Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. TACL 2016

Author: Tal Linzen, Emmanuel Dupoux, Yoav Goldberg

Repoter: Yang Liu

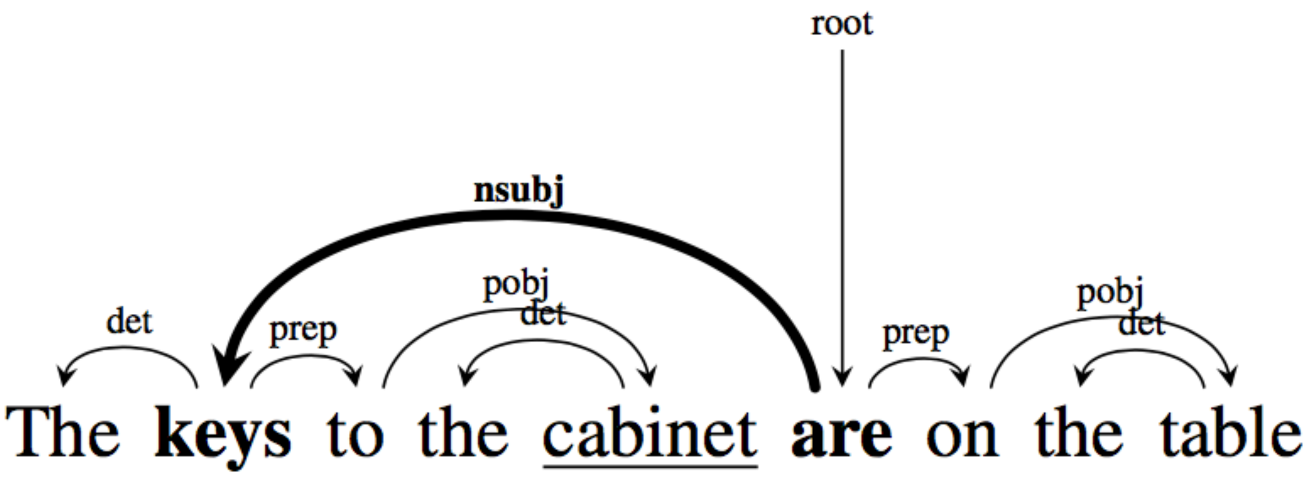
Motivation

word co-occurrence statistics (arbitrary number of words)

Sentence 1: Paris ··· France ··· .   (more likely)
Sentence 1: Penguins ··· France ··· .

	N-gram	RNN
fixed number of words	✔	✔
arbitrary number of words	✘	✔
syntactic structure of the sentence	✘	?

Subject-Verb Agreement as Evidence for Syntactic Structure

1. 

2. The **building** on the far right that's quite old and run down is **the Kilgore Bank Building.**

3. Alluvial **soils** carried in the *floodwaters* add nutrients to the floodplains.
   The **length** of the forewings is 12-13.
   Yet the **ratio** of men who survive to the women and children who survive is not clear in this story.

The Number Prediction Task

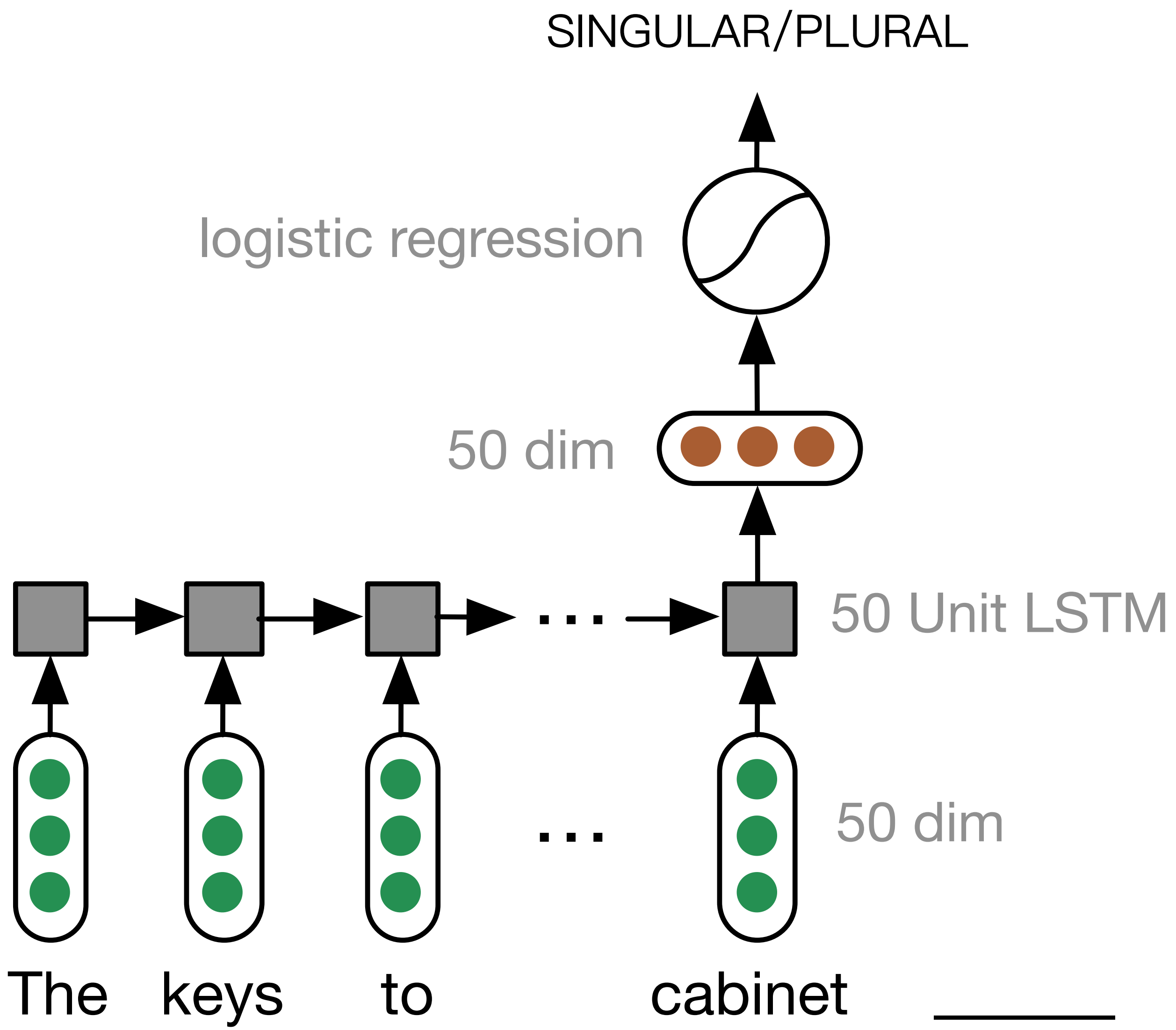
Given: The keys to the cabinet ______
To Predict: PLURAL or SINGULAR

