entropy 2021

Entropy analysis of English n-grams

Anastasia Malashina
HSE University, Moscow, Russia
amalashina@hse.ru

Objective

- ★ We estimate the n-gram entropies of English-language texts, using dictionaries and taking into account punctuation, and find a heuristic method for estimating the marginal entropy
- ★ We propose a method for evaluating the coverage of empirically generated dictionaries and an approach to address the disadvantage of low coverage
- ★ We compare the probability of obtaining a meaningful text by directly iterating through all possible n-grams of the alphabet and conclude that this is only possible for very short text segments

Methods

- ☑ Dictionaries of short length texts (n-grams) are empirically generated on a corpus
- ☑ Theoretical coverage of empirical vocabularies, where K is the initial dictionary volume, and k is the number of n-grams that occur once:

$$coverage = (1 - \frac{k}{K}) \cdot 100\,\%$$

- ☑ Theoretical assessment of saturated vocabulary volume:

$$\tilde{K} = \frac{K}{1 - \frac{k}{K}}$$

- ☑ Entropy of n-grams (bits/character):

$$H_n = \frac{\log_2 K}{n}$$

Length of text segment	Initial vocabulary coverage	Theoretical vocabulary volume
10	51,35 %	22 million
15	32,33 %	149 million
20	20,84 %	386 million
25	15,59 %	606 million

Table 1. Coverage and vocabulary resizing

Dataset

- ○ Corpus is based on text samples from the **iWeb corpus of English** language
- ○ Contains about **100 million characters** collected from web pages
- ○ Alphabet of corpus includes only **29 characters**: the letters of Latin alphabet, space, dot and comma

Results

- ✳ Vocabularies of short English n-grams for length of 10, 15, 20, 25 characters (diagram 1)
- ✳ Coverage of empirical dictionaries and theoretical volume of saturated vocabularies (table 1)
- ✳ Extrapolation results of entropy per character based on a linear system (figure 2)
- ✳ **Marginal entropy** of web English is between **0,65 and 0,8** bits per symbol
- ✳ Approximate assessment of number of meaningful n-grams in a language can be found as:

