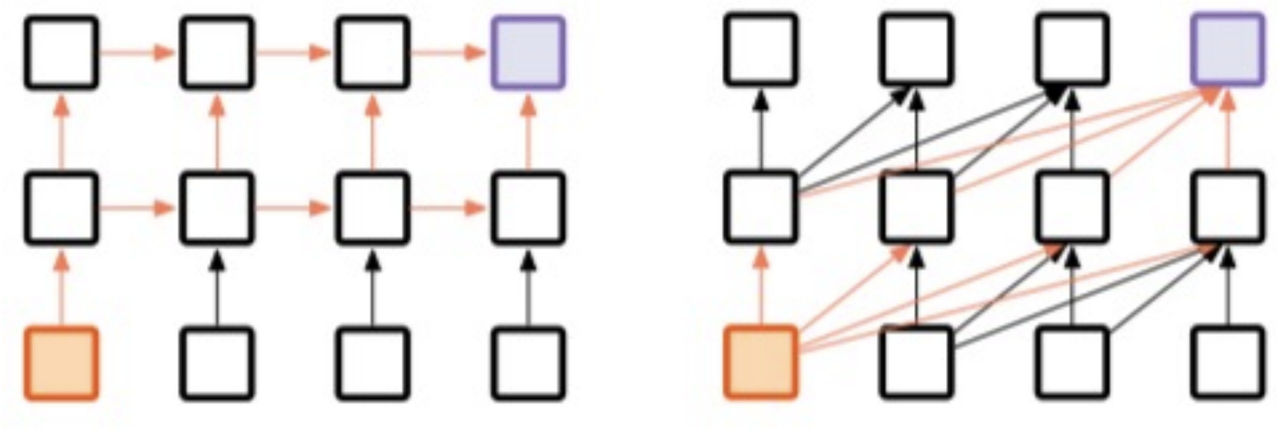
The Importance of Being Recurrent for Modeling Hierarchical Structure

Presenter: Baosong Yang

Motivation

- Do FANs have the same ability to exploit hierarchical structures *implicitly* in comparison to RNNs



(a) LSTM (b) FAN

- Two tasks:
  - (1) subject-verb agreement
  - (2) logical inference.

