SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability

Eneko Agirre[a]*, Carmen Banea[b]*, Claire Cardie[c], Daniel Cer[d], Mona Diab[e]*, Aitor Gonzalez-Agirre[a], Weiwei Guo[f], Iñigo Lopez-Gazpio[a], Montse Maritxalar[a]*, Rada Mihalcea[b], German Rigau[a], Larraitz Uria[a], Janyce Wiebe[g]

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

[b]University of Michigan
Ann Arbor, MI

[c]Cornell University
Ithaca, NY

[d]Google Inc.
Mountain View, CA

[e]George Washington University
Washington, DC

[f]Columbia University
New York, NY

[g]University of Pittsburgh
Pittsburgh, PA

Abstract

In Semantic Textual Similarity (STS), systems rate the degree of semantic equivalence between two text snippets. This year, the participants were challenged with new datasets in English and Spanish. For the English subtask, we exposed the systems to a diversity of testing scenarios, by preparing additional pairs from headlines and image descriptions, as well as introducing new genres, including answer pairs from a tutorial dialogue system, answer pairs from Q&A websites, and pairs from a committed belief dataset. For the Spanish subtask, additional pairs from news and Wikipedia articles were selected. The annotations for both subtasks leveraged crowdsourcing. The English subtask attracted 29 teams with 74 system runs, and the Spanish subtask engaged 7 teams participating with 16 system runs. In addition, this year we ran a pilot task on Interpretable STS, where the systems needed to add an explanatory layer, that is, they had to align the chunks in the sentence pair, explicitly annotating the kind of relation and the score for the chunk pair. The train and test data were manually annotated by an expert, and included headline and image sentence pairs from previous years. 7 teams participated with 29 runs.

1 Introduction and motivation

Given two snippets of text, Semantic Textual Similarity (STS) captures the notion that some texts are more similar than others, measuring their degree of semantic equivalence. Textual similarity can range from complete unrelatedness to exact semantic equivalence, and a graded similarity score intuitively captures the notion of intermediate shades of similarity, as pairs of text may differ from some minor nuanced aspects of meaning to relatively important semantic differences, to sharing only some details, or to simply unrelated in meaning (cf. Section 2).

One of the goals of the STS task is to create a unified framework for combining several semantic components that otherwise have historically tended to be evaluated independently and without characterization of impact on NLP applications. By providing such a framework, STS allows for an extrinsic evaluation of these modules. Moreover, such an STS framework itself could in turn be evaluated intrinsically and extrinsically as a grey/black box within various NLP applications.

STS is related to both Textual Entailment (TE) and Paraphrasing, but it differs in a number of ways and it is more directly applicable to a number of NLP tasks. STS is different from TE inasmuch as it assumes bidirectional graded equivalence between a pair of textual snippets. In the case of TE the equivalence is directional, e.g. *a car is a vehicle*, but *a vehicle is not necessarily a car*. STS also differs from both TE and Paraphrasing (in as far as both tasks have been defined to date in the literature) in that rather than being a binary yes/no decision (e.g. *a vehicle is not a car*), we define STS to be a graded sim-

*Coordinators: e.agirre@ehu.eus, carmennb@umich.edu, mtdiab@gwu.edu, montse.maritxalar@ehu.eus ilarity notion (e.g. *a vehicle* and *a car* are more similar than *a wave* and *a car*). A quantifiable graded bidirectional notion of textual similarity is useful for a myriad of NLP tasks such as MT evaluation, information extraction, question answering, summarization, etc.

In 2012 we held the first pilot task at SemEval 2012, as part of the *SEM 2012 conference, with great success (Agirre et al., 2012). In addition, we held a DARPA sponsored workshop at Columbia University.[1] In 2013, STS was selected as the official Shared Task of the *SEM 2013 conference, with two subtasks: A core task, which was similar to the 2012 task, and a Pilot task on Typed-similarity between semi-structured records. In 2014, new datasets were used, including new genres and the Spanish subtask (Agirre et al., 2014).

This year we presented three subtasks: the English subtask, the Spanish subtask and the Interpretable pilot subtask. The English subtask comprised pairs from headlines and image descriptions, and it also introduced new genres, including answer pairs from a tutorial dialogue system, answer pairs from Q&A websites, and pairs from a dataset tagged with Committed Belief annotations. The first two datasets had training data of the same domain from previous years, while the other three were new. For the Spanish subtask, additional pairs from news and Wikipedia articles were selected. The annotations for both tasks leveraged crowdsourcing.

Finally, with the Interpretable STS pilot subtask we wanted to start exploring whether participant systems are able to explain WHY two sentences are related/unrelated, adding an explanatory layer to the similarity score. As a first step in this direction, given a pair of sentences $s1$ and $s2$, participating systems had to align the chunks in $s1$ to the chunks in $s2$, describing what kind of relation existed between each pair of chunks, and a score for the similarity/relatedness between the pair of chunks. Being a pilot, our goal was to have the final task ready for 2016.

2 Task Description

In this chapter, the three subtasks are described in turn: the English subtask, the Spanish subtask and the pilot on Interpretable STS.

year	dataset	pairs	source
2012	MSRpar	1500	newswire
2012	MSRvid	1500	videos
2012	OnWN	750	glosses
2012	SMTnews	750	MT eval.
2012	SMTeuroparl	750	MT 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	Answers-student	750	student answers
2015	Answers-forum	375	Q&A forum answers
2015	Belief	375	commited belief

Table 2: English subtask: Summary of train (2012, 2013, 2014) and test (2015) datasets.

2.1 English Subtask

The English subtask dataset comprised pairs of sentences from news headlines (HDL), image descriptions (Images), answer pairs from a tutorial dialogue system (Answers-student), answer pairs from Q&A websites (Answers-forum), and pairs from a committed belief dataset (Belief).

