SemEval-2016 Task 2: Interpretable Semantic Textual Similarity

Eneko Agirre*, **Aitor Gonzalez-Agirre**, **Iñigo Lopez-Gazpio**
Montse Maritxalar, **German Rigau** and **Larraitz Uria**
IXA NLP group, University of the Basque Country
Manuel Lardizabal 1, 20.018 Donostia, Basque Country
montse.maritxalar@ehu.eus

Abstract

The final goal of Interpretable Semantic Textual Similarity (iSTS) is to build systems that explain which are the differences and commonalities between two sentences. The task adds an explanatory level on top of STS, formalized as an alignment between the chunks in the two input sentences, indicating the relation and similarity score of each alignment. The task provides train and test data on three datasets: news headlines, image captions and student answers. It attracted nine teams, totaling 20 runs. All datasets and the annotation guideline are freely available[1]

1 Introduction

Semantic Textual Similarity (STS) (Agirre et al., 2015) measures the degree of equivalence in the underlying semantics of paired snippets of text. The idea of Interpretable STS (iSTS) is to explain *why* two sentences may be related/unrelated, by supplementing the STS similarity score with an explanatory layer.

Our final goal would be to enable interpretable systems, that is, systems that are able to explain which are the differences and commonalities between two sentences. For instance, let's assume the following two sentences drawn from a corpus of news headlines:

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

The output of such a system would be something like the following:

> The two sentences talk about accidents with casualties in Pakistan, but they differ in the number of people killed (12 vs. 10) and level of detail: the first one specifies that it is a *bus* accident, and the second one specifies that the location is *NW* Pakistan.

While giving such explanations comes naturally to people, constructing algorithms and computational models that mimic human level performance represents a difficult Natural Language Understanding (NLU) problem, with applications in dialogue systems, interactive systems and educational systems.

In the iSTS 2015 pilot task (Agirre et al., 2015), we defined a first step of such an ambitious system, which we follow in 2016. Given the input (a pair of sentences), participant systems need first to identify the chunks in each sentence, and then, align chunks across the two sentences, indicating the relation and similarity score of each alignment. The relation can be one of equivalence, opposition, specificity, similarity or relatedness, and the similarity score can range from 1 to 5. Unrelated chunks are left unaligned. An optional tag can be added to alignments for the cases where there is a difference in factuality or polarity. See Figure 1 for the manual alignment of the two sample sentences. The alignments between chunks in Figure 1 can be used to produce the kind of explanations shown in the previous example.

In previous work, Brockett (2007) and Rus et al. (2012) produced a dataset where corresponding

* Authors listed in alphabetical order

[1] http://at.qrci.org/semeval2016/task2/

Proceedings of SemEval-2016, pages 512–524,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

```
[12] <=> [10] : (SIMILAR 4)
[killed] <=> [killed] : (EQUIVALENT 5)
[in bus accident] <=> [in road accident] : (MORE-SPECIFIC 4)
[in Pakistan] <=> [in NW Pakistan] : (MORE-GENERAL 4)
```

Figure 1: Example of a manual alignment of two sentences: “12 killed in bus accident in Pakistan” and “10 killed in road accident in NW Pakistan”. Each aligned pair of chunks included information on the type of alignment, and the score of alignment.

words (including some multiword expressions like named-entities) were aligned. Although this alignment is useful, we wanted to move forward to the alignment of segments, and decided to align chunks (Abney, 1991). Brockett (2007) did not provide any label to alignments, while Rus et al. (2012) defined a basic typology. In our task, we provided a more detailed typology for the aligned chunks as well as a similarity/relatedness score for each alignment. Contrary to the mentioned works, we first identified the segments (chunks in our case) in each sentence separately, and then aligned them.

In a different strand of work, Nielsen et al. (2009) defined a textual entailment model where the “facets” (words under some syntactic/semantic relation) in the response of a student were linked to the concepts in the reference answer. The link would signal whether each facet in the response was entailed by the reference answer or not, but would not explicitly mark which parts of the reference answer caused the entailment. This model was later followed by Levy et al. (2013). Our task was different in that we identified the corresponding chunks in both sentences. We think that, in the future, the aligned facets could provide complementary information to chunks.

