Modeling Semantic Plausibility by Injecting World Knowledge

Su Wang[1,2] **Greg Durrett**[3] **Katrin Erk**[1]
[1]Department of Linguistics
[2]Department of Statistics and Data Science
[3]Department of Computer Science
The University of Texas at Austin
shrekwang@utexas.edu gdurrett@cs.utexas.edu katrin.erk@mail.utexas.edu

Abstract

Distributional data tells us that a man can swallow candy, but not that a man can swallow a paintball, since this is never attested. However both are physically plausible events. This paper introduces the task of semantic plausibility: recognizing plausible but possibly novel events. We present a new crowdsourced dataset of semantic plausibility judgments of single events such as *man swallow paintball*. Simple models based on distributional representations perform poorly on this task, despite doing well on selection preference, but injecting manually elicited knowledge about entity properties provides a substantial performance boost. Our error analysis shows that our new dataset is a great testbed for semantic plausibility models: more sophisticated knowledge representation and propagation could address many of the remaining errors.

1 Introduction

Intuitively, a *man* can *swallow* a *candy* or *paintball* but not a *desk*. Equally so, one cannot plausibly *eat* a *cake* and then *hold* it. What kinds of semantic knowledge are necessary for distinguishing a physically plausible event (or event sequence) from an implausible one? *Semantic plausibility* stands in stark contrast to the familiar *selectional preference* (Erk and Padó, 2010; Van de Cruys, 2014) which is concerned with the *typicality* of events (Table 1). For example, *candy* is a typical entity for *man-swallow-** but *paintball* is not, even though both events are plausible physically. Also, some events are physically plausible but are never stated because humans avoid stating the obvious. Critically, semantic plausibility is sensitive to certain properties such as relative object size that are not explicitly encoded by selectional preferences (Bagherinezhad et al., 2016). Therefore, it is crucial that we learn to model these dimensions in addition to using classical distributional signals.

*man-swallow-**	PREFERRED?	PLAUSIBLE?
-candy	✓	✓
-paintball	✗	✓
-desk	✗	✗

Table 1: Distinguishing *semantic plausibility* from *selectional preference*. *candy* is selectionally preferred because it is distributionally common patient in the event *man-swallow-**, as opposed to the bizarre and rarely seen (if at all) patient *paintball*. However both are semantically plausible according to our world knowledge: they are small-sized objects that are swallowable by a man. *desk* is both distributionally unlikely and implausible (i.e. oversized for swallowing).

Semantic plausibility is pertinent and crucial in a multitude of interesting NLP tasks put forth previously, such as narrative schema (Chambers, 2013), narrative interpolation (Bowman et al., 2016), story understanding (Mostafazadeh et al., 2016), and paragraph reconstruction (Li and Jurafsky, 2017). Existing methods for these tasks, however, draw predominantly (if not only) on distributional data and produce rather weak performance. Semantic plausibility over *subject-verb-object* triples, while simpler than these other tasks, is a key building block that requires many of the same signals and encapsulates complex world knowledge in a binary prediction problem.

In this work, we show that world knowledge injection is necessary and effective for the semantic plausibility task, for which we create a robust, high-agreement dataset (details in section 3). Employing methods inspired by the recent work on world knowledge propagation through distributional context (Forbes and Choi, 2017; Wang et al., 2017), we accomplish the goal with minimal effort in manual annotation. Finally, we perform an in-depth error analysis to point to future directions of work on semantic plausibility.

Proceedings of NAACL-HLT 2018, pages 303–308
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

2 Related Work

Simple events (i.e. S-V-O) have seen thorough investigation from the angle of selectional preference. While early works are resource-based (Resnik, 1996; Clark and Weir, 2001), later work shows that unsupervised learning with distributional data yields strong performance (O'Seaghdha, 2010; Erk and Padó, 2010), which has recently been further improved upon with neural approaches (Van de Cruys, 2014; Tilk et al., 2016). Distribution-only models however, as will be shown, fail on the semantic plausibility task we propose.

Physical world knowledge modeling appears frequently in more closely related work. Bagherinezhad et al. (2016) combine computer vision and text-based information extraction to learn the relative sizes of objects; Forbes and Choi (2017) crowdsource physical knowledge along specified dimensions and employ belief propagation to learn relative physical attributes of object pairs. Wang et al. (2017) propose a multimodal LDA to learn the definitional properties (e.g. *animal, four-legged*) of entities. Zhang et al. (2017) study the role of common-sense knowledge in natural language inference, which is inherently between-events rather than single-event focused. Prior work does not specifically handles the (single-event) semantic plausibility task and related efforts do not necessarily adapt well to this task, as we will show, suggesting that new approaches are needed.

