UNIVERSITY OF GOTHENBURG

Sharid Loáiciga

Bringing Together Anaphora Resolution and Linguistic Theory

CLASP centre for linguistic theory and studies in probability

Swedish Research Council

Part 0: Introduction

Introduction

- Work in Anaphora Resolution has distanced itself from linguistic theory in recent years.
- Syntax, for example, has a long tradition in linguistics with aims grounded in cognition.
- Discourse community, on the other hand, is currently very task-oriented, there isn't a clear goal grounded in cognition.
- But that's somewhat ironic because discourse theories such as DRT or Centering have as their main purpose to model the hearer's representation structure.
- The semantics of anaphora is at the very center of it.

- There are still algebraic systems that define their own sets of constrain verification but their success has been limited in comparison with statistical systems.

- We’ve gotten very good at solving the task of anaphora resolution:

SOTA through the years — OntoNotes

(Weischedel et al. 2013)

	MUC			B^3			$CEAF_{\phi 4}$			Average
	P	R	F1	P	R	F1	P	R	F1	
Kirstain, Ram & Levy (2021) *s2e*	86.5	85.1	85.8	80.3	77.9	79.1	76.8	75.4	76.1	80.3
Joshi et al. (2020) *SpanBERT*	85.8	84.8	85.3	78.3	77.9	78.1	76.4	74.2	75.3	79.6
Lee et al. (2017) *e2e*	81.2	73.6	77.2	72.3	61.7	66.6	65.2	60.2	62.6	68.8
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

SOTA through the years — OntoNotes

	MUC			B^3			$CEAF_{\phi 4}$			Average
	P	R	F1	P	R	F1	P	R	F1	
Kirstain, Ram & Levy (2021) *s2e*	86.5	85.1	85.8	80.3	77.9	79.1	76.8	75.4	76.1	80.3
Joshi et al. (2020) *SpanBERT*	85.8	84.8	85.3	78.3	77.9	78.1	76.4	74.2	75.3	79.6
Lee et al. (2017) *e2e*	81.2	73.6	77.2	72.3	61.7	66.6	65.2	60.2	62.6	68.8
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

SOTA through the years — OntoNotes

	MUC			B^3			CEAF$_{\phi4}$			Average
	P	R	F1	P	R	F1	P	R	F1	
Kirstain, Ram & Levy (2021) *s2e*	86.5	85.1	85.8	80.3	77.9	79.1	76.8	75.4	76.1	80.3
Joshi et al. (2020) *SpanBERT*	85.8	84.8	85.3	78.3	77.9	78.1	76.4	74.2	75.3	79.6
Lee et al. (2011)	74.0	90.1	81.3	70.2	89.3	78.6	79.7	53.1	63.7	74.5
Lee et al. (2017) *e2e*	81.2	73.6	77.2	72.3	61.7	66.6	65.2	60.2	62.6	68.8
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Lee et al. (2011)	66.9	63.9	65.4	70.1	71.5	70.8	46.3	49.6	47.9	61.4
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

- We’ve actually gotten very good at solving the task of anaphora resolution
    - at least for **English** OntoNotes.
- More comprehensive learning depends mainly on available corpora.
- We have only started to exploit pre-trained models for discourse phenomena.

- There’s a lot of research confirming that pre-trained language models (LMs) encode syntactic knowledge to different degrees, so we have every reason to believe that they encode discourse knowledge as well.

- The moment is right to look into linguistics theory again and maybe update our discourse theories.

- The annotation of more (diverse) data is crucial

    - Here linguistic theory may come in handy as well.

This talk:

- Information status
- Psycholinguistics
- Multimodal annotation

Part 1:
Information Status

Information Status

Researchers at Plant Genetic Systems N.V. in Belgium said **they**

new indefinite **old pronoun**

have developed a genetic engineering technique for creating hybrid

plants. **The researchers** said **they** have isolated a plant gene that

old definite **old pronoun**

prevents the production of pollen.

Are LMs sensitive to different referring expressions?

Is Incoherence Surprising? Targeted Evaluation of Coherence Prediction from Language Models
Beyer, Loáiciga & Schlangen (2021)

(ARRAU corpus)

condition: pronoun

region 1: And there's a ladder coming out of the tree
and there's **a man** at the top of the ladder

region 2: you can't see **him** yet

vs

condition: repetition

region 1: And there's a ladder coming out of the tree
and there's **a man** at the top of the ladder

region 2: you can't see **the man** yet

Are LMs sensitive to different referring expressions?

- We used surprisal (the mean surprisal for the complete region) to measure if pre-trained LMs prefer the expected condition over the manipulated one.

