Modeling Event Plausibility with Consistent Conceptual Abstraction

Ian Porada[1], **Kaheer Suleman**[2], **Adam Trischler**[2], **and Jackie Chi Kit Cheung**[1]

[1]Mila, McGill University
{ian.porada@mail, jcheung@cs}.mcgill.ca
[2]Microsoft Research Montréal
{kasulema, adam.trischler}@microsoft.com

Abstract

Understanding natural language requires common sense, one aspect of which is the ability to discern the plausibility of events. While distributional models—most recently pre-trained, Transformer language models—have demonstrated improvements in modeling event plausibility, their performance still falls short of humans'. In this work, we show that Transformer-based plausibility models are markedly inconsistent across the conceptual classes of a lexical hierarchy, inferring that "a person breathing" is plausible while "a dentist breathing" is not, for example. We find this inconsistency persists even when models are softly injected with lexical knowledge, and we present a simple post-hoc method of forcing model consistency that improves correlation with human plausibility judgements.

1 Introduction

Of the following events, a human reader can easily discern that (1) and (2) are semantically plausible, while (3) is nonsensical.

(1) The person breathes the air.

(2) The dentist breathes the helium.

(3) The thought breathes the car.

This ability is required for understanding natural language: specifically, modeling *selectional preference*—the semantic plausibility of predicate-argument structures—is known to be implicit in discriminative tasks such as coreference resolution (Hobbs, 1978; Dagan and Itai, 1990; Zhang et al., 2019b), word sense disambiguation (Resnik, 1997; McCarthy and Carroll, 2003), textual entailment (Zanzotto et al., 2006; Pantel et al., 2007), and semantic role labeling (Gildea and Jurafsky, 2002; Zapirain et al., 2013).

More broadly, modeling semantic plausibility is a necessary component of generative inferences such as conditional commonsense inference (Gordon et al., 2011; Zhang et al., 2017), abductive commonsense reasoning (Bhagavatula et al., 2020), and commonsense knowledge acquisition (Zhang et al., 2020a; Hwang et al., 2020).

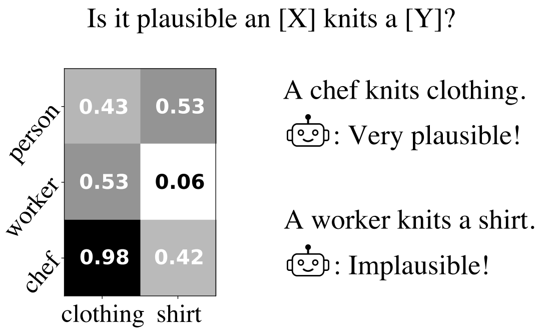


Figure 1: Elements in the matrix are the relative plausibility score for the event "an [X] knits a [Y]" as output by a RoBERTa model fine-tuned to model plausibility. [X] and [Y] correspond to the label of the row and column, respectively. Model scores are inconsistent with respect to the two events shown on the right.

Learning to model semantic plausibility is a difficult problem for several reasons. First, language is sparse, so most events will not be attested even in a large corpus. Second, plausibility relates to likelihood in the world, which is distinct from the likelihood of an event occurring in language. Third, plausibility reflects human intuition, and thus modeling plausibility at its extreme requires "the entire representational arsenal that people use in understanding language, ranging from social mores to naive physics" (Resnik, 1996).

A key property of plausibility is that the plausibility of an event is generally consistent across some appropriate level of abstraction. For example, events of the conceptual form "the [PERSON] breathes the [GAS]" are consistently plausible. Plausibility judgments follow this pattern because people understand that similar concept classes share similar affordances.

Proceedings of the 2021 Conference of the North American Chapter of the
Association for Computational Linguistics: Human Language Technologies, pages 1732–1743
June 6–11, 2021. ©2021 Association for Computational Linguistics

Furthermore, the change in plausibility between levels of abstraction is often consistent. Consider that as we abstract from "person breathes" to "organism breathes" to "entity breathes," plausibility consistently decreases.

In this paper, we investigate whether state-of-the-art plausibility models based on fine-tuning Transformer language models likewise exhibit these types of consistency. As we will show, inconsistency is a significant issue in existing models which results in erroneous predictions (See Figure 1 for an example).

To address this issue, we explore two methods that endow Transformer-based plausibility models with knowledge of a lexical hierarchy—our hypothesis being that these methods might correct conceptual inconsistency without over-generalizing. The first method makes no a priori assumptions as to how the model should generalize and simply provides lexical knowledge as an additional input to the model. The second explicitly enforces conceptual consistency across a lexical hierarchy by taking the plausibility of an event to be a maximum over the plausibility of all conceptual abstractions of the event.

We find that only the second proposed method sufficiently biases the model to more accurately correlate with human plausibility judgments. This finding encourages future work that forces Transformer models to make more discrete abstractions in order to better model plausibility.