3 Data

To study the semantic plausibility of S-V-O events, specifically *physical* semantic plausibility, we create a dataset[1] through Amazon Mechanical Turk with the following criteria in mind: (i) *Robustness*: Strong inter-annotator agreement; (ii) *Diversity*: A wide range of typical/atypical, plausible/implausible events; (iii) *Balanced*: Equal number of plausible and implausible events.

In creating physical events, we work with a fixed vocabulary of 150 concrete verbs and 450 concrete nouns from Brysbaert et al. (2014)'s word list, with a concreteness threshold of 4.95 (scale: 0-5). We take the following steps:

(a) Have Turkers write down plausible or implausible S-V and V-O selections;

(b) Randomly generate S-V-O triples from collected S-V and V-O pairs;

(c) Send resulting S-V-O triples to Turkers to filter for ones with high agreement (by majority vote).

(a) ensures diversity and the cleanness of data (compared with noisy selectional preference data collected unsupervised from free text): the Turkers are instructed (with examples) to (i) consider both typical and atypical selections (e.g. *man-swallow-** with *candy* or *paintball*); (ii) disregard metaphorical uses (e.g. *feel-blue* or *fish-idea*). 2,000 pairs are collected in the step, balancing typical and atypical pairs. In (b), we manually filter error submissions in triple generation. For (c), 5 Turkers provide labels, and we only keep the ones that have $\geq$ 3 majority votes, resulting with 3,062 triples (of 4,000 annotated triples, plausible-implausible balanced), with **100% $\geq$ 3 agreement, 95% $\geq$ 4 agreement, and 90% 5 agreement**.

To empirically show the failure of distribution-only methods, we run Van de Cruys (2014)'s neural net classifier (hereforth NN), which is one of the strongest models designed for selectional preference (Figure 1, left-box). Let $\boldsymbol{x}$ be the concatenation of the embeddings of the three words in an S-V-O triple. The prediction $P(y|\boldsymbol{x})$ is computed as follows:

$$P(y=1|\boldsymbol{x}) = \sigma_2(W_2\sigma_1(W_1\boldsymbol{x})) \qquad (1)$$

where σ is a nonlinearity, W are weights, and we use 300D pretrained GloVe vectors (Pennington et al., 2014). The model achieves an accuracy of 68% (logistic regression baseline: 64%) after fine-tuning, verifying the intuition that distributional data alone cannot satisfactorily capture the semantics of physical plausibility.

4 World Knowledge Features

Recognizing that a distribution-alone method lacks necessary information, we collect a set of world knowledge features. The feature types derive from inspecting the high agreement event triples for knowledge missing in distributional selection (e.g. relative sizes in *man-swallow-paintball/desk*). Previously, Forbes and Choi (2017) proposed a three level (3-LEVEL) featurization scheme, where an object-pair can take 3

[1] Link: `https://github.com/suwangcompling/Modeling-Semantic-Plausibility-NAACL18/tree/master/data`.

values for, e.g. relative size: $\{-1, 0, 1\}$ (i.e. lesser, similar, greater). This method, however, does not explain many cases we observed. For instance, *man-hug-cat/ant*, *man* is larger than both *cat* and *ant*, but the latter event is implausible. 3-LEVEL is also inefficient: k objects incur $O(k^2)$ elicitations. We thus propose a binning-by-landmark method, which is sufficiently fine-grained while still being efficient and easy for the annotator: given an entity n, the Turker decides to which of the landmarks n is closest to. E.g., for SIZE, we have the landmarks $\{$*watch, book, cat, person, jeep, stadium*$\}$, in ascending sizes. If $n =$ *dog*, the Turker may put n in the bin corresponding to *cat*. The features[2] are listed with their landmarks as follows:

- SENTIENCE: *rock, tree, ant, cat, chimp, man*.
- MASS-COUNT: *milk, sand, pebbles, car*.
- PHASE: *smoke, milk, wood*.
- SIZE: *watch, book, cat, person, jeep, stadium*.
- WEIGHT: *watch, book, dumbbell, man, jeep, stadium*.
- RIGIDITY: *water, skin, leather, wood, metal*.

