Can LLMs Disambiguate Grounded Language? The Case of PP Attachment

John Blackmore
Rutgers University
john.blackmore@rutgers.edu

Matthew Stone
Rutgers University
matthew.stone@rutgers.edu

Abstract

We study resolution of ambiguity in prepositional phrase attachment by Large Language Models in the zero-shot setting. We evaluate a strong "plausibility" baseline derived from token probabilities of descriptions encoding alternative attachments, and explore possible improvements using additional token probabilities that reflect aspects of information structure. Error analysis suggests directions for more sophisticated tools, common-sense reasoning, world knowledge, and additional context to better resolve ambiguity.

1 Introduction

Thanks to improved curation and training processes and ever-larger datasets, Large Language Models (LLMs) can make increasingly sophisticated generalizations about the regularities exhibited in English text. In this work, we use these generalizations as a lens to gain a deeper understanding of ambiguity and the factors involved in resolving it.

We focus on prepositional phrase attachment—in particular, the ambiguous attachments exhibited in English in descriptions such as (1), where a noun phrase X is followed by successive prepositional phrases p_1 Y and p_2 Z.

(1) Brown and white dog with brown and black dog on gray sheet

Here X is *Brown and white dog*, p_1 is *with*, Y is *brown and black dog*, p_2 is *on* and Z is *gray sheet*. Vividly—but uncharacteristically—the attachment ambiguity of (1) is salient to human readers: is dog X on sheet Z, is dog Y on the sheet, or are they both? In fact, (1) is the caption in the dataset of Young et al. (2014) for the image shown in Figure 1.

PP attachment ambiguity is a microcosm of the syntactic, semantic and pragmatic challenges of reconstructing the interpretation of an utterance that

Figure 1: Brown and white dog with brown and black dog on gray sheet. (CC BY-SA 2.0 via flickr.com)

the speaker had in mind. The open-ended considerations involved mean that as recently as 2016, in a blog post describing Google's SyntaxNet as "The world's most accurate parser," Slav Petrov wrote "the major source of errors at this point are examples such as the prepositional phrase attachment ambiguity described above, which require real-world knowledge (e.g. that a street is not likely to be located in a car) and deep contextual reasoning."[1] Researchers rarely focus on such precise problems nowadays, but the resources and tools now available in NLP open up possibilities for understanding the fundamental nature of these ambiguities and how to resolve them. Concretely, in this paper we explore the following problem. Suppose you had a perfect knowledge source about what happens in the world and what people tend to say. How would you query that knowledge source to resolve PP attachment?

As Petrov suggests, real-world plausibility of the situation described is a crucial factor. Plausibility demands Y attachment for (2), for example, since cakes get frosted and children do not.

(2) A child with a cake with rainbow frosting.

[1] "https://opensource.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html"

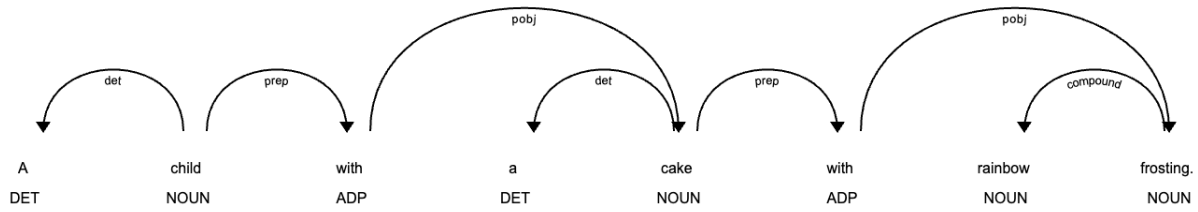

Figure 2: Dependency parse tree of *A child with a cake with rainbow frosting*. This shows *Y* attachment, but the alternate *X* attachment would replace the arrow from *cake* to *with* with an arrow from *child* to *with*

.

But this is not the only factor. A separate question is whether the resolution makes for an interesting and relevant contribution. Take (3).

