Supplementary Material

Modeling Physical Semantic Plausibility by Injecting World Knowledge

1 World Knowledge Features

The six feature types on which we collected annotation are defined as follows:

- SENTIENCE. The "livingness" and intelligence of an entity. For example, the non-living *rock* is lower on the scale than the plant *flower*, which is in turn lower than the animal *cat*.
- MASS-COUNT. The "countability" or "separatability" of an entity. For example, *milk* is less countable/separable than *sand*, which is in turn less countable/separable than *pebbles*.
- PHASE. The common physical state of an entity, i.e. gas, liquid and solid.
- SIZE. The size of an entity with respect to the absolute scale marked with landmark entities[1].
- WEIGHT. Similar to size, only differs in landmarks.
- RIGIDITY. The material hardness of an entity. For example, a *cloth* is softer than a piece of *wood*, which is in turn softer than a *metal* bar.

While the feature types are grounded in scientific terms, primarily they are intended to be about subject perception about the physical attributes of *prototypes* of entities. For example, a *couch* is not necessarily larger in size than a *television*, but prototypically speaking it is in general.

While the landmark entities for the feature types read reasonable and by intuition one may expect them to work well in plausibility classification, they have been constructed by introspection together with trial-and-error. It is reasonable to believe there must be an optimal way in which the landmarks as boundaries are drawn, however we refrain from inching deeper into the subject in this work, as it pertains more to cognitive psychology, although we would project interesting work in this direction.

2 Models and Configurations

NN [Van de Cruys, 2014] is a simple three-layer feedforward network. Let $\boldsymbol{v}, \boldsymbol{s}, \boldsymbol{o}$ be the looked-up GloVe embeddings [Pennington et al., 2014] of a triple (i.e the verb and its subject and object), we first concatenate the vectors

$$\boldsymbol{x} = [\boldsymbol{v}; \boldsymbol{s}; \boldsymbol{o}] \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^{900}$ with 300D word embeddings. We then make two affine transformations

$$\boldsymbol{a}_1 = \sigma(W_1\boldsymbol{x} + \boldsymbol{b}_1) \tag{2}$$

$$\boldsymbol{a}_2 = W_2\boldsymbol{a}_1 + \boldsymbol{b}_2 \tag{3}$$

where σ is a non-linearity for which we use ReLU activation, $W_1 \in \mathbb{R}^{d\times 900}, W_2 \in \mathbb{R}^{2\times d}$ (and the bias vectors according dimensions. The original work proposes a max-margin loss function, but on our task the Adam [Kingma and Ba, 2014] optimizer works better empirically. Further, while [Van de Cruys, 2014] trains the model with negative sampling (i.e. randomly generated triples for negative instances for selectional

[1] Section 4 has more details on landmark entities

preference), our training is straight-out supervision with "true" negative cases. Hyperparameter-wise, the hidden dimension $d = 10$; We train the model by 20 epochs with a batch size of 20, initial learning rate 0.0001, and a learning rate decay factor of 0.95 per 100 global steps (i.e. batches).

NN + WK has the identical settings for the **NN** component, i.e. it produces the vector $\boldsymbol{a}_1$ as in Eq. 2. For the **WK** component (**W**orld **K**nowledge features), we have three configuratory choices:

(i) *Feature encoding scheme*: 3-LEVEL or BIN-DIFF;

(ii) *Featurization method*: raw feature vectors or feature embeddings;

(iii) *Transformation*: affine or bilinearity.

We elaborate on each now. The feature encoding is of a subject-object pair, i.e. a function $f_{\text{FEATURE-TYPE}}(s, o)$ that takes the subject s and the direct object o in a triple *s-v-o* as the input and returns a feature value. For example, for the triple *man-hug-ant*, $f_{\text{SIZE}}(man, ant)$ is its corresponding feature value for the feature type SIZE. To explain 3-LEVEL vs. BIN-DIFF schemes, again take SIZE as the running example. Suppose *man* and *ant* are in the 4^{th} and 1^{st} SIZE bin respectively,

$$f_{\text{SIZE}}^{\text{3-LEVEL}}(man, ant) = 1$$
$$f_{\text{SIZE}}^{\text{BIN-DIFF}}(man, ant) = 3$$

where in 3-LEVEL, $f_{\text{SIZE}}^{\text{3-LEVEL}}$ says *man* is larger in size than *ant*, and in BIN-DIFF scheme $f_{\text{SIZE}}^{\text{BIN-DIFF}}$ says *man* is 3 "degrees" larger than *ant*. Essentially BIN-DIFF subsumes 3-LEVEL and is more fine-grained.

