Probing for semantic evidence of composition by means of simple classification tasks

Allyson Ettinger[1], **Ahmed Elgohary**[2], **Philip Resnik**[1,3]
[1]Linguistics, [2]Computer Science, [3]Institute for Advanced Computer Studies
University of Maryland, College Park, MD
{aetting,resnik}@umd.edu,elgohary@cs.umd.edu

Abstract

We propose a diagnostic method for probing specific information captured in vector representations of sentence meaning, via simple classification tasks with strategically constructed sentence sets. We identify some key types of semantic information that we might expect to be captured in sentence composition, and illustrate example classification tasks for targeting this information.

1 Introduction

Sentence-level meaning representations, when formed from word-level representations, require a process of composition. Central to evaluation of sentence-level vector representations, then, is evaluating how effectively a model has executed this composition process.

In assessing composition, we must first answer the question of what it means to do composition well. On one hand, we might define effective composition as production of sentence representations that allow for high performance on a task of interest (Kiros et al., 2015; Tai et al., 2015; Wieting et al., 2015; Iyyer et al., 2015). A limitation of such an approach is that it is likely to produce overfitting to the characteristics of the particular task.

Alternatively, we might define effective composition as generation of a meaning representation that makes available all of the information that we would expect to be extractable from the meaning of the input sentence. For instance, in a representation of the sentence "The dog didn't bark, but chased the cat", we would expect to be able to extract the information that there is an event of chasing, that a dog is doing the chasing and a cat is being chased, and that there is no barking event (though there is a semantic relation between *dog* and *bark*, albeit modified by negation, which we likely want to be able to extract as well). A model able to produce meaning representations that allow for extraction of these kinds of key semantic characteristics—semantic roles, event information, operator scope, etc—should be much more generalizable across applications, rather than targeting any single application at the cost of others.

With this in mind, we propose here a linguistically-motivated but computationally straightforward diagnostic method, intended to provide a targeted means of assessing the specific semantic information that is being captured in sentence representations. We propose to accomplish this by constructing sentence datasets controlled and annotated as precisely as possible for their linguistic characteristics, and directly testing for extractability of semantic information by testing classification accuracy in tasks defined by the corresponding linguistic characteristics. We present the results of preliminary experiments as proof-of-concept.

2 Existing approaches

The SICK entailment dataset (Marelli et al., 2014) is a strong example of a task-based evaluation metric, constructed with a mind to systematic incorporation of linguistic phenomena relevant to composition. SICK is one of the most commonly used benchmark tasks for evaluating composition models (Kiros et al., 2015; Tai et al., 2015; Wieting et al., 2015). However, conclusions that we can draw from this dataset are limited for a couple of reasons. First, certain cues in this dataset allow for strong performance without composition (for example, as Bentivogli et al. (2016) point out, 86.4% of sentence pairs labeled as CONTRADICTION can be identified simply by detecting the presence of negation; a similar obser-

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 134–139,
Berlin, Germany, August 12, 2016.

vation is made by Lai and Hockenmaier (2014)), which means that we cannot draw firm composition conclusions from performance on this task. Furthermore, if we want to examine the extent to which specific types of linguistic information are captured, SICK is limited in two senses. First, SICK sentences are annotated for transformations performed between sentences, but these annotations lack coverage of many linguistic characteristics important to composition (e.g., semantic roles). Second, even within annotated transformation categories, distributions over entailment labels are highly skewed (e.g., 98.9% of the entailment labels under the "add modifier" transformation are ENTAILMENT), making it difficult to test phenomenon- or transformation-specific classification performance.

In an alternative approach, Li et al. (2015) use visualization techniques to better examine the particular aspects of compositionality captured by their models. They consider recurrent neural network composition models trained entirely for one of two tasks—sentiment analysis and language modeling—and employ dimensionality reduction to visualize sentiment neighborhood relationships between words or phrases before and after applying modification, negation, and clause composition. They also visualize the saliency of individual tokens with respect to the prediction decision made for each of their tasks.

In comparison, our proposal aims to provide generic (task-independent) evaluation and analysis methods that directly quantify the extractability of specific linguistic information that a composition model should be expected to capture. Our proposed evaluation approach follows a similar rationale to that of the diagnostic test suite TSNLP (Balkan et al., 1994) designed for evaluating parsers on a per-phenomenon basis. As highlighted by Scarlett and Szpakowicz (2000) the systematic fine-grained evaluation of TSNLP enables precise pinpointing of parsers' limitations, while ensuring broad coverage and controlled evaluation of various linguistic phenomena and syntactic structures. Our proposal aims at initiating work on developing similar test suites for evaluating semantic composition models.

