NLP A1

Tim Mooren

April 2023

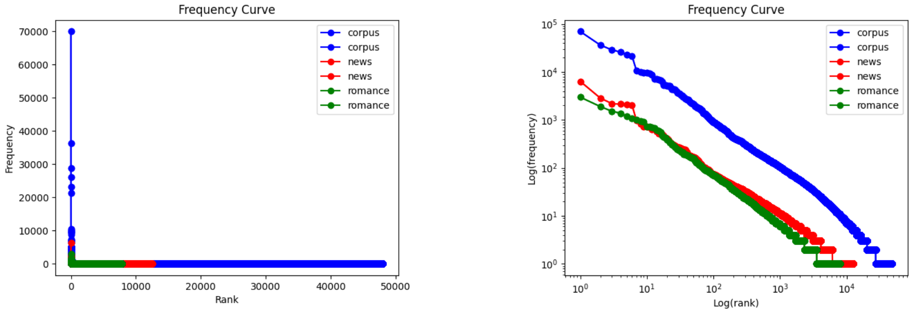
1 Inspecting the data & Zipf's law

Research the origin of the corpus and provide a brief discussion of the findings in light of what you find out. Be sure to highlight any key important linguistic observations that relate to class discussions.

The Brown Corpus, formally known as the ”Brown University Standard Corpus of Present-Day American English,” was created at Brown University in the 1960s. It was the first modern, computer-readable corpus of general English. The corpus was collected from a wide ranges of sources, aiming to represent a wide variety of genres and subject matter. The complete corpus consists of approximately one million words.

The Brown Corpus played a crucial role in the development of computational linguistics and corpus linguistics. By analyzing the corpus, linguists could make empirical observations on lexical, grammatical, and semantic patterns in English.

An example of such a linguistic observation that relates to our classes is the discovery of Zipf's Law, which states that word frequencies in a natural language corpus are inversely proportional to their rank in the frequency distribution. This phenomenon is evident in the Brown Corpus and can be verified through the frequency curve plots included in this document.



(a) Frequency Curve

(b) Frequency Curve Logarithmic

Figure 1: Frequency Curves

2 Unigram model

Estimate (just by eyeballing) the proportion of the word types that occurred only once in this corpus and explain your estimate. Do you think the proportion of words that occur only once would be higher or lower if we used a larger corpus (e.g., all 57000 sentences in Brown)? Use concepts discussed in class with examples for your answer. Eyeballing the printed counts vector, the vast majority of entries appear to have a value of 1, indicating that these word types occurred only once in the 100-sentence corpus. We would approximate the proportion of the word types that occurred only once to be around 90%. If we were to use the complete 57,000 sentences in the Brown Corpus, the proportion of words that occur only once would likely be lower. This is because a larger corpus provides more opportunities for words to occur multiple times. As the corpus size increases, the probability of observing each word type at least once also increases. This could lead to a more accurate representation of the word distribution in the relevant language.

3 Bigram Model

Why is smoothing useful when calculating probabilities related to language? Smoothing is a crucial technique in natural language processing, particularly when dealing with language models. It is used to address the issue of data sparsity and the inevitable fact that many possible word combinations are absent from the training data. Smoothing assigns non-zero probabilities to unseen events by redistributing some probability mass from observed events to unobserved ones. This prevents the model from assigning zero probability to novel sequences, allowing for better generalization and handling of unseen data.

Why did all four probabilities go down in the smoothed model? All four probabilities in the smoothed model decreased because of the before mentioned redistribution of probability mass. In the add-α smoothing method, a constant value α is added to the counts of all bigram occurrences, including unseen ones. This ensures that every possible bigram has a non-zero probability. To maintain a valid probability distribution, the normalization step ensures that the sum of probabilities for each conditioning event is equal to one. As a result, some probability mass is taken away from observed events and distributed among unobserved ones. Consequently, the probabilities of the observed bigrams decrease after smoothing.

