Tabula nearly *rasa*: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text

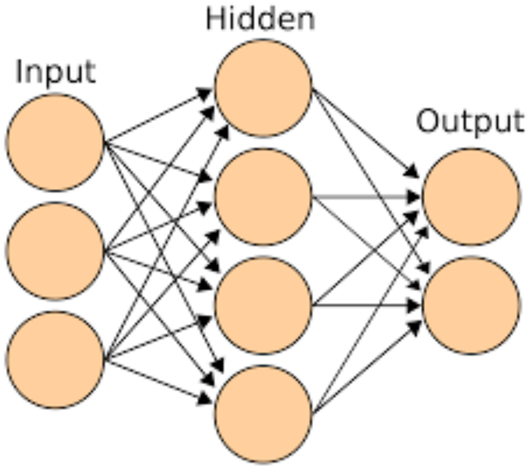
Marco Baroni and Michael Hahn

Facebook AI Research

Outline

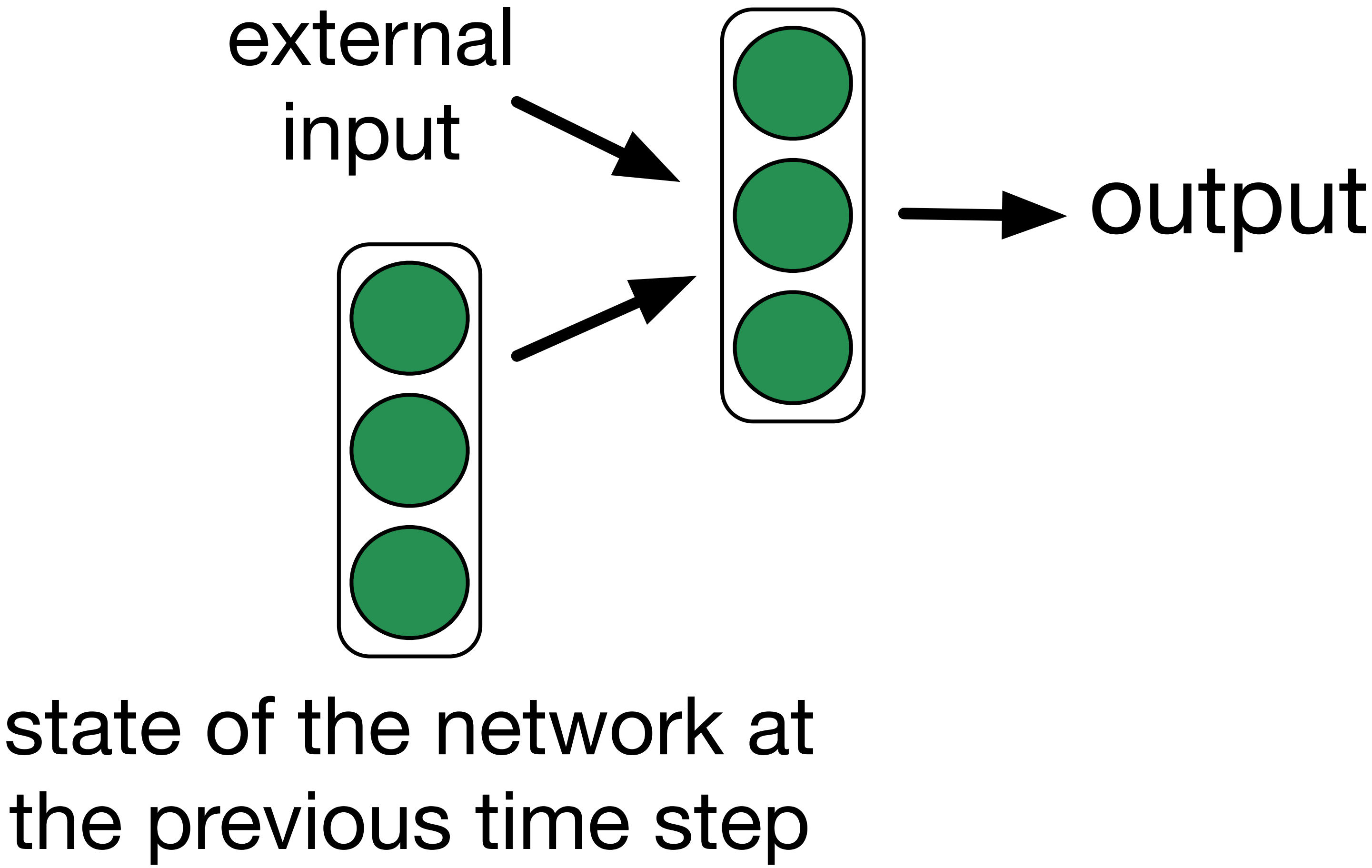
- Motivation
- Linguistic challenges for near-tabula-rasa RNNs
- Discussion

Probing neural networks as comparative psychology



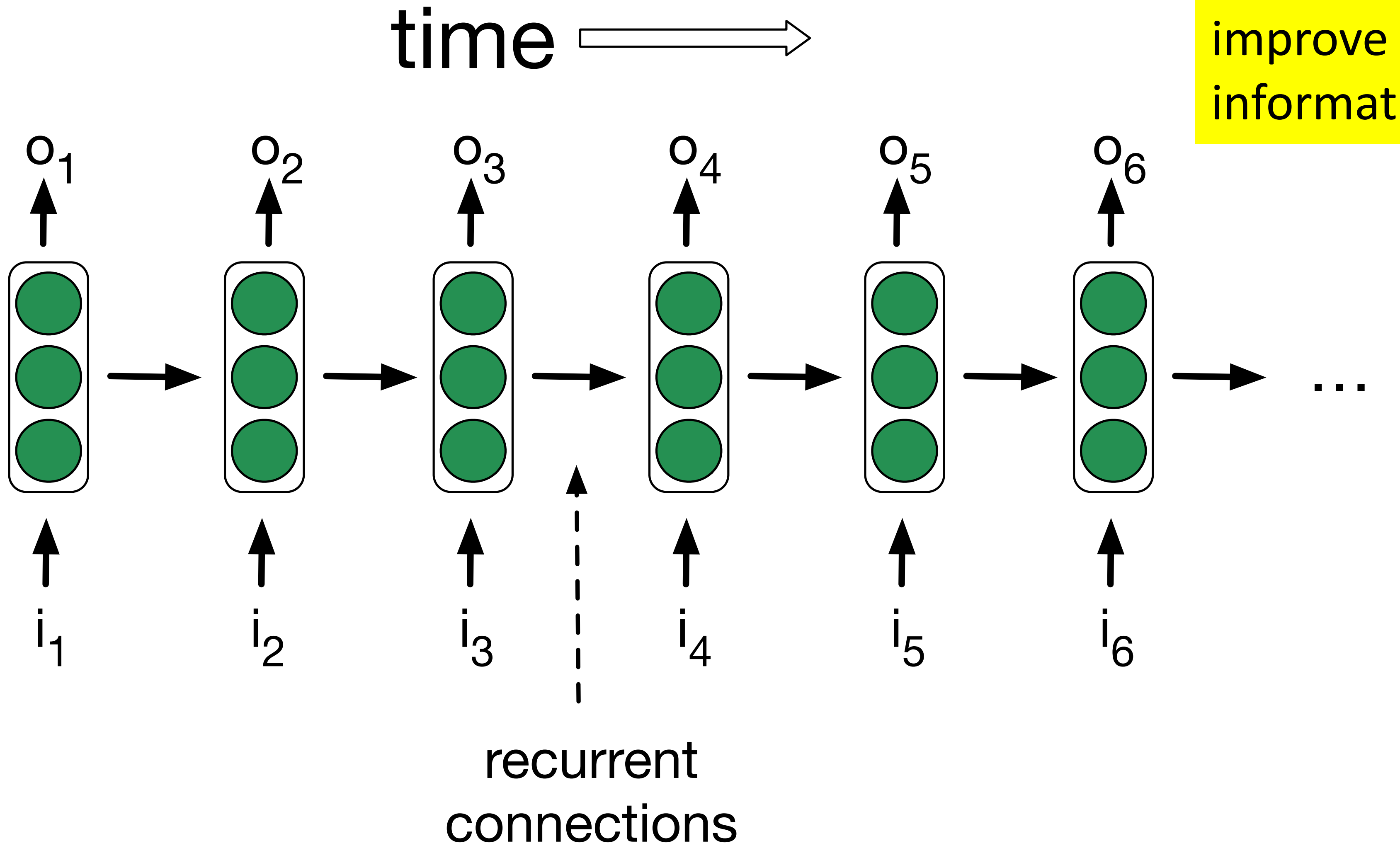
This is the "good cop" talk,
come back on Wednesday for the "bad cop"