- ARRAU corpus data

3 genres (ARRAU corpus)

	WSJ	VPC	Dialogue	Fiction
GPT-2	0.53	0.56	0.47	0.42
DIALOGPT	0.44	0.51	0.47	0.36
#items	512	75	68	98

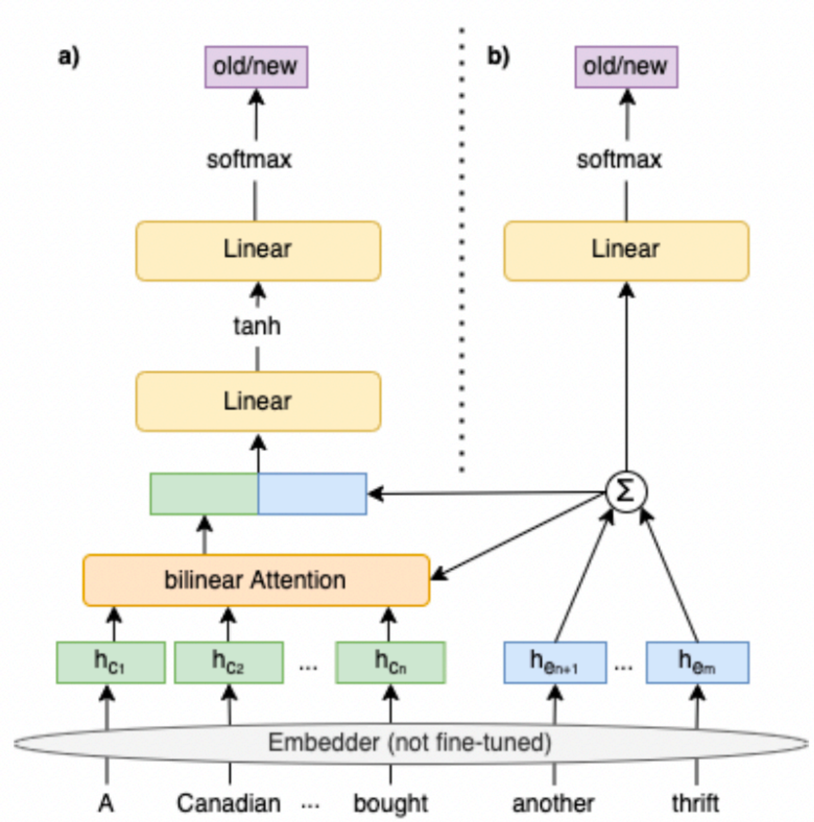
2 LMs

Accuracy scores: how many times the LMs preferred the expected condition.

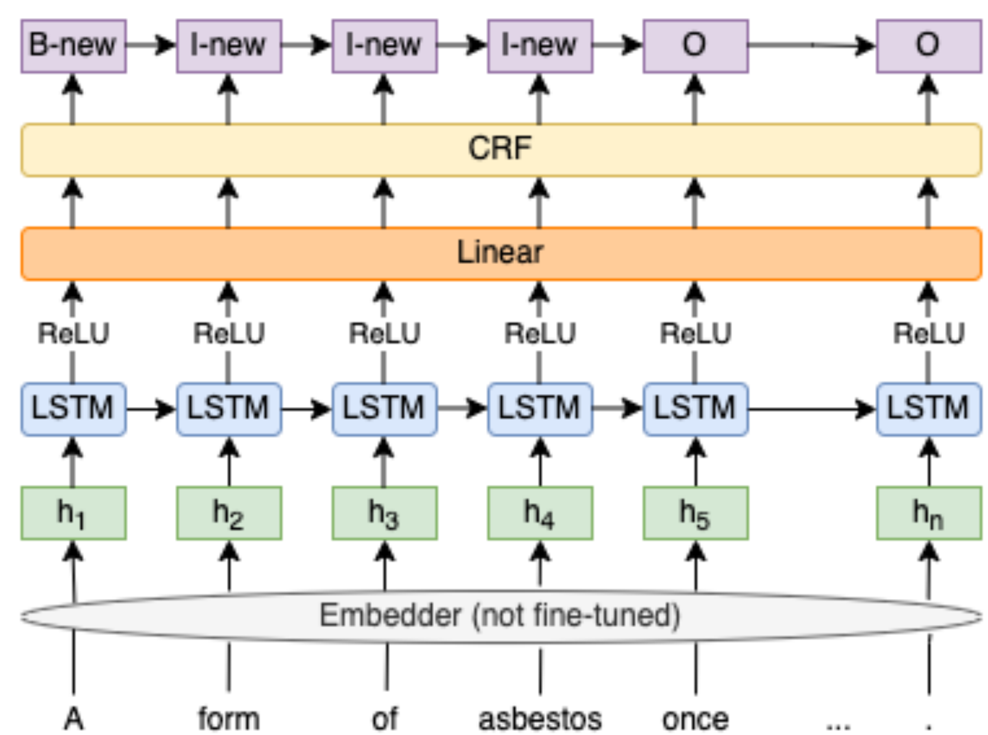
Do LMs encode entity knowledge?