3 Probing for semantic information with targeted classification tasks

The reasoning of our method is as follows: if we take a variety of sentences—each represented by a composed vector—and introduce a classification scheme requiring identification of a particular type of semantic information for accurate sentence classification, then by testing accuracy on this task, we can assess whether the composed representations give access to the information in question. This method resembles that used for decoding human brain activation patterns in cognitive neuroscience studies of language understanding (Frankland and Greene, 2015), as well as work in NLP that has previously made use of classification accuracy for assessing information captured in vector representations (Gupta et al., 2015).

In order to have maximum confidence in our interpretation of performance in these tasks, our sentences must have sufficient diversity to ensure that there are no consistently correlating cues that would allow for strong performance without capturing the relevant compositional information. Relatedly, we want to ensure that the classification tasks cannot be solved by memorization (rather than actual composition) of phrases.

3.1 Dataset construction

The goal in constructing the sentence dataset is to capture a wide variety of syntactic structures and configurations, so as to reflect as accurately as possible the diversity of sentences that systems will need to handle in naturally-occurring text—while maintaining access to detailed labeling of as many relevant linguistic components of our data as possible. Ideally, we want a dataset with enough variation and annotation to allow us to draw data for all of our desired classification tasks from this single dataset.

For our illustrations here, we restrict our structural variation to that available from active/passive alternations, use of relative clauses at various syntactic locations, and use of negation at various syntactic locations. This allows us to demonstrate decent structural variety without distracting from illustration of the semantic characteristics of interest. Many more components can be added to increase complexity and variation, and to make sentences better reflect natural text. More detailed discussion of considerations for construction of the actual dataset is given in Section 5.

3.2 Semantic characteristics

There are many types of semantic information that we might probe for with this method. For our purposes here, we are going to focus on two basic types, which are understood in linguistics to be fundamental components of meaning, and which have clear ties to common downstream applications: semantic role and scope.

The importance of semantic role information is well-recognized both in linguistics and in NLP—for the latter in the form of tasks such as abstract meaning representation (AMR) (Banarescu et al., 2013). Similarly, the concept of scope is critical to many key linguistic phenomena, including negation—the importance of which is widely acknowledged in NLP, in particular for applications such as sentiment analysis (Blunsom et al., 2013; Iyyer et al., 2015). Both of these information types are of course critical to computing entailment.

3.3 Example classification tasks

Once we have identified semantic information of interest, we can design classification tasks to target this information. We illustrate with two examples.

Semantic role If a sentence representation has captured semantic roles, a reasonable expectation would be extractability of the entity-event relations contained in the sentence meaning. So, for instance, we might choose *professor* as our entity, *recommend* as our event, and AGENT as our relation—and label sentences as positive if they contain *professor* in the AGENT relation with the verb *recommend*. Negative sentences for this task could in theory be any sentence lacking this relation—but it will be most informative to use negative examples containing the relevant lexical items (*professor*, *recommend*) without the relation of interest, so that purely lexical cues cannot provide an alternative classification heuristic.

Examples illustrating such a setup can be seen in Table 1. In this table we have included a sample of possible sentences, varying only by active/passive alternation and placement of relative clauses, and holding lexical content fairly constant. The verb *recommend* and its agent have been bolded for the sake of clarity.

An important characteristic of the sentences in Table 1 is their use of long-distance dependencies, which cause cues based on linear order and word adjacency to be potentially misleading. Notice, for instance, that sentence 5 of the positive label column contains the string *the school recommended*, though *school* is not the agent of *recommended*—rather, the agent of *recommended* is located at the beginning of the sentence. We believe that incorporation of such long-distance dependencies is critical for assessing whether systems are accurately capturing semantic roles across a range of naturally-occurring sentence structures (Rimell et al., 2009; Bender et al., 2011).

This example task can obviously be extended to other relations and other entities/events as desired, with training and test data adjusted accordingly. We will remain agnostic here as to the optimal method of selecting relations and entities/events for classification tasks; in all likelihood, it will be ideal to choose and test several different combinations, and obtain an average accuracy score. Note that if we structure our task as we have suggested here—training and testing only on sentences containing certain selected lexical items—then the number of examples at our disposal (both positive and negative) will be dependent in large part on the number of syntactic structures covered in the dataset. This emphasizes again the importance of incorporating broad structural diversity in the dataset construction.

Negation scope Negation presents somewhat of a challenge for evaluation. How can we assess whether a representation captures negation properly, without making the task as simple as detecting that negation is present in the sentence?

