Lexicosyntactic Inference in Neural Models

Aaron Steven White
University of Rochester

Rachel Rudinger
Johns Hopkins University

Kyle Rawlins
Johns Hopkins University

Benjamin Van Durme
Johns Hopkins University

Abstract

We investigate neural models' ability to capture *lexicosyntactic inferences*: inferences triggered by the interaction of lexical and syntactic information. We take the task of event factuality prediction as a case study and build a factuality judgment dataset for all English clause-embedding verbs in various syntactic contexts. We use this dataset, which we make publicly available, to probe the behavior of current state-of-the-art neural systems, showing that these systems make certain systematic errors that are clearly visible through the lens of factuality prediction.

1 Introduction

The formal semantics literature has long been concerned with the complex array of inferences that different open class lexical items trigger (Kiparsky and Kiparsky, 1970; Karttunen, 1971a,b; Horn, 1972; Karttunen and Peters, 1979; Heim, 1992; Simons, 2001, 2007; Simons et al., 2010; Abusch, 2002, 2010; Gajewski, 2007; Anand and Hacquard, 2013, 2014). For example, why does (1a) give rise to the inference (2a), while the structurally identical (1b) triggers the inference (2b)?

(1) a. Jo doesn't believe that Bo left.
b. Jo doesn't know that Bo left.

(2) a. Jo believes that Bo didn't leave.
b. Bo left.
c. Bo didn't leave.

A major finding of this literature is that lexically triggered inferences are conditioned by surprising aspects of the syntactic context that a word occurs in. For example, while (3a), (3b), and (4a) trigger the inference (2b), (4b) triggers the inference (2c).

(3) a. Jo remembered that Bo left.
b. Jo didn't remember that Bo left.

(4) a. Bo remembered to leave.
b. Bo didn't remember to leave.

Accurately capturing such interactions – e.g. between clause-embedding verbs, negation, and embedded clause type – is important for any system that aims to do general natural language inference (MacCartney et al. 2008 *et seq*; cf. Dagan et al. 2006) or event extraction (see Grishman and Sundheim 1996 *et seq*), and it seems unlikely to be a trivial phenomenon to capture, given the complexity and variability of the inferences involved (see, e.g., Karttunen, 2012, 2013; Karttunen et al., 2014; van Leusen, 2012; White, 2014; Baglini and Francez, 2016; Nadathur, 2016, on implicatives).

In this paper, we investigate how well current state-of-the-art neural systems for a subtask of general event extraction – event factuality prediction (EFP; Nairn et al., 2006; Saurí and Pustejovsky, 2009, 2012; de Marneffe et al., 2012; Lee et al., 2015; Stanovsky et al., 2017; Rudinger et al., 2018) – capture inferential interactions between lexical items and syntactic context – *lexicosyntactic inferences* – when trained on current event factuality datasets. Probing these particular systems is useful for understanding neural systems' behavior more generally because (i) the best performing neural models for EFP (Rudinger et al., 2018) are simple instances of common baseline models; and (ii) the task itself is relatively constrained.

To do this, we substantially extend the MegaVeridicality1 dataset (White and Rawlins, 2018) to cover all English clause-embedding verbs in a variety of the syntactic contexts covered by recent psycholinguistic work (White and Rawlins, 2016), and we use the resulting dataset – MegaVeridicality2 – to probe these models' behavior. We focus on clause-embedding verbs because they show effectively every possible patterning of lexicosyntactic inference (Karttunen, 2012).

We discuss three findings: (i) Tree biLSTMs (T-biLSTMs) are better able to correctly predict lexicosyntactic inferences than linear-chain biLSTMs

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4717–4724
Brussels, Belgium, October 31 - November 4, 2018.

(L-biLSTMs); (ii) L-biLSTMs and T-biLSTMs capture different lexicosyntactic inferences, and thus ensembling their predictions can reliably improve performance; and (iii) even when ensembled, these models show systematic errors – e.g. performing well when the polarity of the matrix clause matches the polarity of the true inference, but poorly when these polarities mismatch.