New or Old? Exploring How Pre-Trained Language Models Represent Discourse Entities
Loáiciga, Beyer & Schlangen (2021)

- Idea of taking a step back:
    - Build a probe able to predict discourse status of entities as new or old.
- Probe 1: Binary classification



Probe 2: Sequence labeling



Classification Probe

Data — from ARRAU (Uryupina et al., 2020)

Original

[The researchers]$_t$ said **[they]$_t$** have isolated **[a plant gene that prevents [the production of pollen]$_i$]$_j$]$_m$. [The gene]$_m$** thus can prevent **[a plant]$_y$** from fertilizing **[itself]$_y$**.

Spans

The researchers said they have isolated **[a plant gene that prevents the production of pollen]**. ⟶ new

The researchers said they have isolated a plant gene that prevents the production of pollen. **[The gene]** ⟶ old

Heads

The researchers said they have isolated a plant gene that prevents the production of **[pollen]**. ⟶ new

The researchers said they have isolated a plant gene that prevents the production of pollen. The **[gene]** ⟶ old

Classification Probe

Results — averaged over 5 runs

	Heads							Spans						
	Discourse New			Discourse Old			Acc.	Discourse New			Discourse Old			Acc.
	Prec.	Rec.	F1	Prec.	Rec.	F1		Prec.	Rec.	F1	Prec.	Rec.	F1	
Probing Transformer-XL														
Attention-based	0.86	0.92	**0.89**	**0.88**	0.80	**0.84**	**0.87**	**0.88**	0.91	**0.89**	**0.86**	**0.81**	**0.83**	**0.87**
Entity-based	**0.87**	0.91	**0.89**	0.87	**0.81**	**0.84**	**0.87**	0.85	**0.92**	0.88	0.86	0.76	0.80	0.85
Baselines fastText 300														
Attention-based	0.76	0.86	0.81	0.76	0.62	0.68	0.76	0.82	0.89	0.85	0.81	0.71	0.75	0.82
Entity-based	0.70	**0.93**	0.80	0.82	0.46	0.59	0.73	0.76	**0.92**	0.83	0.82	0.56	0.67	0.78
Baselines w/o embeddings														
POS-based	0.66	0.83	0.73	0.63	0.40	0.49	0.65	0.74	0.80	0.77	0.66	0.57	0.61	0.71
Majority class	0.58	1.00	0.73	0.00	0.00	0.00	0.58	0.60	1.00	0.75	0.00	0.00	0.00	0.60

1. Context doesn't make a difference.
2. New and Old are equally easy/difficult.
3. Spans and Heads are equally easy/difficult.

Sequence Labeling Probe

Data — same gold labels as before, IOB format

Spans

The researchers said they have isolated a plant gene
B-old I-old O **B-old** O O **B-new I-new I-new**

that prevents the production of pollen . The gene thus
I-new I-new I-new I-new I-new I-new . B-old I-old O

can prevent a plant from fertilizing itself .
O O **B-new I-new** O O **B-old .**

Heads

The researchers said they have isolated a plant gene
O **B-old** **O** **B-old** O O O O **B-new**

that prevents the production of pollen . The gene thus
O O O **B-new** O **B-new .** O **B-old** O

can prevent a plant from fertilizing itself .
O O O **B-new** O O **B-old** .

Sequence Labeling Probe

Results —averaged over 5 runs

	Heads							Spans						
	Discourse New			Discourse Old			Avg.F1	Discourse New			Discourse Old			Avg.F1
	Prec.	Rec.	F1	Prec.	Rec.	F1		Prec.	Rec.	F1	Prec.	Rec.	F1	
Transformer-XL														
LSTM + Linear + CRF	**0.75**	0.79	**0.77**	**0.80**	**0.78**	**0.79**	**0.78**	0.59	0.59	0.59	**0.80**	**0.72**	**0.75**	0.66
Linear + CRF	0.70	0.70	0.70	0.75	0.69	0.72	0.71	0.43	0.38	0.41	0.69	0.63	0.66	0.51
Baselines fastText 300														
LSTM + Linear + CRF	0.67	0.76	0.71	0.75	0.63	0.68	0.70	0.50	0.50	0.50	0.76	0.60	0.67	0.57
Linear + CRF	0.55	0.63	0.59	0.69	0.45	0.55	0.57	0.25	0.19	0.22	0.63	0.41	0.50	0.33
Baselines														
Simple CRF	0.57	0.70	0.62	0.71	0.45	0.55	0.59	0.32	0.28	0.29	0.64	0.44	0.52	0.38
POS baseline	0.65	0.51	0.57	0.51	0.58	0.55	0.56	**0.77**	0.61	0.68	0.62	0.71	0.66	**0.67**
Majority class	0.50	**1.00**	0.74	0.00	0.00	0.00	0.43	0.60	**1.00**	**0.75**	0.00	0.00	0.00	0.45

1. The LSTM is able to contextualize the representations further.
2. New is harder than Old.
3. Spans are harder than Heads.

Maybe this task is too easy, or maybe it's not a discourse task at all.