Recurrent neural networks

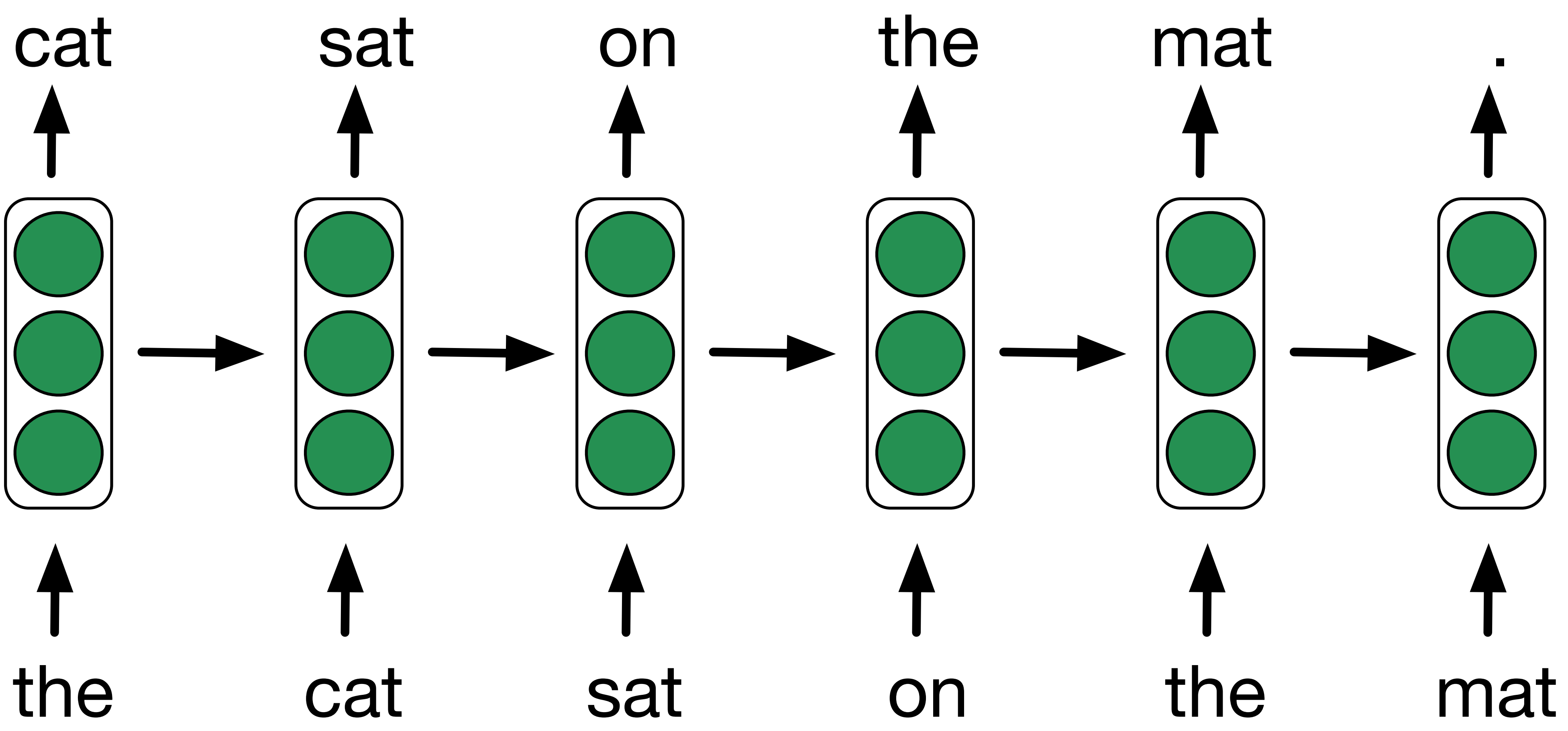


Recurrent neural networks
The "unfolded" view

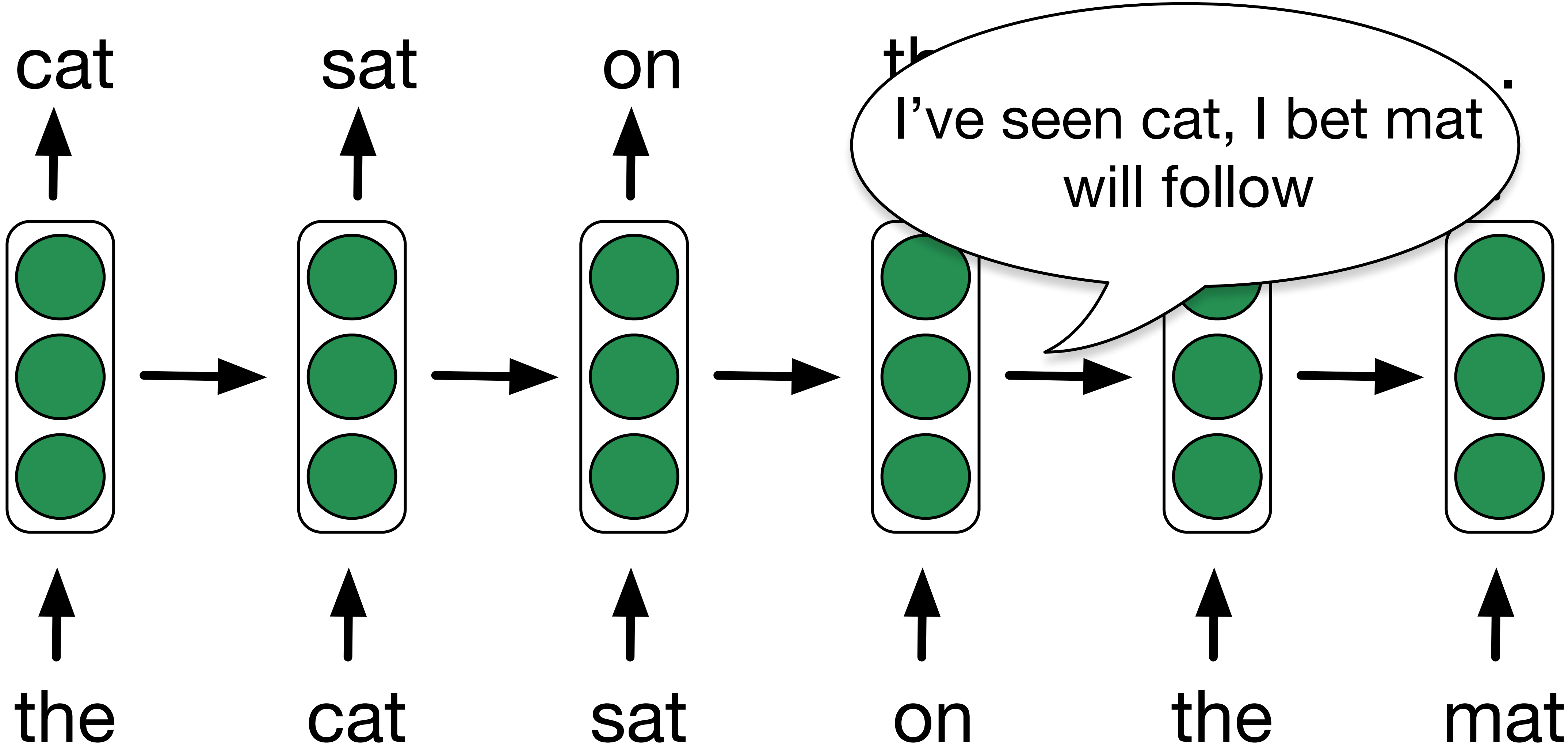
Modern RNNs (e.g., LSTMs) possess gating mechanism that improve temporal information flow



