Lexicosyntactic Inference in Neural Models

Anonymous EMNLP submission

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 predicates 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 MegaVeridicality 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 it to probe the behavior of these models. 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 (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 – 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 our new dataset at `url:anon` as a benchmark for probing the ability of neural systems – whether systems for factuality prediction or for more general natual language inference – to capture lexicosyntactic inference.

2 Data collection

We substantially extend the MegaVeridicality dataset (White and Rawlins, 2018), which contains factuality judgments for all English clause-embedding verbs that take finite 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*.

An important aspect of these contexts is that 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.

We extend White and Rawlins's dataset by collecting 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 – see, e.g., (3)-(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. A particular person {was, wasn't} overjoyed to {do, have} a particular thing.
d. A particular person {managed, didn't manage} to {do, have} a particular thing.

[1] See Appendix A for further details.

Frame	# verbs	Ex.
NP _ed that S	375	(5a)
NP was _ed that S	169	(5b)
NP _ed for NP to VP	184	(6a)
NP _ed NP to VP[+ev]	197	(6b)
NP _ed NP to VP[-ev]	128	(6b)
NP was _ed to VP[+ev]	278	(6c)
NP was _ed to VP[-ev]	256	(6c)
NP _ed to VP[+ev]	217	(6d)
NP _ed to VP[-ev]	165	(6d)
Total	1,969	

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*. The first two rows derive from White and Rawlins 2018; the remainder derive from this work.

For each sentence, we also collected judgments from 10 different annotators, using slightly modified questions, depending on the sentence but the same response options. Table 1 shows the number of verb types for each syntactic context.

To build a factuality prediction test set from these sentences, we combine MegaVeridicality 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.

3 Model and evaluation

We use our lexicosyntactic inference dataset to evaluate the performance of three neural models for 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 regession parameters specific to each dataset. We

[2] See Appendix B for further details.

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.[3] 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.

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. (2018) find the opposite pattern of performance, with the H-biLSTM outperforming the L-biLSTM, and the L-biLSTM 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 T- and H-biLSTMs.

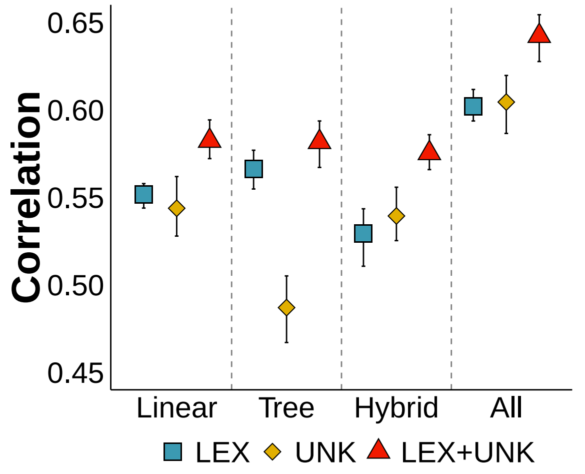

Figure 1: Mean correlation between model predictions and true factuality in nested cross-validation. Error bars show bootstrapped (iter=1,000) 95% confidence intervals for mean correlation across 10 outer folds.

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 have only 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 either the All-LEX or All-UNK ensembles alone.

5 Analysis

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 ma-

[3] We use the same UNKing method used by Rudinger et al. (2018): a single UNK vector is randomly generated at train time, and all OOV items are mapped to it. For the UNK models, we map all the embedding verbs to this vector at test.

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.

trix clause polarity. This inability to predict mismatching inferences is perhaps unsurprising since the majority of inferences match the matrix clause polarity, evidenced in Figure 2.

Figure 2 plots the factuality predicted by the the best performing ensemble (All-LEX+UNK) against the true factuality, broken out by frame and polarity. Table 3 shows the corresponding correlations for each biLSTM.

	Linear		Tree		Hybrid	
	pos	neg	pos	neg	pos	neg
NP _ed that S	**0.25**	-0.02	0.19	**0.12**	0.19	0.10
NP was _ed that S	0.11	0.20	0.08	0.17	**0.23**	**0.24**
NP _ed for NP to VP	**0.26**	**-0.02**	-0.00	-0.06	-0.00	-0.04
NP _ed NP to VP[+ev]	-0.04	-0.20	**0.04**	0.20	-0.08	**0.22**
NP _ed NP to VP[-ev]	-0.09	-0.01	**0.00**	**0.24**	-0.08	0.16
NP was _ed to VP[+ev]	0.21	0.24	**0.29**	0.38	0.26	**0.41**
NP was _ed to VP[-ev]	0.36	0.13	**0.44**	0.43	0.40	**0.55**
NP _ed to VP[+ev]	0.09	0.14	**0.20**	**0.23**	-0.06	0.02
NP _ed to VP[-ev]	0.24	0.13	**0.25**	**0.22**	0.12	0.06

Table 3: Correlation between predictions from LEX+UNK model and true factuality in nested cross-validation by biLSTM, frame, and polarity. Bolding shows best performance on positive and best performance on negative in each row.

