University of Ljubljana, Faculty of Computer and Information Science

ELMo embeddings

Prof Dr Marko Robnik-Šikonja

Natural Language Processing, Edition 2022

Contents

- subword input for NNs
- ELMo embeddings

Sub-word inputs for NNs

- good for morphologically rich languages
- Byte Pair Encoding
  - Most frequent byte pair↦a new byte
  - in NLP, we use character ngrams instead of bytes
- Start with a unigram vocabulary of all (Unicode) characters in data
- Most frequent ngram pairs ↦a new ngram
- Have a target vocabulary size and stop when you reach it
- Do deterministic longest piece segmentation of words
- Segmentation is only within words identified by some prior tokenizer
- Automatically decides vocabulary for system
- No longer strongly “word” based in conventional way

Word/sentence piece encoding

- Google NMT uses a variant of Byte Pair Encoding
- wordpiece model
- sentencepiece model
- Rather than char *n*-gram count, uses a greedy approximation to maximizing language model log likelihood to choose the pieces
- Add *n*-gram that maximally reduces perplexity

- Wordpiece model tokenizes inside words
- Sentencepiece model works from raw text
- Whitespace is retained as special token (_) and grouped normally
- You can reverse things at end by joining pieces and recoding them to spaces

ELMo embeddings

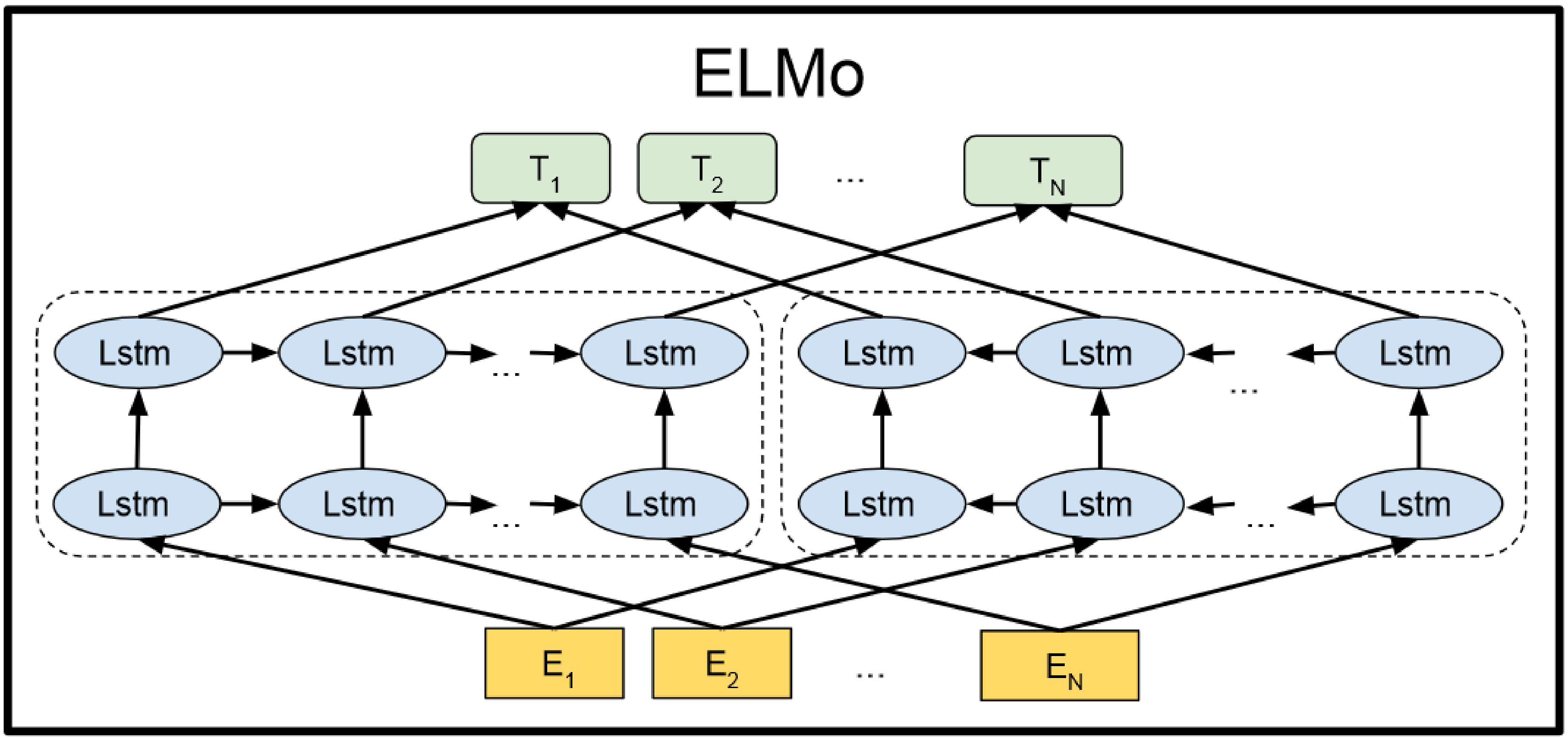
- Embeddings from Language Models
- Learn word token vectors using long contexts (whole sentence, could be longer)
- Learn a deep bidirectional neural language model (Bi-NLM) and use all its layers for the prediction or extract fixed-size vectors