Error Analysis

- Many errors concern *it*, *this*, *that* and *which* — known to be problematic.
- The most common error is predicting a mention when there isn't one (gold is O).
- Most of the errors with spans are about identifying the boundaries of the entity.

[an environmental cleanup]
gold: B-new I-new I-new
predicted: O O B-new

Part 2:
(Large Scale)
Psycholinguistics

Event vs entity

Event and entity coreference across five languages: Effects of context and referring expression
Bevacqua, Loáiciga, Hardmeier & Rohde (2021)

The snow that was covering the fields was melting down.

Event vs entity

<u>The snow</u> that was covering the fields <u>was melting down</u>.

It was a welcome sight, after the harsh winter. *Event*

It had turned into slush and mud. *Entity*

This was as dependable as the sun rising each morning. *Event*

This was always on time. *Entity*

Story Continuation Task

The snow that was covering the fields was melting down. It ____

The snow that was covering the fields was melting down. This____

EN, FR, DE, IT & ES

The colonial building was collapsing slowly. It/This ...

Le bâtiment colonial a croulé sous la neige. Il/Cela/C’est ...

Das prachtvolle Gebäude zerfiel über die Jahre. Es/Das/Dies ...

Il palazzo coloniale è collassato improvvisamente. Questo/Ciò/null ...

El edificio colonial implosionó lentamente. Esto/Este/null …

Experiments

- Human monolingual speakers recruited from Amazon's Mechanical Turk.
- 50 participants per language, 24 experimental items.
- Collected continuations are annotated as referring to

the event or entity.

- Sentences controlled for **verb type** and aspect.
- Modeled using mixed effects logistic regression.

Verb Alternation

Hannah **popped** the balloon.

participant 1 participant 2

agent patient/theme

The balloon **popped**.

participant 1

patient/theme

Verb Alternation

The train from the Highlands **arrived** promptly.

participant 1

agent

* Hannah arrived the train from the Highlands.

What we found

- **It** yields more entity readings and **This** more event readings, but the distinction is not categorical.
- Alternating verbs with more participants trigger more event readings.
- Aspect doesn't make a difference.

Finding Alternating Verbs

Unsupervised Discovery of Unaccusative and Unergative Verbs
Loáiciga, Bevacqua & Hardmeier (2021)

1. Vocabulary V of **Glove embeddings**, list of **subjects** S and **objects** O.
2. Disjoint sets of **seed words** are created

$$S' = V \cap S \setminus O \text{ and } O' = V \cap O \setminus S$$

3. We **expand sets S+ and O+** from S′ and O′ respectively:

   (a) We draw 20 samples of 10 items from the seed words.

   (b) For each sample, we find the 50 nearest neighbors in the embedding space. The union of these 20 sets of nearest neighbors forms the **expansion candidates**.

   (c) **Disjoint sets S+ and O+** are created by taking the 30 highest-scoring expansion candidates generated from S′ and O′ respectively.

Finding English Alternating Verbs

We test GPT-2 using probing sentences with the pattern:

<s> The NOUN VERBs . </s>

The train from the Highlands arrived.

The balloon popped.

Finding English Alternating Verbs

Work in progress… sort of.

	#	Alt	Non-alt
Constructed	*20*		
Expanded EP		0.78	0.71
Expanded Leff		0.78	0.71
FAVA (Kann et al. 2019)	120		
Expanded EP		0.45	0.62
Expanded Leff		0.42	0.65
FrameNet (Baker et al. 1998)	329		
Expanded EP		0.24	0.22
Expanded Leff		0.16	0.20

Why to do this?

- Event anaphora is an understudied area.
- Humans and other languages might offer alternative clues about anaphora.
- Expectation-driven models of processing discourse, for example, have been shown to be relevant for anaphora resolution in a QUD context (e.g., verb in question determines coreference pattern of response, Kehler & Rohde 2016).

Part 3:
Multimodal Annotation

From Text to Image

John gave Mary **five dollars**. **It** was more than he gave Sue.

John gave Mary **five dollars**. **One** of **them** turned out to be counterfeit.

Example from Bonnie Webber

From Text to Image

- Deep neural networks and pre-trained LMs have figured out a lot of the textual semantics or "meaning" that traditional discourse theories sought to solve (cf Piantadosi & Hill, 2022, Meaning without reference in large language models).
- However, to start capturing reference, vision & language data is a good starting point.

