Deep neural networks in language models

Szymon Łęski

SAMSUNG

Language models

- Language model assigns probability to a sequence of words

The cat is walking in the bedroom.

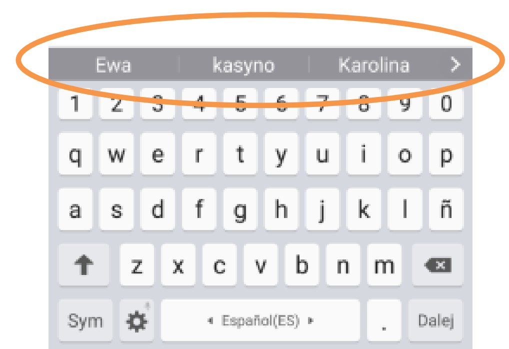
The elephant is walking in the bedroom.

Walking cat is bedroom the the in.

- Applications: choosing the right word/sentence from a list of candidates

Language models – applications

- Speech recognition: *hard rock ↔ card rock*
- Text or handwriting recognition (OCR):
  you are right ↔ you ave right
- Machine translation:
  which translation is better?
- Typo correction
  Are you happy now?
  Happy now → Happy new
- Text input prediction
- …



Language models – how to build

The cat is walking in the bedroom.

- Counting sentences in a corpus?

The cat is walking in the bedroom.	*~ 76400*
Three days ago I saw my cat walking in the bedroom.	*0*

- The sentence for which we evaluate the model is most likely not present in the corpus

Language models – how to build

- $p(w_1, w_2, \dots w_N) =$

  $p(w_1) \cdot p(w_2|w_1) \cdot p(w_3|w_1, w_2) \cdot \cdots \cdot p(w_N|w_1, w_2, \dots w_{N-1})$
- Close words are more statistically dependent
- Unigram model: $\approx p(w_1) \cdot p(w_2) \cdot p(w_3) \cdot \cdots \cdot p(w_N)$
- Bigram model: $\approx p(w_1) \cdot p(w_2|w_1) \cdot p(w_3|w_2) \cdot \cdots \cdot p(w_N|w_{N-1})$
- Probabilities estimated *from corpus:*

  $p(w_3|w_1, w_2) = n(w_1, w_2, w_3)/n(w_1, w_2)$

Three days ago I saw my cat walking in the bedroom.

n-grams – example

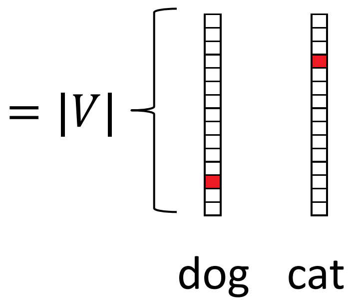
- serve as the inspiration 1390
- serve as the input 1323
- serve as the information 838
- serve as the instrument 614
- serve as the industry 607
- …
- serve as the indispensible 40

Google Research Blog

n-gram language models

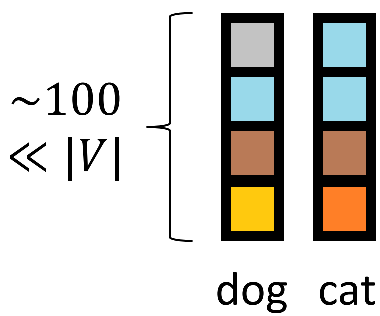
- Pros: good scalability, speed
- Cons:
  - sparsity / short context:
    for larger *n* many *n*-grams will not be present in the corpus (*serve as the insight*)
  - word similarity is not taken into account:
    e.g. went $\leftrightarrow$ gone; cat $\leftrightarrow$ dog

Word representations



The cat is walking in the bedroom.

A dog was running in a room.

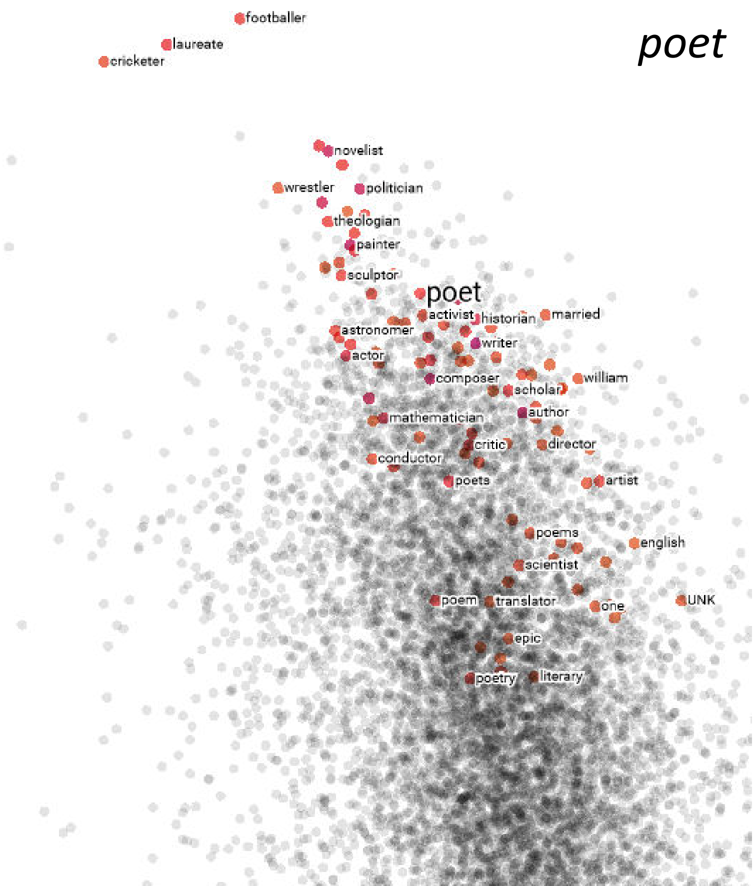


The cat is running in a room.
A dog is walking in a bedrom.
The dog was walking in the room.
...

