CAN NEURAL NETWORK LANGUAGE MODELS LEARN SPATIAL PERSPECTIVE FROM TEXT ALONE?

Carolyn Jane Anderson & Tessa Patapoutian
Department of Linguistics
University of Massachusetts, Amherst
Amherst, MA 01060, USA
carolynander@umass.edu

ABSTRACT

Verbs like *come* describe motion relative to a particular spatial perspective. Because spatial perspective is implicit, context-sensitive, and grounded in the physical world, these verbs are particularly challenging for text-based language models. We present a new American English dataset for assessing the ability of language models to learn spatial perspective, using perspectival motion verbs as a test case. We explore the performance of text-trained neural network language models and find that some models are able to infer spatial perspective despite lacking grounded input: BERT predicts the correct perspectival motion verb for 91.3% of automatically scraped and 74.4% of manually annotated examples in our dataset.

1 THE CHALLENGE OF SPATIAL PERSPECTIVE

A growing body of work has explored the kinds of linguistic knowledge that neural network models acquire from text (Linzen et al., 2016; McCoy et al., 2018; van Schijndel et al., 2019; Tenney et al., 2019), focusing for the most part on latent syntactic structure. This paper focuses instead on a challenging kind of latent semantic information: **spatial perspective**. Expressions that encode spatial perspective depend on an individual's relative spatial point-of-view for their meaning. For instance, Thelma's destination in Example 1 depends on whether Sam is using his perspective or Lucy's.

1. *Context: Sam, in Northampton, is talking on the phone to Lucy, in Amherst.*
   Sam: Thelma is coming in 15 minutes.

Spatial perspective is particularly challenging because it is implicit, context-sensitive, and grounded in the physical world. Situational and perceptual information is available to human language users, and has been hypothesized to play an important role in child language acquisition of perspective (Glenberg & Gallese, 2012) and even in adult processing of motion descriptions (Kaschak & Therriault, 2005). By contrast, computational models of language are often trained on text data, which is by nature **ungrounded**: it does not come with information about the real-world situation in which it occurs. If humans require access to such information in order to acquire perspectival aspects of language, then text-based models might be unable to acquire such meanings. On the other hand, human reliance on grounded information might be an artifact of the way human cognition works.

This paper explores the ability of text-based neural network language models to distinguish between the perspectival motion verbs *go* and *come* in context. We explore the performance of several popular pre-trained neural network models on a new dataset for evaluating grounded linguistic terms, composed of a large set of automatically extracted examples and a small set of manually annotated examples. We find that despite lacking access to grounded information, some neural language models are able to successfully predict perspectival motion verbs from textual context alone: BERT performs particularly well, and some, but not all GPT2 models perform above a random baseline.

1.1 THE SEMANTICS OF PERSPECTIVAL MOTION VERBS

The perspectival motion verbs *come* and *go* provide an ideal test case for probing understanding of spatial perspective because their meanings are the same apart from their **perspectival** components:

come requires a perspective-holder to be at the destination of motion (Fillmore, 1966; Barlew, 2017). By contrast, *go* cannot be used if the perspective-holder is at the destination (# indicates infelicity).[1]

2. **Speaker-anchoring**:
   You need to come meet me right away. # You need to go meet me right away.
3. **Listener-anchoring**:
   I will come meet you right away. #I will go meet you right away.
4. **Attitude-holder anchoring**:
   John wants us to come to the cafe where he works.

As shown above, in English[2], there are multiple possible perspective-holders: the speaker, the listener, and subjects of attitude verbs (**attitude-holders**). Therefore, in order to correctly use *come* and *go*, a conversational agent must be able to track and understand multiple spatial perspectives.

2 DATASET

We present a new corpus for assessing the ability of language models to learn grounded linguistic terms. The corpus is composed of two subsets: a set of automatically extracted examples (n=47385) and a manually collected set of annotated examples (n=600). By providing a larger scraped dataset and a smaller manually annotated dataset, we balance two goals: to provide enough data for robust assessment, and to provide annotations of linguistic features for detailed error analysis. The dataset contains instances of 5 verbs: the perspectival motion verbs *come* and *go* and three non-perspectival motion verbs for comparison: *walk*, *arrive*, and *drive*.