- Model syntactic number and syntactic subject-hood
- sensitivity to hierarchical syntax

Data

- generate practically **unlimited** training and testing examples
- based on **Wikipedia**
- ~1.35 million number prediction problems
- ~121,500 (9%) for training
- ~13,500 (1%) for validation
- ~1.21 million (90%) for test (enough for less common constructions)

Model


SINGULAR/PLURAL
logistic regression
50 dim
50 Unit LSTM
50 dim
The
keys
to
cabinet

Baseline (*noun-only baselines*)

- only receives common nouns (*dogs*, *pipe*)
- also receives pronouns (*he*) and proper nouns (*France*).

Results-Overall

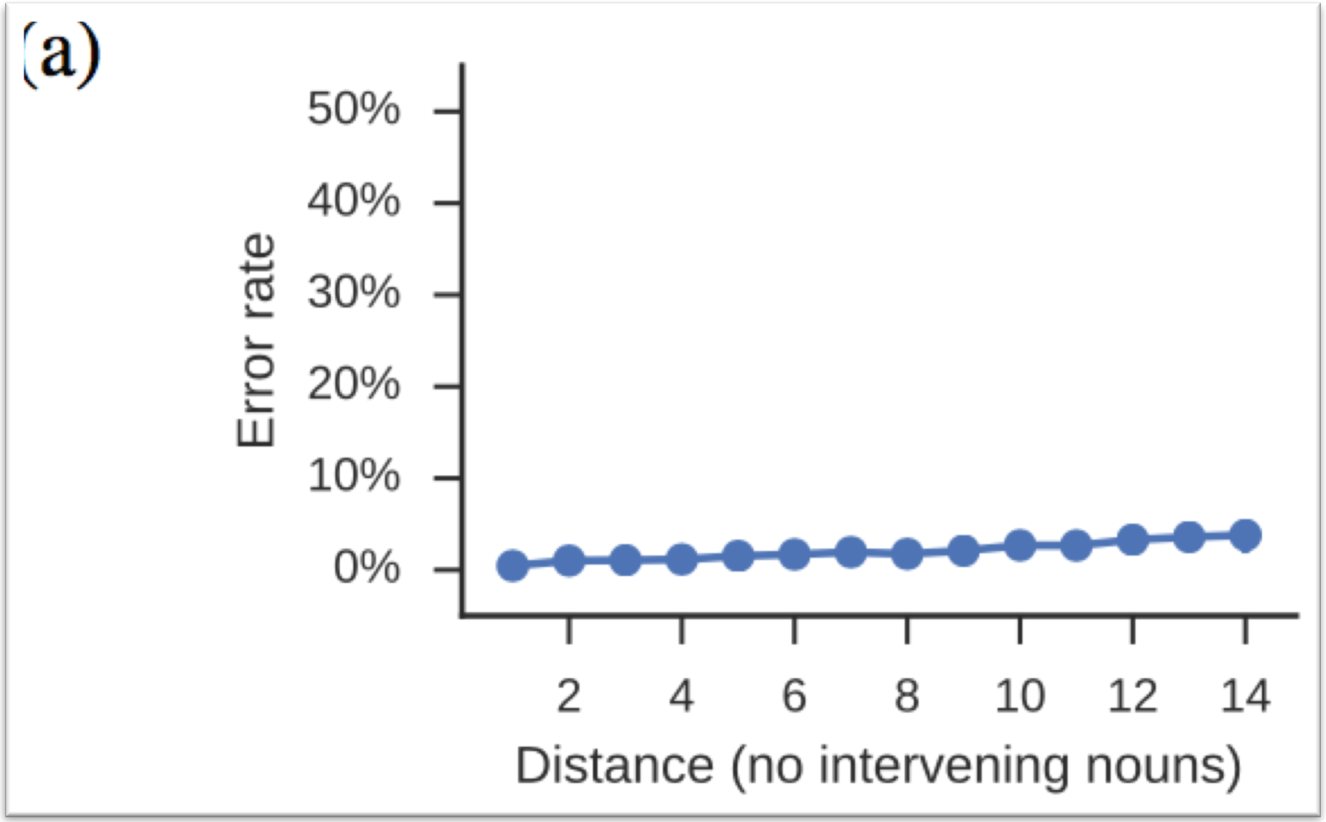
	All-words	Common-nouns	All-nouns
Error	0.83%	4.2%	4.5%

