## A Data collection

We manipulate two aspects of the subordinate clause in our extension of the MegaVeridicality1 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).

## 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).

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

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.

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

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

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

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 Out of vocabulary

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 described in §3, we map all the embedding verbs to this vector at test.

### B.5 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 $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1., 2., 5., 10., 100.\}$.

## C Regression analysis

We regress the absolute error of the predictions from the All-LEX+UNK ensemble (logged and standardized) against true factuality, matrix polarity, and frame (as well as all of their two- and three-way interactions) using a linear mixed effects model with random intercepts for verb and by-verb random slopes for matrix polarity. Table 3 summarizes the fixed effect coefficients based on a sum coding of matrix polarity (negative = -1, positive = 1) and context (NP was _ed that S = -1).

| | Coef $\beta$ | SE($\beta$) | t |
|---|---|---|---|
| (Intercept) | 0.00 | 0.03 | 0.1 |
| polarity | 0.15 | 0.02 | 6.2 |
| factuality | 0.00 | 0.03 | 0.1 |
| NP _ed to VP[+ev] | −0.07 | 0.05 | −1.3 |
| NP _ed to VP[-ev] | −0.04 | 0.06 | −0.6 |
| NP was _ed to VP[+ev] | 0.02 | 0.05 | 0.3 |
| NP was _ed to VP[-ev] | 0.23 | 0.05 | 4.7 |
| NP _ed NP to VP[+ev] | −0.01 | 0.07 | −0.1 |
| NP _ed NP to VP[-ev] | −0.30 | 0.08 | −3.8 |
| NP _ed for NP to VP | −0.34 | 0.07 | −5.2 |
| NP _ed that S | 0.09 | 0.04 | 2.1 |
| polarity:factuality | 0.02 | 0.03 | 0.7 |
| polarity:NP _ed to VP[+ev] | −0.05 | 0.05 | −0.8 |
| polarity:NP _ed to VP[-ev] | 0.03 | 0.06 | 0.4 |
| polarity:NP was _ed to VP[+ev] | −0.20 | 0.05 | −4.0 |
| polarity:NP was _ed to VP[-ev] | −0.09 | 0.05 | −1.8 |
| polarity:NP _ed NP to VP[+ev] | −0.06 | 0.07 | −0.8 |
| polarity:NP _ed NP to VP[-ev] | 0.28 | 0.08 | 3.4 |
| polarity:NP _ed for NP to VP | 0.01 | 0.07 | 0.1 |
| polarity:NP _ed that S | 0.08 | 0.04 | 1.8 |
| factuality:NP _ed to VP[+ev] | −0.04 | 0.05 | −0.9 |
| factuality:NP _ed to VP[-ev] | −0.04 | 0.06 | −0.7 |
| factuality:NP was _ed to VP[+ev] | 0.09 | 0.05 | 1.7 |
| factuality:NP was _ed to VP[-ev] | 0.06 | 0.05 | 1.2 |
| factuality:NP _ed NP to VP[+ev] | 0.17 | 0.08 | 2.1 |
| factuality:NP _ed NP to VP[-ev] | −0.18 | 0.10 | −1.7 |
| factuality:NP _ed for NP to VP | 0.13 | 0.10 | 1.3 |
| factuality:NP _ed that S | 0.03 | 0.05 | 0.5 |
| polarity:factuality:NP _ed to VP[+ev] | 0.06 | 0.05 | 1.3 |
| polarity:factuality:NP _ed to VP[-ev] | 0.02 | 0.06 | 0.3 |
| polarity:factuality:NP was _ed to VP[+ev] | 0.07 | 0.05 | 1.4 |
| polarity:factuality:NP was _ed to VP[-ev] | −0.14 | 0.05 | −3.0 |
| polarity:factuality:NP _ed NP to VP[+ev] | −0.05 | 0.08 | −0.6 |
| polarity:factuality:NP _ed NP to VP[-ev] | 0.28 | 0.10 | 2.7 |
| polarity:factuality:NP _ed for NP to VP | 0.12 | 0.10 | 1.2 |
| polarity:factuality:NP _ed that S | −0.17 | 0.05 | −3.2 |

Table 3: Fixed effects from regression analysis

The estimated standard deviation for the verb random intercepts is 0.30, and the estimated standard deviation for the by-verb random slopes for polarity is 0.22. Their estimated correlation between the two is 0.30. The marginal $R^2$ is 0.05 and the conditional $R^2$ is 0.20 (Nakagawa and Schielzeth, 2013).