A SICK cure for the evaluation of compositional distributional semantic models

M. Marelli[1], S. Menini[1,2], M. Baroni[1], L. Bentivogli[2], R. Bernardi[1], R. Zamparelli[1]
[1]University of Trento, [2]Fondazione Bruno Kessler
marco.marelli@unitn.it, menini@fbk.eu, marco.baroni@unitn.it,
bentivo@fbk.eu, raffaella.bernardi@unitn.it, roberto.zamparelli@unitn.it

Abstract

Shared and internationally recognized benchmarks are fundamental for the development of any computational system. We aim to help the research community working on compositional distributional semantic models (CDSMs) by providing SICK (Sentences Involving Compositional Knowldedge), a large size English benchmark tailored for them. SICK consists of about 10,000 English sentence pairs that include many examples of 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 (idiomatic multiword expressions, named entities, telegraphic language) that are not within the scope of CDSMs. By means of crowdsourcing techniques, each pair was annotated for two crucial semantic tasks: relatedness in meaning (with a 5-point rating scale as gold score) and entailment relation between the two elements (with three possible gold labels: entailment, contradiction, and neutral). The SICK data set was used in SemEval-2014 Task 1, and it freely available for research purposes.

Keywords: Compositional Distributional Semantic Models, Data Sets, Semantic Relatedness, Textual Entailment

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 (Compositional DSMs, or 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 (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011; Socher et al., 2012). Despite the ever increasing interest in the field, 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 (2011), are limited to a few hundred instances of very short sentences with a fixed structure. While CDSMs could be evaluated on data sets developed for other purposes, such as Semantic Text Similarity (`STS`)[1] or Recognizing Textual Entailment (`RTE`),[2] in these settings CDSMs can only act as components of more complex systems, since they require dealing with issues, such as identifying multiword expressions, recognizing named entities, processing dates and digits, or accessing encyclopedic knowledge about individuals, that are orthogonal to the generic meaning representations that CDSMs are supposed to produce. At the same time, the challenging phenomena that CDSMs must handle for a satisfactory account of sentence-level semantics (e.g., contextual synonymy and other lexical variation phenomena, active/passive and other syntactic alternations, impact of negation, quantifiers and other grammatical elements) do not occur very frequently in the `STS` and `RTE` data sets, since the latter were designed for other purposes.

With these considerations in mind, we developed SICK (Sentences Involving Compositional Knowledge), a data set aimed at filling this void. SICK includes 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 (multiword expressions, named entities, telegraphic language) that are not within the domain of compositional distributional semantics.

The SICK data set consists of around 10,000 English sentence pairs, each annotated for relatedness in meaning. The sentence relatedness score provides a direct way to evaluate CDSMs, insofar as their outputs are meant to quantify the degree of semantic relatedness between sentences. Since detecting the presence of entailment is one of the traditional benchmarks of a successful semantic system, each sentence pair is also annotated for the entailment relation between the two elements of the pair.

2. Related Work

Starting from the assumption that understanding a sentence means knowing when it is true, being able to verify whether an entailment is valid is a crucial challenge of any computational semantic system. The development of evaluation data sets on this task can help Computational Semantics make tangible progress.

A first interesting test suite of entailment problems, `FraCaS` (Framework for Computational Semantics), was developed already in the mid-'90 by a group of formal semanticists (Cooper et al., 1996), but a cleanly processable version of it has been made available only recently.[3] The data set contains entailment problems in which a conclusion has to be derived from one or more premises and not necessarily all premises are needed to verify the entailment. Premises and conclusions are simple, lab-made, English sentences involving semantic phenomena frequently ad-

[1] `http://www.cs.york.ac.uk/semeval-2012/task6/`

[2] `http://www.nist.gov/tac/2011/RTE/`

[3] `http://www-nlp.stanford.edu/~wcmac/downloads/fracas.xml`

dressed by formal semanticists, such as generalized quantifiers (GQs), ellipsis and temporal reference. No checking of their frequency in naturally occurring setting was carried out, and cross-annotator agreement was not considered. Still, the data set could be important for anyone interested in the phenomena it covered. For instance, FraCaS data was used to evaluate NatLog, which aims to tackle entailment problems based on monotonicity of GQs (MacCartney and Manning, 2007).