We find that there is high variability in which model best predicts inferences in particular syntactic contexts. This may be why the ensemble of all biLSTMs is able to outperform any particular model, and it suggests that particular biLSTMs are better at representing interactions between negation, lexical items, and certain syntactic structures.

The is corroborated in analysis of particular items. For each biLSTM we extracted the items that that model showed the lowest absolute error on in comparison to the other models. For the L-biLSTM, this list was dominated by sentences like (7a), which the L-biLSTM does best on overall (see Table 3). In contrast, the T-biLSTM shows more variety in the interactions it captures – including sentences like (7b), which the H-biLSTM tended to perform better on overall.

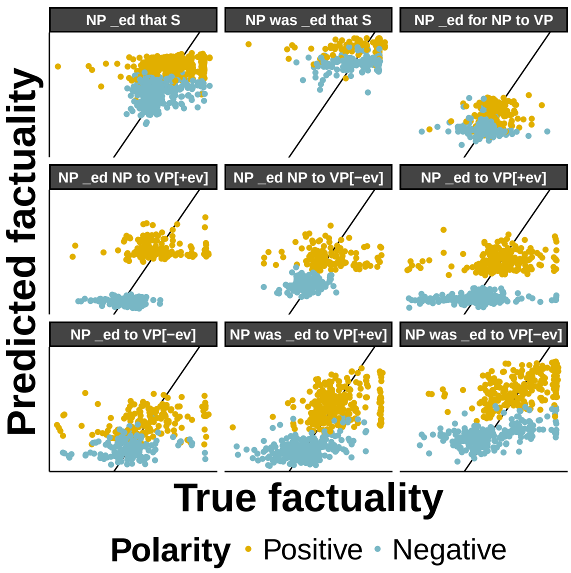

Figure 2: Factuality by syntactic context and polarity, each point a verb. Diagonals show perfect prediction.

(7) Someone...

a. didn't mandate for something to happen.

b. wasn't excited to do something.

This suggests that L-biLSTMs might fruitfully be used to target specific lexicosyntactic inferences, while others T-biLSTMs might be used to capture more general patterns of lexicosyntactic inference. A remaining question is whether other forms of lexicosyntactic inference show similar patterns.

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; 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. It is also related to work that uses hypothesis-only baselines for a similar purpose (Gururangan et al., 2018; Poliak et al., 2018b; Tsuchiya, 2018).

7 Conclusion

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

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.

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, 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.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

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 [cs]*. 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.

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*, 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.

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

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018a. Towards a Unified Natural Language Inference Framework to Evaluate Sentence Representations. *arXiv:1804.08207 [cs]*. ArXiv: 1804.08207.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis Only Baselines in Natural Language Inference. *arXiv:1805.01042 [cs]*. ArXiv: 1805.01042.

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*, 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.

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 [cs]*. 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*.

A Data collection

We manipulate two aspects of the subordinate clause in our extension of the MegaVeridicality dataset: (i) whether and how an NP embedded subject is introduced; and (ii) whether the embedded clause contains an eventive predicate (*do*, *happen*) or a stative predicate (*have*).

The first manipulation is known to give rise to different inferential interactions for predicates that take different kinds of infinitival subordinate clauses – e.g. *remember*, *forget*. For example, while (8a), (8b), and (9a) trigger the inference (11a), (9b) triggers the inference (11b). And just a slight tweak to (9a) and (9b) can make these inferences go away completely: neither (10a) nor (10b) trigger an inference to either (11a) or (11b).

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

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

(10) a. Jo remembered Bo to have left.
b. Jo didn't remember Bo to have left.

(11) a. Bo left.
b. Bo didn't leave.

The second manipulation is known to give rise to importantly different temporal interpretations, which also seem to affect factuality judgments (White, 2014). For instance, *believe* is generally rated more acceptable in sentences with stative embedded predicates, like (12a), and less acceptable in sentences with eventive embedded predicates, like (12b).

