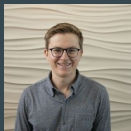


Language Models Have Implicit Entailment Semantics

...



Will Merrill



Alex Warstadt



Tal Linzen

March 30th @ MIT

Background:

Semantics in Language Models

Supervised Learning vs. Pretrained Language Models (LMs)

Supervised training

Boston and NYC are cities Boston is a city	Entails	x10,000
---	---------	---------

(Pre)training a LM

Founded in response to the increasing industrialization of the United States, MIT adopted [MASK] as a polytechnic university model and stressed laboratory instruction in applied [MASK] engineering. The institute has an urban campus...	x1,000,000,000
--	----------------

Extracting Semantic Relations from LMs

Task	Prompt Template	Prompt found by AUTOPROMPT	Label Tokens
Sentiment Analysis	{sentence} [T]... [T] [P].	unflinchingly bleak and desperate Writing academicswhere overseas will appear [MASK].	pos: partnership, extraordinary, ##bla neg: worse, persisted, unconstitutional
NLI	{prem}[P][T]... [T]{hyp}	Two dogs are wrestling and hugging [MASK] concretepathic workplace There is no dog wrestling and hugging	con: Nobody, nobody, nor ent: ##found, ##ways, Agency neu: ##ponents, ##lary, ##uated
Fact Retrieval	<i>X plays Y music</i> {sub}[T]... [T][P].	Hall Overton fireplacemade antique son alto [MASK].	
Relation Extraction	<i>X is a Y by profession</i> {sent}{sub}[T]... [T][P].	Leonard Wood (born February 4, 1942) is a former Canadian politician. Leonard Wood gymnasium brotherdicative himself another [MASK].	

Distributional Semantics

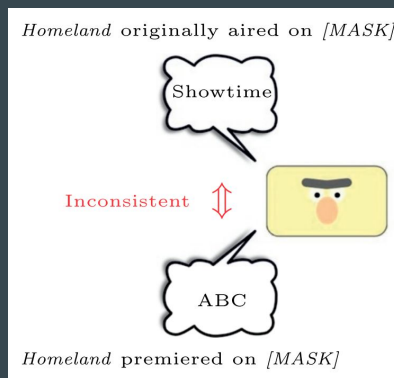
- Hypothesis (Firth, 1951) that:
 - Pattern between a word and its context encodes its meaning
 - Therefore learning cooccurrence patterns \Rightarrow semantic representations
- But why?
 - This talk can be viewed as a theoretical explanation supporting the distributional hypothesis

Issues with Claiming LMs Encode Semantics

1. Facts generated by LMs are often not **truthful** (Lin et al., 2021)

GPT-3: The Earth is flat

2. LMs generate **inconsistent output** given minor paraphrases of the same input (Elazar et al., 2021)



Is it Impossible for LMs to Learn Semantics *Fully*?

In other words, are there limits on learning semantics distributionally

“Scaling up any LM-like model ... may eventually run into (or could already be running into) the limits of the pretraining objective.”

GPT-3 paper (Brown et al., 2020)

“We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning.”

Bender and Koller (2020)

