SemEval-2016 Task 2: Interpretable Semantic Textual Similarity

UPV/EHU

Eneko Agirre, Iñigo Lopez-Gazpio, Aitor Gonzalez-Airre,
Montse Maritxalar, German Rigau, Larraitz Uria

inigo.lopez@ehu.eus

IXA NLP Group, University of the Basque Country (UPV/EHU)

Motivation

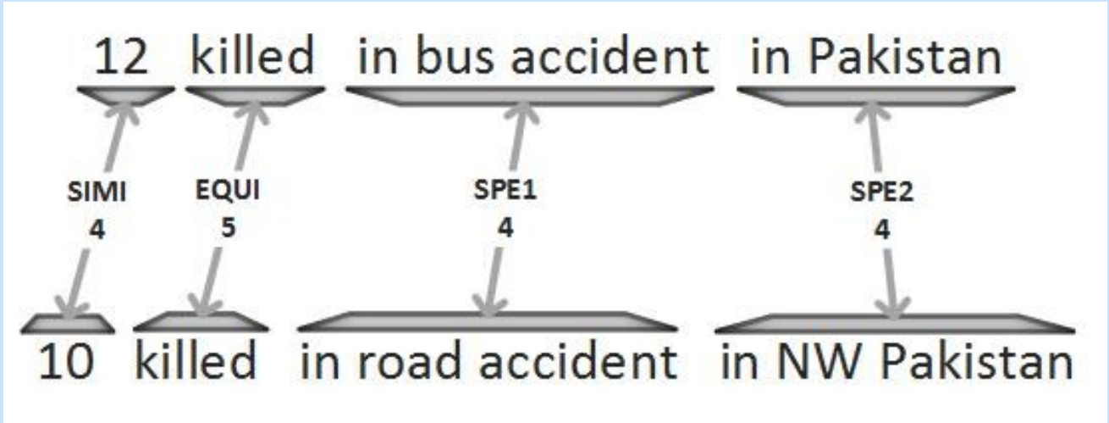
Semantic Textual Similarity (STS)

- Measures the **semantic equivalence** between a pair of sentences, using a graded similarity scale
- The scale ranges from 0 (**complete unrelatedness**) to some maximum value (**semantic equivalence**) and captures the notion that some sentences are more **similar** than others
- STS is useful in many NLP areas, but **to explain in detail** why a sentence pair is related / unrelated and to identify their differences and commonalities, we need to supplement the STS score with **an explanatory layer on top of STS**

Interpretable Semantic Textual Similarity (iSTS)

- iSTS forks from STS and **formalizes the explanatory layer** used to describe the relations across the sentence pair

12 killed in bus accident in Pakistan
10 killed in road accident in NW Pakistan



- iSTS provides **fine-grained information** that is of valuable use for NLU systems, in order to increase their ability **to explain reasoning**
- The **long term goal** consists on using this information to **produce valuable feedback in the educational domain**

They are quite similar, but...

Note that 'in Pakistan' is a bit more general than 'in NW Pakistan' in this context

Note also that 'in bus accident' is a bit more specific than 'in road accident' in this context

Note also that '12' and '10' are very similar in this context

Task Definition

The **overall outline** is as follows:

- Given input **paired sentences**
- **Identify chunks** or use gold standard chunks
- **Align** chunks across sentence pairs
- Indicate a **similarity score** for each aligned pair
- Indicate a **relation label** for each aligned pair

- The **similarity score** is a value bounded by [0 5]
- The **relation label** can be one of the following:

EQUI: The chunks are semantically equivalent (similarity score must be 5)

OPPO: The chunks are in opposition to each other

SPE1: The first chunk is more specific than the second chunk

SPE2: The second chunk is more specific than the first chunk

SIMI: Both chunks have similar meanings, but there is no EQUI, OPPO, SPE1 or SPE2 relation

REL: Both chunks are related, but there is no SIMI relation

NOALI: The chunk has no corresponding segment in the pair (similarity score must be 0)

In addition, there are two extra qualifiers to state **factuality (FACT)** or **polarity (POL)** nuances

There are **two possible evaluation scenarios**:

- **Syschunks**: Participants need to identify chunks across input sentences
- **Goldchunks**: Participants are provided with gold chunk marks

Example

Syschunk scenario:
12 killed in bus accident in Pakistan
10 killed in road accident in NW Pakistan

Goldchunk scenario:
[12] [killed] [in bus accident] [in Pakistan]
[10] [killed] [in road accident] [in NW Pakistan]

Alignment
[12] <=> [10] : SIMI 4
[killed] <=> [killed] : EQUI 5
[in bus accident] <=> [in road accident] : SPE1 4
[in Pakistan] <=> [in NW Pakistan] : SPE2 4

Official evaluation metrics

There are four official evaluation metrics, based on MT alignment

F: F score ignoring relation types and similarity scores. Just **the alignment**.