(3) dog with head in snow.

Even if it's unlikely for dogs to bury their heads in snow, we shouldn't favor an *X* attachment in (3). That reading would implicate that dogs sometimes don't have heads. Similarly, a sensible image caption should focus on what is in the foreground, so regardless of *X* and *Y* you should predict *Y* attachment in cases such as these:

(4) a. A cake with a child in the background.
b. A boy with a child in the background.

A further consideration is whether the description is fluent and natural. For example, *of* usually precedes *in* if they modify the same noun. Therefore, we should expect *Y* attachment in (5) regardless of real-world plausibility.

(5) picture in gallery of sculpture.

X attachment would have been realized by *picture of sculpture in gallery*. In linguistics, the decisive factors beyond real-world knowledge required to disambiguate the PPs of (3–5) fall under the category of **information structure** (Féry and Ishihara, 2016; Krifka, 2008).

In this paper, we explore the relationship of real-world knowledge and information structure in disambiguation both qualitatively and quantitatively. We use image captions to develop a corpus of potentially ambiguous PP attachments whose interpretations can be assessed by visual grounding.[2] We explore the use of zero-shot probes derived using LLMs to resolve these ambiguities, using continuation probabilities to track plausibility and completion probabilities to track information structure. While continuation probability from Llama 3.1 (Touvron et al., 2023) is highly reliable (89.4% correct in our experiments), adding information structure seems to further reduce errors (to 92.9%). Comparable performance can be obtained by calibration to address reporting bias (Gordon and Van Durme, 2013; Liu et al., 2023). Corpus and error analysis underscores that multiple cues align in most natural examples. We therefore recommend that future work refine and assess these queries using probing studies with minimal pairs.

[2] https://github.com/depthfirst/pp-ambiguity

2 Background

A preposition in English, to employ a standard definition, is "a word governing (and usu. preceding) a noun or pronoun and expressing a relation to another word or element" (Fowle and Burchfield, 2000).

Prepositions express varied meanings: locating entities in space and time (*meet at Heathrow Airport at six o'clock*); indicating roles in events (*open with a knife*); describing physical and functional relationships (*book on the table with the red lamp in the living room*)—and, as with *of*, underspecifying relationships that must be inferred in context (*hair of gold, photo of a dog, piece of furniture*) (McArthur et al., 2018).

English allows consecutive prepositional phrases. This creates a structural ambiguity about what phrase the second PP modifies or "attaches" to. Alternative attachments do not always lead to a difference in meaning. For example, in *people on a ride at a fair*, if people are on a ride, and the ride is at a fair, then the people must also be at the fair—and conversely, if the people are at the fair, then the ride must be too. In cases like these, we say that the attachment choice doesn't matter. In our data, such cases are surprisingly frequent.

Recent syntactic probe studies such as Zhou et al. (2023) have noted that LLMs do not perform well on prompts with prepositional phrase ambiguity. Our specific focus in this work, however, is on **grounded** language, that is, utterances that presume a shared contextual understanding with the reader. In our study, this contextual grounding comes from photographs. Practically, this makes it possible to be confident about the intended interpretation in cases such as (1). Despite our focus on grounded language, however, we are not taking a multimodal approach to understanding. We are less focused on the cases where disambiguation depends on visual context (surprisingly rare in our data); rather, we use the clear, physical interpretation of our examples as a test case to explore the interactions between common-sense knowledge, plausibility, and speaker goals.

PP attachment has long been a proving ground for the use of data-driven methods in language technology, going back to Brill and Resnik (1994). Work on PP attachment by Collins and Brooks (1995) and Ratnaparkhi et al. (1994) helped catalyze statistical parsing by demonstrating the power of corpus frequency and machine learning to give probabilistic approximations to common-sense plausibility. Many more sophisticated methods have since been proposed, including "compositional" neural network architectures (Belinkov et al., 2014; Socher et al., 2013), word embeddings (Dasigi et al., 2017), and ontological and knowledge-based methods (Allen et al., 2020; Dasigi et al., 2017; Nakashole and Mitchell, 2015). Nevertheless, as Petrov's quote indicates, a certain amount of pessimism has remained about the general effectiveness of such solutions. Since then, despite the general improvements of LLMs and the importance of PP attachment in practical dialogue—in dialogue with robots, for example Tellex et al. (2011)—no dedicated evaluation of current PP attachment performance has been available.