The SemEval Semantic Textual Similarity (STS) task in 2015 contained a subtask on Interpretable STS (Agirre et al., 2015), showing that the task is feasible, with high inter-annotator agreement and system scores well above baselines. The datasets comprised news headlines and image captions.

For 2016, the pilot subtask has been updated into a standalone task. The restriction from the iSTS 2015 task to allow only one-to-one alignments has been now lifted, and we thus allow any number of chunks to be aligned to any number of chunks. Annotation guidelines have been revised accordingly, including an updated chunking criterium for subordinate clauses and a better explanation of the instructions.

The 2015 datasets were re-annotated and released as training data. New pairs from news headlines and image captions have been annotated and used for test. In addition, a new dataset of sentence pairs from the education domain has been produced, including train and test data.

The paper is organized as follows. We first provide the description of the task, followed by the evaluation metrics and the baseline system. Section 5 describes the participation, Section 6 the results, and Section 7 comments on the systems, tools and resources used.

2 Task Description

The dataset was produced using sentence pairs from news headlines, image captions and answers from students. Headlines have been mined from several news sources by European Media Monitor, and collected by us using their RSS feed[2]. We saw a pair of headlines from this corpus in the introduction.

The Image descriptions dataset is a subset of the Flickr dataset presented in (Rashtchian et al., 2010), which consisted of 8108 hand-selected images from Flickr, depicting actions and events of people or animals, with five captions per image. The image captions of the dataset are released under a Creative Commons Attribution-Share Alike license. This is a sample pair from this dataset:

A man sleeps with a baby in his lap
A man asleep in a chair holding a baby

The Answer-Students corpus consists of the interactions between students and the BEETLE II tutorial dialogue system. The BEETLE II system is an intelligent tutoring engine that teaches students in basic electricity and electronics. At first, students

[2] http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html

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

Table 1: Details of the datasets, including number of pairs, source, and relation to STS datasets. HDL stands for Headlines, and Student to Student-Answers.

spend from three to five hours reading the material, building and observing circuits in the simulator and interacting with a dialogue-based tutor. They used the keyboard to interact with the system, and the computer tutor asked them questions and provided feedback via a text-based chat interface. The data from 73 undergraduate volunteer participants at south-eastern US university were recorded and annotated to form the BEETLE human-computer dialogue corpus (Dzikovska et al., 2010; Dzikovska et al., 2012), and later used in a SemEval 2015 task (Dzikovska et al., 2013). In the present corpus, we include sentence pairs composed of a student answer and the reference answer of a teacher. We have rejected those answers containing pronouns whose antecedent is not in the sentence (pronominal coreference), as the question is not included in the train data and, therefore, it is not possible to deduce which is the antecedent. There are also some dataset-specific details that are mentioned in the same section. The next pair sentences are an example of the Answer-Students corpus.

> because switch z is in bulb c's closed path
> there is a path containing both Z and C

All datasets have been previously used in STS tasks. Table 1 shows details of the datasets, including train-test splits. The Headlines and Images datasets are tokenized, as in the STS release. The Answer-Students dataset was not tokenized, and was used as in the STS release.

2.1 Annotation

The manual annotation has been performed following the annotation guidelines [3]. Please refer to those guidelines for further details. The general annotation procedure is as follows:

1. First identify the chunks in each sentence separately.
2. Align chunks in order, from the clearest and strongest correspondences to the most unclear or weakest ones.
3. For each alignment, provide a similarity/relatedness score.
4. For each alignment, choose one (or more) alignment label.

Chunk annotation was based on those used in the CoNLL 2000 chunking task (Tjong Kim Sang and Buchholz, 2000). The annotators were provided with the output of an automatic chunker[4] trained on the CoNLL corpora[5], which they corrected manually.

Independently of the labels, and before assigning any label, the annotators need to provide a similarity/relatedness score for each alignment from 5 (maximum similarity/relatedness) to 0 (no relation at all), as follows:

5 if the meaning of both chunks is equivalent

[4,3] if the meaning of both chunks is very similar or closely related

[2,1] if the meaning of both chunks is slightly similar or somehow related

0 (represented as NIL) if the meaning of the chunk is completely unrelated.

