LightRNN: Memory and Computation-Efficient Recurrent Neural Networks
NIPS16

Xiang Li, Tao Qin, Jian Yang, Tie-Yan Liu

Problems with traditional RNN-LM

- Huge and slow model with large vocabularies.

- Approximation solution: Hierarchical Softmax, Negative Sampling

- Limits: Still huge (HS, NS)/ slow at test (NS).

Reconsider the organization of vocabulary

- Vocabulary is an ordered list of words ?

- Reduce model size and computational complexity by redefining vocabulary.

List to table

embedding	word
x_1	January
x_2	February
…	…
x_{15}	one
x_{16}	two

shared embedding word table

embedding	$x_1^{\mathbf{c}}$	$x_2^{\mathbf{c}}$	$x_3^{\mathbf{c}}$	$x_4^{\mathbf{c}}$
x_1^{r}	January	February	…	…
x_2^{r}	one	two	…	…
x_3^{r}	…	…	…	…
x_4^{r}	…	…	…	…

embedding	word
$(x_1^{\mathrm{r}}, x_1^{\mathrm{c}})$	January
$(x_1^{\mathrm{r}}, x_2^{\mathrm{c}})$	February
$(x_2^{\mathrm{r}}, x_1^{\mathrm{c}})$	one
$(x_2^{\mathrm{r}}, x_2^{\mathrm{c}})$	two
…	…

Model size: from $|V|$ to $2\sqrt{V}$

Use them in RNN-LM

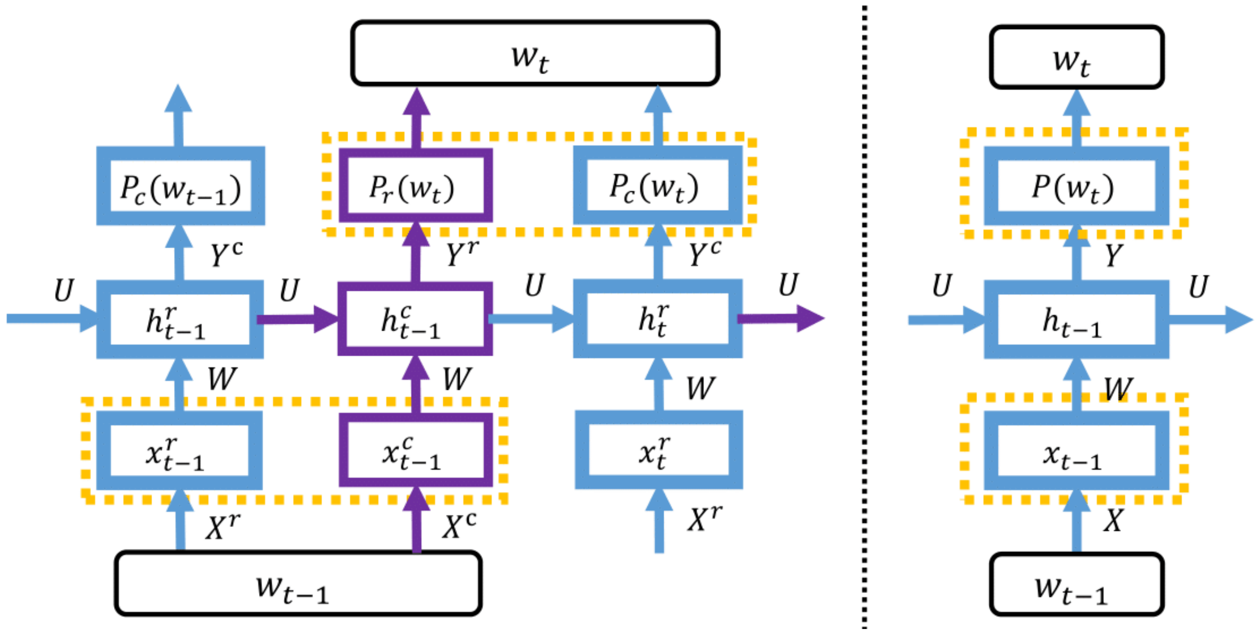


Figure 2: LightRNN (left) vs. Conventional RNN (right).