In fact, progress in LLMs allows us not only to document improved performance but also to revisit the conceptual underpinnings of data-driven strategies for disambiguation. The basic objective of LLMs—for better or worse—is to predict the probability of word sequences (Bender and Koller, 2020). Such objectives conflate what happens in the world with what people think is worth mentioning. This can make it difficult to find written evidence for obvious and therefore uninteresting truths: this problem of **reporting bias** is a significant constraint on the acquisition of knowledge from text (Gordon and Van Durme, 2013). The effectiveness of zero-shot disambiguation of PPs with LLMs will depend on how reporting bias affects LLM estimates of real-world plausibility.

This potential conflation of semantics and pragmatics also complicates the effort to assess alternative interpretations on pragmatic grounds. In general, **information structure** refers to the articulation of an assertion into subject and predicate, topic and comment, old and new information, or question and answer, as well as the reconstruction of implicit assumptions, comparisons, and contrasts from the linguistic choices of the author (Féry and Ishihara, 2016; Krifka, 2008). We might expect some of this information to be factored into the continuation-probability predictions of LLMs, but we need to find the right reference points and queries to successfully probe this knowledge.

3 Approach

3.1 Data

We chose to use image caption data for our study because it provides a good source of grounded language with potential for syntactic ambiguity that can be resolved through a variety of approaches.

FlickR 30k (Young et al., 2014) is a dataset of over 30,000 images that have been given over 150,000 captions by crowdsourced workers. They are all taken from outdoor photos. We have performed some unsupervised corpus mining of this dataset for our study.

Using spaCy[3] for part-of-speech tagging and chunking, we extracted all image captions with a noun phrase (NP) followed by two consecutive prepositional phrases (PP). Further curation was done to check for tagging errors, for example for any verbs that were tagged as nouns.

Using these heuristics, we extracted 730 of what we refer to as *5-tuples* containing ambiguity in the prepositional phrase attachment. All examples are image captions with two consecutive prepositional phrases following a noun phrase. Symbolically, we write these tuples in the form $X\ p_1\ Y\ p_2\ Z$. The ambiguity is in the attachment of $p_2\ Z$, where the valid choices are X or Y. We then labeled each example with the correct attachment (X or Y) or "doesn't matter" (N), and whether or not the image is needed for disambiguation.

[3] https://spacy.io/

Prep	p1	p2	both
with	85	36	0
in	64	67	4
on	21	51	0
of	16	33	1
at	9	2	0

Table 1: Distribution of most common prepositions in ambiguous attachment cases. *p1* - appears as the first preposition. *p2* - appears as the second preposition. *both* appears in the caption as both prepositions.

We excluded the majority of these captions (532/730), because the attachment selection had no effect on the grounded meaning. In most of these cases (322), it was because of the common-sense transitive property of some prepositional relations:

(6) a. Young girls at a pool on a bright sunny day.
b. A man in green pants on a bicycle.

In other cases (138), the X noun denoted a group or collective and p_1 was *of*. In such cases, it doesn't matter whether the second PP describes the collection or its members.

(7) a. A bunch of boats in the nice blue water.
b. A crowd of people around a bike rail.

We leave the identification of such spurious ambiguities as a problem for future work. While the group or collective noun cases should be easy to detect, predicting common-sense transitivity is more challenging.

After exclusions, we were left with 198 captions where attachment choice affects grounded meaning. Of these, 82 were X attachments and 116 were Y attachments. Several (19/198) required visual context (i.e., the image) to determine the correct attachment. The most common prepositions were *with*, *in*, *on*, and *of*. See Table 1 for more details.

