Noname manuscript No.
(will be inserted by the editor)

SICK Through the SemEval Glasses

Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment

Luisa Bentivogli · Raffaella Bernardi · Marco Marelli · Stefano Menini · Marco Baroni · Roberto Zamparelli

Received: date / Accepted: date

Abstract This paper is an extended description of SemEval-2014 Task 1, the task on the evaluation of Compositional Distributional Semantics Models on full sentences. Systems participating in the task were presented with pairs of sentences and were evaluated on their ability to predict human judgments on *(i)* semantic relatedness and *(ii)* entailment. Training and testing data were subsets of the SICK (Sentences Involving Compositional Knowledge) data set. SICK was developed with the aim of providing a proper benchmark to evaluate compositional semantic systems, though task participation was open to systems based on any approach. Taking advantage of the SemEval experience, in this paper we analyze the SICK data set, in order to evaluate the extent to which it meets its design goal and to shed light on the linguistic phenomena that are still challenging for state-of-the-art computational semantic systems. Qualitative and quantitative error analyses show that many systems are quite sensitive to changes in the proportion of sentence pair types, and degrade in the presence of additional lexico-syntactic complexities which do not affect human judgements. More compositional systems seem to perform better when the task proportions are changed, but the effect needs further confirmation.

1 Introduction

Distributional Semantic Models (DSMs) approximate the meaning of words with vectors summarizing their patterns of co-occurrence in corpora. Recently, several compositional extensions of DSMs (CDSMs) have been proposed, with the purpose of representing the meaning of phrases and sentences by composing the distributional representations of the words they contain (Baroni and Zamparelli, 2010; Mitchell and Lapata, 2010; Grefenstette and Sadrzadeh, 2011; Socher et al, 2012). Despite the ever increasing interest of the field in this domain, the development of adequate benchmarks for CDSMs, especially at the sentence level, is still lagging behind. Existing data sets, such as those introduced by Mitchell and Lapata (2008) and Grefenstette and Sadrzadeh

L. Bentivogli, S. Menini
Fondazione Bruno Kessler (FBK), Via Sommarive, 18 Povo (TN), 38123, Italy
Tel.: +39-0461314444
E-mail: bentivo,menini@fbk.eu

M. Baroni, R. Bernardi, M. Marelli, R. Zamparelli
University of Trento, Corso Bettini, 31 Rovereto (TN), 38068 , Italy +39-0464-808615
E-mail: name.surname@unitn.it

(2011), are limited to a few hundred instances of very short sentences with a fixed structure. In the last ten years, several large data sets have been developed for various computational semantics tasks, such as Semantic Text Similarity (STS) (Agirre et al, 2012) or Recognizing Textual Entailment (RTE) (Dagan et al, 2006). Working with such data sets, however, requires dealing with issues, such as identifying non-compositional multiword expressions, recognizing named entities or accessing encyclopedic knowledge, which have little to do with compositionality *per se*. CDSMs should instead be evaluated on data that is challenging for reasons due to semantic compositionality, such as context-cued synonymy resolution and other lexical variation phenomena, active/passive and other syntactic alternations, impact of negation at various levels, operator scope, and other effects linked to the functional lexicon. These issues do not occur frequently in, e.g., the STS and RTE data sets.

With these considerations in mind, we developed SICK (Sentences Involving Compositional Knowledge), a data set aimed at filling this gap, including a large number of sentence pairs that are rich in the lexical, syntactic and semantic phenomena that CDSMs are expected to account for, but do not require dealing with other aspects of existing sentential data sets that are not within the scope of compositional distributional semantics. Moreover, we distinguished between generic semantic knowledge about general concept categories (such as knowledge that a couple is formed by a bride and a groom) and encyclopedic knowledge about specific instances of concepts (e.g., the fact that the current president of the US is Barack Obama). The SICK data set contains many examples of the former, but none of the latter.

SICK was used in the SemEval 2014 shared task on the evaluation of Compositional Distributional Semantics Models on full sentences (Marelli et al, 2014a), where systems were presented with pairs of sentences and were evaluated on two tasks: *(i)* predicting the degree of *semantic relatedness* between the two sentences, and *(ii)* detecting the *entailment* relation holding between them. These two tasks are not new in the literature. With respect to the first one, previous works often used tasks of *semantic similarity*, rather than *semantic relatedness*. In the Semantic Text Similarity Task of Agirre et al (2012), for instance, 'similarity' is defined as the degree of equivalence between two sentences. Budanitsky and Hirst (2006) highlighted how the concept of 'relatedness' is more general than that of 'similarity'; the relatedness between two entities can be expressed by many relations, and similarity (e.g. money-cash) is just one of them. Other possible relations are meronymy (e.g. car-wheel), antonymy (e.g hot-cold) or even just frequent association (e.g. rain-flood). This explains why, according to Gabrilovich and Markovitch (2007), the assessment of *semantic relatedness* involves a deeper understanding of the text than the evaluation of semantic similarity, one which requires common sense and domain-specific knowledge.

The *entailment* task consists of deciding, given two text fragments, whether the meaning of one text fragment is entailed (i.e. can be inferred) by the other. The notion of entailment is defined in terms of truth values: a text t entails another text h if, typically, a human reading t would infer that h is most likely true (Dagan et al, 2006). The entailment relation differs from traditional text similarity, and can only be applied to declarative sentences, *qua* truth-value bearing elements. Both tasks can be formally defined, or left at an intuitive level: in the SICK data set gold-standard annotation was obtained via crowdsourcing by giving participants examples of the various relations, rather than the exact definition of the two tasks.

The SICK-based SemEval evaluation exercise was especially targeted to developers of CDSMs, but participation was open to systems based on any approach. The rationale behind this choice was that—besides being of intrinsic interest—the performance of the latter systems would situate CDSM's performance within the broader landscape of computational semantics. The SemEval results highlighted the need for a deeper analysis of the various performances, with the aim of understanding the specific characteristics of the systems and the main difficulties encountered in addressing the various phenomena represented in SICK.

This paper attempts to carry out this analysis, using various orthogonal approaches. First, we try to distinguish, when possible, the contribution to the final result of the more compositional features used by some of the systems, especially in terms of robustness. In order to study this aspect, we check how the various systems performed on a subset of the results where the task's conditions have been balanced. Finally, we carry out a qualitative evaluation of those cases that were easy for humans but hard for machines, looking for linguistic generalizations, but also for better ways to design future releases of the data set or alternative methods to study compositionality in computational semantics.

The paper is organized as follows. In Section 2 we introduce the SICK data set, describing how it was built and giving detailed statistics about the type of data it contains. In Section 3 we present the SemEval 2014 shared task where SICK was proposed, and its outcomes in terms of community response, results achieved, and approaches adopted. Finally, in Section 4 we present the new analyses carried out on the data set and on the performances of the participating systems.

2 The SICK data set

The SICK data set[1] consists of about 10,000 English sentence pairs annotated for relatedness in meaning and entailment. The main characteristics of the data set are outlined in the following subsections, while all the details about the procedure followed in its creation can be found in Marelli et al (2014b).

2.1 Data set creation

SICK was built starting from two existing data sets: the 8K ImageFlickr data set[2] (Hodosh et al, 2013) and the SemEval-2012 STS MSR-Video Descriptions data set[3] (Chen and Dolan, 2011). The 8K ImageFlickr data set is a data set of images, where each image is associated with five descriptions. To derive SICK sentence pairs we randomly chose 750 images and we sampled two descriptions from each of them. The SemEval-2012 STS MSR-Video Descriptions data set is a collection of sentence pairs sampled from the short video snippets which comprise the Microsoft Research Video Description Corpus. A subset of 750 sentence pairs were randomly chosen from this data set to be used in SICK.

In order to generate SICK data from the 1,500 sentence pairs taken from the source data sets, a 3-step process was applied to each sentence composing the pair, namely *(i) normalization*, *(ii) expansion* and *(iii) pairing*. Table 1 presents an example of the output of each step in the process.

The *normalization* step was carried out on the original sentences (*Orig_a, Orig_b*) to exclude or simplify instances that contained lexical, syntactic or semantic phenomena that CDSMs are currently not expected to account for. For instance, named entities were replaced with words standing for the class (e.g. *Ferrari* with *car*); numbers not acting as determiners were removed (e.g. "*A football player is wearing a green jersey with the number 4 on it*" was normalized as "*A football player is wearing a green jersey with a number on it*"), while numbers used as determiners were turned into letters (e.g. *3 people* into *Three people*); multiword expressions - i.e. sequences of words which habitually co-occur and whose meaning cannot be derived compositionally - were

[1] The original SICK data set and all the derived versions used for the analyses carried out in this paper can be downloaded at http://clic.cimec.unitn.it/composes/sick.html.