Goal

- **Larger Research Program:**

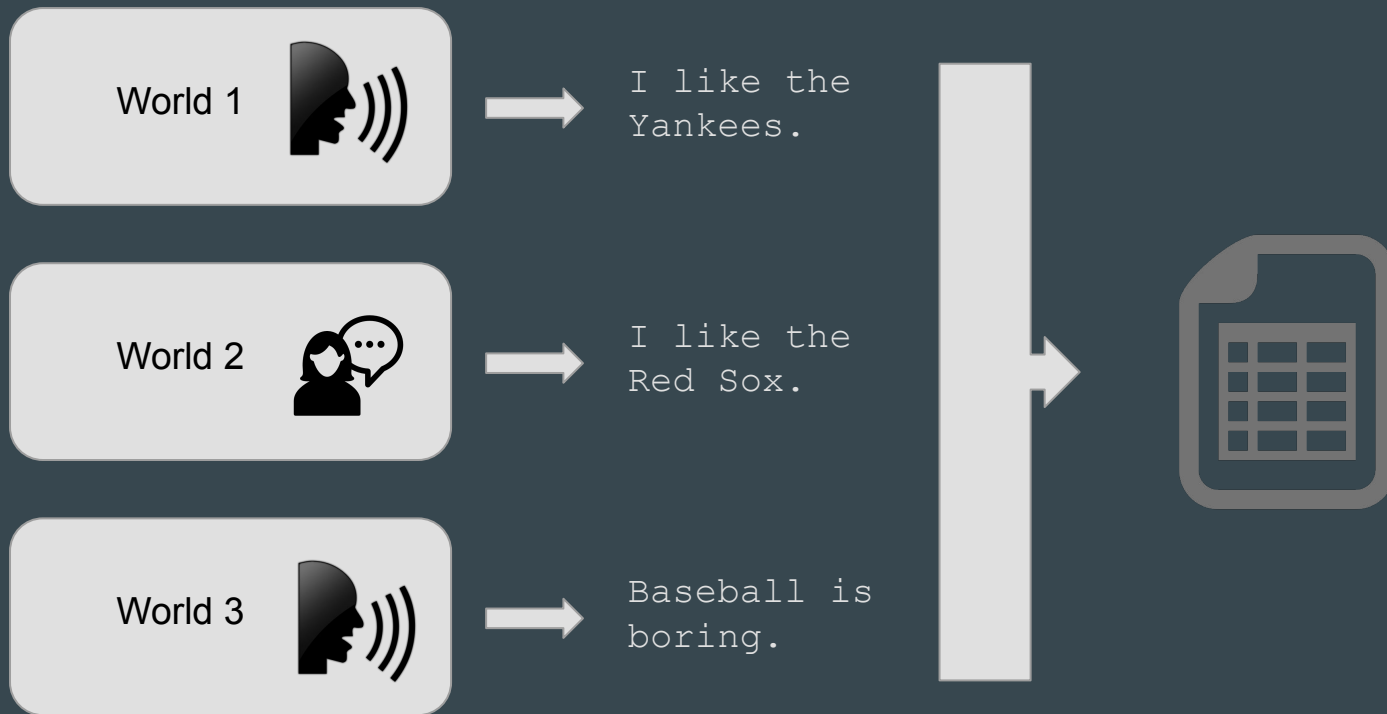
Principled theoretical understanding of how semantics can be learned distributionally

- **This Talk:**

Classifying entailment is reducible to language modeling on **natural data**

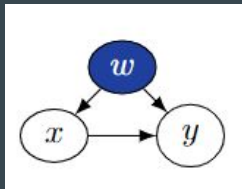
Understanding LM Training Data

Data Comes from Humans in Different “World States”



Formalizing this Setup

- Document = pair of sentences from same speaker/world



I live in Boston. I like
the Red Sox.

- Different documents come from different speakers/worlds

$$p(xy) = \sum_w p(xy, w) = \sum_w p(x \mid w)p(y \mid x, w)p(w).$$

- Conditional distribution $p(x_t \mid x_{<t}, w)$ models a “speaker”

How Do Speakers Produce Sentences?

- Speakers don't produce sentences arbitrarily; they respect Gricean maxims:

- **Truthfulness**

$$p(\text{I am in New York} | w) > p(\text{I am in Boston} | w)$$

- **Brevity**

$$p(\text{Rain!} | w) > p(\text{Look, a rain drop. Look a rain drop...} | w)$$

- **Informativeness**

$$p(\text{Boston has 5 trains} | w) > p(\text{Boston is Boston} | w)$$

Training Data Has Semantic Signal

- Training distribution encodes **truthfulness**
- Training distribution encodes **informativeness**
- Therefore learning p requires learning *something* about truthfulness and informativeness

Training Data Summary

1. LM target distribution is “superposition” of different worlds:

$$p(xy) = \sum_w p(xy, w) = \sum_w p(x | w)p(y | x, w)p(w).$$

2. Gricean maxims suggest natural text will have semantic signals about truth and informativeness

Defining Entailment

The Meanings of Sentences

- The meaning of a sentence is the set of worlds where it is true

$[[\text{John likes the Yankees}]] = \{w : \text{John likes the Yankees in } w\}$

Entailment

Sentence x entails sentence y iff:

- y is true in every world where x is

For sentences $x, y \in \mathcal{X}$, x entails y iff $\llbracket x \rrbracket \subseteq \llbracket y \rrbracket$.