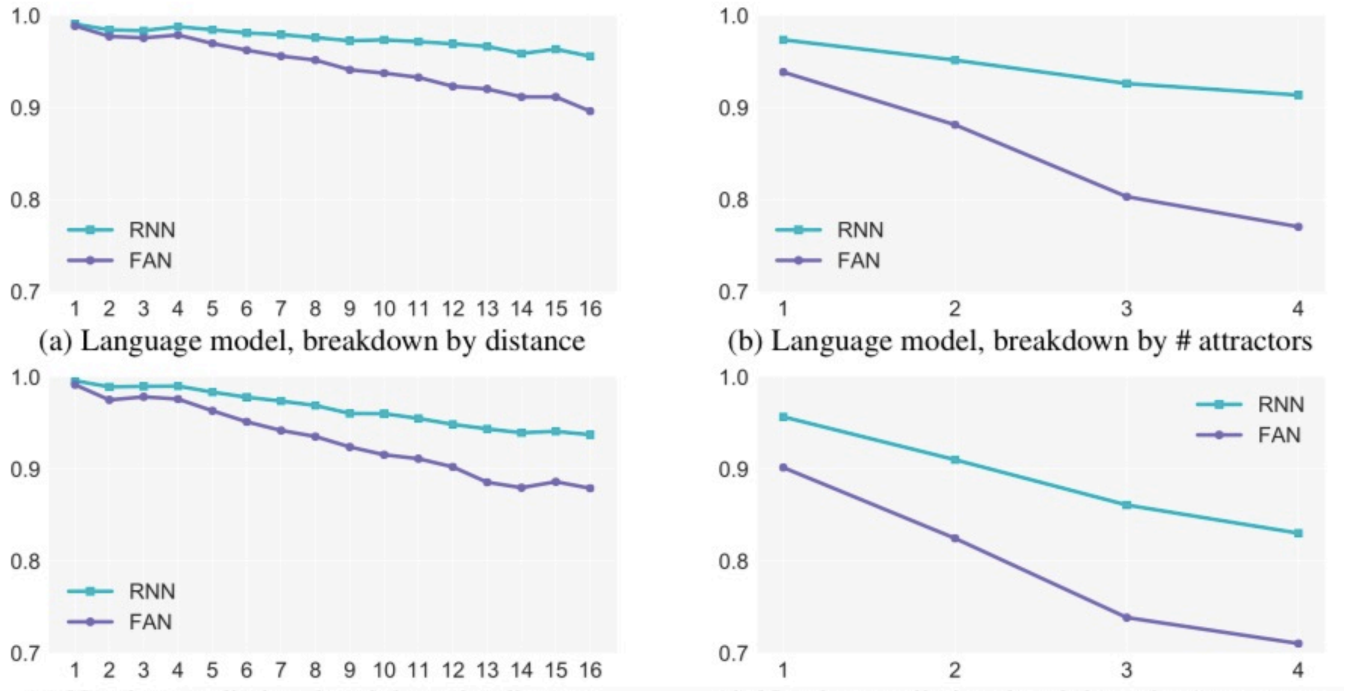
Subject-Verb Agreement

- Predicting number agreement between subject and verb in naturally occurring English sentences:
  - a) a general language model
  - 2) predict the number of the verb given its sentence history.

	Input	Train	Test
(a)	the keys to the cabinet	are	$p(\text{are}) > p(\text{is})$?
(b)	the keys to the cabinet	plural	plural/singular?

- Settings:
  - 10% of the data for training, 1% for validation, and the rest for testing.
  - 4 layers, the dropout rate is 0.2, and word-embeddings and hidden sizes are set to 128.
  - 2 attention head for Transformer.

Subject-Verb Agreement



(a) Language model, breakdown by distance

(b) Language model, breakdown by # attractors

(c) Number prediction, breakdown by distance

(d) Number prediction, breakdown by # attractors

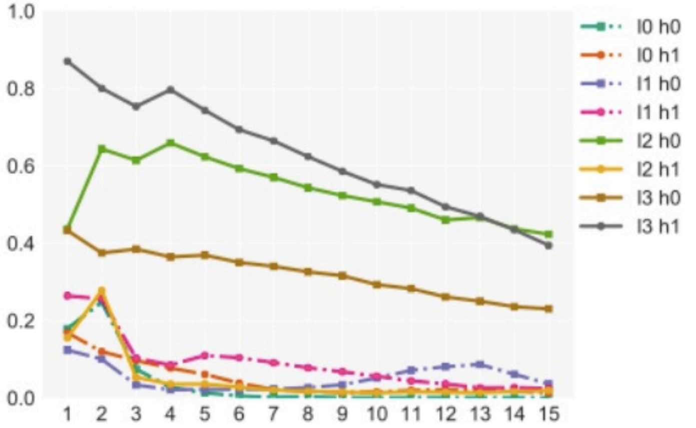


Figure 3: Proportion of times the subject is the most attended word by different heads at different layers (l3 is the highest layer). Only cases where the model made a correct prediction are shown.

- LSTMs are clearly more robust than FANs with respect to task difficulty
  - (a) word distance
  - (b) number of agreement attractors: intervening nouns with the opposite number from the subject.

Logical Inference

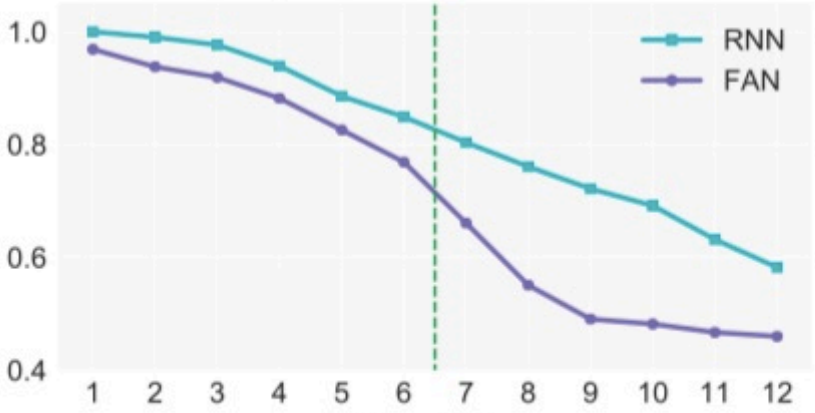
(d (or f)) ⊐ (f (and a))

(d (and (c (or d)))) # (not f)