Why did add-α smoothing cause probabilities conditioned on 'the' to fall much less than these others? Add-α smoothing causes probabilities conditioned on 'the' to fall less than those conditioned on other words due to the relatively high frequency of 'the' in the training data. 'The' is a common word and serves as a context for many other words, leading to a larger number of bigrams starting with 'the'. When applying add-α smoothing, we introduce a constant value α to the counts of all bigrams, which has a less significant impact on the probability distribution for high-frequency words like 'the' compared to lower-frequency words.

And why is this behavior (causing probabilities conditioned on 'the' to fall less than the others) a good thing? This behavior is beneficial because it maintains the relative importance of high-frequency words while still ensuring non-zero probabilities for unseen bigrams. By decreasing the probabilities conditioned on 'the' to a lesser extent, the smoothed model preserves the natural prominence of 'the' as a frequent context, allowing it to still capture the overall structure and patterns in the language while also assigning non-zero probabilities to novel utterances.

4 Using n-gram models

Compare the performance of the different models. What do you notice? How can we evaluate the performance of each in relation to another?

Perplexity serves as an evaluation metric that quantifies how well a probability model predicts a given data set. Lower perplexity values indicate better model performance as they correspond to higher probabilities assigned to the validation data.

Model	**Perplexity**
Unigram	153.032
Bigram	7.570
Smoothed Bigram	54.236

The table displays the performance of the unigram, bigram, and smoothed bigram model for the second sentence of the toy corpus text. The unsmoothed bigram model was expected to have a better performance than the unigram model, as it takes word pairs into account and could better capture local dependencies between words. This is in accordance with the results, which show that the bigrams perplexity is about 20 times lower. The smoothed bigram performed worse than the unsmoothed bigram, for which the reasons will be discussed in the following sections. Since the same testing set (toy corpus) is used to compare the models with we can equally evaluate the models performance in relation to one another. Yet, for a more accurate comparison it would be advised to use multiple different test sets instead of just two sentences.

Did smoothing help or hurt the model's 'performance' when evaluated on this corpus? Why might that be?

Smoothing is used to adjust the probabilities of rare or missing words in a language model, with the goal of improving its accuracy. Normally, many rare word sequences do not occur in the training data and would be assigned a probability of zero by the model. This can lead to poor performance when testing data contains these previously unobserved word sequences.

Intuitively, you would expect smoothing to improve the perplexity of the language model, as it is measurement of how well the model predicts a sequence of words. A lower perplexity indicates that the language model assigns higher probabilities to the sequence of words.

However, in the present case, smoothing actually decreased the performance of the model. Specifically, the perplexity increased by around 100% for bigram models and almost 700% for trigram models. This is because the sentences used to measure perplexity in this toy corpus were already present in the training data of the larger Brown corpus. Therefore, all the words in these sentences already existed in the model with a certain probability, and smoothing only served to decrease the toy corpus relative probabilities by adding 0.1 to zero occurrences of other word orders.

The negative effect of smoothing was higher for the bigrams and trigram, than for the unigram. This is because the number of possible word combinations increases for higher dimensional n-grams, and therefore more of these combinations are likely to be rare or missing in the training data and be set to 0.1 by smoothing. This causes the relative probability mass of the toy corpus to decrease even further, causing the perplexity to be even higher.

Compare the models qualitatively based on their generation. How does each perform? Cite examples that illustrate linguistic principles discussed in class (ambiguity, compositionality, creativity, etc.) and compare these across models. Attempt to explain the performance of the models given what you know about how they work.

The unigrams sentences are mostly nonsense in terms of nonsense and there appears to be no clear coherence in the combination and order of words. For example the sentence *governments . “ there grady a seek been opposed byrd’s these city berry to on and petition costs bonds property , when of place ”* lacks grammatical structure, meaning and context.