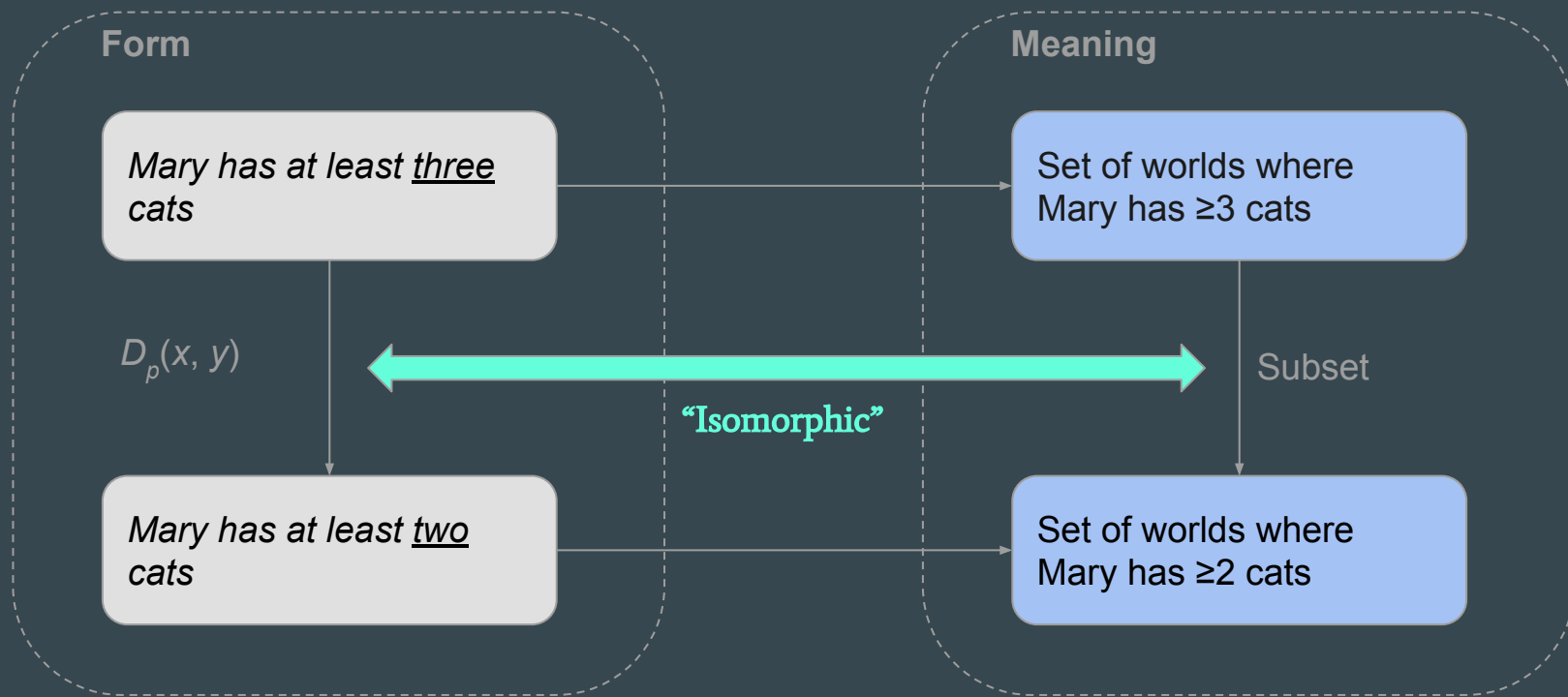
Distributional Relations

- A relation over pairs of sentences defined in terms of document probabilities
- **Example:**

$$D_p(x, y) \Leftrightarrow p(x) = p(y)$$

“x and y are equally likely in the training data”

How Can a Distribution “Encode” Entailment?



Research Question

For some speaker $p(x \mid w)$, (with $p(x)$ being marginal variant of $p(x \mid w)$)

is there a distributional relation $D_p(x, y)$ for sentences x, y

s.t., for all x, y , $D_p(x, y) \Leftrightarrow x$ entails y ?

Consequences of Form/Meaning Isomorphism

1. Natural distributions of text encode entailment semantics
2. Learning entailment **reducible** to language modeling
 - a. An LM that fully masters its training objective would fully understand entailments

Uniformly Truthful Speaker

Uniform Truthful Speaker

- Overly simple model of NL speakers:
 - a. Never say false sentences
 - b. All true sentences equally likely

Example:

World w where Mary has 2 cats. Mary chooses one of two sentences:

$p(\textit{I have at least one cat} \mid w)$	=	1/2
$p(\textit{I have at least two cats} \mid w)$	=	1/2
$p(\textit{I have at least three cats} \mid w)$	=	0