In 2005, the PASCAL RTE (Recognizing Textual Entailment) challenge was launched, to become a task organized year after year by the associations responsible for the main international evaluation campaigns. The development process of the RTE data sets was steadily improved, increasing the difficulties of the entailment problems involved. In 2008, the RTE-4 committee made the task more fine-grained by requiring a classification of the pairs as "entailment", "contradiction" (the negation of the conclusion is entailed from the premise) and "unknown" (Giampiccolo et al., 2008). All RTE data sets shared the approach of the original one, namely looking at the entailment problem as a sub-task of NLP real-life applications, like Question Answering, Information Retrieval or Information Extraction. Given this focus, the RTE data sets, differently from FraCaS, contain real life natural language sentences and the sort of entailment problems that occur in corpora collected from the web. As such, additional NLP tools such as Named Entity Recognizers or Word Sense Disambiguation are always needed.

Somewhere in between the FraCaS and RTE data sets lies the work by Toledo et al. (2012). They chose to annotate RTE data sets (RTE 1-3) with respect to semantic phenomena familiar to formal semanticists which have also been shown to occur frequently in real-life texts, are intuitive and hence yield high annotation consistency and do not require sophisticated abstract representations.

The semantic annotation efforts mentioned so far have been done by experts. Interestingly, crowdsourcing services have proved to be useful for textual entailment annotation. Snow et al. (2008) show high agreement between non-expert annotations of the RTE-1 dataset and existing gold standard labels assigned by expert annotators. This annotation method has been used in several international evaluation campaigns, among which Cross-Lingual Textual Entailment (Negri et al., 2012; Negri et al., 2013) and Semantic Textual Similarity (STS) (Agirre et al., 2012). The latter is related to both textual entailment (TE) and paraphrase tasks, but differs from them in a number of ways, notably the fact that, rather than being a binary yes/no decision, it is a graded similarity notion. Moreover, differently from TE, it assumes a bidirectional relation.

None of these existing data sets are suitable for evaluating CDSMs in isolation, highlighting the weakness and strengths of their main current focus: deriving a plausible meaning for the whole from the meaning of the parts. As we mentioned, a few data sets (e.g., Mitchell and Lapata (2008) and Grefenstette and Sadrzadeh (2011)) have already been developed specifically for CDSM evaluation, but none of them was geared toward evaluation campaigns, or took in consideration issues such as annotation quality or the data set size, variety of considered phenomena, etc.

Against this background, we have developed SICK. Similarly to FraCaS we use simplified English sentences, but they are derived from natural examples and we focus on entailment pairs, rather than on a conclusion following a set of premises. Furthermore, we use crowdsourcing services to annotate our data set. Following STS annotations, we provide a graded relatedness score to which we add the three class annotation "entailed" vs. "contradiction" vs. "unknown" (neutral) introduced in RTE-4.

3. Data Set Creation Process

The data set was built starting from two existing sets: the 8K ImageFlickr data set[4] and the SemEval 2012 STS MSR-Video Description data set.[5] These two data sets seemed particularly appropriate as starting points, as they contain sentences (as opposed to paragraphs) that describe the same picture or video and are thus near paraphrases, are lean on named entities and rich in generic terms. In order to generate SICK sentence pairs, we randomly selected a subset of sentence pairs from each source data set (750 + 750, S0), and we applied a 3-step process. First, the original sentences were *normalized* to remove unwanted linguistic phenomena; the normalized sentences were then *expanded* to obtain up to three new sentences with specific characteristics suitable to CDSM evaluation; as a last step, all the sentences generated in the expansion phase were *paired* to the normalized sentences in order to obtain the final data set. While all the details about the three steps are given in the following subsections, Table 1 presents a complete example of the output of this process.

