Improving Text Categorization

with Semantically Enriched LSTM

overview

- categorization of Slovak texts
- extraction of keywords
- novel LSTM architecture with latent feature vectors

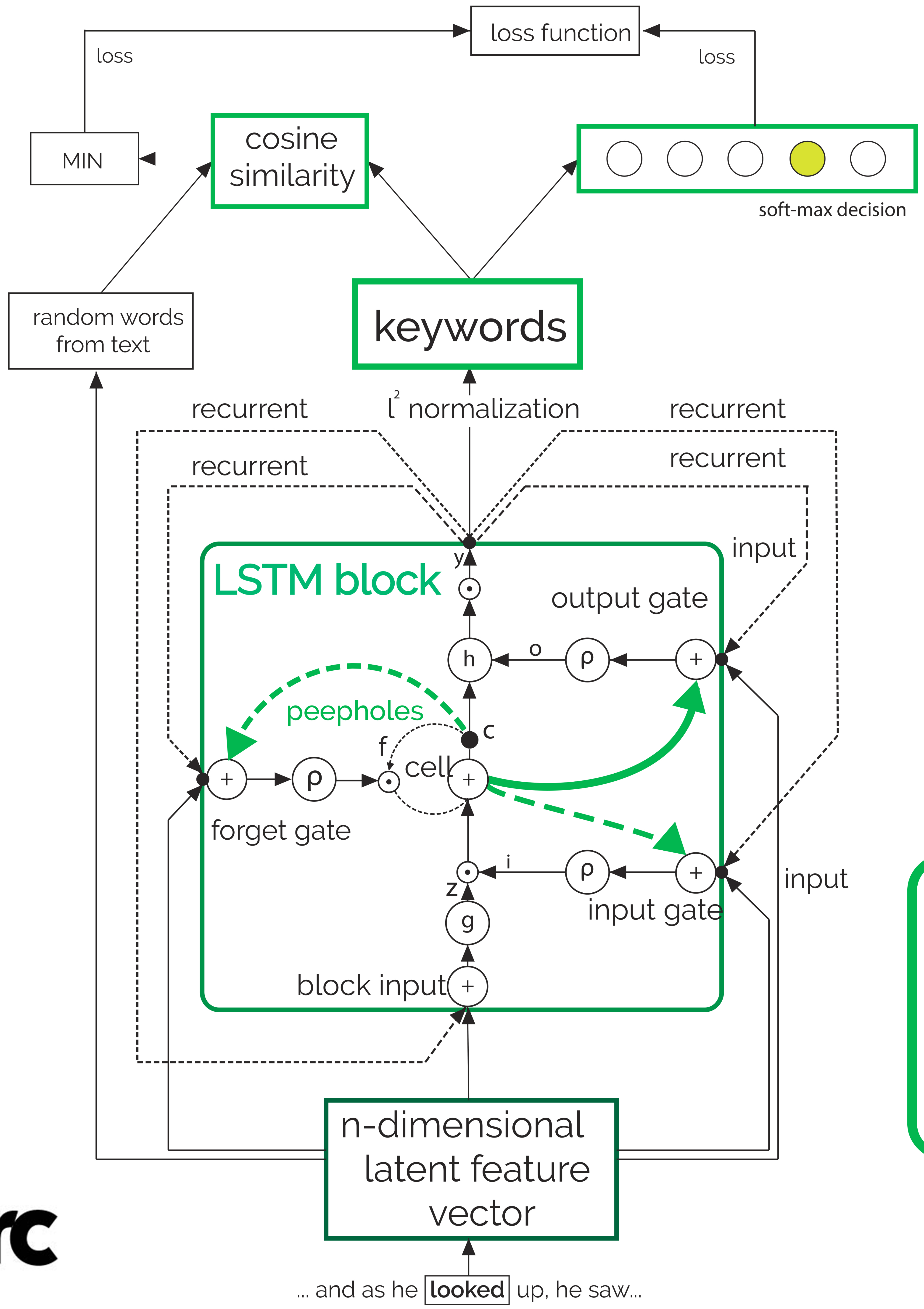
our approach

Our architecture is separated into 3 main layers:

- Transformation of input words into latent feature vectors using Word2Vec
- The LSTM - Long Short-Term Memory module
  - memory cell can maintain its state overtime
  - gating units can regulate the information flow
- - softmax function for categorization
  - cosine similarity between random words and output, which will transform into keywords