5 Turkers provide annotations for all 450 nouns, and we obtained **93% $\geq$ 3 agreement, 85% $\geq$ 4 agreement, and 79% 5 agreement**.

Our binning is sufficiently granular, which is crucial for semantic plausibility of an event in many cases. E.g. for *man-hug-cat/ant*, *man, cat* and *ant* fall in the $4^{th}, 3^{rd}$ and 1^{st} bin, which suffices to explain why *man-hug-cat* is plausible while *man-hug-ant* is not. Compared to past work (Forbes and Choi, 2017), it is efficient. Each entity only needs one assignment in comparison to the landmarks to be located in a "global scale" (e.g. from the smallest to the largest objects), and even for extreme granularity, it only takes $O(k \log k)$ comparisons. It is also intuitive: differences in bins capture the intuition that one can *hug* smaller objects as long as those objects are not too small.

5 Models

We answer two questions: (i) Does world knowledge improve the accuracy of semantic plausibility classification? (ii) Can we minimize effort in knowledge feature annotation by learning from a small amount of training data?

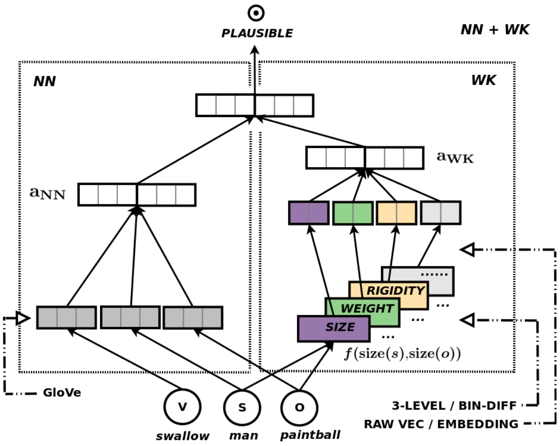

Figure 1: Model architecture (example input: *man-swallow-paintball*). Left box: Van de Cruys (2014)'s neural net (**NN**, embeddings only); Right box: world knowledge feature net **WK** with different modeling choices (Section 5). Only SIZE, WEIGHT, RIGIDITY are shown; the rest receive the same treatment. **NN + WK**: embedding and world knowledge combined.

For question (i), we experiment with various methods to incorporate the features on top of the embedding-only **NN** (Section 3). Our architecture[3] is outlined in Figure 1, where we ensemble the **NN** (left-box) and another feedforward net for features (**WK**, right-box) to produce the final prediction. For the feature net, the relative physical attributes of the subject-object pair can be encoded in **3-LEVEL** (Section 4) or the bin difference (**BIN-DIFF**) scheme.[4] For BIN-DIFF, given the two entities in an S-V-O event (i.e. S, O) *ant* and *man*, which are in the bins of the landmark *watch* (i.e. the 1^{st}) and that of *person* (i.e. the 4^{th}), the pair *ant-man* gets a BIN-DIFF value of $1 - 4 = -3$. Exemplifying the featurization function $f(s, o)$ with SIZE:

$$f_{\text{3-L}}(\text{SIZE}(s), \text{SIZE}(o)) \in \{-1, 0, 1\} \quad (2)$$

$$f_{\text{BIN}}(\text{SIZE}(s), \text{SIZE}(o)) = \text{BIN}(s) - \text{BIN}(o) \quad (3)$$

Then, given a featurization scheme, we may feed raw feature values (**RAW VEC**, for 3-LEVEL, e.g. concatenation of -1, 0 or 1 of all feature types, in that order, and in one-hot format), or feature embeddings (**EMBEDDING**, e.g. concatenation of embeddings looked up with feature values). Fi-

[2] We experimented with numerous feature types, e.g. size, temperature, shape, etc. and kept the subset that contributes most substantially to semantic plausibility classification. More details on the feature types in supplementary material (`https://github.com/suwangcompling/Modeling-Semantic-Plausibility-NAACL18/tree/master/supplementary`).

[3] More configuration details in supplementary material.

[4] We also tried using *bin numbers* directly, however it does not produce ideal results (classification accuracy between 3-LEVEL and BIN-DIFF). Thus for brevity we drop this setup.

