MACHINE TRANSLATION

Converting German Text to English

Mayank Jha(16bcs023) & Shreya Pandita(16bcs046)
Mentor: Dr. Naveen Gondhi

OVERVIEW

1. INTRODUCTION
2. DATA PREPARATION
3. THE MODEL
4. RESULTS

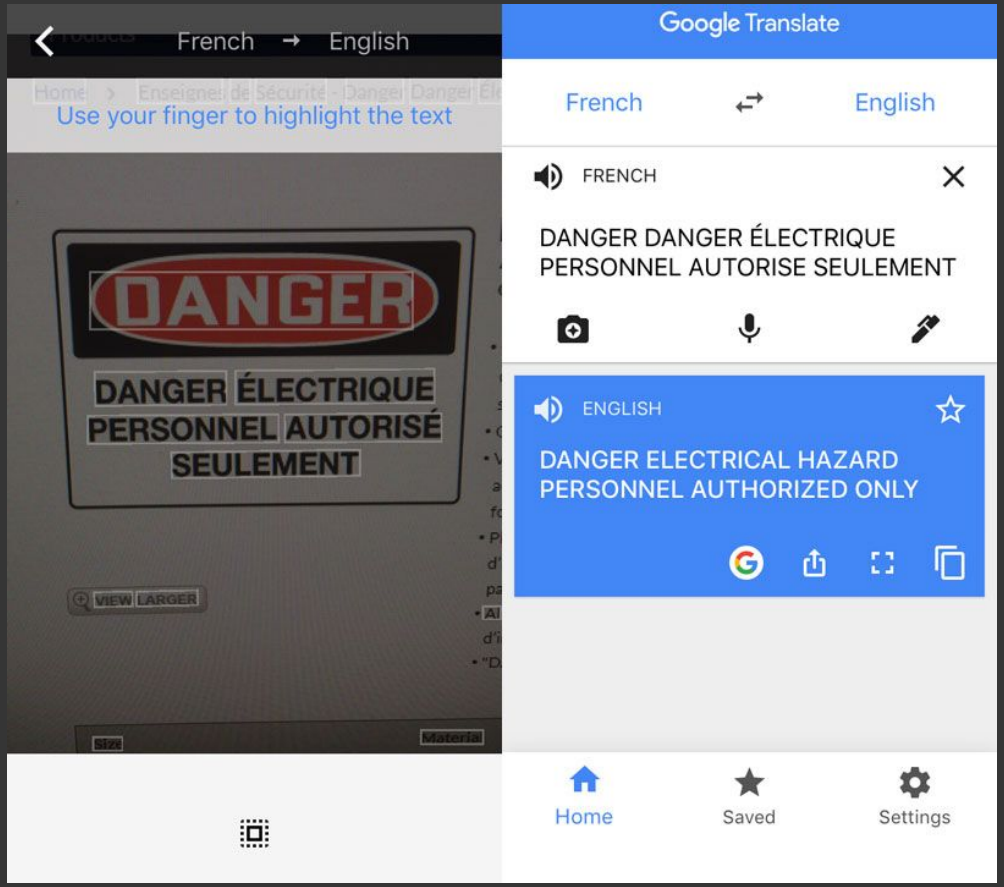
INTRODUCTION

How many languages do you need to know to communicate with the rest of the world?

Just one!

Your own.