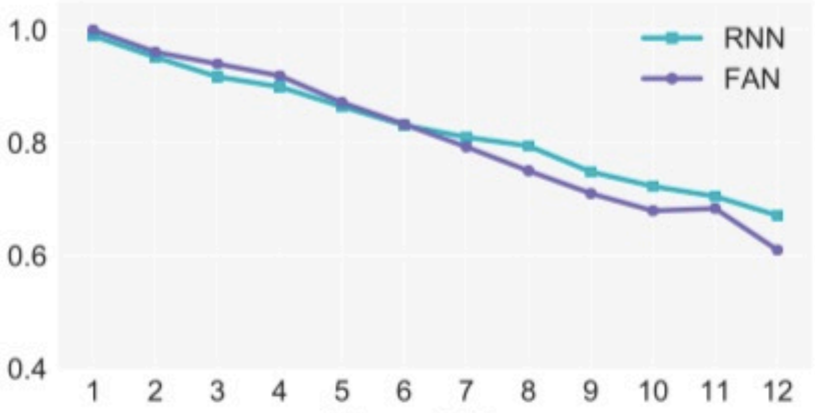
(not (d (or (f (or c))))) ⊏ (not (c (and (not d))))

- 6 word types {*a*, *b*, *c*, *d*, *e*, *f*} and 3 logical operations {*or*, *and*, *not*}.
- 7 mutually exclusive logical relations between two sentences:
  - entailment (,)
  - equivalence (≡)
  - exhaustive and non-exhaustive contradiction (∧, |)
  - two types of semantic independence (#,)
- Settings:
  - train/dev/test dataset ratios are set to 80%/10%/10% from 60k samples.

Logical Inference



(a) $n \leq 6$



(b) $n \leq 12$

- n: Number of logical operations
- (a): training on n<=6
- (b): training on n<=12

What we learn?

- LSTMs slightly but consistently outperform FANs with respect to the ability of capturing hierarchical structure.
  - more robust
- Problem:
  - Small training set.
  - Small model.
  - FANs might capture other aspects of language better than LSTMs.