3.1. Sentence Normalization

The pairs of sentences in the original data sets were normalized when necessary to exclude or simplify instances containing lexical, syntactic or semantic phenomena that CDSMs are currently not expected to account for. The normalization criteria adopted—resulting from a preparatory analysis of the original data sets—are described and exemplified in Table 2. To ensure the quality of the normalization phase, each sentence in the original pairs was normalized by two different annotators. Since there are several ways to normalize a sentence, only for around 50% of the sentences the two normalized sentences turned out to be the same. In those cases where the two normalized sentences were different, a third judge chose the most suitable one, according to criteria such as being grammatically correct, natural and respectful of the normalization rules. When both alternatives were correct and suitable, one of them was randomly chosen. When the normalization was not possible, both sentences of the pair were discarded from the data set.

3.2. Sentence Expansion

From the normalized pool of pairs we considered a subset of 500 pairs from each source data for a total number

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

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

Original pair	
S0a: *A sea turtle is hunting for fish*	**S0b:** *The turtle followed the fish*
Normalized pair	
S1a: *A sea turtle is hunting for fish*	**S1b:** *The turtle is following the fish*
Expanded pair	
S2a: *A sea turtle is hunting for food*	**S2b:** *The turtle is following the red fish*
S3a: *A sea turtle is not hunting for fish*	**S3b:** *The turtle isn't following the fish*
S4a: *A fish is hunting for a turtle in the sea*	**S4b:** *The fish is following the turtle*

Table 1: Example of output of data set creation process.

Rule	Example
Replace possessive pronouns with the word they stand for or with a determiner.	S0: "*A man is standing outside his house*" S1: "*A man is standing outside the house*"
Replace Named Entities with a word that stands for the class.	S0: "*A woman is playing Mozart*" S1: "*A woman is playing classical music*"
In order to avoid generic sentences , transform all non-stative verb tenses into present continuous.	S0: "*Birds land on clothes lines*" S1: "*Birds are landing on clothes lines*"
Replace complex verb constructions into simpler ones.	S0: "*A man is attempting to surf down a hill made of sand*" S1: "*A man is surfing down a hill made of sand*"
Simplify verb phrases with modals and auxiliaries.	S0: "*A kid has to eat a vegetable soup*" S1: "*A kid is eating a vegetable soup*"
Replace phrasal verbs with a synonym if verb and preposition are not adjacent.	S0: "*A man is sorting the documents out*" S1: "*A man is organizing the documents*"
Remove multiword expressions.	S0: "*A person is playing guitar right now*" S1: "*A person is playing guitar*"
Remove dates and numbers; if the number is a determiner write it in letters.	S0: "*3 people are on a small boat enjoying the view*" S1: "*Three people are on a small boat enjoying the view*"
Turn subordinates into coordinates.	S0: "*A faucet is running while a bird is standing in the sink below*" S1: "*A faucet is running and a bird is standing in the sink below*"
Turn non-sentential descriptions into sentences.	S0: "*An airplane in the air*" S1: "*An airplane is flying in the air*"
Remove indirect interrogative and parenthetical phrases.	We did not find any instance in the data sets

Table 2: Normalization rules

of 2000 sentences (S1). Each of these sentences was expanded to create up to three new sentences. In this step we applied syntactic and lexical transformations with predictable effects in order to obtain, with respect to the normalized sentence, *(i)* a sentence with a similar meaning, *(ii)* a sentence with a logically contradictory or at least highly contrasting meaning, and *(iii)* a sentence that contains most of the same lexical items, but has a different meaning.

To obtain sentences with a similar meaning (S2) we applied meaning preserving transformations, while to obtain sentences with a contradictory or strongly contrasting meaning (S3) we used "negative" transformations. Finally, in order to get a sentence with a different meaning using the same lexical items as much as possible (S4), a set of word-scrambling rules were applied ensuring that the resulting sentence was still meaningful. A further requirement was that all the sentences created had the same overall syntactic complexity (see Table 1 for an example of a triplet of modified sentences produced for each original item).

The rationale behind this approach is that of building a data set so that understanding when two sentences have close meanings or entail each other crucially requires a compositional semantics step, and not simply considering the individual lexical items involved, the syntactic complexity of the two sentences, or world knowledge. The complete list of expansion rules is presented in Table 3.