For **HDL**, we used naturally occurring news headlines gathered by the Europe Media Monitor (EMM) engine (Best et al., 2005) from several different news sources (from April 2nd, 2013 to July 28th, 2014). EMM clusters together related news. Our goal was to generate a balanced dataset across the different similarity ranges. Therefore, we built two sets of headline pairs: a set where the pairs come from the same EMM cluster and another set where the headlines come from a different EMM cluster. Then, we computed the string similarity between those pairs. Accordingly, we sampled 1000 headline pairs of headlines that occur in the same EMM cluster, aiming for pairs equally distributed between minimal and maximal similarity using simple string similarity as a metric. We sampled another 1000 pairs from the different EMM cluster in the same manner.

[1] http://www.cs.columbia.edu/~weiwei/workshop/

Score	English	Spanish
5/4	*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.	El pájaro se esta bañando en el lavabo. El pájaro se está lavando en el aguamanil.
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.	John dijo que él es considerado como testigo, y no como sospechoso. "Él ya no es un sospechoso," John dijo.
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.	Ellos volaron del nido en grupos. Volaron hacia el nido juntos.
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.	La mujer está tocando el violín. La joven disfruta escuchar la guitarra.
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.	Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. La salida del sol al amanecer es una magnífica vista que puede presenciar si usted se despierta lo suficientemente temprano para verla.

Table 1: Similarity scores with explanations and examples for the English and Spanish subtasks, where the sentences in Spanish are translations of the English ones. A similarity score of 5 in English is mirrored by a maximum score of 4 in Spanish; the definitions pertaining to scores 3 and 4 in English are collapsed under a score of 3 in Spanish, with the definition "*The two sentences are mostly equivalent, but some unimportant details differ*."

The **Images** dataset was a subset of the PASCAL VOC-2008 dataset (Rashtchian et al., 2010), which consists of 1000 images with around 10 descriptions each, and has been used by a number of image description systems.

The source of the **Answers-student** pairs was the BEETLE corpus (Dzikovska et al., 2010), which is a question-answer dataset collected and annotated during the evaluation of the BEETLE II tutorial dialogue system. The BEETLE II system is an intelligent tutoring engine that teaches students basic electricity and electronics. The corpus was used in the student response analysis task of Semeval-2013. Given a question, a known correct "reference answer" and the "student answer", the goal of the task was to assess whether student answers were correct, contradictory or incorrect (partially correct, irrelevant or not in the domain). For STS, we selected pairs of answers made up by single sentences. The pairs were sampled from string similarity values between 0.6 and 1. We organized two buckets with 1000 pairs each: one with pairs of descriptions from the same image, and the other one with pairs of descriptions from different images.

The **Answers-forums** dataset consisted of paired answers collected from the Stack Exchange question and answer websites (http://stackexchange.com/). Some of the paired answers were responses to the same question, while others were responses to different questions. Each answer in the pair consisted of a statement composed of a single sentence or sentence fragment. For multi-sentence answers, we extracted the single sentence from the larger answer that appears to best summarize the answer.

The **Belief** pairs were collected from the DEFT Committed Belief Annotation dataset (LDC2014E55). All source documents are English Discussion Forum data. We sampled 2000 pairs using string similarity values between 0.5 and 1. It is worth noting that the similarity values were skewed, with very few pairs above 0.8 similarity.

In an attempt to improve the quality of the data, we selected 2000 pairs from each dataset and annotated them. This "raw" data was automatically filtered in order to achieve the following three (partially conflicting) goals: (1) To obtain a more uniform distribution across scores; (2) To select pairs with high inter-annotator agreement; (3) To select pairs which were difficult for a string-matching baseline. The filtering process was purely automated and involved no manual selection of pairs. The raw annotations and the Perl scripts that generated the final gold standard are available at the task website. See Table 2 for the number of selected pairs per dataset.

Table 1 shows the explanations and values associated with each score between 5 and 0. As in prior years, we used Amazon Mechanical Turk (AMT)[2] to crowdsource the annotation of the English pairs. Detailed instructions (Figure 1) were provided to the annotators, and they were asked to label each STS sentence pair on our six point scale, selecting from a dropdown box. Five sentence pairs were presented to each annotator at once, per human intelligence task (HIT), at a payrate of $0.20. We collected five separate annotations per sentence pair. Annotators were only eligible to work on the task if they had the Mechanical Turk Master Qualification, a special qualification conferred by AMT (using a priority statistical model) to annotators who consistently maintain a very high level of quality across a variety of tasks from numerous requesters. Access to these skilled workers entails a 20% surcharge.

To monitor the quality of the annotations, we used a gold dataset of 105 pairs that were manually annotated by the task organizers during STS 2013. We included one of these gold pairs in each set of five sentence pairs, where the gold pairs were indistinguishable from the rest. Unlike when we ran on CrowdFlower for STS 2013, the gold pairs were not used for training purposes, neither were workers automatically banned from the task if they made too many mistakes annotating the pairs. Rather, the gold pairs were only used to help in identifying and removing the data associated with poorly performing annotators. With few exceptions, 90% of the answers from each individual annotator fell within +/-1 of the answers selected by the organizers for the gold dataset.

2.2 Spanish Subtask

The Spanish subtask follows a setup similar to the English subtask, except that the similarity scores were adapted to fit a range from 0 to 4 (see Table 1). We thought that the distinction between a score of 3 and 4 for the English task would pose more difficulty for us in conveying into Spanish, as the sole difference between the two lies in how the annotators perceive the importance of additional details or missing information with respect to the core semantic interpretation of the pair. As this aspect entails a subjective judgement, we casted the annotation guidelines into straightforward and unambiguous instructions, and thus opted to use a similarity range from 0 to 4.