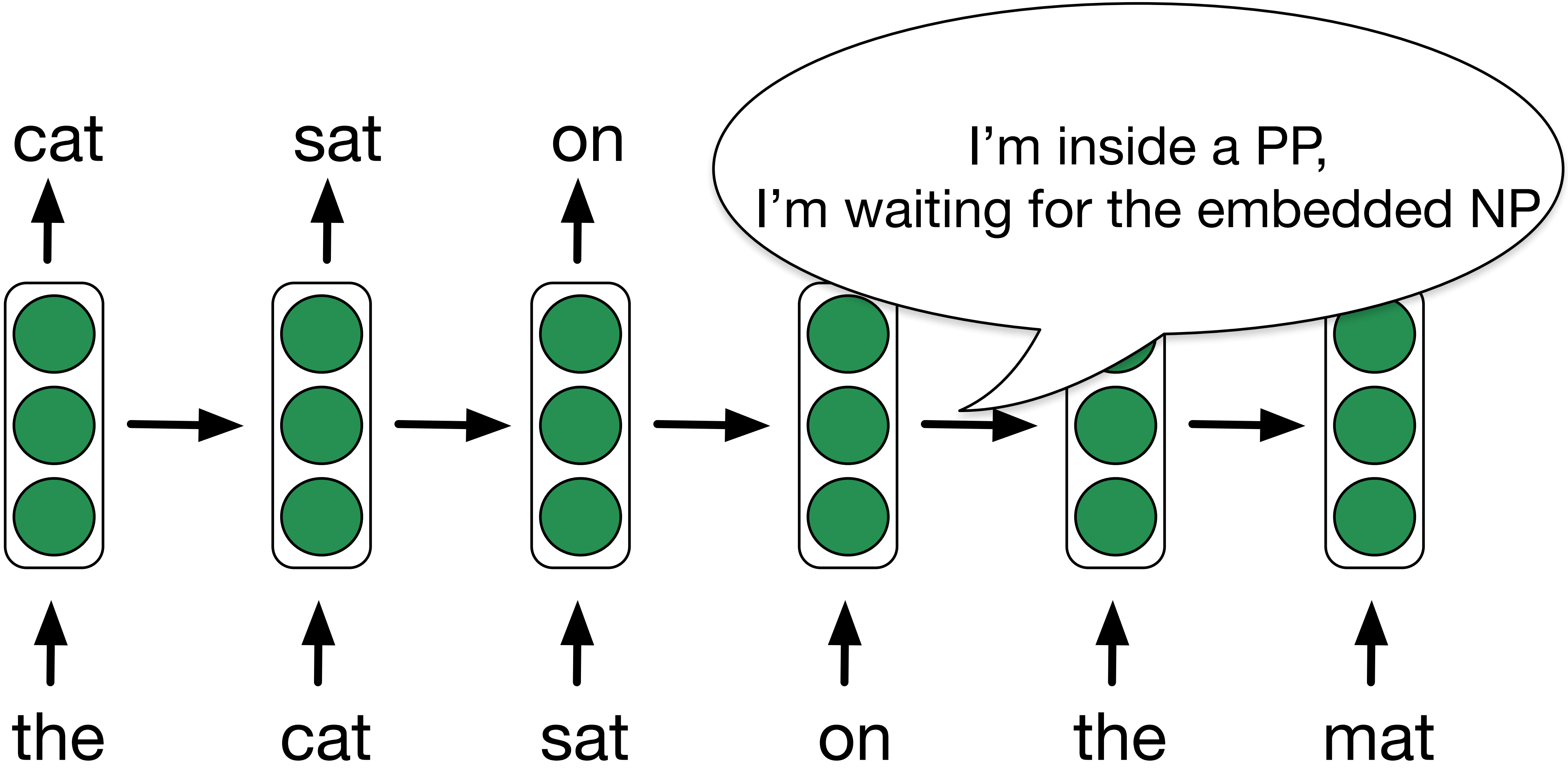
The language modeling training objective



What are LM-trained RNNs learning about language?



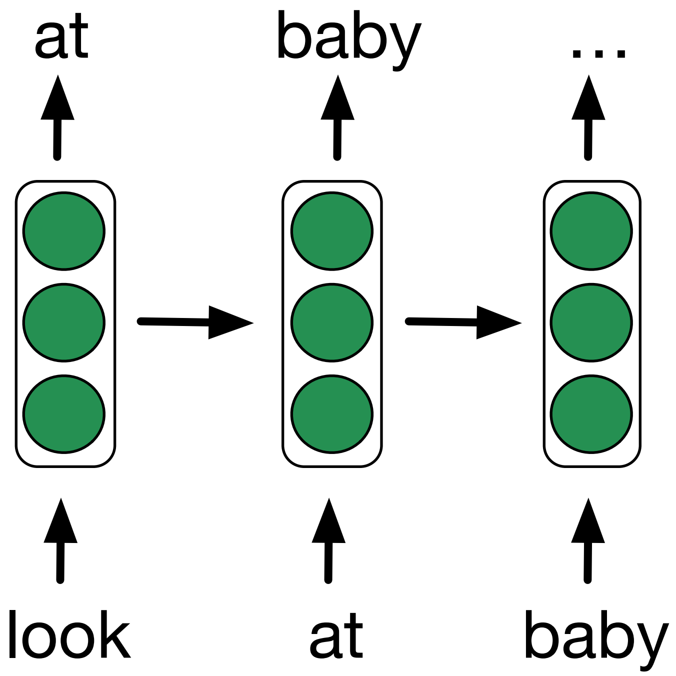
What are LM-trained RNNs learning about language?



Words as prior knowledge?

lookat...ba..by?

lookatbaby



A finite set of words as primitives?