+T: F plus **type penalty**. **Types** between alignments **need to match**.

+S: F plus **score penalty**. **Scores** between alignments **need to match**.

+TS: F plus **type and score penalty**.

Datasets

The iSTS task released **3 datasets divided in train and test splits**

News Headlines (H)

News headlines mined from news sources by the European Media Monitor

Image Captions (I)

Image captions mined from a subset of the FLICKR dataset

Answer-Students (AS)

Student answers mined from interactions of students with the BEETLE II tutorial dialog system

All the datasets have been **re-annotated** from previous STS tasks

Datasets were annotated by **one expert annotator**, but employed a second annotator to calculate **ITA: 0.73 (H), 0.75 (I) and 0.70 (AS)**

dataset	pairs	source	STS
HDL train	750	news headlines	2013
HDL test	375	news headlines	2014
Images train	750	image captions	2014
Images test	375	image captions	2015
Student train	333	student answers	2015
Student test	344	student answers	2015

Overall Test Results for Type and Score

+TS Syschunks					
System	**I**	**H**	**AS**	**Mean**	**R**
Baseline	.404	.438	.443	.428	
DTSim_r3	.610	.545	.503	.552	1
DTSim_r2	.599	.547	.507	.551	2
DTSim_r1	.587	.538	.505	.543	3
FBK-HLT-NLP_r1	.548	.510	.542	.533	4
FBK-HLT-NLP_r3	.535	.505	.555	.532	5
FBK-HLT-NLP_r2	.497	.503	.541	.513	6
Inspire_r1	.563	.520	.452	.512	7
IISCNLP_r2	.487	.492	.520	.500	8
IISCNLP_r3	.474	.469	.545	.496	9
IISCNLP_r1	.474	.469	.540	.494	10
Inspire_r2	.536	.495	.419	.483	11
Inspire_r3	.450	.446	.338	.411	12
Venseseval_r1	.462	.453	-	-	
∗iUBC_r2	.550	.476	.559	.528	
∗iUBC_r3	.516	.498	.559	.524	
∗iUBC_r1	.477	.423	.449	.450	
AVG	.525	.499	.497	.510	
MAX	.610	.547	.555	.552	

+TS Goldchunks					
System	**I**	**H**	**AS**	**Mean**	**R**
Baseline	.480	.546	.557	.528	
UWB_r1	.667	.621	.625	.638	1
UWB_r3	.671	.630	.611	.637	2
DTSim_r2	.636	.649	.546	.610	3
DTSim_r3	.648	.641	.537	.609	4
Inspire_r1	.613	.696	.510	.606	5
DTSim_r1	.624	.639	.543	.602	6
Inspire_r2	.588	.663	.479	.576	7
VRep_r3	.547	.597	.580	.575	8
VRep_r2	.543	.597	.579	.573	9
FBK-HLT-NLP_r3	.566	.562	.589	.572	10
FBK-HLT-NLP_r1	.574	.559	.581	.571	11
UWB_r2	.621	.601	.475	.566	12
IISCNLP_r2	.509	.556	.617	.560	13
IISCNLP_r1	.485	.551	.639	.558	14
IISCNLP_r3	.492	.541	.639	.557	15
VRep_r1	.548	.596	.523	.556	16
FBK-HLT-NLP_r2	.525	.555	.571	.551	17
Rev_r1	.493	.562	.410	.489	18
Inspire_r3	.487	.579	.386	.484	19
Venseseval_r1	.574	.573	-	-	
∗iUBC_r2	.612	.587	.644	.614	
∗iUBC_r3	.578	.592	.644	.604	
∗iUBC_r1	.513	.505	.499	.506	
AVG	.570	.598	.549	.573	
MAX	.671	.696	.639	.638	