How is the performance on more challenging dependencies?

Results-Distance

- no nouns intervened between the subject and the verb.

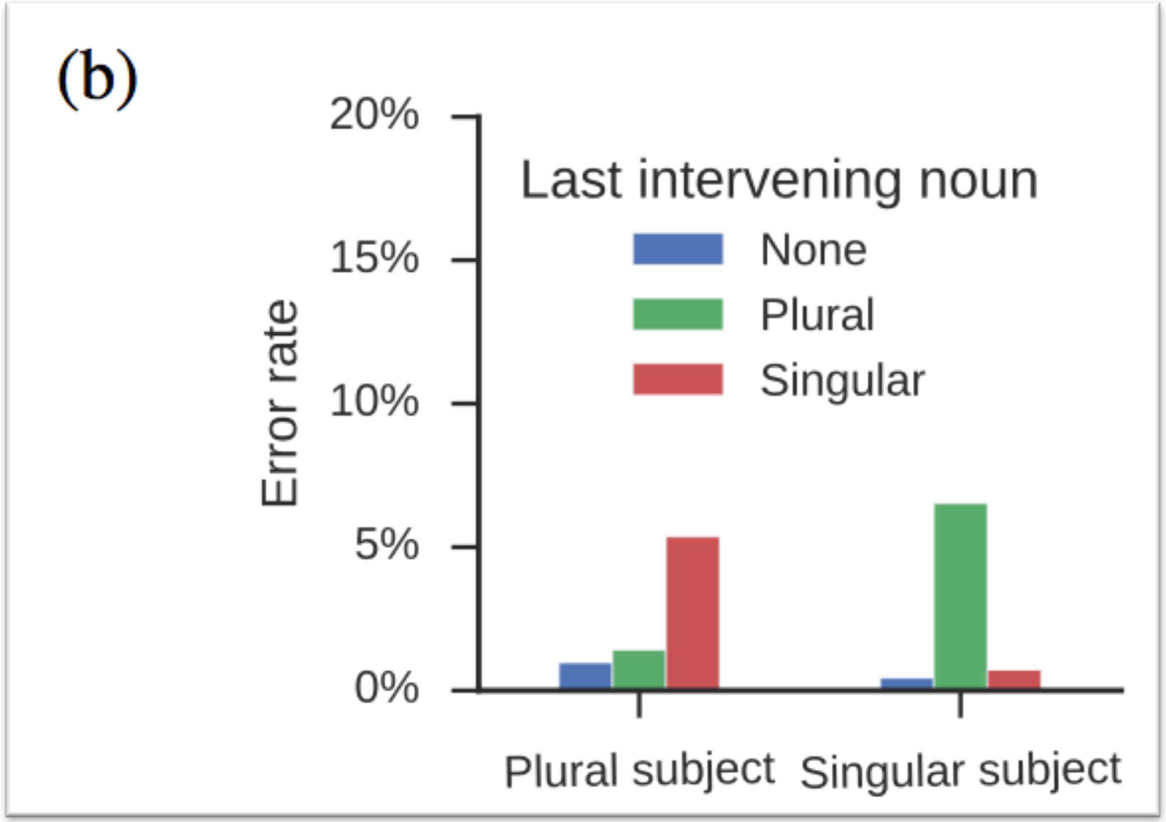
- the network generalized the dependency from the common distances of 0 and 1 to rare distances of 10 and more.