On the other hand, the bigram generated sentences are generally more coherent and grammatically correct. The difference in performance compared to the unigram model is very notable. The bigram was able to produce the sentence *a dispute with city council in fulton county purchasing departments which new management takes charge of the petition as the jury further said*, that is both grammatically correct as coherent in terms of meaning. Yet, the bigram model also often makes mistakes. For example, the sentence *the exception of commerce is one of keeping the house in its 1961 session monday* is grammatically correct, but lacks clear context. It is not clear what ”the exception of commerce” refers to, or how it relates to the rest of the sentence.

In terms of ambiguity and compositionality the unigram model seems to perform badly on all three aspects. Many of the sentences lack compositional structure and not able to combine words or sentence parts in such a way that they accuire meaning. This also causes ambiguity to occur more often, since most ambiguous words lack sensible context that can help clarify their meaning. The bigram model outperforms the unigram model in the sense that it has a wider understanding of how certain words need to be combined. Since it makes use of two dimensional probabilities and thus takes the combination/placement of two words into account instead of each word individually. Since placement and context are major parts in understanding and composing language this provides a great advantage. While the model is certainly not perfect yet, one can image that the performance would only increase with higher order models.

In terms of creativity, both models are still very limited in the sense that they base their produce their sentences based on statistical probabilities. This does not allow for new rare words and word combinations to occur much, and even makes it impossible for new words to ever be created. Expanding the models would be necessary to enhance their creativity.

5 Bonus

Discuss the validity of the independence assumption for unigram models. The independence assumption for unigram models states that the probability of each word occurring in a given sequence is independent of the words that precede or follow it. However, this simplifying assumption does not necessarily reflect the true relationships between words in natural language.

In this analysis, we calculated the Pointwise Mutual Information (PMI) for all successive word pairs (w1, w2) in the Brown corpus. PMI is a measure of the statistical dependence between two events, in this case, the occurrence of consecutive words in a text. A PMI value close to zero indicates that the presence of one word has little or no impact on the presence of the other word, while a positive or negative PMI value suggests a significant relationship between the words.

Our results, which can be found in the tables below, indicate that certain word pairs exhibit very high PMI values, which implies a strong association between the words. These pairs are often collocations or multi-word expressions (e.g., "hong kong", "viet nam", "puerto rico"). On the other hand, some word pairs display very low PMI values, indicating that the occurrence of one word has little influence on the probability of the other word. The lowest PMI values are primarily observed for frequent words like articles, punctuation, prepositions, and conjunctions (e.g., 'the', '.', 'and', 'of').

Table 1: Word Pairings

(a) Lowest PMI Scores

Word 1	Word 2	PMI
.	,	-11.2755
the	.	-10.5379
the	a	-10.4488
and	.	-10.2599
of	of	-10.1571
and	and	-9.4857
the	in	-9.3284
the	is	-9.2506
the	,	-9.1944
the	and	-9.1788
a	in	-8.7354
his	the	-8.7196
of	to	-8.6799
the	i	-8.2816
of	he	-8.2259
he	of	-8.2259
of	for	-8.217
,	;	-8.1273
the	not	-8.1178
the	have	-7.892

(b) Highest PMI Scores

Word 1	Word 2	PMI
hong	kong	16.6877
viet	nam	16.1472
pathet	lao	16.0597
simms	purdew	16.0597
herald	tribune	15.7947
el	paso	15.6877
lo	shu	15.6877
wtv	antigen	15.6691
puerto	rico	15.5622
drainage	ditch	15.4467
anionic	binding	15.3107
unwed	mothers	15.2403
kohnstamm	reactivity	15.1988
carbon	tetrachloride	15.1623
phonologic	subsystems	15.0883
willie	mays	14.9873
peaceful	coexistence	14.9873
computing	allotments	14.9508
presiding	elder	14.9184
nineteenth-century	immigration	14.8252

6 Contributions

It was decided that everybody would complete the assignments independently to maximize learning. Afterwards, we would review each other's work and identify any errors. Tim completed all the assignments, including the bonus task, while Nikki completed assignments I, II, III and IV.5 Agnes worked on assignment I, III.4, and completed assignment IV. Regarding the report, Tim had already written the majority of it, and Agnes completed the remaining sections.