					0.7684	0.1464		
compiLIG					0.8302	0.155		
compiLIG					0.791	0.1494		
cosine baseline	0.5370	0.6045	0.5155	0.7117	0.6220	0.0320	0.7278	0.5456

Table 8: STS 2017 Rankings: Late or corrected systems are marked with a * symbol.

5.5 Rankings

The rankings for the 2017 STS evaluation are in Tables 8. On the primary evaluation, which averaged correlation scores across all tracks, the best overall system is ECNU with an averaged correlation score of 0.7316. ECNU also obtained the best performance on the track 2 Arabic-English data (r: 0.7493), the track 3 Spanish-Spanish data (r: 0.8559) and the track 6 Turkish-English surprise language data (r: 0.7706). On track 1's Arabic pairs, BIT achieved the best performance (r: 0.7543). Team compiLIG placed first on the SNLI sourced Spanish-English pairs (r: 0.8302), while SEF@UHH performed best on the quality estimation sourced pairs from WMT (r: 0.3407). RTV performed best on the track 5 English data (r: 0.8547). The baseline system would attain a rank of 23 on the primary track.

The results show that the most challenging tracks with data sourced from SNLI are track 1 with Arabic pairs and track 2 with Arabic-English pairs. The Arabic tracks are even more challenging than the English-Turkish surprise language pairs from track 6. Spanish-English performance is very strong for the SNLI sourced data, but very challenging for data drawn from the WMT quality estimation data. We believe this highlights the importance of making fine grained distinctions for particular downstream applied tasks (Reimers et al., 2016). The systems attained lower results on the cross-lingual tracks compared to their monolingual counterparts, but although the decrease is of around 10 points for the baseline, some systems have a smaller decrease. For instance the best primary system (ECNU run1) has the same results for monolingual Arabic and the crosslingual Arabic tracks, and only a 4 point drop from monolingual Spanish (and English) to crosslingual Spanish.

5.6 Methods

Participating teams used a diverse set of techniques ranging from deep learning models that only make use of a large amount of unannotated training data to elaborate feature engineered systems that build on and improve established signals for accurately estimating STS scores. Techniques include surface similarity scores such edit distance or matching n-grams, scores derived from monolingual word alignments, assessment by MT evaluation metrics, estimates of conceptual similarity as well as the similarity between word and sentence level embeddings. For the cross-lingual tracks, machine translation was widely used to convert the two sentences being compared into the same language. Below we highlight interesting and successful methods and approaches for this year's shared task.

ECNU The best overall system is submitted by ECNU (Tian et al., 2017). ECNU makes use of a strong set of STS features including: n-gram overlap; string matching scores such as edit distance, longest common prefix/suffix/substring; tree kernel similarity (Moschitti, 2006); monolingual alignment (Sultan et al., 2015); a suite of MT evaluation metrics (BLEU, GTM-3, NIST, WER, METEOR, ROUGE); kernel similarity of over vectors defined by bags-of-words, bags-of-dependency-triples and pooled word-embeddings. Features are combined with RandomForest (RF), Gradient Boosting (GB) and XGBoost (XGB). Separate deep learning similarity scores are computed using paraphrastic sentence embeddings (Wieting et al., 2015) computed using either averaged word-embeddings, projected word embeddings, a deep averaging network (DAN) or LSTM. The core system is English only with MT being used to participate in the non-English and cross-lingual tracks. They submit one run that ensembles all three classifier types with the deep learning similarity scores (run3). The two other runs use a single classifier, either RF or GB, and all features except the similarity scores from the deep learning models. The ensemble system that includes the deep learning scores has a performance advantage across all tracks. The ensemble also exhibits a low performance drop when comparing monolingual and cross-lingual tracks, suggesting improved robustness to language changes. In addition to the primary track, ECNU took first place on Arabic-English (Track 2), Spanish (Track 3) and Turkish-English (Track 7).

BIT Second place overall is achieved by the BIT team (Wu et al., 2017). BIT focused on the development of a single strong WordNet based information content (IC) feature for accurately predicting STS scores. BIT developed three systems one that exclusively made use of the IC feature. Another ensembles this feature with Sultan et al. (2015)'s word alignment based similarity method, while the third system ensembles the IC feature with cosine similarity from summed word embeddings with an

IDF derived weighting scheme. The combination of the IC feature with weighted word embedding similarity provides the best performance. However, the IC feature in isolation is able to out perform every other system except those submitted by ECNU. The BIT team took 1st place on Arabic (Track 1)

HCTI Third place overall is obtained by HCTI (Shao, 2017) 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 is composed of twin convolutional neural networks (CNNs) that generate sentence level embeddings. The sentence level embedding vectors are compared using cosine similarity and element wise difference with the resulting layers being feed to another deep neural network to generate the final classification score. A similar approach was also taken by UMDeep (Barrow and Peskov, 2017) but using LSTMs rather than CNNs to generate the sentence level embedding vectors. HCTI augmented the word embeddings provided as input to the CNN with a Boolean bit for matching numeric values and 1-hot encoding of the POS tag for each word.

MITRE Fourth place overall goes to MITRE that, like ECNU, took an ambitious feature engineering approach with some of the features being based on deep learning models. Features include the cosine similarity of aligned word embeddings, the output of the TakeLab STS system (Šarić et al., 2012), MT features (BLEU, WER, PER, ROUGE), an RNN over similarity signals and an Enhanced BiLSTM model that represents the current state-of-the-art for the SNLI entailment task (Chen et al., 2016). The resulting core system is English only. MT was used to participate in the non-English and cross-lingual tracks.

