

92586 Computational Linguistics

Lesson 19. LSTM: characters and generation

Alberto Barrón-Cedeño

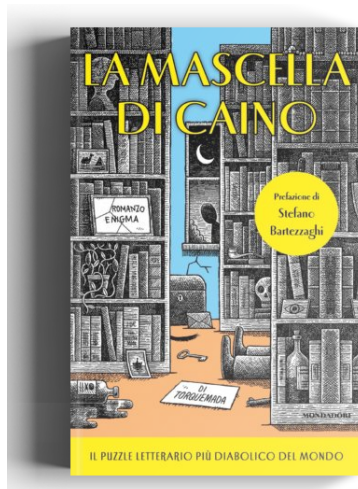
Alma Mater Studiorum-Università di Bologna
a.barron@unibo.it @albarron_

09/05/2022



A couple of ads

Yet another project proposal (with a high thesis potential)



CLEF 2022 is coming to Bologna!

CLEF 2022 Conference and Labs of the Evaluation Forum
Information Access Evaluation meets Multilinguality, Multimodality, and Visualization
5 - 8 September 2022, Bologna - Italy

Welcome

CLEF 2022 is the 13th CLEF conference continuing the popular CLEF campaigns which have run since 2000 contributing to the systematic evaluation of information access systems, primarily through experimentation on shared tasks.

Building on the format first introduced in 2010, CLEF 2022 consists of an independent peer-reviewed conference on a broad range of issues in the fields of multilingual and multimodal information access evaluation, and a set of labs and workshops designed to test different aspects of mono and cross-language information retrieval systems. Together, the conference and the lab series will maintain and expand upon the CLEF tradition of community-based evaluation and discussion on evaluation issues.

CLEF 2022 will be hosted by Università di Bologna, Italy, 5-8 September 2022

Home
Programme
Conference Sessions
Lab Sessions
Keynote Talks
Conference
Workshop Papers
Call for Papers
Labs
Call for Lab Proposals
Call for Lab Participation
Labs Registration
Venue
About Bologna
Accommodation
Schedule
Organisation

Tweets by @clef_initiative

CLEF Retweeted
@CraigMacdonald
@clef_initiative conference research paper track - deadline extended until 08 May - CIP webpage will be updated shortly https://twitter.com/craig_macdonald/status/1519280912713797633

May 2, 2022

CLEF Retweeted
@IA_Hub
@ia_initiative
As part of the #SearchEngines course at @UnibPadova, Wed 23 Mar at 10:30 CET, @albarron_ will talk about 'You're at #CLEF2022' @clef_initiative

When 5–8 September
Where Complessi Belmeloro & Santa Cristina/Cantina
Bentivoglio
Web <https://clef2022.clef-initiative.eu/>

Previously

- ▶ Convolutional neural networks
- ▶ Recurrent neural networks
- ▶ Bidirectional Recurrent neural networks
- ▶ Long short-term memory networks

Table of Contents

Out of Vocabulary

Characters

Text generation

Chapter 9 of Lane et al. (2019)

Out of Vocabulary

The curse of OOV

Out-of-vocabularies cause big trouble

The Mexico City Metro, operated by the Sistema de Transporte Colectivo, is the second largest metro system in North America after the New York City Subway.

The Mexico_City Metro, operated by the · de · · , is the second largest metro system in North America after the New_York City Subway.

Alternatives

- ▶ Replace the unknown with a random word, from the embedding space
- ▶ Replace the unknown word wit UNK, and produce a random vector
- ▶ **Turn into characters**

https://en.wikipedia.org/wiki/Mexico_City_Metro (2021)


Characters

Into Characters

Words are *just* a sequence of characters

By modeling the representations at the character level...

- ▶ We end up with a close vocabulary
- ▶ We get rid of OOVs
- ▶ We can learn patterns at a lower level
- ▶ We reduce the variety of input vectors drastically

 Let us see

Into Characters: outcome

- ▶ The training takes no less than 30 minutes (it took me 36 last time)¹

epoch	seconds	acc	acc _{val}
1	208	0.5206	0.5934
2	190	0.6832	0.5900
3	184	0.7534	0.5826
4	183	0.8029	0.5664
5	182	0.8371	0.5654
6	182	0.8633	0.5652
7	182	0.8908	0.5672
8	179	0.9086	0.5774
9	178	0.9212	0.5744
10	179	0.9346	0.5898

¹2.5GHz Quad-Core Intel Core i7 with 16GB of RAM

Into Characters: outcome

- ▶ The training accuracy is quite promising: ~ 93.00
- ▶ The validation accuracy is terrible: ~ 59.00

▶ Overfitting

Reasons/Solutions

- ▶ The model might be *memorising* the dataset
- ▶ Increase the dropout (try!)
- ▶ Add more labeled data (hard!)