Bengio et al. 2003

word2vec

poet



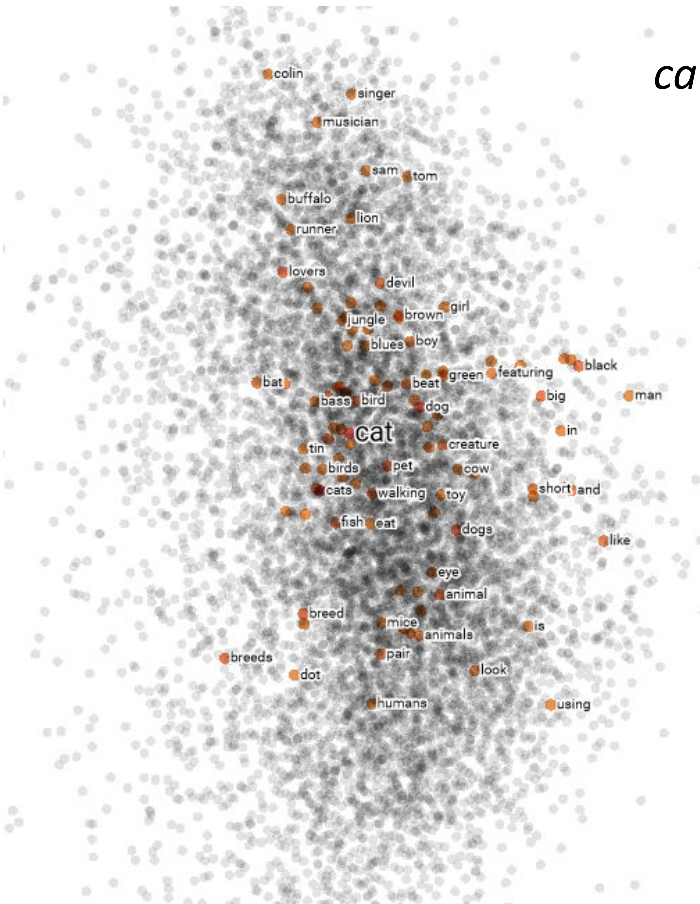
Nearest points in the original space:

writer	0.171
painter	0.202
author	0.206
composer	0.207
novelist	0.214
politician	0.222
philosopher	0.230
playwright	0.233
journalist	0.247
historian	0.258
mathematician	0.264
actor	0.265
musician	0.280
actress	0.286
poets	0.287
laureate	0.289
statesman	0.290

http://projector.tensorflow.org/

word2vec

cat

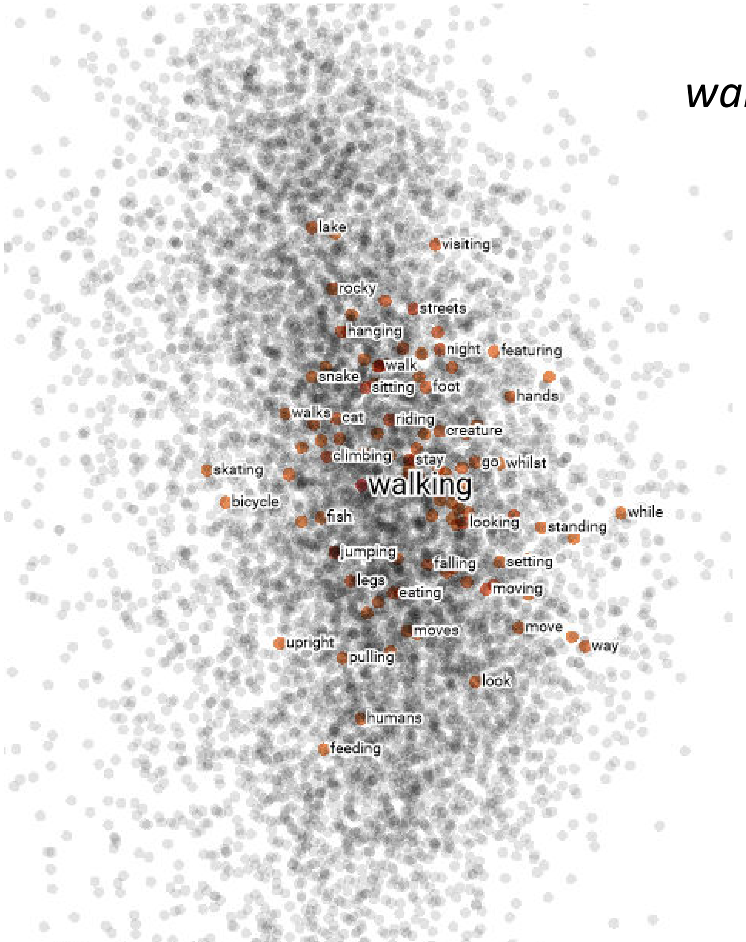


Nearest points in the original space:

cats	0.290
dog	0.352
dogs	0.388
pet	0.395
lovers	0.401
animal	0.416
breed	0.416
black	0.418
fish	0.423
breeds	0.425
big	0.428
walking	0.428
bat	0.436
bird	0.438
like	0.444
devil	0.447
hat	0.448

