Contrafactives, Learnability, and Production

David Strohmaier, Simon Wimmer
University of Cambridge, University of Düsseldorf

Background

x factives p entails that x believes p and presupposes p is true.
x contrafactives p entails that x believes p and presupposes p is false.
Natural languages appear to universally feature factive verbs like *know* (Goddard, 2010). No clear example of a contrafactive has been found yet (see, e.g., Glass, 2023; Holton, 2017; Roberts and Özyildiz, 2023).

Previous Proposal

Contrafactives are harder to learn.

Evidence

Small effects when training transformer models to predict the truth value of sentences using factives and contrafactives (Strohmaier and Wimmer, 2022, 2023).

New Experiment

A full transformer-model trained on an artificial language including factives, contrafactives, and non-factives.

Improvements over previous work

- Model production: generate sentences from input
- Account for presupposition failure

The model is implemented as taking an input sequence and returning an output sequence. The input sequence specifies the conditions for producing an attitude ascription sentence (the output).

INPUT

Truth value of ascription — *Correspondence of attitude content to world*

a) True True = eat rory tomato basil soup lunch today
Assumed truth value of attitude content — *Attitude content*

b) True False != buy ahab carrot oregano pie dinner yesterday

c) P-failure False ? order lane mushroom pepper stew breakfast tomorrow

OUTPUT: Attitude Ascription

a) factive rory eat tomato-basil soup for lunch today

b) contrafactive ahab bought carrot-oregano pie for dinner yesterday

c) factive lane will-order mushroom-pepper stew for breakfast tomorrow

- Input and output vocabulary differ and correspondence must be learned.
- For some inputs more than one output is evaluated as correct.
  E.g. when the conditions require a false ascription, while the attitude content is assumed to be true, and corresponds to the world, *any* factive ascription that does not match the attitude content is correct.

Setup

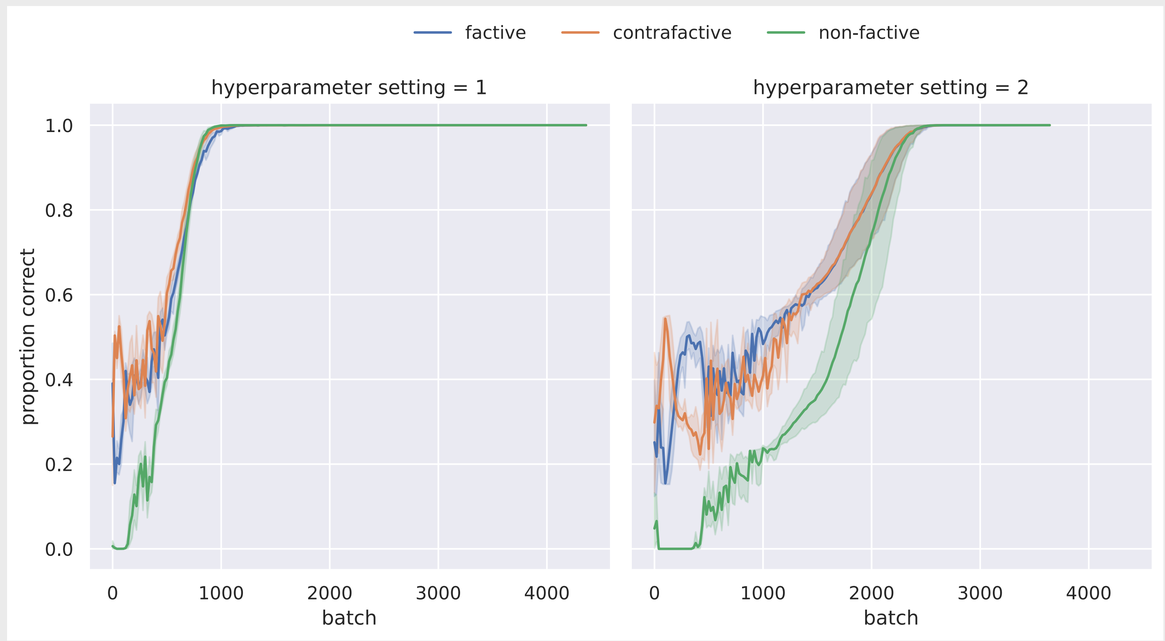
After an initial hyperparameter search, we selected two sets of hyperparameters. For each set, we trained and evaluated the model multiple times with different random seeds. To document the learning process, the evaluation is run after every 20 batches of training.

Funding Info

This paper reports on research supported by Cambridge University Press & Assessment. We thank the NVIDIA Corporation for the donation of the Titan X Pascal GPU used in the exploration phase of this research.