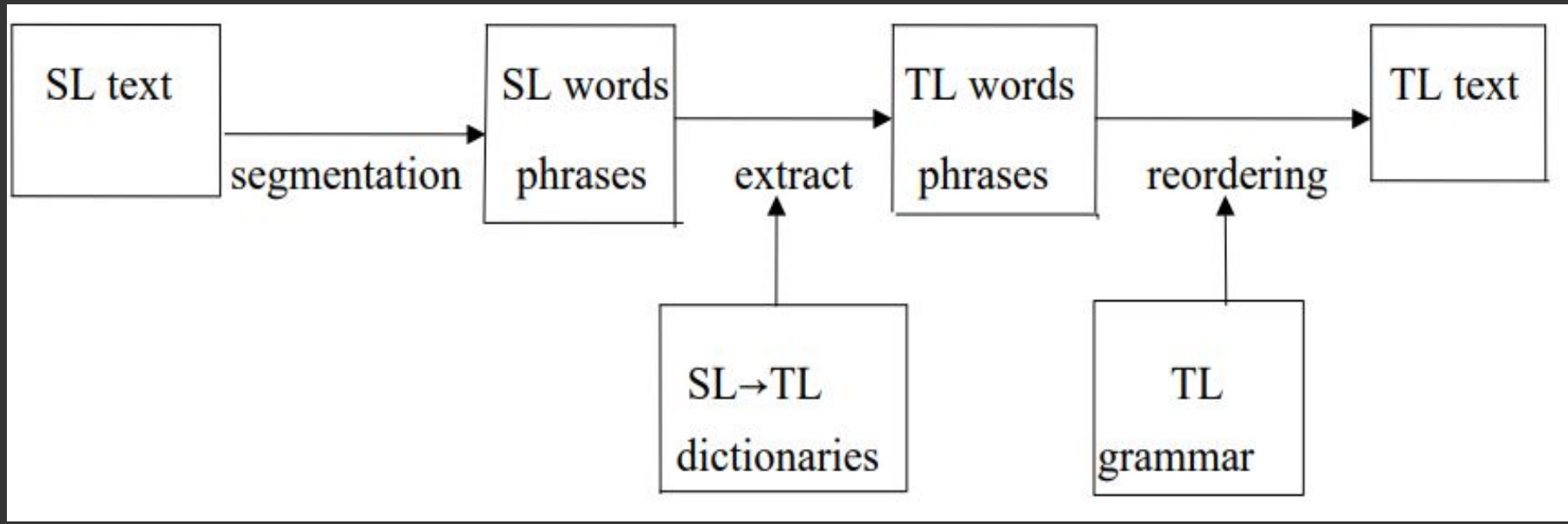
(With a little help from your smartphone)



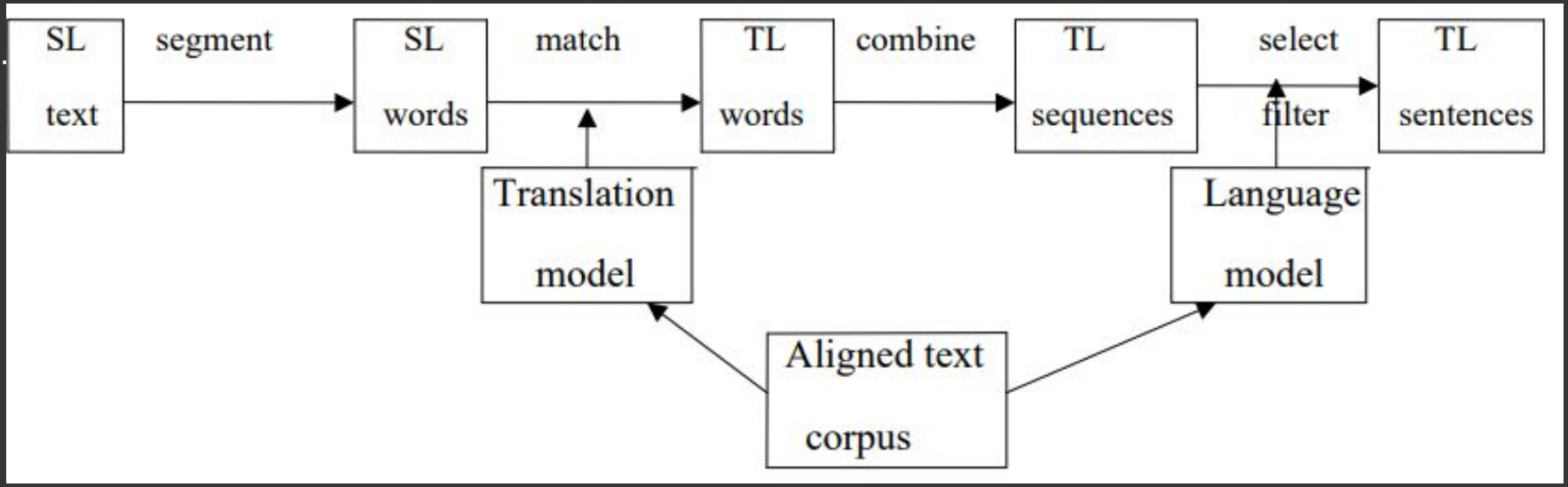
MACHINE TRANSLATION

The translation of text by a computer, with no human involvement.