We furthermore release MegaVeridicality2 at MegaAttitude.io as a benchmark for probing the ability of neural systems – whether for factuality prediction or for general natural language inference – to capture lexicosyntactic inference.

2 Data collection

We substantially extend the MegaVeridicality1 dataset (White and Rawlins, 2018), which contains factuality judgments for all English clause-embedding verbs that take tensed subordinate clauses. In White and Rawlins's annotation protocol, all verbs that are grammatical with such subordinate clauses – based on the MegaAttitude dataset (White and Rawlins, 2016) – are slotted into contexts either like (5a) or (5b), depending on whether they take a direct object or not.

(5) a. Someone {knew, didn't know} that a particular thing happened.
b. Someone {was, wasn't} told that a particular thing happened.

For each sentence generated in this way, 10 different annotators are asked to answer the question *did that thing happen?*: *yes*, *maybe or maybe not*, *no*.

There are two important aspects of these contexts to note. First, all lexical items besides the embedding verbs are semantically bleached to ensure that the measured lexicosyntactic inferences are only due to interactions between the embedding predicate – e.g. *know* or *tell* – and the syntactic context. Second, the matrix polarity – i.e. the presence or absence of *not* as a direct dependent of the embedding verb – is manipulated to create two sentences for each verb-context pair.

Our extension, MegaVeridicality2, includes judgments for a variety of infinitival subordinate clause types, exemplified in (6).[1] We investigate infinitival clauses because they can give rise to different lexicosyntactic inferences than finite subordinate clauses – e.g. compare (3) and (4).

(6) a. Someone {needed, didn't need} for a particular thing to happen.
b. Someone {wanted, didn't want} a particular person to do, have a particular thing.
c. Someone {wanted, didn't want} a particular person to have a particular thing.
d. A particular person {was, wasn't} overjoyed to do a particular thing.
e. A particular person {was, wasn't} overjoyed to have a particular thing.
f. A particular person {managed, didn't manage} to do a particular thing.
g. A particular person {managed, didn't manage} to have a particular thing.

Syntactic context	# verbs	# sents	Ex.
NP _ed that S	375	750	(5a)
NP was _ed that S	169	338	(5b)
NP _ed for NP to VP	184	368	(6a)
NP _ed NP to VP[+ev]	197	394	(6b)
NP _ed NP to VP[-ev]	128	256	(6c)
NP was _ed to VP[+ev]	278	556	(6d)
NP was _ed to VP[-ev]	256	512	(6e)
NP _ed to VP[+ev]	217	434	(6f)
NP _ed to VP[-ev]	165	330	(6g)
Total	1,969	3,938	

Table 1: Contexts and number of verbs for which annotations were collected: S = *something happened*, NP = *someone*, VP = *happen*, VP[+ev] = *do something*, VP[-ev] = *have something*. First two rows: MegaVeridicality1. All rows: MegaVeridicality2. The number of sentences is always twice the number of verbs, since matrix polarity is manipulated.

For each sentence, we also collect judgments from 10 different annotators, using the same question as White and Rawlins for context (6a) and modified questions for contexts (6b)-(6g): *did that person do that thing?* for (6b), (6d), and (6f); and *did that person have that thing?* for for (6c), (6e), and (6g). Table 1 shows the number of verb types for each syntactic context. With the polarity manipulation, this yields a total of 3,938 sentences.

To build a factuality prediction test set from these sentences, we combine MegaVeridicality1 with our dataset and replace each instance of *a particular person* or *a particular thing* with *someone* or *something* (respectively). Then, following White and Rawlins, we normalize the 10 responses for each sentence to a single real value using an ordinal mixed model-based procedure. We refer to the resulting dataset as MegaVeridicality2.

