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

Marco Marelli[1] **Luisa Bentivogli**[2] **Marco Baroni**[1]
Raffaella Bernardi[1] **Stefano Menini**[1,2] **Roberto Zamparelli**[1]
[1] University of Trento, Italy
[2] FBK - Fondazione Bruno Kessler, Trento, Italy
{name.surname}@unitn.it, {bentivo,menini}@fbk.eu

Abstract

This paper presents the task on the evaluation of Compositional Distributional Semantics Models on full sentences organized for the first time within SemEval-2014. Participation was open to systems based on any approach. Systems were presented with pairs of sentences and were evaluated on their ability to predict human judgments on *(i)* semantic relatedness and *(ii)* entailment. The task attracted 21 teams, most of which participated in both subtasks. We received 17 submissions in the relatedness subtask (for a total of 66 runs) and 18 in the entailment subtask (65 runs).

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; Grefenstette and Sadrzadeh, 2011; Mitchell and Lapata, 2010; 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. 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. 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 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 are challenging for reasons due to semantic compositionality (e.g. 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 the void, 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., knowing 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.

2 The task

The Task involved two subtasks. (i) *Relatedness*: predicting the degree of semantic similarity between two sentences, and (ii) *Entailment*: detecting the entailment relation holding between them (see below for the exact definition). Sentence relatedness scores provide a direct way to evaluate CDSMs, insofar as their outputs are able to quantify the degree of semantic similarity between sentences. On the other hand, 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 for semantic systems.

In the semantic relatedness subtask, given two sentences, systems were required to produce a relatedness score (on a continuous scale) indicating the extent to which the sentences were expressing a related meaning. Table 1 shows examples of sentence pairs with different degrees of semantic relatedness; gold relatedness scores are expressed on a 5-point rating scale.

In the entailment subtask, given two sentences A and B, systems had to determine whether the meaning of B was entailed by A. In particular, systems were required to assign to each pair either the ENTAILMENT label (when A entails B, viz., B cannot be false when A is true), the CONTRADICTION label (when A contradicted B, viz. B is false whenever A is true), or the NEUTRAL label (when the truth of B could not be determined on the basis of A). Table 2 shows examples of sentence pairs holding different entailment relations.

Participants were invited to submit up to five system runs for one or both subtasks. Developers of CDSMs were especially encouraged to participate, but developers of other systems that could tackle sentence relatedness or entailment tasks were also welcome. Besides being of intrinsic interest, the latter systems' performance will serve to situate CDSM performance within the broader landscape of computational semantics.

3 The SICK data set

The SICK data set, consisting of about 10,000 English sentence pairs annotated for relatedness in meaning and entailment, was used to evaluate the systems participating in the task. The data set creation methodology is outlined in the following subsections, while all the details about data generation and annotation, quality control, and inter-annotator agreement can be found in Marelli et al. (2014).

3.1 Data set creation

SICK was built starting from two existing data sets: the 8K ImageFlickr data set[1] and the SemEval-2012 STS MSR-Video Descriptions data set.[2] The 8K ImageFlickr dataset is a dataset 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 snip-

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

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

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 1: Examples of sentence pairs with their gold relatedness scores (on a 5-point rating scale).

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*"

Table 2: Examples of sentence pairs with their gold entailment labels.

pets which compose 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 3 presents an example of the output of each step in the process.

The *normalization* step was carried out on the original sentences (S0) to exclude or simplify instances that contained lexical, syntactic or semantic phenomena (e.g., named entities, dates, numbers, multiword expressions) that CDSMs are currently not expected to account for.

The *expansion* step was applied to each of the normalized sentences (S1) in order to create up to three new sentences with specific characteristics suitable to 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 (S2), *(ii)* a sentence with a logically contradictory or at least highly contrasting meaning (S3), and *(iii)* a sentence that contains most of the same lexical items, but has a different meaning (S4) (this last step was carried out only where it could yield a meaningful sentence; as a result, not all normalized sentences have an (S4) expansion).

Finally, in the *pairing* step each normalized sentence in the pair was combined with all the sentences resulting from the expansion phase and with the other normalized sentence in the pair. Considering the example in Table 3, *S1a* and *S1b* were paired. Then, *S1a* and *S1b* were each combined with *S2a*, *S2b*,*S3a*, *S3b*, *S4a*, and *S4b*, leading to a total of 13 different sentence pairs.