RBMT (Rule Based)
SMT (Statistical)
NMT (Neural)



Rule Based MT



Statistical MT

Data Preparation

The Dataset

```
line=open('deu.txt','r',encoding='utf-8')
text=line.read()
text
```

```
'Hi.\tHallo!\nHi.\tGrüß Gott!\nRun!\tLauf!\nFire!\tFeuer!\nHelp!\tHilfe!\nHelp!\tZu Hülf!\nStop!\tStopp!\nWait!\tWarte!\nGo o
n.\tMach weiter.\nHello!\tHallo!\nI ran.\tIch rannte.\nI see.\tIch verstehe.\nI see.\tAha.\nI try.\tIch probiere es.\nI won!
\tIch hab gewonnen!\nI won!\tIch habe gewonnen!\nSmile.\tLächeln!\nCheers!\tZum Wohl!\nFreeze!\tKeine Bewegung!\nFreeze!\tSte
henbleiben!\nGot it?\tVerstanden?\nGot it?\tEinverstanden?\nHe ran.\tEr rannte.\nHe ran.\tEr lief.\nHop in.\tMach mit!\nHug m
e.\tDrück mich!\nHug me.\tNimm mich in den Arm!\nHug me.\tUmarme mich!\nI fell.\tIch fiel.\nI fell.\tIch fiel hin.\nI fell.\t
Ich stürzte.\nI fell.\tIch bin hingefallen.\nI fell.\tIch bin gestürzt.\nI know.\tIch weiß.\nI lied.\tIch habe gelogen.\nI lo
st.\tIch habe verloren.\nI'm 19.\tIch bin 19 Jahre alt.\nI'm 19.\tIch bin 19.\nI'm OK.\tMir geht's gut.\nI'm OK.\tEs geh
t mir gut.\nI'm up.\tIch bin wach.\nI'm up.\tIch bin auf.\nNo way!\tUnmöglich!\nNo way!\tDas gibt’s doch nicht!\nNo way!\tA
usgeschlossen!\nNo way!\tIn keinster Weise!\nReally?\tWirklich?\nReally?\tEcht?\nReally?\tIm Ernst?\nThanks.\tDanke!\nTry i
t.\tVersuch’s!\nWhy me?\tWarum ich?\nAsk Tom.\tFrag Tom!\nAsk Tom.\tFragen Sie Tom!\nAsk Tom.\tFragt Tom!\nBe cool.\tEntspann
dich!\nBe fair.\tSei nicht ungerecht!\nBe fair.\tSei fair!\nBe nice.\tSei nett!\nBe nice.\tSeien Sie nett!\nBeat it.\tGeh we
g!\nBeat it.\tHau ab!\nBeat it.\tVerschwinde!\nBeat it.\tVerdufte!\nBeat it.\tMach dich fort!\nBeat it.\tZieh Leine!\nBeat i
t.\tMach dich vom Acker!\nBeat it.\tVerzieh dich!\nBeat it.\tVerkrümele dich!\nBeat it.\tTroll dich!\nBeat it.\tZisch ab!\nBe
at it.\tPack dich!\nBeat it.\tMach ’ne Fliege!\nBeat it.\tSchwirr ab!\nBeat it.\tMach die Sause!\nBeat it.\tScher dich weg!\n
Beat it.\tScher dich fort!\nCall me.\tRuf mich an.\nCome in.\tKomm herein.\nCome in.\tHerein!\nCome on!\tKomm!\nCome on!\tKom
mt!\nCome on!\tMach schon!\nCome on!\tMacht schon!\nCome on.\tKomm schon!\nGet Tom.\tHol Tom.\nGet out!\tRaus!\nGet out.\tGeh
raus.\nGo away!\tGeh weg!\nGo away!\tHau ab!\nGo away!\tVerschwinde!\nGo away!\tVerdufte!\nGo away!\tMach dich fort!\nGo awa
y!\tZieh Leine!\nGo away!\tMach dich vom Acker!\nGo away!\tVerzieh dich!\nGo away!\tVerkrümele dich!\nGo away!\tTroll dich!\n
Go away!\tZisch ab!\nGo away!\tPack dich!\nGo away!\tMach ’ne Fliege!\nGo away!\tSchwirr ab!\nGo away!\tMach die Sause!\nGo a
```