http://projector.tensorflow.org/

word2vec



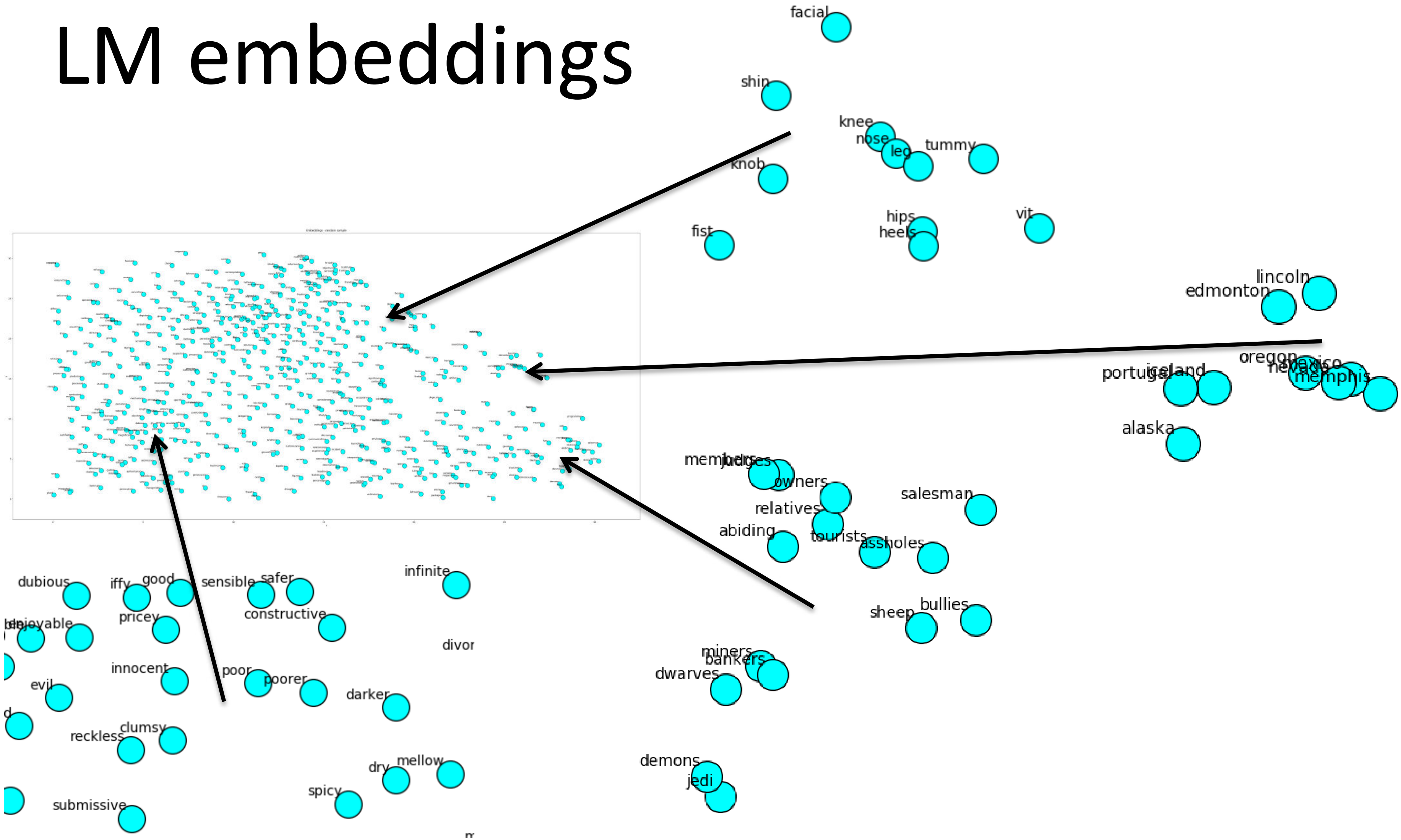
walking

Nearest points in the original space:

walk	0.349
jumping	0.386
sitting	0.386
streets	0.395
moving	0.399
riding	0.406
climbing	0.406
looking	0.407
stay	0.418
eating	0.426
legs	0.427
night	0.427
falling	0.428
cat	0.428
nearby	0.428
going	0.430
bed	0.430

http://projector.tensorflow.org/

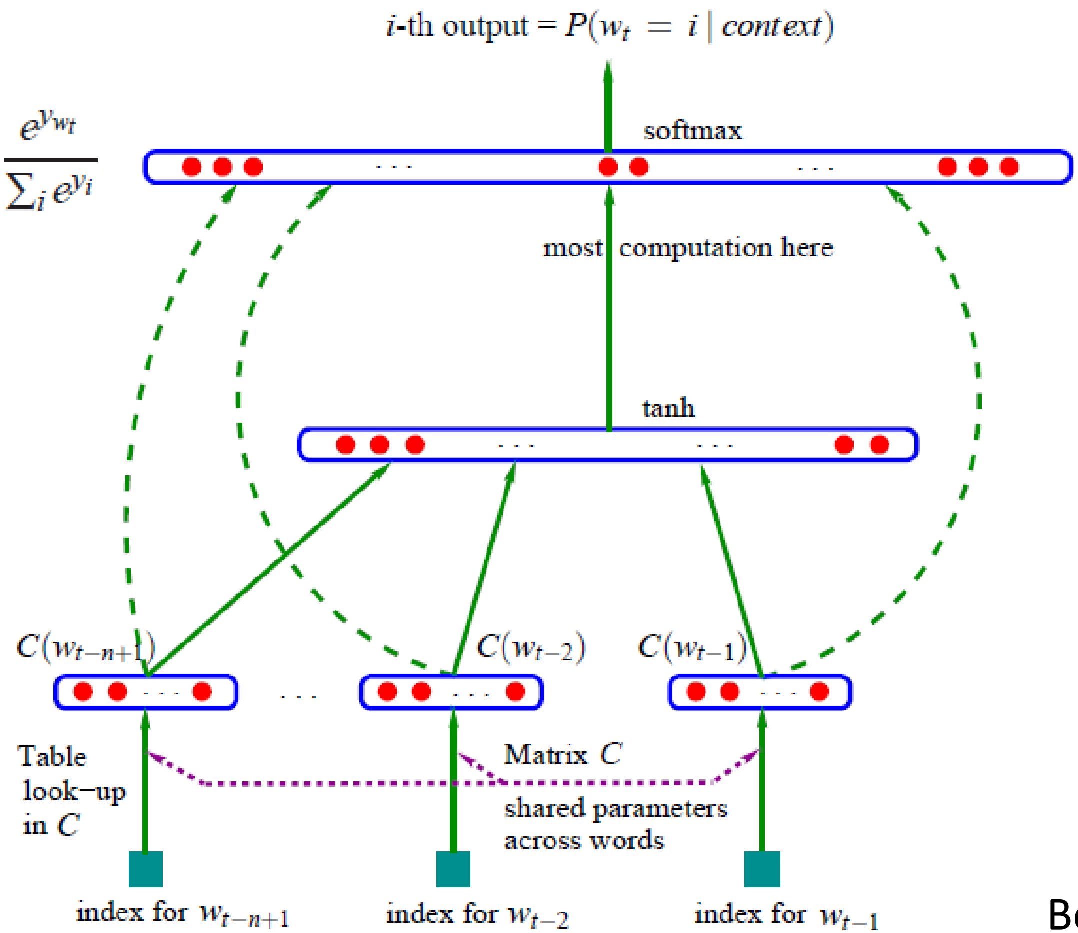
LM embeddings



NN LM

$$\hat{P}(w_t | w_{t-1}, \cdots w_{t-n+1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

$C: |V| \times m$



Bengio et al. 2003

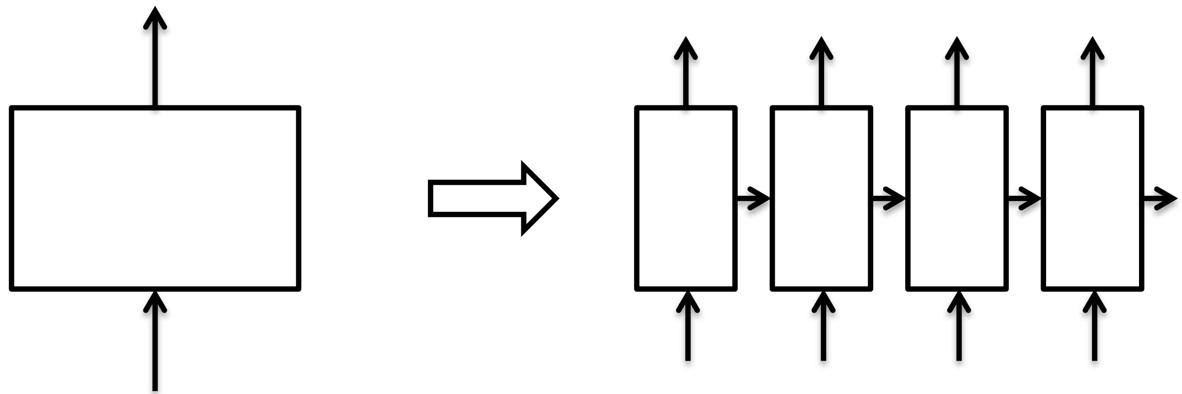
Training NN LM

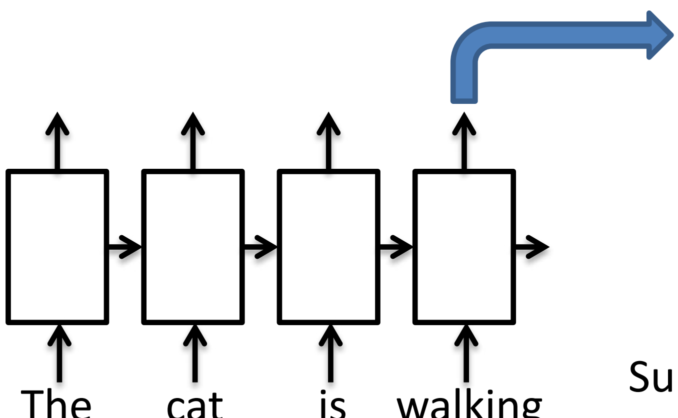
- We want the model to learn the probability distribution $p(w_t|w_{t-1}, w_{t-2}, \dots w_{t-N+1})$
- Backpropagation
- Regularization

