Language Engineering Project
Word Prediction

Axel Larsson
Kamil Mellouk

Task

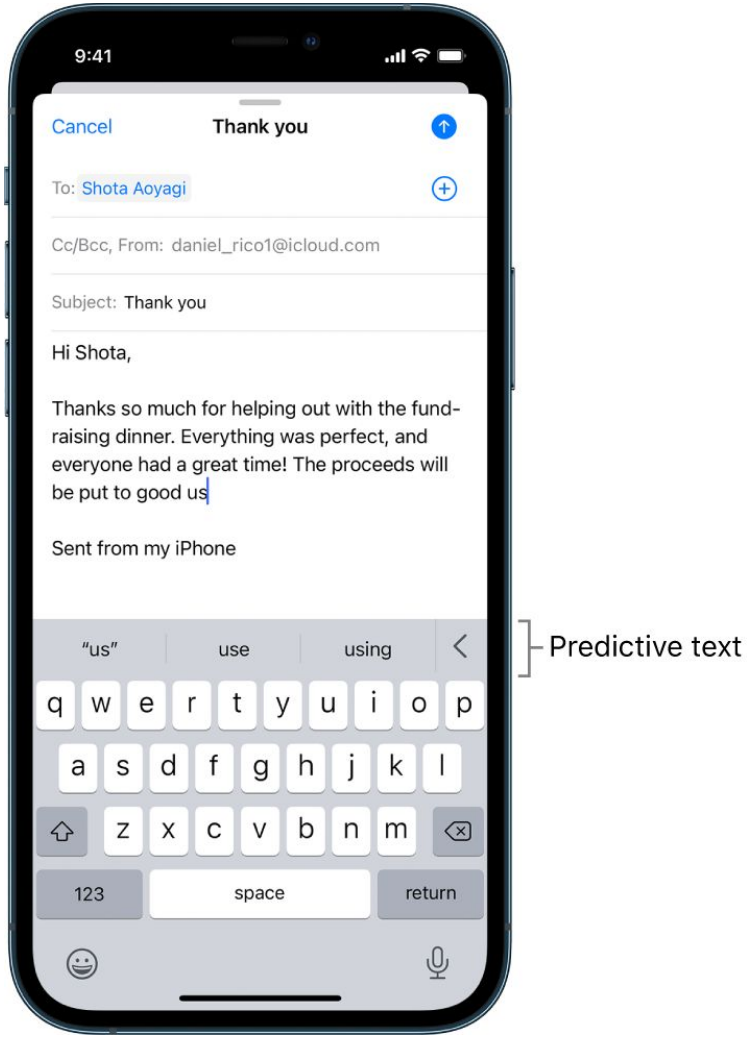


image source

Data

News dataset (~2M lines)

```
Hollywood also wants to avoid having a single company like Apple enticing people to buy only from its own closed digital system, and ending up wit
"I feel like I have a good chance, but I don't feel like I'm anywhere close to the favorite," Ligety said after a downhill training run last week.
The supermarket group will also point to a bonanza in non-food sales, having capitalised on the collapse of Woolworths and Zavvi to sell DVD box s
Canada's Old Bear mauls Norway to win gold
President Nicolas Sarkozy has made fighting unemployment a priority and his government has told Total it must guarantee jobs, in particular since
Here's how semi-legalised piracy works: you wait until the cargo has been offloaded — the cargo's owner and the boat's owner are rarely one and th
If the trend holds, then the 2010 holiday driving season should be equally safe, experts say.
Trikes are a fairly big and therefore more visible," said Jim McGrath, 75, of Chula Vista, whose bright red, low-riding Rewaco trike measures 12 1
And a great human being.
Within a few months, she would be shooting in her own house as the main set from just a six-page outline -- the film itself has no writing credit
```