CLEANING THE DATA

- ➔ Removing all non printable characters
- ➔ Removing all punctuations
- ➔ Normalising unicode characters to ASCII
- ➔ Converting all text to lowercase
- ➔ Removing numerical tokens

```
clean_pairs
```

```
array([['hi', 'hallo'],
       ['hi', 'hallo'],
       ['hi', 'gru gott'],
       ...,
       ['if someone who doesnt know your background says that you sound like a native speaker it means they probably noticed so
mething about your speaking that made them realize you werent a native speaker in other words you dont really sound like a nati
ve speaker',
        'wenn jemand der deine herkunft nicht kennt sagt dass du wie ein muttersprachler sprichst bedeutet das dass man wahrsch
einlich etwas an deiner sprechweise bemerkt hat das erkennen lie dass du kein muttersprachler bist mit anderen worten du horst
dich nicht wirklich wie ein muttersprachler an'],
       ['if someone who doesnt know your background says that you sound like a native speaker it means they probably noticed so
mething about your speaking that made them realize you werent a native speaker in other words you dont really sound like a nati
ve speaker',
        'wenn jemand fremdes dir sagt dass du dich wie ein muttersprachler anhorst bedeutet das wahrscheinlich er hat etwas an
deinem sprechen bemerkt dass dich als nichtmuttersprachler verraten hat mit anderen worten du horst dich nicht wirklich wie ein
muttersprachler an'],
```

```
def clean_pairs(lines):
        cleaned = list()
        # prepare regex for char filtering
        re_print = re.compile('[^%s]' % re.escape(string.printable))
        # prepare translation table for removing punctuation
        table = str.maketrans('', '', string.punctuation)
        for pair in lines:
                clean_pair = list()
                for line in pair:
                    # normalize unicode characters
                    line = normalize('NFD', line).encode('ascii', 'ignore')
                    line = line.decode('UTF-8')
                    # tokenize on white space
                    line = line.split()
                    # convert to lowercase
                    line = [word.lower() for word in line]
                    # remove punctuation from each token
                    line = [word.translate(table) for word in line]
                    # remove non-printable chars form each token
                    line = [re_print.sub('', w) for w in line]
                    # remove tokens with numbers in them
                    line = [word for word in line if word.isalpha()]
                    # store as string
                    clean_pair.append(' '.join(line))
                cleaned.append(clean_pair)
        return array(cleaned)
```

PREPARING THE DATA

- 339626 phrase pairs reduced to 14000, split into test, train and validation data
- Prepared tokenizers, vocabulary sizes and maximum lengths for both phrases.
- All sequences encoded to integers and padded to maximum phrase length. Word embedding for input and one-hot encoding for output.

```
print(X_train.shape)
print(X_val.shape)
print(X_test.shape)
print(Y_train.shape)
print(Y_val.shape)
print(Y_test.shape)
```