- iPad, covfefe, hipsterical...
- pre-, hyper-, -ment, -wise, Hong Kong, hot dog, kill the breeze, spend the night, the X-er the Y-er...
- t-ə-meyŋ-ə-levt-pəɣt-ə-rkən

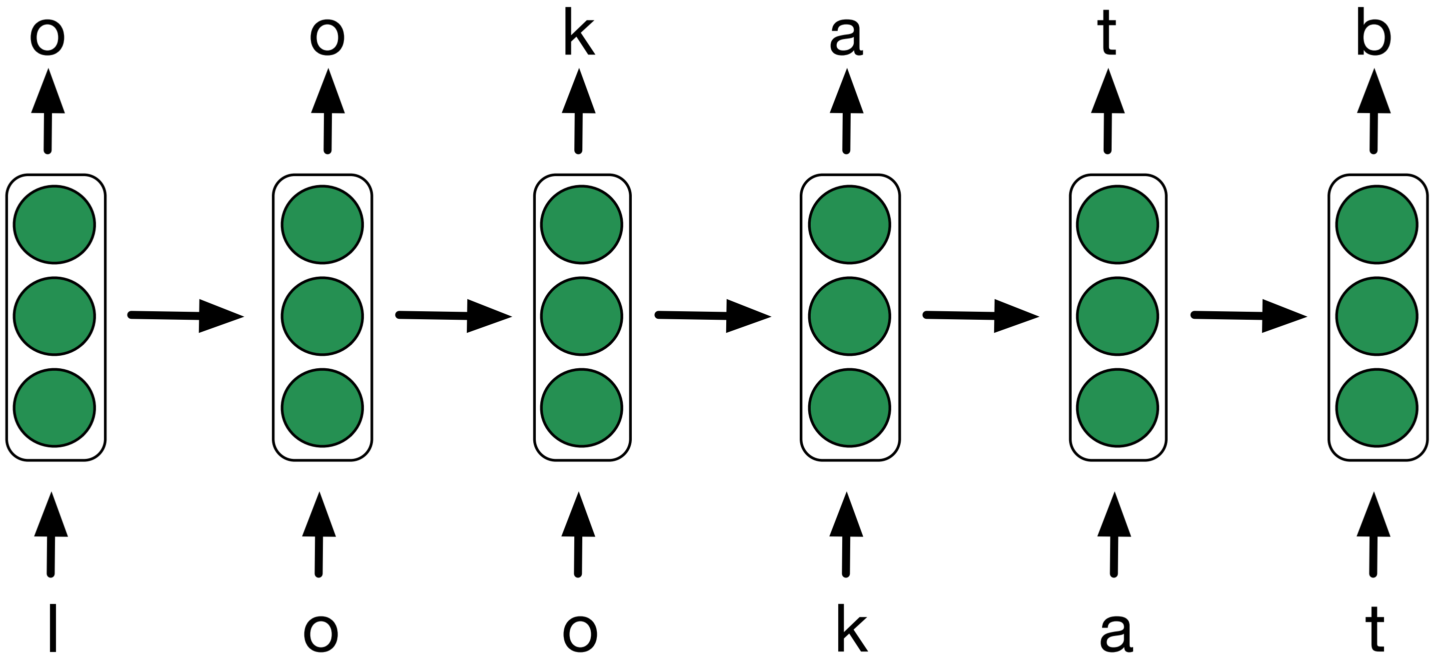
  1.SG.SUBJ-great-head-hurt-PRES.1

  "I have a fierce headache"

  (Chukchi, from Wikipedia)

Our study

- Train a character-level RNN on language model objective, feeding it input without spaces



- Test the trained RNN to probe its linguistic knowledge at different levels

Linguistic challenges for character-based RNNs

Models and training regime

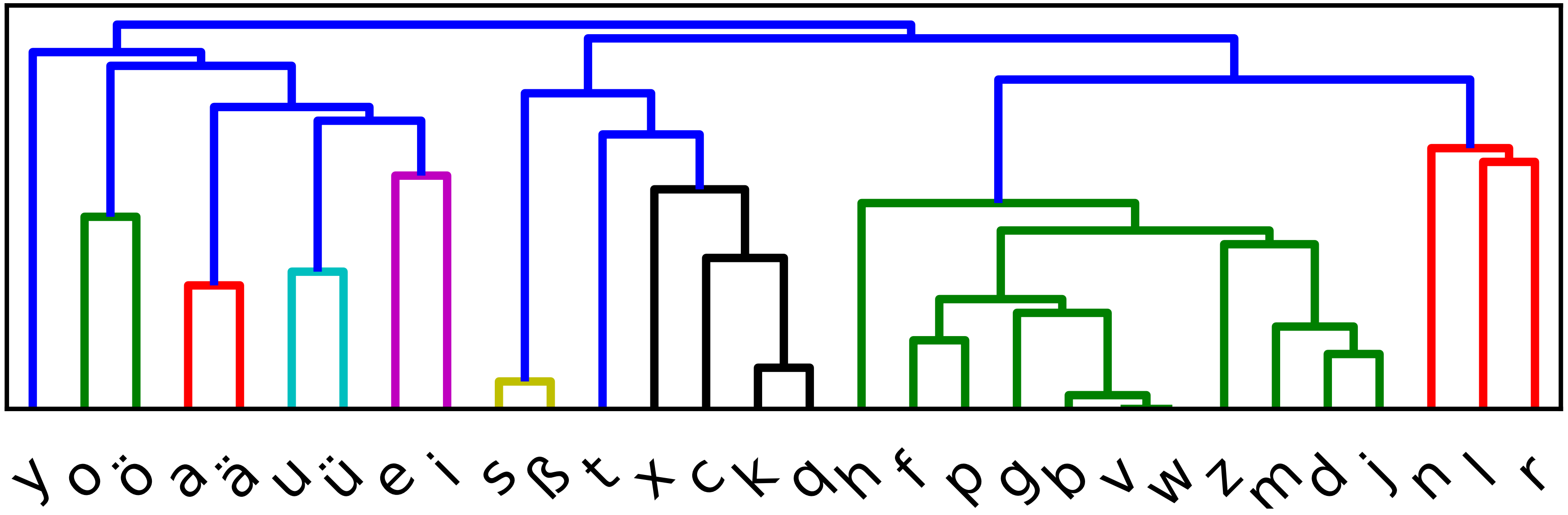
- **LSTM**: an LSTM trained at the character level on unsegmented text
- **RNN**: a "vanilla" RNN trained at the character level on unsegmented text
- **WordNLM:** an LSTM trained at the word level on segmented text

- Models trained on Wikpedia fragments containing 819M (German), 463M (Italian) and 2,333M (English) words
- Training for 72 hours
- Best hyperparameters determined on Wikipedia-based validation set
- All best models attain reasonable language modeling performance on Wikipedia-based test set