Furthermore, a number of pairs composed

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 pairs	
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 3: Data set creation process.

of completely unrelated sentences were added to the data set by randomly taking two sentences from two different pairs.

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 that of building a data set which encouraged the use of a compositional semantics step in understanding when two sentences have close meanings or entail each other, hindering methods based on individual lexical items, on the syntactic complexity of the two sentences or on pure world knowledge.

3.2 Relatedness and Entailment annotation

Each pair in the SICK dataset 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 or contradicts the other (considering both directions). The ratings were collected through a large crowdsourcing study, 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. Once all the annotations were collected, the relatedness gold score was computed for each pair as the average of the ten ratings assigned by participants, whereas a majority vote scheme was adopted for the entailment gold labels.

3.3 Data set statistics

For the purpose of the task, 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 4 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.

4 Evaluation metrics and baselines

Both subtasks were evaluated using standard metrics. In particular, the results on entailment were evaluated using accuracy, whereas the outputs on relatedness were evaluated using Pearson correlation, Spearman correlation, and Mean Squared Error (MSE). Pearson correlation was chosen as the official measure to rank the participating systems.

Table 5 presents the performance of 4

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

Table 4: Distribution of sentence pairs across the Training and Test Sets.

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.

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

Table 5: Performance of baselines. Figure of merit is Pearson correlation for relatedness and accuracy for entailment. NA = *Not Applicable*

5 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 received 17 submissions to the relatedness subtask (for a total of 66 runs) and 18 for the entailment subtask (65 runs).

We asked participants to pre-specify a primary run to encourage commitment to a theoretically-motivated approach, rather than post-hoc performance-based assessment. Interestingly, some participants used the non-primary runs to explore the performance one could reach by exploiting weaknesses in the data that are not likely to hold in future tasks of the same kind (for instance, run 3 submitted by The Meaning Factory exploited sentence ID ordering information, but it was not presented as a primary run). Participants could also use non-primary runs to test smart baselines. In the relatedness subtask six non-primary runs slightly outperformed the official winning primary entry,[3] while in the entailment task all ECNU's runs but run 4 were better than ECNU's primary run. Interestingly, the differences between the ECNU's runs were due to the learning methods used.

We present the results achieved by primary runs against the Entailment and Relatedness subtasks in Table 6 and Table 7, respectively.[4] We witnessed a very close finish in both subtasks, with 4 more systems within 3 percentage points of the winner in both cases. 4 of these 5 top systems were the same across the two subtasks. Most systems performed well above the best baselines from Table 5.

The 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, SICK entailment is "easier" than RTE. Con-

[3] They were: The_Meaning_Factory's run3 (Pearson 0.84170) ECNU's runs2 (0.83893) run5 (0.83500) and StanfordNLP's run4 (0.83462) and run2 (0.83103).

[4] ITTK's primary run could not be evaluated due to technical problems with the submission. The best ITTK's non-primary run scored 78,2% accuracy in the entailment task and 0.76 *r* in the relatedness task.

sidering the first five RTE challenges (Bentivogli et al., 2009), the median values ranged from 56.20% to 61.75%, whereas the average values ranged from 56.45% to 61.97%. The entailment scores obtained on the SICK data set are considerably higher, being 77.06% for the median system and 75.36% for the average system. On the other hand, the relatedness task is more challenging than the one run on MSRvid (one of our data sources) at STS 2012, where the top Pearson correlation was 0.88 (Agirre et al., 2012).

ID	Compose	ACCURACY
Illinois-LH_run1	P/S	84.6
ECNU_run1	S	83.6
UNAL-NLP_run1		83.1
SemantiKLUE_run1		82.3
The_Meaning_Factory_run1	S	81.6
CECL_ALL_run1		80.0
BUAP_run1	P	79.7
UoW_run1		78.5
Uedinburgh_run1	S	77.1
UIO-Lien_run1		77.0
FBK-TR_run3	P	75.4
StanfordNLP_run5	S	74.5
UTexas_run1	P/S	73.2
Yamraj_run1		70.7
asjai_run5	S	69.8
haLF_run2	S	69.4
RTM-DCU_run1		67.2
UANLPCourse_run2	S	48.7

Table 6: Primary run results for the entailment subtask. The table also shows whether a system exploits composition information at either the phrase (P) or sentence (S) level.