Note that not all the transformations were applicable to each sentence, as some transformations would have returned ungrammatical or incorrect sentences or implausible meanings. In particular, it was not possible to create an S4

Meaning Preserving Transformations	
Rule	Example
Turn active sentences into passive sentences and viceversa.	S1: "*A man is driving a car* " S2: "*The car is being driven by a man*"
Replace words with near synonyms or similar words.	S1: "*A young boy is jumping into water* " S2: "*A young kid is jumping into water*" S1: *A man and two women in a darkened room are sitting at a table with candles* S2: *A man and two women in a dark room are sitting at a table with candles*
Add modifiers that do not radically alter the meaning of the sentence.	S1: "*A deer is jumping a fence* " S2: "*A wild deer is jumping a fence*" S1: "*A woman is tapping her fingers*" S2: "*A woman is tapping her fingers nervously*"
Expand agentive nouns.	S1: "*A soccer player is kicking a ball into the goal*" S2: "*A person who plays soccer is kicking a ball into the goal*"
Turn compounds into relative clauses.	S1: "*A woman is using a sewing machine*" S2: "*A woman is using a machine made for sewing*"
Turn adjectives into relative clauses.	S1: "*Two men are taking a break from a trip on a snowy road* " S2: "*Two men are taking a break from a trip on a road covered by snow*"
Replace quantifiers with others that have a similar meaning.	S1: "*The surfer is riding a big wave*" S2: "*A surfer is riding a big wave*"

Meaning Altering Transformations	
Rule	Example
Insert or remove negations to produce contradictions.	S1: "*The boy is playing the piano*" S3: "*The boy is not playing the piano*"
Change determiners with their opposite. $\{$*the, a, all, every, some, a few*$\} \Rightarrow \{$*no*$\}$, $\{$*no*$\} \Rightarrow \{$*every, each*$\}$, $\{$*many*$\} \Leftrightarrow \{$*few*$\}$, $\{$*much*$\} \Leftrightarrow \{$*little*$\}$.	S1: "*A dog is walking along a snowdrift*" S3: "*There is no dog walking along a snowdrift*"
Replace words with semantic opposites.	S1: "*The girl is spraying the plants with water*" S3: "*The boy is spraying the plants with water*" S1: "*A plane is taking off*" S3: "*A plane is landing*"
Scramble words: switch the arguments of a transitive verb, switch and mix modifiers, exploit verb transitive/intransitive alternations, exploit homonymy and polysemy.	S1: "*The turtle is following the fish*" S4: "*The fish is following the turtle*" S1: "*A man with a jersey is dunking the ball at a basketball game*" S4: "*The game of basketball consists of a ball being dunked by a man with a jersey*"

Table 3: Expansion rules

for each normalized sentence in the data set, due to the difficulty of getting meaningful sentences after changing the order of the words, especially in the case of very short sentences, which are frequent in the data set.

The fact that not all the transformations were always suitable to the sentences results into an inhomogeneous distribution of the usage of the rules in the dataset. Table 4 shows the distribution of the expansion rules within the data set. Note that since it was possible to apply more than one rule at the same time to expand a sentence, the sum of the frequency of each rule is higher than the number of the sentences we expanded.

3.3. Generation of the SICK sentence pairs

In order to produce the final data set, the sentences resulting from the normalization and expansion phases were paired. More precisely, each normalized sentence in the pair was combined with all the sentences resulting from the expan-

Freq.	Expansion rules to S2
18	Turn passive sentences into active
303	Turn active sentences into passive
865	Replace words with synonyms
294	Add modifiers
28	Expand agentive nouns
58	Turn compounds into relative clauses
191	Turn adjectives into relative clauses
272	Replace quantifiers
Freq.	Expansion rules to S3
422	Insert a negation
648	Change determiners with opposites
974	Replace words with semantic opposites
Freq.	Expansion rules to S4
380	Scramble words

Table 4: Frequency of application of expansion rules

sion phase and with the other normalized sentence in the pair. Considering the example in Table 1, we paired each S1 with all the other seven sentences in the group. Moreover, we added to the data set a number of pairs composed of completely unrelated sentences by randomly taking two sentences from two different pairs. For example:

- *A sea turtle is hunting for fish*
- *A young woman is playing the guitar*

We constructed in this way a set of around 10,000 new sentence pairs. To ensure the quality of the data set, all the sentences were checked for grammatical or lexical mistakes and disfluencies by a native English speaker. Our goal was to obtain sentences that sound natural to a native speaker despite all the restrictions adopted during the normalization and expansion phases.