Blogs dataset(~900k lines)

```
In the years thereafter, most of the Oil fields and platforms were named after pagan “gods”.
We love you Mr. Brown.
Chad has been awesome with the kids and holding down the fort while I work later than usual! The kids have been busy together playing Skylander or
so anyways, i am going to share some home decor inspiration that i have been storing in my folder on the puter. i have all these amazing images st
With graduation season right around the corner, Nancy has whipped up a fun set to help you out with not only your graduation cards and gifts, but
If you have an alternative argument, let's hear it! :)
If I were a bear,
Other friends have similar stories, of how they were treated brusquely by Laurelwood staff, and as often as not, the same names keep coming up. Ab
Although our beloved Cantab can't claim the international recognition afforded the Station Inn, otherwise these two joints feel like twins separat
Peter Schiff: Hard to tell. It will look pretty bad for most Americans when prices will go way up and they can't afford to buy stuff. It could als
```

n-gram approach

Takes into account the n-1 preceding words when computing statistics:

$$P(w_i|w_1,\ldots,w_{i-1}) = P(w_i|w_{i-n+1},\ldots,w_{i-1})$$

Iterates through the text to compute the counts, and saves the probabilities to a model file.

Optionally applies Laplace smoothing to transfer some of the probability mass from the seen sequences to the unseen sequences:

$$P_{Laplace}(w_i|w_{i-1}) = \frac{c(w_{i-1},w_i)+1}{c(w_{i-1})+V}$$

Implementation

```
def predict(self, w0 = "", w1 = None, w2 = None):
    """
    Gives the top 3 predictions from the model given, if available, the current word and the previous two
    """
    predictions = []

    if w1 and w2:
        prev_options = self.trigram_prob.get(self.w2i.get(w2, -1), None)
        if prev_options:
            options = prev_options.get(self.w2i.get(w1, -1), None)
            if options:
                predictions = [self.i2w[i] for (i, _) in options if self.i2w[i][:len(w0)] == w0][:3]
    elif w1 and not w2:
        options = self.bigram_prob.get(self.w2i.get(w1, -1), None)
        if options:
            predictions = [self.i2w[i] for (i, _) in options if self.i2w[i][:len(w0)] == w0][:3]
    else:
        predictions = [self.i2w[i] for (i, _) in self.unigram_count if self.i2w[i][:len(w0)] == w0][:3]

    return predictions
```

Test procedure

```
def check_predictions(self, w0 = "", w1 = None, w2 = None):
    for p in self.predict(w0, w1, w2):
        l0 = len(w0)
        if p[:l0] == w0:
            return len(p) - l0
    return 0

def compute_keystrokes(self, test_filename):
    saved = 0
    total = 0

    for line in tqdm(self.text_gen(test_filename), desc="computing proportion of saved keystrokes", total=2000):
        for i, w in enumerate(line):
            w1 = line[i-1] if i >= 1 else None
            w2 = line[i-2] if i >= 2 else None

            saved += self.check_predictions(w0=w, w1=w1, w2=w2)
            total += len(w)

            for j in range(len(w)):
                saved += self.check_predictions(w0=w[:j], w1=w1, w2=w2)
                total += len(w)

    print("proportion of saved keystrokes: %.6f" % (saved / total))
```

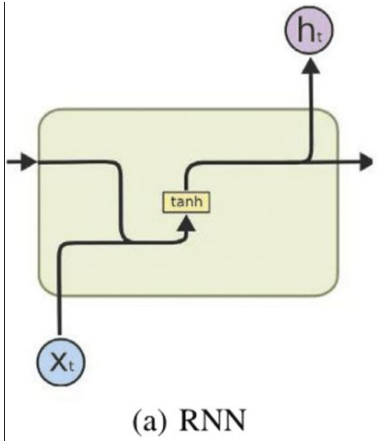
Results using the trigram model

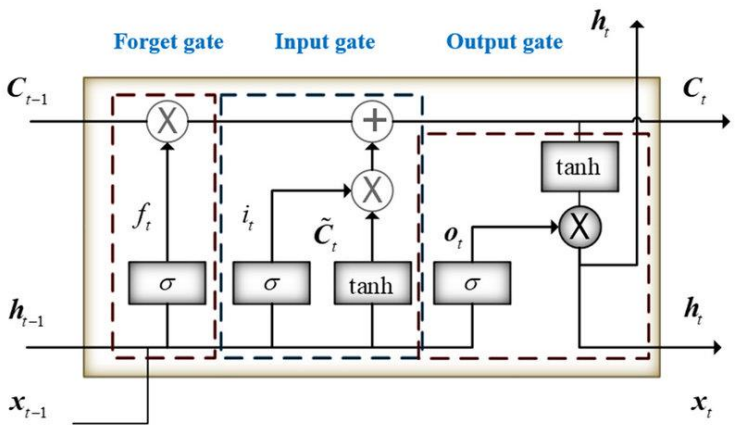
Training data	blogs		news	
Test data	blogs	news	blogs	news
raw	36.60%	25.66%	26.28%	31.63%
laplace smoothing	36.60%	25.66%	26.29%	31.63%
lowercase	43.24%	26.42%	26.58%	31.76%