A character-level model shines at its best when modeling/generating language

Text generation

Predicting the next word

- An LSTM can learn

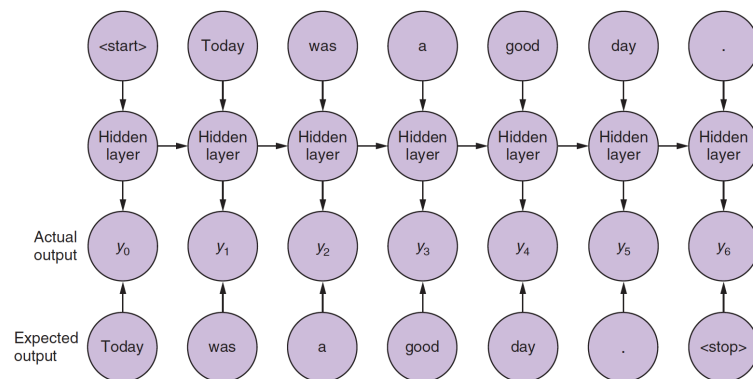
$$p(w_t \mid w_{t-1}, w_{t-2}, \dots, w_{t-n}) \quad (1)$$

- It can do so **with a memory** (full context)
- It can do so at the **character level**

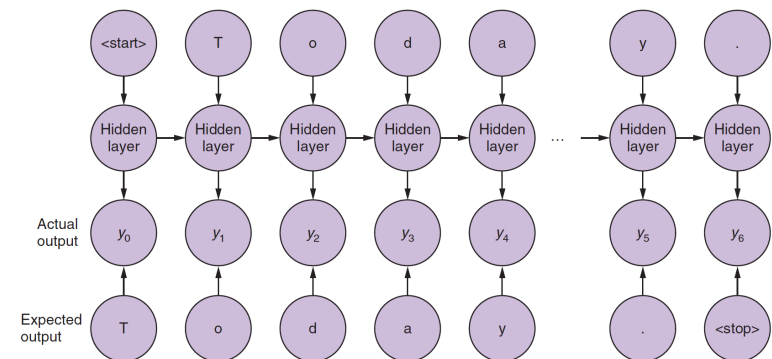
From classification to generation

- No more classification layer at the end
- Now we want to predict the next word (\sim word2vec?)
- We want to learn a *general* representation of language

Unrolling the next-word prediction (word 2-grams)



Unrolling the next-word character prediction

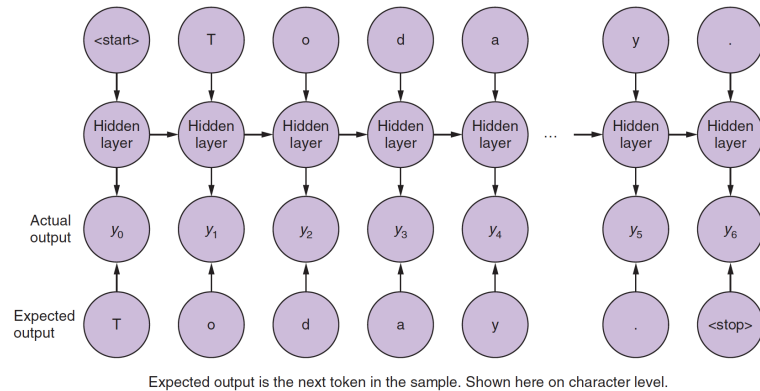


Expected output is the next token in the sample. Shown here on character level.

- Now the error is computed for every single output
- We still back-propagate until visiting a full instance

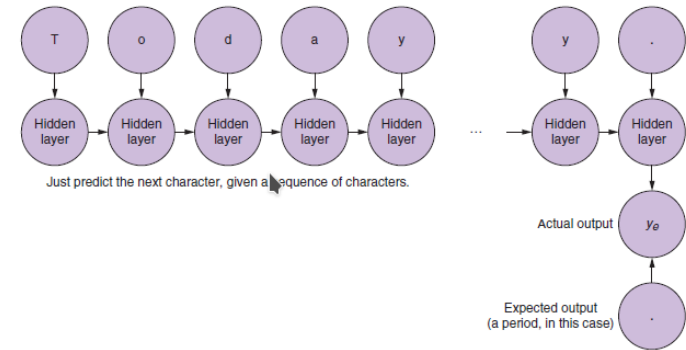
New target labels

New output: a one-hot encoding (again) of the next character



(Lane et al., 2019, 299)

Predict after having looked at a sequence




(Lane et al., 2019, 300)

Generation example

Since we are interested in *style* and in creating a consistent model, we won't use IMDB (multi-authored and small).

Let us try to *mimic* William Shakespeare

 Let us see

References

Lane, H., C. Howard, and H. Hapkem
2019. *Natural Language Processing in Action*. Shelter Island, NY: Manning Publication Co.