ID	Compose	r	ρ	MSE
ECNU_run1	S	0.828	0.769	0.325
StanfordNLP_run5	S	0.827	0.756	0.323
The_Meaning_Factory_run1	S	0.827	0.772	0.322
UNAL-NLP_run1		0.804	0.746	0.359
Illinois-LH_run1	P/S	0.799	0.754	0.369
CECL_ALL_run1		0.780	0.732	0.398
SemantiKLUE_run1		0.780	0.736	0.403
RTM-DCU_run1		0.764	0.688	0.429
UTexas_run1	P/S	0.714	0.674	0.499
UoW_run1		0.711	0.679	0.511
FBK-TR_run3	P	0.709	0.644	0.591
BUAP_run1	P	0.697	0.645	0.528
UANLPCourse_run2	S	0.693	0.603	0.542
UQeResearch_run1		0.642	0.626	0.822
ASAP_run1	P	0.628	0.597	0.662
Yamraj_run1		0.535	0.536	2.665
asjai_run5	S	0.479	0.461	1.104

Table 7: Primary run results for the relatedness subtask (r for Pearson and ρ for Spearman correlation). The table also shows whether a system exploits composition information at either the phrase (P) or sentence (S) level.

6 Approaches

A summary of the approaches used by the systems to address the task is presented in Table 8. In the table, systems in bold are those for which the authors submitted a paper (Ferrone and Zanzotto, 2014; Bjerva et al., 2014; Beltagy et al., 2014; Lai and Hockenmaier, 2014; Alves et al., 2014; León et al., 2014; Bestgen, 2014; Zhao et al., 2014; Vo et al., 2014; Biçici and Way, 2014; Lien and Kouylekov, 2014; Jimenez et al., 2014; Proisl and Evert, 2014; 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, and divide the former into 'fully compositional' (systems that compositionally computed the meaning of the full sentences, though not necessarily by assigning meanings to intermediate syntactic constituents) and 'partially compositional' (systems that stop the composition at the level of phrases). 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.

Let us focus on such compositional methods. Concerning the relatedness task, the fine-grained analyses reported for several systems (Illinois-LH, The Meaning Factory and ECNU) shows that purely compositional systems currently reach performance above 0.7 r. In particular, ECNU's compositional feature gives 0.75 *r*, The Meaning Factory's logic-based composition model 0.73 *r*, and Illinois-LH compositional features combined with Word Overlap 0.75 *r*. While competitive, these scores are lower than the one of the best purely non-compositional system (UNAL-NLP) which reaches the 4th position (0.80 *r* UNAL-NLP vs. 0.82 *r* obtained by the best system). UNAL-NLP however exploits an ad-hoc "negation" feature discussed below.

In the entailment task, the best non-compositional model (again UNAL-NLP) reaches the 3rd position, within close reach of the best system (83% UNAL-NLP vs. 84.5% obtained by the best system). Again, purely compositional models have lower performance. haLF CDSM reaches 69.42% accuracy, Illinois-LH Word Overlap combined with a compositional feature reaches 71.8%. The fine-grained analysis reported by Illinois-LH (Lai and Hockenmaier, 2014) shows that a full compositional system (based on pointwise multiplication) fails to capture contradiction. It is better than partial phrase-based compositional models in recognizing entailment pairs, but worse than them on recognizing neutral pairs.

Given our more general interest in the distributional approaches, in Table 8 we also classify the different DSMs used as 'Vector Space Models', 'Topic Models' and 'Neural Language Models'. Due to the impact shown by learning methods (see ECNU's results), we also report the different learning approaches used.

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 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 negation had been optionally combined with word-rearranging in the creation of S4 sentences, see Section 3.1 above).