One solution that we propose is to incorporate negation at various levels of syntactic structure (corresponding to different negation scopes), which allows us to change sentence meaning while holding lexical content relatively constant. One way that we might then assess the negation information accessible from the representation would be to adapt our classification task to focus not on a semantic role relation *per se*, but rather on the event described by the sentence meaning. For instance, we might design a task in which sentences are labeled as positive if they describe an event in which a professor performs an act of recommending, and negative otherwise.

The labeling for several sentences under this as well as the previous classification scheme are given in Table 2. In the first sentence, when negation falls in the relative clause (*that did not like the school*)—and therefore has scope only over *like the school*—*professor* is the agent of *recommend*,

Positive label	Negative label
the professor recommended the student	**the student recommended** the professor
the administrator was **recommended** by **the professor**	the professor was **recommended** by **the administrator**
the school hired the researcher that **the professor recommended**	the school hired the professor that **the researcher recommended**
the school hired **the professor** that **recommended** the researcher	the school hired the professor that was **recommended** by **the researcher**
the professor that liked the school **recommended** the researcher	**the school** that hired the professor **recommended** the researcher

Table 1: Labeled data for professor-as-agent-of-recommend task (*recommend* verb and its actual agent have been bolded).

and the professor entity does perform an act of recommending. In the second sentence, however, negation has scope over *recommend*, resulting in a meaning in which the professor, despite being agent of *recommend*, is not involved in performing a recommendation. By incorporating negation in this way, we allow for a task that assesses whether the effect of negation is being applied to the correct component of the sentence meaning.

4 Preliminary experiments

As proof-of-concept, we have conducted some preliminary experiments to test that this method yields results patterning in the expected direction on tasks for which we have clear predictions about whether a type of information could be captured.

Experiments Settings

We compared three sentence embedding methods: 1) Averaging GloVe vectors (Pennington et al., 2014), 2) Paraphrastic word averaging embeddings (Paragram) trained with a compositional objective (Wieting et al., 2015). and 3) Skip-Thought embeddings (ST) (Kiros et al., 2015).[1] For each task, we used a logistic regression classifier with train/test sizes of 1000/500.[2] The classification accuracies are summarized in Table 4.

We used three classification tasks for preliminary testing. First, before testing actual indicators of composition, as a sanity check we tested whether classifiers could be trained to recognize the simple presence of a given lexical item, specifically *school*. As expected, we see that all three models are able to perform this task with 100% accuracy, suggesting that this information is well-captured and easily accessible. As an extension of this sanity check, we also trained classifiers to recognize sentences containing a token in the category of "human". To test for generalization across the category, we ensured that no human nouns appearing in the test set were included in training sentences. All three models reach a high classification performance on this task, though Paragram lags behind slightly.

Finally, we did a preliminary experiment pertaining to an actual indicator of composition: semantic role. We constructed a simple dataset with structural variation stemming only from active/passive alternation, and tested whether models could differentiate sentences with *school* appearing in an agent role from sentences with *school* appearing as a patient. All training and test sentences contained the lexical item "school", with both active and passive sentences selected randomly from the full dataset for inclusion in training and test sets. Note that with sentences of this level of simplicity, models can plausibly use fairly simple order heuristics to solve the classification task, so a model that retains order information (in this case, only ST) should have a good chance of performing well. Indeed, we see that ST reaches a high level of performance, while the two averaging-based models never exceed chance-level performance.

5 Discussion

We have proposed a diagnostic method for directly targeting and assessing specific types of semantic information in composed sentence representations, guided by considerations of the linguistic information that one might reasonably expect to be extractable from properly composed sentence meaning representations.

Construction of the real dataset to meet all of our desired criteria promises to be a non-trivial task, but we expect it to be a reasonable one. A carefully-engineered context-free-grammar-based

[1] We used the pretrained models provided by the authors. GloVe and Paragram embeddings are of size 300 while Skip-Thought embeddings are of size 2400.

[2] We tuned each classifier with 5-fold cross validation.

sentence	prof-ag-of-rec	prof-recommends
the professor that *did not* like the school **recommended** the researcher	TRUE	TRUE
the professor that liked the school *did not* **recommend** the researcher	TRUE	FALSE
the school that liked the professor **recommended** the researcher	FALSE	FALSE

Table 2: Sentence labeling for two classification tasks: "contains *professor* as AGENT of *recommend*" (column 2), and "sentence meaning involves professor performing act of recommending" (column 3).