Results-Agreement attractors

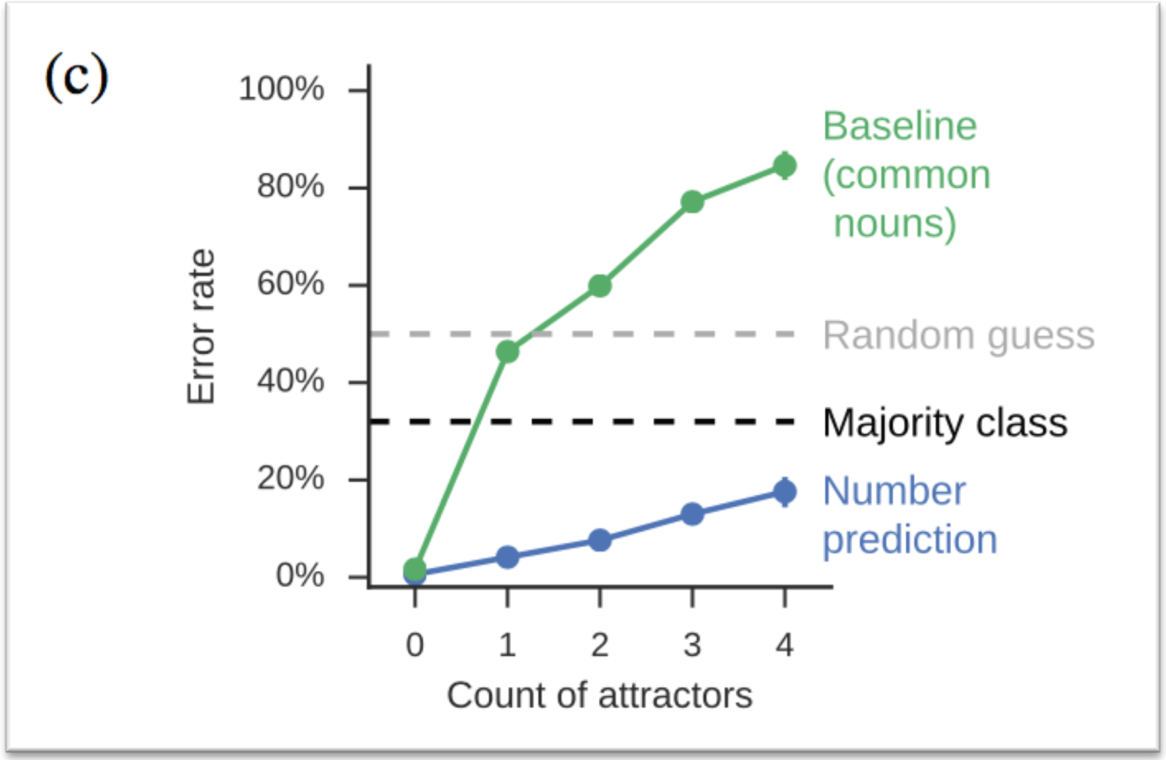
- Last intervening noun of the
  - same number +0.3-0.4%
  - differ number x10

- Baseline with error rates of
  - 46.4% (common nouns)
  - 40% (all nouns).



Results-Attractors' effect cumulative?

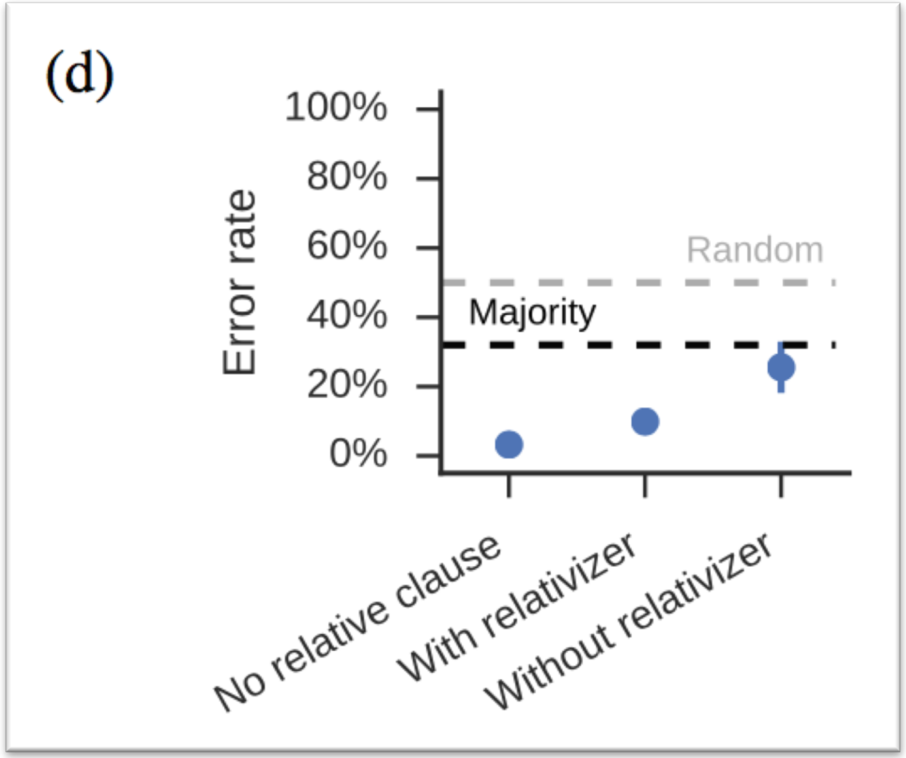
- ※homogeneous intervention
  - The **roses** in the vase by the door are red.
  - The **roses** in the vase by the ~~chairs~~ are red.
- Attractors with number of
  - 4 word 17.6%
- Baseline with error rates of
  - 84% (common nouns)
- confirms that syntactic cues are critical