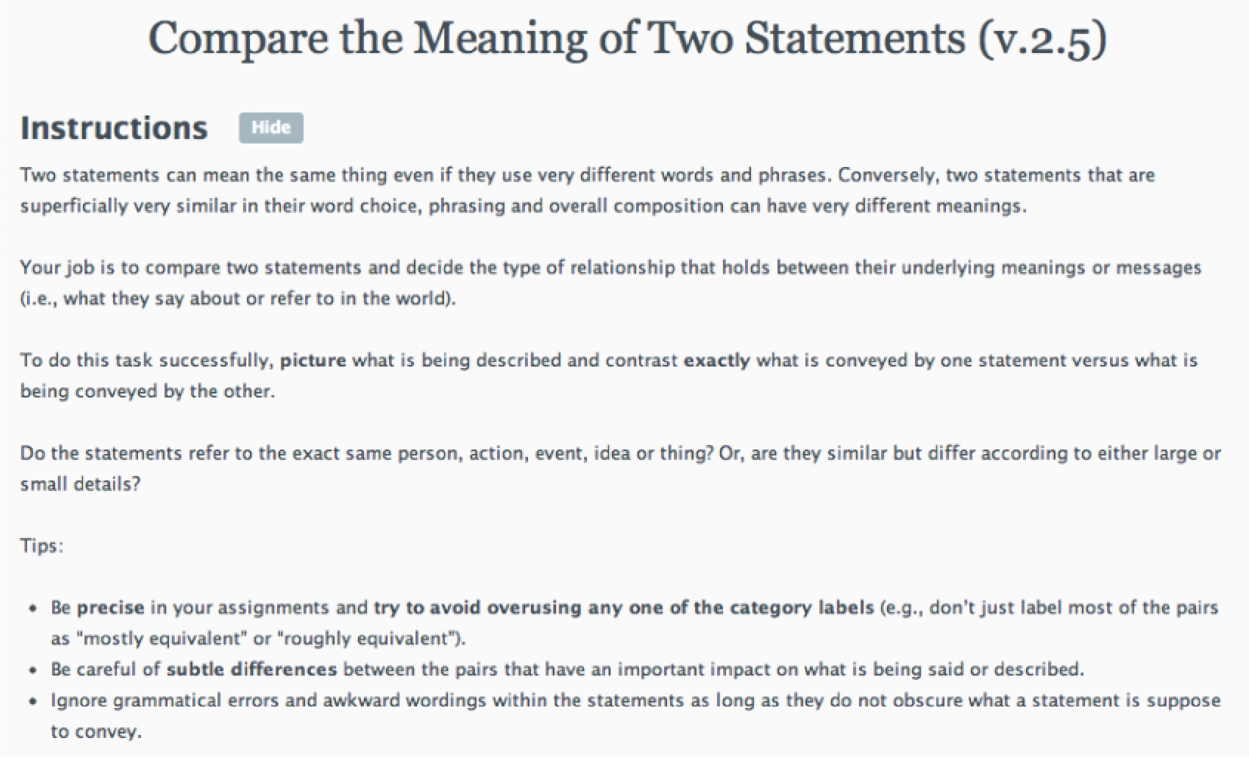
Prior to the evaluation window, the participants had access to a trial dataset consisting of 65 sentence pairs annotated for similarity and the test data released as part of SemEval 2014 Task 10 (Agirre et al., 2014), consisting of approximately 800 sentence pairs extracted from Spanish newswire and encyclopedic content. For the evaluations, we constructed two datasets, one extracted from the Spanish Wikipedia[3] (December 2013 dump) consisting of 251 sentence pairs, and the other one from contemporary news articles collected from news media in Spanish (November 2014) of 500 pairs.

2.2.1 Spanish Wikipedia

The Wikipedia dump was processed using the Parse::MediaWikiDump Perl library. We removed all titles, html tags, wiki tags and hyperlinks (keeping only the surface forms). Each article was split into paragraphs, where the first paragraph was considered to be the article's abstract, while the remaining ones were deemed to be its content. Each of these were split into sentences using the Perl library Uplug::PreProcess::SentDetect, and only the sentences longer than eight words were used. We

[2] www.mturk.com

[3] es.wikipedia.org



Compare the Meaning of Two Statements (v.2.5)

Instructions

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

Your job is to compare two statements and decide the type of relationship that holds between their underlying meanings or messages (i.e., what they say about or refer to in the world).

To do this task successfully, **picture** what is being described and contrast **exactly** what is conveyed by one statement versus what is being conveyed by the other.

Do the statements 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:

- Be **precise** in your assignments and **try to avoid overusing any one of the category labels** (e.g., don't just label most of the pairs as "mostly equivalent" or "roughly equivalent").
- Be careful of **subtle differences** between the pairs that have an important impact on what is being said or described.
- Ignore grammatical errors and awkward wordings within the statements as long as they do not obscure what a statement is suppose to convey.

Figure 1: Annotation instructions for English subtask.

iteratively computed the lexical similarity[4] between every sentence in the abstract and every sentence in the content, and retained those pairs whose sentence length ratio was higher than 0.5, and their similarity scored over 0.35.

The final set of sentence pairs was split into five bins, and their scores were normalized to range from 0 to 1. The more interesting and difficult pairs were found, perhaps not surprisingly, in bin 0, where synonyms/short paraphrases were more frequent, and 251 sentence pairs were manually selected from this bin in order to ensure a diverse and challenging set.

We then proceeded to annotate the sentence pairs for textual similarity by designing an AMT task, following a similar structure as in 2014, namely creating HITs consisting of seven sentence pairs, where six of them were a subset of the newly developed dataset, and one of them was reused from 2014 data with the purpose of control and to enable annotation quality comparisons.[5] As in the previous year, AMT providers were eligible to complete a task if they had more than 500 accepted HITs, with an over 90% acceptance rate. Each HIT was annotated by five AMT providers, and the remuneration was of $0.30 per HIT.[6] The final sentence pair similarity scores was computed by averaging over the judgments of the five AMT providers.