Note that 0 is not possible for an aligned pair, as that would mean that the two chunks would be left unaligned. Note also that if the score is 5, then the label assigned later should be equivalence (EQUI, see below). After assigning the label, the annotator should check for the following: if a chunk is not aligned it should have NIL score, equivalent chunks

[3] http://alt.qcri.org/semeval2016/task2/data/uploads/annotationguidelinesinterpretablests2016v2.2.pdf

[4] https://github.com/ixa-ehu/ixa-pipe-chunk

[5] http://www.clips.ua.ac.be/conll2000/chunking/

(EQUI) should have a 5 score. The rest of the labels should have a score larger than 0 but lower than 5.

We will now describe the alignment types, but first note that the interpretation of the whole sentence, including common sense inference, has to be taken into account. This means that we need to take into account the context in order to know whether the aligned chunks refer to the same instance (or set of instances) or not. Instances may refer to physical or abstract object instances (for NPs) or real world event instances (for verb chains):

- EQUI: both chunks have the same meaning, they are semantically equivalent in this context.
- OPPO: the meanings of the chunks are in opposition to each other, lying in an inherently incompatible binary relationship.
- SPE1: both chunks have similar meanings, but chunk in sentence 1 is more specific.
- SPE2: like SPE1, but it is the chunk in sentence 2 which is more specific.

In addition, the meaning of the chunks can be very close, either because they have a similar meaning, or because their meanings have some other relation. In those cases, we use SIMI or REL as follows:

- SIMI: both chunks have similar meanings, they share similar attributes and there is no EQUI, OPPO, SPE1 or SPE2 relation.
- REL: both chunks are not considered similar but they are closely related by some relation not mentioned above (i.e. no EQUI, OPPO, SPE1, SPE2, or SIMI relation).
- NOALI: this chunk has not any corresponding chunk in the other sentence. Therefore, it is left unaligned.

The above seven labels are exclusive, and each alignment should have one such label.

In addition to one of the labels above, there are two labels which can be used either in isolation or together, that is, you can use none, one or both:

- FACT: the factuality in the aligned chunks (i.e. whether the statement is or is not a fact or a speculation) is different.
- POL: the polarity in the aligned chunks (i.e. the expressed opinion, which can be positive, negative, or neutral) is different.

Note that NOALI can also be FACT or POL, meaning that the respective chunk adds a factuality or polarity nuance to the sentence.

Listing 1 shows the annotation format for a given sentence pair from the training set (note that each alignment is reported in one line as follows: token-id-sent1 <==> token-id-sent2 // label // score // comment).

Finally, there are some specific criteria related to the Answer-Students corpus that have been followed during the annotation process. For instance, in the Answer-Students example in the previous section, *switch z* (first sentence) and *Z* (second sentence) are considered equivalent as, in this dataset, X, Y, and Z always refer to switches X, Y, and Z. The same criteria is followed when annotating *bulb c* and *C* as equivalent, as A, B and C are always used to refer to bulb A, B and C. In the same way *closed path* and *a path* are equivalent, as paths are always considered to be closed. For further details related to such a corpus specific criteria refer to the annotation guidelines.

3 Evaluation Metrics

The official evaluation is based on (Melamed, 1998), which uses the F1 of precision and recall of token alignments (in the context of alignment for Machine Translation). Fraser and Marcu (2007) argue that F1 is a better measure than other alternatives such as the Alignment Error Rate. The idea is that, for each pair of chunks that are aligned, we consider that any pairs of tokens in the chunks are also aligned with some weight. The weight of each token-token alignment is the inverse of the number of alignments of each token (so-called fan out factor, Melamed, 1998). Precision is measured as the ratio of token-token alignments that exist in both system and gold standard files, divided by the number of alignments in the system. Recall is measured similarly, as the ratio of token-token alignments that exist in both system and gold-standard, divided by the number of alignments in the gold standard. Precision and recall are evaluated separately for all alignments of all pairs.

