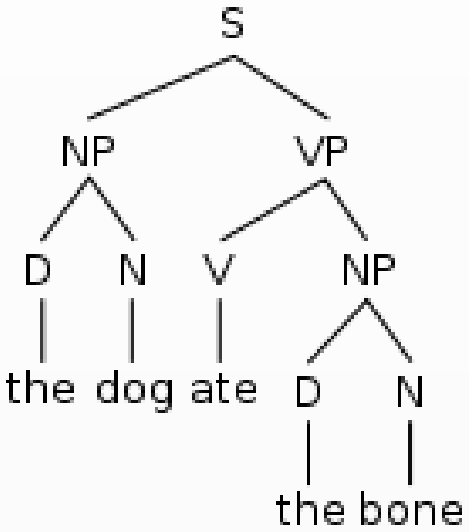
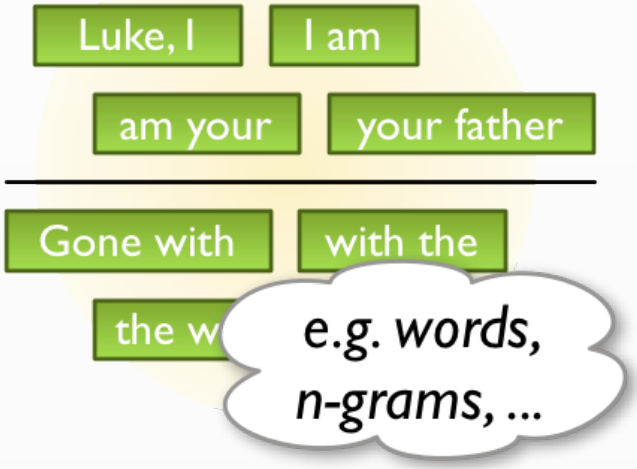
Words learning using deep structures

19.6.2015

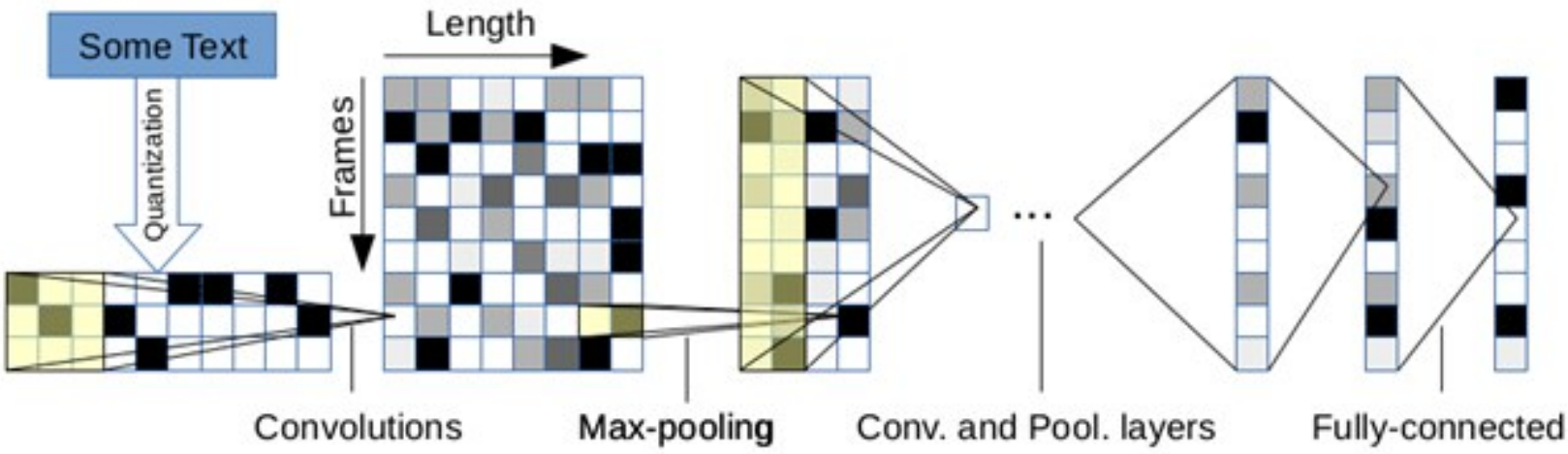
- Grammatical models
- Bag of words
- Vectorization and complex ideas



