Language modeling and N-grams applied in the reconstruction of history

Eva Pérez S.

ABSTRACT

Our objective from this research paper is to get a more deep understanding about the accuracy while completing texts from our ancestors that have been lost in time.
Another subject to cover that have a lot of importance, is if other human languages (Chinese, French, Spanish, etc.) would have the same accuracy completing these texts as well as in English.
This research gathers information from diverse sources such as PDF, articles and a very rich information from a book from the MIT, all these from reliable sources that confirm each others perspective on the subject.
Our findings have been that simply there's not enough data collection as well as in English, some languages have more complex grammatical structures that can make accurate modeling more challenging.
Reconstructing lost texts is often a collaborative effort that draws on the expertise of individuals from various disciplines, having an AI that could do it automatically wouldn't be accurate, more than using it as a substitute of the professionals of this field, it can be used more to backup the recent findings of certain historical piece.
The implications of the fundamentals of NLP and ML have a field of improvement in the AI area, with enough research of how to make easier for a computer to follow certain human languages or to have a new language applied to their configuration, our options are far from just a few.

Keywords: Natural Language Processing (NLP), Machine Learning (ML), N-grams, Language Modeling

INTRODUCTION

In the field of Artificial Intelligence, there is a lot of enigmas to consider when we have all this freedom to do what we want with a machine that can learn for itself. Entering in the subject of history, one of the questions that have been going through a lot is if it's possible to complete texts lost in time with just a partial example of it, could a language model be that powerful to complete those files and be compared against professionals and scholars who work in fields such as history, archaeology, linguistics, paleography, and digital humanities? Even if it's the case, realistically speaking, not all the texts can be in the English language, there could be missing pieces from other languages like Latin, German, Chinese, etc. Just to mention a few. How could it work in those cases? The language model formula that we know today could work regardless of any human language? Is it possible to have the same accuracy?

As we know, history has been of great importance to help us make sense of the world we live in, our heritage, culture, evolution,etc. And this is a vital discipline that improves our understanding of the human experience and the world, there could be so many things we don't know or are limited because of documents lost in time, if we can make a difference to get to know more about our roots using high technology it would meant we are in the right path and we can achieve so much more with it.

LANGUAGE MODELS AND N-GRAMS

Language Models are basically a concept in NLP and ML where it's possible to predict certain phrases or next word in a sequence based on a data set where words have already been uploaded. Uses a probability of word occurrence and its based in a gigantic database (preferable a good language model). The data that enters the model have a lot of importance, it determines the quality of the sequence and how will it collect their data based on it, don't expect a high quality model with poor collected data texts.

N-grams are part of the language models, they facilitate how to extract the information, used to represent text or speech. For example: ”I love eating watermelon” can be divided into 2-grams it would be like this: ”I love”, ”love eating”, ”eating watermelon”. All of these conjugation of words mean different

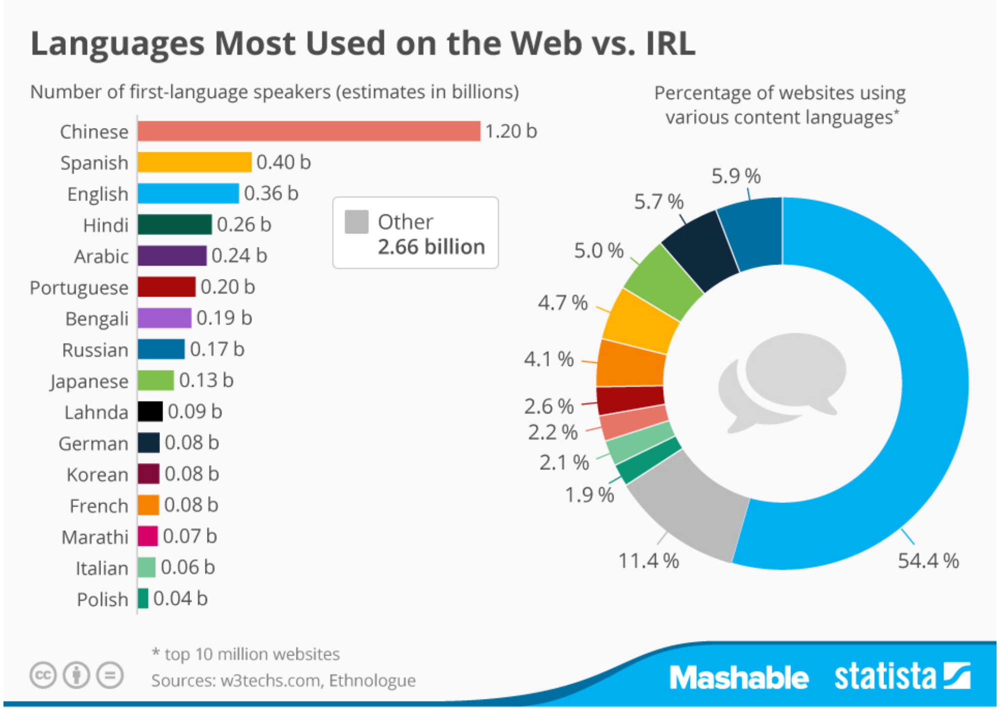

Figure 1. Languages Most Used On the Web vs. IRL

things, but can be implied that someone loves, eats something, or someone loves eating something and can be used in sentiment analysis, this way, the model can associate this 2-gram with the objective of some correlation between the words put together and what they meant to express. Brill (NA)