Participating runs were evaluated using four different metrics: F1 where alignment type and score are ignored (alignment F1, F for short); F1 where alignment types need to match, but scores are ignored (type F1, +T for short); F1 where alignment type is ignored, but each alignment is penalized

Listing 1: Annotation format

```
<sentence id=”6” status=””>
 12 killed in bus accident in Pakistan
 10 killed in road accident in NW Pakistan
 ...
 <alignment>
  1 <==> 1 // SIMI // 4 // 12 <==> 10
  2 <==> 2 // EQUI // 5 // killed <==> killed
  3 4 5 <==> 3 4 5 // SPE1 // 4 // in bus accident <==> in road accident
  6 7 <==> 6 7 8 // SPE2 // 4 // in Pakistan <==> in NW Pakistan
 </alignment>
</sentence>
```

when scores do not match[6] (score F1, +S for short); and, F1 where alignment types need to match, and each alignment is penalized when scores do not match (type and score F1, +TS for short). The type and score F1 is the main overall metric.

Note that our evaluation procedure does not explicitly evaluate the chunking results. The method implicitly penalizes chunking errors via the induced token-token alignments, using a soft penalty.

4 Baseline System

The baseline system consists of a cascade concatenation of several procedures. First, input sentences are tokenized using simple regular expressions. Additionally, we collect chunks coming either from the gold standard or from the chunking done by *ixa-pipes-chunk* (Agerri et al., 2014). This is followed by a lower-cased token aligning phase, which consists of aligning (or linking) identical tokens across the input sentences. Then we use chunk boundaries as token regions to group individual tokens into groups, and compute all links across groups. The weight of the link across groups is proportional to the number of links counted between within-group tokens. The next phase consists of an optimization step in which groups x,y that have the highest link weight are identified, as well as the chunks that are linked to either x or y but not with a maximum alignment weight (thus enabling us to know which chunks were left unaligned). Finally, in the last phase, the baseline system uses a rule-based algorithm to directly assign labels and scores: to chunks with the highest link weight assign label = “EQUI” and score = 5, to the rest of aligned chunks (with lower weights) assign label = “NOALI” and score = NIL, and, to unaligned chunks assign label = “NOALI” and score = NIL.

5 Participation

The pilot task presented two scenarios: raw text and gold standard chunks. In the first scenario, given a pair of sentences, participants had to identify the composing chunks, and then align them; after that they would assign a relatedness tag and a similarity score to each alignment. In the gold standard scenario, participants were provided with the gold standard chunks.

In both scenarios the datasets were provided with tokenized text, with exception of Answer-Students, which was not tokenized[7].

The task allowed up to a total of three submissions for each team on each of the evaluation scenarios. The organizers provided a script to check if the run files are well formed.

Nine teams participated on the gold chunks scenario, and out of them six teams also participated in the system chunks scenario. Regarding the datasets, all the teams gave their results for the three datasets,

[6] The penalization is the difference between the scores divided by five.

[7] In fact The Answer-Students dataset was only partially tokenized. In order to be consistent with the gold standard, participants had to follow the partial tokenization, separating tokens at blanks alone.

except *Venseseval* who sent results only for Headlines and Images.

The iUBC team includes some of the organizers of the interpretable STS task. It is marked by the symbol $*$ in the result tables, and it is not taken into account in the rankings. The organizers took measures to prevent developers of that team to access the test data or any other information, so the team participated in identical conditions to the rest of participants.

6 Results

Table 2 provides the overall type and score (+TS) performance per dataset, and the mean accross the three datasets. Results for Headlines, Images and Answer-Students datasets are shown in the Appendix, tables 3, 4 and 5, respectively. Each row of the tables corresponds to a run configuration named *TeamID_RunID*. Note that task results are separately written with respect to the scenario. A unique baseline was used for both evaluation scenarios and its performance is jointly presented with the scores obtained by participants.

The results of the present edition corroborate last years' results regarding the difficulty of the system chunks scenario. Indeed, it is considerably more challenging than the gold chunks scenario.