We focus our analysis on simple events in English represented as subject-verb-object (s-v-o) triples, and we evaluate models by correlation with two datasets of human plausibility judgements. Our models build off of RoBERTa (Liu et al., 2019), a pre-trained Transformer masked language model.[1] We use WordNet 3.1 (Miller, 1995) hypernymy relations as a lexical hierarchy.

Concretely, our contributions are:

- We evaluate the state of the art in modeling plausibility, both in terms of correlation with human judgements and consistency across a lexical hierarchy.
- We propose two measures of the consistency of plausibility estimates across conceptual abstractions.
- We show that injecting lexical knowledge into a plausibility model does not overcome conceptual inconsistency.
- We present a post-hoc method of generalizing plausibility estimates over a lexical hierarchy that is necessarily consistent and improves correlation with human plausibility judgements.

2 Related Work

While plausibility is difficult to define precisely, we adopt the following useful distinctions from the literature:

- Plausibility is a matter of degree (Wilks, 1975; Resnik, 1993). We therefore evaluate models by their ability to estimate the relative plausibility of events.
- Plausibility describes non-surprisal conditioned on some context (Resnik, 1993; Gordon et al., 2011). For example, conditioned on the event "breathing," it is less surprising to learn that the agent is "a dentist" than "a thought" and thus more plausible.
- Plausibility is dictated by likelihood of occurrence in the world rather than text (Zhang et al., 2017; Wang et al., 2018). This discrepancy is due to reporting bias—the fact that people do not state the obvious (Gordon and Van Durme, 2013; Shwartz and Choi, 2020); e.g., "a person dying" is more likely to be attested than "a person breathing" (Figure 2).

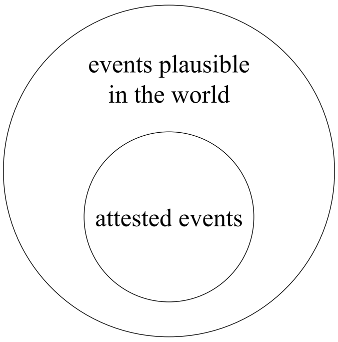


Figure 2: An attested event is necessarily plausible in the world, but not all plausible events are attested. By the world we refer to some possible world under consideration—in this sense plausibility is an epistemic modality.

Wang et al. (2018) present the problem formulation that we use in this work, and they show that

[1] Our implementation and data is available at https://github.com/ianporada/modeling_event_plausibility

static word embeddings lack the world knowledge needed for modeling plausibility.

The state of the art is to take the conditional probability of co-occurrence as estimated by a distributional model as an approximation of event plausibility (Zhang et al., 2020a). Our fine-tuned RoBERTa baseline follows this approach.

Similar in spirit to our work, He et al. (2020) extend this baseline method by creating additional training data using the Probase taxonomy (Wu et al., 2012) in order to improve conceptual generalization; specifically, for each training example they swap the event's arguments with its hypernym or hyponym, and they take this new, perturbed example to be an implausible event.

There is also recent work focusing on monotonic inferences in semantic entailment (Yanaka et al., 2019; Goodwin et al., 2020; Geiger et al., 2020). Plausibility contrasts with entailment in that plausibility is not strictly monotonic with respect to hypernymy/hyponymy relations: the plausibility of an entity is not sufficient to infer the plausibility of its hyponyms (i.e., not downward entailing: it is plausible that a person gives birth but not that a man gives birth) nor hypernyms (i.e., not upward entailing: it is plausible that a baby fits inside a shoebox but not that a person does).

Non-monotonic inferences have recently been explored in the context of defeasible reasoning (Rudinger et al., 2020): inferences that may be strengthened or weakened given additional evidence. The change in plausibility between an event and its abstraction can be formulated as a type of defeasible inference, and our findings may contribute to future work in this area.

2.1 Selectional Preference

Modeling the plausibility of single events is also studied in the context of *selectional preference*—the semantic preference of a predicate for taking an argument as a particular dependency relation (Evens, 1975; Resnik, 1993; Erk et al., 2010); e.g., the relative preference of the verb "breathe" for the noun "dentist" as its nominal subject.

Models of selectional preference are sometimes evaluated by correlation with human judgements (Ó Séaghdha, 2010; Zhang et al., 2019a). The primary distinction between such evaluations and those of semantic plausibility, as in our work, is that evaluations of semantic plausibility emphasize the importance of correctly modeling atypical yet plausible events (Wang et al., 2018).

Closely related to our work are models of selectional preference that use the WordNet hierarchy to generalize co-occurrence probabilities over concepts. These include the work of Resnik (1993), related WordNet-based models (Li and Abe, 1998; Clark and Weir, 2002), and a more recent experiment by Ó Séaghdha and Korhonen (2012) to combine distributional models with WordNet. Notably, these methods make a discrete decision as to the right level of abstraction—if the most preferred subject of "breathe" is found to be "person," for example, then all hyponyms of "person" will be assigned the same selectional preference score.