The distribution of the SICK sentence pairs with respect to the transformations performed during data set creation is presented in Table 5, 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.[6]

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 is left to human judges, as illustrated in the next section.

[6]The number of planned sentence pairs was 10,000 but at the end of the data set creation process 160 pairs had to be excluded: 126 pairs (1.3% of the set) were repeated more than once due to converging modifications in the normalization and expansion phases, and 34 (0.4% of the set) resulted in identical elements.

Expected relation	N. of pairs
Similar meaning	4366 (44.4%)
S1aS2a, S1aS2b, S1bS2a, S1bS2b, S1aS1b	
Opposite/contrasting meaning	3574 (36.3%)
S1aS3a, S1aS3b, S1bS3a, S1bS3b	
Similar lexicon, different meaning	703 (7.1%)
S1aS4a, S1aS4b, S1bS4a, S1bS4b	
Unrelated	1197 (12.2%)
Total	9840 (100%)

Table 5: Distribution of the SICK sentence pairs with respect to the transformations performed during data set creation.

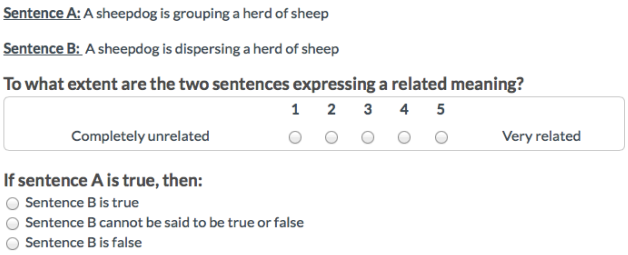

Figure 1: Example of the annotation task as displayed on the CrowdFlower interface.

3.4. Relatedness and Entailment Annotation

Each pair in the SICK data set was annotated to mark *(i)* the degree to which two sentence meanings are related (on a 5-point scale), and *(ii)* whether one entails the other. Human ratings were collected through a large crowdsourcing study, exploiting the CrowdFlower (CF) platform[7] to reach the Amazon Mechanical Turk (MTurk) marketplace.[8] Figure 1 shows the double annotation task as it was displayed to MTurk contributors through the CF interface.

In order to clarify the task to non-expert participants, while avoiding biasing their judgments with strict definitions, the instructions described the task only through examples of relatedness and entailment, as shown below:

- *A and B are completely unrelated*:
  - A: Two girls are playing outdoors near a woman
  - B: The elephant is being ridden by the man
- *A and B are very related*:
  - A: A man is cooking pancakes
  - B: The man is cooking pancakes
- *If A is true, then B is true*:
  - A: A dog is running in a field
  - B: An animal is running in a field

[7]www.crowdflower.com
[8]www.mturk.com

- *If A is true, then B cannot be said to be true or false*:
  - A: A man is breaking three eggs in a bowl
  - B: A girl is pouring some milk in a bowl
- *If A is true, then B is false*:
  - A: A man is playing golf
  - B: No man is playing golf

Control items (i.e., pairs with known correct answers) and geographical restrictions were added to the crowdsourcing task as a quality control mechanism to ensure that participants were committed to the task and exclude non-proficient English speakers.

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.

As far as time and cost are concerned, we collected 200,000 judgments in about three months, with a total cost of 2,030 dollars (40 jobs were run in total, each costing 50.75 dollars and taking 3 days as an average to complete).