MODELS	5%	20%
Label Spreading (Zhu et al., 2004)	0.56	0.59
Factor Graph (Forbes and Choi, 2017)	0.69	0.71
Multi-LDA (Wang et al., 2017)	0.64	0.72
Logistic Regression	0.72	0.83
Factor Graph (initialized with our LR)	0.72	0.84
Ordinal-LR	**0.76**	**0.88**

MODELS	5%		20%	
	3-L	BIN	3-L	BIN
Logistic Regression	0.61	0.21	0.68	0.26
Ordinal-LR	**0.66**	**0.32**	**0.76**	**0.40**

Table 2: Feature Propagation. Top-table: results on Forbes and Choi (2017)'s 2.5k object pair data; Bottom-table: results on our 10k object pair data.

nally, let $\boldsymbol{a}_{\text{NN}}, \boldsymbol{a}_{\text{WK}}$ be the penultimate-layer vectors of **NN** and **WK** (see Figure 1), we affine transform their concatenation to predict label $\hat{y}$ with argmax on the final softmax layer:

$$\hat{y} = \underset{y}{\text{argmax}}\ \text{softmax}(\sigma(W[\boldsymbol{a}_{\text{NN}}; \boldsymbol{a}_{\text{WK}}] + \boldsymbol{b})) \quad (4)$$

where σ is a ReLU nonlinearity. We will only report the results from the best-performing model configuration, which has BIN-DIFF + EMBEDDING. The model will be listed below as **NN + WK-GOLD** (i.e. with **GOLD**, Turker-annotated **W**orld **K**nowledge features).

For question (ii), we select a data-efficient feature learning model. Following Forbes and Choi (2017) we evaluate the models with 5% or 20% of training data. We experiment with several previously proposed techniques: (a) *label spreading*; (b) *factor graph*; (c) *multi-LDA*. As a baseline we employ a simple but well-tuned logistic regressor (LR). We also initialize the factor graph with this LR, on account of its unexpectedly strong performance.[5] Finally, observing that the feature types are inherently ordinal (e.g. SIZE from small to large), we also run *ordinal logistic regression* (Adeleke and Adepoju, 2010). For model selection we first evaluate the object-pair attribute data collected by Forbes and Choi (2017), 2.5k pairs labeled in the 3-LEVEL scheme. We then compared the the LR and Ordinal-LR (our strongest models[6] in this experiment) on 10k randomly generated object-pairs from our annotated nouns. The results are summarized in Table 2, where we see

MODELS	ACCURACY			
Random	0.50			
LR baseline	0.64			
NN (Van de Cruys, 2014)	0.68			
NN + WK-GOLD	**0.76**			
	5%		20%	
NN + WK-PROP	3-L	BIN	3-L	BIN
	0.69	0.70	0.71	0.74

Table 3: Semantic Plausibility (binary) Classification. The average of 10-fold CV (splitting on the total 3,062 entries). The neural classifier injected with full annotation of world knowledge (i.e. **NN + WK-GOLD**) performs substantially better, and the performance retainment is rather strong with propagated features (by Ordinal-LR) from small fractions of gold annotation (i.e. in **NN + WK-PROP**).

(i) 3-LEVEL propagation is much easier; (ii) our object-pairs are more challenging, likely due to sparsity with larger vocabulary size; (iii) ordinality information contributes substantially to performance. The model that uses propagated features (w/ Ordinal-LR) will be listed as **NN + WK-PROP**.

6 Semantic Plausibility Results

We evaluate the models on the task of classifying our 3,062 S-V-O triples by semantic plausibility (10-fold CV, taking the average over 20 runs with the same random seed). We compare our three models in the 3-LEVEL and BIN-DIFF schemes, with **NN + WK-PROP** evaluated in 5% and 20% training conditions. The results are outlined in Table 3. Summarizing our findings: (i) world knowledge undoubtedly leads to a strong performance boost (~8%); (ii) BIN-DIFF scheme works much better than 3-LEVEL — it manages to outperform the latter even with much weaker propagation accuracy; (iii) the accuracy loss with propagated features seems rather mild with 20% labeled training and the best scheme.

7 Error Analysis

To understand what challenges remain in this task, we run the models above 200 times (10-fold CV, random shuffle at each run), and inspect the top 200 most frequently misclassified cases. The percentage statistics below are from counting the error cases.

In the cases where **NN** misclassifies while **NN + WK-GOLD** correctly classifies, 60% relates to SIZE and WEIGHT (e.g. missing *man-hug-ant*

[5] We verified our setup with the authors and they attributed the higher performance of our LR to hyperparameter choices.

[6] Because the factor graph + LR gives very slight improvement, for simplicity we choose LR instead.