Results- Relative clauses

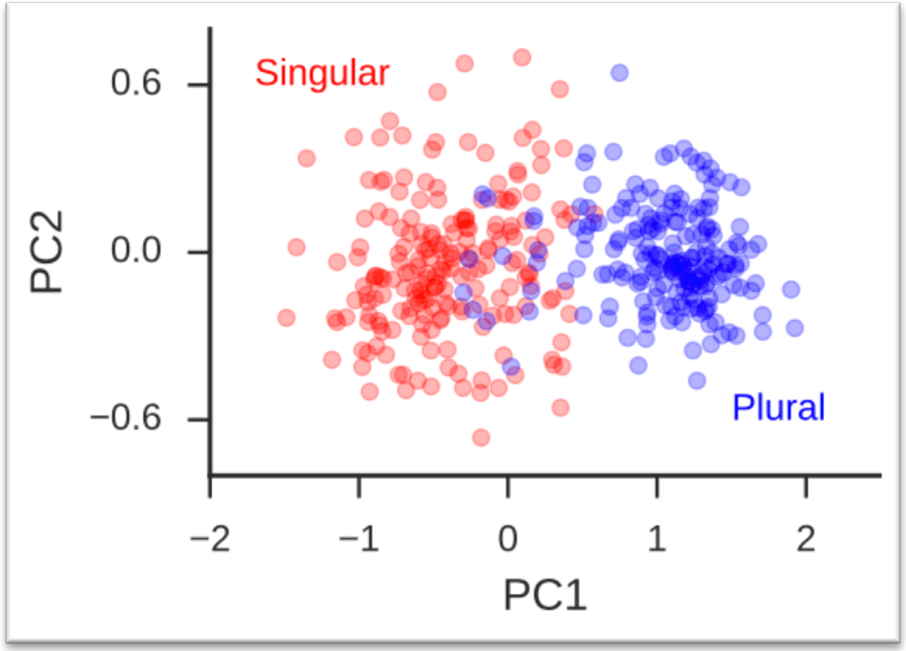
- E.g.
  - The **landmarks** (that) this <u>article</u> lists here are also run-of-the-mill and not notable.

- Control only one attractor.
- No clauses 3.2%
- Clauses
  - With relativizer(that, which etc.) 9.9%
  - Without elativizer 25%



Results- ~~Word representations~~

- PCA on Word-Embedding (50 dims)
- PC1 corresponded number of the noun
- Note that:
  - Model not have access to suffixes such as *-s*

