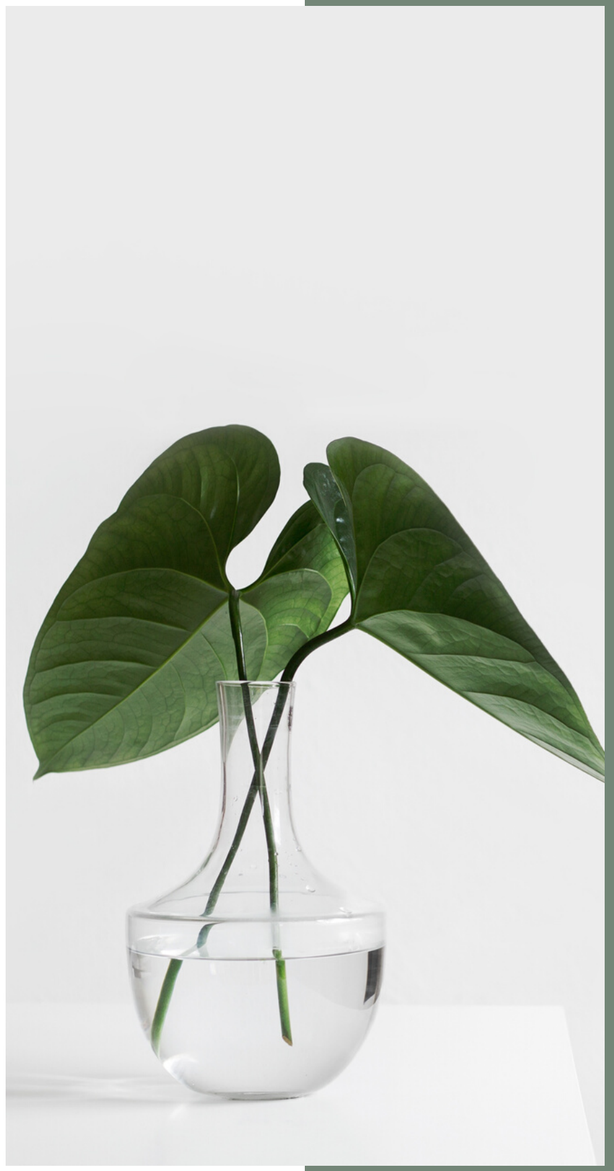
01

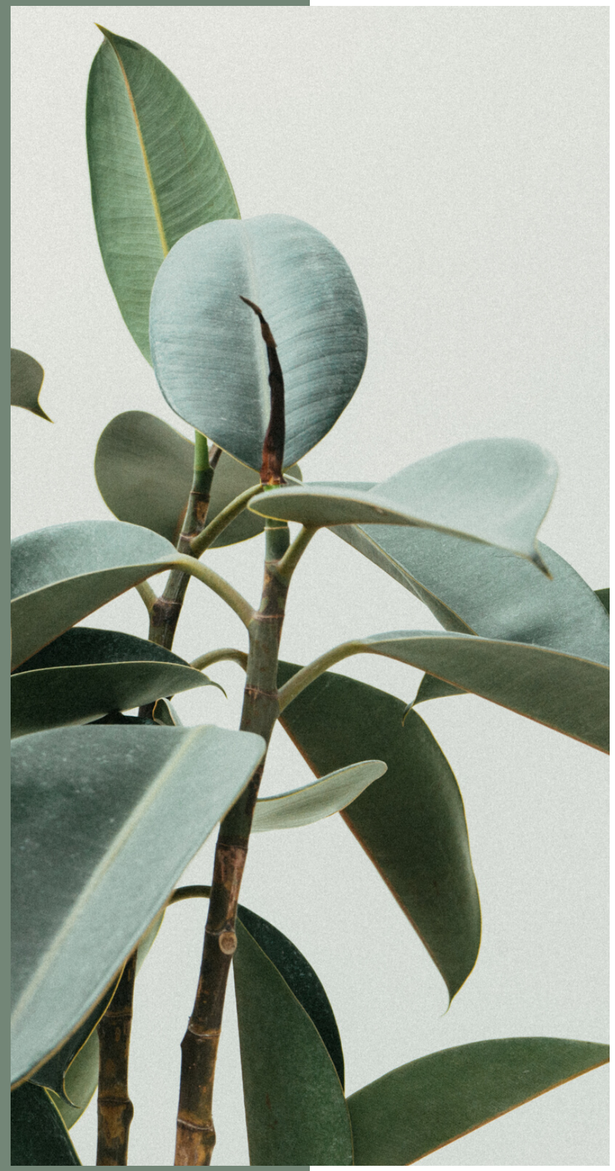


Machine Language Translation

MINI PROJECT V

ROBYN MUNDLE

02



Comtrans Corpora

German to English

Wiederaufnahme der Sitzungsperiode

Resumption of the session

0-0 1-1 1-2 2-3

German to French

Wiederaufnahme der Sitzungsperiode

Reprise de la session

0-0 1-1 1-2 2-3

English to French

Resumption of the session

Reprise de la session

0-0 1-1 2-2 3-3

German > English

Recurrent Neural Networks (RNN)

-- Long Short-Term Memory (LSTM)

- Explore Dataset
- Preprocessing
- Modeling
- Prediction

Sampling the Corpora

German sample 1: ['Wiederaufnahme', 'der', 'Sitzungsperiode']
English sample 1: ['Resumption', 'of', 'the', 'session']

German sample 2: ['Ich', 'erkläre', 'die', 'am', 'Freitag', ',', 'dem', '17.', 'Dezember', 'unterbrochene', 'Sitzungsperiode', 'des', 'Europäischen', 'Parlaments', 'für', 'wiederaufgenommen', ',', 'wünsche', 'Ihnen', 'nochmals', 'alles', 'Gute', 'zum', 'Jahreswechsel', 'und', 'hoffe', ',', 'daß', 'Sie', 'schöne', 'Ferien', 'hatten', '.']
English sample 2: ['I', 'declare', 'resumed', 'the', 'session', 'of', 'the', 'European', 'Parliament', 'adjourned', 'on', 'Friday', '17', 'December', '1999', ',', 'and', 'I', 'would', 'like', 'once', 'again', 'to', 'wish', 'you', 'a', 'happy', 'new', 'year', 'in', 'the', 'hope', 'that', 'you', 'enjoyed', 'a', 'pleasant', 'festive', 'period', '.']