[2] http://nlp.cs.illinois.edu/HockenmaierGroup/data.html

[3] http://www.cs.york.ac.uk/semeval-2012/task6/index.php?id=data

Table 1 Data set creation process. Examples of the normalization and expansion of a sentences pair.

Original pair	
Orig_a: *"A sea turtle is hunting for fish"*	**Orig_b:** *"The turtle followed the fish"*
Normalized pair	
Norm_a: *"A sea turtle is hunting for fish"*	**Norm_b:** *"The turtle is following the fish"*
Expanded pairs	
Sim_a: *"A sea turtle is hunting for food"*	**Sim_b:** *"The turtle is following the red fish"*
Contr_a: *"A sea turtle is not hunting for fish"*	**Contr_b:** *"The turtle isn't following the fish"*
Diff_a: *"A fish is hunting for a turtle in the sea"*	**Diff_b:** *"The fish is following the turtle"*

removed. The general idea behind normalization was to remove unwanted phenomena without changing the original sentence significantly and make it easily processable by state-of-the-art parsers. To ensure the quality of the normalization phase, each sentence in the original pairs was normalized by two different annotators with a strong background in linguistics, and a third judge chose the most suitable one. A post-check of the data set revealed that some multiword expressions are actually present in the data. These are mostly of two types: *(i)* they have the function of subject in the sentence, and were thus not removable without considerably modifying it; *(ii)* they are not clearly non compositional and can be processed by systems. Related to this latter characteristics, it has to be acknowledged that the boundary between multiwords and free (i.e. completely compositional) combinations of words is not clear-cut, and distinguishing between expressions that lie in the middle of a continuum is a very subtle and sometimes controversial task.

The *expansion* step was applied to each of the normalized sentences (*Norm_a*, *Norm_b*) in the pair, in order to create up to three new sentences for each normalized one with specific characteristics suitable for CDSM evaluation. In this step, syntactic and lexical transformations with predictable effects were applied to each normalized sentence, in order to obtain *(i)* a sentence with a similar meaning *(Sim)*, *(ii)* a sentence with a logically contradictory or at least highly contrasting meaning *(Contr)*, and *(iii)* a sentence that contains most of the same lexical items, but has a different meaning *(Diff)*. The latter transformation was carried out mainly by scrambling the words of the normalized sentence, but only where this scrambling could yield a meaningful sentence; as a result, not all normalized sentences have a *Diff* expansion. Note that as a result of the expansion process, two expansion sets are created for each pair, one for each of the normalized sentences.

Finally, in the *pairing* step each normalized sentence (i.e. *Norm_a*, *Norm_b*) was combined with all the sentences resulting from the expansion phase as well as with the other normalized sentence in the pair, leading to a total of 13 different sentence pairs. Among them, 6 pairs belong to the same expansion set, i.e. the paired sentences originate from the same source sentence (*Norm_a-Sim_a,Norm_a-Contr_a,Norm_a-Diff_a,Norm_b-Sim_b,Norm_b-Contr_b,Norm_b-Diff_b*) – we will refer to them as the *"same set"* pairs – while 7 pairs are composed of sentences belonging to the different expansion sets (*Norm_a-Norm_b, Norm_a-Sim_b, Norm_b-Sim_a, Norm_a-Contr_b, Norm_b-Contr_a, Norm_a-Diff_b, Norm_b-Diff_a*) – we will refer to them as the *"cross set"* pairs.

Furthermore, a number of pairs composed of completely unrelated sentences were added to the data set by randomly taking two sentences from two different pairs. For example: *"A sea*

Table 2 Distribution of the SICK sentence pairs with respect to the transformations performed during data set creation.

Expected relation	N. of pairs	
Similar meaning	4366	(44.4%)
Norm_a-Sim_a, Norm_a-Sim_b, Norm_b-Sim_a, Norm_b-Sim_b, Norm_a-Norm_b		
Contrasting meaning	3574	(36.3%)
Norm_a-Contr_a, Norm_a-Contr_b, Norm_b-Contr_a, Norm_b-Contr_b		
Similar lexicon, different meaning	703	(7.1%)
Norm_a-Diff_a, Norm_a-Diff_b, Norm_b-Diff_a, Norm_b-Diff_b		
Unrelated	1197	(12.2%)
Total	9840	(100%)

turtle is hunting for fish" and *"A young woman is playing the guitar"*. The result is a set of about 10,000 new sentence pairs, in which each sentence is contrasted with either a (near) paraphrase, a contradictory or strongly contrasting statement, another sentence with very high lexical overlap but different meaning, or a completely unrelated sentence. The rationale behind this approach was to build a data set which hindered methods based on individual lexical items, on the syntactic complexity of the two sentences or on pure world knowledge, thus encouraging the use of a compositional semantics step in understanding when two sentences have close meanings or entail each other.

The distribution of the SICK sentence pairs with respect to the transformations performed during data set creation is presented in Table 2, which summarizes the type of relation predicted to hold between the sentences in the pair, all the pairing combinations and their frequencies in the data set. We stress that we constructed the pairs by following the procedure outlined in order to generate a balanced distribution of possible sentence relations. However, the ultimate assessment of semantic relatedness and entailment between sentence pairs was left to human judges, as illustrated in the next section.

2.2 Relatedness and Entailment annotation

Each pair in the SICK data set was annotated to mark *(i)* the degree to which the two sentence meanings are related (on a 5-point scale), and *(ii)* whether one entails the other. In particular, for the entailment task three labels were considered: ENTAILMENT (if sentence A is true, sentence B is true), CONTRADICTION (if A is true, then B is false), NEUTRAL (the truth of B cannot be determined on the basis of A).

The ratings were collected through a large crowdsourcing study (see Marelli et al (2014b) for all details), where each pair was evaluated by 10 different subjects, and the order of presentation of the sentences was counterbalanced (i.e., 5 judgments were collected for each presentation order). Swapping the order of the sentences within each pair served a two-fold purpose: (i) evaluating the entailment relation in both directions and (ii) controlling possible bias due to priming effects in the relatedness task. In order to clarify the task to non-expert participants, while avoiding biasing their judgments with strict definitions, the instructions described the task through examples of relatedness and entailment. Furthermore for the entailment task participants

Table 3 Examples of sentences pairs with their gold relatedness scores (on a 5-point rating scale).

Relatedness score	Example
1.6	A: "*A man is jumping into an empty pool*"
	B: "*There is no biker jumping in the air*"
2.9	A: "*Two children are lying in the snow and are making snow angels*"
	B: "*Two angels are making snow on the lying children*"
3.6	A: "*The young boys are playing outdoors and the man is smiling nearby*"
	B: "*There is no boy playing outdoors and there is no man smiling*"
4.9	A: "*A person in a black jacket is doing tricks on a motorbike*"
	B: "*A man in a black jacket is doing tricks on a motorbike*"

Table 4 Examples of sentences pairs with their gold entailment labels.

Entailment label	Example
ENTAILMENT	A: "*Two teams are competing in a football match*"
	B: "*Two groups of people are playing football*"
CONTRADICTION	A: "*The brown horse is near a red barrel at the rodeo*"
	B: "*The brown horse is far from a red barrel at the rodeo*"
NEUTRAL	A: "*A man in a black jacket is doing tricks on a motorbike*"
	B: "*A person is riding the bicycle on one wheel*"

were explicitly asked to assume that both sentences referred to the same situation or event. As we shall see, this instruction – crucial for correctly interpreting the sentences and judging their entailments – was not always followed.

Once all the annotations were collected, the gold labels were calculated with two different methodologies. For each pair, the relatedness gold score was computed as the average of the 10 ratings assigned by the participants. Table 3 shows examples of sentence pairs with different degrees of semantic relatedness. As a measure of (inverse) inter–rater agreement, we computed the average of the standard deviation of relatedness scores for each sentence pair, resulting in $SD = 0.84$.[4] This means that, on average, participants' judgments varied ± 0.84 rating points around the final score assigned to each pair.

With regards to entailment gold labels, a majority vote schema was adopted. Pairs were classified as CONTRADICTION when most participants indicated that "if sentence A is true, sentence B is false" in both presentation orders; pairs were classified as ENTAILMENT when most participants indicated that "if sentence A is true, sentence B is true" for the corresponding presentation order; the remaining pairs were classified as NEUTRAL. Thus, a pair was classified as NEUTRAL in two conditions: when most participants indicated that "the truth of B cannot be determined on the basis of A" for the corresponding presentation order, and when the majority label was CONTRADICTION only in one presentation order. Table 4 shows examples of sentence pairs with different entailment relations. Inter–rater agreement for the entailment task was 0.87,

[4] Inter–rater agreement figures given in this paper for both relatedness and entailment slightly differ from those reported in (Marelli et al, 2014b), due to a small bug that has been fixed.