- Formalizes Gricean maxim of truthfulness, but not informativeness or brevity

Isomorphism for Entailment

Theorem 1 *If p is a uniformly truthful speaker, then entailment is isomorphic to $D_p(x, y) \iff p(xy) = p(xx)$, i.e., for all sentences $x, y \in \mathcal{X}$,*

$$\llbracket x \rrbracket \subseteq \llbracket y \rrbracket \iff p(xy) = p(xx).$$

- Semantic signal from a truthful speaker would allow a strong LM to encode entailment relations
- Very crude model of human speakers; we will move to fully Gricean speakers

Informative Speaker

Informative Speaker

- Speaker who attempts to convey information to their listener
- Unlikely to be untruthful or redundant

$x = I \text{ have } \underline{two} \text{ cats}$

$p(I \text{ have } \underline{no} \text{ cats} \mid x, w)$ = low

$p(I \text{ have } \underline{a} \text{ cat} \mid x, w)$ = low

$p(I \text{ also have } \underline{a \ dog} \mid x, w)$ = higher

Formalizing Information

- **Information of text:** Listener's reduction in uncertainty about the speaker's belief world after hearing the text

Definition 6 The information content of a text $z \in \mathcal{X}^*$ to a listener $\ell(w \mid z)$ is

$$I_\ell(z; w) = -\log \ell(w) + \log \ell(w \mid z).$$

Formalizing Informativeness

- An *informative* speaker tries to convey information to some imagined listener
 - Imagined listener must “absorb” all the information

Definition 8 A speaker p is *informative* if there exists a listener $\ell(w \mid z)$, an invertible function $f : \mathbb{R} \rightarrow \mathbb{R}$, and a cost function $c : \mathcal{X} \rightarrow \mathbb{R}$ such that, for all sentences $z \in \mathcal{X}^*$:

$$p(z \mid w) \propto \exp (f(I_{\ell}(z; w) - c(y)) .$$

Further, ℓ must satisfy the following for all texts x, y :

$$\forall w [I_{\ell}(y \mid x; w) = 0] \iff \llbracket x \rrbracket \subseteq \llbracket y \rrbracket .$$

- Pretty decent model of human speakers!
 - Matches linguistic theories like Rational Speech Acts (Goodman et al., 2016)

Isomorphism for Entailment

Theorem 3 *Under any informative speaker p , entailment is isomorphic to a distributional relation. Specifically, for all sentences $x, y \in \mathcal{X}$,*

$$\llbracket x \rrbracket \subseteq \llbracket y \rrbracket \iff \frac{p(y \mid x)}{p(\epsilon \mid x)} - \frac{p(y \mid y)}{p(\epsilon \mid y)}.$$

- **Intuition:** look at the numerators
- **Potential limitation:** human speakers are mildly redundant in certain cases
 - E.g., teaching a class, giving a talk
 - Equivalent to assuming listener doesn't fully absorb information

Learnability of Entailment

Is Entailment *Efficiently* Learnable By an LM?

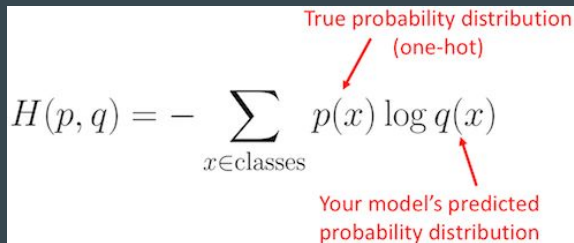
- Perfectly learning language modeling implies perfectly learning entailment

$$\llbracket x \rrbracket \subseteq \llbracket y \rrbracket \iff \frac{p(y \mid x)}{p(\epsilon \mid x)} = \frac{p(y \mid y)}{p(\epsilon \mid y)}$$

- Do current LMs learn to estimate $p(x)$ well enough for this equation to hold?