Once all the annotations were collected, the gold labels were calculated following two different methodologies. The relatedness gold score was computed for each pair as the average of the ten ratings assigned by participants. 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 $SD = 0.76$. This means that, as an average, participants' judgments varied ± 0.76 rating points around the final score assigned to each pair.

With regards to entailment gold labels, a majority vote scheme was adopted. Pairs were classified as CONTRADICTION when most participants indicated that "if sentence A is true, sentence B is false" for 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. Inter–rater agreement for the entailment task was .84, computed as the average proportion of the majority vote across pairs and indicating that, as an average, 84% of participants agreed with the majority vote in each pair.

4. The SICK Data Set

The distributions of gold scores for both relatedness and entailment in the data set are summarized in Table 6.

Relatedness		Entailment	
[1-2) range	923 (10%)	NEUTRAL	5595 (57%)
[2-3) range	1373 (14%)	CONTRADICTION	1424 (14%)
[3-4) range	3872 (39%)	ENTAILMENT	2821 (29%)
[4-5] range	3672 (37%)		

Table 6: Distribution of SICK sentence pairs for each gold relatedness level and entailment label.

As a further specification of the ENTAILMENT class, note that we obtained 1300 pairs with mutual entailment ("if sentence A is true, sentence B is true" for both presentation orders) and 1521 pairs with unidirectional entailment ("if sentence A is true, sentence B is true" for only one presentation order).

The two-way Table 7 represents the gold data when considering the relatedness and entailment results together, that is, each cell in the table reports the number of sentence pairs for each combination between relatedness classes and entailment labels. The gold scores and labels are clearly associated ($\chi^2(11) = 12976$; $p = .0001$): pairs labeled as ENTAILMENT and CONTRADICTION belong to the highest relatedness levels, whereas NEUTRAL pairs are more evenly distributed. This is also evident when considering the average relatedness score for the sentence pairs in each of the entailment label: ENTAILMENT pairs have an average relatedness score of 4.57 ($SD = .34$), CONTRADICTION pairs have an average relatedness score of 3.59 ($SD = .44$), and NEUTRAL pairs have an average relatedness score of 2.98 ($SD = .93$). In fact, these average scores are all significantly different from each other ($p < .0001$) at the Tukey HSD multi-comparisons tests (Abdi and Williams, 2010).

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

Table 7: Distribution of sentence pairs across the two tasks.

The resulting gold relatedness scores and entailment labels were further analyzed by investigating how they are distributed across the pair types generated through the pairing process.

The relatedness score distribution is summarized in Table 8. The S1S2 pairs (similar meaning) were judged to be maximally related, followed by S1S3 (contrasting) and S1S4 (lexical overlap only). This confirms the observation we made above that pairs cueing an opposite meaning (S1S3) are judged as more related than pairs that have no strong meaning relation but contain the same words (S1S4). This trend could be observed both when comparing sentences belonging to the same expansion set, that is, originating from the same source sentence (S1aS2a, S1bS2b, S1aS3a, S1bS3b, S1aS4a, S1bS4b), and in pairs containing sentences from different sets (S1aS2b, S1bS2a, S1aS3b, S1bS3a, S1aS4b, S1bS4a), although the latter case is characterized by generally lower ratings and higher variance. This was expected, since the STS source pairs were already capturing different degrees of relatedness. Unrelated pairs were assigned the lowest average ratings.

Distributions of the entailment labels are reported in Table 9 (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 S1S2 pairs

Type of pair	Average relatedness (SD)
S1S2 same set	4.65 (.29)
S1S3 same set	3.59 (.44)
S1S4 same set	3.40 (.57)
S1S1 cross set	3.82 (.70)
S1S2 cross set	3.75 (.70)
S1S3 cross set	3.15 (.65)
S1S4 cross set	2.94 (.66)
Unrelated pairs	1.78 (.85)

Table 8: Average relatedness scores (and corresponding standard deviations) for each pair type.