FCICU Similar to BIT, the fifth place system, FCICU, (Hassan et al., 2017) focuses on the development of a single strong signal for predicting STS labels. FCICI propose a new multilingual sense-base alignment that operates over BabelNet synset neighborhoods (Navigli and Ponzetto, 2010). The alignment scores are used with two runs: one that uses the BabelNet similarity scores within a string kernel and another that uses them with a weighted variant of Sultan et al. (2015) method. By using BabelNet, scores can be computed for non-English and cross-lingual sentences without the need for machine translation. Both of FCICU's runs average the Babelnet-based scores with STS scores computed using soft-cardinality (Jimenez et al., 2012).

CompiLIG The best Spanish-English performance on SNLI data was achieved by CompiLIG (Ferrero et al., 2017), 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 The LIM-LIG (Nagoudi et al., 2017) system is notable for achieving second place on Arabic with a system based strictly on weighted word embeddings. LIM-LIG trains word embeddings over a large collection of data sources and combines them into sentence level embeddings using one of three different weighting schemes. The submitted runs explore POS and IDF based weighting. The results show that POS weighting achieves the best performance. After the evaluation, the team explored the multiplicative combination of IDF and POS weights with the resulting combination achieving a correlation score of 0.7667, a higher score than any system submitted to the official evaluation on Arabic.

DT_Team Second place on English (Track 5) is achieved by DT_Team (Maharjan et al., 2017) using a feature engineering based approach combined with the following deep learning models: DSSM (Huang et al., 2013), CDSSM (Shen et al., 2014), and skip-thoughts (Kiros et al., 2015).[6] 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. Variants of select features include a multiplicative penalty for unaligned words. Different runs combine features with either Linear SVM Regressions, linear regression, or gradient boosted regression.

[6]We did not receive a system description paper for the team that achieved first place on English, RTV.

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

Team	Method	Dev	Test
ECNU	Ensemble	0.85	0.81
BIT	WordNet+Embeddings	0.83	0.81
DT_TEAM	Ensemble	0.83	0.79
HCTI	CNN	0.83	0.78
SEF@UHH	Doc2Vec	0.62	0.59

Table 10: STS Benchmark. Results of selected participants.

SEF@UHH First place on the challenging Spanish-English MT pairs (Track 4b) was achieved by SEF@UHH (Duma and Menzel, 2017). The team made use of no supervised training data. Rather similarity scores are derived by comparing paragraph vectors (Le and Mikolov, 2014) using either cosine, negation of Bray-Curtis dissimilarity or vector correlation. While the team notes that Bray-Curtis performs well overall, their best performing system on the Spanish-English MT pairs uses cosine similarity.

6 STS benchmark

STS Benchmark comprises a selection of the English datasets used in the STS tasks organized by us in the context of SemEval between 2012 and 2017. Table 9 gives details of the datasets. In order to provide a standard benchmark to compare among systems, it is organized into train, development and test. The development part can be used to develop and tune hyperparameters of the systems, and the test part should be only used once for the final system. Our goal is that this benchmark is used in the future for setting the state-of-the-art in Semantic Textual Similarity for English, and we are already setting a leaderboard which includes results of some selected systems. Table 10 shows the results of some of the best systems in Track 5 (EN-EN)[7]. Note that the ranking between systems in Track 5 EN-EN is very similar to that in Table 10, except for DT_TEAM and HCTI, who swapped their positions. Please find all details and an updated leaderboard in the official website. [8]

[7]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

7 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 was shifted from English to multilingual and cross-lingual STS involving four different languages: Arabic, Spanish, English and Turkish. Even with this substantial shift 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. We observe that 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 judgements. However, for both pure and cross-lingual Arabic-English pairs and cross-lingual Turkish-English, we observe weaker correlations from participating systems. 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 only a smaller quantities of supervised training data with STS labels exists for the language.

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 the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, CO.

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

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

lingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 497–511. 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 **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada, pages 385–393. 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 *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Atlanta, Georgia, USA, pages 32–43. http://www.aclweb.org/anthology/S13-1004.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *COLING '98 Proceedings of the 17th international conference on Computational linguistics - Volume 1*.

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 the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Alexandre Berard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. Multivec: a multilingual and multilevel representation learning toolkit for nlp. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

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 *Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, WMT, pages 12–58. http://www.aclweb.org/anthology/W/W14/W14-3302.

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 the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 1–44. 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 the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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

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. https://www.microsoft.com/en-us/research/publication/learning-bidirectional-intent-embeddings-by-convolutional-deep-structured-semantic-models-for-spoken-language-understanding/.

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

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

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 the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

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 the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

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 the*

11th International Workshop on Semantic Evaluation (SemEval 2017). Vancouver, Canada.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. ACM International Conference on Information and Knowledge Management (CIKM). 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. 2012. Soft cardinality: A parameterized similarity function for text comparison. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Stroudsburg, PA, USA, SemEval '12, pages 449–453.

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.

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

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 the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

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

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

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *Proceedings of the 17th European Conference on Machine Learning*. Springer-Verlag, Berlin, Heidelberg, ECML'06, pages 318–329. https://doi.org/10.1007/11871842$_3$2.

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 the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '10, pages 216–225. http://dl.acm.org/citation.cfm?id=1858681.1858704.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 87–96. http://aclweb.org/anthology/C16-1009.

David Graff Junbo Kong Ke Chen Robert Parker and Kazuaki Maeda. 2011. Gigaword fifth edition ldc2011t07. Linguistic Data Consortium.

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 the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

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

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

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 Task 1: A global model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. TakeLab: Systems for measuring semantic text similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada, pages 441–448. http://www.aclweb.org/anthology/S12-1060.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *CoRR* abs/1511.08198. 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 the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.