- Forgets sequential information
- Sparse
- TF-IDF (Terms frequency, inverse document frequency)
- unigrams / bigrams / trigrams
- Latent dirichlet allocation
- Latent semantic indexing



- word2vec - assisted to group similar meanings
- Text understanding from scratch



- Text classification (sentinent analysis, emotion recognition)
- clustering (for indexing documents, for groupping similar texts)

The idea

”Gsues wichh of tshee tehre scenetens mghit be mroe rdbleaae?”

$$r : \mathcal{A}^{\infty} \to \mathcal{B} \subset \mathbb{R}^{2 \cdot |\mathcal{A}|} : word \in \mathcal{A}^{\infty} \to r(word) \in \mathcal{B} : r(word) =$$

$$\left(\sum_{c \in |\mathcal{A}|} vec(c) \cdot \sum_{i=0}^{len(word)} \delta_{vec(word_{[i]}), vec(c)}, \right.$$

$$\left. vec(word_{[0]}) + vec(word_{[len(word)]}) \right)$$

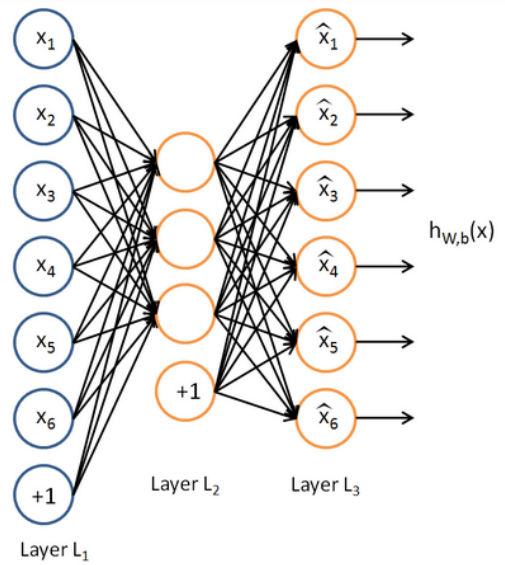
- doorstops doorposts
- doorstops doorposts
- kleig klieg
- noncasual noncausal
- organization's organizations
- regains reginas
- snakes sneaks
- teazles teazels

(To try another language - from a Czech dictionary consisting of 300000 words, the number of collisions are 43.)

- VS Gramatic models - different plugins for different languages
- VS bag of words - big input space
- every word has a fixed length
- mostly sparse
- to reconstruct the word from its code
- humans system

Architectures tried and to try

- autoencoders
- convolutional neural networks
- Moving window
- (RBM)
- (RBM + autoenc)



Datasets tried

- orders (failed)
- DBpeadia

some libraries for NNs

- FANN
- Theano
- Pylearn2

Experiments

- speed of 47 words
- Learning rate 0.1, L2 regularization 0.0001
- batch sizes were chosen to be 361, size of training dataset was 649800 and testing and validation sets 238260 items each.
- smaller convNN: 20 frames+kernel 5x5+pool2x2; 30 frames+kernel 5x5+pool2x2; 20 frames+kernel 5x5+pooling2x2; 500 fully connected sigmoid.
- larger convNN: 30 frames+kernel 5x5+pool2x2; 40 frames+kernel 5x5+pool2x2; 50 frames+kernel 5x5+pooling2x2; 1000 fully connected sigmoid.
- autoencoder (sigmoid layers), number of neurons: 3000, 4000,3000,2000,1000,500,200,100

- 51.54% validation error, 52.76% test error for bigger convolutional network
- and 53.80% validation, 55.23% for smaller network on dbpeadia dataset

”A Walk in the Sun is a World War II war film released in 1945 based on the novel by Harry Brown who was a writer for Yank the Army Weekly based in England...”

- ”A Walk to Beautiful is a 2007 American documentary film produced and distributed by Engel Entertainment about women who suffer from childbirth injuries in Ethiopia. In 2007 it premiered in film festivals and was chosen for the International Documentary...”(dist: 2.94E-06)
- ”A Walk in the Woods: Rediscovering America on the Appalachian Trail is a 1998 book by travel writer Bill Bryson describing his attempt to walk the Appalachian Trail with his friend Stephen Katz. The book is written...”(dist: 0)

Some of the far-away (dist > 7) :

- She Married for Love is a 1914 silent comedy film featuring Oliver Hardy.
- The Scroafa River is a tributary of the Architа River in Romania.

Tkahns for aotnitten!