3.2 Models

3.2.1 Llama

We chose Llama 3.1 8B (Touvron et al., 2023) because it is a recent and left-to-right generative large language model (LLM) and offers a stable, reproducible API for obtaining token probabilities for zero-shot probes.

3.2.2 VERA

VERA-T5 7B (Liu et al., 2023) predicts the plausibility of a statement. Training data consists of 7M correct and incorrect statements converted from 19 QA datasets and 2 KBs. We chose VERA because the authors have shown promising evidence that VERA has made progress in overcoming the reporting bias that affects many language models. In particular, cooperative speakers don't belabor the obvious (Grice, 1975), so the more common or expected something is, the less likely it is to be mentioned as the primary intent of the utterance (Gordon and Van Durme, 2013). The result is that descriptions of very rare events can seem much more likely than descriptions of commonplace but unremarkable ones.

VERA computes a score from the internal representation of a statement by taking the last hidden state (**h**) of the EOS token. Then a linear layer projects **h** to a scalar logit followed by a sigmoid function. The result is a real-valued score $s \in [0, 1]$. The model is trained with a linear combination of three loss functions: a binary classification loss, a multi-class loss, and a supervised contrastive loss. It is calibrated so that its confidence in its predictions is approximately equal to the actual frequency of correctness.

3.3 Experiments

Our experiments quantify the capability of simple LLM computations to disambiguate prepositional phrase attachments in grounded language. We analyze the error cases for further insights into disambiguation.

3.3.1 Plausibility

To assess the plausibility of a possible attachment, we construct a prompt that captures the context for the attachment decision and makes the intended resolution explicit. Concretely, take a caption X p_1 Y p_2 Z as in (8a). First, we supply a context sentence *There (is|are)* X p_1 Y. to ensure the coherence of disambiguating continuations. (The verb is chosen to agree in number with X.) Then we continue with a disambiguating followup: either with the characterization X^* *(is|are)* p_2 Z. as in (b), or Y^* *(is|are)* p_2 Z. as in (c). (To compute X^* and Y^* we remove any determiner and add *The*; the verb is chosen to agree with the subject.)