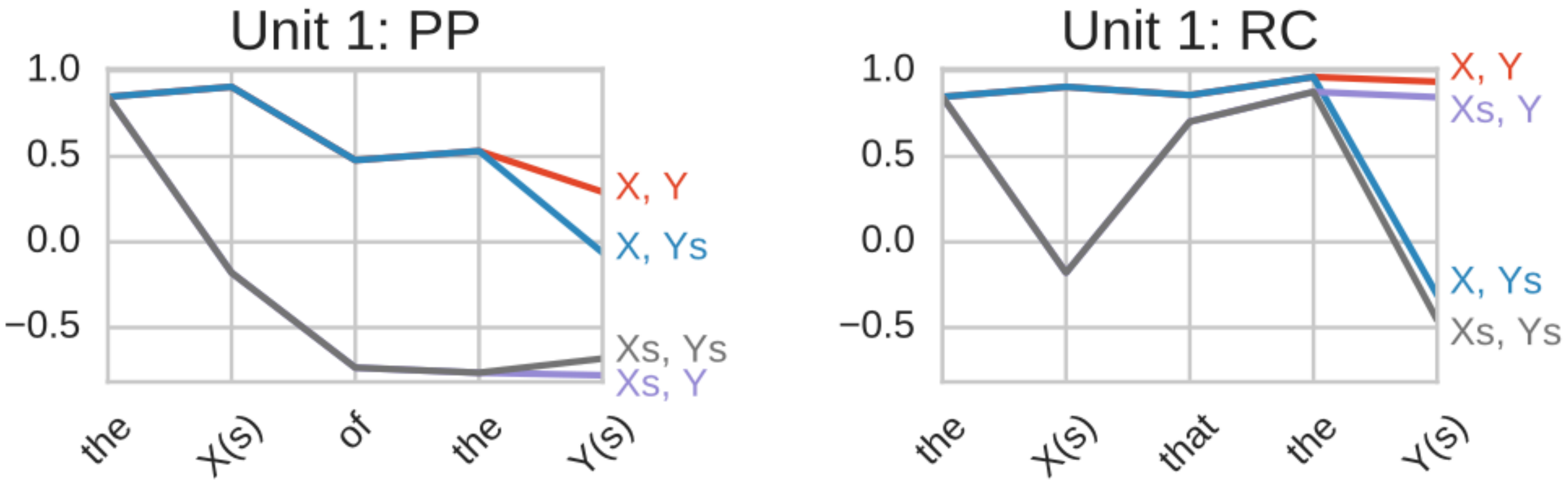


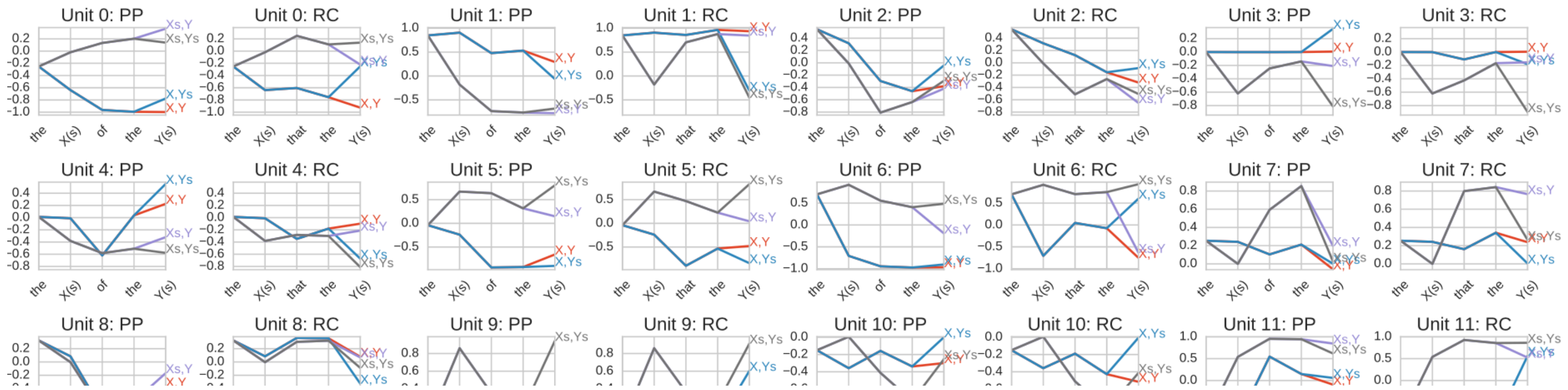
Results-Visualizing the network's activations

- Use constructed sentences simplify.
  - **PP:** The **toy(s)** of the boy(s)…
  - **RC:** The toy(s) that the **boy(s)**…
  - (2*2) * (10 diff. n-n relation) * (2 rc,pp)= 80

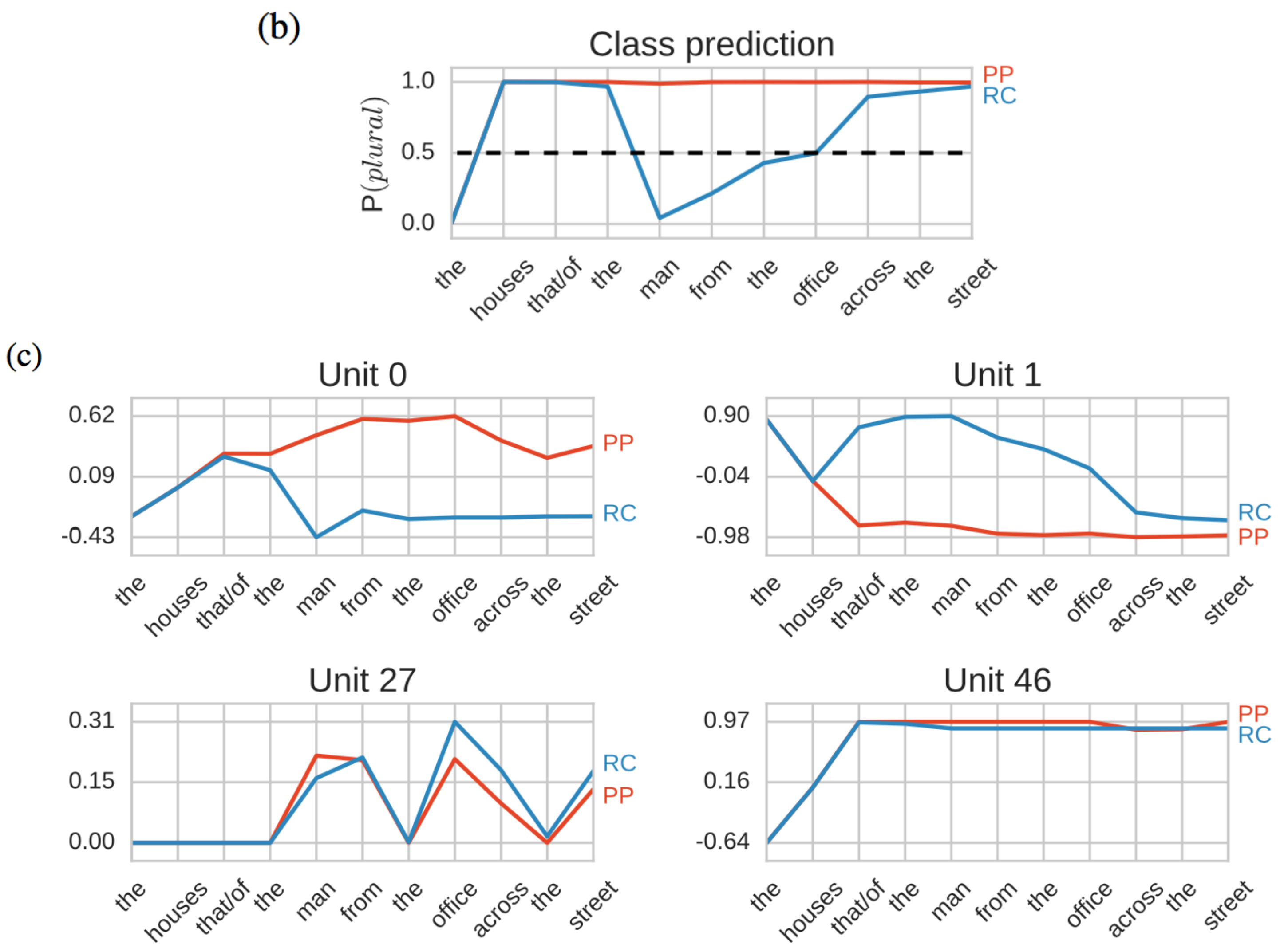
Results-Visualizing the network's activations