Table 5 Distribution of sentences pairs across the two tasks according to the relatedness and entailment annotations.

relatedness	NEUTRAL		CONTRADICTION		ENTAILMENT		TOTAL	
1-2 range	922	(10%)	0	(0%)	1	(0%)	923	(10%)
2-3 range	1253	(13%)	118	(1%)	2	(0%)	1373	(14%)
3-4 range	**2742**	(28%)	**994**	(10%)	136	(1%)	3872	(39%)
4-5 range	678	(7%)	312	(3%)	**2682**	(27%)	3672	(37%)
TOTAL	5595	(58%)	1424	(14%)	2821	(28%)	9840	(100%)

computed as the average proportion of the majority vote across pairs and indicating that, as an average, 87% of participants agreed with the majority vote in each pair.

In the relatedness task, participants' ratings present a certain degree of variability. One may thus challenge the reliability of these data, claiming that the responses are largely random. Certainly, this seems unlikely for the overall data set, but may be more reasonable when considering the ratings in specific entailment classes. In CONTRADICTION pairs, for example, the relatedness-response variance is 1.198, as opposed to the 0.718 variance of the full data set. To clarify this point we ran a simulation study to estimate the rating distribution under a random-response assumption. We computed the variance of 10 points randomly sampled from the 1-5 range (representing participants' responses to a given pair) in 10,000 examples (i.e., the number of pairs in the whole data set). The range of the resulting average variances, based on a total of 20,000 of these simulations, was from 1.965 to 2.032. That is, if responses were given at the chance level on a 5-point scale in a data set of this size, we would expect the average variance value to be between 1.965 and 2.032.[5] The observed variance values are out of this interval, indicating that the distribution of the actual responses is virtually impossible at the chance level, and supporting the reliability of participants' responses.

2.3 Data set statistics

The resulting data set is presented in Table 5, which shows the gold data when considering the relatedness and entailment results together: each cell in the table reports the number of sentence pairs for each combination between relatedness classes and entailment labels. As the table shows, the great majority of pairs in the ENTAILMENT relation are highly related (2682/2821 pairs are in the highest relatedness range), and also CONTRADICTION pairs have high relatedness scores (994/1424 contradictions are in the 3-4 range).

A further analysis of the gold relatedness scores and entailment labels was carried out to investigate how they are distributed across the various pair types. The relatedness score distribution is summarized in Table 6. The *Norm-Sim* pairs (similar meaning) were judged to be maximally related, followed by *Norm-Contr* (contrasting) and *Norm-Diff* (lexical overlap only). This confirms what can be gleaned from Table 5: pairs conveying an opposite/contrasting meaning (*Norm-Contr*) are judged as more related than pairs that have no strong meaning relation but contain the same words (*Norm-Diff*). The high relatedness scores obtained for opposite sentences also highlights the difference between using similarity and relatedness (see Section 1). This trend

[5] A comparable variance range is obtained running the same simulation on the number of CONTRADICTION pairs (1424).

Table 6 Average relatedness scores (and corresponding standard deviations) for each pair type.

Type of pair	Average relatedness (SD)
Norm-Sim same set	4.65 (.29)
Norm-Contr same set	3.59 (.44)
Norm-Diff same set	3.40 (.57)
Norm-Norm cross set	3.82 (.70)
Norm-Sim cross set	3.75 (.70)
Norm-Contr cross set	3.15 (.65)
Norm-Diff cross set	2.94 (.66)
Unrelated pairs	1.78 (.85)

Table 7 Distribution of entailment annotations across pair types.

Type of pair	ENTAILMENT	CONTRADICTION	NEUTRAL
Norm-Sim same set	94.2%	0%	5.8%
Norm-Contr same set	0.9%	58.2%	40.9%
Norm-Diff same set	9.8%	1.9%	88.3%
Norm-Norm cross set	37.8%	0.2%	62%
Norm-Sim cross set	35%	0%	65%
Norm-Contr cross set	2.5%	16.9%	80.6%
Norm-Diff cross set	4%	0%	96%
Unrelated pairs	0.9%	0.3%	98.8%

could be observed both when comparing sentences belonging to the same expansion set, that is, originating from the same source sentence (*Norm_a-Sim_a, Norm_b-Sim_b, Norm_a-Contr_a, Norm_b-Contr_b, Norm_a-Diff_a, Norm_b-Diff_b*), and in pairs containing sentences from different sets (*Norm_a-Sim_b, Norm_b-Sim_a, Norm_a-Contr_b, Norm_b-Contr_a, Norm_a-Diff_b, Norm_b-Diff_a*), although the latter case is characterized by generally lower ratings and higher variance. This was expected, since the STS source pairs were already capturing some degree of relatedness. Unrelated pairs were assigned the lowest average ratings.

Distributions of the entailment labels are reported in Table 7 (percentage of assigned label to pair type). The results generally match our expectations when considering pairs of sentences from the same expansion set. The ENTAILMENT label is mostly assigned in case of *Norm-Sim* pairs (similar meaning), the CONTRADICTION label in case of *Norm-Contr* pairs (contrast/contradiction), and the NEUTRAL label in case of *Norm-Diff* pairs (lexical overlap only). We observe however a relatively high proportion of *Norm-Contr* pairs labeled NEUTRAL. Inspection of the NEUTRAL *Norm-Contr* pairs reveals a significantly higher incidence of pairs of sentences where subjects contained indefinite articles (72% vs. 19% in the CONTRADICTION pairs). Our explanation is that for these pairs - despite the specific instructions requiring to assume that both sentences refer to the same entity, situation or event - subjects tend to think that, for instance, *"A woman is wearing an Egyptian headdress"* does not contradict *"A woman*

is wearing an Indian headdress", since one could easily imagine both sentences truthfully uttered in a single scene, where two different women are wearing different headdresses. In the future, a higher proportion of CONTRADICTION labels could be elicited by using grammatical and possibly visual cues (pictures), to encourage co-indexing of the entities in the two sentences.

We observe a weaker link between expected and assigned labels among the cross-set pairs, as most of the pairs belong to the NEUTRAL group. The influence of pair type can still be observed, though: ENTAILMENT is assigned to 35% of the *Norm-Sim* cross-set pairs, whereas CONTRADICTION is assigned to 16.9% of the *Norm-Contr* cross-set pairs. The preponderance of NEUTRAL is not surprising either, as in the cross-set condition the original pairs were already different to start with. The transformation process brought them further apart, making it less likely that the new pairs would describe situations similar enough to trigger contradiction/contrast intuitions (indeed, we observed above that for the cross-set cases we have lower relatedness ratings).

3 SemEval 2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment

The SICK-based SemEval challenge[6] involved two subtasks: predicting a continuous *relatedness* score for SICK sentence pairs (a direct way to evaluate the extent to which CDSMs can quantify the degree of semantic relatedness between sentences) and (ii) stating their *entailment* status (ENTAIL, CONTRADICT or NEUTRAL).

3.1 SICK for SemEval

For the purpose of the tasks, the data set was randomly split into training and test set (50% and 50%), ensuring that each relatedness range and entailment category was equally represented in both sets. Table 8 shows the distribution of sentence pairs considering the combination of relatedness ranges and entailment labels. The "TOTAL" column indicates the total number of pairs in each range of relatedness, while the "TOTAL" row contains the total number of pairs in each entailment class.

3.2 Evaluation metrics and baselines

Both subtasks were evaluated using standard metrics. In particular, the official measures chosen to rank the participating systems were *Pearson correlation* (r) for the relatedness subtask and *accuracy* for the entailment subtask. Furthermore, systems' results were additionally evaluated using Spearman correlation and Mean Squared Error (MSE) for relatedness and Precision, Recall, and F measure for entailment.

Table 9 presents the performance of 4 baselines. The Majority baseline always assigns the most common label in the training data (NEUTRAL), whereas the Probability baseline assigns labels randomly according to their relative frequency in the training set. The Overlap baseline measures word overlap, again with parameters (number of stop words and ENTAILMENT/NEUTRAL/CONTRADICTION thresholds) estimated on the training part of the data. Since the Majority and Probability baselines require discrete distributions, they cannot be computed for the relatedness task, where continuous scores are used. The code for computing the

[6] http://alt.qcri.org/semeval2014/task1/

Table 8 Distribution of sentence pairs across the Training and Test Sets according to the relatedness and entailment annotations.