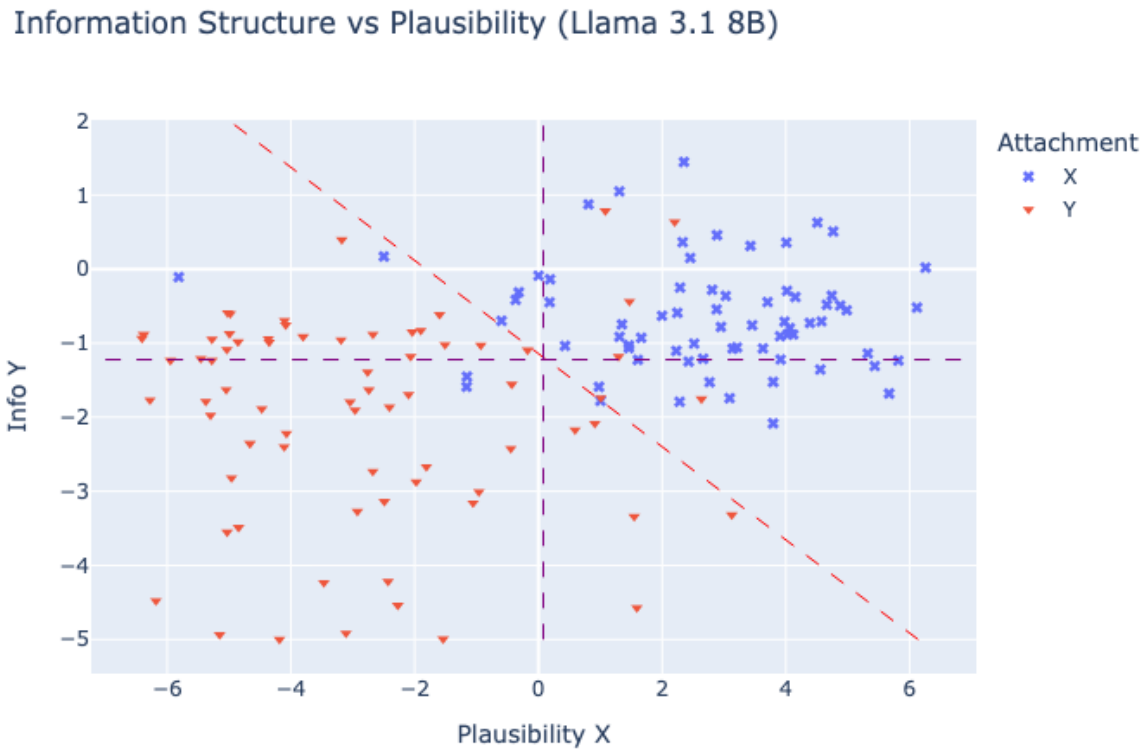

Figure 3: Information Structure (Info Y) vs Plausibility (X) from Llama (equations 3 and 1, respectively) on a sample of 198 image captions from Flickr30K (Young et al., 2014). Dashed lines represent decision boundaries from the model trained from one or both metrics - the horizontal line depends on IS only, the vertical on P only, and the diagonal on P+IS.

(8) a. A toddler with eggs in a bowl.
b. There is a toddler with eggs. The toddler is in a bowl.
c. There is a toddler with eggs. The eggs are in a bowl.

For Llama, we get estimates of $P(p_2Z|X)$ and $P(p_2Z|Y)$ by multiplying the softmax of the logits from the top hidden layer of Llama's output for the tokens of the second PP for the respective prompts. The log-likelihood ratio of the scores indicates which attachment is more likely.

$$P_X = ln(\frac{P(p_2Z|X)}{P(p_2Z|Y)} \quad (1)$$

The scatterplot in Figure 3 shows this quantity as the X-coordinate.

For VERA, we score the two prompts in their entirety. The output will be a score in the interval $[0,1]$, where a higher score indicates a more plausible statement. We again take the log of the ratio of the two scores.

$$P_X = ln(\frac{s(Xp_2Z)}{s(Yp_2Z)} \quad (2)$$

3.3.2 Information Structure

Intuitively, information structure correlates with attachment in captions of the form $X\ p_1\ Y\ p_2\ Z$ because of the two different possible functions of the phrase $p_1\ Y$. If X is the topic, then $p_1\ Y\ p_2\ Z$ is a joint or combined comment, in which case $p_1\ Y$ and $p_2\ Z$ both attach to X. If Y is the topic, X provides framing context, and $p_2\ Z$, the main comment, attaches to Y.

Thus, if $X\ p_1\ Y$ has a low probability as a standalone caption (e.g., *dog with head*), this suggests that the description is incomplete and further comment is required—in other words, that subsequent PPs (e.g., *in snow*) should prefer Y attachment. The hypothesis we want to address is whether this model of information structure can improve disambiguation in cases where $X\ p_2\ Z$ may look more plausible than $Y\ p_2\ Z$ but X attachment imposes an

anomalous pragmatic interpretation for $X\ p_1\ Y$, as in (3).

We compute information structure score for Llama using the logits associated with final punctuation in $X\ p_1\ Y.$ and $X\ p_1\ Y\ p_2\ Z.$:

$$IS_Y = ln(\frac{P(.|Xp_1Y)}{P(.|Xp_1Yp_2Z)}) \quad (3)$$

(Note: we add a '.' at the end of every caption if it does not have one.) A lower value would indicate that the complete caption ($X\ p_1\ Y\ p_2\ Z$) is more plausible than $X\ p_1\ Y$, suggesting a frame–topic reading and thus Y attachment.