Specific Test Results for each dataset

Headlines Syschunks					
System	**F**	**+T**	**+S**	**+TS**	**R**
Baseline	.649	.438	.591	.438	
DTSim_r2	.837	.561	.760	.547	1
DTSim_r3	.838	.560	.759	.545	2
DTSim_r1	.837	.561	.739	.538	3
Inspire_r1	.704	.526	.659	.520	4
FBK-HLT-NLP_r1	.808	.523	.737	.510	5
FBK-HLT-NLP_r3	.805	.519	.737	.505	6
FBK-HLT-NLP_r2	.797	.514	.731	.503	7
Inspire_r2	.759	.503	.691	.495	8
IISCNLP_r2	.821	.508	.740	.492	9
IISCNLP_r1	.811	.489	.723	.469	10
IISCNLP_r3	.811	.494	.721	.469	11
Venseseval_r1	.708	.468	.649	.453	12
Inspire_r3	.769	.455	.687	.446	13
∗iUBC_r3	.809	.507	.739	.498	
∗iUBC_r2	.809	.486	.738	.476	
∗iUBC_r1	.809	.431	.714	.423	
AVG	.793	.514	.718	.499	
MAX	.838	.561	.760	.547	

Headlines Goldchunks					
System	**F**	**+T**	**+S**	**+TS**	**R**
Baseline	.846	.546	.761	.546	
Inspire_r1	.819	.703	.787	.696	1
Inspire_r2	.892	.673	.832	.663	2
DTSim_r2	.907	.665	.836	.649	3
DTSim_r3	.907	.658	.833	.641	4
DTSim_r1	.907	.665	.819	.639	5
UWB_r3	.899	.641	.838	.630	6
UWB_r1	.898	.632	.835	.621	7
UWB_r2	.890	.615	.815	.601	8
VRep_r3	.893	.602	.805	.597	9
VRep_r2	.901	.603	.808	.597	10
VRep_r1	.891	.602	.803	.596	11
Inspire_r3	.897	.589	.818	.579	12
Venseseval_r1	.873	.593	.810	.573	13
Rev_r1	.866	.571	.784	.562	14
FBK-HLT-NLP_r3	.885	.577	.809	.562	15
FBK-HLT-NLP_r1	.879	.574	.810	.559	16
IISCNLP_r2	.913	.576	.829	.556	17
FBK-HLT-NLP_r2	.886	.564	.802	.555	18
IISCNLP_r1	.914	.573	.820	.551	19
IISCNLP_r3	.914	.567	.821	.541	20
∗iUBC_r3	.928	.602	.858	.592	
∗iUBC_r2	.928	.600	.861	.587	
∗iUBC_r1	.928	.512	.830	.505	
AVG	.892	.612	.816	.598	
MAX	.914	.703	.838	.696	

Images Syschunks					
System	**F**	**+T**	**+S**	**+TS**	**R**
Baseline	.713	.404	.625	.404	
DTSim_r3	.843	.628	.781	.610	1
DTSim_r2	.843	.615	.781	.599	2
DTSim_r1	.843	.615	.759	.587	3
Inspire_r1	.754	.564	.704	.563	4
FBK-HLT-NLP_r1	.843	.566	.786	.548	5
Inspire_r2	.817	.543	.742	.536	6
FBK-HLT-NLP_r3	.842	.554	.785	.535	7
FBK-HLT-NLP_r2	.843	.518	.781	.497	8
IISCNLP_r2	.846	.499	.777	.487	9
IISCNLP_r1	.834	.486	.765	.474	10
IISCNLP_r3	.834	.486	.765	.474	11
Venseseval_r1	.743	467	.695	.463	12
Inspire_r3	.811	.453	.735	.450	13
∗iUBC_r2	.856	.561	.796	.550	
∗iUBC_r3	.856	.523	.794	.516	
∗iUBC_r1	.856	.489	.770	.477	
AVG	.822	.538	.758	.525	
MAX	.846	.628	.786	.610	

