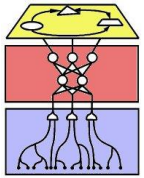
A Study in Recurrency: Generating a Sherlock Holmes Story with an LSTM RNN

Amy Bryce & Theresa Pekarek-Rosin



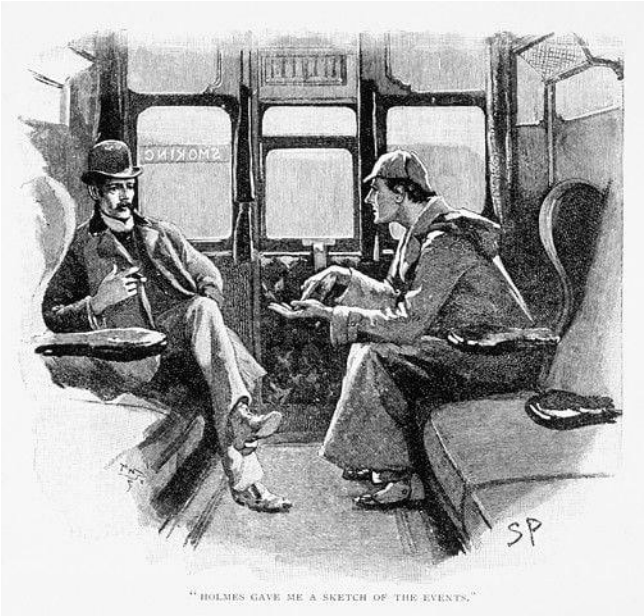
University of Hamburg
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics

Knowledge Technology, WTM

23. January 2019

The Enduring Popularity of Sherlock Holmes

- Published as short stories in Strand magazine by Sir Arthur Conan Doyle beginning in 1887.
- Had a unique format for the time:
  - Stories connected through one main recurring character.
  - Kept every individual story self-contained.
- Source material remains popular.



Sidney Paget, ‘Sherlock Holmes and Watson’, 1893

Why Build a Story Generator?

- There is a high demand for original Sherlock Holmes stories.
- Arthur Conan Doyle is no longer around to write more stories.
- Can we use AI as solution?

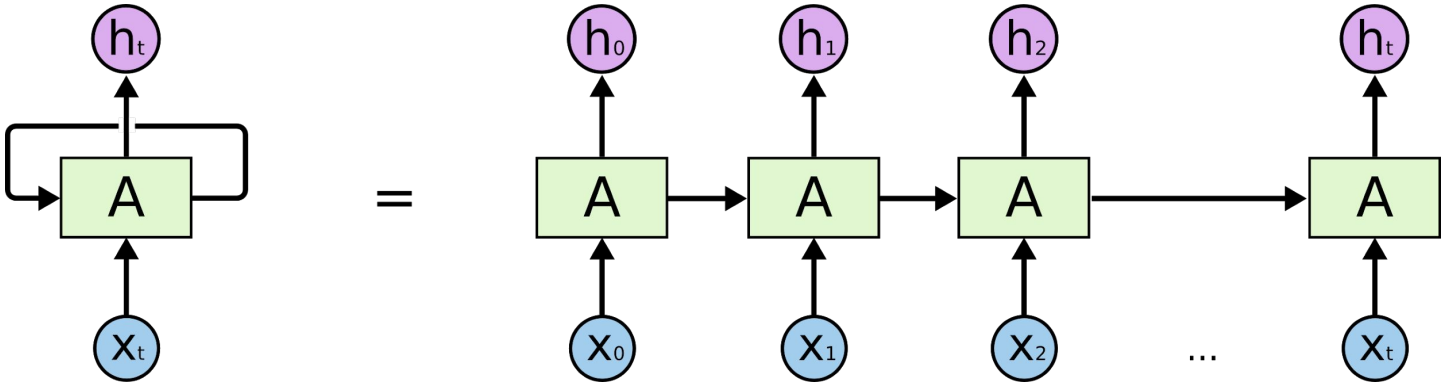
"Sherlock", BBC 2017

Outline

1. Background
   - RNN
   - LSTM
   - Word2Vec
2. Implementation
   - Sherlock Holmes Data
   - Architecture
3. Discussion & Evaluation
4. Summary

Background
Recurrent neural networks (RNN)

