UNIVERSITY OF CAMBRIDGE

Evaluating Character-Level Recurrent Neural Networks

Leonie Weißweiler, Daniela Gerz, Anna Korhonen

Language and Technology Lab

Advantages and Disadvantages of Char-Level Models

- Potential advantage
  - the model could learn about character-level dependencies
  - This would hopefully capture some of the language’s morphology
  - The model could then generate some correct unseen words
- Potential disadvantage
  - The generated words aren’t guaranteed to be words
  - There is always a chance of some of the output being gibberish

Motivation

- Character-level models are normally evaluated with perplexity
- Perplexity captures how well the model adapted to the data, not the chances and risks versus word-level models
- There are little to no studies comparing different character-level RNNs

Analysis of Generated Text

a) Words that have been observed in the training data and were repeated by the model

b) Words that were not present in the training data, but which do exist in the language

c) Words that do not exist in the language

→ Correctness Percentage $\frac{a+b}{a+b+c}$

→ New Sensible Word Count b

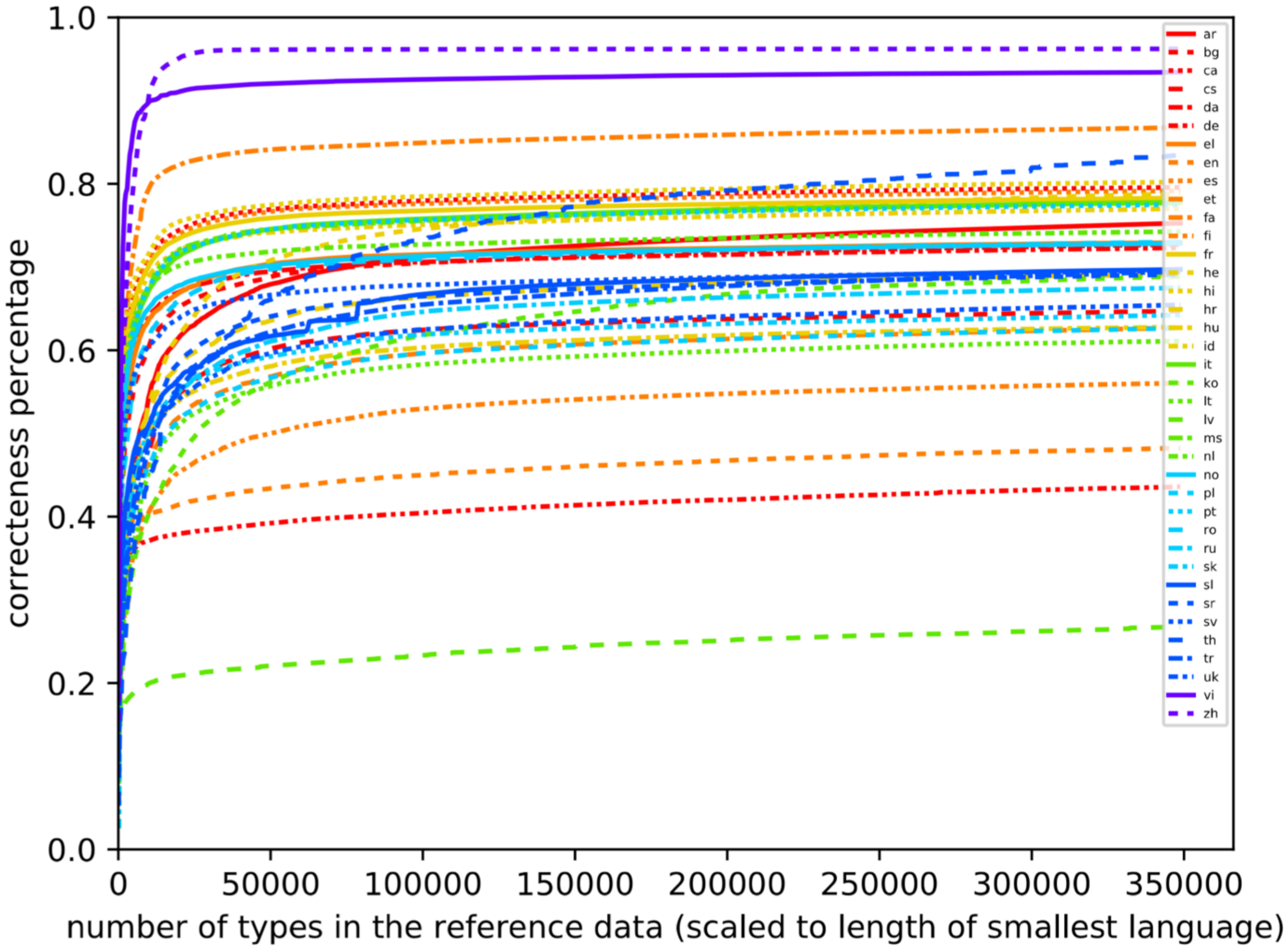
New Metrics

- Correctness Percentage: percentage of the words in the output of the character-level model that exist in the language
  - → measures how many non-words are produced
  - → measures the disadvantage in comparison to word-level models
- New Sensible Word Count: number of words that were not observed in the training data, but which do exist in the language
  - → measures how well the model captures character-level dependencies
  - → measures the advantage in comparison to word-level models

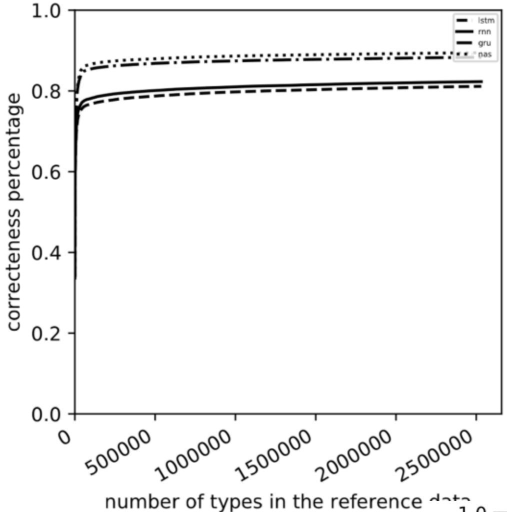
Evaluation Setup

- Wikipedia Corpus in 38 languages
- Used small chunks as training data and the rest for checking which words are in the language
- Tokenized with the OpenNLP and the Polyglot Tokenizer
- Tested LSTM, RNN, GRU, NAS
- Standard parameters, no tuning
- 5 million characters training data, 500,000 characters sampled for each language