SICK Training Set								
relatedness	CONTRADICTION		ENTAILMENT		NEUTRAL		TOTAL	
1-2 range	0	(0%)	0	(0%)	471	(10%)	471	(10%)
2-3 range	59	(1%)	2	(0%)	638	(13%)	699	(14%)
3-4 range	498	(10%)	71	(1%)	1344	(27%)	1913	(38%)
4-5 range	155	(3%)	1344	(28%)	352	(7%)	1851	(38%)
TOTAL	712	(14%)	1417	(29%)	2805	(57%)	4934	(100%)
SICK Test Set								
relatedness	CONTRADICTION		ENTAILMENT		NEUTRAL		TOTAL	
1-2 range	0	(0%)	1	(0%)	451	(9%)	452	(9%)
2-3 range	59	(1%)	0	(0%)	615	(13%)	674	(14%)
3-4 range	496	(10%)	65	(1%)	1398	(28%)	1959	(39%)
4-5 range	157	(3%)	1338	(28%)	326	(7%)	1821	(38%)
TOTAL	712	(14%)	1404	(29%)	2790	(57%)	4906	(100%)

Table 9 Performance of baselines. Figure of merit is Pearson (r) correlation for relatedness and accuracy for entailment. NA = *Not Applicable*.

Baseline	Relatedness (r)	Entailment (accuracy)
Chance	0	33.3%
Majority	NA	56.7%
Probability	NA	41.8%
Overlap	0.63	56.2%

baselines - including full documentation - is freely available and can be downloaded from the SemEval website.[7]

3.3 Submitted runs and results

Overall, 21 teams participated in the task. Participants were allowed to submit up to 5 runs for each subtask and had to choose the primary run to be included in the comparative evaluation. We asked participants to pre-specify a primary run to encourage commitment to a theoretically-motivated approach, rather than a post-hoc performance-based assessment. Interestingly, some participants used the non-primary runs to explore the performance level that could be reached by exploiting weaknesses in the data that are not likely to hold in future tasks of the same kind (for instance, the non-primary run 3 submitted by The Meaning Factory exploited sentence ID ordering information). Participants also used non-primary runs to test smart baselines.

[7] http://alt.qcri.org/semeval2014/task1/index.php?id=data-and-tools.

Table 10 Statistics for the relatedness and entailment subtasks. Relatedness values indicate the Pearson (r) correlation while the entailment is shown as the percentage of accuracy.

	Relatedness (r)		Entailment (accuracy)	
	All runs	Primary runs	All runs	Primary runs
Highest	0.842	0.828	84.6%	84.6%
Median	0.713	0.714	75.7%	77.1%
Average	0.707	0.719	74.7%	75.4%
Lowest	0.412	0.479	48.7%	48.7%

We received 17 primary submissions to the relatedness subtask (for a total of 66 runs) and 18 to the entailment subtask (65 runs). Table 10 gives some overall statistics about the task results, calculated both *(i)* over all the submitted runs and *(ii)* considering only the primary run of each participating group. In the relatedness subtask, 6 non-primary runs slightly outperformed the

Table 11 Table left: Primary run results for the relatedness subtask. Official ranking metric: Pearson correlation (r); additional metrics: Spearman correlation (ρ) and Mean Squared Error (MSE). Table right: Primary run results for the entailment subtask according to the official ranking measure Accuracy. Systems that are significantly better with respect to the next-highest ranked system at p-value$\leq$0.05 are marked with *. The table also shows whether a system exploits composition information at either the phrase (P) or sentence (S) level, see Section 3.4 for a detailed explanation.

Relatedness Task					Entailment Task		
ID	Comp.	r	ρ	MSE	ID	Comp.	Acc. (%)
ECNU_run1	S	0.828	0.769	0.325	Illinois-LH_run1	P/S	84.6%
StanfordNLP_run5	S	0.827	0.756	0.323	ECNU_run1	S	83.6%
The_Meaning_Factory_run1	S	0.827*	0.772	0.322	UNAL-NLP_run1		83.1%
UNAL-NLP_run1		0.804	0.746	0.359	SemantiKLUE_run1		82.3%
Illinois-LH_run1	P/S	0.799*	0.754	0.369	The_Meaning_Factory_run1	S	81.6%*
CECL_ALL_run1		0.780	0.732	0.398	CECL_ALL_run1		80.0%
SemantiKLUE_run1		0.780*	0.736	0.403	BUAP_run1	P	79.7%
RTM-DCU_run1		0.764*	0.688	0.429	UoW_run1		78.5%
UTexas_run1	P/S	0.714	0.674	0.499	Uedinburgh_run1	S	77.1%
UoW_run1		0.711	0.679	0.511	UIO-Lien_run1		77.0%
FBK-TR_run3	P	0.709	0.644	0.591	FBK-TR_run3	P	75.4%
BUAP_run1	P	0.697	0.645	0.528	StanfordNLP_run5	S	74.5%
UANLPCourse_run2	S	0.693*	0.603	0.542	UTexas_run1	P/S	73.2%*
UQeResearch_run1		0.642	0.626	0.822	Yamraj_run1		70.7%
ASAP_run1	P	0.628*	0.597	0.662	asjai_run5	S	69.8%
Yamraj_run1		0.535*	0.536	2.665	haLF_run2	S	69.4%*
asjai_run5	S	0.479	0.461	1.104	RTM-DCU_run1		67.2%*
					UANLPCourse_run2	S	48.7%

official winning primary entry, namely: The_Meaning_Factory's run3 (Pearson 0.842), ECNU's run2 (0.839) and run5 (0.835), and StanfordNLP's run4 (0.835) and run2 (0.831). In the entailment subtask, the first ranked run was indeed primary. However, the second-top primary run by the ECNU team was preceded by two non-primary runs by the same team.

The official ranking of primary runs are presented in Table 11.[8] For the relatedness subtask, the table reports Pearson (r) correlation results as well as additional results calculated using Spearman (ρ) correlation and Mean Squared Error (MSE). For the entailment subtask, results are reported in terms of Accuracy. We can see that in both tasks most systems performed well above the best baselines from Table 9. As for the top-scoring systems, we witnessed a very close finish in both subtasks, with 4 systems within 3 percentage points from the first-ranked one in both cases. 4 of these 5 top systems were the same across the two subtasks. To better understand the SemEval results, Table 11 also reports when the difference between an adjacent pairing of systems is statistically significant. For the relatedness task we applied the Fisher's *r-to-z* transformation, while for entailment we used the chi-squared test (two-tailed tests). Differences were considered statistically significant at p-value$\leq$0.05. Note that significance was not calculated for all pairs of systems, but only between systems that are adjacent in the ranking, i.e. we started from the best scoring system and evaluated each run only with respect to the following one in the rank. We can see that, both for relatedness and entailment, the difference between the top-scoring systems are not statistically significant, meaning that it is impossible to draw a clear individual "winner" for the evaluation exercise.

As regards the entailment task, Table 12 presents additional results in terms of Precision, Recall, F-measure for each class. We can see that, overall, systems achieve the best results in detecting the NEUTRAL pairs, while the identification of the ENTAILMENT pairs appears to be the most challenging.

3.4 Approaches

A summary of the approaches used by the systems to address the task is presented in Table 13. In the table, systems in bold are those for which the authors submitted a paper: haLF (Ferrone and Zanzotto, 2014), The Meaning Factory (Bjerva et al, 2014), UTexas (Beltagy et al, 2014), Illinois-LH (Lai and Hockenmaier, 2014), ASAP (Alves et al, 2014), BUAP (León et al, 2014), CECL (Bestgen, 2014), ECNU (Zhao et al, 2014), FBK-TR (Vo et al, 2014), RTM-DCU (Biçici and Way, 2014), UIO-Lien (Lien and Kouylekov, 2014), UNAL-NLP (Jimenez et al, 2014), SemantiKLUE (Proisl and Evert, 2014) and UoW (Gupta et al, 2014). For the others, we used the brief description sent with the system's results, double-checking the information with the authors. In the table, "E" and "R" refer to the entailment and relatedness task respectively, and "B" to both.

Almost all systems combine several kinds of features. To highlight the role played by composition, we draw a distinction between 'compositional' and 'non-compositional' features, to be understood as follows. The standard definition of compositionality (see e.g. Pagin and Westersthl (2010)) is that the meaning of a complex expression is a function of the meaning of its immediate constituents and their syntactic relations. Since not all compositional systems reach the final sentential level, we use the term "phrase compositional" to refer to systems that stop their composition at the level of phrases, and we use the term "sentence- compositional" to refer to systems which compute a meaning for the whole sentence, but don't necessarily assign a meaning to intermediate syntactic units. Among the non-compositional features we count word

[8] ITTK's primary run could not be evaluated due to technical problems with the submission. The best ITTK's non-primary run scored 0.76 r in the relatedness task and 78.2% accuracy in the entailment task.

Table 12 Primary run results for the entailment subtask in terms of Precision (P), Recall (R), F-measure (F1) on Entailment, Contradiction, and Neutral classes.

