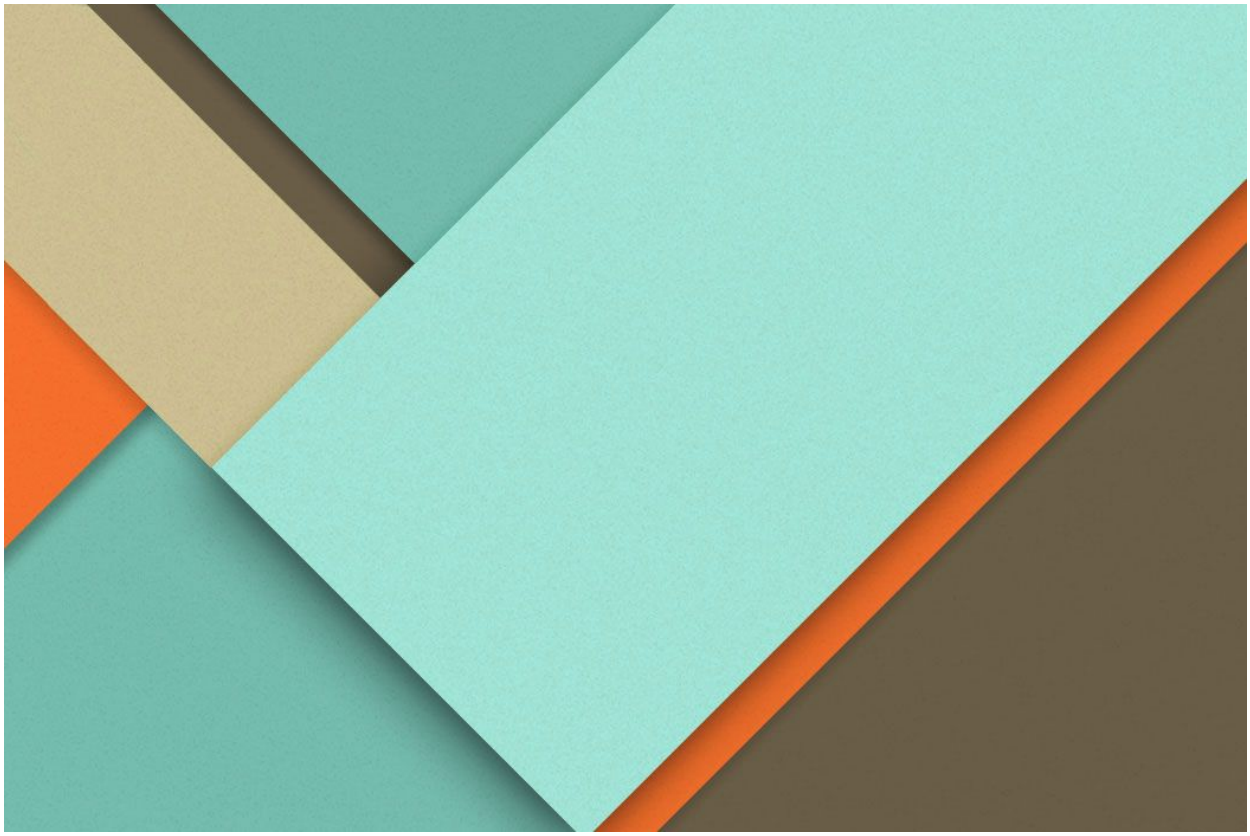

Project Proposal

09.29.2017

Li Breite, Kevin Allen, Mitko Nikolov
Northeastern University
Boston, MA

Introduction

We would like to create an Android application that can help improve life for people who are blind or visually impaired. The product would be able to take a picture and then answer questions such as “What is in this picture?” “Is there a cat in this picture?” “What color is this shirt?” or “What does this text/sign say?” The software can prove to be extremely beneficial to people who cannot see as they are trying to navigate their day-to-day lives in an unfamiliar setting. For example, the act of choosing matching clothes might appear to be a rather mundane task for a seeing person, but could be a real challenge for someone who cannot see. With this app, they can just take a picture of the shirt they are planning to put on, and ask “What color is this shirt?” The app would then be able to answer something along the lines of “This shirt is red,” “That is not a shirt, it is a dress,” or “I do not see any articles of clothing in this picture.” It is important that this problem is studied and worked on because visually impaired people are a significant and important part of the human race. Just as any other person, they deserve opportunities and support and creating something that can assist them in the described way would be a step on this path. We hope that the described software will be able to go a long way in assisting people who need a virtual pair of eyes in navigating the world.

Related work

There are of course many applications that can respond to speech, such as Siri, or Amazon Echo. Additionally, there is technology that can look at a picture and return what is happening in it. However, the combination of the two technologies is, as far as we know, rather novel. Our application would need to be able to hear human speech, parse the sound into meaning, look in the picture for the answer, and respond (with very human-like tonality and cadence) with the correct answer.

In terms of competition, there is a company called pivothead that more than a year ago expressed their interest in creating sunglasses with an embedded camera that could do something similar. However, they designed their product as a tool for people to share where they are and generally read signs around them. We are targeting specifically blind people and are focusing on serving as many of their needs as possible. Alongside with this, to use our product, people would only need their phone.

We do not know of any other applications or software that can do all of these things. While we will rely heavily on APIs that can analyze photography, we will be implementing the core natural language processing components ourselves.

Datasets

We intend to use the WikiLinks Corpus by Google in order to create a Language Model that will allow us to process human speech and turn it into different queries so that our users can be assisted. The corpus consists of web pages that contain at least one hyperlink that points to English Wikipedia. For every mention of a string, the corpus provides the actual mention string, the byte offset of the mention from the start of the page and the target url all separated by a tab. The corpus contains approximately 59 million mentions across 13 million Wikipedia articles. Providing a vast amount of data that would give us millions of sentences to work with, the corpus would enable us to create a model that can successfully identify the meaning behind a modern day expression.

We might also use WordNet in order to tag words as different parts of speech. WordNet contains 206,941 word-sense pairs. The total number of unique strings, regardless of the part of speech the strings belongs to, is 147,278 and taking into account the part of speech classifications of each unique string, leading to multiple counting of the same string, raises this number to 155,287. It identifies relations between different words and then identifies interrelations between the different sets. WordNet is suitable for the task because it is currently known as one of the best developed data sources as far as POS tagging goes and it is available to use for free.

Evaluation

We intend to use both extrinsic and intrinsic evaluation. Our initial method would be intrinsic: we would calculate perplexity by evaluating to what extent our model is capable of generating the test corpus with various N-grams.

On the next stage we will start performing extrinsic evaluation by checking whether the application performs correct API queries based on the speech input we have provided. As part of our development cycle, we are going to use a development corpus in order to fine tune our language model and continuously refine it.