(similar meaning), the CONTRADICTION label in case of S1S3 pairs (contrast/contradiction), and the NEUTRAL label in case of S1S4 pairs (lexical overlap only). We observe however a relatively high proportion of S1S3 pairs labeled NEUTRAL. Inspection of the NEUTRAL S1S3 pairs reveals a significantly higher incidence of pairs of sentences with subjects with indefinite articles (72% vs. 19% in the CONTRADICTION pairs). Not unreasonably, subjects found that, say, *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 to refer to 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) encouraging 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 S1S2 cross-set pairs, whereas CONTRADICTION is assigned to 16.9% of the S1S3 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, and 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 lower relatedness ratings for the cross-set cases).

Type of pair	ENTAILMENT	CONTRADICTION	NEUTRAL
S1S2 same set	94.2%	0%	5.8%
S1S3 same set	0.9%	58.2%	40.9%
S1S4 same set	9.8%	1.9%	88.3%
S1S1 cross set	37.8%	0.2%	62%
S1S2 cross set	35%	0%	65%
S1S3 cross set	2.5%	16.9%	80.6%
S1S4 cross set	4%	0%	96%
Unrelated pairs	0.9%	0.3%	98.8%

Table 9: Distribution of entailment labels across pair types.

Finally, Table 10 report performance of some baselines on the portion of SICK that has been used as test data in SemEval 2014 Task 1 (see next section). Note that the Probability baseline assigns labels randomly according to their relative frequency, and that the parameters of this and the word Overlap baseline were estimated on the SemEval training data.

Baseline	Relatedness	Entailment
Chance	0	33.3%
Majority	NA	56.7%
Probability	NA	41.8%
Overlap	0.63	56.2%

Table 10: Performance of baselines on SICK tasks. Figure of merit is Pearson correlation for relatedness and percentage accuracy for entailment.

5. Conclusion

SICK is a large data set on compositional meaning, annotated with subject ratings for both relatedness and entailment relation between sentences. As it includes a large number of phenomena that CDSMs are expected to account for (lexical and syntactic variations, effect of negation and functional structure, relatedness and entailment), we think SICK constitutes the ideal resource to assess systems that attempt to handle compositional semantics. SICK was used in SemEval 2014 Task 1[9]; details on the shared task results are reported in Marelli et al. (2014). The whole data set can be downloaded from the SICK website[10] under a Creative Commons Attribution-NonCommercial-ShareAlike license.

Acknowledgments

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

This paper is dedicated to the memory of our colleague and friend Emanuele Pianta.

6. References

Abdi, H. and Williams, L. (2010). Newman-Keuls and Tukey test. In Salkind, N., Frey, B., and Dougherty, D., editors, *Encyclopedia of Research Design*, pages 897–904. Sage, Thousand Oaks, CA.

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre., A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, volume 2.

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA.

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

[10] http://clic.cimec.unitn.it/composes/sick/

Cooper, R., Crouch, D., Eijck, J. V., Fox, C., Genabith, J. V., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., and Pulman, S. (1996). Using the framework. Technical Report Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Giampiccolo, D., Dang, H. T., Bernardo, M., Dagan, I., and Cabrio, E. (2008). The fourth pascal recognising textual entailment challenge. In *n Proceedings of the TAC 2008 Workshop on Textual Entailment*.

Grefenstette, E. and Sadrzadeh, M. (2011). Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP*, pages 1394–1404, Edinburgh, UK.

MacCartney, B. and Manning, C. D. (2007). Natural logic for textual inference. In *ACL Workshop on Textual Entailment and Paraphrasing*.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014). 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*. In press.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of ACL*, pages 236–244, Columbus, OH.

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

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., and Giampiccolo, D. (2012). Semeval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., and Giampiccolo, D. (2013). Semeval-2013 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Socher, R., Huval, B., Manning, C., and Ng, A. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*, pages 1201–1211, Jeju Island, Korea.

Toledo, A., Katrenko, S., Alexandropoulou, S., Klockmann, H., Stern, A., Dagan, I., and Winter, Y. (2012). Semantic annotation for textual entailment recognition. In Springer-Verlag, editor, *Proceedings of the 11th Mexican International Conference on Artificial Intelligence*, Lecture Notes in Artificial Inteligence, pages 12–25.