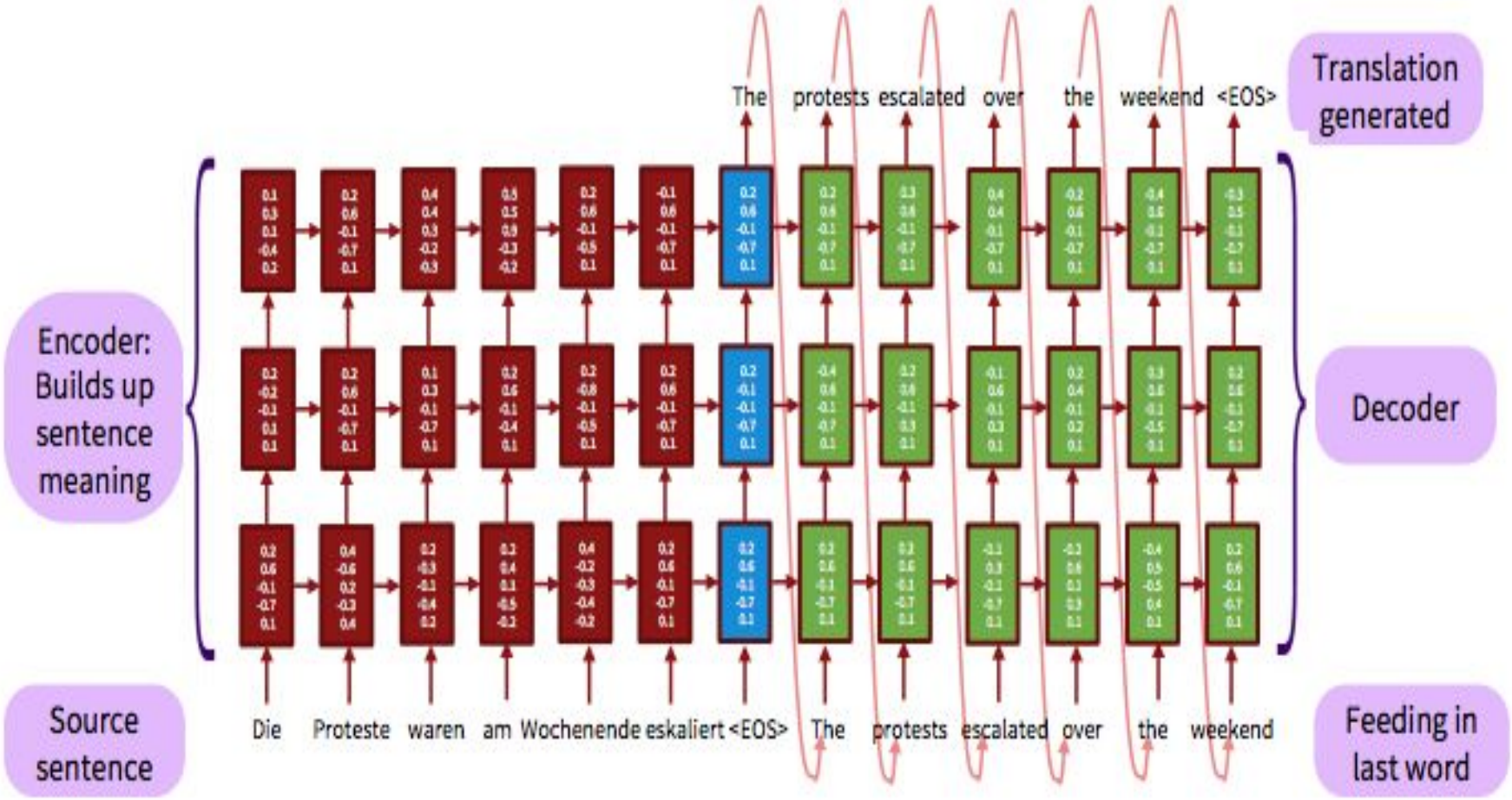
```
(10000, 9)
(2000, 9)
(2000, 9)
(10000, 5)
(2000, 5)
(2000, 5)
```

```
Maximum length in English is 5
Maximum length in German is 9
```

```
English Vocab Size 1797
German Vocab Size 2895
```

THE MODEL

(ENCODER-DECODER LSTM)


Translation generated
The protests escalated over the weekend <EOS>
Encoder: Builds up sentence meaning
Decoder
Source sentence
Die Proteste waren am Wochenende eskaliert <EOS>
Feeding in last word
The protests escalated over the weekend