Probability $P(w_t)$

- Now there are two parts of it:

$$P_r(w_t) = \frac{\exp(h^c_{t-1} \cdot y^r_{r(w)})}{\sum_{i \in S_r} \exp(h^c_{t-1} \cdot y^r_i)} \qquad P_c(w_t) = \frac{\exp(h^r_t \cdot y^c_{c(w)})}{\sum_{i \in S_c} \exp(h^r_t \cdot y^c_i)},$$

$$P(w_t) = P_r(w_t) \cdot P_c(w_t),$$

Computational complexity: from a |V|-way normalization to two $\sqrt{|V|}$ -way normalization.

Remained question

- How to allocate the words into appropriate positions?

$$
\begin{aligned}
NLL_w &= \sum_{t \in S_w} -\log P(w_t) = l(w, r(w), c(w)) \\
&= \sum_{t \in S_w} -\log P_r(w_t) + \sum_{t \in S_w} -\log P_c(w_t) = l_r(w, r(w)) + l_c(w, c(w)),
\end{aligned}
$$

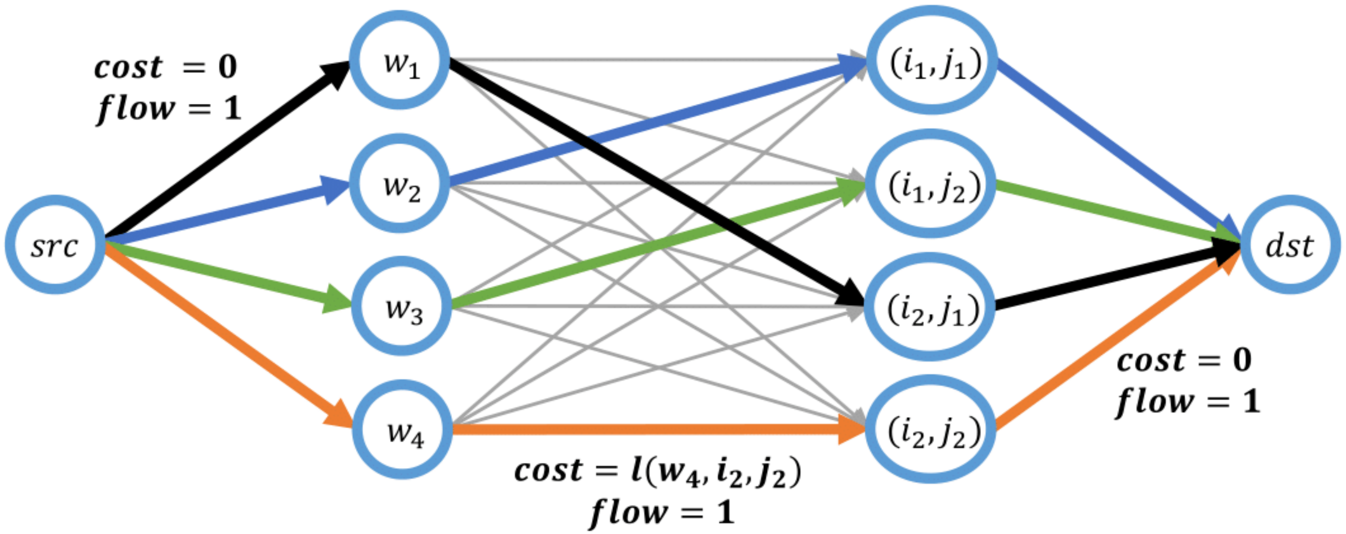
A Minimum Cost Maximum Flow Question

$$\min_a \sum_{(w,i,j)} l(w,i,j)a(w,i,j) \quad \text{subject to}$$

$$\sum_{(i,j)} a(w,i,j) = 1 \ \ \forall w \in V, \quad \sum_{w} a(w,i,j) = 1 \ \ \forall i \in S_r, j \in S_c,$$