2.2 Conceptual Abstraction

Our second proposed method can be thought of as finding the right level of abstraction at which to infer plausibility. This problem has been broadly explored by existing work.

Van Durme et al. (2009) extract abstracted commonsense knowledge from text using WordNet, obtaining inferences such as "A [PERSON] can breathe." They achieve this by first extracting factoids and then greedily taking the WordNet synset that dominates the occurrences of factoids to be the appropriate abstraction.

Gong et al. (2016) similarly abstract a verb's arguments into a set of prototypical concepts using Probase and a branch-and-bound algorithm. For a given verb and argument position, their algorithm finds a small set of concepts that has high coverage of all nouns occurring in said position.

Conceptual abstractions are captured to some extent in pre-trained language models' representations (Ravichander et al., 2020; Weir et al., 2020).

3 Problem Formulation

Given a vocabulary of subjects $\mathcal{S}$, verbs $\mathcal{V}$, and objects $\mathcal{O}$, let an event be represented by the s-v-o triple $e \in \mathcal{S} \times \mathcal{V} \times \mathcal{O}$.

We take g to be a ground-truth, total ordering of events expressed by the ordering function $g(e) > g(e')$ iff e is more plausible than e'. Our objective is to learn a model $f : \mathcal{S} \times \mathcal{V} \times \mathcal{O} \rightarrow \mathbf{R}$ that is monotonic with respect to g, i.e., $g(e) > g(e') \implies f(e) > f(e')$.

This simplification follows from previous work (Wang et al., 2018), and the plausibility score for a given triple can be considered the relative plausibility of the respective event across all contexts and realizations.

While meaning is sensitive to small linguistic perturbations, we are interested in cases where one event is more plausible than another marginalized over context. Consider that *person-breathe-air* is more plausible than *thought-breathe-car* regardless of the choice of determiners or tense of the verb.

In practice, we would like to learn f without supervised training data, as collecting a sufficiently large dataset of human judgements is prohibitively expensive (Zhang et al., 2020b), and supervised models often learn dataset-specific correlations (Levy et al., 2015; Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019). Therefore, we train model f with distant supervision and evaluate by correlation with human ratings of plausibility which represent the ground-truth ordering g.

3.1 Lexical Hierarchy

We define $\mathcal{C}$ to be the set of concepts in a lexical hierarchy, in our case *synsets* in WordNet, with some root concept $c^{(1)} \in \mathcal{C}$. The *hypernym chain* of concept $c^{(h)} \in \mathcal{C}$ at depth h in the lexical hierarchy is defined to be the sequence of concepts $\alpha(c^{(h)}) = (c^{(1)}, c^{(2)}, \dots, c^{(h)})$ where $\forall i, c^{(i)}$ is a direct hypernym of $c^{(i+1)}$. A lexical hierarchy may be an acyclic graph in which case concepts can have multiple hypernyms, and it follows that there may be multiple hypernym chains to the root. In this case, we take the hypernym chain $\alpha(c^{(h)})$ to be the shortest such chain.

3.2 Consistency Metrics

Based on our intuition as to how we expect plausibility estimates to be consistent across abstractions in a hypernym chain, we propose two quantitative metrics of *inconsistency*, Concavity Delta (CCΔ) and Local Extremum Rate (LER). These metrics provide insight into the degree to which a model's estimates are inconsistent.

3.2.1 Concavity Delta

For a given event, as we traverse up the hypernym chain to higher conceptual abstractions, we expect plausibility to increase until we reach some maximally appropriate level of abstraction, and then decrease thereafter. In other words, we expect that consistent estimates will be concave across a sequence of abstractions.

For example, in the sequence of abstractions "penguin flies" → "bird flies" → "animal flies," plausibility first increases and then decreases. Our intuition is that plausibility increases as we approach the most appropriate level of abstraction, then decreases beyond this level.

A concave sequence is defined to be a sequence $(a_1, a_2, a_3, \dots)$ where $\forall i, 2a_i > a_{i-1} + a_{i+1}$.

Let a_{i-1}, a_i, and a_{i+1} be the plausibility estimates for three sequential abstractions of an event. We define the *divergence from concavity* to be

$$\delta = \begin{cases} \frac{1}{2}(a_{i-1} + a_{i+1}) - a_i & 2a_i < a_{i-1} + a_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

We then define the *Concavity Delta*, CCΔ, to be the average δ across all triples of conceptually sequential estimates. Ideally, a model's estimates should have low CCΔ. A higher CCΔ reflects the extent to which models violate our intuition.

3.2.2 Local Extremum Rate

LER simply describes how often a conceptual abstraction is a local extremum in terms of its plausibility estimate. Most often, the change in plausibility between sequential abstractions is consistently in the same direction. For example, from "bird flies" → "animal flies" → "organism flies," plausibility consistently decreases. The majority of abstractions will not be the most appropriate level of abstraction and therefore not a local extremum.