Data from Tell-me-more

5,701 image-document pairs, Ilinykh, Zarrieß & Schlangen, (2019)

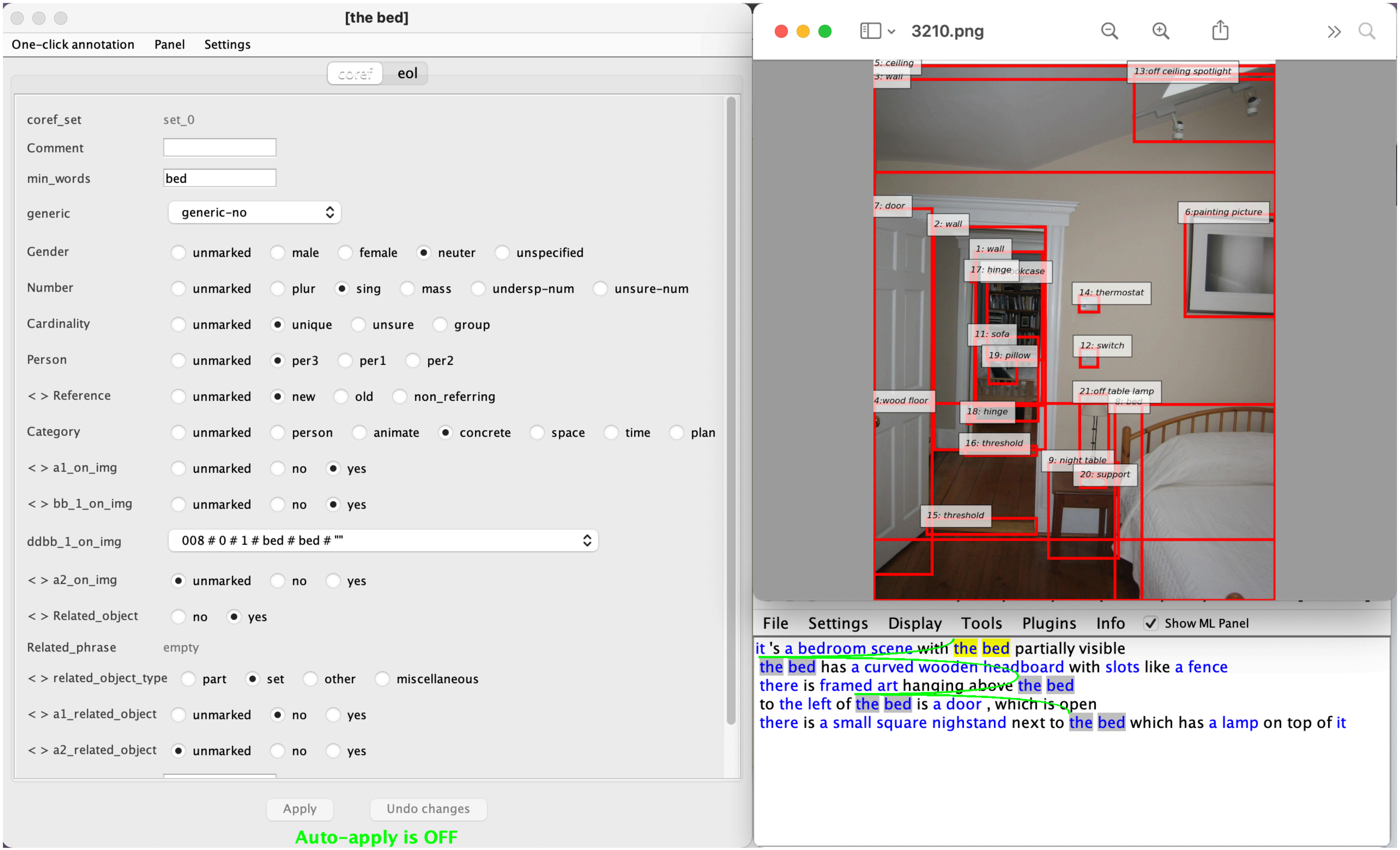
1) There is a **four chair red lacquer dining set** shown in the image.