In order to assess the robustness of the AMT annotations, we computed the Pearson correlation between the similarity scores newly assigned to the control sentences, and those assigned in 2014. We obtained a measure of over 0.92, indicating a high resemblance between the two sets of judgements and highlighting the consistency of crowd wisdom, which is able to produce coherent outcomes irrespective of the individuals participating in the decision process.

2.2.2 Spanish News

The second Spanish dataset was extracted from news articles published in Spanish language media from around the world in November 2014. The hyperlinks to the articles were obtained by parsing

[4] Algorithm based on the Linux *diff* command (Algorithm::Diff Perl module).

[5] The control pair appeared randomly within each HIT.

[6] For additional information, we refer the reader to (Agirre et al., 2014).

the ”International” page of Spanish Google News,[7] which aggregates or clusters in real time articles describing a particular event from a diverse pool of news sites, where each grouping is labeled with the title of one of the predominant articles. By leveraging these clusters of links pointing to the sites where the articles were originally published, we were able to gather raw text that had a high probability to contain semantically similar sentences. We encountered several difficulties while mining the articles, ranging from each article having its own formatting depending on the source site, to advertisements, cookie requirements, to encoding for Spanish diacritics. We used the *lynx text-based browser*,[8] which was able to standardize the raw articles to a degree. The output of the browser was processed using a rule based approach taking into account continuous text span length, ratio of symbols and numbers to the text, etc., in order to determine when a paragraph is part of the article content. After that, a second pass over the predictions corrected mislabeled paragraphs if they were preceded and followed by paragraphs identified as content. All the content pertaining to articles on the same event was joined, sentence split, and *diff* pairwise similarities were computed. The set of candidate sentences followed the same constraints as those enforced for the Wikipedia dataset, namely a length ratio higher than 0.5 and a similarity score over 0.35. From these, we manually extracted 500 sentence pairs, which were annotated in an AMT task mirroring the same setup as used for the encyclopedic data annotation. The correlation between this year’s annotations and those of the 2014 STS task using the control sentence pairs remained high, at 0.886.

Since historically many of the text-to-text similarity algorithms have relied heavily on lexical matching, this year’s Spanish datasets featured sentence pairs with a higher degree of difficulty. This was achieved by handpicking pairs which shared some common vocabulary, yet carried completely different meanings at the sentence level.

2.3 Interpretable Subtask

The Semeval 2015 STS task offered a new pilot subtask on Interpretable STS. With this pilot task we wanted to start exploring whether participant systems were able to explain WHY two sentences are related/unrelated, adding an explanatory layer to the similarity score. As a first step in this direction, given a pair of sentences, participating systems needed to align the chunks between both sentences, describing what kind of relation exists between each pair of aligned chunks, and a score for the similarity/relatedness between them.

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

In Interpretable STS the similarity scores range from 0 to 5, as in the English subtask. With respect to the relation between the aligned chunks, the present pilot only allowed 1:1 alignments. As a consequence, we had to include an special alignment context tag (ALIC) to simulate those chunks which had some semantic similarity or relatedness in the other sentence, but could not have been aligned because of the 1:1 restriction. In the case of the aligned chunks, the following relatedness tags were defined:

- EQUI, for chunks which are semantically equivalent in the context.

[7] news.google.es

[8] lynx.browser.org

- OPPO, for chunks which are in opposition to each other in the context.
- SPE1 and SPE2, for chunks which have similar meanings, but which include different level of detailed information, chunk in sentence1 more specific than chunk in sentence2, or vice versa.
- SIMI, for chunks with similar meanings, but no EQUI, OPPO, SPE1, or SPE2.
- REL, for chunks which have related meanings, but no EQUI, OPPO, SPE1, SPE2, or SIMI.

In addition, a pair of chunks could be annotated with factuality (FACT) and polarity (POL), if there was a phenomena associated to those which made the meaning of the two chunks different. Finally, in the case of chunks which had not any similarity/relatedness in the other sentence, they were tagged as NOALI.

The pilot presented two scenarios: sentence raw text and gold standard chunks. In the sentence raw text scenario, given the input (pairs of sentences) participants first had to identify the chunks in each sentence, and then, align the corresponding chunks and assign a similarity score and a relatedness tag to them. Otherwise, in the gold standard scenario, participants were provided with the gold standard chunks, which were based on those used in the CoNLL 2000 chunking task (Tjong Kim Sang and Buchholz, 2000), with some adaptations (see annotation guidelines).

The training and test datasets consisted of 1500 and 753 sentence pairs respectively, extracted from the HDL and Images datasets used in 2014. 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.

3 System evaluation for STS

This Section reports the results for the English and Spanish subtasks. Note that participants could submit a maximum of three runs per subtask.

3.1 Evaluation metrics

As in previous exercises, we used Pearson product-moment correlation between the system scores and the GS scores.

Listing 1: STS interpretable - annotation format

```
<sentence id="6" status="">
 A woman riding a brown horse
 A young girl riding a brown horse
 ...
 <alignment>
  1 2 <==> 1 2 3 // SIMI // 4 // A woman <==>
    A young girl
  4 5 6 <==> 5 6 7 // EQUI // 5 // a brown
    horse <==> a brown horse
  3 <==> 4 // EQUI // 5 // riding <==> riding
 </alignment>
</sentence>
```

3.2 Baseline system

In order to provide a simple word overlap baseline (Baseline-tokencos), we tokenized the input sentences splitting on white spaces, and then each sentence was represented as a vector in the multidimensional token space. Each dimension had 1 if the token was present in the sentence, 0 otherwise. Vector similarity was computed using the cosine similarity metric.

We also ran the TakeLab system (Šarić et al., 2012) from STS 2012, which yielded strong results in previous years evaluations.[9]. The system was trained on all previous datasets STS12, STS13 and STS14, and tested on each subset of STS15.

