



Wikipedia job extractor

by John Helbrink and Love Malmros

Introduction



- Text classification is a powerful tool in a lot of areas.
- Creating a deep learning algorithm for classifying people's profession based on first paragraph.
- Utilized tools such as Python, Keras, Wikidata, Wikipedia and Sparql.

What we are going to talk about today



- Related work
- What was the problem?
- How did we solve the problem?
- Preprocessing
- Architecture
- Results
- Demo
- Further improvements

Related work



- Earlier project.
- Solved the problem part-of-speech and dependency relations.
- We wanted to solve it by using deep learning.

How did we solve the problem?



- **Keras** is a high-level neural networks API, written in python.
- **Wikidata** is a central database storage that contains structured wikipedia data. Extraction using SPARQL
- **Wikipedia** contained the relevant first paragraph for each person.

Extracting data



Extract occupations from wikidata

Used SPARQL to extract occupations from 900 000 people. Converted to a json file for easier processing.

Extract first paragraph from wikipedia

Combined the json file with people and corresponding jobs with docria file containing the first paragraph from Wikipedia.

Combining the jobs with the paragraphs.

Generated the final corpus used to train our model containing all the relevant information needed. ~900k people.

Preprocessing


The texts:

- Tokenize, max length = 150
- To integer sequences
- Pad sequences with 0s

The occupations:

- One hot encoded [1 0 0 1 0 0 0]





Scaling the input data in order to make it less imbalanced.



- Imbalanced with a majority (10%) of the jobs being politicians.

2 new sets:

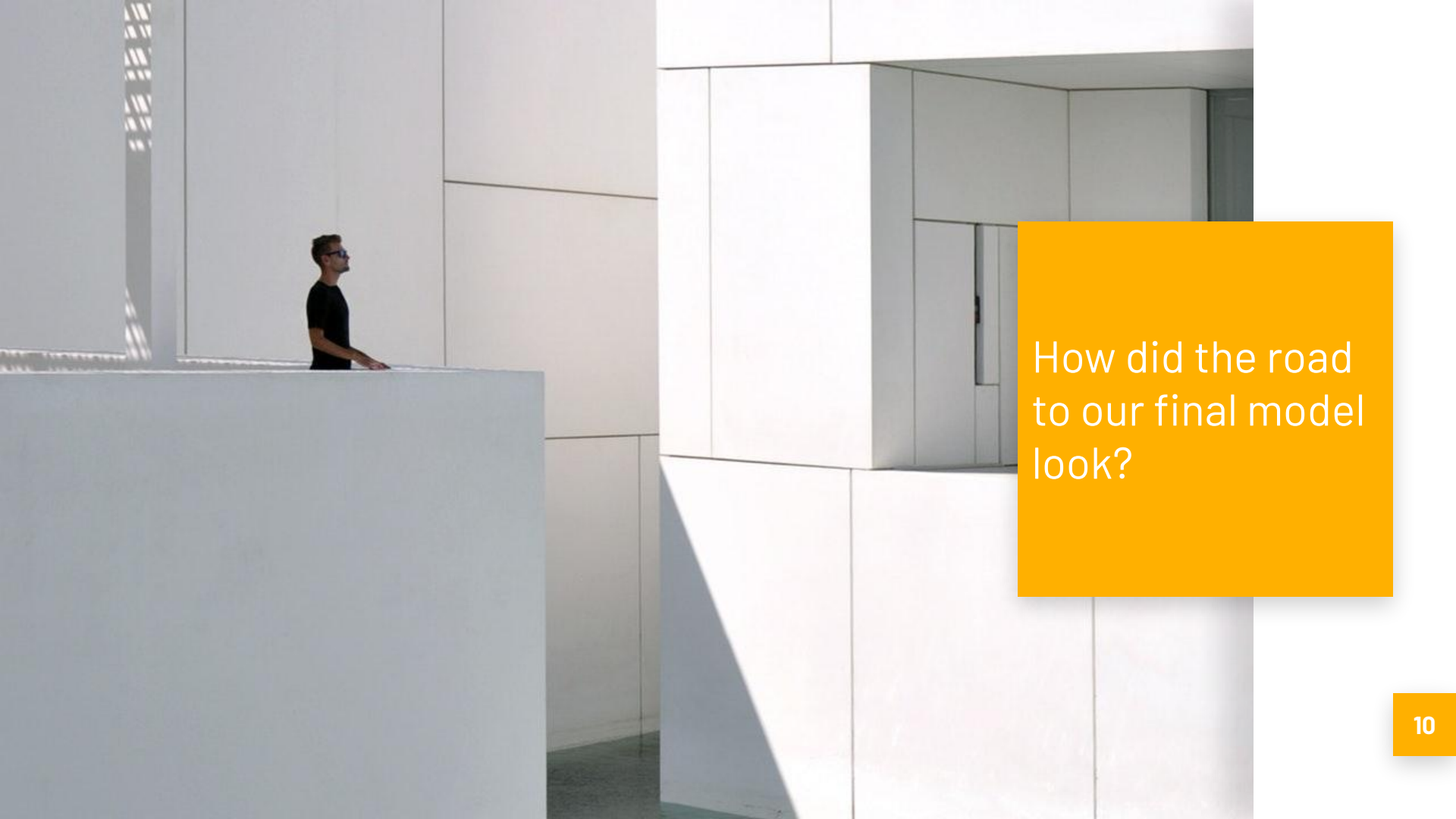
- 50 - 500 (438 jobs)
- 100 - 500 (247 jobs)