Finally, Table 8 reports about the use of external resources in the task. One of the reasons we created SICK was to have a compositional semantics benchmark that would not require too many external tools and resources (e.g., named-entity recognizers, gazetteers, ontolo-

Participant ID	Non composition features									Comp features		Learning Methods							External Resources				
	Vector Semantics Model	Topic Model	Neural Language Model	Denotational 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
ASAP	R	R				R	R	R	R		R			R					R				
ASJAI	B		B	B	B	B	B	B		B		E		B				R		B			
BUAP	B		B			B	B		E		B	E							B				
UEdinburgh	B				B		B		B	B		E	R								B		
CECL	B				B	B		B										B			B		
ECNU	B		B		B	B	B	B		B		B	B	B	B				B		B		
FBK-TR		R			R	R	E	B	E	E	B	R		E					R		R		E
haLF	E					E				E		E											
IITK	B				B	B	B	B	B		B	B							B				
Illinois-LH	B			B	B	B	B		B	B	B							B	B			B	B
RTM-DCU	B						B					B			B						B		
SemantiKLUE	B				B	B	B		B			B							B		B		
StandfordNLP	B		B		R	R			R	B							E						
The Meaning Factory	R		R	R	R		R	R		B		E			R	E			B	B	R		
UANLPCourse	B				B	B				B		B											
UIO-Lien									E										E				
UNAL-NLP						B	B		B									B	R	B	B		
UoW				B	B		B		B			B								B			
UQeRsearch					R	R	R	R	R									R	R				
UTexas	B			B					B	B	B					B					B		
Yamarj	B				B	B						B											

Table 8: Summary of the main characteristics of the participating systems on R(elatedness), E(ntailment) or B(oth)

gies). By looking at 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 knowledge resources (mostly, WordNet and paraphrase corpora) were used.

7 Conclusion

We presented the results of the first task on the evaluation of compositional distributional semantic models and other semantic systems on full sentences, organized within SemEval-2014. 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. Moreover, the top-ranking systems in both tasks use compositional features. However, it must be noted that all systems also exploit non-compositional features and most of them

use external resources, especially WordNet. Almost all the participating systems outperformed the proposed baselines in both tasks. Further analyses carried out by some participants in the task show that purely compositional approaches reach accuracy above 70% in entailment and 0.70 *r* for relatedness. These scores are comparable with the average results obtained in the task.

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

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 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.

Ana O. Alves, Adirana Ferrugento, Mariana Lorenço, and Filipe Rodrigues. 2014. ASAP: Automatica semantic alignment for phrases. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

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

Islam Beltagy, Stephen Roller, Gemma Boleda, Katrin Erk, and Raymon J. Mooney. 2014. UTexas: Natural language semantics using distributional semantics and probablisitc logic. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

Luisa Bentivogli, Ido Dagan, Hoa T. Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *The Text Analysis Conference (TAC 2009)*.

Yves Bestgen. 2014. CECL: a new baseline and a non-compositional approach for the Sick benchmark. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

Ergun Biçici and Andy Way. 2014. RTM-DCU: Referential translation machines for semantic similarity. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

Johannes Bjerva, Johan Bos, Rob van der Goot, and Malvina Nissim. 2014. The Meaning Factory: Formal Semantics for Recognizing Textual Entailment and Determining Semantic Similarity. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. Evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, pages 177–190. Springer.

Lorenzo Ferrone and Fabio Massimo Zanzotto. 2014. haLF:comparing a pure CDSM approach and a standard ML system for RTE. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

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

Rohit Gupta, Ismail El Maarouf Hannah Bechara, and Costantin Orasăn. 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*.

Sergio Jimenez, George Duenas, Julia Baquero, and Alexander Gelbukh. 2014. UNAL-NLP:

Combining soft cardinality features for semantic textual similarity, relatedness and entailment. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

Alice Lai and Julia Hockenmaier. 2014. Illinois-lh: A denotational and distributional approach to semantics. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

Saúl León, Darnes Vilarino, David Pinto, Mireya Tovar, and Beatrice Beltrán. 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*.

Elisabeth Lien and Milen Kouylekov. 2014. UIO-Lien: Entailment recognition using minimal recursion semantics. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*, Reykjavik.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL*, pages 236–244, Columbus, OH.

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

Thomas Proisl and Stefan Evert. 2014. SemantiKLUE: Robust semantic similarity at multiple levels using maximum weight matching. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

Richard Socher, Brody Huval, Christopher Manning, and Andrew Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*, pages 1201–1211, Jeju Island, Korea.

An N. P. Vo, Octavian Popescu, and Tommaso Caselli. 2014. FBK-TR: SVM for Semantic Relatedness and Corpus Patterns for RTE. In *Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*.

Jiang Zhao, Tian Tian Zhu, and Man Lan. 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*.