3 Model and evaluation

We use MegaVeridicality2 to evaluate the performance of three state-of-the-art neural models of

[1] We also explicitly manipulate two aspects of the subordinate clause in our extension of the MegaVeridicality dataset: (i) how NP embedded subjects are introduced; and (ii) whether the embedded clause contains an eventive predicate (*do*, *happen*) or a stative predicate (*have*). See Appendix A for details on the reasoning behind these manipulations.

event factuality (Rudinger et al., 2018): a linear-chain biLSTM (L-biLSTM), a dependency tree biLSTM (T-biLSTM), and a hybrid biLSTM (H-biLSTM) that ensembles the two. To predict the factuality of the event referred to by a particular predicate, these models pass the output state of the biLSTM at that predicate through a two-layer regression. In the case of the H-biLSTM, the output state of both the L- and T-biLSTMs are simply concatenated and passed through the regression.[2]

Following the multi-task training regime described by Rudinger et al. (2018), we train these models on four standard factuality datasets – FactBank (Saurí and Pustejovsky, 2009, 2012), UW (Lee et al., 2015), MEANTIME (Minard et al., 2016), and UDS (White et al., 2016; Rudinger et al., 2018) – with tied biLSTM weights but regression parameters specific to each dataset. We then use these trained models to predict the factuality of the embedded predicate in our dataset.

To understand how much of these models' performance on our dataset is really due to a correct computation of lexicosyntactic inferences, we also generate predictions for the sentences in our dataset with the embedding verbs UNKed. In this case, the model can rely only on the syntactic context surrounding the predicate to make its inferences. We refer to the models with lexical information as the LEX models and the ones without lexical information as the UNK models.

Each model produces four predictions, corresponding to the four different datasets it was trained on. We consider three different ways of ensembling these predictions using a cross-validated ridge regression: (i) ensembling the four predictions for each specific model (LEX or UNK); (ii) ensembling the predictions for the LEX version of a particular model with the UNK version of that same model (LEX+UNK); and (iii) ensembling the predictions across all models (LEX, UNK, or LEX+UNK). Each ensemble is evaluated in a 10-fold/10-fold nested cross-validation (see Cawley and Talbot, 2010). In each iteration of the outer cross-validation, a 10% test set is split off, and a 10-fold cross-validation to tune the regularization is conducted on the remaining 90%.

4 Results

Figure 1 shows the mean correlation between model predictions and true factuality on the outer fold test sets of the nested cross-validation described in §3. We note three aspects of this plot.

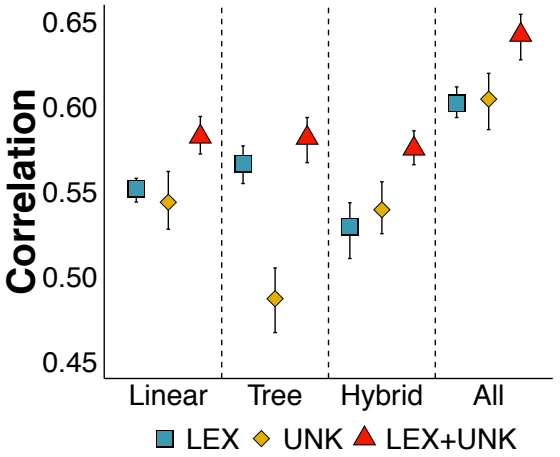

Figure 1: Mean correlation between model predictions and true factuality in nested cross-validation. Error bars show bootstrapped (iter=1,000) 95% confidence intervals for mean correlation across 10 outer folds.

First, among the LEX models, the T-biLSTM performs best, followed by the L-biLSTM, then the H-biLSTM. This is somewhat surprising, since Rudinger et al. find the opposite pattern of performance: the L- and H-biLSTMs vie for dominance, both outperforming the T-biLSTM. This indicates that T-biLSTMs are better able to represent the lexicosyntactic inferences relevant to this dataset, even though they underperform on more general datasets. This possibility is bolstered by the fact that, in contrast to the L- and H-biLSTMs, the LEX version of the T-biLSTMs performs significantly better than the UNK version, suggesting that the T-biLSTM is potentially more reliant on the lexical information than the other two.

Second, when the LEX and UNK version of each model is ensembled (LEX+UNK), we find comparable performance for all three biLSTMs – each outperforming the LEX version of the T-biLSTM. This indicates that each model captures similar amounts of information about lexicosyntactic inference, but this information is captured in the models' parameterizations in different ways.