$$a(w,i,j) \in \{0,1\}, \quad \forall w \in V, i \in S_r, j \in S_c,$$

A Minimum Cost Maximum Flow Question



Results

Table 1: Statistics of the datasets

Dataset	#Token	Vocabulary Size
ACLW-Spanish	56M	152K
ACLW-French	57M	137K
ACLW-English	20M	60K
ACLW-Czech	17M	206K
ACLW-German	51M	339K
ACLW-Russian	25M	497K
BillionW	799M	793K

PPL on ACLW test						
Method	Spanish/#P	French/#P	English/#P	Czech/#P	German/#P	Russian/#P
KN[4]	219/–	243/–	291/–	862/–	463/–	390/–
HSM[13]	186/61M	202/56M	236/25M	701/83M	347/137M	353/200M
C-HSM[13]	*169*/48M	*190*/44M	*216*/20M	*578*/64M	*305*/104M	*313*/152M
LightRNN	**157**/**18M**	**176**/**17M**	**191**/**17M**	**558**/**18M**	**281**/**18M**	**288**/**19M**

More interesting results

row 832	Karwan Narok Cocodrie Noja Anambra Alaska. Lantau Willmar Zululand Tianmen …
row 852	281-211 3-6-0 17-of-44 21-for-27 100-64 1,173-767 10-to-2 7-and-5 15,350 of-15 …
row 861	103-run 12-way 23-hit 151-game 13-ball 105-meter 302-minute 189-yard 67-foot …
row 872	totaled hunted rigged scored vetoed inflicted froze swam won dried raged smiled …
row 877	plods riles hankers misbehaves contrives utilizes disbands computes propagates …
row 887	www.angiotech.com www.huntsman.com media.floridarealtors.org 2010.census.gov …
row 889	years. decade evening hours. weeks spring summer. day-and-a-half April-to-June …
row 891	44kg 63pc 170mph 18cm 22C 12A 150bp 17st 656ft 2Mbps 680g 10x 13ph. 2M …

My experiments

- 1D, 2D, 3D, …, ND.

- Not really faster because the MCMF algorithm.

- 2D is best.

Dependency based word embedding
ACL14. Omer Levy and Yoav Goldberg

- What does word embedding learn from? Co-occurrence of words

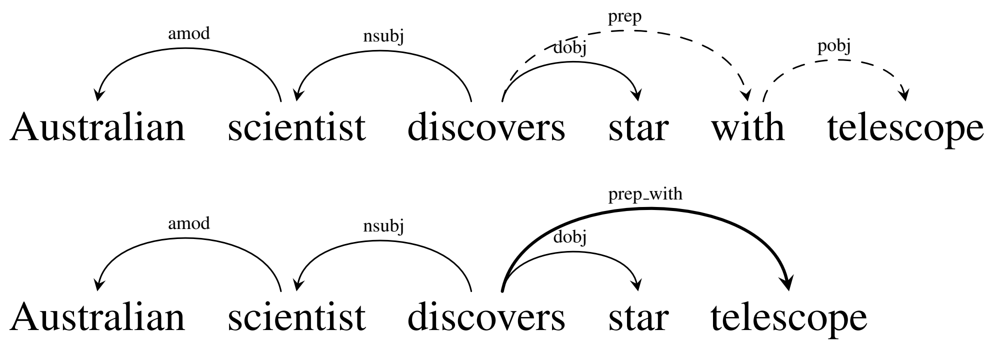
- How to define a context? Linear?

Problems with linear context

- Close words are not necessarily really related.

- Distant word sometimes are important. Consider long dependencies relations within a sentence.