n-gram language models

- Pros: good scalability, speed
- Cons:
  - sparsity / **short context:**
    for larger *n* many *n*-grams will not be present in the corpus
  - word similarity is not taken into account:
    e.g. went ↔ gone; cat ↔ dog

RNN-based language models

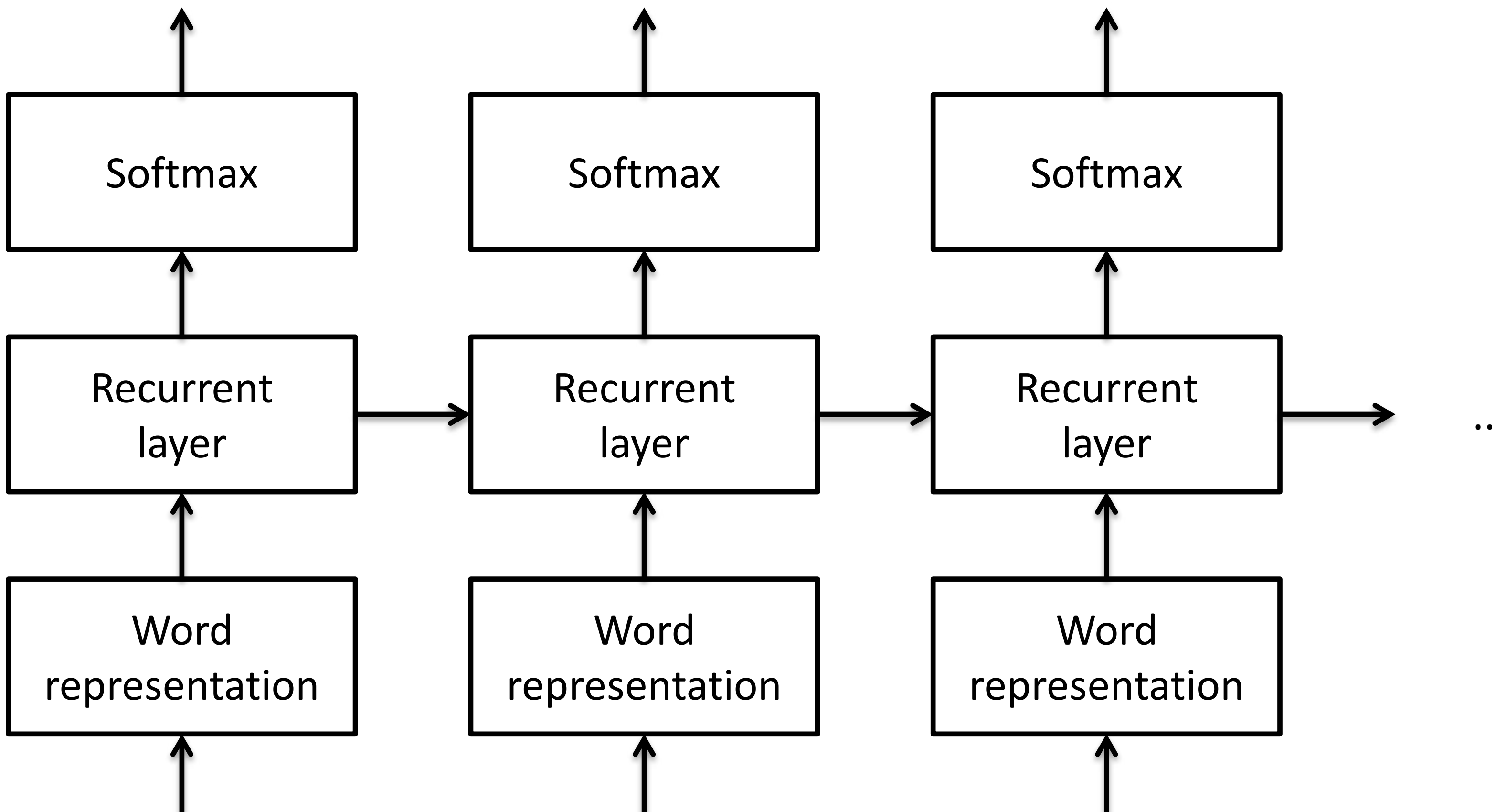




Probability distribution of the next word

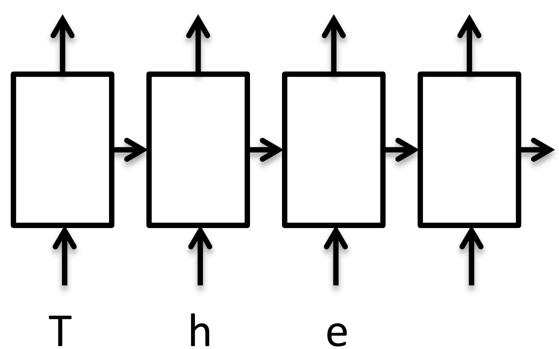
Subsequent words are RNN input in consecutive time steps

RNN-based language models



RNN-based language models (LSTM)