(a)





Results-Visualizing the network's activations



Alternative Training Objectives

Training objective	Sample input	Training signal	Prediction task	Correct answer
Number prediction	*The keys to the cabinet*	PLURAL	SINGULAR/PLURAL?	PLURAL
Verb inflection	*The keys to the cabinet [is/are]*	PLURAL	SINGULAR/PLURAL?	PLURAL
Grammaticality	*The keys to the cabinet are here.*	GRAMMATICAL	GRAMMATICAL/UNGRAMMATICAL?	GRAMMATICAL
Language model	*The keys to the cabinet*	are	$P(are) > P(is)$?	True

Verb inflection Task

Training objective	Sample input	Training signal	Prediction task	Correct answer
Number prediction	*The keys to the cabinet*	PLURAL	SINGULAR/PLURAL?	PLURAL
Verb inflection	*The keys to the cabinet [is/are]*	PLURAL	SINGULAR/PLURAL?	PLURAL
Grammaticality	*The keys to the cabinet are here.*	GRAMMATICAL	GRAMMATICAL/UNGRAMMATICAL?	GRAMMATICAL
Language model	*The keys to the cabinet*	are	$P(are) > P(is)$?	True

- Verb is known. ([be] in the example)
- Subject – verb. **Semantics information**
  - Eg. **People** from the capital often eat pizza.
  - (only *people* is a plausible subject for *eat*)

Grammaticality judgments

Training objective	Sample input	Training signal	Prediction task	Correct answer
Number prediction	*The keys to the cabinet*	PLURAL	SINGULAR/PLURAL?	PLURAL
Verb inflection	*The keys to the cabinet [is/are]*	PLURAL	SINGULAR/PLURAL?	PLURAL
Grammaticality	*The keys to the cabinet are here.*	GRAMMATICAL	GRAMMATICAL/UNGRAMMATICAL?	GRAMMATICAL
Language model	*The keys to the cabinet*	are	$P(are) > P(is)$?	True

- Whole sentence is known.
- Verb position / syntactic clause boundaries

Language modeling (LM)

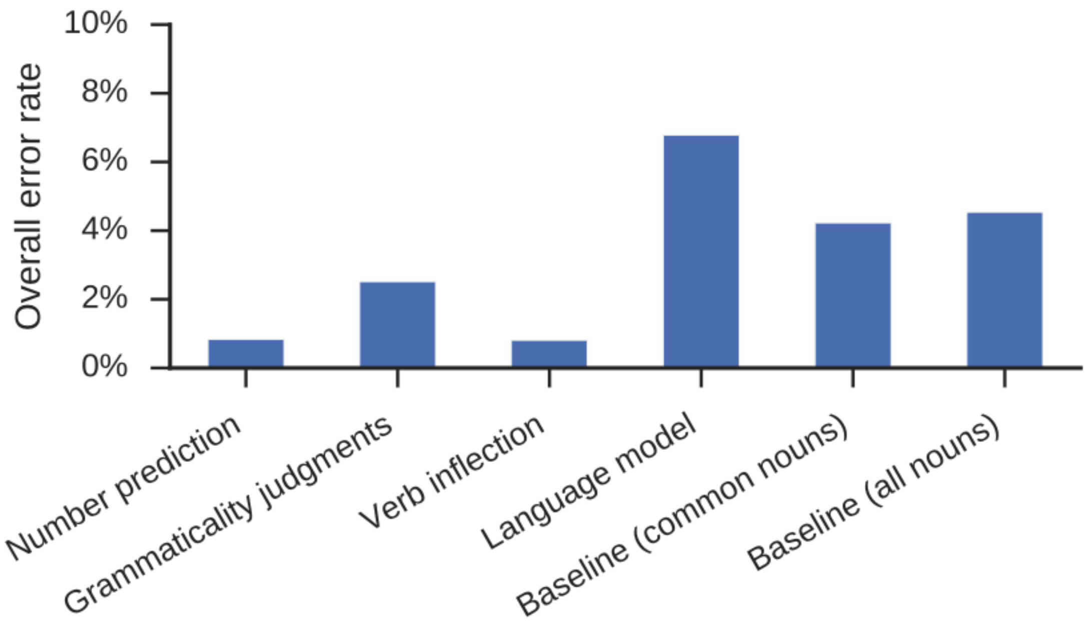
Training objective	Sample input	Training signal	Prediction task	Correct answer
Number prediction	*The keys to the cabinet*	PLURAL	SINGULAR/PLURAL?	PLURAL
Verb inflection	*The keys to the cabinet [is/are]*	PLURAL	SINGULAR/PLURAL?	PLURAL
Grammaticality	*The keys to the cabinet are here.*	GRAMMATICAL	GRAMMATICAL/UNGRAMMATICAL?	GRAMMATICAL
Language model	*The keys to the cabinet*	are	$P(are) > P(is)$?	True

- No grammatically relevant supervision
- Model:
  - WordEMB=>RNN=>activate=>fully connected layer=>softmax