As in §3.2.1, we consider all triples of conceptually sequential estimates of the form a_{i-1}, a_i, and a_{i+1}. Formally, LER is the number of triples where $a_i > max(a_{i-1}, a_{i+1})$ or $a_i < min(a_{i-1}, a_{i+1})$ divided by the total number of triples.

A high LER signifies that plausibility estimates have few monotonic subsequences across abstractions. Therefore, a more consistent model should have a lower LER. There are, of course, exceptions to our intuition, and this metric is most insightful when it varies greatly between models.

4 Models

The models that we consider are all of the same general form. They take as input an event and output a relative plausibility score.

4.1 RoBERTa

Our proposed models are structured on top of a RoBERTa baseline. We use RoBERTa in the standard sequence classification framework. We format an event in the raw form as '`[CLS] subject verb object [SEP]`' where the s-v-o triple is tokenized

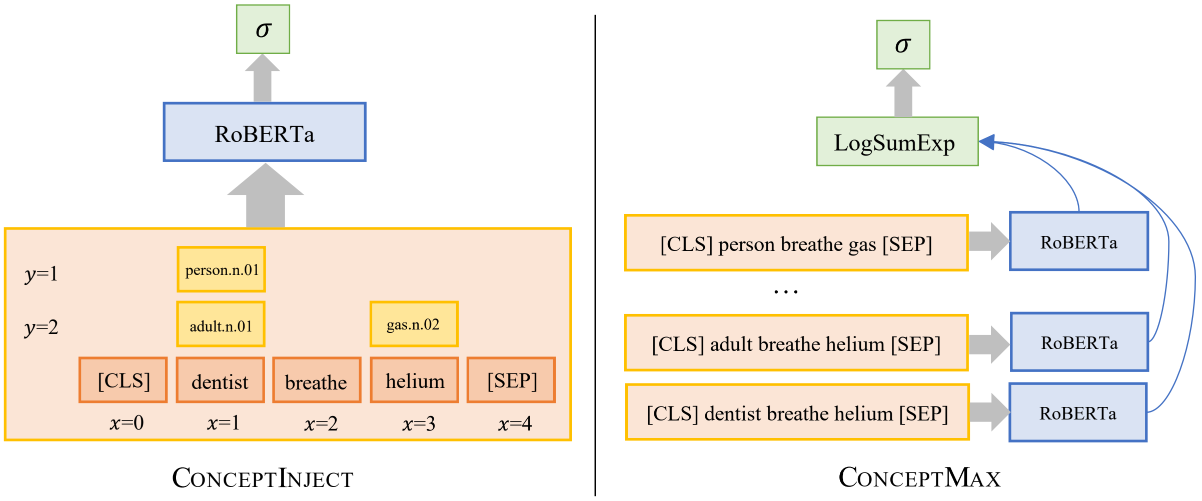


Figure 3: Left: The general formulation of ConceptInject; this model takes as input an event and the full hypernym chains of each argument. Right: ConceptMax which calculates a plausibility score for each abstraction of an event using RoBERTa, and then takes the ultimate output to be the maximum of these abstractions. σ represents an element-wise sigmoid function.

using a byte pair encoding.[2] These tokens are used as input to a pre-trained RoBERTa model, and a linear layer is learned during fine-tuning to project the final-layer `[CLS]` token representation to a single logit which is passed through a sigmoid to obtain the final output, $f(e)$.

We use the HuggingFace Transformers library PyTorch implementation of RoBERTa-base with 16-bit floating point precision (Wolf et al., 2020).

4.2 ConceptInject

ConceptInject is an extension of the existing state-of-the-art plausibility models. This model takes as input, in addition to an event, the hypernym chains of the synsets corresponding to each argument in the event. We propose this model to explore how injecting simple awareness of a lexical hierarchy affects estimates.

ConceptInject is similar in principle to Onto-LSTM (Dasigi et al., 2017), which provides the entire hypernym chains of nouns as input to an LSTM for selectional preference, and also similar to K-BERT (Liu et al., 2020), which injects knowledge into BERT during fine-tuning by including relations as additional tokens in the input. K-BERT has demonstrated improved performance over Chinese BERT on several NLP tasks.

The model extends our vanilla RoBERTa baseline (§4.1). We add an additional token embedding to RoBERTa for each synset $c \in \mathcal{C}$. We initialize the embedding of c as the average embedding of the sub-tokens of c's lemma.[3] We refer to RoBERTa's positional embedding matrix as the x-position and randomly initialize a second positional embedding matrix, the y-position.

The model input format follows that used for RoBERTa (§4.1), with the critical distinction that we also include the tokens for the hypernyms of the subject and object as additional input.

For the subject s, we first disambiguate the synset c of s using BERT-WSD (Yap et al., 2020). Then for each hypernym $c^{(i)}$ in the hypernym chain $\alpha(c)$, the token of $c^{(i)}$ is included in the model input: this token takes the same x-position as the first sub-token of s and takes its y-position to be i, the depth in the lexical hierarchy. Finally, the x-position, y-position, and token embedding are summed for each token to compute its initial representation (Figure 3).