To derive a metric for information structure using VERA, we construct prompts that isolate $X\ p_1\ Y$ and the original caption, $X\ p_1\ Y\ p_2\ Z$ and insert the phrase *Here we have* before the start of the caption to make it a complete sentence.

(9) a. Here we have dog with head in snow.
b. Here we have dog with head.

We again take the log of the ratio of the two plausibility scores.

$$IS_Y = ln(\frac{s(Xp_1Y)}{s(Xp_1Yp_2Z)}) \quad (4)$$

3.4 Metrics

We use the quantities from equations 1 through 4 to classify attachments. To measure the contribution of information structure **over and above** plausibility, we compare the predictions of a joint model (P+IS) to predictions plausibility (P) alone. To build a joint P+IS model, we use linear SVC with weights trained by 5-fold cross validation.

4 Results

The scatterplot and decision boundaries for the Llama model are shown in Figure 3. The vertical boundary is based on the plausibility dimension alone. We can thus see precisely where the contribution of information structure comes into play.

The accuracy of the model trained from the plausibility and information structure metrics from VERA was 92.4% overall. The model trained from the plausibility metric alone was 92.9% accurate, and from information structure alone, 63.6%.

The accuracy of the model trained from the plausibility and information structure metrics from Llama 3.1 was 93.5% overall. The model trained from the plausibility metric alone was 88.9% accurate, and from information structure alone, 73.2%.

5 Analysis

For the joint model experiment with Llama, as Figure 3 shows, in cases where the plausibility score is near 0 (neutral) and the Xp_1Y probability is much lower than the $X\ p_1\ Y\ p_2\ Z$ probability, the attachment is predominantly Y, as expected. Conversely, when the $X\ p_1\ Y$ probability is higher than the $X\ p_1\ Y\ p_2\ Z$ probability, the attachment is predominantly X.

These results show that the plausibility of the respective attachments is a strong factor in disambiguation, but there is a noteworthy contribution from our measure of information structure. The captions that IS corrects are shown in Table 2.

With the help of some simple patterns and rules, we investigate where we need more sophisticated tools, common-sense reasoning, world knowledge, or something other than context-free pattern recognition. The patterns and results are summarized in Table 3. The third-person pronoun pattern matches examples with a pronoun in the last noun phrase (Z).

(10) a. A woman with feathers in her hair.

The *p2='of'* pattern matches when 'of' appears the second preposition; *in/on* when the prepositions are *in* and *on*, respectively, and *(fore|background)* when the caption ends with *in the foreground* or *in the background*.

On the subset of cases that were matched by the simple rules and patterns, the plausibility model for Llama was 93% accurate, and on the rest, 85%. The joint model for Llama was 97% accurate on the cases matched by simple rules and patterns, 90% on the others. The cases that this model classified incorrectly are shown in Tables 4 and 5, respectively.

On the subset of cases that were matched by the simple rules and patterns, the plausibility model for VERA was 99% accurate, and on the rest, 87%. The joint model for VERA was 98% accurate on the cases matched by simple rules and patterns, 87% on the others. The cases that this model classified incorrectly are shown in Tables 6 and 7, respectively.

6 Conclusions and Future Work

Many of the cases that the language models could not disambiguate correctly had been labeled as needing visual context to resolve. Among next

Caption	P	IS	Att
Some patrons at a some sort of diner.	0.43	-6.28	Y
Two boys with blue swim caps in the murky water.	0.18	-0.45	X
A child with his mouth on a red plastic toy.	0.51	-5.73	Y
Kids at a water park on steps.	0.19	-0.14	X
Two children at a medieval picture with face cutouts.	0.91	-2.08	Y
A busy downtown New York City with a taxi in the foreground.	1.55	-3.34	Y
A winter landscape with two people in the foreground.	0.59	-2.17	Y
Dog with head in snow.	3.13	-3.32	Y
a man by himself in a building.	-0.60	-0.70	X
Brown and white dog with brown and black dog on gray sheet.	-0.32	-0.31	X
Lady in black hoolahoops on the street.	-0.37	-0.42	X
A street shot of people in an Asian country.	1.59	-4.57	Y

Table 2: Captions classified incorrectly with plausibility (P) alone, corrected by the joint model of plausibility (P) and information structure (IS) (equations 1 and 3, respectively) from Llama. The Att column shows the correct attachment.