- Can capture long-term dependencies (not limited to *n* words of context)
- Can work on the level of words, morphemes, characters...



```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, ?? and the fact that any U affine, see Morphisms, Lemma ??. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma ?? we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longrightarrow (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example ??. It may replace S by $X_{spaces,étale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma ??. Namely, by Lemma ?? we see that R is geometrically regular over S.

Andrej Karpathy blog

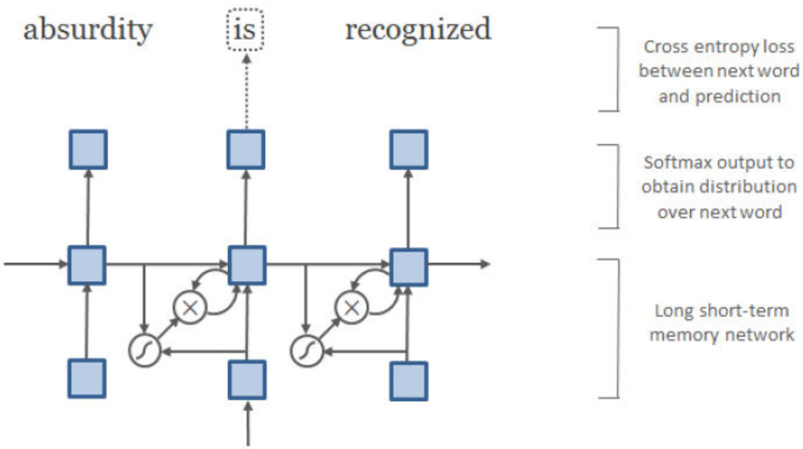
$< S >$ With even more new technologies coming onto the market quickly during the past three years , an increasing number of companies now must tackle the ever-changing and ever-changing environmental challenges online . $< S >$ Check back for updates on this breaking news story . $< S >$ About 800 people gathered at Hever Castle on Long Beach from noon to 2pm , three to four times that of the funeral cortège . $< S >$ We are aware of written instructions from the copyright holder not to , in any way , mention Rosenberg 's negative comments if they are relevant as indicated in the documents , " eBay said in a statement . $< S >$ It is now known that coffee and cacao products can do no harm on the body . $< S >$ Yuri Zhirkov was in attendance at the Stamford Bridge at the start of the second half but neither Drogba nor Malouda was able to push on through the Barcelona defence .