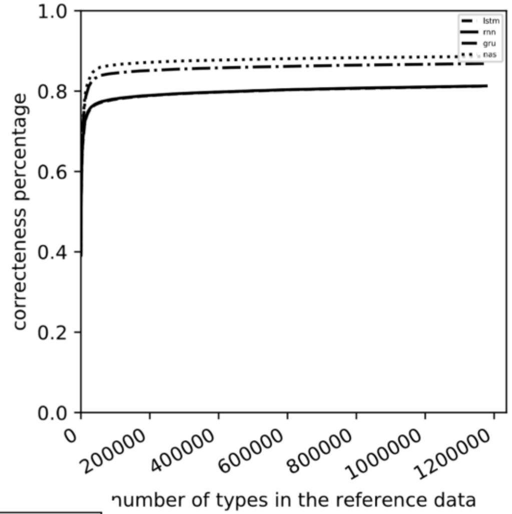
Correctness Percentages LSTM



Correctness Percentage Groups

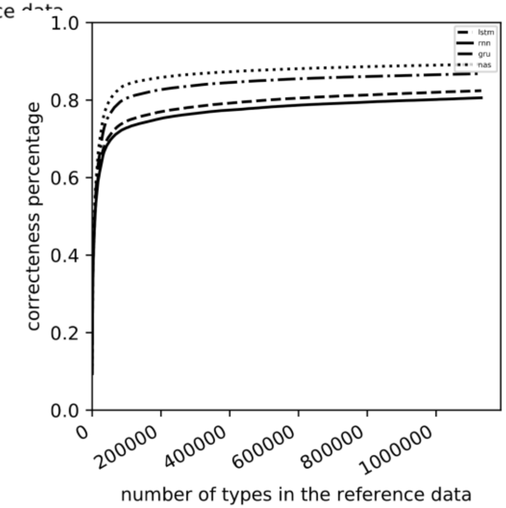


← French

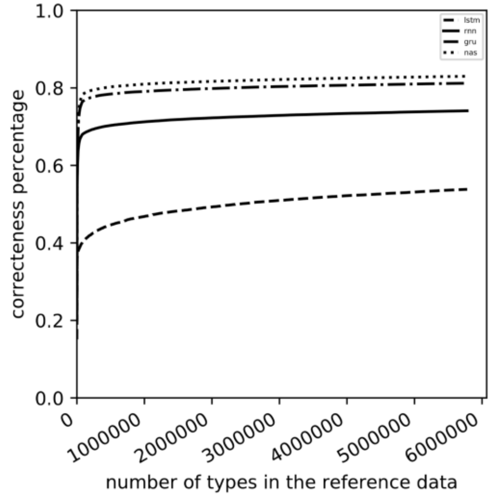


← Catalan

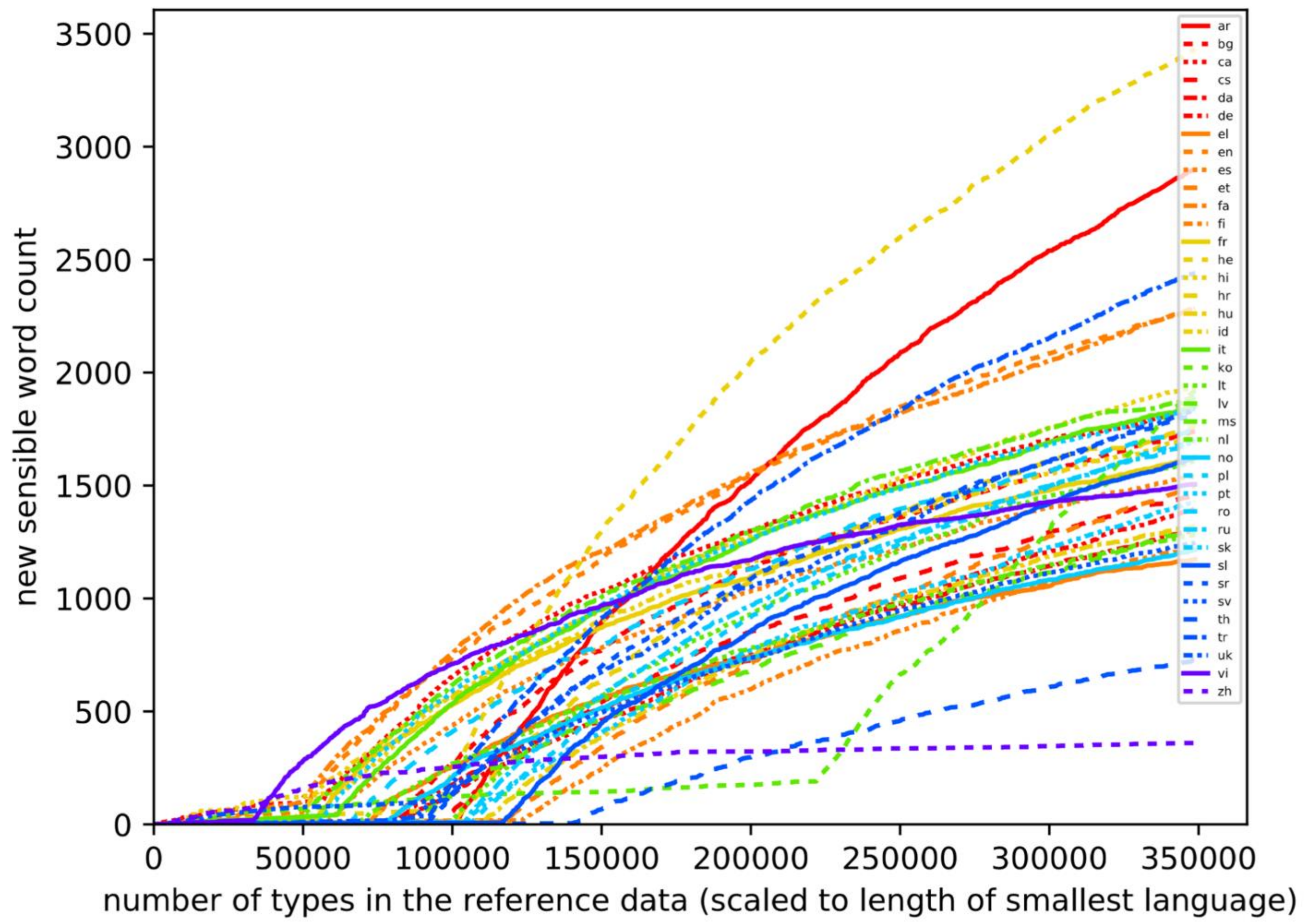
Hebrew →



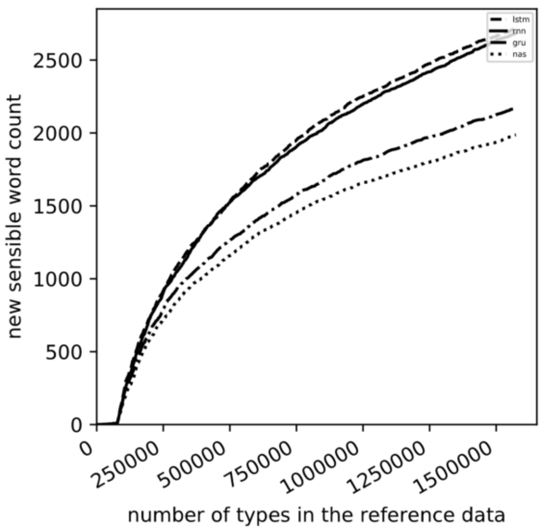
German →



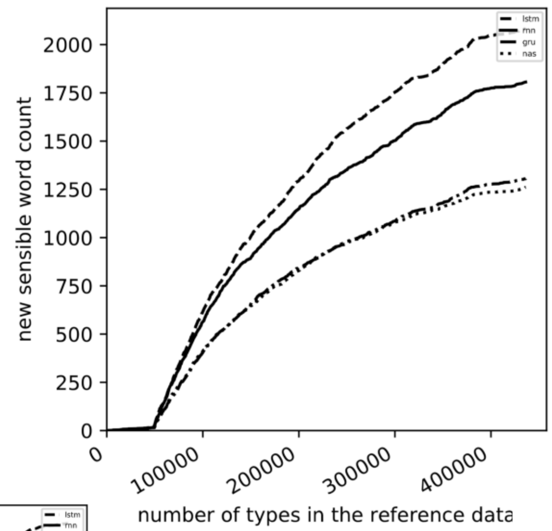
New Sensible Word Counts LSTM



New Sensible Word Count Groups

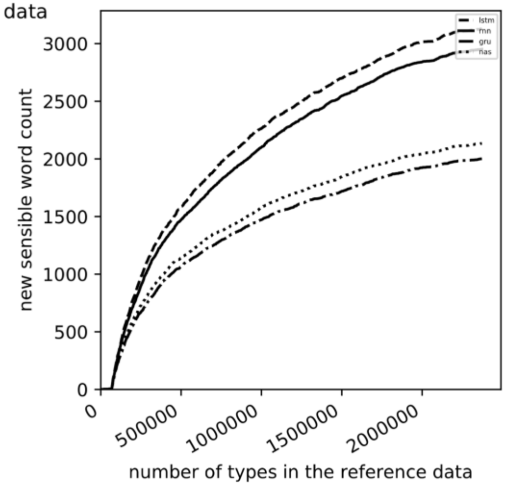


←Norwegian

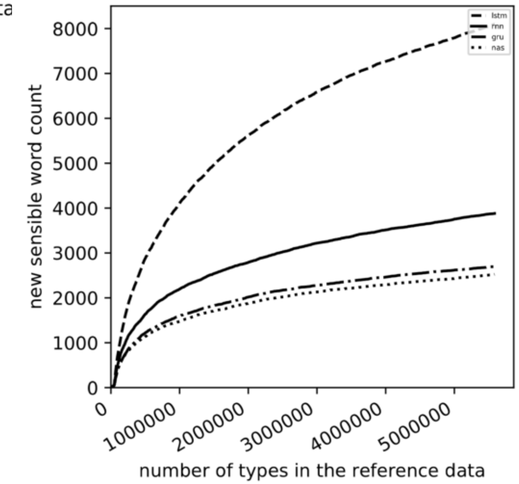


← Malay

Dutch →



English →



General Results

- General trade-off: LSTM → RNN → GRU → NAS if the Correctness Percentage is the priority, otherwise the other way around
- For a number of languages, a few models, like RNN and GRU, should not be used at all
- LSTM and RNN perform better for Correctness Percentage → are better at remembering words
- NAS and GRU perform better for New Sensible Word Count → are better at learning patterns

Conclusion

- We introduced new metrics for evaluating Char-Level RNNs that make relevant statements about their performance
- We introduced the first comprehensive comparison of four Char-Level RNNs
- We gave some recommendations which models are best used for which language using both new metrics

Current and Future Work

- For publication, we are
  - Redoing all the experiments with a smaller, topic-controller corpus
  - Trying different RNN sizes, sample modes, and training data sizes
  - Computing perplexity scores to see if they correlate with correctness percentages