$$\tilde{K}(n) = 2^{H \cdot n}$$

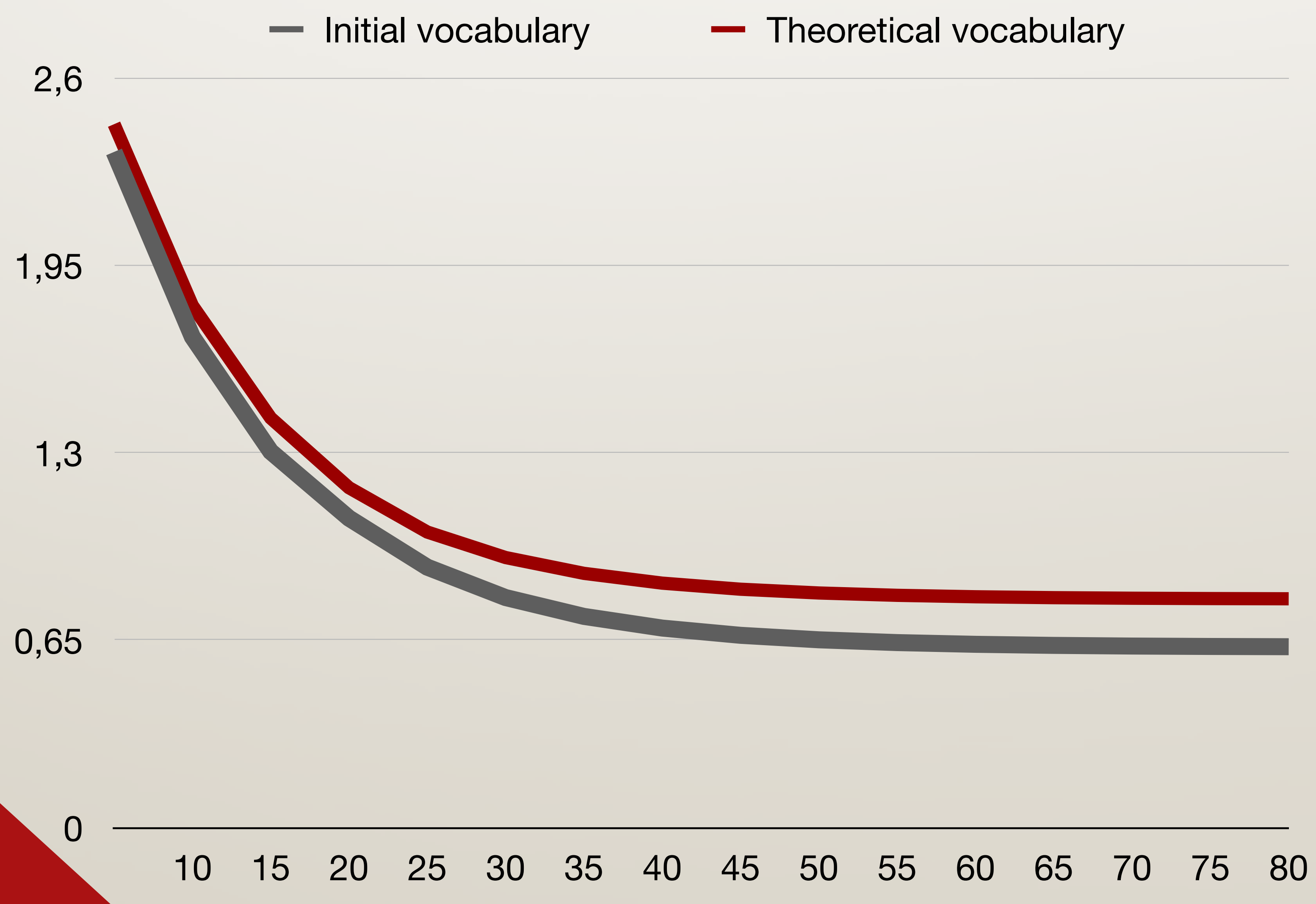

Figure 2. Entropy of n-grams

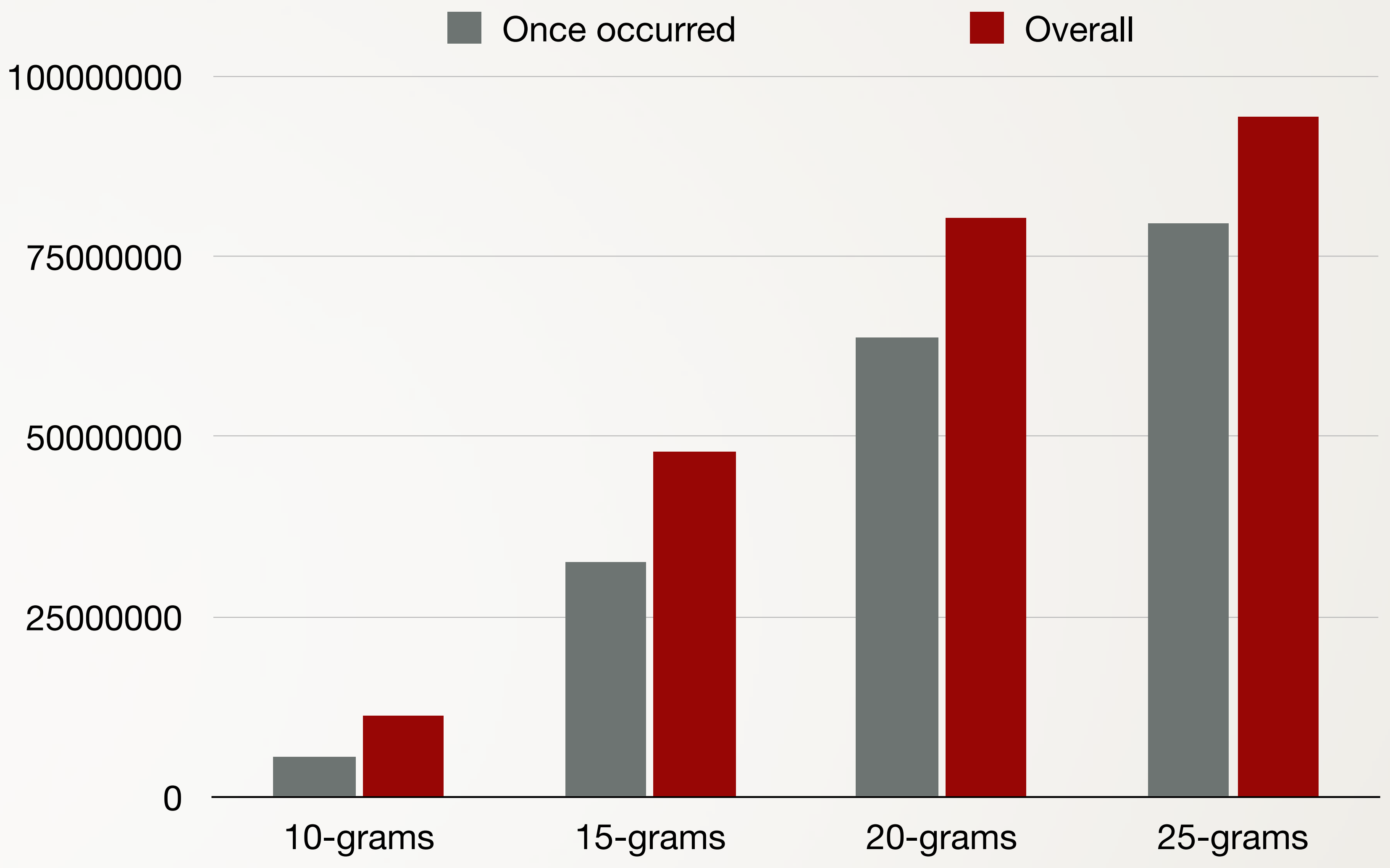

Diagram 1. N-gram vocabularies

Conclusion

We have estimated the n-gram entropies of natural language texts and examined the number of meaningful texts in English. We have found that the empirical method of generating dictionaries can lead to significant type I errors in estimating the number of meaningful n-grams due to low coverage and eliminated this drawback by offering a method for refining the theoretical volume.

By extrapolating the data with a linear recurrent sequence, we have heuristically determined the limiting entropy of our corpus, which is 0.8 bits per character.

Bibliography

- ✓ J.Bellegarda, Robustness in statistical language modeling and speech technology, springer science+ business media dordrecht (2001)
- ✓ L. Chase, R. Rosenfeld and W. Ward, Error-responsive modifications to speech recognizers: Negative n-grams, Third International Conference on Spoken Language Processing, (1994)
- ✓ R. Rosenfeld, Optimizing lexical and n-gram coverage via judicious use of linguistic data, Fourth European Conference on Speech Communication and Technology, (1995)