Colorless green recurrent networks dream hierarchically

Presenter: Baosong Yang

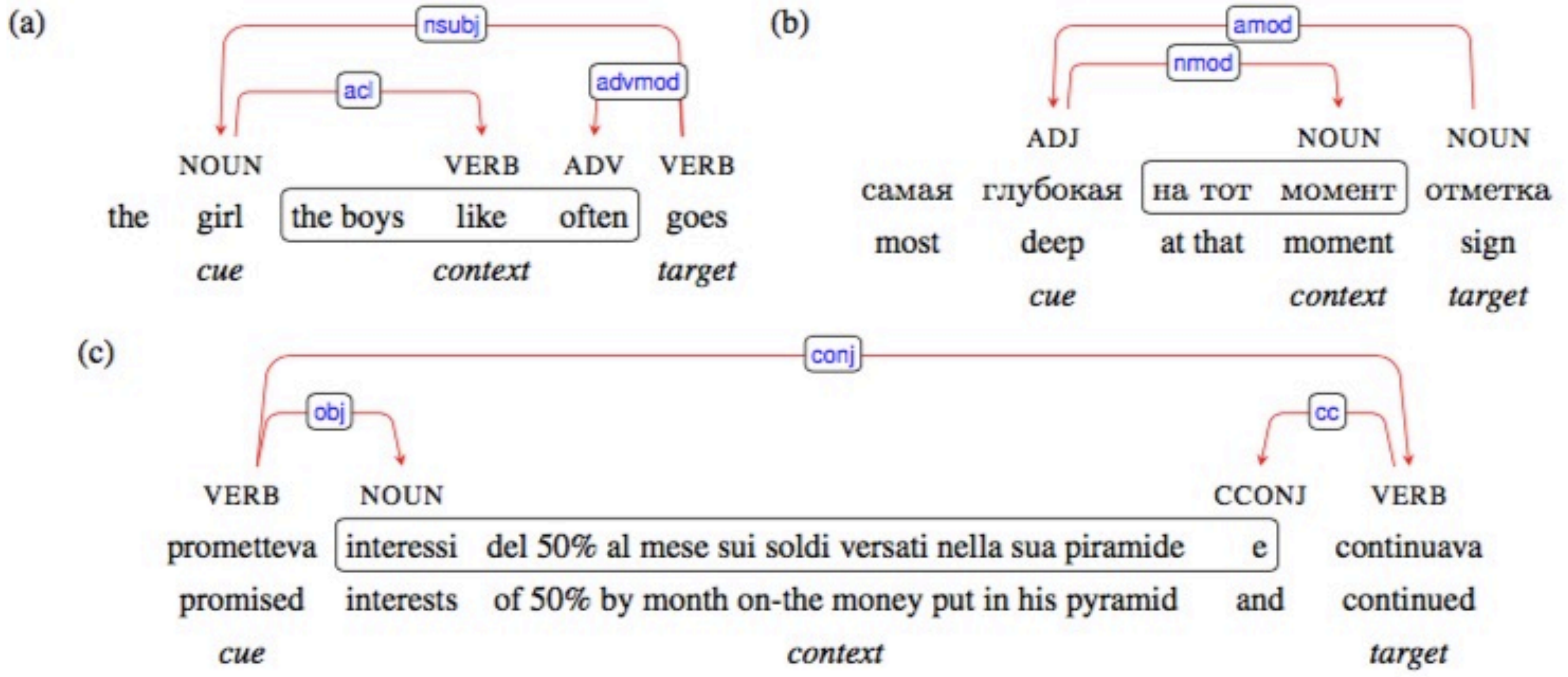
Motivation

- Problem: In “dogs in the neighbourhood often bark”, an RNN might get the right agreement by encoding information about what typically barks (dogs, not neighbourhoods), without relying on more abstract structural cues.

- Hypotheses : Grammaticalness cannot be identified with meaningfulness

- Solusion: A careful architecture/hyperparameter search is crucial to obtain RNNs that are not only good at language modeling, but able to extract syntactic generalizations.

Constructing a long-distance agreement benchmark



- Substituting all content words with random words with the same morphology, resulting in grammatical but nonsensical sequences.

Experiments

	IT	EN	HE	RU
#constructions	8	2	18	21
#original	119	41	373	442
Unigram				
Original	54.6	65.9	67.8	60.2
Nonce	54.1	42.5	63.1	54.0
5-gram KN				
Original	63.9	63.4	72.1	73.5
Nonce	52.8	43.4	61.7	56.8
Perplexity	147.8	168.9	122.0	166.6
5-gram LSTM				
Original	81.8 ±3.2	70.2 ±5.8	90.9 ±1.2	91.5 ±0.4
Nonce	78.0 ±1.3	58.2 ±2.1	77.5 ±0.8	85.7 ±0.7
Perplexity	62.6 ±0.2	71.6 ±0.3	59.9 ±0.2	61.1 ±0.4
LSTM				
Original	92.1 ±1.6	81.0 ±2.0	94.7 ±0.4	96.1 ±0.7
Nonce	85.5 ±0.7	74.1 ±1.6	80.8 ±0.8	88.8 ±0.9
Perplexity	45.2 ±0.3	52.1 ±0.3	42.5 ±0.2	48.9 ±0.6

- Settings: 90M token subsets, training and validation sets (8-to-1 proportion)
- Vocabulary: 50K; Hidden size: 650

		N V V	V NP conj V
Italian	Original	$93.3_{\pm 4.1}$	$83.3_{\pm 10.4}$
	Nonce	$92.5_{\pm 2.1}$	$78.5_{\pm 1.7}$
English	Original	$89.6_{\pm 3.6}$	$67.5_{\pm 5.2}$
	Nonce	$68.7_{\pm 0.9}$	$82.5_{\pm 4.8}$
Hebrew	Original	$86.7_{\pm 9.3}$	$83.3_{\pm 5.9}$
	Nonce	$65.7_{\pm 4.1}$	$83.1_{\pm 2.8}$
Russian	Original	-	$95.2_{\pm 1.9}$
	Nonce	-	$86.7_{\pm 1.6}$

Table 2: LSTM accuracy in the constructions N V V (subject-verb agreement with an intervening embedded clause) and V NP conj V (agreement between conjoined verbs separated by a complement of the first verb).