Sampling the Corpora

666,937 German words

36,146 unique German words

10 most common:

, . die der und in zu den ist daß

710,091 English words

19,231 unique English words

10 most common:

the . , of to and in is a that

Preprocess Pipeline

Clean

tokenize punctuation and lowercase all tokens

Tokenize

convert each word into a number (word IDs)

Filter Length

filtered sentences longer than 20 words out (for either language)

Padding

added padding to the end of the sequences to make sentences between the languages the same length

06

Model: Sequential

Layer	Output Shape
Embedding	(None, 20, 1024)
LSTM [0]	(None, 20, 1024)
LSTM [1]	(None, 20, 512)
LSTM [2]	(None, 20, 512)
LSTM [3]	(None, 20, 512)
Dropout (20%)	(None, 20, 512)
Time Distributed Dense	(None, 20, 512)
Dropout (50%)	(None, 20, 512)
Time Distributed Dense	(None, 20, 17344)

Total Parameters: 59,934,144

Epochs: 25

Batch Size (Batches) 50 (237)

German to English Translation Performance

44%

TRAINING ACCURACY

38%

VALIDATION ACCURACY

<10h

TRAINING TIME

~0.0

BLEU SCORE

07

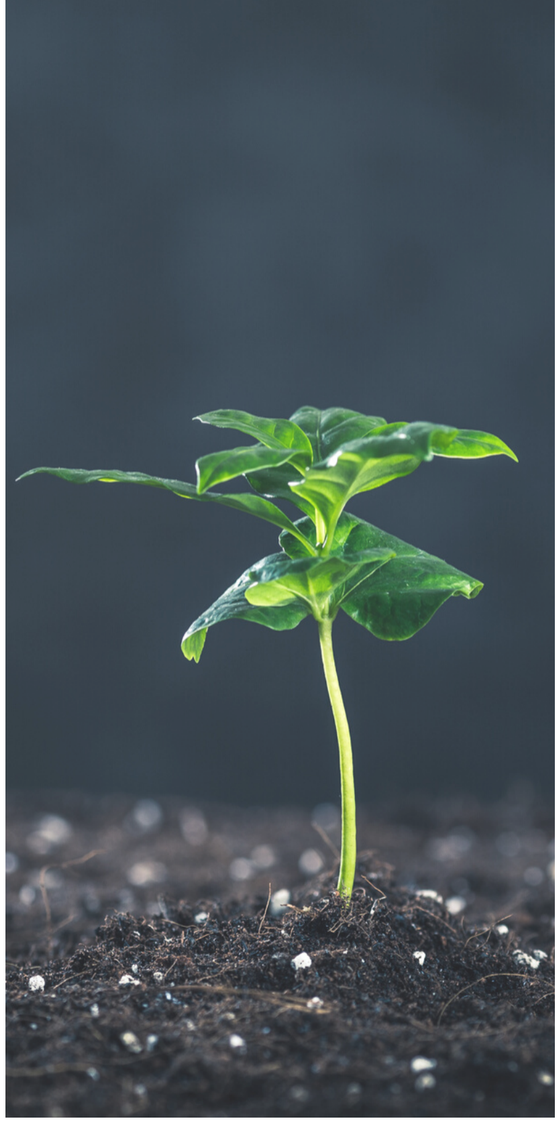
Language Predictions

GER: Könnten Sie mir eine Auskunft zu Artikel 143 im Zusammenhang mit der Unzulässigkeit geben ?
ENG: I would like your advice about Rule 143 concerning inadmissibility .
Pred: thank you very much commissioner the
BLEU score: 1.34830652806260 46e-231

GER: Deswegen beantragt meine Fraktion , diesen Punkt von der Tagesordnung abzusetzen
ENG: That is why my Group moves that this item be taken off the agenda .
Pred: i course , the , to . . . .
BLEU score: 1.3135841289152546e-231

GER: Wir kommen nun zu Herrn Wurtz , der gegen den Antrag spricht .
ENG: We shall now hear Mr Wurtz speaking against this request .
Pred: the is is is the the the the the the the the . .
BLEU score: 1.09316166540311 89e-231

GER: Frau Präsidentin , ich möchte zunächst darauf hinweisen , daß das , was Herr Poettering da sagt , nicht ganz logisch ist .
ENG: Madam President , I would firstly like to point out Mr Poettering ' s lack of logic .
Pred: mr is , , , to to the . .
BLEU score: 1.3319603978104 45e-231



Reflection

Deep learning is slow.

Stacking additional LSTM hidden layers is understood to recombine the learned representation from prior layers and create new representations at high levels of abstraction.

Removing Dropout between LSTM: LSTMs are good for long terms but an important thing about them is that they are not very well at memorising multiple things simultaneously. Dropouts inside the stack leave a chance for forgetting something that should not be forgotten.

Thanks