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT* (pp. 2227-2237).

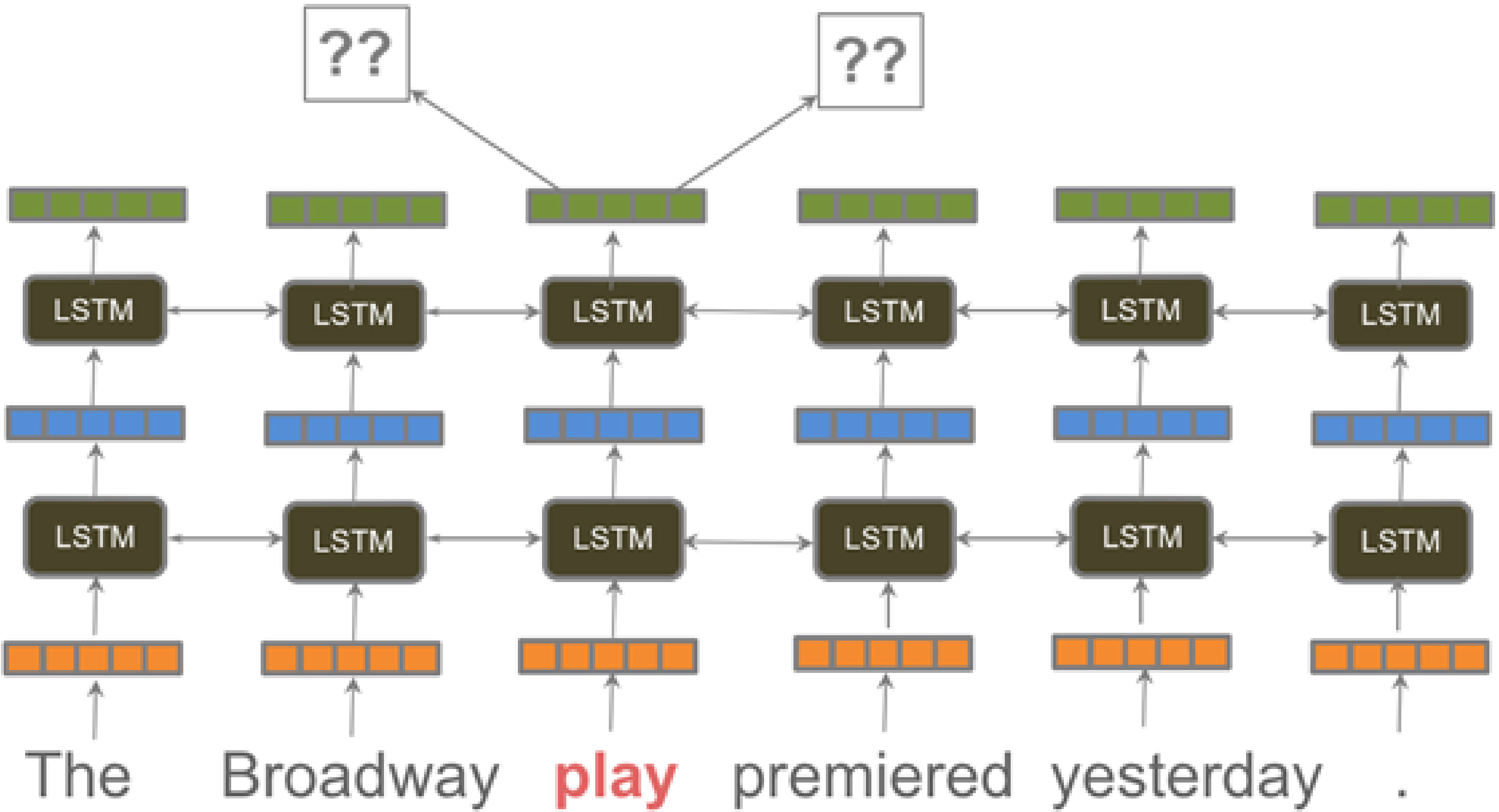
ELMo embeddings details

- Train a bidirectional LM
- Aim at performant but not overly large LM:
- Use 2 biLSTM layers
- Use character CNN to build initial word representation (only)
  - 2048 char n-gram filters and 2 highway layers, 512 dim projection
- User 4096 dim hidden/cell LSTM states with 512 dim projections to next input
- Use a residual connection
- Tie parameters of token input and output (softmax) and tie these between forward and backward LMs