3.3 Participation

29 teams participated in the English subtask, submitting 74 system runs. One team submitted fixes on one run past the deadline, as explicitly marked in Table 3. After the submission deadline expired, the organizers published the gold standard, the evaluation script, the scripts to generate the gold standard from raw annotation files, and participant submissions on the task website, in order to ensure a transparent evaluation process. As regards the Spanish STS task, it attracted 7 teams, which participated with 16 system runs.

3.4 English Subtask Results

Table 3 shows the results of the English subtask, with runs listed in alphabetical order. The correlation in each dataset is given, followed by the weighted mean correlation (the official measure) and the rank of the run.

[9] Code is available at http://ixa2.si.ehu.eus/stswiki

Run Name	answers-forums	answers-students	belief	headlines	images	Mean	Rank
Baseline-tokencos	0.4453	0.6647	0.6517	0.5312	0.6039	0.5871	61
Baseline-TakeLab	0.5391	0.6176	0.6165	0.7790	0.8115	0.6965	42
A96T-RUN1	0.6686	0.7192	0.7117	0.7357	0.7896	0.7337	29
ASAP-FIRSTRUN	0.2304	0.6503	0.3928	0.6614	0.6548	0.5695	63
ASAP-SECONDRUN	0.2374	0.7095	0.3986	0.7039	0.7294	0.6152	56
*ASAP-THIRDRUN	0.2303	0.6719	0.4342	0.7156	0.7250	0.6112	57
AZMAT-RUNABS	0.3099	0.4282	0.3568	0.5280	0.5118	0.4503	70
AZMAT-RUNCAP	0.2932	0.4282	0.3526	0.5350	0.5186	0.4512	69
AZMAT-RUNSCALE	0.2933	0.4293	0.3587	0.5264	0.5145	0.4490	71
BLCUNLP-1stRUN	0.4231	0.5152	0.5510	0.5651	0.7163	0.5709	62
BLCUNLP-2ndRUN	0.5725	0.6586	0.5510	0.7238	0.8271	0.6928	44
BLCUNLP-3rdRUN	0.5725	0.5753	0.4462	0.7309	0.8070	0.6556	49
BUAP-RUN1	0.5564	0.6901	0.6473	0.7167	0.7658	0.6936	43
DalGTM-run1	0.2902	-0.0534	0.0625	0.0598	0.0663	0.0623	74
DalGTM-run2	0.3537	0.1189	0.0625	0.2354	0.2042	0.1917	72
DalGTM-run3	0.1533	0.1189	-0.1319	-0.0395	0.2021	0.0731	73
DCU-RUN1	0.5556	0.6582	0.5464	0.8284	0.8394	0.7192	34
DCU-RUN2	0.5628	0.6233	0.7549	0.8187	0.8350	0.7340	28
DCU-RUN3	0.6530	0.6108	0.6977	0.8181	0.8434	0.7369	26
DLS@CU-S1	0.7390	0.7725	0.7491	0.8250	0.8644	0.8015	1
DLS@CU-S2	0.7241	0.7569	0.7223	0.8250	0.8631	0.7921	3
DLS@CU-U	0.6821	0.7879	0.7325	0.8238	0.8485	0.7919	5
ECNU-1stSVMALL	0.7145	0.7122	0.7282	0.7980	0.8467	0.7696	19
ECNU-2ndSVMONE	0.6865	0.7329	0.6977	0.8196	0.8358	0.7701	18
ECNU-3rdMTL	0.6919	0.7515	0.6951	0.8049	0.8575	0.7769	16
ExBThemis-default	0.6946	0.7505	0.7521	0.8245	0.8527	0.7878	8
ExBThemis-themis	0.6946	0.7505	0.7482	0.8245	0.8527	0.7873	9
ExBThemis-themisexp	0.6946	0.7784	0.7482	0.8245	0.8527	0.7942	2
FBK-HLT-RUN1	0.7131	0.7442	0.7327	0.8079	0.8574	0.7831	12
FBK-HLT-RUN2	0.7101	0.7410	0.7377	0.8008	0.8545	0.7801	13
FBK-HLT-RUN3	0.6555	0.7362	0.7460	0.7083	0.8389	0.7461	23
FCICU-Run1	0.6152	0.6686	0.6109	0.7418	0.7853	0.7022	41
FCICU-Run2	0.3659	0.6460	0.5896	0.6448	0.6194	0.5970	59
FCICU-Run3	0.7091	0.7096	0.7184	0.7922	0.8223	0.7595	20
IITNLP-FirstRun	0.3728	0.6605	0.7717	0.5996	0.8523	0.6712	47
MathLingBudapest-embedding	0.7039	0.7004	0.7325	0.7690	0.8038	0.7478	22
MathLingBudapest-hybrid	0.7231	0.7513	0.7473	0.8037	0.8442	0.7836	11
MathLingBudapest-machines	0.6977	0.7455	0.7363	0.8046	0.8414	0.7771	15
MiniExperts-Run1	0.6781	0.7304	0.6294	0.6912	0.8109	0.7216	33
MiniExperts-Run2	0.6454	0.7093	0.5165	0.6084	0.7999	0.6746	45
MiniExperts-Run3	0.6179	0.6977	0.3236	0.5775	0.7954	0.6353	55
NeRoSim-R1	0.5260	0.7251	0.6311	0.8131	0.8585	0.7438	24
NeRoSim-R2	0.6940	0.7446	0.7512	0.8077	0.8647	0.7849	10
NeRoSim-R3	0.6778	0.7357	0.7220	0.8123	0.8570	0.7762	17
RTM-DCU-1stPLS.svr	0.5484	0.5549	0.6223	0.7281	0.7189	0.6468	50
RTM-DCU-2ndST.svr	0.5484	0.5549	0.6223	0.7281	0.7189	0.6468	51
RTM-DCU-3rdST.rr	0.5484	0.5549	0.6223	0.7281	0.7189	0.6468	52
Samsung-alpha	0.6589	0.7827	0.7029	0.8342	0.8701	0.7920	4
Samsung-beta	0.6586	0.7819	0.6995	0.8342	0.8713	0.7916	7
Samsung-delta	0.6639	0.7825	0.6952	0.8417	0.8634	0.7918	6
SemantiKLUE-RUN1	0.4913	0.7005	0.5617	0.6681	0.7915	0.6717	46
SopaLipnIimas-MLP	0.6178	0.5864	0.6886	0.8121	0.8184	0.7175	36
SopaLipnIimas-RF	0.6709	0.5914	0.7238	0.8123	0.8414	0.7356	27
SopaLipnIimas-SVM	0.5918	0.5718	0.7028	0.7985	0.8104	0.7070	39
T2a-TrWP-run1	0.6857	0.6618	0.6769	0.7709	0.7865	0.7251	31
T2a-TrWP-run2	0.6857	0.6618	0.7245	0.7709	0.7865	0.7311	30
T2a-TrWP-run3	0.6857	0.6612	0.6772	0.7710	0.7865	0.7250	32
TATO-1stWTW	0.6796	0.6853	0.7206	0.7667	0.8167	0.7422	25
UBC-RUN1	0.4764	0.5459	0.6788	0.6368	0.7852	0.6364	53
UMDuluth-BlueTeam-Run1	0.6561	0.7816	0.7363	0.8085	0.8236	0.7775	14
UQeResearch-AllRuns-run1	0.5923	0.6876	0.5904	0.7521	0.7817	0.7032	40
UQeResearch-AllRuns-run2	0.6132	0.6882	0.6229	0.7602	0.7855	0.7130	37
UQeResearch-AllRuns-run3	0.6188	0.6757	0.7178	0.7549	0.7769	0.7189	35
USAAR_SHEFFIELD-modelx	0.3706	0.3609	0.4767	0.5183	0.5436	0.4616	68
USAAR_SHEFFIELD-modely	0.6264	0.7386	0.7050	0.7927	0.8162	0.7533	21
USAAR_SHEFFIELD-modelz	0.4237	0.6757	0.6994	0.5239	0.6833	0.6111	58
WSL-run1	0.3759	0.5269	0.6387	0.5462	0.5710	0.5379	66
WSL-run2	0.4287	0.6028	0.5231	0.6029	0.4879	0.5424	65
WSL-run3	0.3709	0.5437	0.6478	0.5752	0.6407	0.5672	64
Yamraj-1stRUNNAME	0.5634	0.6727	0.6387	0.6067	0.7425	0.6558	48
Yamraj-2ndRUNNAME	0.4367	0.4716	0.4890	0.5533	0.4799	0.4919	67
Yamraj-3rdRUNNAME	0.5168	0.5835	0.6540	0.5861	0.6097	0.5912	60
yiGou-midbaitu	0.5797	0.6571	0.6473	0.7115	0.8036	0.6964	42
yiGou-xiaobaitu	0.6102	0.6872	0.6065	0.7369	0.8133	0.7114	38
*UBC-RUN1	0.4764	0.5459	0.6788	0.6368	0.7852	0.6364	54