(bad) or *dog-pull-paper* (good)). PHASE takes up 18% (e.g. missing *monkey-puff-smoke* (good)). This validates the intuition that distributional contexts do not encode these types of world knowledge.

For cases often misclassified by *all* the models, we observe two main types of errors: (i) data sparsity; (ii) highly-specific attributes.

Data sparsity (32%). *man-choke-ant*, e.g., is a singleton big-object-choke-small-object instance, and there are no distributionally similar verbs that can help (e.g. *suffocate*); For *sun-heat-water*, because the majority of the actions in the data are limited to solid objects, the models tend to predict implausible for whenever a gas/liquid appears as the object.

Highly-specific attributes (68%). "long-tailed" physical attributes which are absent from our feature set are required. To exemplify a few:[7]

- *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.

The analysis shows that the task and the dataset highlights the necessity for more sophisticated knowledge featurization and cleverer learning techniques (e.g. features from computer vision, propagation methods with stronger capacity to generalize) to reduce the cost of manual annotation.

8 Conclusion

We present the novel task of *semantic plausibility*, which forms the foundation of various interesting and complex NLP tasks in event semantics (Bowman et al., 2016; Mostafazadeh et al., 2016; Li and Jurafsky, 2017). We collected a high-quality dedicated dataset, showed empirically that the conventional, distribution data only model fails on the task, and that clever world knowledge injection can help substantially with little annotation cost, which lends initial empirical support for the scalability of our approach in practical applications, i.e. labeling little but propagating well approximates performance with full annotation. Granted that annotation-based injection method does not cover the full spectrum of leverageable world knowledge information (alternative/complementary sources being images and videos, e.g. Bagherinezhad et al. 2016), it is indeed irreplaceable in some cases (e.g. features such as WEIGHT or RIGIDITY are not easily learnable through visual modality), and in other cases presents a low-cost and effective option. Finally, we also discovered the limitation of existing methods through a detailed error analysis, and thereby invite cross-area effort (e.g. multimodal knowledge features) in the future exploration in automated methods for semantic plausibility learning.

Acknowledgments

This research was supported by NSF grant IIS 1523637. Further, this material is based on research sponsored by DARPA under agreement number FA8750-18- 2-0017. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. We acknowledge the Texas Advanced Computing Center for providing grid resources that contributed to these results. We would also like to thank our reviewers for their insightful comments.

References

K.A. Adeleke and A.A. Adepoju. 2010. Ordinal Logistic Regression Model: An Application to Pregnancy Outcomes. *Journal of Mathematics and Statistics* 6(3):279–285.

Hessam Bagherinezhad, Hannaneh Hjishirzi, Yejin Choi, and Ali Farhadi. 2016. Are Elephants Bigger

[7] Percentages calculated with the 68% as the denominator. Full list in supplementary material.

Than Butterflies? Reasoning about Sizes of Objects. In *Proceedings of AAAI*.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *Proceedings of CoNLL*.

M Brysbaert, A.B. Warriner, and V. Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas. *Behavior Research Methods* 46:904–911.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *Proceedings of EMNLP*.

Stephen Clark and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of NAACL*.

Katrin Erk and Sebastian Padó. 2010. Exemplar-Based Models for Word Meaning in Context. In *Proceedings of ACL*.

Maxwell Forbes and Yejin Choi. 2017. VERB PHYSICS: Relative Physical Knowledge of Actions and Objects. In *Proceedings of ACL*.

Jiwei Li and Daniel Jurafsky. 2017. Neural Net Models of Open-domain Discourse Coherence. In *Proceedings of EMNLP*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of NAACL*.

Diarmiuid O'Seaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of ACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Embeddings. In *Proceedings of EMNLP*.

Philip Resnik. 1996. Selectional Constraints: An Information-Theoretic Model and Its Computational Realization. *Cognition* 61:127–159.

Ottokar Tilk, Vera Demberg, Asad Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event Participant Modeling with Neural Networks. In *Proceedings of EMNLP*.

Tim Van de Cruys. 2014. A Neural Network Approach to Selectional Preference Acquisition. In *Proceedings of EMNLP*.

Su Wang, Stephen Roller, and Katrin Erk. 2017. Distributional Modeling on a Diet: Learning Word Properties from Text Only. In *Proceedings of IJCNLP*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal Common-sense Inference. *TACL* 5:379–395.

Dengyong Zhu, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with Local and Global Consistency. In *Proceedings of NIPS*.