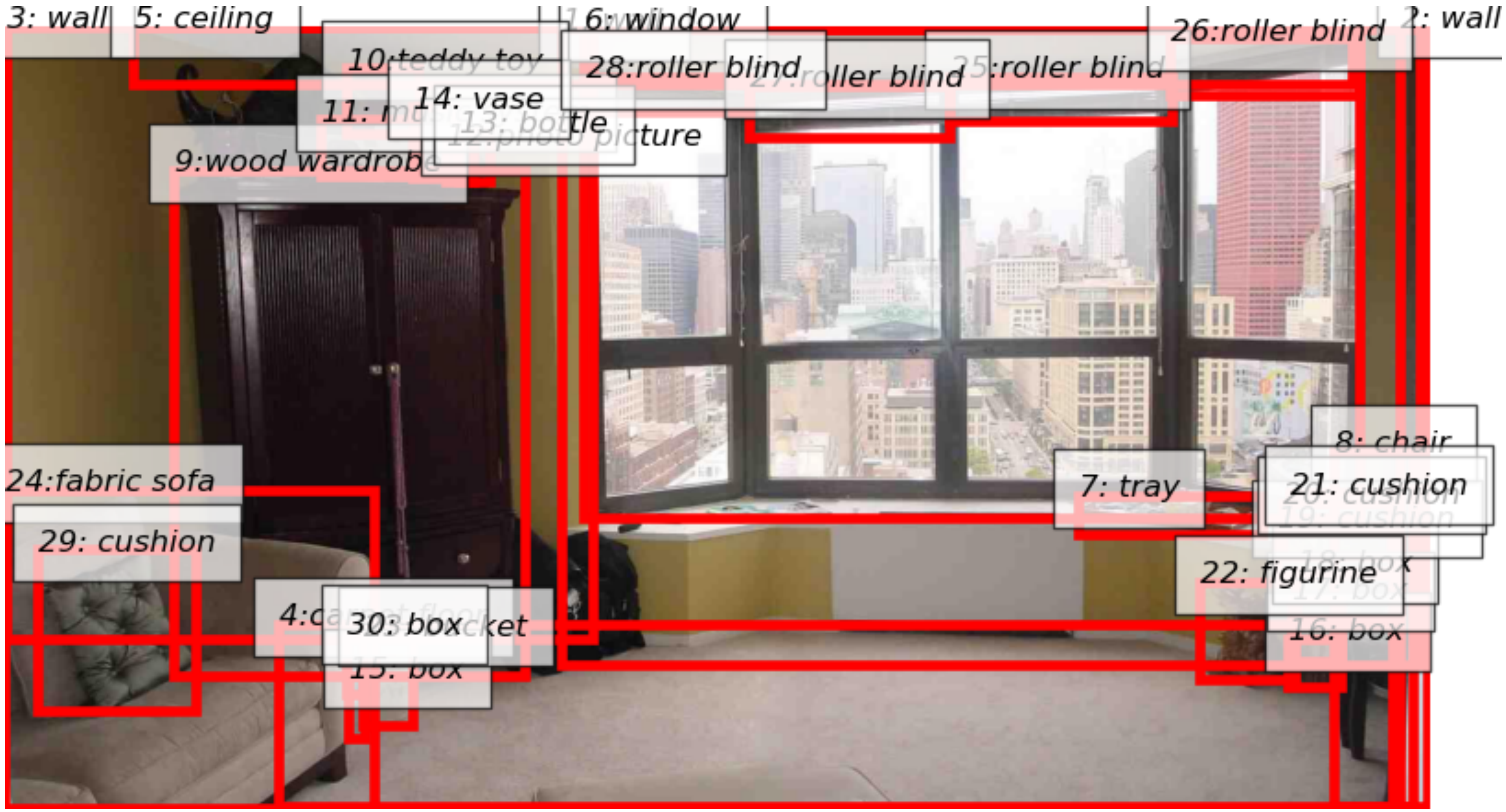
2) There are opened white french doors leading to the outside showing.

3) There is a pool with blue water showing through the french doors.

4) The pool is surrounded by green shrubbery.

5) The wood floor is covered with white paint.


[the bed]
One-click annotation
Panel
Settings
coref
eol
coref_set
set_0
Comment
min_words
bed
generic
generic-no
Gender
unmarked
male
female
neuter
unspecified
Number
unmarked
plur
sing
mass
undersp-num
unsure-num
Cardinality
unmarked
unique
unsure
group
Person
unmarked
per3
per1
per2
< > Reference
unmarked
new
old
non_referring
Category
unmarked
person
animate
concrete
space
time
plan
< > a1_on_img
unmarked
no
yes
< > bb_1_on_img
unmarked
no
yes
ddbb_1_on_img
008 # 0 # 1 # bed # bed # ""
< > a2_on_img
unmarked
no
yes
< > Related_object
no
yes
Related_phrase
empty
< > related_object_type
part
set
other
miscellaneous
< > a1_related_object
unmarked
no
yes
< > a2_related_object
unmarked
no
yes
Apply
Undo changes
Auto-apply is OFF
3210.png
5: ceiling
3: wall
13:off ceiling spotlight
7: door
2: wall
6:painting picture
1: wall
17: hinge
14: thermostat
11: sofa
12: switch
19: pillow
21:off table lamp
4:wood floor
18: hinge
16: threshold
9: night table
20: support
15: threshold
File
Settings
Display
Tools
Plugins
Info
Show ML Panel
it 's a bedroom scene with the bed partially visible
the bed has a curved wooden headboard with slots like a fence
there is framed art hanging above the bed
to the left of the bed is a door , which is open
there is a small square nighstand next to the bed which has a lamp on top of it