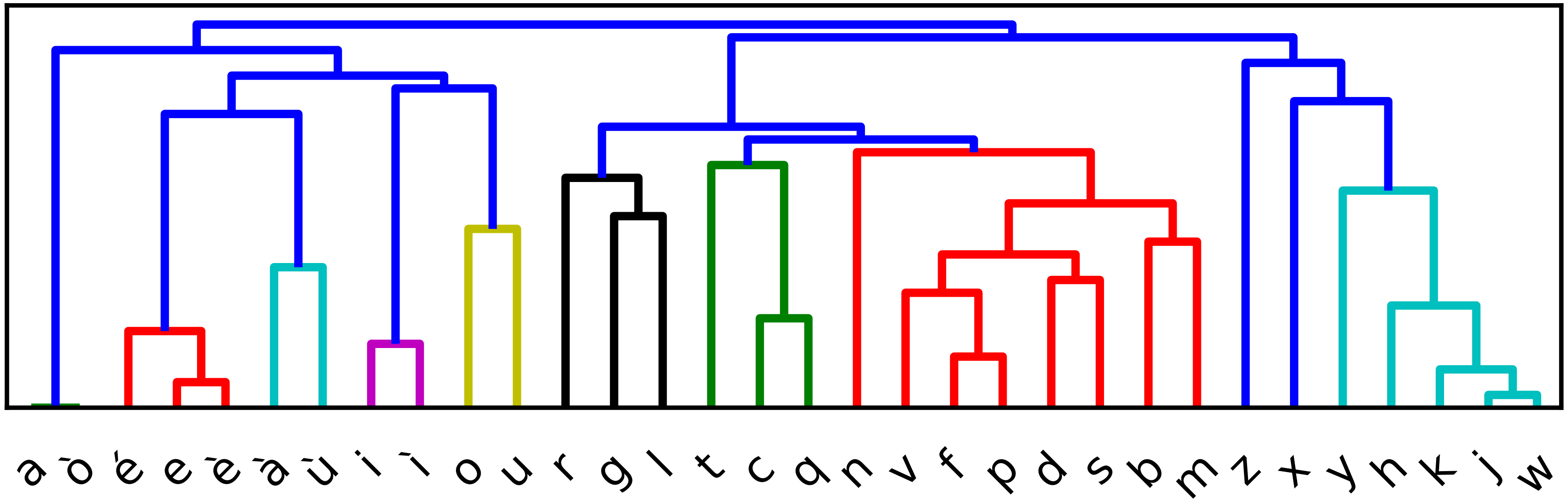
Phonology

Clustering of LSTM output character embeddings

German:



Italian:



Discovering phonotactic constraints

- Create pairs of acceptable and unacceptable letter bigrams such that:
  - They reasonably reflect the language phonology
  - They share the first letter
  - The second letter has larger unigram probability in the unacceptable bigram

tu   *td   (in Italian)

- Re-train the models on versions of the corpora with either bigram removed
- Compute probability assigned by re-trained model to acceptable vs. unacceptable bigrams

Discovering phonotactic constraints

German				*Italian*			
		LSTM	*RNN*			*LSTM*	*RNN*
bu	bt	**4.6**	0.2	bu	bd	$\approx$ **1**	≈ 0
do	dd	**1.9**	0.1	du	dt	**1.3**	≈ 0
fu	ft	**6.5**	≈ 0	fu	ft	**30.5**	≈ 0
po	pt	**6.4**	0.1	pu	pt	**6.8**	≈ 0
tu	tt	**5.4**	≈ 0	tu	td	0.2	≈ 0
zu	zt	**2.4**	0.2	vu	vd	**2.0**	≈ 0
bl	bd	0.8	0.2	zu	zt	**55.7**	≈ 0
fl	fd	**2.1**	0.8	br	bt	$\approx$ **1**	≈ 0
fr	fn	**2.7**	0.1	dr	dt	**2.5**	0.4
kl	kt	**3.8**	0.1	fr	ft	**2.9**	≈ 0
pl	pt	**2.5**	0.9	pr	pt	**5.0**	≈ 0
AM		**3.6**	0.2	AM		**10.7**	≈ 0
GM		**3.0**	0.1	GM		**3.2**	≈ 0

likelihood ratios of acceptable/unacceptable bigrams

Word segmentation

Word segmentation

- Train a classifier to predict if character is word-initial
- Features use probabilities computed by pre-trained models:
  - *surprisal*: log-probability of character given prior context
  - *entropy* of character distribution given prior context
  - context *PMI*, computed as total log-likelihood of next 20 characters considering previous 20 characters context minus unconditioned log-likelihood
- Features computed for 6-character windows, resulting in 21-feature classifier

Segmentation results
precision/recall/F1

- Wikipedia test data:

	LSTM	*RNN*	*8-grams*
English	66/60/63	63/60/61	56/51/53
German	57/52/55	53/49/51	43/36/39
Italian	64/57/60	62/57/60	48/40/44

- Brent child-directed English corpus (with re-training):

	LSTM	Bayesian
Tokens	75.3/76.6/76.0	74.9/69.8/72.3
Lexical	41.2/61.2/49.2	63.6/60.2/61.9
Boundaries	91.3/90.0/90.5	93.0/86.7/89.8

Most frequent **under**segmentations

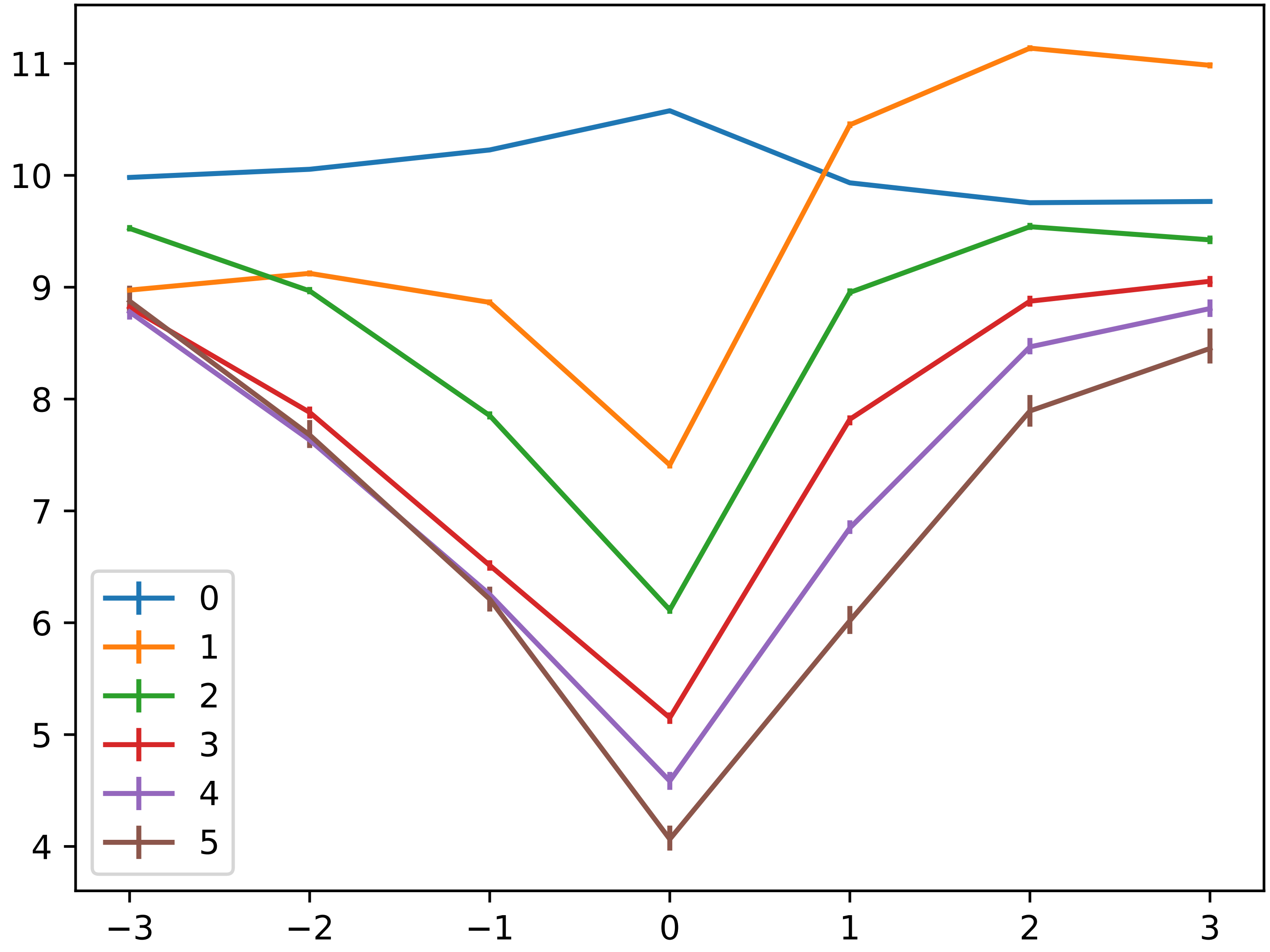
- morethan, aswellas, tothe, basedon, canbe, didnot, accordingto, oneofthe, knownas, tobe, dueto, itis, onthe, itwas, suchas, inthe, isa, asa, atthe, ofthe
- highschool, newyork, unitedstates
- useof, memberof, universityof, numberof, endof, oneof, partof