The hypernyms of the object are included by the same procedure. Non-synset tokens have a y-position of zero. ConceptInject thus sees an event and the full hypernym chains of the arguments when computing a plausibility score.

[2] Technically, RoBERTa's `[CLS]` and `[SEP]` tokens are `<s>` and `</s>`.

[3] We refer to the name of a synset as the synset's lemma, e.g. the lemma of the synset [dog.n.01] is taken to be "dog." For synsets that correspond to multiple lemmas, we randomly sample one.

4.3 ConceptMax

ConceptMax is a simple post-hoc addition to the vanilla RoBERTa model (§4.1). We compute a score for all abstractions of an event e and take the final plausibility $f(e)$ to be a soft maximum of these scores. This method is inspired by that of Resnik (1993) which takes selectional preference to be a hard maximum of some plausibility measure over concepts.

Again, we use BERT-WSD to disambiguate the synset of the subject, $c_s^{(h)}$, and the synset of the object, $c_o^{(l)}$. Using RoBERTa as in §4.1, we then compute a plausibility score for every triple of the form $(c_s^{(i)}, v, c_o^{(j)})$ where $c_s^{(i)}$ and $c_o^{(j)}$ are hypernyms in the hypernym chains $\alpha(c_s^{(h)})$ and $\alpha(c_o^{(l)})$, respectively. Synsets are represented by their lemma when used as input to RoBERTa. Finally, we take the LogSumExp, a soft maximum, of these scores to be the ultimate output of the model (Figure 3).

During training, we sample only three of the abstractions $(c_s^{(i)}, v, c_o^{(j)})$ to reduce time complexity. Thus we only need to compute four total scores instead of $h \times l$. At inference time, we calculate plausibility with a hard maximum over all triples.

4.4 Additional Baselines

RoBERTa$_{\textbf{Zero-shot}}$ We use MLConjug[4] to realize an s-v-o triple in natural language with the determiner "the" for both the subject and object, and the verb conjugated in the indicative, third person tense; e.g., *person-breathe-air* ⟶ "The person breathes the air." We first mask both the subject and object to compute $P(o|v)$, then mask just the subject to compute $P(s|v, o)$. Finally we calculate $f(e) = P(s, o|v) = P(s|v, o) \cdot P(o|v)$. In the case that a noun corresponds to multiple tokens, we mask all tokens and take the probability of the noun to be the geometric mean of its token probabilities.

GloVe+MLP The selectional preference model of Van de Cruys (2014) initialized with GloVe embeddings (Pennington et al., 2014).

n-gram A simple baseline that estimates $P(s, o|v)$ by occurrence counts. We use a bigram model as we found trigrams to correlate less with human judgments.

$$P(s, o|v) \approx \frac{\texttt{Count}(s, v) \cdot \texttt{Count}(v, o)}{\texttt{Count}(v)^2} \quad (1)$$

e	e'
animal-eat-seed	*animal-eat-area*
passenger-ride-bus	*bus-ride-bus*
fan-throw-fruit	*group-throw-number*
woman-seek-shelter	*line-seek-issue*

Table 1: Training examples extracted from Wikipedia. Event e is an attested event taken to be more plausible than its random perturbation e'.

5 Training

Models are all trained with the same objective to discriminate plausible events from less plausible ones. Given a training set $\mathcal{D}$ of event pairs (e, e') where e is more plausible than e', we minimize the binary cross-entropy loss

$$L = - \sum_{(e,e') \in \mathcal{D}} \log(f(e)) + \log(1 - f(e')) \quad (2)$$

In practice, $\mathcal{D}$ is created without supervised labels. For each $(e, e') \in \mathcal{D}$, e is an event attested in a corpus with subject s, verb v, and object o. e' is a random perturbation of e uniformly of the form (s', v, o), (s, v, o'), or (s', v, o') where s' and o' are arguments randomly sampled from the training corpus by occurrence frequency. This is a standard pseudo-disambiguation objective. Our training procedure follows recent works that learn plausibility models with self-supervised fine-tuning (Kocijan et al., 2019; He et al., 2020; Zhang et al., 2020a).

For the models that use WordNet, we use a filtered set of synsets: we remove synsets with a depth less than 4, as these are too broad to provide useful generalizations (Van Durme et al., 2009). We also filter out synsets whose corresponding lemma did not appear in the training corpus.

The WordNet models also require sense disambiguation. We use the raw triple as input to BERT-WSD (Yap et al., 2020) which outputs a probability distribution over senses. We take the argmax to be the correct sense.

We train all models with gradient descent using an Adam optimizer, a learning rate of 2e-5, and a batch size of 128. We train for two epochs over the entire training set of examples with a linear warm-up of the learning rate over the first 10,000 iterations. Fine-tuning RoBERTa takes five hours on a single Nvidia V100 32GB GPU. Fine-tuning ConceptInject takes 12 hours and ConceptMax 24 hours.

