Word Variation Across Occupations

Nicole Mehring
913289895
Year 4
Cognitive Science Major
nsmehring@ucdavis.edu

Shreya Venkatesan
913582534
Year 3
Cognitive Science Major
svenkatesan@ucdavis.edu

Emily Strand
912832609
Year 4
Psychology & Linguistics Major
emmstrand@ucdavis.edu

Abstract

The purpose of this project centered around the hypothesis that word choice differs between occupations. To test this, a corpus was created, consisting of spoken interviews categorized into 8 occupations, ranging from actors to politicians. A text classifier was built to see if spoken words could be classified into the 8 occupations using noteworthy words. Based on the corpus curated, frequency distributions utilized by the classifier, and the word-level analysis conducted, the major takeaway was that word choice was not a distinguishable characteristic across occupations.

1 Project & Report Breakdown

1.1 Project Breakdown

Nicole was responsible for the data collection aspects of this project, searching for and acquiring the interview transcripts across different occupations, transcribing interviews without transcripts, formatting them into text files, annotating them, and sorting them accordingly. Nicole also separated the corpus into the training set, development set, and testing set used in the text classifier.

Shreya was responsible for the technical aspects of this project, specifically building the frequency distributions and coding the text classifier in Python. In addition to building the text classifier, Shreya manually annotated the predictions generated by the classifier to compute its accuracy and reported observations regarding the classifier's performance.

Emily was responsible for the word-level analysis aspects of this project, specifically the POS tagging, determining the proportion of open to closed class words, and word saliency analysis.

1.2 Report Breakdown

The following list breaks down the contributions Nicole, Shreya, and Emily made to the report.

- Nicole: Project & Report Breakdown, Introduction, Corpus Curation
- Shreya: Project & Report Breakdown, Text Classification, References, Appendix
- Emily: Project & Report Breakdown, Abstract, Word-Level Analysis

2 Introduction

Realistically, occupations demand different skill sets and vary drastically in terms of the types of tasks that are completed. For instance, whereas jobs rooted in the arts (i.e. acting) spawn greatly from creativity, jobs that lean more towards the corporate side of things (i.e. business leader) tend to revolve more around practicality and systematicity. Given this initial stance, we might question whether the language used to articulate such occupations might differ through how speakers communicate in accordance to their job. Thus, we decided to test if substantial word variance exists when examining individuals across various occupations and, if so, which words are used more often in those respective occupations.

3 Corpus Curation

For our corpus, we gathered interviews from a variety of sources and them into 8 different occupation categories: actor, author, athlete, musician, director, business leader, politician, and talk show host. We reached just over 600 text files for our final dataset, which was broken down into a training set, development set, and testing set. Roughly half of the corpus was used in the training set, with a

small portion of the training set used for the development set, and the remaining files in the corpus made up the test set.

3.1 The Process

We decided on the above occupation categories for our analysis as a result of those occupations being very public and, consequently, generally having more interviews conducted. The corpus curation process was a long one, as we needed to collect as much corpora as possible in order to give our training and testing sets enough data to work with, and to ensure our classifier could properly do its job with as much accuracy as possible. To accomplish this, transcripts were collected from various interviews from press conferences, radio broadcasts, news broadcasts, and YouTube videos. Making sure the interviews were candid, not scripted, was important, due to the fact that our language question concerns word usage by individuals of a certain occupation, and interview outlines and scripts are often written by other parties or do not capture the language used by individuals in the moment.

Interview transcripts were difficult to find as existing text files, so the following was done in order to get the transcripts we needed in a format easily analyzable by both the human eye and our classifier: YouTube videos were scraped of their captions and outfitted into text files after cleaning up extraneous chatter from people outside of the person of interest, interviews were copied from "The Talks" and NPR's directories and pasted into blank text files, and any interviews that were not possible to get transcripts from were transcribed manually. After our interview transcripts were completed, they were annotated with their occupation label and put into each occupation's respective folder, then sorted into the training, testing, development sets.