	Entailment			Contradiction			Neutral		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Illinois-LH_run1	78.1	**82.7**	**80.3**	89.5	**80.3**	**84.6**	**86.9**	86.6	**86.8**
ECNU_run1	81.0	72.8	76.7	91.3	77.4	83.8	83.2	90.7	**86.8**
UNAL-NLP_run1	76.9	75.7	76.3	93.7	78.9	85.7	83.8	87.8	85.8
SemantiKLUE_run1	80.9	66.9	73.2	94.6	75.0	83.7	80.6	92.0	86.0
The Meaning Factory_run1	92.2	60.8	73.3	96.7	60.3	74.3	76.9	97.6	86.0
CECL_ALL_run1	67.6	80.5	73.5	93.0	73.9	82.4	85.0	81.3	83.1
BUAP_run1	75.1	75.0	75.1	77.3	79.0	78.2	82.6	82.2	82.4
UoW_run1	71.3	65.9	68.5	91.1	71.3	80.0	79.3	86.8	82.9
Uedimburgh_run1	66.4	68.6	67.5	86.3	79.9	83.0	80.5	80.7	80.6
UIO-Lien_run1	93.7	42.9	58.8	84.3	73.6	78.6	72.8	95.2	82.5
FBK-TR_run3	59.4	86.8	70.5	90.2	74.4	81.6	86.1	69.9	77.2
StanfordNLP_run5	66.3	64.1	65.2	76.6	71.5	74.0	77.8	80.5	79.1
UTexas_run1	**98.0**	31.5	47.7	**97.7**	52.8	68.5	68.1	**99.6**	80.9
Yamaraj_run1	69.1	32.3	44.0	55.7	61.1	58.3	74.5	92.7	82.6
asjai_run5	56.8	49.3	52.8	72.7	50.0	59.3	74.3	85.2	79.4
haLF_run2	55.2	63.8	59.2	85.3	64.4	73.4	74.7	73.5	74.1
RTM-DCU_run1	56.7	68.2	62.0	undef.	0.0	undef.	72.7	84.0	78.0
UANLPCourse_run2	43.2	60.1	50.3	9.2	15.8	11.7	83.4	51.5	63.7

overlap, word similarity, the presence of negative markers, etc. (see Table 13 for a schema of the various features; for details, we refer the reader to the articles cited above). As the table shows, thirteen systems used composition in at least one of the tasks; ten used composition for full sentences and six for phrases, only. The best systems are among these thirteen systems. (See also Table 11 which summarizes the results for the two tasks and reports also the information about sentence-compositional (S) and phrase-compositional (P) features.)

Given our more general interest in the distributional approaches, in Table 13 we also classify the different DSMs used as 'Vector Space Models', 'Topic Models' and 'Neural Language Models'. Furthermore, the different learning approaches used by the systems are reported, and the table shows that the most adopted methods are SVM and Kernel methods.

Finally, Table 13 reports on the use of external resources in the task. One of the reasons to create SICK was to have a compositional semantics benchmark that would not require too many external tools and resources (e.g., named-entity recognizers, gazetteers, ontologies). Judging from what the participants chose to use we think we succeeded, as only standard NLP pre-processing tools (tokenizers, PoS taggers and parsers) and relatively few resources (mostly, WordNet and paraphrase corpora) were used.

In general, several participating systems deliberately exploit *ad-hoc* features that, while not helping a true understanding of sentence meaning, exploit some systematic characteristics of SICK that should be controlled for in future releases of the data set. In particular, the Textual Entailment subtask has been shown to rely too much on negative words and antonyms. The

Table 13 Summary of the main characteristics of the participating systems. The table indicates whether a feature or method is used by a system for either the Relatedness Task (R), the Entailment Task (E) or both tasks (B). Participants marked in bold are those who submitted the system description paper. The last two columns report the rank of each system for each of the two tasks.

	Non Composition Features								Comp Features		Learning Methods							External Resources						
Participant ID	Vector Semantics Model	Topic Model	Neural Language Model	Word Overlap	Word Similarity	Syntactic Features	Sentence difference	Negation Features	Sentence Composition	Phrase composition	SVM and Kernel methods	K-Nearest Neighbours	Classifier Combination	Random Forest	FoL/Probabilistic FoL	Curriculum based learning	Other	WordNet	Paraphrases DB	Other Corpora	ImageFlicker	STS MSR-Video Description	Relatedness Task Ranking	Entailment Task Ranking
ASAP	R	R			R	R	R	R		R			R					R					15	-
ASJAI	B		B	B	B	B	B		B		E		B				R		B				17	15
BUAP	B		B		B	B		E		B	E							B					12	7
UEdinburgh	B			B		B		B	B		E	R								B			-	9
CECL	B			B	B		B										B			B			6	6
ECNU	B		B	B	B	B	B		B		B	B	B	B				B		B			1	2
FBK-TR		R		R	R	E	B	E	E	B	R		E					R		R		E	11	11
haLF	E				E				E		E												-	16
IITK	B			B	B	B	B	B		B	B							B					-	-
Illinois-LH	B			B	B	B		B	B	B							B	B			B	B	5	1
RTM-DCU	B					B					B			B						B			8	17
SemantiKLUE	B			B	B	B		B			B							B		B			7	4
StandfordNLP	B		B	R	R			R	B							E							2	12
The Meaning Factory	R		R	R		R	R		B		E			R	E			B	B	R			3	5
UANLPCourse	B			B	B				B		B												13	18
UIO-Lien								E										E					-	10
UNAL-NLP					B	B		B									B	R	B	B			4	3
UoW				B		B		B			B								B				10	8
UQeRsearch				R	R	R	R	R									R	R					14	-
UTexas	B							B	B	B					B					B			9	13
Yamarj	B			B	B						B												16	14

Illinois-LH team reports that just by checking the presence of negative words (the Negation Feature in the table) one can detect 86.4% of the contradiction pairs, and by combining Word Overlap and antonyms one can detect 83.6% of neutral pairs and 82.6% of entailment pairs. This approach, however, is obviously very brittle (it would not have been successful, for instance, if word-rearranging had been optionally combined with negation in the creation of Diff sentences, see Section 2.1 above).

3.5 Lesson learned from SemEval and open questions

As explained in the introduction, SICK was built with the purpose of providing a suitable benchmark to evaluate computational semantic systems on their ability to reach meaning representations of sentences compositionally. To this effect, SICK sentences exhibit many cases of lexical variation phenomena, active/passive and other syntactic alternations, impact of negation and conjunction at various levels and other variations linked to the functional lexicon – all issues that do not occur frequently in existing large data sets of sentences.

SemEval 2014 was our test bed to see whether the way in which participants tackled SICK reflected this goal. Therefore, we especially encouraged developers of compositional (distributional) semantic models to test their methods on SICK, though we welcomed developers of other kinds of systems that could tackle sentence relatedness or entailment tasks (e.g., full-fledged RTE systems).

First of all, it is interesting to compare the results obtained in our evaluation exercise with those reported for other similar tasks offered to the community. For relatedness, we can compare our task to the similarity task run on MSRvid (one of our data sources) at STS 2012 – though the notion of 'relatedness' differs from that of 'similarity' (see Section 1). Judging from the top-scoring results, the relatedness task appears to be more challenging, as the best STS 2012 Pearson correlation was $r = 0.88$, against $r = 0.84$ achieved by our participants. On the other hand, if we consider the average score of all the systems, the performances of the two tasks are aligned, with $r = 0.70$ on STS 2012 and $r = 0.71$ on SICK. As for the entailment task, we can compare our task to the RTE Challenges. In particular, the most similar tasks are RTE-4 and RTE-5 (Bentivogli et al, 2009), since these data sets are annotated with the same three entailment classes as SICK. Though the results of the tasks are not really comparable because the RTE data sets are balanced with respect to the entailment classes while SICK is not, a noticeable difference in the Accuracy results can be observed. The RTE-3 and RTE-4 median values are 54.30% and 52.00% respectively, whereas the average values are 50.65% and 52.91%. The entailment scores obtained on the SICK data set are considerably higher – 77.1% for the median system and 75.4% for the average system. This overall performance pattern suggests that, owing perhaps to the more controlled nature of the sentences, as well as to the purely linguistic nature of the challenges it presents, the SICK task is "easier" than the RTE tasks.

Moreover, as noted in the description of the SemEval 2014 submissions, contrary to our expectations no participant submitted purely compositional systems; all systems also exploited non-compositional features. Still, the majority of the systems (13/21) used composition in at least one of the subtasks. However, as discussed in Section 3.4, several participating systems exploited some systematic characteristics of SICK, such as the presence of negative words and antonyms. Based on this observation, we conjectured that most of the systems overfitted the training data, in the sense that they became overly tuned to the limited syntactic structures of the input, but also to the different proportions of the various types of sentence pairs (e.g. the low number of pairs with similar lexicons but different meanings, or completely unrelated meanings). The obvious question, at this point, is whether we can analyze the performance of the systems more in detail, to better understand if it corresponds to a real ability to cope well with all the different phenomena represented in SICK. To answer this question we carried out a new analysis on a subset of SICK where the various conditions were *balanced* (see Section 4.1), under the assumption that a system optimized to deal correctly with the patterns most commonly represented in the data set would not be as successful when these proportions shift.