[4] https://pypi.org/project/mlconjug/

5.1 Training Data

We use English Wikipedia to construct the self-supervised training data. As a relatively clean, definitional corpus, plausibility models trained on Wikipedia have been shown to correlate with human judgements better than those trained on similarly sized corpora (Zhang et al., 2019a; Porada et al., 2019).

We parse a dump of English Wikipedia using the Stanford neural dependency parser (Qi et al., 2018). For each sentence with a direct object, no indirect object, and noun arguments (that are not proper nouns), we extract a training example (s, v, o): we take s and o to be the lemma of the head of the respective relations (`nsubj` and `obj`), and v to be the lemma of the head of the root verb. This results in some false positives such as the sentence "The woman eats a hot dog." being extracted to the triple *woman-eat-dog* (Table 1).

We filter out triples that occur less than once and those where a word occurred less than 1,000 times in its respective position. We do not extract the same triple more than 1,000 times so as not to over-sample common events. In total, we extract 3,298,396 triples (representing 538,877 unique events).

6 Predicting Human Plausibility Judgements

We evaluate models by their correlation with human plausibility judgements. Each dataset consists of events that have been manually labelled to be plausible or implausible (Table 3). We use AUC (area under the receiver-operating-characteristic curve) as an evaluation metric which intuitively reflects the ability of a model to discriminate a plausible event from an implausible one.

These datasets contain plausible events that are both typical and atypical. While a distributional model should be able to discriminate typical events given that they frequently occur in text, discriminating atypical events (such as *dentist-breathe-helium*) is more difficult.

6.1 PEP-3K

PEP-3K, the crowdsourced **P**hysical **E**vent **P**lausbility ratings of Wang et al. (2018), consists of 3,062 events rated as physically plausible or implausible by five crowdsourced workers. Annotators were instructed to ignore possible metaphorical meanings of an event. We divide the dataset equally into a validation and test set following the split of Porada et al. (2019).

Topic	Question	Answer
cat	Does it lay eggs?	never
carrot	Can you eat it?	always
cocoon	Can it change shape?	sometimes
clock	Can I touch it?	always

Table 2: Example triples from the 20 Questions commonsense dataset. These are those specific examples that contain a simple question with a single s-v-o triple and no modifiers.

PEP-3K	*chef-bake-cookie*	✓
	dog-close-door	✓
	fish-throw-elephant	✗
	marker-fuse-house	✗
20Q	*whale-breathe-air*	✓
	wolf-wear-collar	✓
	cat-hatch-egg	✗
	armrest-breathe-air	✗

Table 3: Representative examples taken from the validation splits of the two plausibility evaluation datasets, PEP-3K and 20Q. For simplicity, we present human judgments as plausible (✓) or implausible (✗). Details are provided in §6.

To evaluate on this dataset, we make the assumption that all events labeled physically plausible are necessarily more plausible than all those labeled physically implausible.

6.2 20Q

The 20 Questions commonsense dataset[5] is a collection of 20 Questions style games played by crowdsourced workers. We format this dataset as plausibility judgments of s-v-o triples similar to PEP-3K.

In the game 20 Questions, there are two players—one who knows a given topic, and the other who is trying to guess this topic by asking questions that have a discrete answer. The dataset thus consists of triples of topics, questions, and answers where the answer is one of: always, usually, sometimes, rarely, or never (Table 2).

We parse the dataset using the Stanford neural dependency parser (Qi et al., 2018). We then extract questions that contain a simple s-v-o triple

[5] `https://github.com/allenai/twentyquestions`

Model	Pep-3k	20Q	Avg.
n-gram	.51	.52	.52
GloVe+MLP	.55	.52	.53
RoBERTa$_{\text{Zero-shot}}$	.56	.57	.56
RoBERTa	.64	.67	.66
ConceptInject	.64	.66	.65
ConceptMax	**.67**	**.74**	**.70**

Table 4: Test set results for predicting human plausibility judgements. Performance is evaluated with AUC with respect to the ground-truth, manually labeled plausibility ratings.

with no modifiers where either the subject or object is a third person singular pronoun. We replace this pronoun with the topic, and otherwise replace any occurrence of a personal pronoun with the word "person." We filter out examples where only two of three annotators labelled the likelihood as never. Finally, we take events labelled "never" to be less plausible than all other events. This process results in 5,096 examples equally divided between plausible and implausible. We split examples into equal sized validation and test sets.

6.3 Quantitative Results

Despite making a discrete decision about the right level of abstraction, ConceptMax has higher AUC on both evaluation sets as compared to ConceptInject and the vanilla RoBERTa baseline (Table 4). The fact that the ConceptMax model aligns with human judgments more than the baselines supports the hypothesis that conceptual consistency improves plausibility estimates.

ConceptInject performs similarly to the RoBERTa baseline even though this model is aware of the WordNet hierarchy. We hypothesize that the self-supervised learning signal does not incentivize use of this hierarchical information in a way that would increase correlation with plausibility judgements. We do find that ConceptInject attends to the hypernym chain, however, by qualitatively observing the self-attention weights.