Table 3: Task 2a: English evaluation results in terms of Pearson correlation.

The results of the best system run were significantly different (p-value < 0.05) from the 11th top scoring system run and below. The top 10 systems did not show statistical significant variation among them. None of these runs was significantly different from any other in the top ten runs, indicating that the best systems performed very close to each other.

3.5 Spanish Subtask Results

The official evaluation results of the Spanish subtask are presented in Table 4. The last row, Baseline-tokencos, shows the results obtained us-

Run Name	System Type	Wikipedia	Newswire	Weighted Mean	Rank
BUAP-run1	unknown	0.489	0.405	0.433	14
ExBThemis-trainEn	supervised	0.676	0.671	0.672	3
ExBThemis-trainEs	supervised	0.705	**0.683**	**0.690**	1
ExBThemis-trainMini	supervised	**0.706**	0.681	0.689	2
RTM-DCU-1stST.tree	supervised	0.582	0.525	0.544	8
RTM-DCU-2ndST.rr	supervised	0.582	0.525	0.544	7
RTM-DCU-3rdST.SVR	supervised	0.582	0.525	0.544	6
SopaLipnIimas-MLP	supervised	0.253	0.534	0.440	12
SopaLipnIimas-RF	supervised	0.564	0.565	0.565	5
SopaLipnIimas-SVM	supervised	0.419	0.401	0.407	15
UMDuluth-BlueTeam-run1	unsupervised	0.594	0.655	0.634	4
MiniExperts-run1	supervised	0.524	0.508	0.513	11
MiniExperts-run2	supervised	0.467	0.544	0.518	9
MiniExperts-run3	supervised	0.440	0.552	0.515	10
Yamraj-1stNoConfidence	unsupervised	0.577	0.365	0.436	13
Yamraj-1stWithConfidence	unsupervised	0.532	0.342	0.405	16
Baseline-tokencos		0.529	0.495	0.506	

Table 4: Task 2b: Spanish evaluation results in terms of Pearson correlation.

ing the same baseline as for the English STS task, which 69% of the system runs were able to surpass. Only about one fifth of the systems were unsupervised, among which, the top performing system, UMDuluth-BlueTeam-run1, was able to come within 0.1 correlation points from the top performing system on Wikipedia and within 0.03 on the Newswire dataset. This relatively narrow gap suggests that unsupervised semantic textual similarity is a viable option for languages with limited resources.

Our efforts for generating closer to real-life textual similarity scenarios, and thus more difficult cases to be discerned by automated systems, were reflected in the lower correlations obtained on this year's datasets in comparison to those of 2014. For Wikipedia, the highest ranking system, ExBThemis-trainMini, achieved a correlation of 0.70, while in 2014, the highest correlation on the same dataset type was of 0.78. This difference was even steeper for the newswire data, where the top system, ExBThemis-trainEs, scored 0.683 in comparison to 2014, where the top ranked system attained a correlation of 0.845.