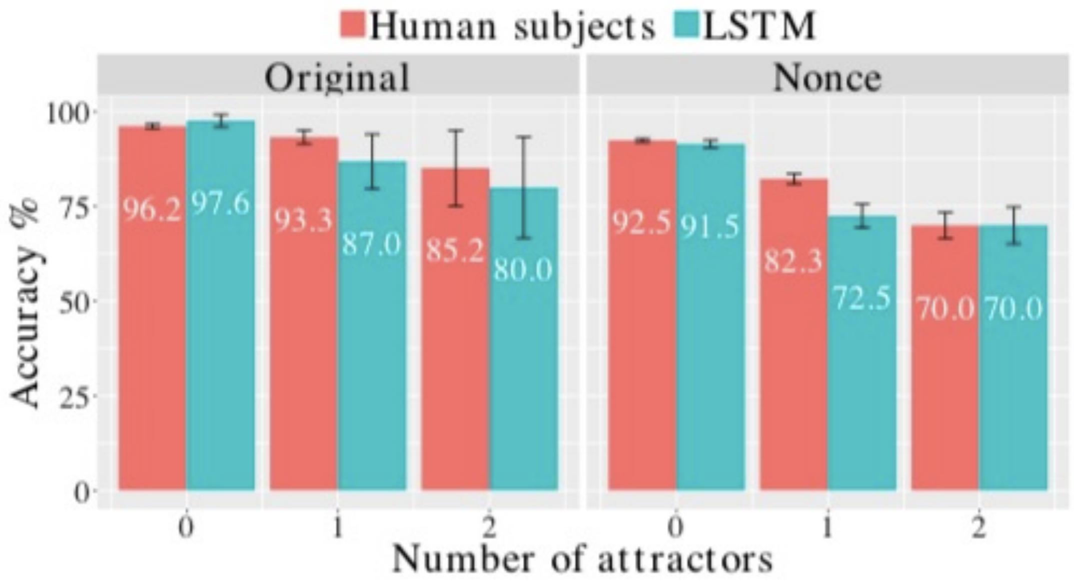
Experiments

- Compare with human being

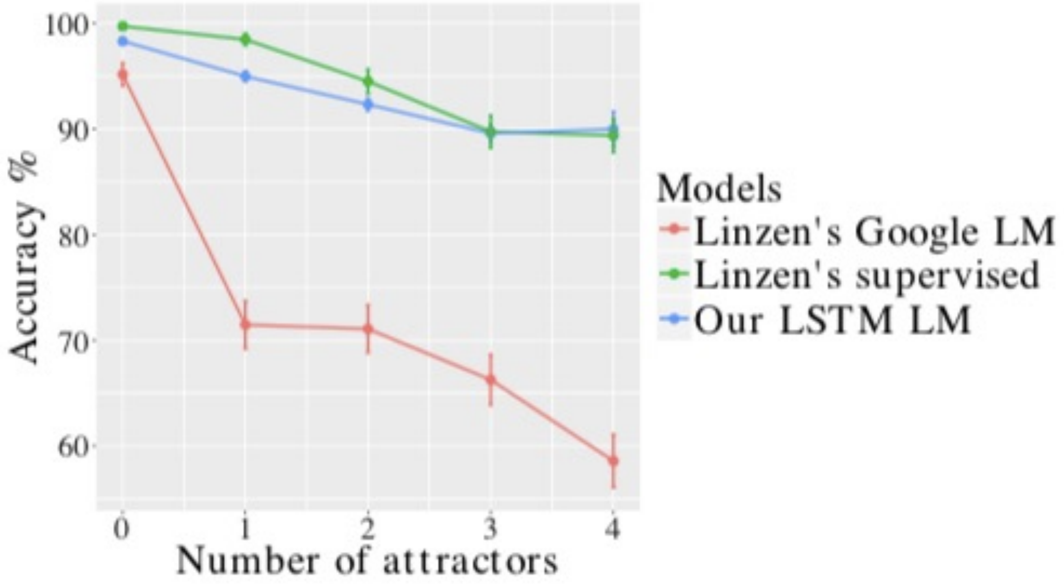
Construction	#original	Original Subjects	Original LSTM	Nonce Subjects	Nonce LSTM
DET [AdjP] NOUN	14	98.7	$98.6_{\pm 3.2}$	98.1	$91.7_{\pm 0.4}$
NOUN [RelC / PartP] clitic VERB	6	93.1	$100_{\pm 0.0}$	95.4	$97.8_{\pm 0.8}$
NOUN [RelC / PartP] VERB	27	97.0	$93.3_{\pm 4.1}$	92.3	$92.5_{\pm 2.1}$
ADJ [conjoined ADJs] ADJ	13	98.5	$100_{\pm 0.0}$	98.0	$98.1_{\pm 1.1}$
NOUN [AdjP] relpron VERB	10	95.9	$98.0_{\pm 4.5}$	89.5	$84.0_{\pm 3.3}$
NOUN [PP] ADVERB ADJ	13	91.5	$98.5_{\pm 3.4}$	79.4	$76.9_{\pm 1.4}$
NOUN [PP] VERB (participial)	18	87.1	$77.8_{\pm 3.9}$	73.4	$71.1_{\pm 3.3}$
VERB [NP] CONJ VERB	18	94.0	$83.3_{\pm 10.4}$	86.8	$78.5_{\pm 1.7}$
(Micro) average		94.5	$92.1_{\pm 1.6}$	88.4	$85.5_{\pm 0.7}$

Experiments

- the overall pattern was comparable



- robust



What we learn

- A new task and data set for analyzing the language model.
- RNNs are not simply memorizing frequent morphosyntactic sequences.