Table 1: Proportion of saved keystrokes using different train/test data and hyperparameter combinations

Neural networks approach

- Classic feedforward network limited when dealing with varying length of sequences
- RNN has loop in connections and can use information from earlier
- LSTM eventually used





Images from Changhui Jiang, Yuwei Chen, Shuai Chen. **"A Mixed Deep Recurrent Neural Network for MEMS Gyroscope Noise Suppressing"**, and Behrooz Mamandipoor, Mahshid Majd, Mostafa Sheikhalishahi. **"Monitoring and detecting faults in wastewater treatment plants using deep learning"**.

Results using the LSTM model

Training data	blogs		news	
Test data	blogs	news	blogs	news
LSTM-all	51.44%	49.97%	49.49%	54.83%
LSTM-lower	54.13%	51.04%	51.25%	55.74%

Table 2: Proportion of saved keystrokes by LSTM models using different train/test data combinations.

Conclusions

Neural Networks seem better suited for the word prediction task

It is important to use the right data:

- Very diverse dataset for a general purpose predictor
- Tailored dataset for word prediction in a specific context

Improvement perspectives

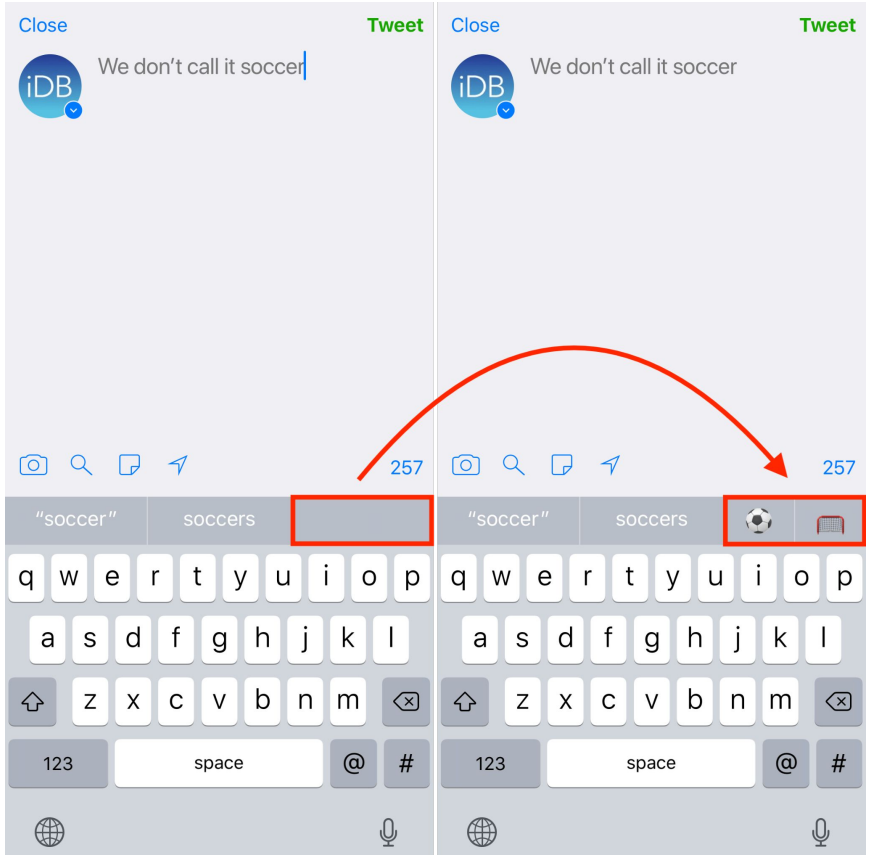


image source