Finally, when all three models are ensembled, we find that both the LEX and UNK version perform significantly better than any specific LEX+UNK model. This may indicate two things: (i) the models that only have access to syntax can perform just as well as ones that have access to both lexical information and syntax; but (ii) these models appear to capture different aspects of inference, since an ensemble of all models (All-LEX+UNK) performs significantly better than ei-

[2] See Appendix B for further details.

Someone ...	True	Pred.
faked that something happened	-3.15	0.86
was misinformed that something happened	-2.62	1.37
neglected to do something	-3.07	-0.02
pretended to have something	-2.96	0.05
was misjudged to have something	-2.46	0.55
forgot to have something	-3.18	-0.17
neglected to have something	-2.93	0.07
pretended that something happened	-2.11	0.86
declined to do something	-3.18	-0.22
was refused to do something	-3.16	-0.22
refused to do something	-3.12	-0.20
pretended to do something	-3.02	-0.11
disallowed someone to do something	-2.56	0.34
was declined to have something	-2.36	0.55
declined to have something	-3.12	-0.23
did n't hesitate to have something	1.84	-0.96
ceased to have something	-2.22	0.57
did n't hesitate to do something	1.86	-0.92
lied that something happened	-1.99	0.78
feigned to have something	-3.07	-0.31

Table 2: Sentences with the highest prediction errors.

ther the All-LEX or All-UNK ensembles alone.

Interestingly, however, even this ensemble performs more than 10 points worse than each model alone on FactBank, UW, and UDS. This raises the question of which lexicosyntactic inferences these models are missing – investigated below.

5 Analysis

We investigate two questions: (i) which inferences do all models do poorly on?; and (ii) what drives the differing strengths of each model?

Where do all models fail? Table 2 shows the 20 sentences with the highest prediction errors under the All-LEX+UNK ensemble. There are two interesting things to note about these sentences. First, most of them involve negative lexicosyntactic inferences that the model predicts to be either positive or near zero. Second, when the true inference is not positive, the matrix polarity of the original sentence is negative. This suggests that the models are not able to capture inferences whose polarity mismatches the matrix clause polarity.

One question that arises here is whether this inability affects all contexts equally. To answer this, we regress the absolute error of the predictions from this same ensemble (logged and standardized) against true factuality, matrix polarity, and context (as well as all of their two- and three-way interactions).[3] We find that the three-way interactions in this regression are reliable ($\chi^2(8)=27.97$, $p < 0.001$) – suggesting that there are nontrivial differences in these state-of-the-art factuality systems' ability to capture inferential interactions across verbs and syntactic contexts. The differences can be verified visually in Figure 2, which

[3] See Appendix C for further details, including a summary of the regression on which the above discussion is based.

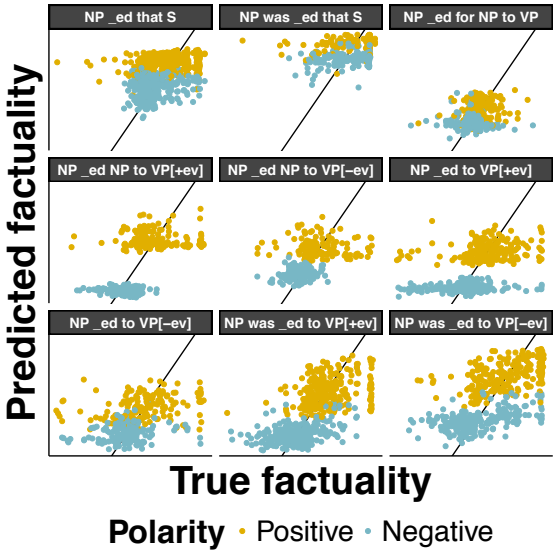

Figure 2: Factuality by syntactic context and polarity, each point a verb. Diagonals show perfect prediction.

plots the factuality predicted by this ensemble against the true factuality from MegaVeridicality2.