Furthermore, we conjectured that a purely compositional system would show more stability in its performance across different benchmarks, since this system should be built in a more principled way and its behavior should thus be independent on the specific aspects of the data set against which it is evaluated. To check this conjecture, we contacted three of the best performing systems which also used compositional features (ECNU, The Meaning Factory, and Illinois), and obtained the output from their compositional features alone. Of course, we did not expect high performance from these modules, but rather more stability when evaluated against the whole SICK data set vs. the balanced one. Our last question is about the mistakes made by participating systems. SICK and SemEval can help us shed light on those phenomena which are still challenging for state-of-the art computational semantic models. To this end, we carried out

a qualitative analysis of the pairs mistaken by the majority of the systems in the two subtasks (see Section 4.2).

4 Learning more about SemEval systems: new analyses on SICK data

In this section we address the questions we raised based on the SemEval experience. In Section 4.1 we look at the stability issue: (a) can we analyse in greater detail the performance of the systems to better understand if they correspond to a true ability to cope well with all the different phenomena represented in SICK? And (b) is it true that a purely compositional system would show more stability than a system that exploits other ad-hoc features when evaluated against the whole SICK data set vs. the balanced one? In Section 4.2, we scrutinize the systems' results in order to detect which semantic phenomena current state-of-the-art systems fail to capture.

4.1 Systems' results on SICK balanced data

As described in Section 2, SICK was created with the goal of offering a data set of sentence pairs where each sentence is contrasted with either a (near) paraphrase (*Norm-Sim*), a contradictory or strongly contrasting statement (*Norm-Contr*), another sentence with very high lexical overlap but different meaning (*Norm-Diff*), or a completely unrelated sentence (*Unrel*). The sentence pairing mechanism adopted led to the creation of 8 classes of pairs addressing different linguistic phenomena, namely *Norm-Sim* same set, *Norm-Contr* same set, *Norm-Diff* same set, *Norm-Norm* cross set, *Norm-Sim* cross set, *Norm-Contr* cross set, *Norm-Diff* cross set, *Unrelated* (see Tables 6 and 7). Since these classes are not uniformly distributed in SICK, we created a balanced subset of the SICK test set, which is composed of 150 sentence pairs for each of the 8 classes, for a total of 1,200 pairs,[9] and we used it to carry out an additional evaluation of the primary runs of the participating systems. The purpose of this new evaluation is to better understand if the performance achieved on the full test set correspond to a real ability of the systems to cope well with all the different phenomena represented in SICK or are due to the fact that the systems adjusted to the unbalanced data set, giving more weight to more common phenomena.

Furthermore, to investigate the hypothesis that a purely compositional system would show more stability in its performance across different benchmarks, we also evaluated the compositional runs contributed by ECNU, The Meaning Factory, and Illinois for the relatedness task and by ECNU and Illinois for the entailment task.

The outcome of this evaluation is presented in Tables 14 and 15, where the official results on the full SemEval test set are given together with the new results on the balanced subset and the difference between the two. For the relatedness task, Table 14 shows a drop across all the systems, demonstrating that they all overfit the training data. Interestingly, when we compare the drops obtained by the compositional runs with their corresponding full-systems (e.g., The_Meaning_Factory_compositional_run v.s. The_Meaning_Factory_run1) we see that the compositional modules are slightly more stable, suggesting that the other features of the systems are more ad-hoc than the compositional one. For the entailment task the situation is somewhat different, as Table 15 shows that 4 full systems out of 18 increase their performance. Again, quite interestingly, the compositional feature of Illinois increases its performance while the full system drops; similarly the compositional module of ECNU has a slightly lower drop than its full system.

[9] Despite the fact that SICK test set contains a total of 4906 sentence pairs, we could not create a larger balanced test set. Each class had to be composed of only 150 pairs since the Norm-Diff cross set class is very small, containing only 168 pairs.

A deeper analysis of the 4 full systems which improved their performance has been carried out both on the full test set and the balanced subset, in order to observe the distribution of results among the 8 sentence pair classes. The analysis of the "full test" results showed that all the systems perform well on phenomena that are over-represented in the data set, namely the "Norm-Diff cross set" and "Unrelated" classes (and also the "Norm-Diff same set" class for UIO-Lien and Utexas systems). The analysis of the "balanced subset" results showed that the accuracy proportion in the various classes is similar to that of the full test set, confirming that the overall improvement in performance on the balanced subset is due to these systems being more adept at dealing with the over-represented classes.

4.2 Qualitative analysis of the most difficult sentence pairs

In order to gain a deeper insight into the systems' difficulties when approaching our tasks and data set, we carried out a qualitative analysis of the most difficult sentence pairs in the SICK test

Table 14 Relatedness Task: systems' Pearson correlation on the SICK balanced test set compared to results obtained on the whole test set. Purely compositional runs and corresponding full systems are in bold and marked with the same symbol.

	r		
ID	Full Data set	Balanced Data set	Variation
asjai_run5	0.479	0.473	-0.006
Yamraj_run1	0.535	0.515	-0.020
The_Meaning_Factory_*compositional*_run ★	**0.608**	**0.583**	**-0.025**
RTM-DCU_run1	0.764	0.734	-0.030
UANLPCourse_run2	0.693	0.658	-0.035
StanfordNLP_run5	0.827	0.787	-0.040
ASAP_run1	0.628	0.586	-0.042
The_Meaning_Factory_run1 ★	0.827	0.783	-0.044
ECNU_*compositional*_run ■	**0.754**	**0.701**	**-0.053**
UTexas_run1	0.714	0.660	-0.054
UQeResearch_run1	0.642	0.585	-0.057
Illinois_*compositional*_run ♣	**0.463**	**0.397**	**-0.066**
CECL_ALL_run1	0.78	0.711	-0.069
SemantiKLUE_run1	0.78	0.711	-0.069
ECNU_run1 ■	0.828	0.758	-0.070
UNAL-NLP_run1	0.804	0.734	-0.070
BUAP_run1	0.697	0.625	-0.072
FBK-TR_run3	0.709	0.633	-0.076
Illinois-LH_run1 ♣	0.799	0.719	-0.080
UoW_run1	0.711	0.618	-0.093

Table 15 Entailment Task: systems' results on the SICK balanced test set compared to results obtained on the whole test set. Purely compositional runs and corresponding full systems are in bold and marked with the same symbol.

	Accuracy (%)		
ID	Full Data set	Balanced Data set	Variation
RTM-DCU_run1	67.2	70.4	+3.2
asjai_run5	69.8	72.8	+3.0
UTexas_run1	73.2	76.1	+2.9
UIO-Lien_run1	77.0	78.3	+1.3
Illinois_*compositional*_run ♣	**65.0**	**65.6**	**+0.6**
The_Meaning_Factory_run1	81.6	81.3	-0.3
Uedinburgh_run1	77.1	76.5	-0.6
Yamraj_run1	70.7	69.7	-1.0
UANLPCourse_run2	48.7	47.3	-1.4
StanfordNLP_run5	74.5	72.8	-1.7
UNAL-NLP_run1	83.1	81.0	-2.1
ECNU_*compositional*_run ■	**72.9**	**70.6**	**-2.3**
FBK-TR_run3	75.4	73.0	-2.4
UoW_run1	78.5	76.0	-2.5
SemantiKLUE_run1	82.3	79.7	-2.6
BUAP_run1	79.7	77.0	-2.7
ECNU_run1 ■	83.6	80.8	-2.8
haLF_run2	69.4	66.0	-3.4
Illinois-LH_run1 ♣	84.6	79.5	-5.1
CECL_ALL_run1	80.0	74.7	-5.3

set, i.e. those which were not not judged correctly by the majority of the systems. In particular, we looked at pairs misjudged by more than 9 systems (viz., evaluated correctly by at most 8 systems). To work with cleaner data, we kept only those pairs which had low variance in the crowdsourced judgments. This human-based criterion ensures that pair difficulty cannot be trivially explained by ambiguities in the corresponding sentences, since this would result in higher disagreement between annotators. In other words, the meanings of the pairs analysed below are clear according to native speakers' intuitions, so the challenge they present to automated systems must depend on the linguistic phenomena they contain. 'Low-variance pairs' were defined by means of quartile points. In the relatedness subtask, we removed those pairs where the *SD* in human judgments was above 1.02. In the entailment subtask, we removed those pairs where the agreement on the majority vote label was below 80%. In what follows, we refer to the pairs thus selected as "reliable" data.