All fine-tuned RoBERTa models correlate better with plausibility judgements than the RoBERTa$_{\text{Zero-shot}}$ baseline, and the n-gram baseline performs close to random—this is perhaps to be expected, as very few of the evaluation triples occur in our Wikipedia training data.

	Pep-3k		20Q	
Model	CCΔ	LER	CCΔ	LER
n-gram	.06	.50	.07	.50
GloVe+MLP	.03	.61	.03	.49
RoBERTa$_{\text{Zero-shot}}$	.13	.70	.12	.65
RoBERTa	.09	.52	.08	.51
ConceptInject	.08	.52	.07	.51
ConceptMax	.02	.00	.02	.00

Table 5: Evaluation of inconsistency. CCΔ describes the degree to which sequences of estimates across a hypernym chain diverge from a concave sequence. LER describes how often conceptual abstractions are local extrema with respect to plausibility.

6.4 Qualitative Analysis

To better understand the performance of these models, we manually inspect 100 examples from each dataset. We find that RoBERTa rarely assigns a high score to a nonsensical event (although this does occur in five cases, such as *turtle-climb-wind* and *person-throw-library*). RoBERTa also rarely assigns a low score to a seemingly typical event, although this is somewhat more common (in cases such as *kid-use-handbag* and *basket-hold-clothes*, for example). This finding confirms our expectation that discerning the typical and nonsensical should be relatively easy for a distributional model.

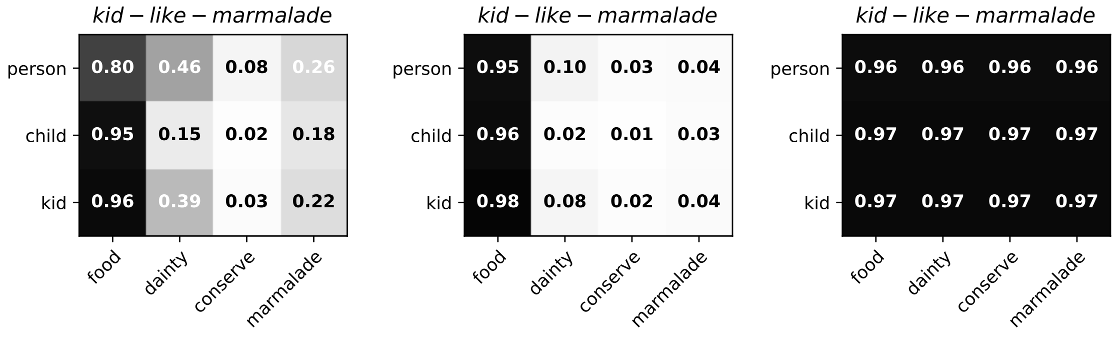
Examples not at the extremes of plausibility are harder to categorize; however, one common failure seems to be when the plausibility of an event hinges on the relative size of the subject and object, such as in the case of *dog-throw-whale*. This finding is similar to the limitations of static word embeddings observed by Wang et al. (2018).

7 Consistency Evaluation

For every event e in the evaluation sets of human plausibility judgments (§6), we disambiguate e using BERT-WSD and then calculate models' estimates for the plausibility of every possible abstraction of e (Figure 4). Based on these estimates, we can analyze the consistency of each model across abstractions.

7.1 Quantitative Results

We use our proposed metrics of consistency (§3.2) to evaluate the extent to which models' estimates are consistent across a hypernym chain (Table 5).



(a) RoBERTa (b) ConceptInject (c) ConceptMax

Figure 4: Outputs across conceptual abstractions for the event *kid-like-marmalade* from the 20Q dataset. This event is taken to be relatively plausible as the ground-truth label was "usually."

RoBERTa$_{\text{Zero-shot}}$, which correlates with plausibility the least of the RoBERTa models, has by far the highest inconsistency.

The fine-tuned RoBERTa and ConceptInject estimates are also largely inconsistent by our metrics. For these models, half of all estimates are a local extrema in the lexical hierarchy. As shown in Figure 4, the space of plausibility estimates is rigid for these models, and most estimates are a local extremum with respect to the plausibility of the subject or object of the event.

ConceptMax is almost entirely consistent by these metrics, which is to be expected as this model makes use of the same WordNet hierarchy that we are using for evaluation. We also evaluated consistency using the longest rather than the shortest hypernym chain in WordNet, but did not find a significant change in results. This is likely because for the consistency evaluation we are using the hypernym chains that have been filtered as described in §3.1.

7.2 Qualitative Results

We qualitatively evaluate the consistency of models by observing the matrix of plausibility estimates for all abstractions as show in Figure 4.

In agreement with our quantitative metrics, we observe that RoBERTa estimates are often inconsistent in that they vary greatly between two abstractions that have similar plausibility. Surprisingly, however, it is also often the case that RoBERTa estimates are similar or identical between abstractions. In some cases, this may be the result of the model being invariant to the subject or object of a given event.