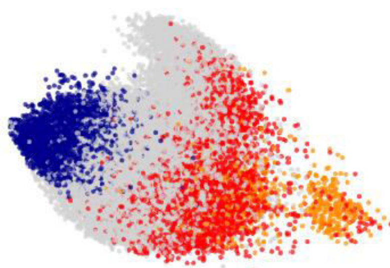
Jozefowicz et al. 2016

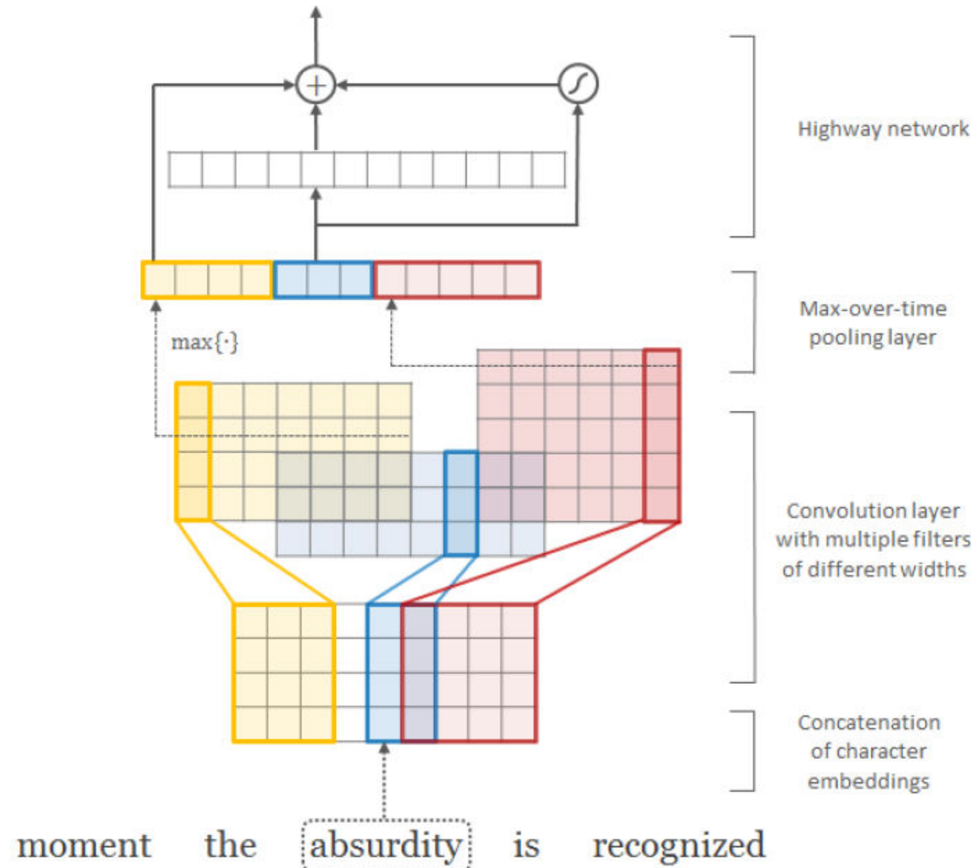
RNN LM – advanced topics

- Character input
- Computing embedding on the fly
- Improving word embeddings using morphological constraints
- Pointer-sentinel model
- Hierarchical character-level model

Character input







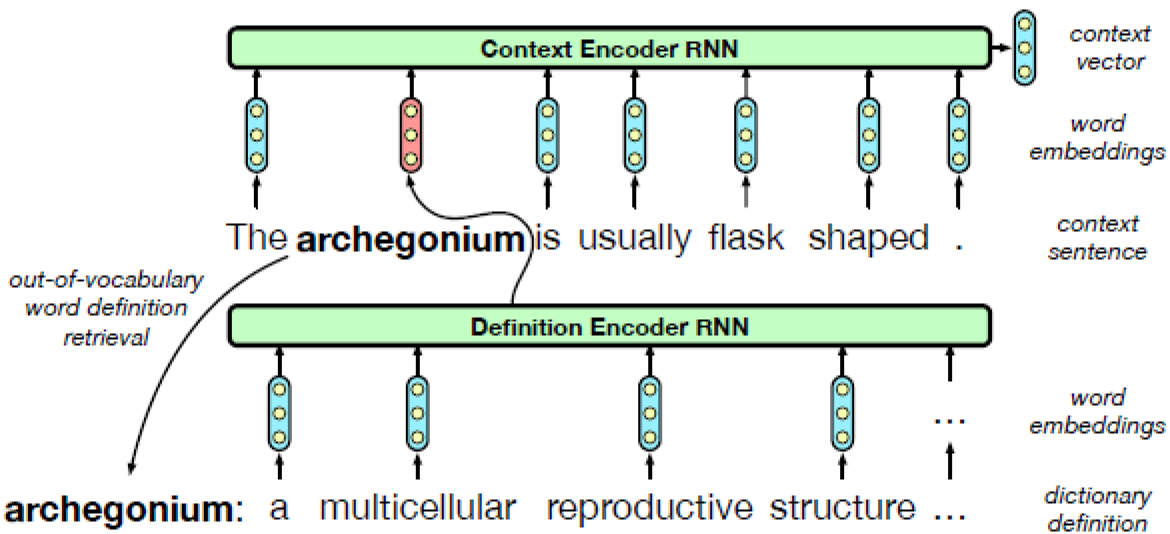
Kim et al. 2016

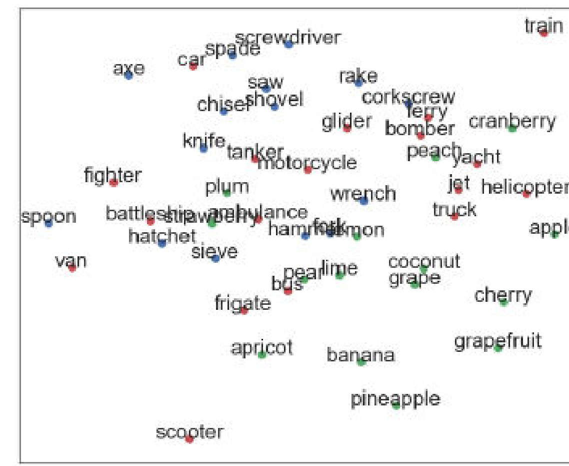
Character input

	In Vocabulary					Out-of-Vocabulary		
	while	*his*	*you*	*richard*	*trading*	*computer-aided*	*misinformed*	*looooook*
LSTM-Word	*although* *letting* *though* *minute*	*your* *her* *my* *their*	*conservatives* *we* *guys* *i*	*jonathan* *robert* *neil* *nancy*	*advertised* *advertising* *turnover* *turnover*	– – – –	– – – –	– – – –
LSTM-Char (before highway)	*chile* *whole* *meanwhile* *white*	*this* *hhs* *is* *has*	*your* *young* *four* *youth*	*hard* *rich* *richer* *richter*	*heading* *training* *reading* *leading*	*computer-guided* *computerized* *disk-drive* *computer*	*informed* *performed* *transformed* *inform*	*look* *cook* *looks* *shook*
LSTM-Char (after highway)	*meanwhile* *whole* *though* *nevertheless*	*hhs* *this* *their* *your*	*we* *your* *doug* *i*	*eduard* *gerard* *edward* *carl*	*trade* *training* *traded* *trader*	*computer-guided* *computer-driven* *computerized* *computer*	*informed* *performed* *outperformed* *transformed*	*look* *looks* *looked* *looking*