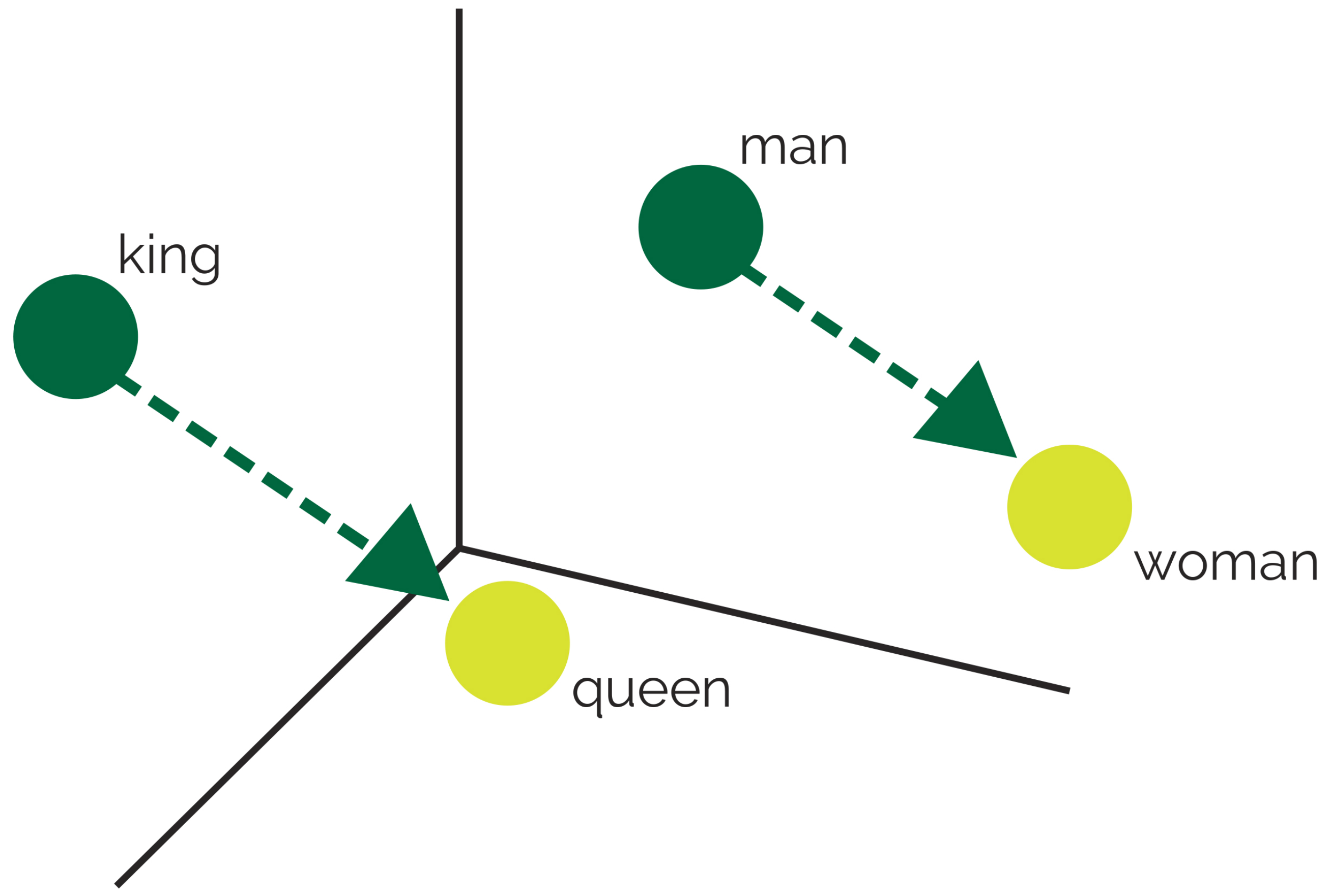
dataset

- consists of Slovak Wikipedia
  - 194 000 articles
  - 72 000 categories
- is separated into three sets
  - training, validation and test
  - ratio of training 80/10/10%
- vectors pre-trained on Slovak National Corpus



latent feature vectors

- using for preprocess morphologically rich Slovak language
- maping words into word vectors
- creating vector space
- clustering semantically close words
- using vector operations



conclusion

- recurrent architecture enables processing of text documents of variable length
- data-driven approach to extract discriminative keywords
- language independent model - requires only pre-trained vectors

Author: Adam Rafajdus

Supervisor: Ing. Márius Šajgalík