We also observe the individual examples with the highest CCΔ. In these cases, it does appear that the variance of model estimates is unreasonable. In contrast, LER is sometimes high for an example where the estimates are reasonably consistent. This is a limitation of the LER metric not taking into account the degree of change between estimates.

Finally, we observe that the BERT-WSD sense is often different from what an annotator primed to rate plausibility would assume. For example, in the case of *dog-cook-turkey*, BERT-WSD takes dog to be a hyponym of person. While this is reasonable in context, it results in a different plausibility than that annotated.

8 Conclusion

While the state of the art in modeling plausibility has improved in recent years, models still fall short of human ability. We show that model estimates are inconsistent with respect to a lexical hierarchy: they correlate less with human judgments as compared to model estimates that are forced to be consistent, and they do not satisfy our intuitively defined quantitative measures of consistency.

In addition, we show that simply injecting lexical knowledge into a model is not sufficient to correct this limitation. Conceptual consistency appears to require a more discrete, hierarchical bias.

Interesting questions for future work are: 1) can we design a *non-monotonic*, consistent model of plausibility that better correlates with human judgements? 2) Can we induce a hierarchy of abstractions rather than using a manually created lexical hierarchy?

Acknowledgments

We would like to thank Ali Emami and Abhilasha Ravichander for useful discussions and comments, as well as the anonymous reviewers for their insightful feedback. This work is supported by funding from Microsoft Research and resources from Compute Canada. The last author is supported by the Canada CIFAR AI Chair program.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

Ido Dagan and Alon Itai. 1990. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*, COLING '90, page 330–332, USA. Association for Computational Linguistics.

Pradeep Dasigi, Waleed Ammar, Chris Dyer, and Eduard Hovy. 2017. Ontology-aware token embeddings for prepositional phrase attachment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2089–2098, Vancouver, Canada. Association for Computational Linguistics.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Martha Walton Evens. 1975. *Semantic Representations for Question-Answering Systems*. Ph.D. thesis, Northwestern University.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Yu Gong, Kaiqi Zhao, and Kenny Q. Zhu. 2016. Representing verbs as argument concepts. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2615–2621. AAAI Press.

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1958–1969, Online. Association for Computational Linguistics.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2011. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *SemEval@NAACL-HLT*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Mutian He, Y. Song, Kun Xu, and Y. Dong. 2020. On the role of conceptualization in commonsense knowledge graph construction. *ArXiv*, abs/2003.03239.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311 – 338.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the Winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842, Florence, Italy. Association for Computational Linguistics.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2901–2908.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Comput. Linguist.*, 29(4):639–654.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444, Uppsala, Sweden. Association for Computational Linguistics.

Diarmuid Ó Séaghdha and Anna Korhonen. 2012. Modelling selectional preferences in a lexical hierarchy. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 170–179, Montréal, Canada. Association for Computational Linguistics.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 564–571, Rochester, New York. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Ian Porada, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. Can a gorilla ride a camel? learning semantic plausibility from text. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 123–129, Hong Kong, China. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

P. Resnik. 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, 61(1-2):127–159.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Tagging Text with Lexical Semantics: Why, What, and How?*

Philip Stuart Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4661–4675, Online. Association for Computational Linguistics.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tim Van de Cruys. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 26–35, Doha, Qatar. Association for Computational Linguistics.

Benjamin Van Durme, Phillip Michalak, and Lenhart Schubert. 2009. Deriving generalized knowledge from corpora using WordNet abstraction. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 808–816, Athens, Greece. Association for Computational Linguistics.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. 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 303–308, New Orleans, Louisiana. Association for Computational Linguistics.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. In *Cognitive Science Society*.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6(1):53 – 74.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, Remi Louf, Patrick von Platen, Tim Rault, Yacine Jernite, Teven Le Scao, Sylvain Gugger, Julien Plu, Clara Ma, Canwei Shen, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, page 481–492, New York, NY, USA. Association for Computing Machinery.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting BERT for word sense disambiguation with gloss selection objective and example sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 41–46, Online. Association for Computational Linguistics.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Maria Teresa Pazienza. 2006. Discovering asymmetric entailment relations between verbs using selectional preferences. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 849–856, Sydney, Australia. Association for Computational Linguistics.

Beñat Zapirain, Eneko Agirre, Lluís Màrquez, and Mihai Surdeanu. 2013. Selectional preferences for semantic role classification. *Computational Linguistics*, 39(3):631–663.

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019a. SP-10K: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 722–731, Florence, Italy. Association for Computational Linguistics.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020a. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4004–4010. International Joint Conferences on Artificial Intelligence Organization. Main track.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020b. Aser: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, WWW '20, page 201–211, New York, NY, USA. Association for Computing Machinery.

Hongming Zhang, Yan Song, Yangqiu Song, and Dong Yu. 2019b. Knowledge-aware pronoun coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 867–876, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.