Background: LM Training Objective

- Minimize perplexity/cross-entropy:

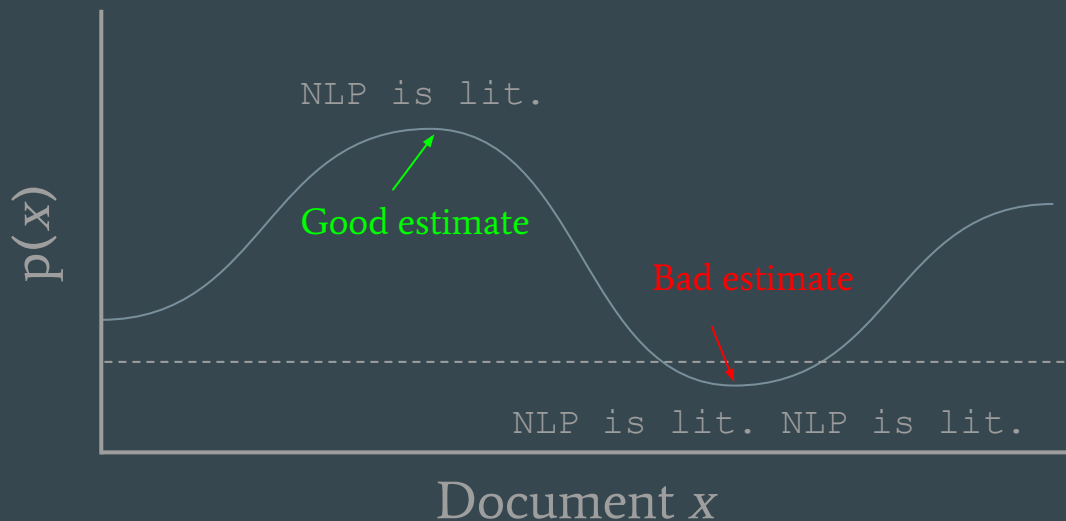
$$H(p, q) = - \sum_{x \in \text{classes}} p(x) \log q(x)$$


The diagram shows the cross-entropy formula $H(p, q) = - \sum_{x \in \text{classes}} p(x) \log q(x)$. Two red arrows point to the terms in the formula: one points from the text "True probability distribution (one-hot)" to $p(x)$, and another points from the text "Your model's predicted probability distribution" to $q(x)$.

- $p(x)$ is good for “the average” document x
 - But not necessarily for all documents

Low Perplexity \nRightarrow Learned Entailment

Because we need good estimate of unlikely document probabilities



Summary

- If an LM is accurate for any x , we can use it to predict entailment via isomorphism results
- But LM with low perplexity won't necessarily be accurate for any x
- Might be possible to regularize the training objective to do better!

Conclusion

Contributions

1. Show sense in which **natural text distributions encode entailment semantics**
2. Explains distributional hypothesis in terms of pragmatics
3. Points to methods to extract entailment predictions from LMs



FIN

Theorem 1 *If p is a uniformly truthful speaker, then entailment is isomorphic to $D_p(x, y) \iff p(xy) = p(xx)$, i.e., for all sentences $x, y \in \mathcal{X}$,*

$$\llbracket x \rrbracket \subseteq \llbracket y \rrbracket \iff p(xy) = p(xx).$$

Proof. The claimed distributional relation can be expanded as

$$\begin{aligned} D_p(x, y) &\iff p(xy) = p(xx) \\ &\iff \sum_w \llbracket x \rrbracket(w) \llbracket y \rrbracket(w) \frac{p(w)}{n(w)^2} = \sum_w \llbracket x \rrbracket(w) \llbracket x \rrbracket(w) \frac{p(w)}{n(w)^2} \\ &\iff \sum_w \llbracket x \rrbracket(w) \llbracket y \rrbracket(w) \frac{p(w)}{n(w)^2} = \sum_w \llbracket x \rrbracket(w) \frac{p(w)}{n(w)^2} \\ &\iff \forall w (\llbracket x \rrbracket(w) \llbracket y \rrbracket(w) = \llbracket x \rrbracket(w)), \end{aligned}$$

where the last step follows by Lemma 1. We can thus conclude that $D_p(x, y) = 1$ if and only if $\llbracket x \rrbracket \subseteq \llbracket y \rrbracket$. \square

Lemma 1 *Let $\mathbb{1}_S$ be the indicator function for set S . For finite sets \mathcal{A}, \mathcal{B} such that $\mathcal{A} \subseteq \mathcal{B} \subseteq \{w \mid 1 \leq w \leq n\}$ and $c \in \mathbb{R}_+^n$, $\mathcal{A} = \mathcal{B}$ if and only if*

$$\sum_{w=1}^n \mathbb{1}_{\mathcal{A}}(w)c_w = \sum_{w=1}^n \mathbb{1}_{\mathcal{B}}(w)c_w.$$