With regard to the datasets, the Answer-Students ended up being more challenging than the other datasets for five out of eight teams, but FBK-HLT-NLP, IISCNLP and iUBC teams give their best results for such a scenario.

Compared to last year, the best results for Images and Headlines in the +TS metric have improved in both SYS and GS scenarios: 4 and 6 points for Headlines (in SYS and GS, respectively), and 5 and 7 points for Images (in SYS and GS, respectively). In order to check whether the datasets where easier this year, we checked the performance of the baseline. The differences are small: this year the Images dataset seems slightly easier (3 and 4 point difference for SYS and GS scenarios), and the Headlines dataset is only slightly more difficult (1 point difference for SYS and GS scenarios). The improvement in results for this year seems to be due to better system performance.

The complexity of the evaluation (cf. tables 3, 4 and 5) was incremental for the four available metrics, which obviously, were lower for the system chunks. Both type and score are bounded by the alignment results and it is thus natural that alignment results are higher. Comparing type and score results, the type results are generally lower, possibly due to the harder task of guessing the correct label. The final results are bounded by both type and score, and the systems doing best in type are the ones doing best overall. From the results we can see that labeling the type was the most challenging.

Regarding the overall test results for type and score (+TS) across datasets, UWB (Konopík et al., 2016) and DTSim (Banjade et al., 2016) obtained the best results for the gold chunks scenario, and DTSim and FBK-HLT-NLP (Magnolini et al., 2016) for the system chunks scenario. In addition, DTSim obtained the best overall results even though they have not good results for the Answer-Students dataset.

7 Systems, tools and resources

Most of the teams reported input text processing such as lemmatization and part of speech tagging, and in some cases named-entity recognition and syntactic parsing. Additional resources such as WordNet, distributional embeddings, paraphrases from PPDB and global STS sentence scores were also used. Participants also revealed that most of their systems were built using some kind of distributional or knowledge-based similarity metrics. We noticed, for instance, that WordNet or word embeddings were used by several teams to compute word similarity.

Looking at the learning approaches, both supervised and unsupervised approaches have been applied, as well are mainly manual rule-based combinations.

Next, we briefly introduce the participant teams, whit slightly more details for the top performing systems.

- UWB (Konopík et al., 2016): UWB used three separate supervised classifiers to perform alignment, scoring and typing. They defined a similarity function based on a distribution similarity paradigm: vector composition, lexical semantic vectors and iDF weighting. They introduced a modified method to create word vectors, and

+TS Syschunks						+TS Goldchunks					
System	**I**	**H**	**AS**	**Mean**	**R**	**System**	**I**	**H**	**AS**	**Mean**	**R**
Baseline	.404	.438	.443	.428		Baseline	.480	.546	.557	.528	
DTSim_r3	.610	.545	.503	.552	1	UWB_r1	.667	.621	.625	.638	1
DTSim_r2	.599	.547	.507	.551	2	UWB_r3	.671	.630	.611	.637	2
DTSim_r1	.587	.538	.505	.543	3	DTSim_r2	.636	.649	.546	.610	3
FBK-HLT-NLP_r1	.548	.510	.542	.533	4	DTSim_r3	.648	.641	.537	.609	4
FBK-HLT-NLP_r3	.535	.505	.555	.532	5	Inspire_r1	.613	.696	.510	.606	5
FBK-HLT-NLP_r2	.497	.503	.541	.513	6	DTSim_r1	.624	.639	.543	.602	6
Inspire_r1	.563	.520	.452	.512	7	Inspire_r2	.588	.663	.479	.576	7
IISCNLP_r2	.487	.492	.520	.500	8	VRep_r3	.547	.597	.580	.575	8
IISCNLP_r3	.474	.469	.545	.496	9	VRep_r2	.543	.597	.579	.573	9
IISCNLP_r1	.474	.469	.540	.494	10	FBK-HLT-NLP_r3	.566	.562	.589	.572	10
Inspire_r2	.536	.495	.419	.483	11	FBK-HLT-NLP_r1	.574	.559	.581	.571	11
Inspire_r3	.450	.446	.338	.411	12	UWB_r2	.621	.601	.475	.566	12
Venseseval_r1	.462	.453	-	-		IISCNLP_r2	.509	.556	.617	.560	13
∗iUBC_r2	.550	.476	.559	.528		IISCNLP_r1	.485	.551	.639	.558	14
∗iUBC_r3	.516	.498	.559	.524		IISCNLP_r3	.492	.541	.639	.557	15
∗iUBC_r1	.477	.423	.449	.450		VRep_r1	.548	.596	.523	.556	16
AVG	.525	.499	.497	.510		FBK-HLT-NLP_r2	.525	.555	.571	.551	17
MAX	.610	.547	.555	.552		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	