Architecture

Constructed our deep learning network by using:

- Linear stack of layers (Sequential)
- Embedding layer
- Bidirectional LSTM (Long short term memory)
- two hidden layer (dense)
- Softmax and categorical cross entropy





How did the road
to our final model
look?

What different models did we create?



First draft (Binary classification)

Created a model with binary classification, determining whether a person had a certain job or not.



Second draft (Feed forward network)

Did not produce good results, however was a step in the right direction!



Final model (Bidirectional LSTM)

Created a few different models using LSTM with varying results. Big improvements when scaling imbalanced data.

F1, precision, recall and accuracy



Limit of 50 - 500
labels.

438 classes

Precision	Recall	F1	Accuracy
0.818	0.809	0.813	0.772

Limit of 100 - 500
labels.

247 classes

Precision	Recall	F1	Accuracy
0.826	0.883	0.853	0.821

Model (50 - 500) sample



Using the model with 438 classes, we supplied the model with the first paragraph corresponding to three people.

The model (50-500) predicted:

- Edmund (or Eadmund; died 1041) was Bishop of Durham from 1021 to 1041.
PREDICTED: PRIEST WIKIDATA: PRIEST
- Elita Krūmiņa (born 21 November 1965 in Jelgava) has been the auditor general of Latvia since 21 January 2013. Prior to becoming auditor general, Krūmiņa was a member of the council of Latvia's state audit office; she became a member in 2005 after spending six years at the Latvian Ministry of Finance.
PREDICTED: POLITICIAN WIKIDATA: BANKER AND ECONOMIST
- Heriberto Andrés Bodeant Fernández (born 15 June 1955 in Young, Río Negro Department) is a Uruguayan Roman Catholic cleric.
PREDICTED: CATHOLIC PRIEST WIKIDATA: CATHOLIC PRIEST

Model (100 - 500) sample



Using the model with 247 classes, we supplied the model with the first paragraph corresponding to three people.

The model (100-500) predicted:

- John A. ("Jack") LaSota is a former Arizona Attorney General (1977-1978). LaSota also served as Bruce Babbitt's Chief of Staff when the former was governor of Arizona. He is a lobbyist for the firm LaSota & Peters, P.L.C.

PREDICTED: LOBBYIST AND OTHER

WIKIDATA: LAWYER AND JUDGE

- Ernesto Alciati (3 December 1901 - November 1984) was an Italian long-distance runner.

PREDICTED: ATHLETICS COMPETITOR AND MARATHON RUNNER

WIKIDATA: MARATHON RUNNER

Future work



- Extend the algorithm to support other languages aswell.
- Create usable web application for classifying texts.
- Retrieve more data to increase accuracy and number of classes.



THANKS!

Any questions?

John Helbrink, mat14jhe@student.lu.se

Love Malmros, sta15lma@student.lu.se