(12) a. Jo believe Mo to be intelligent.
b.?Jo believed Mo to run around the park.

This appears to correlate with certain aspects of the temporal interpretation of such sentences (Stowell, 1982; Pesetsky, 1991; Bošković, 1996, 1997; Martin, 1996, 2001; Grano, 2012; Wurmbrand, 2014).

To accommodate the differences between these contexts and theirs, we use a slightly modified question, depending on the sentence – *did that person do that thing?*; *did that person have that thing?*; or *did that thing happen?* – with the same response options.

B Model and evaluation

We use three models for event factuality prediction proposed by Rudinger et al. (2018): a stacked bidirectional linear-chain LSTM (L-biLSTM), a stacked bidirectional dependency tree LSTM (T-biLSTM), and a simple ensemble of the two that Rudinger et al. refer to as a H(ybrid)-biLSTM. We use the two-layer version of these biLSTMs here.

B.1 Stacked bidirectional linear LSTM

The L-biLSTM we use is a standard extension of the unidirectional linear-chain LSTM (Hochreiter and Schmidhuber, 1997) by adding the notion of a layer $l \in \{1, \ldots, L\}$ and a direction $d \in \{\rightarrow, \leftarrow\}$ (Graves et al., 2013; Sutskever et al., 2014; Zaremba and Sutskever, 2014).

$$\begin{aligned}
\mathbf{f}_t^{(l,d)} &= \sigma\left(\mathbf{W}_{\mathrm{f}}^{(l,d)}\left[\mathbf{h}_{\mathbf{prev}_d(t)}^{(l,d)}; \mathbf{x}_t^{(l,d)}\right] + \mathbf{b}_{\mathrm{f}}^{(l,d)}\right)\\
\mathbf{i}_t^{(l,d)} &= \sigma\left(\mathbf{W}_{\mathrm{i}}^{(l,d)}\left[\mathbf{h}_{\mathbf{prev}_d(t)}^{(l,d)}; \mathbf{x}_t^{(l,d)}\right] + \mathbf{b}_{\mathrm{i}}^{(l,d)}\right)\\
\mathbf{o}_t^{(l,d)} &= \sigma\left(\mathbf{W}_{\mathrm{o}}^{(l,d)}\left[\mathbf{h}_{\mathbf{prev}_d(t)}^{(l,d)}; \mathbf{x}_t^{(l,d)}\right] + \mathbf{b}_{\mathrm{o}}^{(l,d)}\right)\\
\hat{\mathbf{c}}_t^{(l,d)} &= g\left(\mathbf{W}_{\mathrm{c}}^{(l,d)}\left[\mathbf{h}_{\mathbf{prev}_d(t)}^{(l,d)}; \mathbf{x}_t^{(l,d)}\right] + \mathbf{b}_{\mathrm{c}}^{(l,d)}\right)\\
\mathbf{c}_t^{(l,d)} &= \mathbf{i}_t^{(l,d)} \circ \hat{\mathbf{c}}_t^{(l,d)} + \mathbf{f}_t^{(l,d)} \circ \mathbf{c}_{\mathbf{prev}_d(t)}^{(l,d)}\\
\mathbf{h}_t^{(l,d)} &= \mathbf{o}_t^{(l,d)} \circ g\left(\mathbf{c}_t^{(l,d)}\right)
\end{aligned}$$

where $\circ$ is the Hadamard product; $\mathbf{prev}_{\rightarrow}(t) = t - 1$ and $\mathbf{prev}_{\leftarrow}(t) = t + 1$, and $\mathbf{x}_t^{(l,d)} = \mathbf{x}_t$ if $l = 1$; and $\mathbf{x}_t^{(l,d)} = [\mathbf{h}_t^{(l-1,\rightarrow)}; \mathbf{h}_t^{(l-1,\leftarrow)}]$ otherwise. We follow Rudinger et al. in setting g to the pointwise nonlinearity tanh.

B.2 Stacked bidirectional tree LSTM

Rudinger et al. (2018) propose a stacked bidirectional extension to the child-sum dependency tree LSTM (T-LSTM; Tai et al., 2015). The T-LSTM redefines $\mathbf{prev}_{\rightarrow}(t)$ to return the set of indices that correspond to the children of w_t in some dependency tree. In the case of multiple children one defines $\mathbf{f}_{tk}$ for each child index $k \in \mathbf{prev}_{\rightarrow}(t)$ in a way analogous to the equations in §B.1 – i.e. as though each child were the only child – and then sums across k within the equations for $\mathbf{i}_t$, $\mathbf{o}_t$, $\hat{\mathbf{c}}_t$, $\mathbf{c}_t$, and $\mathbf{h}_t$.