Table 2: Overall test results for type and score (**+TS**) across datasets. Each row correspond to a system run, and each column to a dataset: **(I)** for Images, **(H)** for Headlines, **(AS)** for Answer-Students, **Mean** for the mean across the three datasets, and **R** for the rank. The "∗" symbol denotes runs that include task organizers. Additionally, the table shows results for the baseline, average of participants (**AVG**) and maximum score of participants (**MAX**).

combine unique words from the chunks of both sentences into one single vocabulary which is then used to produce similarity measures. They claim that the following three differences have significant influence on the final results: modified lexical semantic vectors (+3% of the mean of T+S F1 scores), shared words (+2%) and POS tags difference (+2%).

- DTSim (Banjade et al., 2016): This team builds on the NeroSim system (Banjade et al., 2015), which participated in the 2015 task with good results using a system based on manual rules blended semantic similarity features. The team explored several chunking algorithms and included new rules. Concretely, they expanded the rules for SIMI and EQUI. They mainly improved the chunker and concluded that a Conditional Random Fields (CRFs) based chunking tool is the best approach for chunking. The input sequence to their chunking model are POS tags, and the chunker yielded the highest average accuracies on both the training and test datasets.
- FBK-HLT-NLP (Magnolini et al., 2016): This teams built a multi-layer perceptron to solve alignment, scoring and typing. The perceptron shares some layers for the three tasks, and other layers are separate. They use a variety of

features, including WordNet and word embeddings. The system performs better in the system chunks scenario than in the gold chunks one. Therefore, there is no specific advantage of using chunked sentence pairs and their system is very powerful. The Answer-Students dataset has better performance than Headlines and Images. They obtain better results training a single system for the three datasets (compared to training a classifier separately for each dataset).

- Inspire (Kazmi and Schüller, 2016): The authors propose a system based on logic programming which extends the basic ideas of NeroSim (Banjade et al., 2015). The rule based system makes use of several resources to prepare the input and uses Answer Set Programming to determine chunk boundaries.
- IISCNLP (Tekumalla and Sharmistha, 2016): The system uses an algorithm, iMATCH, for the alignment of multiple non-contiguous chunks based on Integer Linear Programming (ILP). Similarity type and score assignment for pairs of chunks is done using a supervised multiclass classification technique based on Random Forest Classifier.
- Vrep (Henry and Sands, 2016): features are extracted to create a learned rule-based classifier to assign a label. It uses semantic and syntactic (form of the chunks) relationship features.
- Rev (Ping Ping et al., 2016): The system consists of rules based on the analysis of the Headlines dataset considering lexical overlapping, part of speech tags and synonymy.
- Venseseval: This system is an adaptation of a pre-existing textual entailment system, VENSES, which first performs a semantic analysis of the text including argument structure and then looks for bridging information between chunks using several knowledge resources.
- iUBC (Lopez-Gazpio et al., 2016): A two layer architecture is used to produce the similarity type and score of pairs of chunks. The top layer consists of two models: a classifier and a regressor. The bottom layer consists of a recurrent neural network that processes input and feeds composed semantic feature vectors to the top layer. Both layers are trained at the same time by propagating gradients.

8 Conclusions

Last year, the Interpretable STS task was introduced as a pilot subtask of the STS task. At the present edition, it has been presented as an independent task that has attracted nine teams. In addition to the image caption and news headlines datasets, this year participants were challenged with a new dataset from the Educational area. Concretely, the Answer-Students corpus, which consists of the interactions between students of electronics and the BEETLE II tutorial dialogue system.