To elaborate, the ensemble does best overall on contexts like (7a) and (7b), and worst overall on contexts like (7c). The contrast between (7b) and (7c) is particularly interesting because (i) (7c) is just the passivized form of (7b); and (ii) we do not observe similar behavior for contexts (7d) and (7e), which are analogous to (7b) and (7c), but replace the stative *have* with the eventive *do*.

(7) Someone...
  a. {_ed, didn't _} for something to happen.
  b. {_ed, didn't _} someone to have something.
  c. {was _ed, wasn't _ed} to have something.
  d. {_ed, didn't _} someone to do something.
  e. {was _ed, wasn't _ed} to do something.
  f. {_ed, didn't _} that something happened.

An additional nuance is that the ensemble does reliably better on the negative matrix polarity version of (7b) than on the positive, with the opposite true for (7e). This suggests these models do not capture an important inferential interaction between passivization and eventivity.

This suggestion is further bolstered by the fact that the ensemble's ability to predict cases where the matrix polarity mismatches the true factuality are reliably poorer in context (7c) but not in its minimal pairs (7e) and (7b), where the ensemble performs reliably poorer when the two match. Indeed, it is contexts (7c) and (7f) that drive the polarity mismatch effect evident in Table 2.

What drives differences between models? In §4, we noted two ways that the biLSTMs we in-

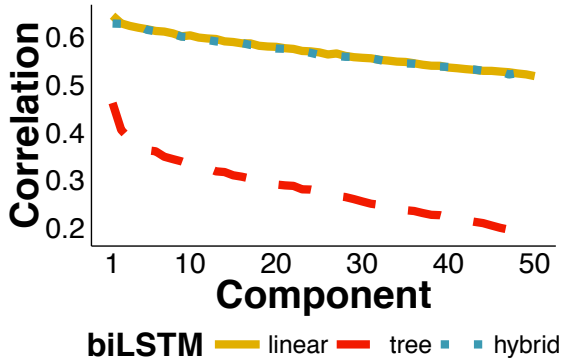

Figure 3: Canonical correlations between embedding verb embeddings and embedded verb hidden states.

vestigate differ: (i) the T-biLSTM appears to be more reliant on lexical information than L- and H-biLSTMs; and (ii) each model appears to encode information about lexicosyntactic inference in its parameterizations in different ways. We hypothesize that these two differences are related – specifically, that the T-biLSTM's heavier reliance on lexical information comes about as a consequence of stronger entanglement between lexical and syntactic information in its hidden states.

To probe this, we ask to what extent the embedding verb's embedding can be recovered from the embedded verb's hidden state using linear functions. If the lexical information is more strongly entangled with the syntactic information, it should be more difficult to construct a homomorphic (linear) function to decode the embedding verb's embedding from the embedded verb's hidden state. To measure this, we conduct a Canonical Correlation Analysis (CCA; Hotelling, 1936) between these two vector space representations for every sentence in our dataset. Given two matrices $\mathbf{X}$ (the embedding verb embeddings column stacked) and $\mathbf{Y}$ (the embedded verb hidden states column stacked), CCA constructs matrices $\mathbf{A}$ and $\mathbf{B}$, such that $\mathbf{a}_i, \mathbf{b}_i = \arg_{\mathbf{a}',\mathbf{b}'}\max\ \text{corr}(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y})$ and $\text{corr}(\mathbf{a}_i\mathbf{X}, \mathbf{a}_j\mathbf{X}) = \text{corr}(\mathbf{b}_i\mathbf{Y}, \mathbf{b}_j\mathbf{Y}) = 0, \forall i < j$. This guarantees that the canonical correlation at component i, $\text{corr}(\mathbf{a}_i\mathbf{X}, \mathbf{b}_i\mathbf{Y})$, is nonincreasing in i, and thus the linearly decodable information about $\mathbf{Y}$ in $\mathbf{X}$ can be assessed using this function.