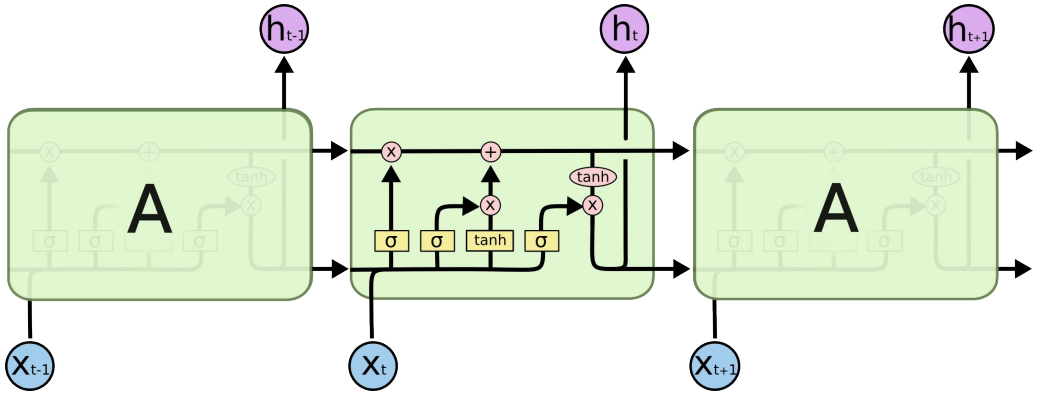
- Introduction of context into the calculation of the output through recurrency
- Ideal for the generation of sequences



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Background
Long short-term memory (LSTM)

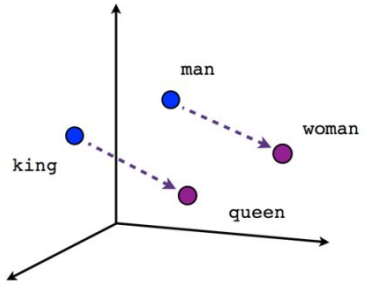
- Special form of RNN
- Consist of input, forget and output gate that regulate the cell state
- LSTM learn structure over long and short-term dependencies
- Work well with word-based generation approach



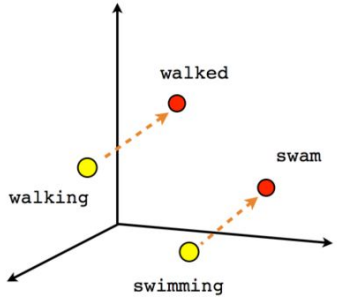
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Background
Word2Vec

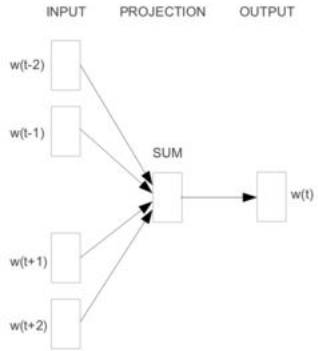
- Grouping words together by distance in the training set
- Text generation not completely at random
- Continuous Bag Of Words vs Skip-gram



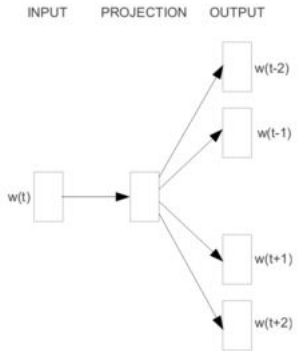
Male-Female



Verb tense



CBOW



Skip-gram

https://towardsdatascience.com/deep-learning-4-embedding-layers-f9a02d55ac12

https://skymind.ai/wiki/word2vec

Implementation

Sherlock Holmes Data: Pre-processing

Input text from the short story, “Silver Blaze”:

```
  "I am afraid, Watson that I shall have to go," said Holmes as
we sat down together to our breakfast one morning.

  "Go! Where to?"

  "To Dartmoor; to King's Pyland."

  I was not surprised. Indeed, my only wonder was that he had
not already been mixed up in this extraordinary case, which was
the one topic of conversation through the length and breadth of
England. For a whole day my companion had rambled about the
room with his chin upon his chest and his brows knitted, charg-
ing and recharging his pipe with the strongest black tobacco, and
absolutely deaf to any of my questions or remarks. Fresh editions
of every paper had been sent up by our news agent, only to be
glanced over and tossed down into a corner.
```