Images Goldchunks					
System	**F**	**+T**	**+S**	**+TS**	**R**
Baseline	.856	.480	.746	.480	
UWB_r3	.892	.687	.841	.671	1
UWB_r1	.894	.683	.840	.667	2
DTSim_r3	.877	.668	.816	.648	3
DTSim_r2	.877	.653	.814	.636	4
DTSim_r1	.877	.653	.796	.624	5
UWB_r2	.871	.635	.808	.621	6
Inspire_r1	.797	.614	.748	.613	7
Inspire_r2	.867	.596	.795	.588	8
FBK-HLT-NLP_r1	.873	.595	.815	.574	9
Venseseval_r1	.844	.579	.805	.574	10
FBK-HLT-NLP_r3	.879	.588	.819	.566	11
VRep_r1	.854	.552	.765	.548	12
VRep_r3	.855	.551	.765	.547	13
VRep_r2	.857	.547	.763	.543	14
FBK-HLT-NLP_r2	.879	.543	.818	.525	15
IISCNLP_r2	.893	.525	.823	.509	16
Rev_r1	.831	.501	.740	.493	17
IISCNLP_r3	.893	.505	.826	.492	18
Inspire_r3	.855	.489	.781	.487	19
IISCNLP_r1	.893	.502	.829	.485	20
∗iUBC_r2	.908	.622	.855	.612	
∗iUBC_r3	.908	.587	.846	.578	
∗iUBC_r1	.908	.520	.816	.513	
AVG	.868	.583	.800	.570	
MAX	.894	.687	.841	.671	

Answer-Students Syschunks					
System	**F**	**+T**	**+S**	**+TS**	**R**
Baseline	.619	.443	.5702	.443	
FBK-HLT-NLP_r3	.817	.561	.757	.555	1
IISCNLP_run3	.756	.560	.710	.545	2
FBK-HLT-NLP_r1	.816	.548	.759	.542	3
FBK-HLT-NLP_r2	.816	.543	.748	.541	4
IISCNLP_r1	.756	.553	.710	.540	5
IISCNLP_r2	.745	.532	.700	.520	6
DTSim_r2	.817	.516	.737	.507	7
DTSim_r1	.817	.516	.725	.505	8
DTSim_r3	.818	.511	.736	.503	9
Inspire_r1	.690	.455	.640	.452	10
Inspire_r2	.725	.424	.653	.419	11
Inspire_r3	.762	.343	.670	.338	12
∗iUBC_r2	.796	.565	.748	.559	
∗iUBC_r3	.796	.565	.748	.559	
∗iUBC_r1	.796	.450	.710	.449	
AVG	.778	.505	.712	.497	
MAX	.818	.561	.759	.555	

Answer-Students Goldchunks					
System	**F**	**+T**	**+S**	**+TS**	**R**
Baseline	.820	.557	.746	.557	
IISCNLP_r1	.868	.651	.825	.639	1
IISCNLP_r3	.868	.651	.825	.639	2
UWB_r1	.864	.630	.809	.625	3
IISCNLP_r2	.868	.627	.826	.617	4
UWB_r3	.859	.617	.804	.611	5
FBK-HLT-NLP_r3	.851	.598	.790	.589	6
FBK-HLT-NLP_r1	.878	.589	.810	.581	7
VRep_r3	.879	.582	.792	.580	8
VRep_r2	.870	.581	.785	.579	9
FBK-HLT-NLP_r2	.860	.576	.791	.571	10
DTSim_r2	.858	.555	.781	.546	11
DTSim_r1	.858	.555	.769	.543	12
DTSim_r3	.861	.547	.780	.537	13
VRep_r1	.772	.525	.701	.523	14
Inspire_r1	.795	.513	.735	.510	15
Inspire_r2	.821	.483	.744	.479	16
UWB_r2	.875	.481	.783	.475	17
Rev_r1	.846	.418	.727	.410	18
Inspire_r3	.874	.391	.770	.386	19
∗iUBC_r2	.892	.651	.843	.644	
∗iUBC_r3	.892	.651	.843	.644	
∗iUBC_r1	.892	.502	.794	.499	
AVG	.854	.556	.781	.549	
MAX	.879	.651	.826	.639	

Participation

6 teams participated in the **syschunk** scenario, totaling 16 runs

9 teams participated in the **goldchuks** scenario, totaling 23 runs

Participant mean AVG well-above baseline

Syschunks +TS wrt baseline: **+.08**

Goldchunks +TS wrt baseline: **+.05**

Best Runs

Mean +TS Syschunks: .552 **DTSim_run3**
Mean +TS Goldchunks: .638 **UWB_run1**

Conclusions

- The present edition of iSTS has **consolidated** the previous year pilot
- **Well-defined** and **feasible** task with guidelines that allow for **high inter-annotator correlations**
  - The annotation guidelines are released publicly in the task website
- **Substantial amount** of sentence pairs from diverse domains
- Top new systems continue to **improve state-of-the-art**
- **High participation with diverse approaches** and good performance
  - Supervised systems, unsupervised systems and rule-based systems

http://alt.qcri.org/semeval2016/task2/

isis-semeval@googlegroups.com