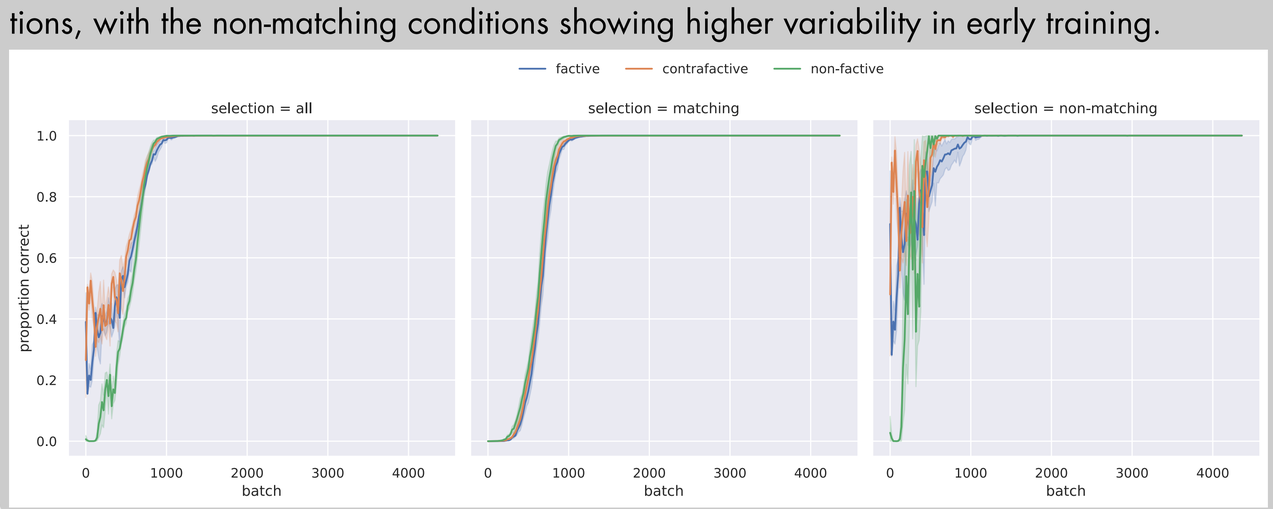
Results: Contrafactives not Harder to Learn

If anything, non-factives are harder to learn.



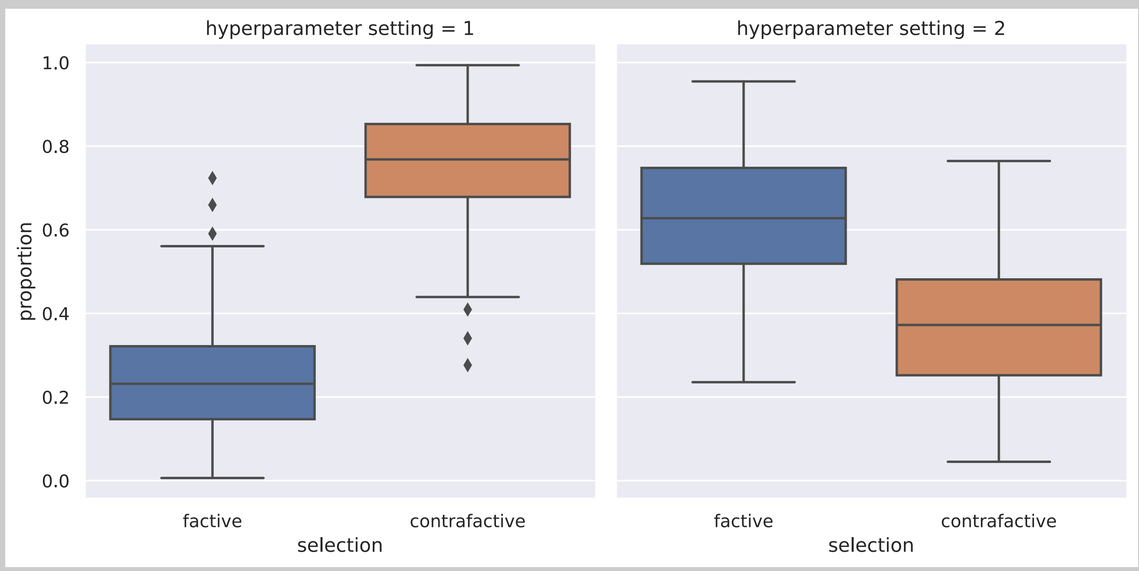
Interpretation: Matching vs. Non-Matching

The conditions for producing a sentence can be distinguished between **(a)** those requiring the production of sentence matching the attitude content and **(b)** those requiring the divergence between sentence and content.
The sentence matches the content iff it describes the same state of affairs with regard to the meal. If the sentence must not match the content, the model has more freedom in choosing which sentence it produces. As a result, the learning dynamics differ between these conditions, with the non-matching conditions showing higher variability in early training.



Selection Preferences

When the truth value of the attitude content is unknown, the use of both factive and contrafactive verbs yields presupposition failure. Thus, the models could exhibit a selection preference between the two verbs. Considering such selection after the model has stabilised (batch >3000), the preference appears to depend on the hyperparameters (see *Setup*).



Bibliography

Glass, L. (2023). "The negatively biased Manadarin belief verg y wéi". In: *Studia Linguistica* 77.1, pp. 1–46. doi: 10.1111/stul.12202.
Goddard, C. (2010). "Universals and Variation in the Lexicon of Mental State Concepts". In: *Words and the Mind: How words capture human experience*. Oxford: Oxford University Press.
Holton, R. (2017). "I—Facts, Factives, and Contrafactives". In: *Aristotelian Society Supplementary Volume* 91.1, pp. 245–266. doi: 10.1093/arisup/akx003.
Roberts, T. and D. Özyildiz (2023). "Bad attitudes: Impossible meanings and the false belief gap".
Strohmaier, D. and S. Wimmer (2022). "Contrafactives and Learnability". In: *Proceedings of the 23rd Amsterdam Colloquium*. Ed. by M. Degano et al. Amsterdam, pp. 298–305.
– (2023). "Contrafactives and Learnability: An Experiment with Propositional Constants". In: *Logic and Engineering of Natural Language Semantics*. Ed. by D. Bekki, K. Mineshima, and E. McCready. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, pp. 67–82. doi: 10.1007/978-3-031-43977-3_5.

david.strohmaier@cl.cam.ac.uk simon.wimmer@hhu.de