4 System evaluation for interpretable STS

4.1 Evaluation metrics

Participating runs were evaluated using four different metrics: F1 where alignment type and score are ignored; F1 where alignment types need to match, but scores are ignored; F1 where alignment type is ignored, but each alignment is penalized when scores do not match; and, F1 where alignment types need to match, and each alignment is penalized when scores do not match.

4.2 Baseline system

The baseline used for the interpretable subtask was a cascade concatenation of the following procedures. To begin with, we considered a brief NLP step in which input sentences were tokenized using simple regular expressions. Additionally, this step collected chunk regions coming either from gold standard or from the chunking done by *ixa-pipes-chunk* (Agerri et al., 2014).

Once the language processing step was finished, the baseline started the lower-cased token aligning phase, in which it aligned tokens between input sentences if they were the same exact word. This alignment was considered as a link between those tokens. After token links were computed, the baseline used chunk boundaries as token regions to group individual tokens into groups; and computed all links between groups. The weight of the link between groups was proportional to the number of links counted between the tokens inside that specific group.

After that, the baseline had an optimization step in which it discovered the groups x,y that had the highest link weight, but also which were the chunks

Run name	H ALI	H TYPE	H SCORE	H T+S	Rank	I ALI	I TYPE	I SCORE	I T+S	Rank
NeRoSim_R3	0.8976	**0.6666**	0.8157	**0.6426**	1	0.8834	0.6035	0.7837	0.5759	4
NeRoSim_R2	0.8972	0.6558	**0.8263**	0.6401	2	0.8800	0.5854	0.7818	0.5619	6
NeRoSim_R1	**0.8984**	0.6543	0.8262	0.6389	3	**0.8870**	**0.6143**	0.7877	0.5841	2
UMDuluth_BlueTeam_1	0.8861	0.5962	0.7960	0.5887	4	0.8853	0.5842	0.7932	0.5729	5
UMDuluth_BlueTeam_2	0.8861	0.5962	0.7968	0.5883	5	0.8853	0.6095	**0.7968**	**0.5964**	1
UMDuluth_BlueTeam_3	0.8861	0.5900	0.7980	0.5834	6	0.8853	0.5964	0.7909	0.5822	3
SimCompass_prefix	0.8360	0.5834	0.7474	0.5338	8	0.8361	0.4708	0.7269	0.4157	12
SimCompass_word2vec	0.8716	0.5806	0.7654	0.5253	9	0.8624	0.4599	0.7405	0.4017	13
SimCompass_combined	0.8710	0.5813	0.7651	0.5239	10	0.8490	0.4555	0.7294	0.3965	14
ExBThemis_avgScorer	0.8146	0.4943	0.7171	0.4885	11	0.8057	0.4413	0.6992	0.4246	11
ExBThemis_mostFreqScorer	0.8146	0.4943	0.7140	0.4884	12	0.8057	0.4413	0.7007	0.4296	9
ExBThemis_regressionScorer	0.8146	0.4943	0.7158	0.4883	13	0.8052	0.4406	0.6989	0.4288	10
FCICU_Run1	0.8455	0.4480	0.7160	0.4325	14	0.8457	0.4740	0.7273	0.4482	7
+RTM-DCU_1stIBM2Alignment	0.4914	0.3712	0.4550	0.3712	15	0.3540	0.2283	0.3187	0.2282	15
∗UBC_RUN2	0.8991	0.6402	0.8211	0.6185	-	0.8846	0.6557	0.8085	0.6159	-
∗UBC_RUN1	0.8991	0.5882	0.8031	0.5882	-	0.8846	0.4749	0.7709	0.4746	-
BASELINE	0.8448	0.5556	0.7551	0.5556	7	0.8388	0.4328	0.7210	0.4326	8

Table 5: STS interpretable results for the gold chunks scenario. Best results have been marked in bold. 'H' stands for Headlines data set and 'I' stands for Images data set. + symbol denotes resubmissions and ∗ symbol denotes task organizers.

Run Name	H ALI	H TYPE	H SCORE	H T+S	Rank	I ALI	I TYPE	I SCORE	I T+S	Rank
UMDuluth_BlueTeam_3	**0.7820**	**0.5154**	**0.7024**	**0.5098**	1	**0.8336**	0.5605	0.7456	0.5473	2
UMDuluth_BlueTeam_2	**0.7820**	0.5109	0.6986	0.5049	2	**0.8336**	**0.5759**	**0.7511**	**0.5634**	1
UMDuluth_BlueTeam_1	**0.7820**	0.5058	0.6968	0.5004	3	**0.8336**	0.5529	0.7498	0.5431	3
ExBThemis_avgScorer	0.7032	0.4331	0.6224	0.4290	5	0.6966	0.3970	0.6068	0.3806	6
ExBThemis_mostFreqScorer	0.7032	0.4331	0.6200	0.4288	6	0.6966	0.3970	0.6106	0.3870	4
ExBThemis_regressionScorer	0.7032	0.4331	0.6209	0.4284	7	0.6966	0.3970	0.6092	0.3867	5
SimCompass_word2vec	0.6461	0.4334	0.5619	0.3878	8	0.5428	0.2831	0.4561	0.2427	8
SimCompass_prefix	0.6310	0.4284	0.5526	0.3872	9	-	-	-	-	-
SimCompass_combined	0.6467	0.4333	0.5636	0.3870	10	0.5433	0.2854	0.4545	0.2421	9
+RTM-DCU_1stIBM2Alignment	0.4914	0.3712	0.4550	0.3712	11	0.3540	0.2283	0.3187	0.2282	10
∗UBC_RUN2	0.7709	0.4865	0.7014	0.4705	-	0.8388	0.6019	0.7634	0.5643	-
∗UBC_RUN1	0.7709	0.5019	0.6892	0.5019	-	0.8388	0.4450	0.7280	0.4447	-
BASELINE	0.6701	0.4571	0.6066	0.4571	4	0.7060	0.3696	0.6092	0.3693	7

Table 6: STS interpretable results for the system chunks scenario. Best results have been marked in bold. 'H' stands for Headlines data set and 'I' stands for Images data set. + symbol denotes resubmissions and ∗ symbol denotes task organizers.

that were linked to either x or y but not with a maximum alignment weight. In consequence, we also knew which chunks were left unaligned.