Table 16 Entailment Task: statistics about the most difficult pairs in the data set and their distribution with respect to *(i)* the data creation methodology (rows 1 to 8), and the *(ii)* gold entailment annotation (3 last rows). The second column contains the distribution of the 4906 sentences pairs in the SICK test set. Out of these 4906 pairs, 3676 have low variance (lv), and their distribution is represented in the third column. The fourth column indicates the number of the low variance pairs correctly classified by 8 or less systems.

	whole-SICK (tot: 4906)	lv-SICK (tot: 3676)	lv, ok by max 8 syst (tot: 366 – 9.9% of lv)	
Norm-Norm cross set	426	320	46	(14%)
Norm-Sim cross set	823	625	97	(15%)
Norm-Contr cross set	793	517	61	(11.7%)
Norm-Diff cross set	168	146	4	(2.7%)
Norm-Sim same set (paraphrases)	954	797	70	(8.7%)
Norm-Contr same set (negations)	985	616	51	(8.2%)
Norm-Diff same set (scrambled words)	180	119	35	(**29.4%**)
Unrel	577	536	2	(0.3%)
CONTRADICTION	712	584	75	(12.8%)
NEUTRAL	2790	1973	80	(4%)
ENTAILMENT	1404	1119	211	(**18%**)

4.2.1 Entailment Task

Out of the 4906 pairs, 3676 pairs (74.9%) have low variance in the Entailment Task. Details about their distribution among the different classes of pairs are reported in Table 16, which shows the Unrelated and the *Norm-Diff* cross pairs to be the easiest pairs to be judged by humans (536 pairs out of 577 and 146 pairs out of 168, respectively, have low variance). Out of the 3676 pairs with low variance, 366 are evaluated correctly by only 8 systems or less. We take them to be the difficult cases for the Entailment Task. Again, the reliable Unrelated pairs and the reliable *Norm-Diff* cross set pairs are among the easiest for the systems (2/536 and 4/146, respectively, are in the set of difficult pairs). As expected, the *Norm-Diff* same-set pairs (where the sentences in the pair have a high lexical overlap but are not connected by an entailment relation) turn out to be the most difficult: 35 pairs out of the 119 with low variance (29.4%) are misjudged by more than 9 systems. Finally, the majority of these 366 difficult pairs stand in the ENTAILMENT relation (211/366, viz. 57.6%), confirming the results obtained by the Precision, Recall and F-measure evaluation (see Table 12) computed over the full test set. Further details about the distribution of the most difficult sentence pairs for each class are reported in Table 17.

A qualitative analysis of the pairs has been carried out to better understand the characteristics of the sentences for each class. Examples of the data scrutinized are given in Table 18.

Contradiction. We have looked into the systems' answers and found out that most of these sentence pairs (71/75) have been considered as NEUTRAL by the majority of the systems. Most of these sentence pairs involve several linguistic phenomena besides the one that generates the contradiction, making it difficult for systems to recognize that between the sentences in the pairs there is indeed a relation, and that one sentence contradicts the other.

Neutral. The sentence pairs in this class are mostly sentence pairs which share many words though describe different events, such as in the examples presented in Table 18.

Table 17 Entailment Task: for each class, the distribution of the low variance pairs correctly classified by 8 systems with respect to the data creation methodology.

	CONT (Tot: 75)	NEUT (Tot: 80)	ENT (Tot: 211)
Norm-Norm cross set	0	2	44
Norm-Sim cross set	0	7	90
Norm-Contr cross set	47	9	5
Norm-Diff cross set	0	4	0
Norm-Sim same set (paraphrases)	0	7	63
Norm-Contr same set (negation)	27	23	1
Norm-Diff same set (scrambled words)	1	28	6
Unrel	0	0	2

Table 18 Examples of the most difficult pairs in the Entailment Task.

	Sentence A	Sentence B
Contradiction	"*A woman is driving a car and is talking to the man who is seated beside her*"	"*The woman and the man are not travelling by car*"
	"*Runners are competing in a race*	"*Runners are not taking part in the race*"
	"*There is no small guinea pig gnawing and eating a piece of carrot on the floor*"	"*A guinea pig is eating a carrot*"
Neutral	*A person in a black jacket is doing tricks on a motorbike*"	"*A person on a black motorbike is doing tricks with a jacket*"
	"*A man is playing soccer*"	"*A soccer man is playing piano*"
	"*The picture of a man is being taken near a lake*"	"*A man is taking pictures of a lake*"
Entailment	"*A man is talking to a woman* "	"*A man and a woman are speaking*"
	"*Two children and an adult are standing next to a tree limb*"	"*Three people are standing next to a tree limb*"
	"*A man and two women in a darkened room are sitting at a table with candle*"	"*The group of people is sitting in a room which is dim*"
	"*Someone is feeding an animal*"	"*Someone is giving food to an animal*"

Entailment. The analysis of this class is the one most revealing of linguistic phenomena. Among the 211 difficult entailing pairs we have identified two main groups. First, the similarity present in these cases are mostly based on the substitution of a noun (15) or verb (30) with a synonym or a hierarchically related word (e.g. a hypernym or hyponym). We also noticed that in the case of the cross set pairs, the substitution often involves both a noun and the verb. The second interesting fact is that there are several cases of *coordination*. While coordination is common in the SICK data set, the problematic cases specifically require reasoning about its relation with *quantity* (e.g. *two children and an adult* amount to *three people*). Another source of confusion is the *comitative* construction (A_{plur} do P $\Rightarrow$ A do(es) P with another A; if *"some women are dancing and singing"* then *"a woman is dancing and singing with another woman"*).

Table 19 Relatedness Task: statistics about the most difficult pairs in the data set and their distribution with respect to *(i)* the data creation methodology (rows 1 to 8), and *(ii)* the gold relatedness scores (3 last rows) The second column contains the distribution of the 4906 sentences pairs in the SICK test set. Out of these 4906 pairs, 3677 have low variance (lv), and their distribution is represented in the third column. The fourth column indicates the number of the low variance pairs correctly predicsted by 8 or less systems.

	whole-SICK (tot: 4906)	lv-SICK (tot: 3677)	lv and ok by max 8 syst (tot: 275 – 7.4% of lv)	
Norm-Norm cross set	426	383	33	(8.6%)
Norm-Sim cross set	823	745	65	(8.7%)
Norm-Contr cross set	793	456	35	(7.6%)
Norm-Diff cross set	168	127	12	(9.4%)
Norm-Sim same set (paraphrases)	954	950	16	(1.6%)
Norm-Contr same set (negation)	985	419	11	(2.6%)
Norm-Diff same set (word scrambled)	180	127	23	(**18.1%**)
Unrel	577	470	80	(17%)
Rel score $x \leq 2$	473	451	97	(**21.5%**)
Rel score $2 < x < 4.5$	3483	2276	133	(5.8%)
Rel score $4.5 \leq x$	950	950	45	(4.73%)

4.2.2 Relatedness task

Out of 4906 pairs, 3677 pairs (74.9%) have low variance and out of these, 275 are predicted correctly by at most 8 systems. A system prediction in a pair was deemed as correct when the absolute difference between the predicted score and the gold standard was lower than 1. As shown in Table 19, the same-set *Norm-Sim* pairs come out as the easiest pairs to be judged by humans (only 4 of them had high variance). As for the entailment task, the pairs obtained by word scrambling (*Norm-Diff* same-set) turn out to be among the most difficult ones for systems, with only 23/127 (18.1%) reliable pairs guessed correctly by at most 8 systems. To better analyze the set of the 275 difficult pairs, we divided them into three groups based on their gold relatedness score: low related pairs (score $\leq$2), pairs whose score is above 2 but below 4.5, and highly related pairs (score $\geq$4.5). It is interesting to note that the low related pairs, while being quite easy to judge by humans (low variance), are the most difficult pairs for systems. Further details about the distribution of the most difficult sentence pairs for each class are reported in Table 20, while examples of the analyzed data are given in Table 21.

Relatedeness score $\leq$2 (tot: 97). These cases are mostly labeled NEUTRAL (96/97) and mostly belong to the "Unrelated" class. (74/97). Among these 97 pairs we identify two cases: (a) sentence A and B share the same syntactic structure, with only the object (and sometimes the subject) or the verb changing; (b) there is a negation in sentence A or B. These aspects make them superficially similar, leading systems that cannot truly capture their meaning to regard them as highly related (See Table 21 for some examples). Interestingly, even the Unrelated pairs (tot: 74) are of these sort due to the high similarity of the images/videos they describe (they share actions and subjects) and the simplicity of the description (a woman/man/child doing something (playing/running etc.).

Table 20 Relatedness Task: for each relatedness score range, the distribution of the low variance pairs correctly classified by 8 systems with respect to the data creation methodology

	Rel score $x \leq 2$ (tot: 97)	Rel score $2 < x < 4.5$ (tot: 133)	Rel score $4.5 \leq x$ (tot: 45)
Norm-Norm cross set	2	24	7
Norm-Sim cross set	5	41	19
Norm-Contr cross set	12	22	1
Norm-Diff cross set	3	7	2
Norm-Sim same set (paraphrases)	0	3	13
Norm-Contr same set (negation)	0	9	2
Norm-Diff same set (scrambled words)	1	21	1
Unrel	74	6	0

Table 21 Examples of the most difficult pairs in the Relatedness Task.