Compared to the results last year (Agirre et al., 2015), the results have improved in the two datasets that happened both years, Images and Headlines. The Answer-Students dataset is the most challenging, and among the three subtasks (alignment, type and score) guessing the correct type of the aligned chunks is the most difficult one. Teams that did best on type get the best overall score.

All datasets and the annotation guideline are available in `http://alt.qcri.org/semeval2016/task2/`.

Acknowledgments

This material is based in part upon work supported a MINECO grant to the University of the Basque Country (TUNER project TIN2015-65308-C5-1-R). Aitor Gonzalez Agirre and Iñigo Lopez-Gazpio are by doctoral grants from MINECO. The IXA group is funded by the Basque Government (A type Research Group).

References

Steven Abney. 1991. Parsing by chunks. In *Principle-based parsing: Computation and psycholinguistics. Robert Berwick and Steven Abney and Carol Tenny(eds.)*, pages 257–278.

Rodrigo Agerri, Josu Bermudez, and German Rigau. 2014. Ixa pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 26–31.

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, June. Association for Computational Linguistics.

Rajendra Banjade, Nobal Bikram Niraula, Nabin Maharjan, Vasile Rus, Dan Stefanescu, Mihai Lintean, and Dipesh Gautam. 2015. Nerosim: A system for measuring and interpreting semantic textual similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 164–171, Denver, Colorado, June. Association for Computational Linguistics.

Rajendra Banjade, Nabin Maharjan, Nobal B. Niraula, and Vasile Rus. 2016. Dtsim at semeval-2016 task 2: Interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Chris Brockett. 2007. Aligning the RTE 2006 corpus. *Microsoft Research*.

Myroslava O. Dzikovska, Diana Bental, Johanna D. Moore, Natalie B. Steinhauser, Gwendolyn E. Campbell, Elaine Farrow, and Charles B. Callaway, 2010. *Sustaining TEL: From Innovation to Learning and Practice: 5th European Conference on Technology Enhanced Learning, EC-TEL 2010, Barcelona, Spain, September 28 - October 1, 2010. Proceedings*, chapter Intelligent Tutoring with Natural Language Support in the Beetle II System, pages 620–625. Springer Berlin Heidelberg, Berlin, Heidelberg.

Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2012. Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 200–210, Stroudsburg, PA, USA. Association for Computational Linguistics.

Myroslava O. Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In **SEM 2013: The First Joint Conference on Lexical and Computational Semantics*, Atlanta, Georgia, USA, 13-14 June. Association for Computational Linguistics.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Sam Henry and Allison Sands. 2016. Vrep at semeval-2016 task 1 and task 2. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June.

Mishal Kazmi and Peter Schüller. 2016. Inspire at semeval-2016 task 2: Interpretable semantic textual similarity alignment based on answer set programming. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Miloslav Konopík, Ondřej Pražák, David Steinberger, and Tomáš Brychcín. 2016. Uwb at semeval-2016 task 2: Interpretable semantic textual similarity with distributional semantics for chunks. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Omer Levy, Torsten Zesch, Ido Dagan, and Iryna Gurevych. 2013. Recognizing partial textual entailment. In *ACL (2)*, pages 451–455.

Iñigo Lopez-Gazpio, Eneko Agirre, and Montse Maritxalar. 2016. iubc at semeval-2016 task 2: Rnns and lstms for interpretable sts. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Simone Magnolini, Anna Feltracco, and Bernardo Magnini. 2016. Fbk-hlt-nlp at semeval-2016 task 2: A multitask, deep learning approach for interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

I Dan Melamed. 1998. Manual annotation of translational equivalence: The blinker project. *arXiv preprint cmp-lg/9805005*.

Rodney D. Nielsen, Wayne Ward, and James H. Martin. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(04):479–501.

Tan Ping Ping, Karin Verspoor, and Tim Miller. 2016. Rev at semeval-2016 task 2: Aligning chunks by lexical, part of speech and semantic equivalence. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT 2010, pages 139–147.