Next consider featurization methods. By raw feature vectors we refer to the feature values $f_t^{\text{3-LEVEL}}$ or $f_t^{\text{BIN-DIFF}}$ for $t \in \{\text{SENTIENCE,MASS-COUNT,PHASE,SIZE,WEIGHT,RIGIDITY}\}$. Take the pair *man-ant* again, its raw feature vector in 3-LEVEL scheme is $\langle 1, 0, 0, 1, 1, 0\rangle$, i.e. *man* is similar to *ant* in MASS-COUNT, PHASE and RIGIDITY, and more SENTIENT, larger in SIZE and WEIGHT. Similar for the BIN-DIFF scheme, only with more granular feature values. For feature embeddings, we map each feature value (wrt. a feature type) to a d-dimensional dense embedding. The embeddings are looked up in a randomly initialized embedding matrix.

Finally transformation. Suppose we opted for feature embedding for featurization, and let $\boldsymbol{x}$ be the concatenation of the embeddings for the six feature types for an *s-o* pair, the affine transformation is simply

$$\boldsymbol{a} = \sigma(W\boldsymbol{x} + \boldsymbol{b}) \tag{4}$$

For bilinearity, we first make transformation for each embedding (corresponding to each feature type)

$$\boldsymbol{a}_t = \boldsymbol{x}_t^\intercal W \boldsymbol{x}_t + b_t, \quad t \in \{\text{SENTIENCE, MASS-COUNT}, \cdots\} \tag{5}$$

and then concatenate the results to feed into a nonlinearity:

$$\boldsymbol{a} = \sigma([\boldsymbol{a}_t; \cdots]), \quad t \in \{\text{SENTIENCE, MASS-COUNT}, \cdots\} \tag{6}$$

Finally the **NN** and the **WK** components merge as follows: we concatenate the final feature vectors $\boldsymbol{a}_{\mathbf{NN}}$ and $\boldsymbol{a}_{\mathbf{WK}}$ and feed it through a softmax layer to produce the prediction $\hat{y}$:

$$\hat{y} = \underset{y}{\text{argmax}}\ \text{softmax}([\boldsymbol{a}_{\mathbf{NN}}; \boldsymbol{a}_{\mathbf{WK}}]) \tag{7}$$

3 Highly-Specific Attributes: Full List

- *edibility* (21%). **-fry-egg* (plausible) and **-fry-cup* (implausible) are hard to distinguish because *egg* and *cup* are similar in SIZE/WEIGHT/..., however introducing large free-text data to help learn edibility misguides our model to mind selectional preference, causing mislabeling of other events.
- *natural vs. artificial* (18%). Turkers often think creating natural objects like *moon* or *mountain* is implausible but creating an equally big (but artificial) object like *skyscraper* is plausible.

- *hollow objects* (15%). *plane-contain-shell* and *purse-contain-scissors* are plausible, but the hollow-object-can-contain-things attribute is failed to be captured.
- *forefoot dexterity* (5%). *horse-hug-man* is implausible but *bear-hug-man* is plausible; For **-snatch-watch*, *girl* is a plausible subject, but not *pig*. Obviously the dexterity of the forefoot of the agent matters here.
- *liquid absorbant* (13%). *sponge* can **-absorb-rain*, so does *cloth* or *soil*, but the objects share little distributional similarity.
- *sharpness* (8%). *knife* and *table* have sharp edges and a *butcher* wields sharp objects, but *knife-trim-tree* often gets misclassified.
- *repeated actions* (10%). *eat* and *swallow* are similar but the latter is a one-go action, thus *worm-eat-infant* is plausible but not *worm-swallow-infant*.
- *multiple readings* (10%). The majority of Turkers judge *smoke-poison-banana* to be implausible and *water-poison-soil* plausible.

References

[Kingma and Ba, 2014] Kingma, D. and Ba, J. (2014). Adam: a method for stochastic optimization. In *Proceedings of ICLR*.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: global vectors for word embeddings. In *Proceedings of EMNLP*.

[Van de Cruys, 2014] Van de Cruys, T. (2014). A neural network approach to selectional preference acquisition. In *Proceedings of EMNLP*.