	Sentence A	Sentence B
Rel score $x \leq 2$	"*A man is playing baseball with a flute*"	"*A man is playing soccer*"
	"*A cat is looking at a store counter*"	"*A dog is looking around*"
	"*Broccoli are being cut by a woman*"	"*A man is cutting tomatoes*"
	"*There is no man playing a game on the grass*"	"*A man is playing the guitar*"
Rel score $2 < x < 4.5$	"*The woman is penciling on eyeshadow*"	"*A woman is putting cosmetics on her eyelid*"
	"*A dog is chasing a ball in the grass*"	"*A dog with a ball is being chased in the grass*"
	"*A man is breaking a wooden hand*" *against a board*"	"*A man is breaking wooden boards with his hand*"
	"*The man is riding a horse*"	"*A horse is riding over a man*"
Rel score $4.5 \leq x$	"*A man is riding on one wheel on a motorcycle*"	"*A person is performing tricks on a motorcycle*"
	"*The man is using a sledgehammer to break*" *a concrete block that is on another man*"	"*A man is breaking a slab of concrete with a sledge hammer*"
	"*Many people are skating in an ice park*"	"*An ice skating rink placed outdoors is full of people*"

Relatedness score between 2 and 4.5 (tot: 133). This class includes a broad range of intermediate relatedness scores, which makes it difficult to highlight regularities or characteristics phenomena. Furthermore, as we can see in Table 20, difficult pairs are not concentrated in some specific type pairs, but are distributed among all the types. However, looking into the systems' responses we found out that – despite the variety of appraches adopted by the participating systems – all of them fail in a similar way, that is, for each case, all systems agree in assigning a lower (or higher) score than the gold score.

Relatedness score ≥4.5 (tot: 45). There are 41 ENTAILMENT, 1 NEUTRAL and 3 CONTRADICTION cases in this group. 40/45 are guessed correctly by a maximum of 8 systems for the Entailment Task as well. They have very little word overlap (e.g., A: "*a man is riding on one wheel on a motorcycle"; B: "a person is performing tricks on a motorcycle"*, see Table 21) or are the same cases discussed for the Entailment Task (Table 18).

5 Conclusion

In this paper we presented new observations on the task "Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Entailment", organized within SemEval-2014 as Task 1. Moreover, we used the SemEval experience to analyse the SICK data set used in the evaluation campaign.

In SemEval Task 1, two subtasks were offered: (i) predicting the degree of relatedness between two sentences, and (ii) detecting the entailment relation holding between them. The task has raised noticeable attention in the community: 17 and 18 submissions for the relatedness and entailment subtasks, respectively, for a total of 21 participating teams. Participation was not limited to compositional models but the majority of systems (13/21) used composition in at least one of the subtasks. Looking at SemEval results we conjectured that most of the participating systems overfitted the training data and that that a purely compositional system would show more stability in its performance across different benchmarks. We tried to verify these conjectures by evaluating the systems and the purely compositional modules which were part of the best performing systems against a balanced subset of SICK test set. The results were suggestive, though not conclusive.

Next, we used SICK and SemEval to shed light on the issue of which phenomena are still challenging for state-of-the-art computational semantic models. Both for the entailment and the relatedness task, pairs generated by words scrambling turned out to be among the most difficult ones, confirming the failure of systems to truly capture sentence meaning. Interestingly, we found out that most of the systems fail to handle cases involving coordination, specifically those cases that require reasoning about its relation with *plurality* (e.g. *two children and an adult* amount to *three people*) and *groups* (e.g. *Two teams are competing ...* vs. *Two groups of people are playing ...*), but also cases of verbs or verb phrases with similar meanings beyond lexical similarity (e.g. *riding on one wheel* vs. *performing tricks*). More generally, it seems to us that the strategy of artificially creating meaningful pairs generated with word scrambles, argument inversions and lexical substitutions (as in our *Norm-Diff*) is a promising avenue to create new data sets that can push the field of computational semantic toward a deeper, more flexible analysis of human language. We believe that this strategy would be further enhanced if the range of constructions is increased (including e.g. parenthetical expressions, comitatives, multiple quantifiers, comparatives) and if negative words are added to the mix, foiling simple word-based strategies, which—despite their apparent good success/complexity rate— ultimately propose a deeply unsatisfactory model of what computational semantics should be.

Acknowledgments

We thank the creators of the ImageFlickr, MSR-Video, and SemEval-2012 STS data sets for granting us permission to use their data for the task. The University of Trento authors were supported by ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

Agirre E, Cer D, Diab M, Gonzalez-Agirre A (2012) SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In: Proceedings of SemEval 2012: the Sixth International Workshop on Semantic Evaluation

Alves AO, Ferrugento A, Lorenço M, Rodrigues F (2014) ASAP: Automatic Semantic Alignment for Phrases. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Baroni M, Zamparelli R (2010) Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In: Proceedings of EMNLP

Beltagy I, Roller S, Boleda G, Erk K, Mooney RJ (2014) UTexas: Natural Language Semantics using Distributional Semantics and Probablisitc Logic. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Bentivogli L, Dagan I, Dang HT, Giampiccolo D, Magnini B (2009) The Fifth PASCAL Recognizing Textual Entailment Challenge. In: Proceedings of The Text Analysis Conference

Bestgen Y (2014) CECL: a new baseline and a non-compositional approach for the SICK benchmark. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Biçici E, Way A (2014) RTM-DCU: Referential Translation Machines for Semantic Similarity. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Bjerva J, Bos J, van der Goot R, Nissim M (2014) The Meaning Factory: Formal Semantics for Recognizing Textual Entailment and Determining Semantic Similarity. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Budanitsky A, Hirst G (2006) Evaluating WordNet-based measures of lexical semantic relatedness. Computational Linguistics 32(1):13–47

Chen DL, Dolan WB (2011) Collecting highly parallel data for paraphrase evaluation. In: Proceedings of ACL

Dagan I, Glickman O, Magnini B (2006) The PASCAL Recognising Textual Entailment challenge. In: Machine Learning challenges. Evaluating predictive uncertainty, visual object classification, and Recognising Textual Entailment, Springer, pp 177–190

Ferrone L, Zanzotto FM (2014) haLF:Comparing a pure CDSM approach and a standard ML system for RTE. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Gabrilovich E, Markovitch S (2007) Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of IJCAI

Grefenstette E, Sadrzadeh M (2011) Experimental support for a categorical compositional distributional model of meaning. In: Proceedings of EMNLP

Gupta R, Hannah Bechara IEM, Orasăn C (2014) UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research pp 853–899

Jimenez S, Duenas G, Baquero J, Gelbukh A (2014) UNAL-NLP: Combining Soft Cardinality Features for Semantic Textual Similarity, Relatedness and Entailment. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Lai A, Hockenmaier J (2014) Illinois-LH: A Denotational and Distributional Approach to Semantics. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

León S, Vilarino D, Pinto D, Tovar M, Beltrán B (2014) BUAP:Evaluating Compositional Distributional Semantic Models on full sentences through Semantic Relatedness and Textual Entailment. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Lien E, Kouylekov M (2014) UIO-Lien: Entailment Recognition using Minimal Recursion Semantics. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Marelli M, Bentivogli L, Baroni M, Bernardi R, Menini S, Zamparelli R (2014a) SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on full sentences through Semantic Relatedness and Textual Entailment. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Marelli M, Menini S, Baroni M, Bentivogli L, Bernardi R, Zamparelli R (2014b) A SICK cure for the evaluation of Compositional Distributional Semantic Models. In: Proceedings of LREC

Mitchell J, Lapata M (2008) Vector-based models of semantic composition. In: Proceedings of ACL

Mitchell J, Lapata M (2010) Composition in distributional models of semantics. Cognitive Science 34(8):1388–1429

Pagin P, Westersthl D (2010) Compositionality i: Definitions and variants. Philosophy Compass 5(3):250–264, DOI 10.1111/j.1747-9991.2009.00228.x, URL http://dx.doi.org/10.1111/j.1747-9991.2009.00228.x

Proisl T, Evert S (2014) SemantiKLUE: Robust Semantic Similarity at multiple levels using maximum weight matching. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Socher R, Huval B, Manning C, Ng A (2012) Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of EMNLP

Vo ANP, Popescu O, Caselli T (2014) FBK-TR: SVM for Semantic Relatedness and Corpus Patterns for RTE. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation

Zhao J, Zhu TT, Lan M (2014) ECNU: One stone two birds: Ensemble of heterogenous measures for Semantic Relatedness and Textual Entailment. In: Proceedings of SemEval 2014: International Workshop on Semantic Evaluation