Dependency based context



WORD	CONTEXTS
australian	scientist/amod^{-1}
scientist	australian/amod, discovers/nsubj^{-1}
discovers	scientist/nsubj, star/dobj, telescope/prep_with
star	discovers/dobj^{-1}
telescope	discovers/prep_with^{-1}

Results

Target Word	BoW5	BoW2	Deps
batman	nightwing aquaman catwoman superman manhunter	superman superboy aquaman catwoman batgirl	superman superboy supergirl catwoman aquaman
hogwarts	dumbledore hallows half-blood malfoy snape	evernight sunnydale garderobe blandings collinwood	sunnydale collinwood calarts greendale millfield
turing	nondeterministic non-deterministic computability deterministic finite-state	non-deterministic finite-state nondeterministic buchi primality	pauling hotelling heting lessing hamming
florida	gainesville fla jacksonville tampa lauderdale	fla alabama gainesville tallahassee texas	texas louisiana georgia california carolina
object-oriented	aspect-oriented smalltalk event-driven prolog domain-specific	aspect-oriented event-driven objective-c dataflow 4gl	event-driven domain-specific rule-based data-driven human-centered
dancing	singing dance dances dancers tap-dancing	singing dance dances breakdancing clowning	singing rapping breakdancing miming busking

Linear contexts yield broad topical similarities.

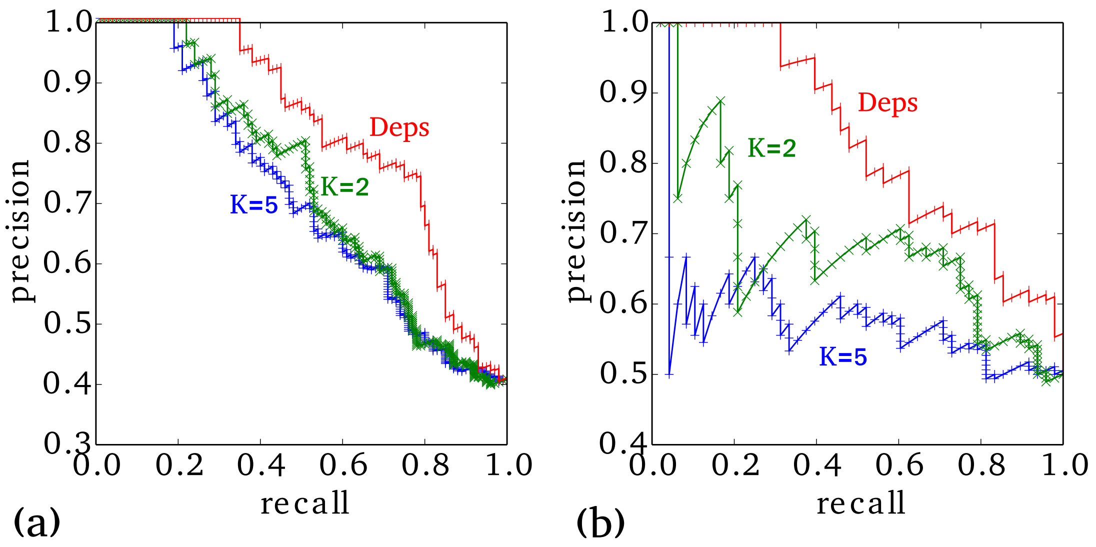
Dependency-based contexts yield more functional similarities.

Datasets

- WordSim535

- This dataset contains pairs of similar words that reflect either *relatedness* (topical similarity) or *similarity* (functional similarity) relations.

Results



precision
recall
Deps
K=2
K=5
(a)
(b)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

Following work

- Prepositional Phrase Attachment

  Exploring Compositional Architectures and Word Vector Representations for Prepositional Phrase Attachment. TACL14

- Knowledge graph (e.g., WordNet)

  Ontology-Aware Token Embeddings for Prepositional Phrase Attachment. ACL17

My experiments

- Treat the dependency relations in a corpus as a (directed/weighted /typed) word graph.
- Run graph embedding methods (e.g., Node2vec, DeepWalk,..)
- Not really work.