Proof. We will prove that $\mathcal{B} \subseteq \mathcal{A}$ by contradiction. Assume there exists $w \in \mathcal{B}$ such that $w \notin \mathcal{A}$. Then the right sum contains the positive term c_w , while the left sum does not. Because all terms in the right sum are positive, the left sum must contain at least one term c_u that the right sum does not. Thus, $c_u \in \mathcal{A}$ but $c_u \notin \mathcal{B}$. But this has violated our assumption that $\mathcal{A} \subseteq \mathcal{B}$. \square

Lemma 2 Under any informative speaker p , for all $x, y \in \mathcal{X}$,

$$\frac{p(\epsilon \mid x)}{p(x \mid x)} = \exp(c(x)).$$

Proof. By construction of an informative speaker, there exist “normalizing constants” given by a function $g(x, w)$ such that, for any $x, y \in \mathcal{X}$,

$$\begin{aligned} p(y \mid x) &= \sum_w p(y \mid x, w) p(w) \\ &= \exp(-c(y)) \sum_w \exp(f(I_\ell(y \mid x; w))) g(x, w) p(w). \end{aligned}$$

We will apply this identity to both sides of the fraction.

$$\begin{aligned} \frac{p(\epsilon \mid x)}{p(x \mid x)} &= \frac{\exp(-c(\epsilon)) \sum_w \exp(f(I_\ell(\epsilon \mid x; w))) g(x, w) p(w)}{\exp(-c(x)) \sum_w \exp(f(I_\ell(x \mid x; w))) g(x, w) p(w)} \\ &= \exp(c(x)) \frac{\sum_w \exp(f(I_\ell(\epsilon \mid x; w))) g(x, w) p(w)}{\sum_w \exp(f(I_\ell(x \mid x; w))) g(x, w) p(w)}. \end{aligned}$$

Since $\llbracket x \rrbracket \subseteq \llbracket \epsilon \rrbracket$ and $\llbracket x \rrbracket \subseteq \llbracket x \rrbracket$, we know that the conditional information of both ϵ and x given x is 0, and, thus,

$$\frac{p(\epsilon \mid x)}{p(x \mid x)} = \exp(c(x)) \frac{\sum_w \exp(f(0)) g(x, w) p(w)}{\sum_w \exp(f(0)) g(x, w) p(w)} = \exp(c(x)).$$

□

Theorem 3 Under any informative speaker p , entailment is isomorphic to a distributional relation. Specifically, for all sentences $x, y \in \mathcal{X}$,

$$\llbracket x \rrbracket \subseteq \llbracket y \rrbracket \iff \frac{p(y \mid x)}{p(\epsilon \mid x)} = \frac{p(y \mid y)}{p(\epsilon \mid y)}.$$

Proof. Recall from the proof of Lemma 2 that there exists a function $g(x, w)$ such that

$$p(y \mid x) \propto f(y)^{-1} \sum_w f(I_\ell(y \mid x; w)) g(x, w) p(w).$$

Thus, the proposed distributional relation can be expanded as

$$\begin{aligned} D_p(x, y) &\iff \frac{p(y \mid x)}{p(\epsilon \mid x)} = \frac{p(y \mid y)}{p(\epsilon \mid y)} \\ &\iff p(y \mid x) \cdot \frac{p(\epsilon \mid y)}{p(y \mid y)} = p(\epsilon \mid x) \cdot \frac{p(x \mid x)}{p(x \mid x)} \\ &\iff p(y \mid x) f(y) = p(x \mid x) f(x) \\ &\iff \sum_w f(I_\ell(y \mid x; w)) g(x, w) p(w) = \sum_w f(I_\ell(x \mid x; w)) g(x, w) p(w). \end{aligned}$$

By Lemma 1, this holds if and only if, for all w ,

$$\begin{aligned} f(I_\ell(y \mid x; w)) &= f(I_\ell(x \mid x; w)) \\ I_\ell(y \mid x; w) &= I_\ell(x \mid x; w) \\ I_\ell(y \mid x; w) &= 0. \end{aligned}$$

By Definition 8, this is equivalent to saying that $\llbracket x \rrbracket \subseteq \llbracket y \rrbracket$.

□

Speculative Way to Improve Distributional Semantic Representations

- LM that is accurate for all x can be understood to encode entailment
- \therefore Improving performance on *unlikely* sentence should lead to LMs with more semantic knowledge
 - Better estimate of $p(\text{My name is Will.} \mid \text{My name is Will.})$
- Inductive bias towards compositionality

Future Work

- Compositional/pragmatic LMs
- Testing theory in simulations/corpus data
- Extending theory to more complex speakers/relations