				Llama		VERA	
Pattern	**N**	**X**	**Y**	**P**	**P+IS**	**P**	**P+IS**
third-person pron	39	3	36	39 (100%)	39 (100%)	39 (100%)	39 (100%)
p2='of'	33	0	33	31 (76%)	32 (97%)	33 (100%)	31 (94%)
in/on	17	17	0	16 (94%)	17 (100%)	16 (94%)	17 (100%)
(fore|back)ground	9	0	9	5 (56%)	7 (78%)	9 (100%)	9 (100%)

Table 3: Classification accuracy for captions matching specific patterns.

steps should be to leverage vision and/or multimodal models in a similar study in disambiguation.

The proportion of ambiguous attachments that did not affect the grounded meaning was surprising. A further study into how well we can predict whether or not the attachment matters would seem to be in order.

We would also like to probe further to determine whether the models are systematic in their ability to handle certain patterns, for example with minimal pairs of *in the foreground* examples, and test how they're attached.

We would like to develop additional tests to cover syntactic restrictions, for example normal order of stacked prepositions and impossibility of multiple attachments to the same node with the same relation. We could try substituting a pronoun for one of the NPs which can't be modified by a PP. For example, *a sculpture on a stack of books* vs. *a stack and a sculpture on it of books* vs *a stack of books and a sculpture on it* as an alternative probe.

LLMs can do some paraphrasing which might enable sense disambiguation for prepositions. For example, paraphrasing a *talk on flowers on Tuesday* as a *talk about flowers happening Tuesday* and therefore concluding that it's two different senses and therefore $X\ p_2\ Z$ is a possible attachment.

7 Limitations

Our study explores the use of zero-shot probes training linear classifiers from VERA's and Llama's outputs, with a focus on showing the contribution of information structure in the resolution of ambiguous preposition phrase attachments. Our results are given on a small sample of labeled data using a cross validation approach with 5 folds. The ground truth labels were provided by the authors of this paper. This study should not be regarded as an unbiased statistical evaluation of performance, but rather a study into how PP attachments are resolved.

This study is focused exclusively on PP attachments in English. Other languages may require alternate prompting strategies for the IS metric, for example in head-final languages such as Japanese.

Acknowledgments

Thanks to Denson George, Rich Magnotti and the reviewers for helpful feedback. Supported by NSF awards 2021628, 2119265, and 2427646.

Caption	P	IS	Att	Explanation
A train engine with a water tower in the background	1.30	-1.18	Y	Better P, IS
A picture of the clarinet section of a band	2.64	-1.75	Y	Better P
An elderly person with a brick wall in the background	1.01	-1.74	Y	Better P, IS

Table 4: Cases matched by simple patterns matched, classified incorrectly by the joint model of Plausibility (P) and Information Structure (IS) from Llama (equations 1 and 3, respectively). Att shows the correct attachment.

Caption	P	IS	Att	Explanation
A man in a suit jacket with a “free word” sign	0.98	-1.59	X	Better IS
A statue of an athletic man on a large stone staircase	-1.16	-1.59	X	Visual context
a girl in a red skirt with some hula hoops	1.00	-1.78	X	Visual context
Woman in black graduation gown with red flower	-2.50	0.17	X	Visual context
a concert with big crowd on a stage	1.47	-0.44	Y	Visual context
A person with a straw hat in the brush	-1.16	-1.45	X	Better P
Woman at kite festival on boardwalk	-5.81	-0.11	X	Visual context
Men on a soccer field with balls	2.20	0.63	Y	Better P, IS
No male construction workers in a work area in a city	-3.18	0.39	Y	Better P
A marathon runner with a dinosaur in the middle	1.08	0.78	Y	Better P, IS

Table 5: Cases not matched by simple patterns, classified incorrectly by the joint model of Plausibility (P) and Information Structure (IS) from Llama (equations 1 and 3, respectively). Att shows the correct attachment.