Task	GloVe	Paragram	ST
Has-school	100.0	100.0	100.0
Has-human	99.9	90.5	99.0
School-as-agent	47.98	48.57	91.15

Table 3: Percentage correct on has-school, has-human, and has-school-as-agent tasks.

generative process should allow us to cover a good deal of ground with respect to syntactic variation as well as annotation of linguistic characteristics. Human involvement in annotation should become necessary only if we desire annotation of linguistic characteristics that do not follow deterministically from syntactic properties.

One example of such a characteristic, which merits discussion of its own, is sentence plausibility. A clear limitation of automated sentence generation in general is the inability to control for plausibility of the generated sentences. We acknowledge this limitation, but would argue that for the purposes of evaluating composition, the presence of implausible sentences is not only acceptable—it is possibly advantageous. It is acceptable for the simple reason that composition seems to operate independently of plausibility: consider, for instance, a sentence such as *blue giraffes interrogate the apple*, which we are able to compose to extract a meaning from, despite its nonsensical nature. Arguments along this vein have been made in linguistics since Chomsky (1957) illustrated (with the now-famous example *colorless green ideas sleep furiously*) that sentences can be grammatical—structurally interpretable—without having a sensible meaning.

As for the potential advantage, the presence of implausible sentences in our dataset may be desirable for the following reason: in evaluating whether a system is able to perform composition, we are in fact interested in whether it is able to compose completely novel phrases. To evaluate this capacity accurately, we will want to assess systems' composition performance on phrases that they have never encountered. Elgohary and Carpuat (2016) meet this need by discarding all training sentences that include any observed bigrams in their evaluation sentences. With implausible sentences, we can substantially reduce the likelihood that systems will have been trained on the phrases encountered during the classification evaluation—while remaining agnostic as to the particulars of those systems' training data. It would be useful, in this case, to have annotation of the plausibility levels of our sentences, in order to ascertain whether performance is in fact aided by the presence of phrases that may reasonably have occurred during composition training. Possible approaches to estimating plausibility without human annotation include using n-gram statistics on simple argument/predicate combinations (Rashkin et al., 2016) or making use of selectional preference modeling (Resnik, 1996; Erk, 2007; Séaghdha, 2010).

A final note: learning low-dimensional vector representations for sentences is bound to require a trade-off between the coverage of encoded information and the accessibility of encoded information—some semantic characteristics may be easily extractable at the cost of others. We have not, in this proposal, covered all semantic characteristics of interest, but it will ultimately be valuable to develop a broad-coverage suite of classification tasks for relevant information types, to obtain an assessment that is both fine-grained and comprehensive. This kind of holistic assessment will be useful for determining appropriate models for particular tasks, and for determining directions for model improvement.

Acknowledgments

This work was supported in part by an NSF Graduate Research Fellowship under Grant No. DGE 1322106. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF. This work benefited from the helpful comments of two anonymous reviewers, and from discussions with Marine Carpuat, Alexander Williams and Hal Daumé III.

References

Lorna Balkan, Doug Arnold, and Siety Meijer. 1994. Test suites for natural language processing. *Translating and the Computer*, pages 51–51.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Emily M Bender, Dan Flickinger, Stephan Oepen, and Yi Zhang. 2011. Parser evaluation over local and non-local deep dependencies in a large corpus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 397–408.

Luisa Bentivogli, Raffaella Bernardi, Marco Marelli, Stefano Menini, Marco Baroni, and Roberto Zamparelli. 2016. SICK through the SemEval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, pages 1–30.

Phil Blunsom, Edward Grefenstette, and Karl Moritz Hermann. 2013. "not not bad" is not "bad": A distributional account of negation. In *Proceedings of the 2013 Workshop on Continuous Vector Space Models and their Compositionality*.

Noam Chomsky. 1957. *Syntactic structures*. Mouton & Co.

Ahmed Elgohary and Marine Carpuat. 2016. Learning monolingual compositional representations via bilingual supervision. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 216–223.

Steven M Frankland and Joshua D Greene. 2015. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, 112(37):11732–11737.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedingsof the Conference on Empirical Methods in Natural Language Processing*.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1681–1691.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3276–3284.

Alice Lai and Julia Hockenmaier. 2014. Illinois-lh: A denotational and distributional approach to semantics. *Proc. SemEval*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

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 *Language Resources and Evaluation*, pages 216–223.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1):127–159.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 813–821.

Elizabeth Scarlett and Stan Szpakowicz. 2000. The power of the tsnlp: lessons from a diagnostic evaluation of a broad-coverage parser. In *Advances in Artificial Intelligence*, pages 138–150. Springer.

Diarmuid O Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 435–444.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.