Most frequent **over**segmentations

- re, de, un, pro, en, co
- ing, ed, ly, er, al, es, ic, ers
- in, to, on, an, the, or
- man, land
- ma, ra, la, le, ta, na, ro, se

Model-based context PMI at constituent boundaries

in German
validation set



Morphological categories

Nouns vs Verbs

- 500 verbs and nouns ending in –*en* (German) and –*re* (Italian) from the training corpus

cantare	**altare**
V	**N**

- 10 verbs and nouns for training, the rest for testing
- Classifier trained on last hidden state of pre-trained language model after it reads a full word

Nouns vs Verbs: results

accuracy and std error over 100 random train/test splits

	German	*Italian*
LSTM	89.0 ($\pm$ 0.14)	95.0 ($\pm$ 0.10)
RNN	82.0 ($\pm$ 0.64)	91.9 ($\pm$ 0.24)
Autoencoder	65.1 ($\pm$ 0.22)	82.8 ($\pm$ 0.26)
WordNLM$_{subs.}$	97.4 ($\pm$ 0.05)	96.0 ($\pm$ 0.06)
WordNLM	53.5 ($\pm$ 0.18)	62.5 ($\pm$ 0.26)

Excluding OOVs

Number across German nominal classes

- Generalize number classifier across pural types
  - E.g., train on *Geschichte* / *Geschichten*, test on *Tochter* / *Töchter*

- Training classes*: -n, -s, -e*
- Test classes: *-r, Umlaut*

- Data from German Universal Dependencies treebank
- 15 singulars and plurals per training class (controlling for length)
- Test on all remaining pairs in training and test classes

Number results

accuracy and std error over 200 random train/test splits

	train classes	test classes	
	-n/-s/-e	*-r*	*Umlaut*
LSTM	77.9 (± 0.8)	88.2 (± 0.3)	52.8 (± 0.6)
RNN	70.3 (± 0.9)	81.3 (± 0.7)	53.3 (± 0.6)
Autoencoder	64.0 (± 1.0)	73.8 (± 0.6)	59.2 (± 0.5)
$\text{WordNLM}_{subs.}$	97.8 (± 0.3)	86.6 (± 0.2)	96.7 (± 0.2)
WordNLM	82.1 (± 0.1)	73.1 (± 0.1)	77.6 (± 0.1)

Excluding OOVs

Syntactic dependencies

German gender agreement

{der, die, das} sehr extrem unglaublich rote Baum

the very extremely incredibly red tree

- Nominal forms from the Universal Dependencies treebank (~ 7k stimuli)
- Pre-trained character-based model fed 3 variations of each sentence without whitespace, lower-cased, delimited by periods

.dersehrextremunglaublichrotebaum.

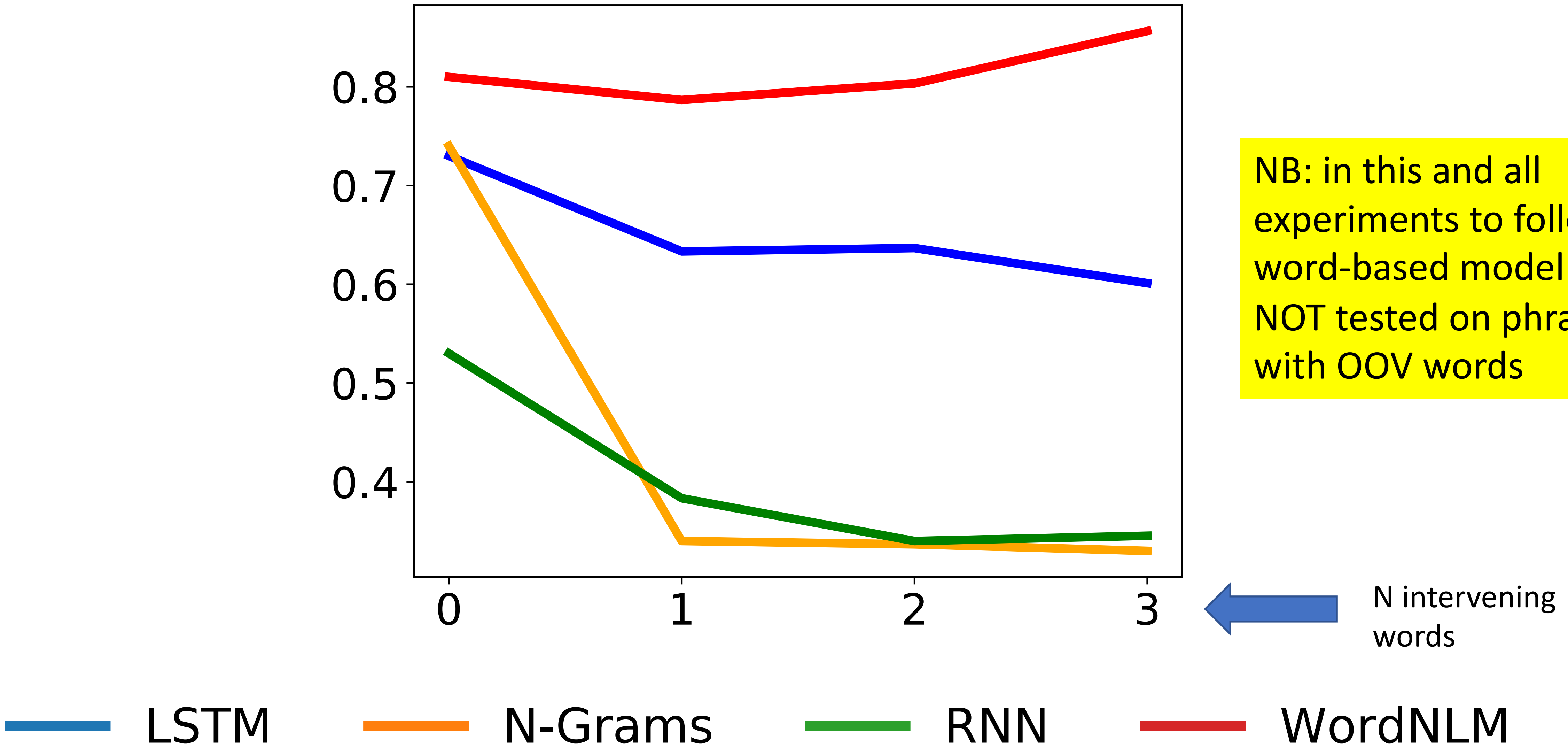
- Model must assign highest probability to version with correct case

NB: "long-distance" for word- vs character-based models

. **das** rote **baum** .