Kim et al. 2016

Calculating embeddings on the fly

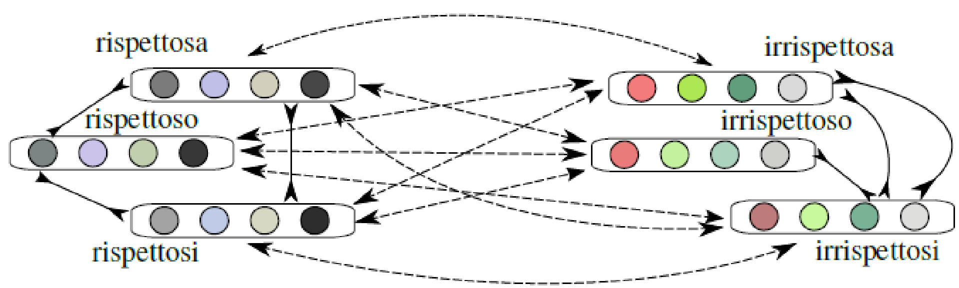




Bahdanau et al. 2017

Improving word embeddings

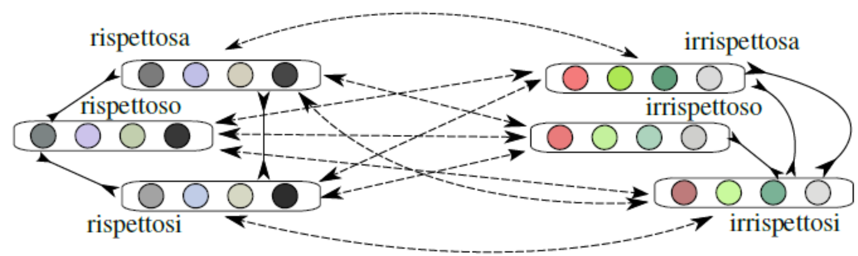
- Problems with word representation in morphologically rich languages:
  – learning representation for infrequent words (each form – separate token)
  – morfology and word meaning



Vulić et al. 2017

Improving word embeddings

- Improving word embeddings with attract-repel pairs
  discuss ⟷ discussed, laugh ⟷ laughing,
  dressed ⟷ undressed, formality ⟷ informality
- Pairs found using morphological rules, e.g. suffix *s, ed, ing;* prefix *dis, un, in, anti*



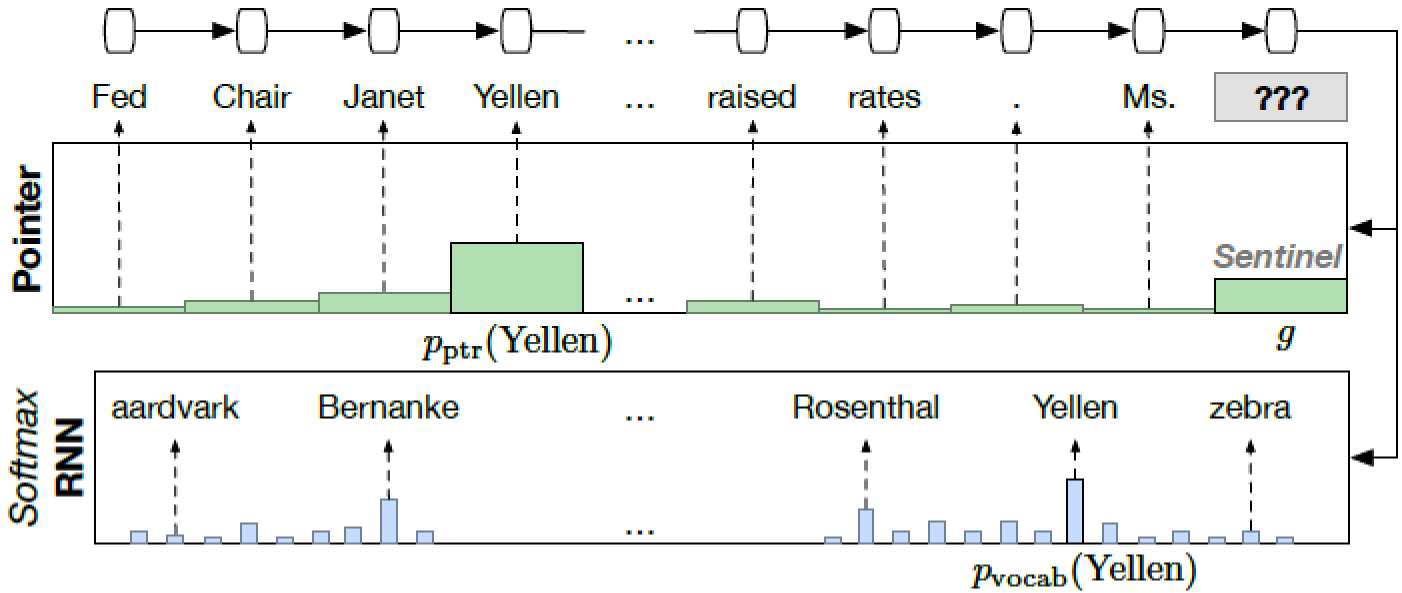
Vulić et al. 2017

Improving word embeddings

en_expensive	de_teure	it_costoso	en_slow	de_langsam	it_lento	en_book	de_buch	it_libro
costly	teuren	dispendioso	fast	allmählich	lentissimo	books	sachbuch	romanzo
costlier	kostspielige	remunerativo	slowness	rasch	lenta	memoir	buches	racconto
cheaper	aufwändige	redditizio	slower	gemächlich	inesorabile	novel	romandebüt	volumetto
prohibitively	kostenintensive	rischioso	slowed	schnell	rapidissimo	storybooks	büchlein	saggio
pricey	aufwendige	costosa	slowing	explosionsartig	graduale	blurb	pamphlet	ecclesiaste
expensiveness	teures	costosa	slowing	langsamer	lenti	booked	bücher	libri
costly	teuren	costose	slowed	langsames	lente	rebook	büch	libra
costlier	teurem	costosi	slowness	langsame	lenta	booking	büche	librare