Figure 3 plots the canonical correlations for the first 50 components for each of the biLSTMs we investigated. We find that the canonical correlations associated with the T-biLSTM are substantially lower than those associated with the L- and H-biLSTMs across these first 50 components. This suggests that the T-biLSTM more strongly entangles lexical and syntactic information, perhaps explaining its apparently heavier reliance on lexical information, observed in §4.

Of note here is that the pattern seen in Figure 3 is probably at least partly a consequence of the different nonlinearities used for the L-biLSTM (tanh) and T-biLSTM (ReLU), and not the architectures themselves. But whether or not this pattern is due to the architectures, nonlinearities, or both, the entanglement hypothesis may still help explain the pattern of results discussed in §4.

6 Related work

This work is inspired by recent work in *recasting* various semantic annotations into natural language inference (NLI) datasets (White et al., 2017; Poliak et al., 2018a,b; Wang et al., 2018) to gain a better understanding of which phenomena standard neural NLI models (Bowman et al., 2015; Conneau et al., 2017) can capture – a line of work with deep roots (Cooper et al., 1996). The experimental setup – specifically, the idea of UNKing the embedding verb – was inspired by recent work that uses hypothesis-only baselines for a similar purpose (Gururangan et al., 2018; Poliak et al., 2018c; Tsuchiya, 2018). This work is also related to the broader investigation of sentence representations – particularly, tasks aimed at probing these representations' content (Pavlick and Callison-Burch, 2016; Adi et al., 2016; Conneau et al., 2018; Conneau and Kiela, 2018; Dasgupta et al., 2018).

7 Conclusion

We investigated neural models' ability to capture lexicosyntactic inference, taking the task of event factuality prediction (EFP) as a case study. We built a factuality judgment dataset for all English clause-embedding verbs in various syntactic contexts and used this dataset to probe current state-of-the-art EFP systems. We showed that these systems make certain systematic errors that are clearly visible through the lens of factuality.

Acknowledgments

This research was supported by the JHU HLTCOE, DARPA LORELEI and AIDA, NSF-BCS (1748969/1749025), and NSF-GRFP (1232825). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Dorit Abusch. 2002. Lexical alternatives as a source of pragmatic presuppositions. *Semantics and Linguistic Theory*, 12:1–19.

Dorit Abusch. 2010. Presupposition triggering from alternatives. *Journal of Semantics*, 27(1):37–80.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France.

Pranav Anand and Valentine Hacquard. 2013. Epistemics and attitudes. *Semantics and Pragmatics*, 6(8):1–59.

Pranav Anand and Valentine Hacquard. 2014. Factivity, belief and discourse. In Luka Crnič and Uli Sauerland, editors, *The Art and Craft of Semantics: A Festschrift for Irene Heim*, volume 1, pages 69–90. MIT Working Papers in Linguistics, Cambridge, MA.

Rebekah Baglini and Itamar Francez. 2016. The implications of managing. *Journal of Semantics*, 33(3):541–560.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Samuel R Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1466–1477, Berlin, Germany. Association for Computational Linguistics.

Željko Bošković. 1996. Selection and the Categorial Status of Infinitival Complements. *Natural Language & Linguistic Theory*, 14(2):269–304.

Željko Bošković. 1997. *The syntax of nonfinite complementation: An economy approach*. 32. MIT Press, Cambridge, MA.

Gavin C. Cawley and Nicola L.C. Talbot. 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.*, 11:2079–2107.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:2126–2136.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the Framework. Technical Report LRE 62-051, The FRACAS Consortium.

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.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2018. Evaluating Compositionality in Sentence Embeddings. *arXiv:1802.04302*.

Jon Robert Gajewski. 2007. Neg-raising and polarity. *Linguistics and Philosophy*, 30(3):289–328.

Thomas Angelo Grano. 2012. *Control and Restructuring at the Syntax-Semantics Interface*. Ph.D. thesis, University of Chicago.

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 273–278. IEEE.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. *arXiv:1803.02324*.

Irene Heim. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of Semantics*, 9(3):183–221.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Laurence Robert Horn. 1972. *On the Semantic Properties of Logical Operators in English*. Ph.D. thesis, UCLA.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 633–644.