Pre-processed text from the short story, “Silver Blaze”:

```
[… '"', 'i', 'am', 'afraid', ',', 'watson', 'that', 'i', 'shall', 'have', 'to', 'go',
',', '"', 'said', 'holmes', 'as', 'we', 'sat', 'down', 'together', 'to', 'our',
'breakfast', 'one', 'morning', '.', '\n', '\n', ' ', ' ', '"', 'go', '!', 'where', 'to',
'?', '"', '\n', '\n', ' ', ' ', '"', 'to', 'dartmoor', ';', 'to', "king's", 'pyland',
'.', '"', '\n', '\n', ' ', ' ', 'i', 'was', 'not', 'surprised', '.', 'indeed', ',',
'my', 'only', 'wonder', 'was', 'that', 'he', 'had', 'not', 'already', 'been', 'mixed',
'up', 'in', 'this', 'extraordinary', 'case', ',', 'which', 'was', 'the', 'one', 'topic',
'of', 'conversation', 'through', 'the', 'length', 'and', 'breadth', 'of', 'england',
'.', 'for', 'a', 'whole', 'day', 'my', 'companion', 'had', 'rambled', 'about', 'the',
'room', 'with', 'his', 'chin', 'upon', 'his', 'chest', 'and', 'his', 'brows', 'knitted',
',', 'charging', '\n', ' ', 'and', 'recharging', 'his', 'pipe', 'with', 'the',
'strongest', 'black', 'tobacco', ',', 'and', 'absolutely', 'deaf', 'to', 'any', 'of',
'my', 'questions', 'or', 'remarks', '.', 'fresh', 'editions', 'of', 'every', 'paper',
'had', 'been', 'sent', 'up', 'by', 'our', 'news', 'agent', ',', 'only', 'to', 'be',
'glanced', 'over', 'and', 'tossed', 'down', 'into', 'a', 'corner', '.', …]
```

Implementation

Sherlock Holmes Data: Post-processing

Pre-processed output for “It was the best of times, it was the worst of times”:

```
['it', 'was', 'the', 'best', 'of', 'times', ',', 'it', 'was', 'the', 'worst',
'of', 'times', '.', 'this', 'man', 'was', 'not', 'an', 'need', '.', 'he', 'is',
'not', 'held', 'to', 'a', 'very', 'plate', ',', 'who', 'had', 'only', 'only',
'a', 'very', 'of', 'his', 'weight', 'hundred', 'during', '.', 'the', 'behind',
',', 'that', 'one', 'of', 'them', 'was', 'that', 'of', 'the', 'perfect', 'of',
'that', 'which', 'was', 'a', 'very', 'to-night', 'but', 'you', 'and', 'set',
'the', 'return', 'to', 'the', 'chair', '.', '"', '\n', '\n', ' ', ' ', '"',
'my', 'dear', 'watson', ',', 'i', 'am', 'inspector', 'that', 'it', 'is', 'not',
'quite', 'so', '.', 'you', 'can', 'see', ',', 'watson', ',', 'that', 'you',
'in', 'your', 'household', 'is', 'believe', 'very', 'age', '.', 'you', 'are',
'inspector', 'that', 'she', 'got', 'mr', '.', 'holmes', '?', …]
```

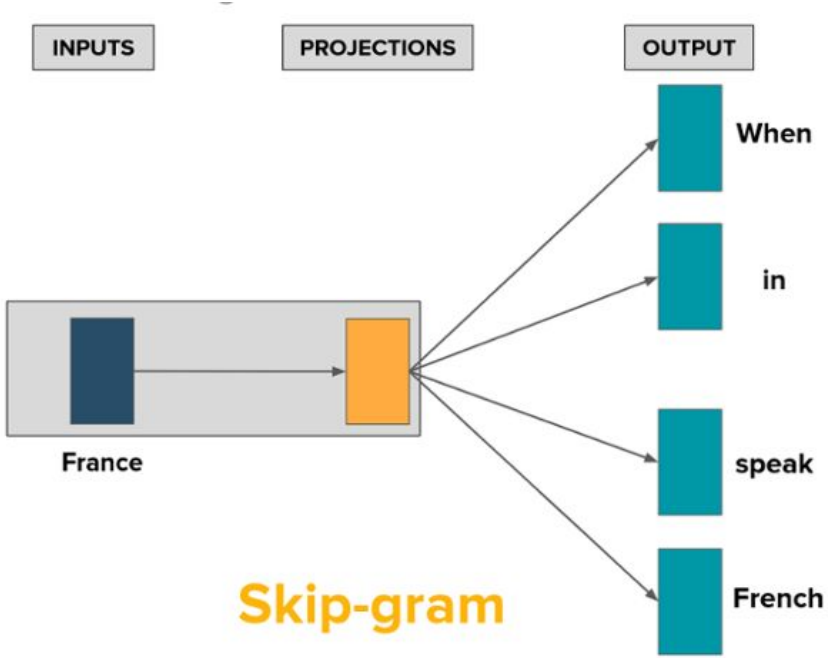
Post-processed output for “It was the best of times, it was the worst of times”:

```
It was the best of times, it was the worst of times. This man was not an need.
He is not held to a very plate, who had only only a very of his weight hundred
during. The behind, that one of them was that of the perfect of that which was a
very to-night but you and set the return to the chair."

  "My dear watson, I am inspector that it is not quite so. You can see, watson,
that you in your household is believe very age. You are inspector that she got
mr. Holmes?" …
```

