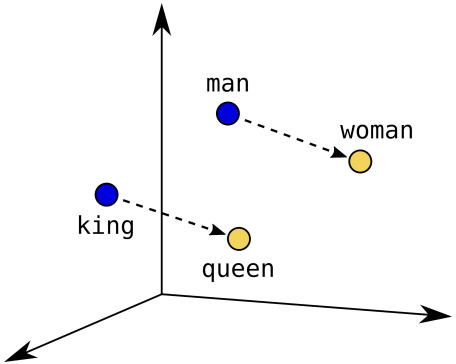
Towards Useful Word Embeddings

Evaluation on Information Retrieval, Text Classification & Language Modeling



Not very useful!

MUNI FI

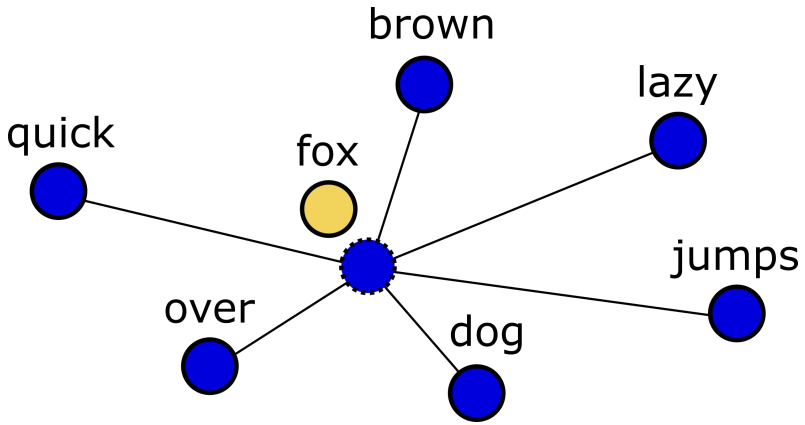
V. Novotný, M. Štefánik, D. Lupták, and P. Sojka

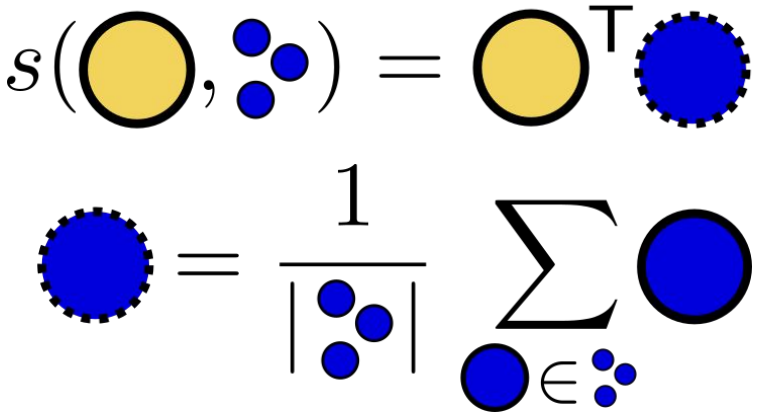
https://mir.fi.muni.cz/

RASLAN 2020

Learning Word Embeddings with FastText

- **Baseline model** predicts a masked word from the mean context word vector:





"The quick brown **???** jumps over the lazy dog"

- **Positional model** makes context word vectors *position-dependent*:
  - Context "Unlike dogs, cats are **???**" has a different vector than "Unlike cats, dogs are **???**".

Information Retrieval & Text Classification

- Word embeddings can be used in both using WMD [8] & SCM [17] measures.
- For inf. retrieval, we preprocessed TREC datasets bought by the NLP Centre.
- For text classification, we repeated the experiment of Kusner et al. (2015) [8]:

Table 1. Classification error of the baseline and positional models with the WMD and SCM measures and the *k*NN classifier on the text classification test sets. For the WMD, we also list the results of Kusner et al. (2015) [8] for comparison. The best results are **emphasized**.

		BBCSPORT	TWITT.	RECIPE	OHSU.	CLASS.	REUTERS	AMAZ.
WMD	Kusner [8]	**4.6%**	29%	43%	44%	**2.8%**	**3.5%**	**7.4%**
	Baseline		**23.78%**	43.47%	46.16%			
	Positional		38.20%	**34.23%**	46.32%			
SCM	Baseline	6.64%	29.03%	45.63%	41.32%	4.85%	7.58%	10.27%
	Positional	5.82%	28.54%	43.52%	**38.93%**	4.40%	8.73%	9.81%

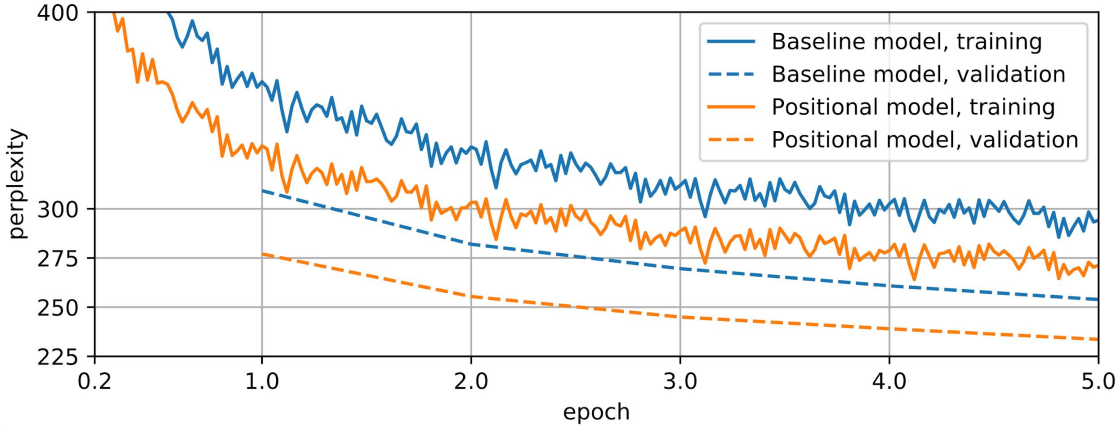
[8]: http://proceedings.mlr.press/v37/kusnerb15.pdf (From Word Embeddings to Document Distances)
[17]: https://dl.acm.org/doi/10.1145/3269206.3269317 (Implementation Notes for the Soft Cosine Measure)

Language Modeling is Also Very Useful™

- Word embeddings can initialize the lookup table of an LSTM language model.
- We trained a single-layer recurrent network with the following architecture:
  1. an input layer with a map from a vocabulary V to frozen word embeddings, followed by
  2. an LSTM unit with a recurrent hidden output of size $D = 300$, followed by
  3. a fully-connected linear layer of size $|V|$, followed by
  4. a softmax output layer that computes probability over the vocabulary V.
- As our datasets, we used the data from [the 2013 ACL WMT Workshop](#).

	Test perplexity	Test loss
Baseline model	270.34	5.60
Positional model	**251.69**	**5.53**

Positional model is just plain better!



Thanks for your Attention!

MUNI
FI

V. Novotný, M. Štefánik, D. Lupták, and P. Sojka
https://mir.fi.muni.cz/
RASLAN 2020