. d a **s** r o t e b a u m **.**

German gender agreement



NB: in this and all experiments to follow, word-based model is NOT tested on phrases with OOV words

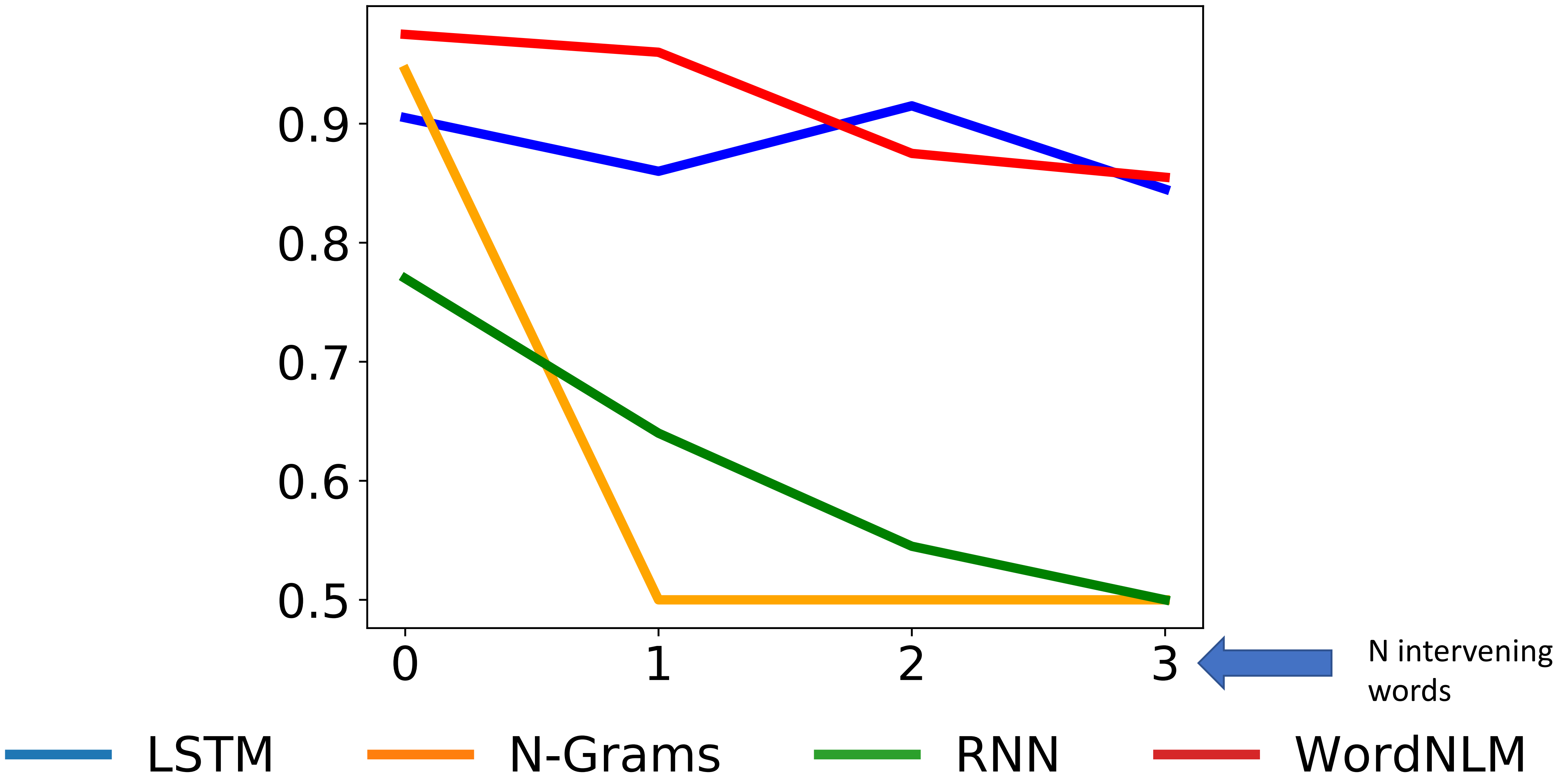
German case agreement

{dem, des}	sehr	extrem	unglaublich	roten	Baum
to/of-the	very	extremely	incredibly	red	tree (dative)

{dem, des}	sehr	extrem	unglaublich	roten	Baums
to/of-the	very	extremely	incredibly	red	tree (genitive)

- Nominal forms from the Universal Dependencies treebank, paradigms from Wiktionary (~ 9k stimuli)
- Model testing as above