Implementation

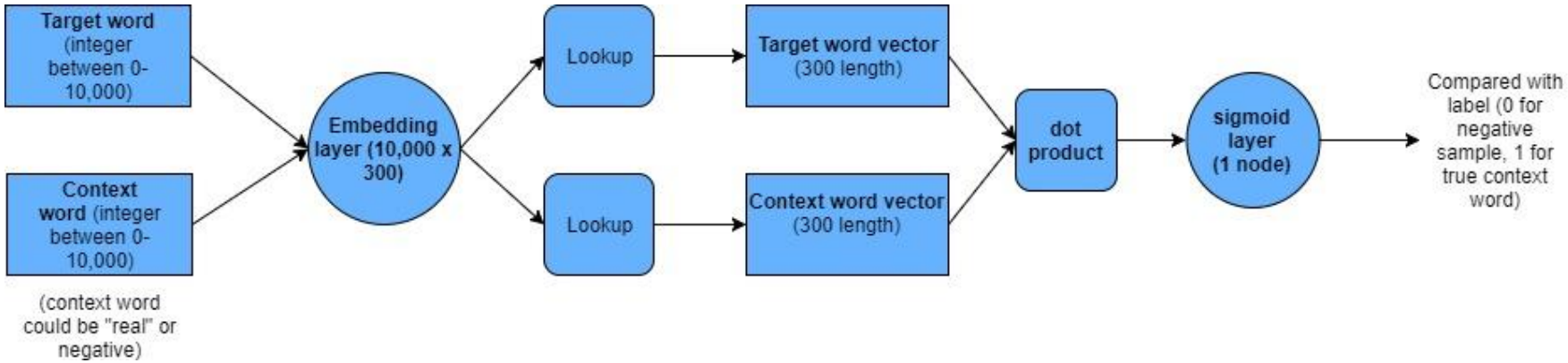
Architecture: Word2Vec; Skip-gram



https://www.datascience.com/blog/word-embeddings-natural-language-processing

Implementation

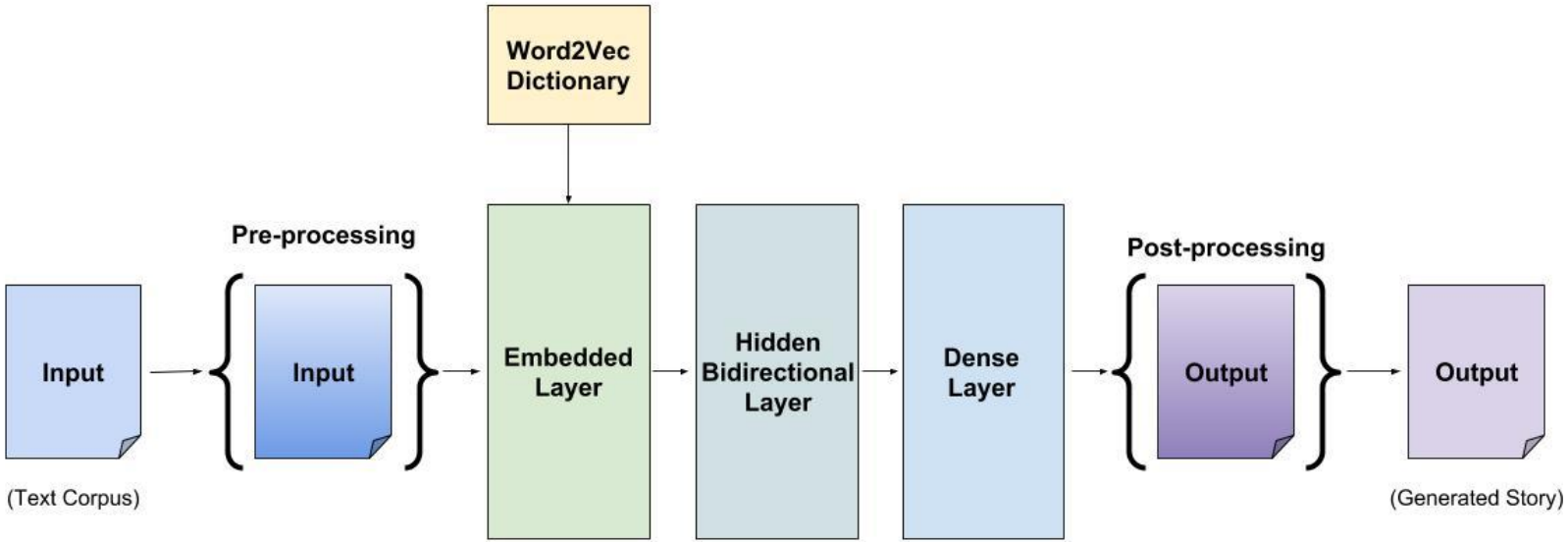
Architecture: Word2Vec; Negative Sampling