Lauri Karttunen. 1971a. Implicative verbs. *Language*, pages 340–358.

Lauri Karttunen. 1971b. Some observations on factivity. *Papers in Linguistics*, 4(1):55–69.

Lauri Karttunen. 2012. Simple and phrasal implicatives. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 124–131. Association for Computational Linguistics.

Lauri Karttunen. 2013. You will be lucky to break even. In Tracy Holloway King and Valeria dePaiva, editors, *From Quirky Case to Representing Space: Papers in Honor of Annie Zaenen*, pages 167–180.

Lauri Karttunen and Stanley Peters. 1979. Conventional implicature. *Syntax and Semantics*, 11:1–56.

Lauri Karttunen, Stanley Peters, Annie Zaenen, and Cleo Condoravdi. 2014. The Chameleon-like Nature of Evaluative Adjectives. In *Empirical Issues in Syntax and Semantics 10*, pages 233–250. CSSP-CNRS.

Paul Kiparsky and Carol Kiparsky. 1970. Fact. In Manfred Bierwisch and Karl Erich Heidolph, editors, *Progress in Linguistics: A collection of papers*, pages 143–173. Mouton, The Hague.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event Detection and Factuality Assessment with Non-Expert Supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal. Association for Computational Linguistics.

Noor van Leusen. 2012. The Accommodation Potential of Implicative Verbs. In *Logic, Language and Meaning*, volume 7218 of *Lecture Notes in Computer Science*, pages 421–430. Springer.

Bill MacCartney, Michel Galley, and Christopher D Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the conference on empirical methods in natural language processing*, pages 802–811. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Roger Martin. 2001. Null case and the distribution of PRO. *Linguistic Inquiry*, 32(1):141–166.

Roger Andrew Martin. 1996. *A minimalist theory of PRO and control*. Ph.D. thesis, University of Connecticut, Storrs.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 23–28, Paris, France. European Language Resources Association (ELRA).

Makoto Miwa and Mohit Bansal. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Prerna Nadathur. 2016. Causal necessity and sufficiency in implicativity. *Semantics and Linguistic Theory*, 26:1002–1021.

Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. 2006. Computing relative polarity for textual inference. In *Proceedings of the Fifth International Workshop on Inference in Computational Semantics (ICoS-5)*, pages 20–21, Buxton, England. Association for Computational Linguistics.

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional Entailment in Adjective-Nouns. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:2164–2173.

David Pesetsky. 1991. Zero syntax: vol. 2: Infinitives.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018a. On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 513–523, New Orleans.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White,

and Benjamin Van Durme. 2018b. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018c. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 180–191, New Orleans. Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural Models of Factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 731–744, New Orleans. Association for Computational Linguistics.

Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Mandy Simons. 2001. On the conversational basis of some presuppositions. *Semantics and Linguistic Theory*, 11:431–448.

Mandy Simons. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117(6):1034–1056.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. *Semantics and linguistic theory*, 20:309–327.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1):207–218.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating Deep Linguistic Features in Factuality Prediction over Unified Datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–357. Association for Computational Linguistics.

Tim Stowell. 1982. The tense of infinitives. *Linguistic Inquiry*, 13(3):561–570.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

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 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv:1804.07461*.

Aaron Steven White. 2014. Factive-implicatives and modalized complements. In *Proceedings of the 44th annual meeting of the North East Linguistic Society*, pages 267–278, University of Connecticut.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is Everything: Recasting Semantic Resources into a Unified Evaluation Framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 996–1005.

Aaron Steven White and Kyle Rawlins. 2016. A computational model of S-selection. *Semantics and Linguistic Theory*, 26:641–663.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, page to appear, Amherst, MA. GLSA Publications.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, TX. Association for Computational Linguistics.

Susi Wurmbrand. 2014. Tense and aspect in English infinitives. *Linguistic Inquiry*, 45(3):403–447.

Wojciech Zaremba and Ilya Sutskever. 2014. Learning to execute. *arXiv preprint arXiv:1410.4615*.