Finally, in the last phase, the baseline directly assigned labels and scores depending on the results of the previous step. The algorithm was the following: to chunks with the highest link weight assign label = "EQUI" and score = 5, to the rest of aligned chunks (with lower weights) assign label = "ALIC" and score = NIL, and, to not aligned chunks assign label = "NOALI" and score = NIL. To conclude, the baseline produced the output in .wa (word alignment) format as specified in the guidelines.

4.3 Participation

The interpretable subtask allowed up to a total of three submissions for each team on each of the evaluation scenarios. As previously mentioned, the first evaluation scenario provided gold standard chunks for all input sentence pairs. This way, participating systems only had to worry about making correct alignments and providing them with appropriate labels and scores. The second evaluation scenario consisted of using only raw text as input, and so, each system was also responsible for segmenting the input.

Seven teams participated on the gold chunks scenario, and out of them five teams also participated in the system chunks scenario as it was more challenging. The UBC system participation, marked with a ∗, corresponds to the organizer team for the interpretable STS subtask. However, it should be noted that the actual participating team was an independent sub team that was not involved in the task organization. Moreover, one more team is marked with + as their results reflect a resubmission.

4.4 Interpretable Subtask Results

Results for the gold chunks scenario and the system chunks scenario are shown in Table 5 and Table 6, respectively. Each row of the tables corresponds to a run configuration named *TeamID_RunID*, and each column corresponds to a evaluation result.

Note that task results are separately written with respect to the scenario, but distinct datasets that pertain to the same scenario have been collapsed in the corresponding table so that 'H' corresponds to the Headlines dataset and 'I' corresponds to the Images dataset. A unique baseline was used for both evaluation scenarios and its performance is jointly presented with the scores obtained by participants.

Results clearly show that the system chunks scenario was considerably more challenging than the gold chunks scenario. Actually, the complexity of the evaluation was incremental for the four available metrics, and, the most challenging F Type+Score metric performance seems bounded by the performance obtained in the F alignment metric, which obviously, was lower for the system chunks.

As regards both datasets, we can also conclude that the Images dataset resulted to be more challenging than the Headlines dataset. For instance, in the gold chunks scenario, the participant average F Type+Score metric reached 0.4748 for the Images dataset, whereas it reached 0.5381 for Headlines.[10] The maximum value obtained by participants was also higher, as it reached 0.6426 and 0.5964 respectively for Headlines and Images. For the case of the system chunks scenario, the average results followed the same tendency as the participant average F Type+Score metric reached 0.3912 for the Images dataset and 0.4335 for Headlines; both values under the ones obtained for the gold chunks. In contrast, the maximum value obtained by participants was in this case greater for Images, as it reached 0.5634, whereas it reached 0.5098 for Headlines.

4.5 Tools and resources

The majority of the systems used the same kind of tools for both scenarios despite integrating an auxiliary chunker for system chunks runs. The most used NLP tools for preprocessing are Stanford's NLP parser and the OpenNLP framework. Actually, all of the teams confirmed that they performed some kind of input text processing such as lemmatization, part of speech tagging or syntactic parsing. Additional resources such as named-entity recognition and acronym repositories, ConceptNet, NLTK, time and date resolution or PPDB were also used by most of the participants. 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 Mikolov embeddings were used by several teams to compute word similarity.

5 Conclusion

This year participants were challenged with new datasets for English and Spanish, including image captions, news headlines, Wikipedia articles, news, and new genres like answers from a tutorial dialogue system, answers from Q&A websites, and commited belief. The crowdsourced annotations had a high inter-tagger agreement. The English subtask attracted 29 teams, while the Spanish subtask had 7 teams.

In addition, we succesfully introduced a new subtask on interpretability, where systems add a explanatory layer, in the form of alignments between text segments, explicitly annotating the kind of relation and the score for each segment pair. The interpretable subtask attracted 7 teams.

Acknowledgements

This work was partially funded by MINECO (CHIST-ERA READERS project – PCIN-2013-002-C02-01, SKaTeR project – TIN2012-38584-C06-02), and the European Commission (QTLEAP – FP7-ICT-2013.4.1-610516). Aitor Gonzalez-Agirre and Iñigo Lopez-Gazpio are supported 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 psycholin-*

[10] The team involved on the organization of the task, marked by the symbol $*$ is not taken into account for the following statistics.

guistics. Robert Berwick and Steven Abney and Carol Tenny(eds.), pages 257–278. Springer Science & Business Media.

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 (LREC2014)*, pages 26–31.

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)*, pages 385–393, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

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)*, pages 81–91, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Clive Best, Erik van der Goot, Ken Blackler, Teófilo Garcia, and David Horby. 2005. Europe Media Monitor - System Description. In *EUR Report 22173-En*, Ispra, Italy.

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

Omer Levy, Torsten Zesch, Ido Dagan, and Iryna Gurevych. 2013. Recognizing Partial Textual Entailment. In *ACL (2)*, pages 451–455.

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

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 '10, pages 139–147, Stroudsburg, PA, USA.

V Rus, M Lintean, C Moldovan, W Baggett, N Niraula, and B 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.

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*, pages 127–132. Association for Computational Linguistics.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. TakeLab: Systems for Measuring Semantic Text Similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 441–448, Montréal, Canada, 7-8 June. Association for Computational Linguistics.