2.1 AUTOMATICALLY EXTRACTED CORPUS

Examples were scraped from the Manually Annotated Sub-Corpus of the American National Corpus (MASC) and the Open American National Corpus (OANC) using all lemmas of *come*, *go*, *walk*, *arrive*, and *drive*. A full breakdown is available in Table 4 in the Appendix.

2.2 ANNOTATED CORPUS

Annotated examples were drawn from several publicly available corpora of American English: The Corpus of Contemporary American English (Davies, 2008); the Corpus of Online Registers of English (Davies, 2016); and The Corpus of American Soap Operas (Davies, 2011). These corpora provide a genre-balanced sample, which is important since the availability of the different perspectives varies by genre. Examples were selected to avoid non-perspectival uses of *come* and *go* (like *Come on, man!*). The examples were manually annotated for the following linguistic features:

Perspective-holder

Examples containing *come* were annotated for the perspective-holder: speaker, listener, attitude holder, protagonist, home-base, accompaniment, or other. In addition, we distinguish between examples where the perspective-holder is the speaker at event time, and those where the perspective-holder is the speaker at utterance time (Ex. 5).

5. **Event time:** When I was working at the cafe yesterday, Sue came and bought a coffee.
   Utterance time: John is coming here now.

Subject

The perspective-holder is often impossible to determine in sentences that do not contain perspectival content. Instead, for the other 4 verbs, the subject of the motion verb was recorded.[3]

[1] The exact reason for this infelicity is a subject of debate. See Wilkins & Hill (1995) for more discussion.

[2] There is a rich body of cross-linguistic work on *come* and *go* (Gathercole, 1987; Nakazawa, 2007; 2009). We use American English data for two reasons: (1) there are many widely used pre-trained language models available and (2) the set of licit perspective-holders for motion verbs is comparatively rich in English.

[3] These fields should not be confused: the perspective-holder is very rarely the subject of *come*, since *come* requires the perspective-holder to already be located at the destination of motion (Barlew, 2017).

Figure 1: Example item from dataset, target in bold

Bidirectional presentation: Later, my father knocked at my door. "We're going to Swaziland," he said when I let him in, his voice high with excitement. A weekend site visit, he explained, to a clinic in Mbabane. They didn't want to assume anything, but he and Betsy hoped I would **come** along to look after Ernest." "Your efforts with him haven't been lost on us," he said. "Dad, he's fourteen. He doesn't need me to watch him. Trust me."

Syntactic embedding

Because attitude verbs like *say* introduce their subject as a potential perspective-holder, the perspectival verb prediction task is more challenging in embedded clauses. Examples were annotated for the motion verb's embedding environment: no embedding, speech verb, thought verb, other embedding environment, or quotation. A quota was used to select sufficient numbers of each environment: 25 examples of *come* and 15 of every other verb under each kind of embedding.

Destination of motion

Examples were also annotated for the destination of motion.

Tense

All examples in both the annotated and scraped subsets of the corpus were shallowly annotated for tense/aspect in order to select the correct form of the competitor verbs.

3 TASK

We assess the ability of text-trained language models to correctly predict perspectival motion verbs. For unidirectional models, we take the sentence up until the critical verb and compare the predicted probabilities of each of the 5 verbs in our dataset: *come*, *go*, *walk*, *arrive*, and *drive*. For bidirectional models, we mask the critical verb (Figure 1). We look both at overall accuracy (a 5-way comparison), and accuracy in predicting *come* versus *go* (a 2-way comparison).[4]

3.1 MODELS

We compare the performance of publicly available pre-trained neural network language models: an LSTM (Hochreiter & Schmidhuber, 1997) from Verwimp et al. (2018) and several models from the HuggingFace Transformers library (Wolf et al., 2019): Transformer XL (Dai et al., 2019), BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), GPT (Radford, 2018), and GPT2 (Radford et al., 2018). We also provide a trigram model as a non-neural baseline, trained with Kneser-Ney smoothing on English Wikipedia (Bird et al., 2009; Rescia, 2015).