THE MODEL

```
def define_model(src_vocab, tar_vocab, src_timesteps, tar_timesteps, n_units):
    model = Sequential()
    model.add(Embedding(src_vocab, n_units, input_length=src_timesteps, mask_zero=True))
    model.add(LSTM(n_units))
    model.add(RepeatVector(tar_timesteps))
    model.add(LSTM(n_units, return_sequences=True))
    model.add(TimeDistributed(Dense(tar_vocab, activation='softmax')))
    return model
```

```
model = define_model(ger_vocab_size, eng_vocab_size, 9,5, 256)
model.compile(optimizer='adam', loss='categorical_crossentropy')
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 9, 256)            741120
_________________________________________________________________
lstm_3 (LSTM)                (None, 256)               525312
_________________________________________________________________
repeat_vector_2 (RepeatVecto (None, 5, 256)            0
_________________________________________________________________
lstm_4 (LSTM)                (None, 5, 256)            525312
_________________________________________________________________
time_distributed_1 (TimeDist (None, 5, 1797)           461829
=================================================================
Total params: 2,253,573
Trainable params: 2,253,573
Non-trainable params: 0
```

- Hyperparameter epoch tuned to 30.
- Adam optimizer used to optimize updation of weight
- Categorical loss entropy function used to minimize errors

```
Train on 10000 samples, validate on 2000 samples
Epoch 1/30
 - 14s - loss: 4.0065 - val_loss: 3.2777
Epoch 2/30
 - 10s - loss: 3.1595 - val_loss: 3.1122
Epoch 3/30
 - 10s - loss: 3.0035 - val_loss: 2.9901
Epoch 4/30
 - 10s - loss: 2.8346 - val_loss: 2.8344
Epoch 5/30
 - 9s - loss: 2.6435 - val_loss: 2.6450
Epoch 6/30
 - 10s - loss: 2.4467 - val_loss: 2.4869
Epoch 7/30
 - 10s - loss: 2.2655 - val_loss: 2.3424
```

```
Epoch 24/30
 - 10s - loss: 0.3898 - val_loss: 0.8994
Epoch 25/30
 - 10s - loss: 0.3454 - val_loss: 0.8694
Epoch 26/30
 - 10s - loss: 0.3075 - val_loss: 0.8302
Epoch 27/30
 - 10s - loss: 0.2781 - val_loss: 0.8152
Epoch 28/30
 - 10s - loss: 0.2517 - val_loss: 0.7956
Epoch 29/30
 - 10s - loss: 0.2291 - val_loss: 0.7721
Epoch 30/30
 - 10s - loss: 0.2059 - val_loss: 0.7487
```

RESULTS

- Best model selected during validation
- Evaluated on training and test dataset
- Prediction is a series of integers that we perform reverse mapping on to get the corresponding word, resulting in a string of words for the entire sentence.
- Result compared to the existing value in English.

```
evaluate_model(model, eng_tokens, X_test, test)
```

```
src=[ihr konnt nicht gehen], target=[you cant go], predicted=[you cant go]
src=[es tut uns leid], target=[were sorry], predicted=[youre won]
src=[tom wurde geschnappt], target=[tom got busted], predicted=[tom got busted]
src=[ich bin im netz], target=[i am online], predicted=[im in]
src=[es war nicht meines], target=[it wasnt mine], predicted=[it wasnt mine]
src=[das wurde mir wohl gefallen], target=[id like that], predicted=[id like that]
src=[unterschreibe hier], target=[sign here], predicted=[sign here]
src=[tom kann nicht lesen], target=[tom cant read], predicted=[tom cant try]
src=[niemand ist glucklich], target=[nobodys happy], predicted=[fishing is fun]
src=[tom ist eingebrochen], target=[tom broke in], predicted=[tom broke in]
src=[mach mal schluss damit], target=[give it a rest], predicted=[give it a rest]
src=[versuchs doch einfach], target=[just try it], predicted=[just it on]
src=[melden sie sich], target=[come forward], predicted=[take them]
src=[ich schwimme gern], target=[i like to swim], predicted=[i like singing]
```

THANK YOU