ruinously	teurer	dispendioso	slows	langsamem	veloce	rebooked	büches	libre
unaffordable	teurerer	dispendiose	idle	langsamen	rapido	books	büchen	librano

Vulić et al. 2017

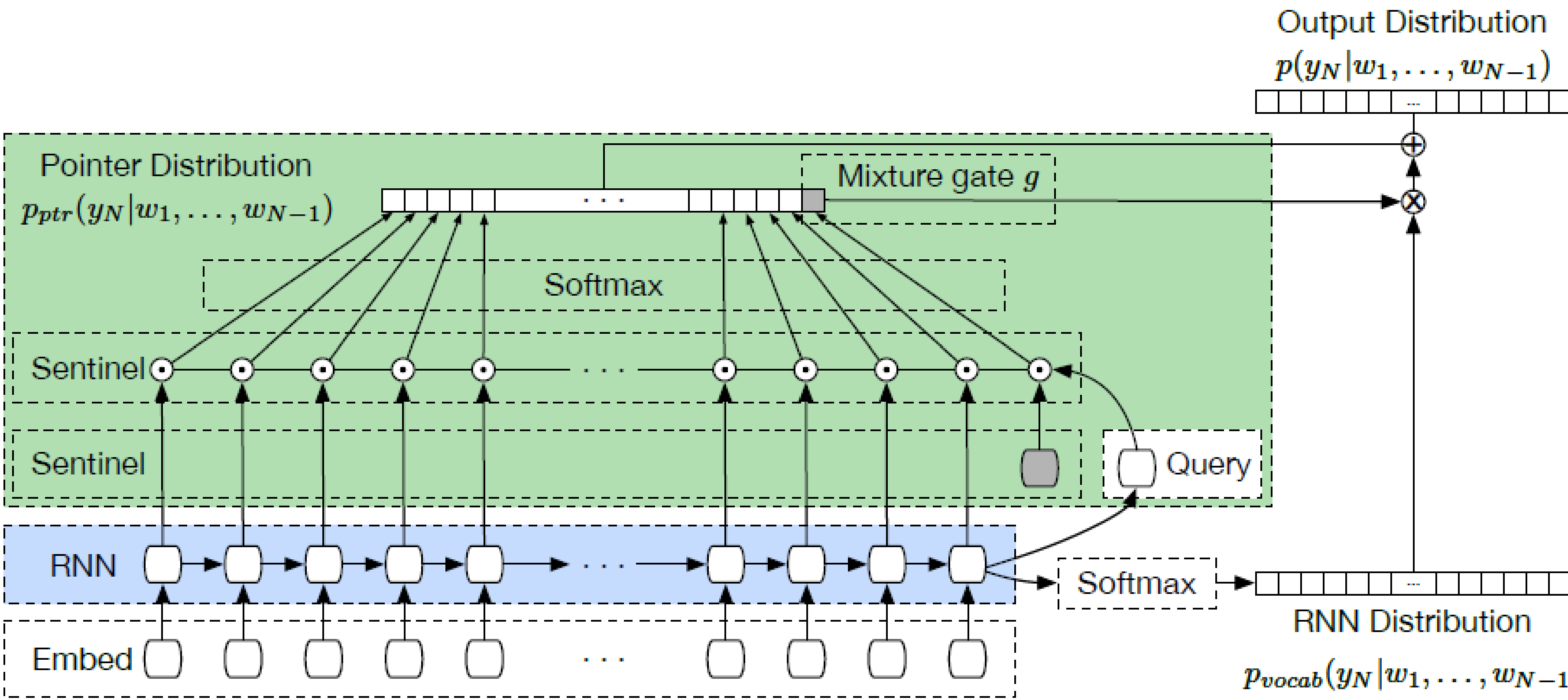
Pointer-sentinel model



$$p(\text{Yellen}) = g\, p_{\text{vocab}}(\text{Yellen}) + (1-g)\, p_{\text{ptr}}(\text{Yellen})$$

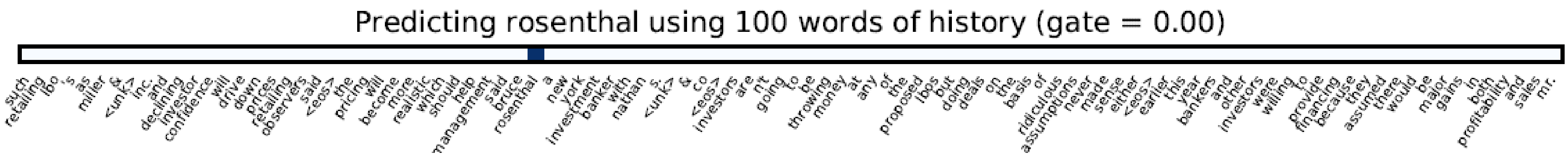
Merity et al. 2016

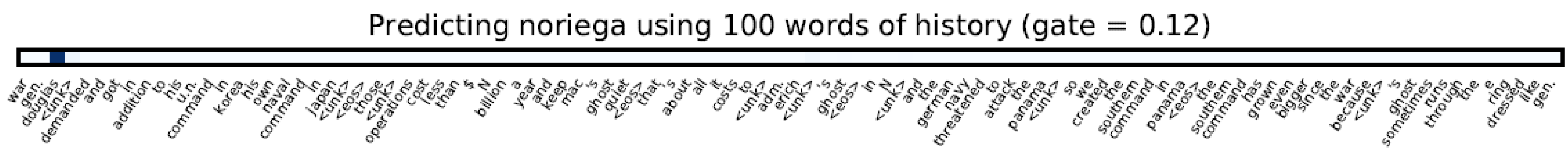
Pointer-sentinel model



Merity et al. 2016

Pointer-sentinel model

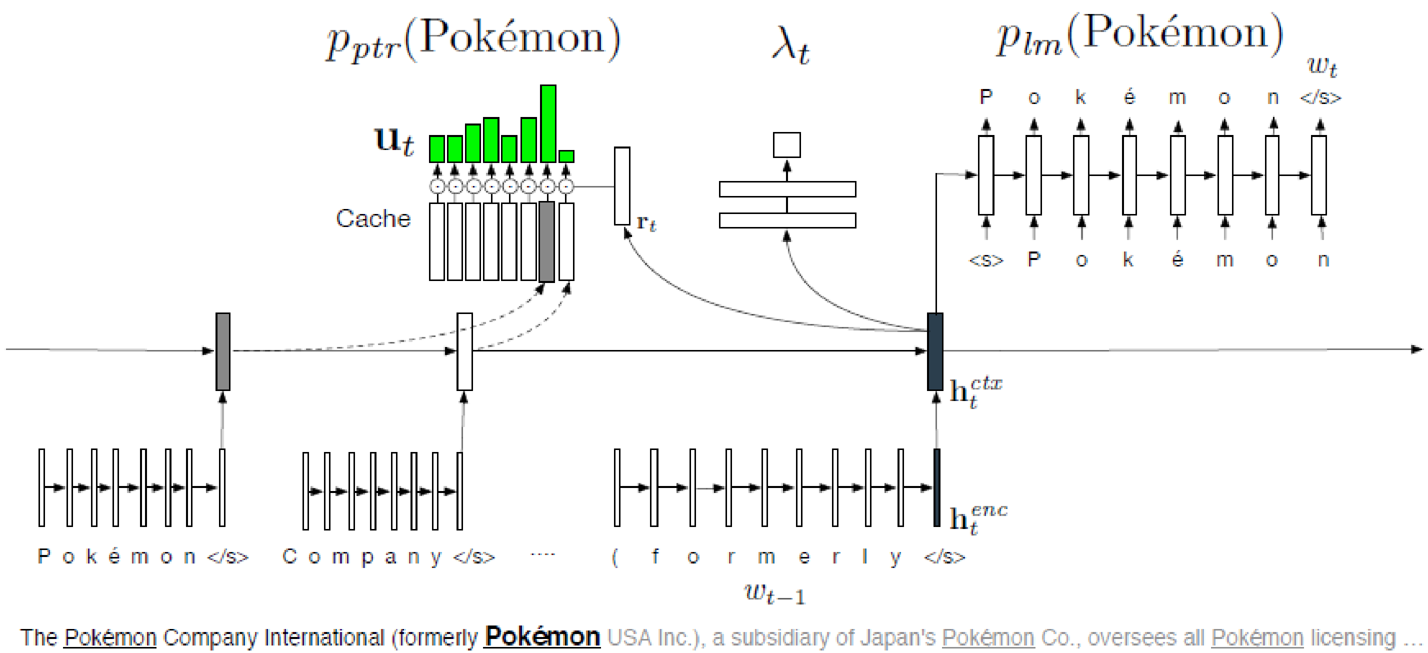




Merity et al. 2016

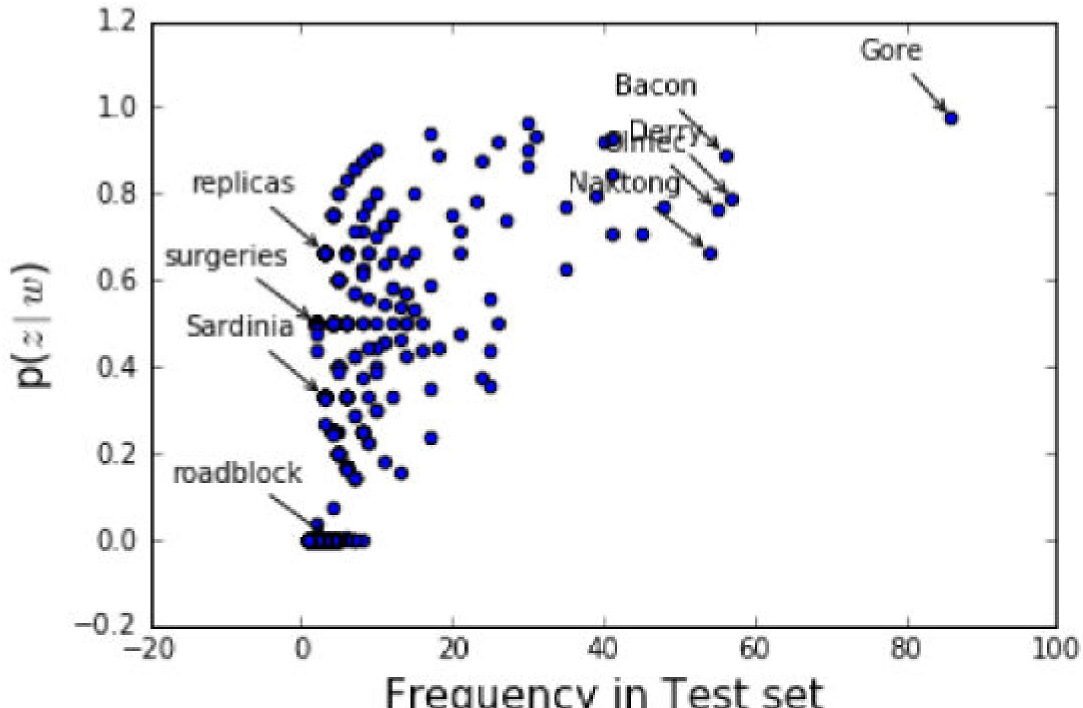
Hierarchical character-level model

$$p(\text{Pokémon}) = \lambda_t p_{lm}(\text{Pokémon}) + (1 - \lambda_t) p_{ptr}(\text{Pokémon})$$



Kawakami et al. 2017

Hierarchical character-level model



Kawakami et al. 2017

References

1. A neural probabilistic language model, Bengio et al., 2003
2. Exploring the Limits of Language Modeling, Jozefowicz et al., 2016, arXiv:1602.02410
3. Character-Aware Neural Language Models, Kim et al., 2016
4. Learning to Compute Word Embeddings on the Fly, Bahdanau et al., 2017, arXiv:1706.00286
5. Morph-fitting: Fine-Tuning Word Vector Spaces with Simple Language-Specific Rules, Vulić et al., 2017, arXiv:1706.00377
6. Pointer Sentinel Mixture Models, Merity et al., 2016, arXiv:1609.07843
7. Learning to Create and Reuse Words in Open-Vocabulary Neural Language Modeling, Kawakami et al., 2017, arXiv:1704.06986