Bryan Holcomb
CS 410: Text Information Systems
15 November 2020

Tech Review: Statistical Language Model Use Cases in Speech Recognition

Introduction

One of the more widespread technological advances of the last few years has been the advent of voice recognition software. From digital assistants such as Amazon Alexa and iPhone's Siri to the complete automation of customer support phone systems, almost every American utilizes state-of-the-art speech recognition on a routine basis. This speech recognition is driven by statistical language models, which are continuing to evolve and thus become more accurate in comprehending human speech (Raju, 2019).

Statistical Language Models – An Overview

A statistical language model (LM) is simply a probability distribution over a sequence of words (Zhai, 2020). Essentially, the LM assumes that text is created by randomly choosing a word out of a distribution based on that probability. However, the probability of a given word being chosen is context-dependent. For instance, an article about baseball is far more likely to contain the words "bat" or "mound" than the word "keyboard". Using this context, the model adjusts the expected probabilities of each new word as it continues to generate text.

One example that readers might be familiar with is the "suggested word" feature in modern smartphone messaging (Dehdari, 2015). When the user begins typing a text message, the phone "suggests" words that the user may want to type next, based on what was previously typed in the conversation. For instance, if somebody texts the user "Are you free tonight to chat?" and the user begins to reply "Sure, give" – the phone may suggest the words "me", "five", and "minutes" as the user continues to write. The phone may even adjust its suggestions based on the user, weighting phrases or words the user types more often with a higher probability.

According to Amazon's Anirudh Raju, most "conventional language models are *n*-gram based, meaning that they model the probability of the next word given the past *n*-1 words", where *n* is typically around four (Raju, 2019). This enables the model to select words that are related to other words in the phrase.

LM Use in Speech Recognition

The first step in speech recognition is the sensing of speech, which is registered by the computer as sound waves. The computer then compares these sound waves to that of known *phonemes*, or phonetic sound. Once it compiles a sequence of phonemes, the computer compares this sequence to known words that contain those phonemes (Grabianowski, 2006). This is where the LM comes into play.

From here, the LM assigns a probability estimate over a distribution of words. This is repeated for every subsequent phoneme-word, and the probabilities are adjusted based on the distributions before and after them. From here, the computer can narrow down the likelihood of a specific word out of that distribution.

Advantages of the *n*-Gram Model

The *n*-gram model, while simple relative to other models, tends to be very useful on the whole for a number of reasons. As Dehdari notes, *n*-gram models train very quickly on new data and require little to no manual annotation (Dehdari, 2015). MIT's Glass and Zue also note that the probabilities are based on lots of data and "incorporate local syntax" – specifically, that "many languages have a strong tendency toward a standard word order" (Glass & Zue, 2003). For this reason, *n*-gram models – especially trigrams ($n = 3$) – have been used since the mid-1970s and remain popular to this day (Glass & Zue, 2003).

Disadvantages of the *n*-Gram Model and Alternative Approaches

On the other hand, modern advances are starting to move away from the *n*-gram model and incorporate "deep learning" to overcome some drawbacks of the *n*-gram approach. First and foremost, the *n*-gram model depends on a short-range context, specifically trigrams or four-word sets as mentioned previously by Glass & Zue (2003) and Raju (2019). Unfortunately, most sentences – especially in English – are longer than four words. Thus, the *n*-gram is not able to use the context of two related concepts that may be mentioned on either side of a longer sentence. Additionally, Glass & Zue (2003) note prior literature that *n*-gram models may have sparse data – for instance, while the English language contains millions of words, many of those words are very rarely used. This limits the ability of the models to recognize the relation within context, especially within less-common subject matters.

Recurrent neural networks, however, address these issues. Although they are far more difficult to implement and take far longer to train, they are able to "learn such long-range dependencies, and they represent words as points in a continuous space, which makes it easier to factor in similarities between words" (Raju, 2019). In fact, Raju and colleagues found that their neural model reduced the word recognition error rate by 6.2% over the conventional *n*-gram model, a number that will surely grow as neural network models continue to evolve.

Conclusion

As our daily lives become more automated, speech recognition becomes a more integral and useful part of our daily lives. While *n*-gram statistical language models continue to be effective in powering the automatic speech recognition, the rapid advance of deep learning – specifically with regard to neural networks – will serve to both enhance existing speech recognition systems as well as open up new possibilities.

References

Dehdari, Jon. “A Short Overview of Statistical Language Models.” The Workshop on Data Mining. jon.dehdari.org/tutorials/lm_overview.pdf.

Glass, James, and Victor Zue. “Language Modelling for Speech Recognition.” Automatic Speech Recognition. ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-345-automatic-speech-recognition-spring-2003/lecture-notes/lecture1112.pdf.

Grabianowski, Ed. “How Speech Recognition Works.” *HowStuffWorks*, 10 Nov. 2006, electronics.howstuffworks.com/gadgets/high-tech-gadgets/speech-recognition1.htm.

Raju, Anirudh. *How to Make Neural Language Models Practical for Speech Recognition*. Amazon Science, 6 Dec. 2019, www.amazon.science/blog/how-to-make-neural-language-models-practical-for-speech-recognition.

Zhai, ChengXiang. “Probabilistic Retrieval Model: Statistical Language Model.” Text Retrieval and Search Engines. d18ky98rnyall9.cloudfront.net/_2e926c62f63841ddb2902e65241cee69_4.2-Statistical-Language-Models.pdf?Expires=1605657600&Signature=cahugrNtJAS~1Bjvzp3-toeQMSGvpesHkKKGMghLaA3Dspm6HZEJoRjDrClsgUHsX1wNYs0tmGVkLIbnqobGlgFr~LCeSP3EYDc1z30Z1lDH1e4J9yo0B1AVUVihkRaLzY4perTRtu3zkq4MYko50Jgwqx6SybTSXC9uN~MwesY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A.