Using pre-trained models has a few disadvantages. One is that the BERT family of models uses the whole context, while the other models are unidirectional and use only the forward context. Another is that the models differ in complexity and are not trained on the same data. Model complexity and training data are summarized in Tables 5 and 6 in the Appendix.

4 RESULTS

The results (Table 1) suggest that some of the neural language models do acquire information about perspective. BERT does very well on the scraped dataset: BERT Large achieves 91.3% *come*/*go* accuracy, as well as 87.7% overall accuracy. BERT also performs well above chance on the more challenging annotated dataset. The GPT and GPT2 models perform well above chance on the scraped dataset, but struggle on the annotated dataset, suggesting that their performance may be inflated by successful completion of non-perspectival uses of *come* and *go*.

[4]For *drive*, *walk*, and *arrive*, a 2-way comparison is also reported: *drive* and *walk* are compared with each other; *arrive* is compared with *come*.

Table 1: Model performance

Family	Model	Corpus	Accuracy	*come*/*go* accuracy	Other motion accuracy
random guess	-	-	20%	50%	50%
trigram	forward	annotated	12.2%	23.3%	11.9%
		scraped	40.7%	47.3%	38.6%
	backward	annotated	27.5%	53.8%	27.0%
		scraped	40.7%	46.9%	33.6%
RNN	wiki	annotated	29.5%	59.0%	39.2%
		scraped	21.6%	50.3%	48.2%
TransformerXL	base	annotated	26.2%	45.9%	41%
		scraped	63.0%	72.1%	64.5%
BERT	base	annotated	58.0%	**74.4%**	74.7%
		scraped	84.6%	88.8%	85.4%
	large	annotated	**60.0%**	73.4%	**77.5%**
		scraped	**87.7%**	**91.3%**	**88.7%**
RoBERTa	base	annotated	36.7%	62.6%	34.5%
		scraped	66.3%	74.8%	51.0%
DistilBERT	base	annotated	37.7%	53.1%	53.6%
		scraped	68.9%	76.6%	69.8%
GPT	base	annotated	39.5%	63.7%	59.5%
		scraped	67.4%	74.8%	66.7%
GPT2	base	annotated	26.5%	46.7%	54.1%
		scraped	45.8%	70.2%	65.1%
	medium	annotated	34.2%	57.5%	50.3%
		scraped	45.9%	71.8%	67.3%
	large	annotated	33.8%	55.6%	63.3%
		scraped	45.8%	70.2%	65.1%
	extra-large	annotated	37.7%	60.5%	63.9%
		scraped	52.5%	74.3%	72.2%

It is important to note that the BERT and GPT/GPT2 families of models cannot be compared directly, because BERT uses the full sentence context, while the GPT/GPT2 models are unidirectional. It is possible that the sentence postfix is more helpful than the prefix for this task, though a backward trigram model performs no better than a forward trigram model.

While several models do not outperform a random baseline on the more challenging annotated dataset, the high accuracy rate of the BERT models on the perspectival motion verb prediction task suggests that language models can extract some information about spatial perspective despite lacking access to situational and perceptual information. Our results, while preliminary and limited, present a challenge to grounded theories of language acquisition, since they suggest that it is possible to acquire understanding of spatial perspective without access to grounded information.

5 Future work

Our finding that text-based language models can succeed in predicting perspectival expressions opens up several interesting avenues for future work. An immediate goal is to undertake a more extensive error analysis. First, we plan to compare the model performance against human behavioral data. We will conduct crowdsourced behavioral experiments using the annotated corpus and a sample of the most challenging scraped examples. Second, we plan to use the annotated linguistic features to explore whether there are systematic differences between what is challenging for the models and for humans. Third, we hope to explore whether neural network models are learning simpler heuristics for this task, and if so, whether they resemble the heuristics documented in child acquisition of perspectival expressions (Clark & Garnica, 1974; Winston, 1988).