Rudinger et al.'s stacked bidirectional T-biLSTM extends the T-LSTM with a *downward* computation in terms of a $\mathbf{prev}_{\leftarrow}(t)$ that returns the set of indices that correspond to the *parents* of w_t in some dependency tree.[4] The same method for combining children in the upward computation is then used for combining parents in the downward computation.

$$\begin{aligned}
\mathbf{f}_{tk}^{(l,d)} &= \sigma\left(\mathbf{W}_{\mathrm{f}}^{(l,d)}\left[\mathbf{h}_k^{(l,d)}; \mathbf{x}_t^{(l,d)}\right] + \mathbf{b}_{\mathrm{f}}^{(l,d)}\right)\\
\hat{\mathbf{h}}_t^{(l,d)} &= \sum_{k \in \mathbf{prev}_d(t)} \mathbf{h}_k^{(l,d)}\\
\mathbf{i}_t^{(l,d)} &= \sigma\left(\mathbf{W}_{\mathrm{i}}^{(l,d)}\left[\hat{\mathbf{h}}_t^{(l,d)}; \mathbf{x}_t^{(l,d)}\right] + \mathbf{b}_{\mathrm{i}}^{(l,d)}\right)\\
\mathbf{o}_t^{(l,d)} &= \sigma\left(\mathbf{W}_{\mathrm{o}}^{(l,d)}\left[\hat{\mathbf{h}}_t^{(l,d)}; \mathbf{x}_t^{(l,d)}\right] + \mathbf{b}_{\mathrm{o}}^{(l,d)}\right)\\
\hat{\mathbf{c}}_t^{(l,d)} &= g\left(\mathbf{W}_{\mathrm{c}}^{(l,d)}\left[\hat{\mathbf{h}}_t^{(l,d)}; \mathbf{x}_t^{(l,d)}\right] + \mathbf{b}_{\mathrm{c}}^{(l,d)}\right)\\
\mathbf{c}_t^{(l,d)} &= \mathbf{i}_t^{(l,d)} \circ \hat{\mathbf{c}}_t^{(l,d)} + \sum_{k \in \mathbf{prev}_d(t)} \mathbf{f}_{tk}^{(l,d)} \circ \mathbf{c}_k^{(l,d)}\\
\mathbf{h}_t^{(l,d)} &= \mathbf{o}_t^{(l,d)} \circ g\left(\mathbf{c}_t^{(l,d)}\right)
\end{aligned}$$

We follow Rudinger et al. in using a ReLU pointwise nonlinearity for g, and in contrast to other dependency tree-structured T-LSTMs (Socher et al., 2014; Iyyer et al., 2014), not using the dependency labels in any way to make the L- and T-biLSTMs as comparable as possible.

B.3 Regression model

To predict the factuality v_t for the event referred to by a word w_t, we follow Rudinger et al. (2018) in using the hidden states from the final layer of the stacked L- or T-biLSTM as the input to a two-layer regression model.

$$\mathbf{h}_t^{(L)} = [\mathbf{h}_t^{(L,\rightarrow)}; \mathbf{h}_t^{(L,\leftarrow)}]$$
$$\hat{v}_t = \mathbf{V}_2\, g\left(\mathbf{V}_1 \mathbf{h}_t^{(L)} + \mathbf{b}_1\right) + \mathbf{b}_2$$

where $\hat{v}_t$ is passed to a loss function $\mathbb{L}(\hat{v}_t, v_t)$. we follow Rudinger et al. (2018) in using smooth L1 for $\mathbb{L}$ and a ReLU pointwise nonlinearity for g.

We also use the simple ensemble method proposed by Rudinger et al. (2018), which they call the H(ybrid)-biLSTM. In this hybrid, the hidden states from the final layers of both the stacked L-biLSTM and the stacked T-biLSTM are concatenated and passed through the same two-layer regression model (cf. Miwa and Bansal, 2016; Bowman et al., 2016).

B.4 Ensemble model

We use a ridge regression to ensemble the predictions from various models. The regularization hyperparameter was tuned in the inner fold of the nested cross-validation described in §3 using exhaustive grid search over $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1., 2., 5., 10., 100.\}$.

[4] Miwa and Bansal (2016) propose a similar extension for constituency trees.