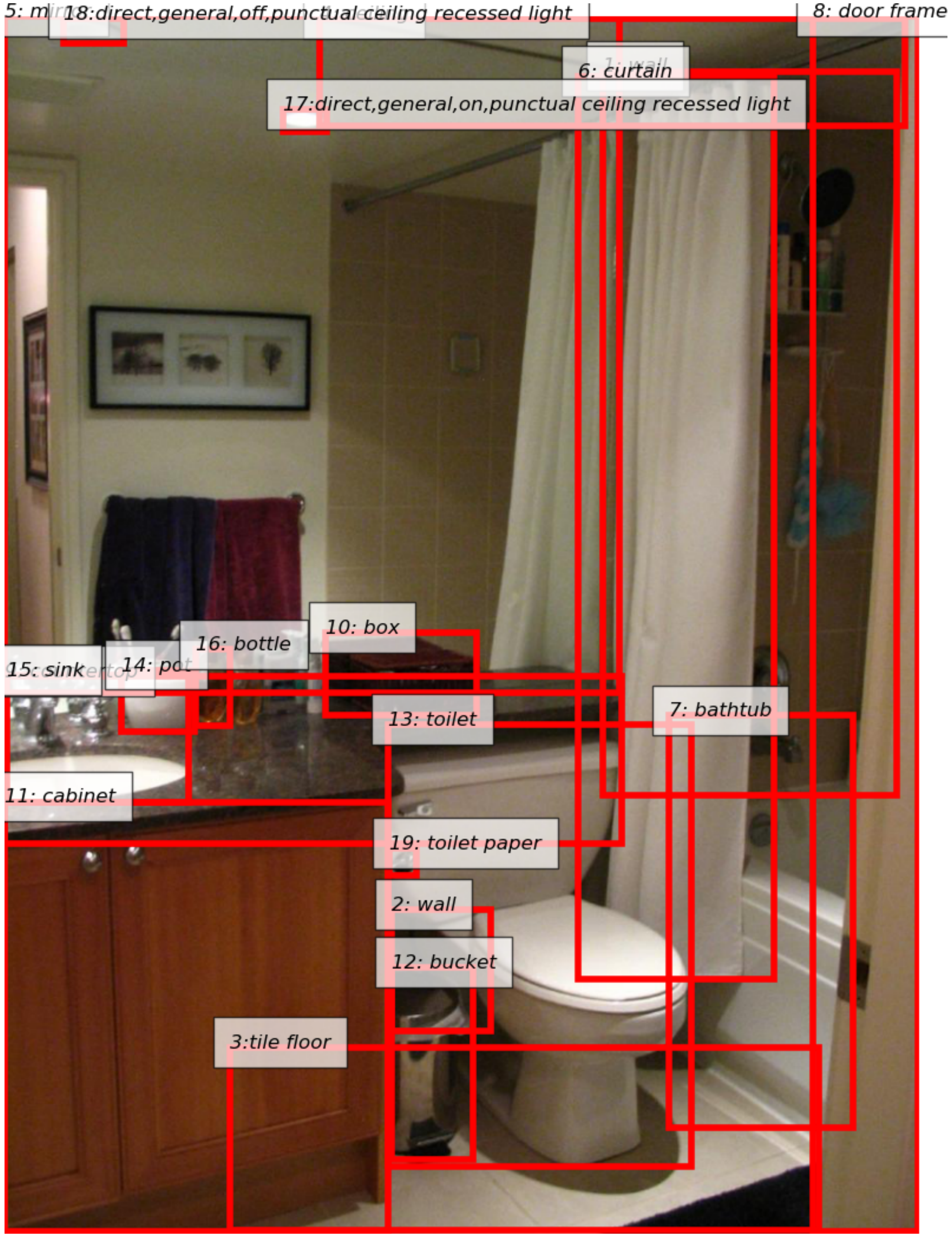
1. a city can be seen outside a large window.

2. there is a fancy tall dark brown cabinet on the left.

3. part of a light brown sofa can be seen on the left.

4. many small items are on the cabinet.

5. what looks like a white heater is straight ahead, near the floor.