REFERENCES

Jefferson Barlew. *The semantics and pragmatics of perspectival expressions in English and Bulu: The case of deictic motion verbs*. Dissertation, The Ohio State University, 2017.

Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

Eve V. Clark and Olga K. Garnica. Is he coming or going? on the acquisition of deictic verbs. *Journal of Verbal Learning and Verbal Behavior*, 13, 1974.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860, 2019. URL http://arxiv.org/abs/1901.02860.

Mark. Davies. *The Corpus of Contemporary American English (COCA): 600 million words, 1990-present*. Available online at https://www.english-corpora.org/coca, 2008.

Mark. Davies. *Corpus of American Soap Operas: 100 million words*. Available online at https://www.english-corpora.org/soap, 2011.

Mark. Davies. *Corpus of Online Registers of English (CORE)*. Available online at https://www.english-corpora.org/core, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Charles Fillmore. Deictic categories in the semantics of 'come'. *Foundations of Language*, 2, 1966.

Virginia Gathercole. Towards a Universal for Deictic Verbs of Motion. In *Kansas Working Papers in Linguistics*, volume 3, 1987.

Arthur. M. Glenberg and Vittorio Gallese. Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, 48, 2012.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

Michael P. Kaschak and David J. Therriault. Perception of motion affects language processing. *Cognition*, 94, 2005.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Richard McCoy, Robert Frank, and Tal Linzen. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. *ArXiv*, abs/1802.09091, 2018.

Tsuneko Nakazawa. A typology of the ground of deictic motion verbs as path-conflating verbs: the speaker, addressee, and beyond. *Poznan Studies in Contemporary Linguistics*, 43(2), 2007.

Tsuneko Nakazawa. A typology of the ground of deictic motion verbs as path-conflating verbs: the entailment of arrival and the deictic center. *Poznan Studies in Contemporary Linguistics*, 45(3), 2009.

Alec Radford. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Giovanni Rescia, 2015. URL https://github.com/giovannirescia/PLN-2015.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *ACL*, 2019.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. Quantity doesn't buy quality syntax with neural language models. In *EMNLP/IJCNLP*, 2019.

Lyan Verwimp, Hugo Van hamme, and Patrick Wambacq. Tf-lm: Tensorflow-based language modeling toolkit. *Proceedings of LREC*, 2018.

David P. Wilkins and Deborah Hill. When 'go' means 'come': questioning the basicness of basic motion verbs. *Cognitive Linguistics*, 6, 1995.

Millicent Winston. *Deictic verbs: use and acquisition*. Dissertation, University of Connecticut, 1988.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

A Appendix

Figure 2: Sample of examples from annotated dataset, critical word bolded

1. Abruptly he turned and swam toward the diving boards, grabbing the end of the low board with both hands. He bounced twice, then did a chin-up and hoisted himself all the way up, walked down the length of the board and back in the direction we came. Mom was waiting in the car when I walked back outside, towel around my waist. It had stopped raining. "I didn't see you," she said. "I thought maybe you'd walk **home**." "I didn't have an umbrella." "Where were you?" "Nowhere," I said.
2. Miguel: Come on, Charity. Everything's fine. Look, tonight is going to be nothing but fun. Charity: You're right. I'm sorry. Miguel: It's all right. Ivy: Do you like the jasmine, Sam? I wore it especially for you. Do you remember the first night I wore it for you? It was on the beach, and we made a fire. You remember it, don't you? Pilar: I still can't believe that Grace let Sam **come** to this party without her.
3. Here his mind became altogether distracted from classic lore, by the appearance of a very unclassic boy, clad in a suit of brown corduroys and wearing hob-nailed boots a couple of sizes too large for him, who, **coming** suddenly out from a box-tree alley behind the gabled corner of the rectory, shuffled to the extreme verge of the lawn and stopped there, pulling his cap off, and treading on his own toes from left to right, and from right to left in a state of sheepish hesitancy.
4. Alone. The word was a lead balloon. After brunch ended, I was acutely aware of all the couples I passed while **walking** home. They were holding hands, fingers interlaced. My palms dangled at my sides, empty.
5. The storekeeper might say to someone, "Well, here's old Duck. Wonder what he's up to today?" Someone might say, "Duck, you been up on the hill yet?" or "Seen any new birds lately?" But the questions were meant for each other. They didn't wait for Duck to answer. In this way he had learned many things about himself. He learned that his neighbors believed he **went** up on the hill to wait for airplanes to pass or to watch birds. He learned that people still spoke of him as a first-rate fiddler although he hadn't played in years.

