Identity word completion

Elmar H. Langholz

elmarhl2@illinois.edu

Problem definition

- Online anonymity and privacy is at risk due to massively amounts of data recollected and stylometry which helps identify your writing and therefore you.

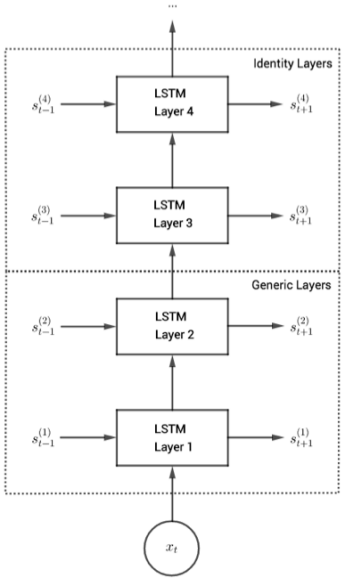
- Is it possible to sound like someone else while writing?

Novelty of my project

- Post-process:
  - Data anonymization tools try to generically remove the personal writing style from a provided text
    - Anonymouth
    - ARX
- Real-time:
  - Autocomplete using Language Models
    - Generalization vs. Personalization
    - **What if we want to mimic someone else but still retain some generic style?**

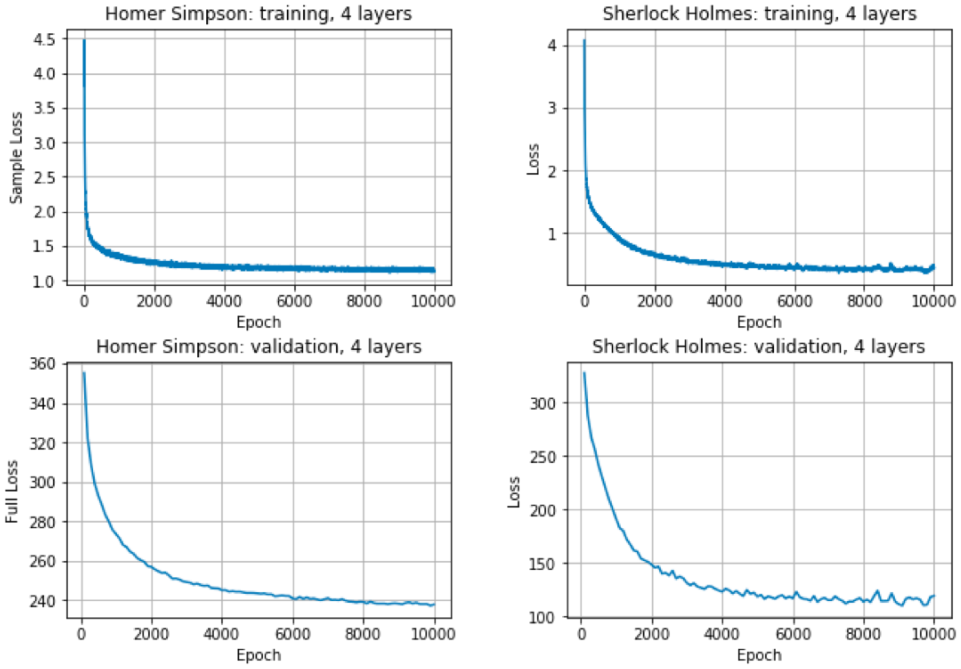
General approach

- Personalized Character RNN-LSTM
  - Generic layers: train first *N* layers with generic data
  - Add *M* more layers and train while freezing the first *N* layers



Training and evaluation results

Model	Layers count	Model Size	Training time	Test loss
Homer	4 (2, 2)	3 MB	154m 25s	237.87
Sherlock	4 (2, 2)	3 MB	155m 35s	119.51



Generation results

Model	Temp	Context	Sentence prediction
Homer	0.8	“<eos>i thought ”	“you'd live in here”
Sherlock			“I've got a text or knump out for the confession”
Homer	0.8	“<eos>marge ”	“three Dad would never do that”
Sherlock		“<eos>watson ”	“is uh the contrast at the girl on the plane and”
Homer	0.5	“<eos>i ”	“nutless separate laws”
Sherlock			“obviously his feet agents of distract me derail me”
Homer	0.9	“<eos>you ”	“alse in my own words you saved my marriage”
Sherlock			“when there's an idiot a shooting”
Homer	0.8	“<eos> can you tell me how”	“thou imm hot anymore”
Sherlock			“long talking to all this time”

Further extension

Model improvements

- Try beam search to improve text generation
- Increase the length of the training sequences
- Optimize so that we can learn faster with fewer amount of data
- Allow further personalization from the user

Application improvements

- Create a mobile device application that leverages these models
- Demonstrate that the model also performs word completion

Demo and code base

- Demo: https://goo.gl/maEFuL
- Code: https://github.com/langholz/identity-word-prediction