https://adventuresinmachinelearning.com/word2vec-keras-tutorial/

Implementation

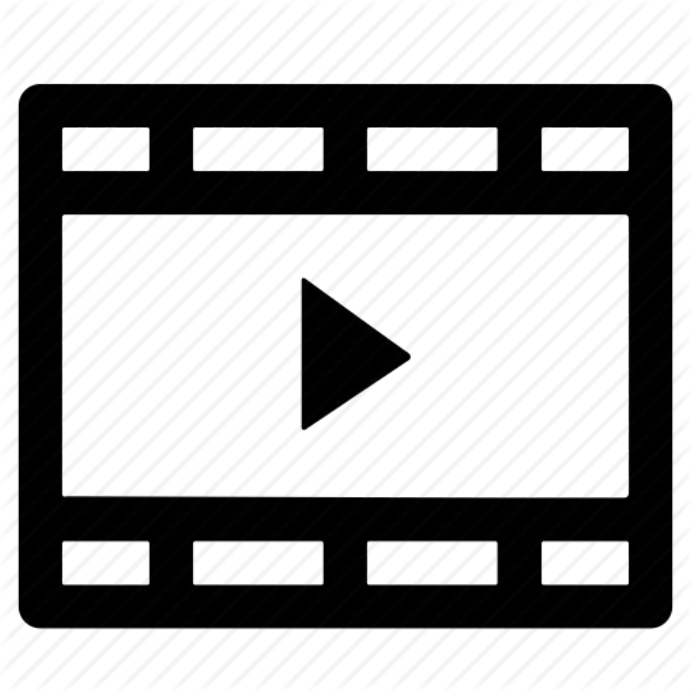
Architecture: LSTM



Network Architecture for Sherlock Holmes LSTM Story Generator

Implementation

Architecture: LSTM; Inference



Discussion & Evaluation

- Achieved recognizable structure of Sherlock Holmes story through:
  - Detailed pre- and post-processing
  - Inclusion of punctuation in training data

"The murder was on the same and the place?"

"Yes, yes, it was a man who is front. It is a very bottom one of the one of the ran who has left her second. In this way you have made up it? I think that, I know, and I have no doubt that I have seen it before."

"I think it is on which you have been very circle."

"Help, my dear fellow,"said he.

"You are a man of a very singular scandal. Your myself is a circus."

Discussion & Evaluation

- Coherence still lacking, could be improved through:
  - Entity recognition [Clark 2018]
  - Connection of inference and training process [Goodfellow 2018]

- Further improvement of output quality through:
  - Higher number of hidden layers
  - Exclusion of low-appearance words

- For the future:
  - Evaluation through non-biased observers
  - Plagiarism checker for overfitting

"I will have you the matter, watson,"said he."You are your shook in the room?"

"Yes, sir,"said he."I am investigation that it is south that the fell decent to be a man who has been left. With a until strong show would have gone to do it you."

"I do not know what you would say. If the matter is not in your own shut and. But I am brought that I have not seen it all."

"Help, mrs! What do you order?"He cried.

"I have the first words the appear of the drew, who was at the very night,"said he, as he else in the england; you see him a few fitzroy of paper."

"The off of them have you floor to me in a most leaving, or you to give them down at the rushed and the very deal of the let lady, and you know the whose of that professor, and they were just after above unless before. They found it out for a time we heard the sound of them found it the means unique. Perhaps it was only a very trust body that he had been in the case from which it had been so poor in his fashion.

"'There is a narrative. It is that you may have been here in the matter from the same certainty by the always that stood. Gasped! Is you will young us give the myself into the matter."

Holmes lead,"i think, just as I had seen a reply powerful as that the night and thing in the that's and had been walked. It was clear and dr it. We had been iron for the mad of the evening.

Summary

- Built a LSTM RNN with Word2Vec word embedding
- Trained both models on the collection of Sherlock Holmes short stories
- Achieved a recognizable structure through detailed pre- and postprocessing

Thank you for your attention.