Caption	P	IS	Att	Explanation
Old ruins in the background of a city street	-0.44	0.74	Y	Near boundary
two white geese on the surface of the ocean	-0.40	1.72	Y	Better IS

Table 6: Cases matched by simple patterns, classified incorrectly by the joint model of Plausibility (P) and Information Structure (IS) from VERA (equations 2 and 4, respectively). Att shows the correct attachment.

Caption	P	IS	Att	Explanation
Woman in black graduation gown with red flower	-0.18	-0.15	X	Visual context
a concert with big crowd on a stage	2.94	0.87	Y	Visual context
A young female in the air with rollerskates	0.43	-0.94	X	Better P
Graffiti of a weasel on a wall	0.13	-0.026	X	Better P
Woman at kite festival on boardwalk	-0.07	-0.15	X	Visual context
Men on a soccer field with balls	0.17	0.04	Y	Better P, IS
No male construction workers in a work area in a city	0.19	0.36	Y	Better IS
A woman with her nose in a book	1.00	-1.07	Y	Ideomatic
Dog with head in snow	0.41	-0.39	Y	Better IS
A marathon runner with a dinosaur in the middle	0.62	0.41	Y	Better P, IS
Brown and white dog with brown and black dog on gray sheet	-0.79	0.24	X	Visual context
Display of cartoon characters in museum	0.077	0.15	Y	Visual context
Two people in Africa with a homemade toy car	0.37	-0.82	X	Better IS

Table 7: Cases not matched by the simple patterns, classified incorrectly by the joint model of Plausibility (P) and Information Structure (IS) from VERA (equations 2 and 4, respectively). Att shows the correct attachment.

References

James Allen, Hannah An, Ritwik Bose, Will de Beaumont, and Choh Man Teng. 2020. A broad-coverage deep semantic lexicon for verbs.

Yonatan Belinkov, Tao Lei, Regina Barzilay, and Amir Globerson. 2014. Exploring compositional architectures and word vector representations for prepositional phrase attachment. *Transactions of the Association for Computational Linguistics*, 2:561–572.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Eric Brill and Philip Resnik. 1994. A rule-based approach to prepositional phrase attachment disambiguation. *arXiv preprint cmp-lg/9410026*.

Michael Collins and James Brooks. 1995. Prepositional phrase attachment through a backed-off model. *arXiv preprint cmp-lg/9506021*.

Pradeep Dasigi, Waleed Ammar, Chris Dyer, and Eduard Hovy. 2017. Ontology-aware token embeddings for prepositional phrase attachment. *arXiv preprint arXiv:1705.02925*.

C. Féry and S. Ishihara. 2016. *The Oxford Handbook of Information Structure*. Oxford Handbooks Series. Oxford University Press.

Henry Watson Fowle and Robert William Burchfield. 2000. *The new Fowler's modern English usage*. Oxford University Press.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Manfred Krifka. 2008. Basic notions of information structure. *Acta Linguistica Hungarica*, 55(3-4):243–276.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. *arXiv preprint arXiv:2305.03695*.

Tom McArthur, Jacqueline Lam-McArthur, and Lise Fontaine. 2018. *Oxford companion to the English language*. Oxford University Press.

Ndapandula Nakashole and Tom Mitchell. 2015. A knowledge-intensive model for prepositional phrase attachment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 365–375.

Adwait Ratnaparkhi, Jeff Reynar, and Salim Roukos. 1994. A maximum entropy model for prepositional phrase attachment. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465.

Stefanie A Tellex, Thomas Fleming Kollar, Steven R Dickerson, Matthew R Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 1507–1514.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. 2023. How well do large language models understand syntax? an evaluation by asking natural language questions.