Vasile Rus, Mihai Lintean, Cristian Moldovan, William Baggett, Nobal Niraula, and Brent Morgan. 2012. The SIMILAR corpus: A resource to foster the qualitative understanding of semantic similarity of texts. In *Semantic Relations II: Enhancing Resources and Applications, The 8th Language Resources and Evaluation Conference (LREC 2012), May*, pages 23–25.

Lavanya Sita Tekumalla and Sharmistha. 2016. IISC-NLP at semeval-2016 task 2: Interpretable STS with ILP based multiple chunk aligner. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7 (CoNLL 2000)*, pages 127–132.

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	

Table 3: Test results in Headlines for both scenarios. Each row correspond to a system run, and each column to one evaluation metric: F alignment (**F**), F alignment with type penalty (**+T**), F alignment with score penalty (**+S**) and F alignment with type and score penalty (**+TS**), and **R** for the rank. The "∗" symbol denotes runs that include task organizers. Additionally, the table shows results for the baseline, average of participants (**AVG**) and maximum score of participants (**MAX**).

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	

Table 4: Test results in Images for both scenarios. Each row correspond to a system run, and each column to one evaluation metric: F alignment (**F**), F alignment with type penalty (**+T**), F alignment with score penalty (**+S**) and F alignment with type and score penalty (**+TS**), and **R** for the rank. The "∗" symbol denotes runs that include task organizers. Additionally, the table shows results for the baseline, average of participants (**AVG**) and maximum score of participants (**MAX**).

Answer-Students Syschunks						Answer-Students Goldchunks					
System	F	+T	+S	+TS	R	System	F	+T	+S	+TS	R
Baseline	.619	.443	.5702	.443		Baseline	.820	.557	.746	.557	
FBK-HLT-NLP_r3	.817	.561	.757	.555	1	IISCNLP_r1	.868	.651	.825	.639	1
IISCNLP_run3	.756	.560	.710	.545	2	IISCNLP_r3	.868	.651	.825	.639	2
FBK-HLT-NLP_r1	.816	.548	.759	.542	3	UWB_r1	.864	.630	.809	.625	3
FBK-HLT-NLP_r2	.816	.543	.748	.541	4	IISCNLP_r2	.868	.627	.826	.617	4
IISCNLP_r1	.756	.553	.710	.540	5	UWB_r3	.859	.617	.804	.611	5
IISCNLP_r2	.745	.532	.700	.520	6	FBK-HLT-NLP_r3	.851	.598	.790	.589	6
DTSim_r2	.817	.516	.737	.507	7	FBK-HLT-NLP_r1	.878	.589	.810	.581	7
DTSim_r1	.817	.516	.725	.505	8	VRep_r3	.879	.582	.792	.580	8
DTSim_r3	.818	.511	.736	.503	9	VRep_r2	.870	.581	.785	.579	9
Inspire_r1	.690	.455	.640	.452	10	FBK-HLT-NLP_r2	.860	.576	.791	.571	10
Inspire_r2	.725	.424	.653	.419	11	DTSim_r2	.858	.555	.781	.546	11
Inspire_r3	.762	.343	.670	.338	12	DTSim_r1	.858	.555	.769	.543	12
∗iUBC_r2	.796	.565	.748	.559		DTSim_r3	.861	.547	.780	.537	13
∗iUBC_r3	.796	.565	.748	.559		VRep_r1	.772	.525	.701	.523	14
∗iUBC_r1	.796	.450	.710	.449		Inspire_r1	.795	.513	.735	.510	15
AVG	.778	.505	.712	.497		Inspire_r2	.821	.483	.744	.479	16
MAX	.818	.561	.759	.555		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	

Table 5: Test results in Answer-Students for both scenarios. Each row correspond to a system run, and each column to one evaluation metric: F alignment (**F**), F alignment with type penalty (**+T**), F alignment with score penalty (**+S**) and F alignment with type and score penalty (**+TS**), and **R** for the rank. The "∗" symbol denotes runs that include task organizers. Additionally, the table shows results for the baseline, average of participants (**AVG**) and maximum score of participants (**MAX**).