Alternative Training Objectives Results

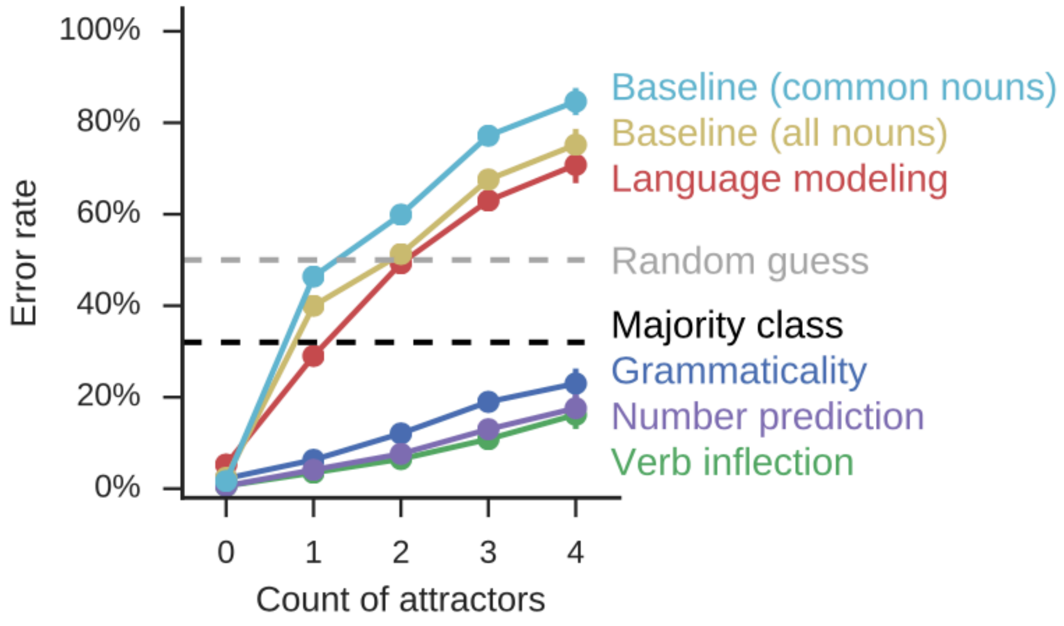
1. verb semantics helps (0.8%=>0.83%)

2. Grammaticality judgments better than Baseline (show to learn syntactic dependencies)



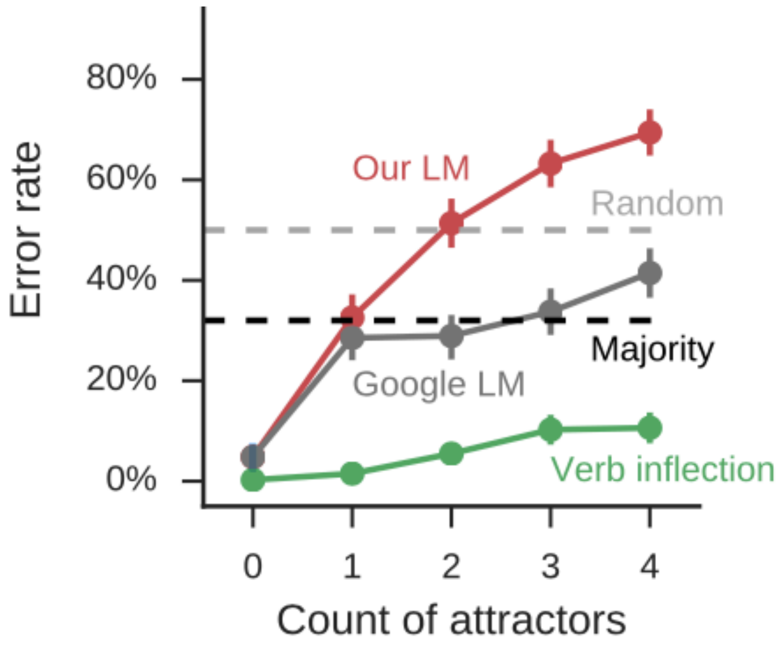
Alternative Training Objectives Results

- Grammaticality is more difficult
- Conclusion
  - LSTM is capable of learning syntax-sensitive agreement dependencies
  - the language-model alone is not sufficient for learning such dependencies



Alternative Training Objectives Results

- LM faced a much harder objective?
- Google LM.
  - vocabulary of 800,000 words
  - two-layer LSTM with 8192 units in each layer
  - 300 times as many units as our LM



Additional Experiments

- Comparison to simple recurrent networks
  - success of the network is due to the LSTM cells?
  - twice errors, not qualitative different.
- Training only on difficult dependencies

Error Analysis

- Singular vs. plural subjects
  - Violate prior probability experience when using SRN model
- Qualitative analysis
  - 1. n-n compounds. 2. v/n word. 3. hard to recognize subject

Conclusion

- LSTMs can learn to approximate structure-sensitive dependencies fairly well given explicit supervision
- more expressive architectures may be necessary to eliminate errors altogether.
- language modeling objective is not by itself sufficient for learning structure-sensitive dependencies

Summary of the reporter

- Baseline model is ingenious.
- Homogeneous intervention. Variables control.
- Interpretability
- The whole work begin with the easy and efficiency function to build the large dataset.

Thanks and Q&A.