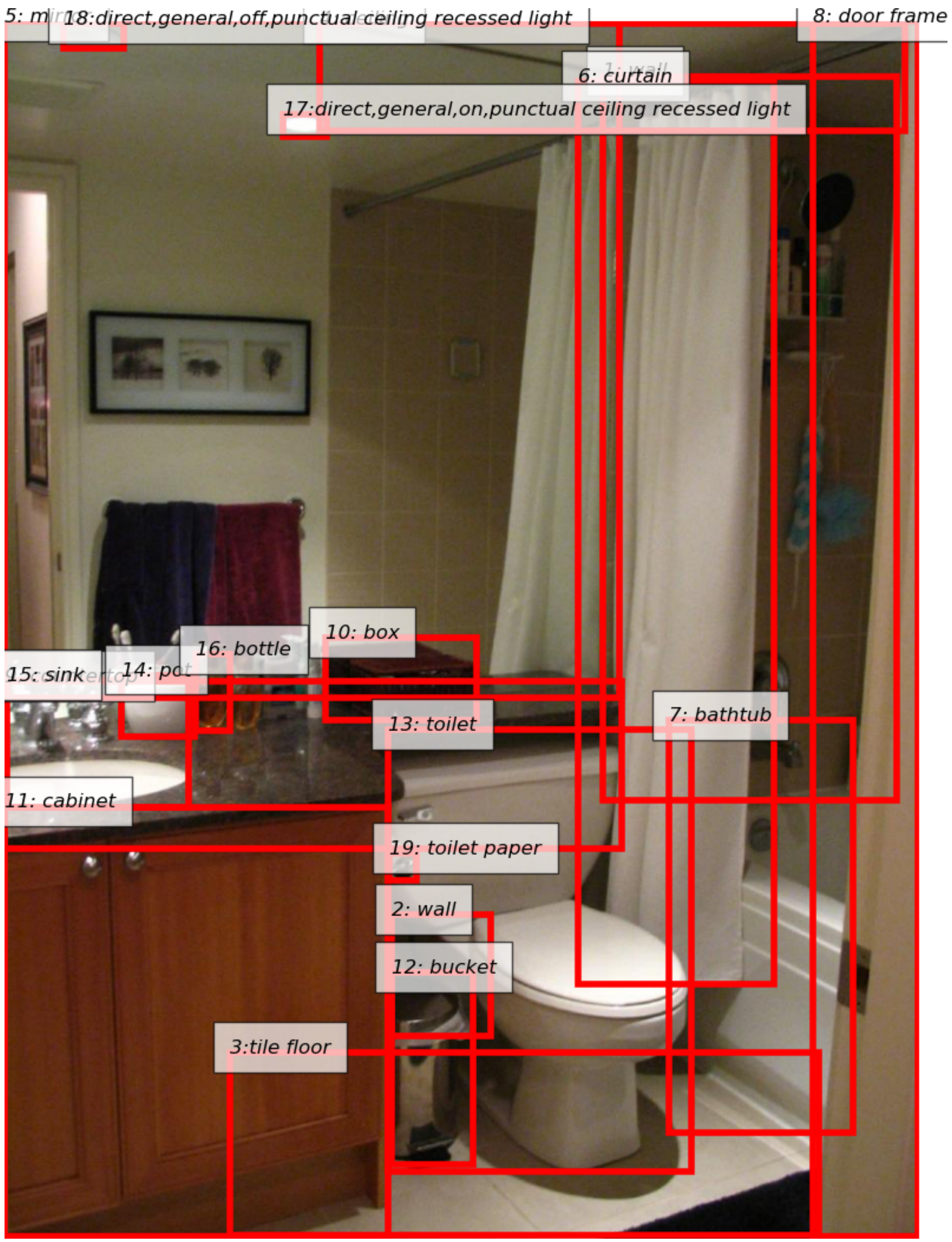
1. this is a bathroom

2. there is a bathtube with white drapes

3. the toilet is white with wide top and is located between the sink and bathtub

4. the sink has wooden cabinet underneath

5. the floor is white tiles.



1. this is **a bathroom**

2. there is **a bathtube** with white drapes

3. **the toilet** is white with wide top and is located between **the sink** and **bathtub**

4. **the sink** has wooden cabinet underneath

5. the floor is white tiles.

Part 4:
Conclusions

Conclusions

- Part 1: Pre-trained Language Models
  - Big opportunity to learn how far can we get away with expectations for anaphora resolution.
  - Pre-trained LMs encode discourse knowledge, so let's exploit that to update our discourse theories.
- Part 2: Psycholinguistics
  - Very few items but highly controlled conditions: type of verb, it / this, aspect.
  - Depth rather than breadth of analysis: study took 3 years.
  - Still, we can take inspiration from how humans solve the task.

Conclusions

- Part 3: Linguistic annotation
  - This annotation makes explicit the relations between text and image.
  - In realistic terms, it’s how linguistic theory reaches the systems we train.
  - Corpus annotation is a thankless job, but extremely necessary to advance the field.
- All parts:
  - We need to look at more languages than just English.

Thank you

Extra Material

Comparison with corpus statistics

ParcorFull, Lapshinova-Koltunski et al. (2018)

	Antecedent	English		German			French		
		this	**it**	**es**	**das**	**dies**	**c'**	**il**	**cela**
Human responses	Entity	4	52	35	1	5	6	31	6
	Event	38	6	3	22	34	24	1	32
Corpus annotation	Entity	7	61	20	28	1	27	23	6
	Event	15	17	5	44	2	33	0	11

All the cells are percentages

Language models and surprisal

- Current pre-trained language models (LMs) are ubiquitous. They are the backbone of a lot of applications in computational linguistics, natural language processing, and AI.
- Sentences can be seen as sequences w1…wn, where n is the length of the sentence.
- The language modeling task is to predict an unseen wi, where 1 ≤ n ≤ i.
- This is expressed as the probability p(wi|w1…wi-1),
- where w1…wi-1 is the left context.
- This also makes it very smooth to compute surprisal
- s(wi) = –log(p(wi|w1…wi-1)).

Conclusions Part 1

- Pre-trained representations encode discourse knowledge about entities, but this is not a hard task.
- LSTMs are able to further contextualize pre-trained embeddings for this task at the sentence level, suggesting that part of the information is encoded at the sentence level.
- Localizing the entity within the sentence is difficult, implying that identifying referring discourse entities from scratch is a problem.