ACCURACY IN VARIOUS LANGUAGES

It doesn't all depend on the human languages used, main part for an algorithm to learn a pattern correctly is from the data collected. "It also depends on the type of task your algorithm is supposed to fulfill and the method you use to achieve AI."(MARUSARZ, 2022)

"In general, traditional machine learning algorithms will not need as much data as deep learning models. 1000 samples per category are considered a minimum for simplest machine learning algorithms, but it won't be enough to solve the problem in most cases." (MARUSARZ, 2022) The main problem with looking for accuracy in a certain language will always depend on the amount of data the internet have collected, meanwhile doing something like recollecting data for a language would take years to improve, the vast amount of data in English is what we can use more frequently with the knowledge that it's quality data and not just random words on a paper. Shareghi (NA)

We can have a better visualisation of how the human languages in the world work with a chart. Figure1

Here we can see more clearly how the English language is more biased to communicate with other people worldwide and how little difference does it make for Chinese speakers to be at the top of the most spoken languages in the world.

Data Recollection

Without going too deep into this subject, for a model to be good it requires at least the sufficient amount to divide it into training and testing purposes. To optimize this resources, the usual is to split it in the proportion of 80 % for **training data** and 20 % for **testing data**.(MARUSARZ, 2022) This with the objective to verify the accuracy of the previous trained model. Now, as the question as how can this be applied in a human language depends on the population that uses it daily (preferable native), have a good grammar, has a good lexicon, etc. This to be used as the main collection data and quality samples for our model to be trained in the best way possible, resembling the language as accurate as it can.

Bayes' Rule

If we would want to see how translation works for this cases without having to get a new database for each language, Bayes' Rule is our solution. This rule help us make the translation more optimal for getting the data from the original text. This is called the **noisy channel model** for a reason, it envisions English text turning into another language by passing through a noisy channel, p$_s|_e$. Why not modeling language directly instead of using a translation model to turn it into English? One way to get into this issue is by knowing that these two models are different from each other, p$_s|_e$ the translation model and p$_e$ the language model have to be estimated from different data. The **translation model** needs examples from grammatically correct translations that make sense, meanwhile the **language model** only needs text in English to function properly.(Einstein, 2019) Such monolingual data is more worldwide available, as we saw in Figure1.

Bayes' rule is given by:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{1}$$

Cross-Linguistic Data Formats

The volume of digital data available for languages worldwide is steadily increasing. However, much of this data is presented in diverse formats, making it challenging to compare and utilize effectively. To address this issue, the Cross-Linguistic Data Formats initiative has introduced new standards for two fundamental types of data used in historical and typo-logical language comparison: word lists and structural data-sets. Additionally, they've developed a framework to incorporate other data types like parallel texts and dictionaries. This new specification for cross-linguistic data formats is accompanied by software tools for validation and manipulation, a basic framework that connects to broader standards, and practical examples of recommended practices.International Union for Conservation of Nature (2017)

RESULTS

Simply there's not enough data collection as well as in English, some languages have more complex grammatical structures that can make accurate modeling more challenging. In some cases translation wouldn't be possible or simply not the same as the original. Reconstructing lost texts is often a collaborative effort that draws on the expertise of individuals from various disciplines, each contributing their unique skills and knowledge to piece together the puzzle of the past. Having an AI that could do it automatically wouldn't be accurate, more than using it as a substitute of the professionals of this field, it can be used more to backup the recent findings of certain historical piece. As a reaffirmation of the previous sentence, it is also needed a similar environment from the specific era in research, such as culture, mannerisms, grammar, social context, etc. All of them very unlikely to find in sources as the internet if technology wasn't as advanced as today.

CONCLUSION

While it is possible to achieve high accuracy in various languages with language models, the availability of training data and the complexity of the language can impact the level of accuracy that can be achieved. English models tend to have a natural advantage due to the abundance of English text data. More data is needed in other languages to have more flexibility in the resources and not be so English-biased. This same data is necessary to be confirmed as good quality to be used in future AI models.Fonseca (2022)

Accuracy in a language model depends on the database, so if any type of historical text is wanted to be completed, there should be a vast amount of texts or resources from that specific era, it can be possible using similar patterns as Bayes' rule for a translation, but context is needed to have a more clear understanding of the historical piece in mind.

There are a lot of implications in this field, but more importantly there's a lot we can work on in this topic, we now know it's possible to have the same accuracy in other languages as well as in English with the appropriate data, the hard part is collecting it.

REFERENCES

Brill, E. (NA). Beyond n-grams: Can linguistic sophistication improve language modeling?

Einstein, J. (2019). *Introduction to Natural Language Processing*. The Massachusets Institute of Technology.

Fonseca, J. (2022). Building wikipedia n-grams with apache spark.

International Union for Conservation of Nature (2017). Deforestation and forest degradation (issues brief). https://www.nature.com/articles/sdata2018205.

MARUSARZ, W. (2022). How much data does ai need? what to do when you have limited datasets? [Online; accessed 8-September-2023].

Shareghi, E. (NA). Show some love to your n-grams: A bit of progress and stronger n-gram language modeling baselines.