German case agreement

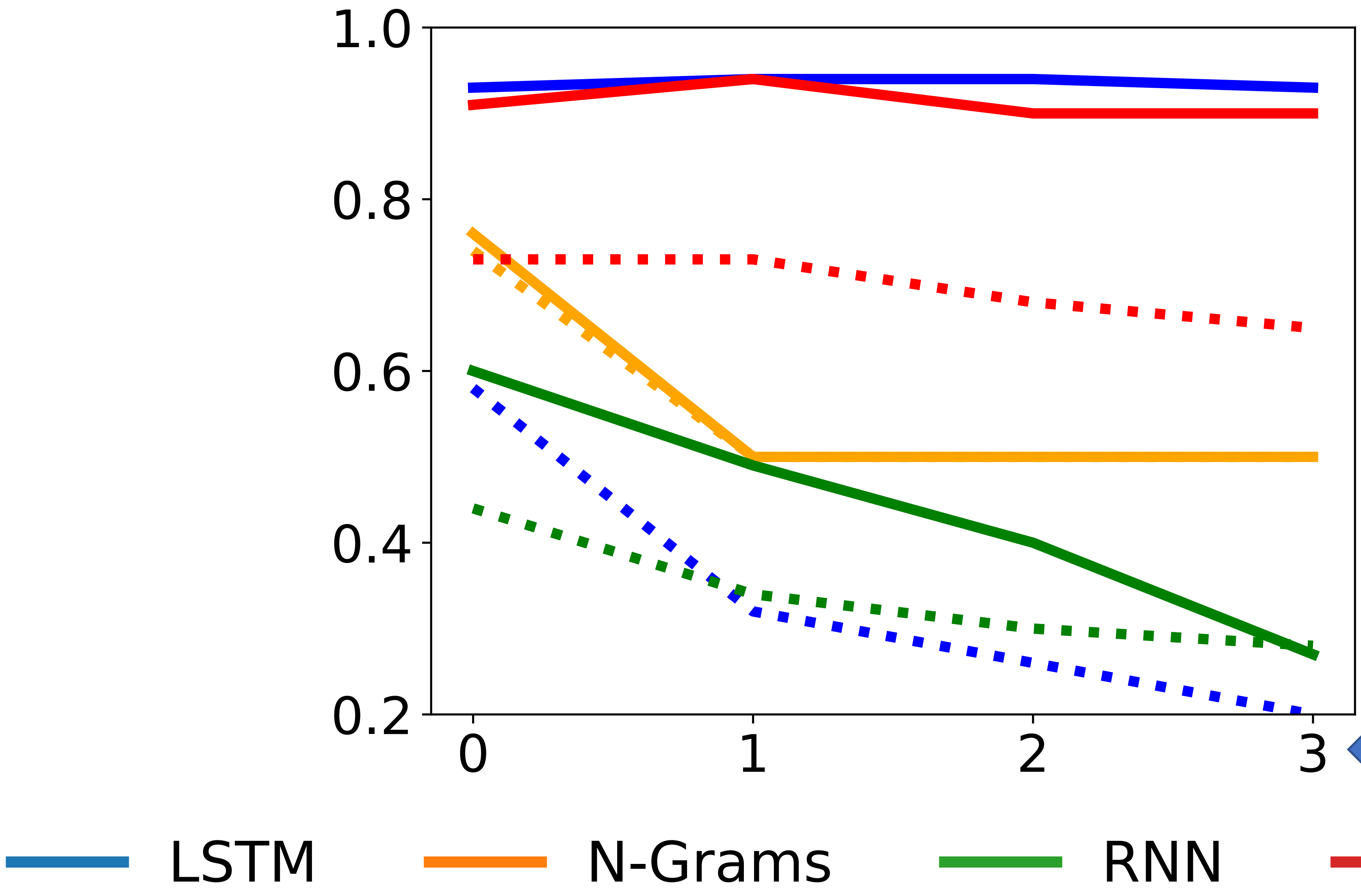


German case subcategorization

mit der sehr extrem unglaublich {rote, <u>roten</u>}
with the very extremely incredibly red one (dat.)

- Embedded in sentences for more natural context, extracted from Universal Dependencies treebank (~1.6k stimuli)
- Model testing as above

German case subcategorization



dashed lines:
der sehr {rote, roten}

N intervening words

Italian article-noun gender agreement

il congeniale {candidato, candidata}
the (m.) congenial candidate

la congeniale {candidato, candidata}
the (f.) congenial candidate

- ~30k stimuli, selected based on corpus frequency and checked for semantic well-formedness
- No adjective-noun combination attested in training corpus
- Model testing as in German

Italian article-adjective gender agreement

il meno {<u>alieno</u>, aliena}
the (m.) less alien

la meno {alieno, <u>aliena</u>}
the (f.) less alien

- ~200 stimuli, with similar selection conditions as above

Italian article-adjective number agreement

la meno {aliena, aliene}

the (s.) less alien

le meno {aliena, aliene}

the (p.) less alien

- ~200 stimuli, with similar selection conditions as above

Italian syntactic dependency results

	CNLM		WordNLM
	LSTM	*RNN*	
Noun Gender	93.1	79.2	97.4
Adj. Gender	99.5	98.9	99.5
Adj. Number	99.0	84.5	100.0

Semantics

Microsoft Research Sentence Completion

Zweig and Burgess 2011

Was she his ____, his friend, or his mistress?

client
musings
discomfiture
choice
opportunity

- ~1k sentences from Sherlock Holmes novels
- Chosen to be hard for language models

Microsoft Research Sentence Completion

- Evaluate pre-trained models by feeding sentence with each variant, picking most likely one as model guess

- Big gap between Wikipedia and Sherlock Holmes

- Also re-trained models with provided training data from 19^{th} century novels (~ 41.5M words)
  - No further hyperparameter tuning

MSR Sentence Completion: Results

(accuracies)

From the literature

Our models with out/in domain training

LSTM	34.1/59.0
RNN	24.3/24.0
WordNLM	37.1/63.3

KN5	40.0	Skipgram	48.0
Word RNN	45.0	Skipgram + RNNs	58.9
Word LSTM	56.0	PMI	61.4
LdTreeLSTM	60.7	Context Embeddings	65.1

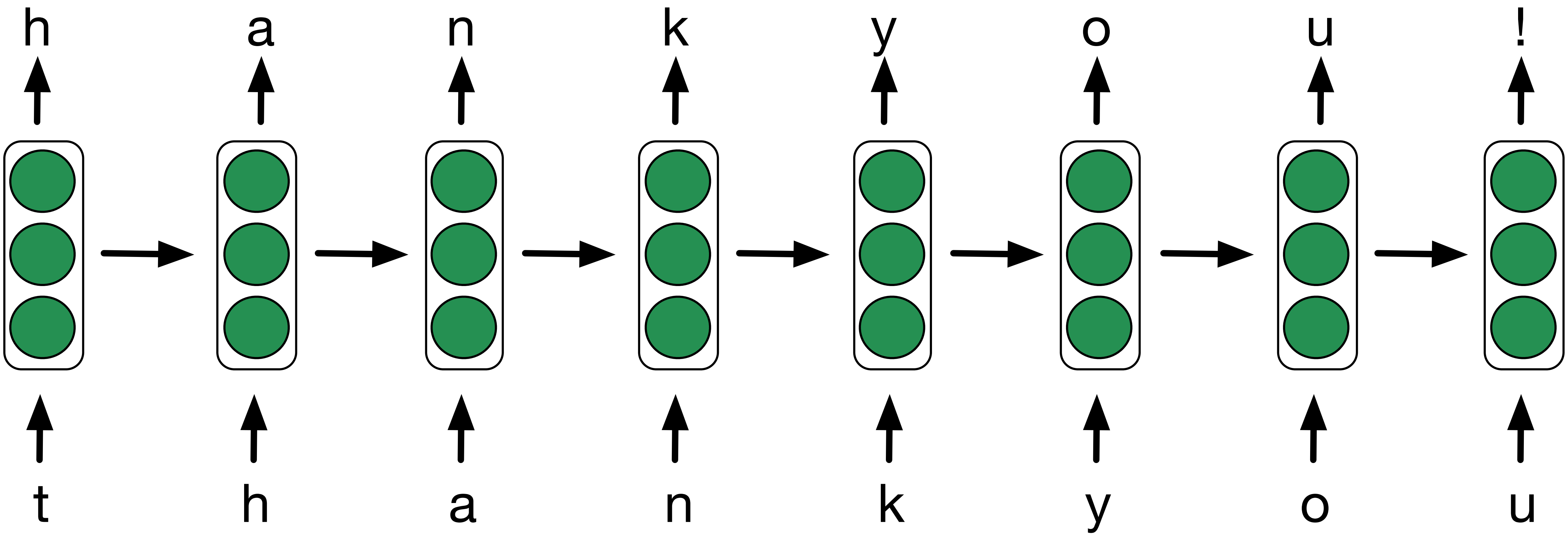
What have we learned?

Summary

- LSTMs trained to predict next character in unsegmented large corpus implicitly discover phonological, lexical, morphological, syntactic, semantic generalizations
- Systematically better than n-gram controls (thus, not only relying on shallow co-occurrence statistics)
- Not as good as word-trained model, but not much worse either, suggesting words are helfpul prior but not fundamental
- LSTMs generally outperform RNNs: better (or faster) learners in character domain, where information has to be carried through longer stretches of time

Where next?

- How much does training corpus size matter?
  - See bad-cop talk on Wednesday
- How is lexical knowledge implicitly encoded in the weights of the character-based LSTM language model?
- Can we use character-based models for better accounts of domains where word-centric view fails?
  - Polysynthetic, agglutinative languages
  - Morphemes, compounds, idioms, constructions...


h a n k y o u !
t h a n k y o u