ELMO: biLSTM architecture



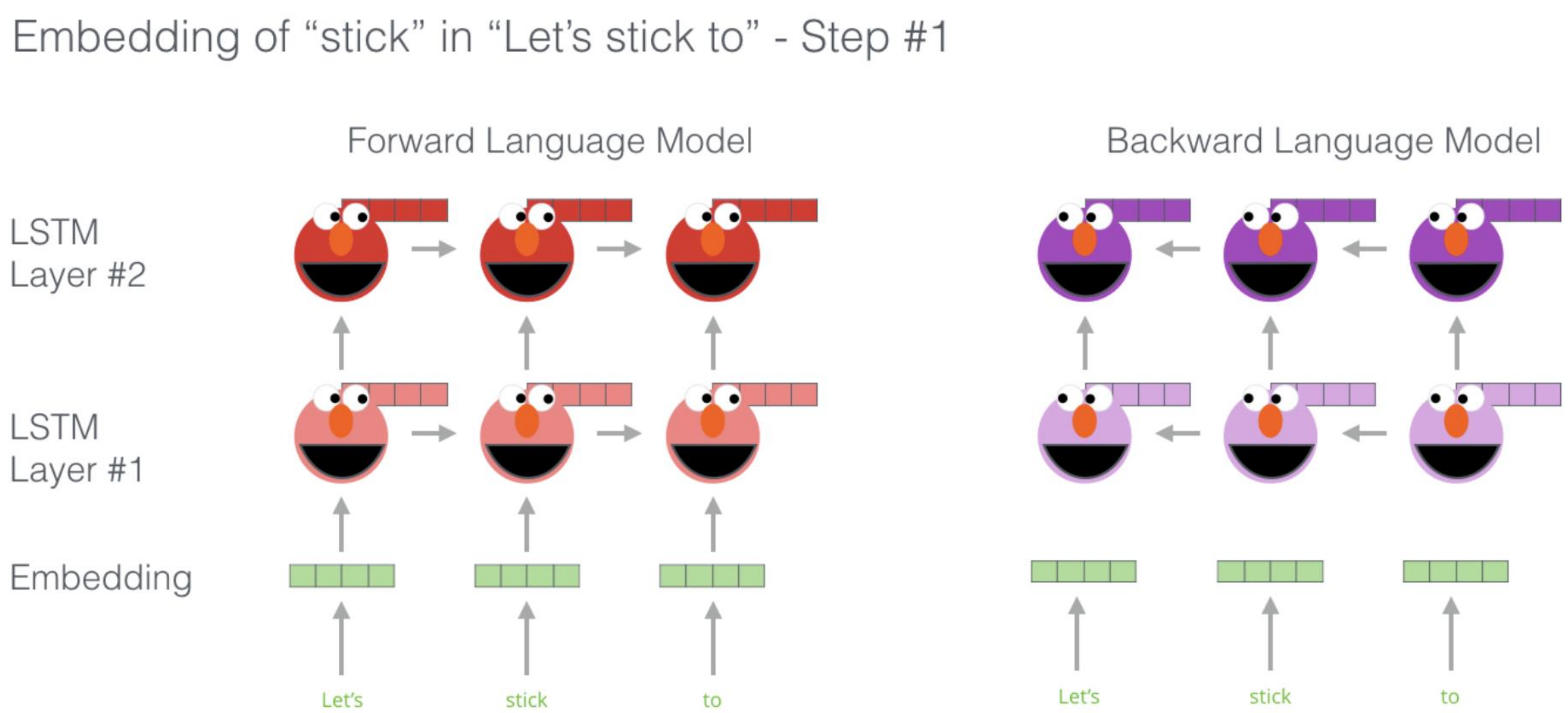
How ELMo works



ELMo weights

- The two biLSTM NLM layers have differentiated uses/meanings
- Lower layer is better for lower-level syntax: Part-of-speech tagging, syntactic dependencies, NER
- Higher layer is better for higher-level semantics: sentiment, semantic role labeling, question answering, SNLI

Producing contextualized embeddings 1/2



the illustrations by Jay Alammar

Producing contextualized embeddings 2/2