Figure 3: Sample of examples from scraped dataset, critical word bolded

1. I always had to get up early, milk the cows, and, uh, run, run them, as we say, because it's a, to the pastures, until times got pretty bad, and one day, I sent my daughter to, to the pasture to bring in the cows. We brought them back in the afternoon, when I saw that, behind her there **came** a big group of, they looked like soldiers, but in street clothes.
2. i can see why children do drop out. yeah i think that's right i i can remember as a child you know nobody ever worried about me wondering out at night and **going** where i wanted to.
3. Armed guards kept round-the-clock watch. Because McLaren's neighbors spotted armed men **arriving** in trucks with Idaho license plates, it is believed that some of the 12 men and women currently residing in the "embassy" may belong to
4. coast and back at the small cove. To get to Cala de Biniparratx, park your car in the small lot and **walk** along a beautiful path through a canyon and shrubs to a narrow cove. The beach is sheltered by steep
5. But there was Mother and Jackie halfway between town and halfway between the country house having to walk all the way, and, of course, they were all scuffed up and their clothes were all dirty and they were reprimanded when they uh **arrived** back home.

Table 2: Summary of manually selected *come* examples

	None	*say*	*believe*	Quote	Other	Total
Speaker@ET	4	4	4	4	4	20
Speaker@UT	4	4	4	4	4	20
Protagonist	4	1	1	2	2	12
Listener	5	3	3	7	3	21
Attitude-holder	0	8	8	0	7	21
Home-base	1	1	1	1	1	5
Accompaniment	1	1	1	1	1	5
Other	6	3	3	6	3	21
Total	25	25	25	25	25	125

Table 3: Summary of manually selected non-*come* examples

	None	*say*	*believe*	Quote	Other	Total
Speaker	4	3	3	4	3	17
Protagonist	3	0	0	1	1	5
Listener	3	3	3	5	2	16
Attitude-holder	0	5	5	0	4	14
3rd-person	3	2	2	3	3	13
Home-base	1	1	1	1	1	5
Accompaniment	1	1	1	1	1	5
Total	15	15	15	15	15	75

Table 4: Summary of scraped corpus by source and genre

Source	Genre	*come*	*go*	*walk*	*arrive*	*drive*	Total
OANC	Spoken	5222	19337	812	23	1182	26576
	Written	6854	8471	1100	770	1681	18876
MASC	Spoken	142	427	12	0	6	587
	Written	432	606	174	50	84	1346
Total		12650	28841	2098	843	2953	47385

Table 5: Model complexity

Family	**Model**	**# layers**	**# attention heads**	**Embedding size**
TransformerXL	base	18	16	1024
BERT	base	12	12	768
	large	24	16	1024
RoBERTa	base	12	12	768
DistilBERT	base	6	12	768
GPT	base	12	12	768
GPT2	base	12	12	768
	medium	24	16	1024
	large	36	20	1280
	extra-large	48	25	1600

Table 6: Model training data

Family	**Training data**	**# tokens (M)**	**Vocabulary**
TransformerXL	WikiText-103	100	26735
BERT	English Wikipedia, BooksCorpus	3300	30522
RoBERTa	English Wikipedia, BooksCorpus CC-News, Open Web Text, Stories	> 3300	50266
DistilBERT	English Wikipedia, BooksCorpus	3300	30522
GPT	BooksCorpus	800	40478
GPT2	WebText	unknown	50257