3.2 Challenges

A particular challenge in curating this corpus was finding interview material that had all of the following qualities for all of the occupation categories: candidness, limited interruptions from the interviewer or audience, understandability, and enough length to offer sufficient data. Some interviews were much easier to find than others, examples being people in the actor and musician categories. On the other hand, finding enough data for the politician and business leader categories was difficult and, in the end, multiple interviews from singular individuals from each respective occupation were required in order to meet all of the above requirements. Even then, many of the interviews varied in length, resulting in some categories having much more data to work with than others.

What this meant for our classifier, and its subsequent accuracy, is that when training, there may not have been enough data to properly give it a sense of the words used by certain occupations. On the other hand, when looking at our results, it is plausible that there just aren't many salient words spoken in interviews. After all, the majority of the most salient words across occupations were occupation neutral, such as "people", "time", "work", and "years".

4 Text Classification

With our language question centered around word usage in various occupations, we wanted to see if we could build a text classifier that would be able to classify interviews into occupations based on salient words found in each occupation. Keeping this question in mind, we built a text classifier similar to the one built in our third homework assignment and created a training/development set and a test set. The distinction between the classifier used in our homework assignment and the occupation classifier was the number of categories/classes. For the occupation classifier, there were 8 categories for the text to be classified as, hence when computing the argmax for the conditional probabilities, each if condition had to have 7 comparisons. Once the classifier was trained, it was run using the test set and manual annotation to check accuracy was done.

4.1 Observations

One observation from the predictions our classifier made was that "business leader" was never classified by the classifier, despite there being "business leader" files in the training data and the test data. This could have occurred for 2 reasons: (1) there were not enough "business leader" files in our corpus, and (2) nothing is really salient or distinct in the "business leader" interviews.

Another interesting observation from how our classifier performed was that when given a few interviews from the same individual, the classifier classified each of those interviews differently. This occurred for some of the classifications, while others were correctly identified. For example, there were three "David Lynch" files, and the correct classification for all three would be "director." However, the classifier classified the three as "actor", "musician", and "director." So, one of the three was correctly classified, while the other two were incorrect. In the other scenario, though, there were three "Harrison Ford" files, and the classifier correctly identified all three of those to be "actor."

4.2 Accuracy & Pitfalls

The total number of predictions, also known as the number of files in the test set, was 302. Of the 302, 113 of the predictions were incorrect. In other words, 189 of the predictions were correctly classified. Thus, the accuracy of our classifier was $189/302 = 62.58278\% \approx 63\%$, which is quite low. There are a few reasons that we believe could account for this low accuracy. One of these reasons, which will be discussed in the section 5, is the lack of salient words. Contrary to our hypothesis, maybe there aren't distinct enough words in the interviews after all. Though we did find some unique words in each occupation category, it might not have been enough for the classifier to distinguish each occupation from one another.

The other possible reason for the low accuracy was not having enough data in our corpus. The total corpus was around 600 files. The 600 files was split among 8 categories; however, the split was uneven, with there being a lot more files in the actor and musician categories and less files in the business leader and politician categories. Additionally, the 600 files was also split into 2 major data sets: training and test. Because the corpus was quite small, roughly half the corpus was in the training set, with a small portion of this being in the development set, and the other half was in the test set. A mere 300 files is not enough to train the text classifier, especially when there are 8 categories the interviews can be classified into. Adding onto the lack of data in the corpus was the fact that the size of the interviews was also inconsistent. The politician interviews tended to be quite long, much longer than the interviews for the other occupations. This could have played a part in the training of our classifier and ultimately affected the accuracy of classification.

4.3 Next Steps

To improve our classifier, there are a few next steps that can be taken. These include removing stop words, lowercasing tokens, increasing our corpus size, improving proportionality of files in each category, and ensuring the content in each file is as even as possible. To improve the overall occupation classification, the next step would be to increase the number of categories. Currently, there are 8, but this is not representative of all occupations. Hence, increasing the number of categories will encapsulate a wide range of occupations.

5 Word-Level Analysis

Given our interest in determining whether language differs in terms of our eight specified occupations, we focused primarily on the specific words/types used. We decided to this generally and specifically: (1) by comparing the proportion of open to closed class words, and (2) by comparing the salient words.

5.1 POS Tagging and Word Classification

Initially, we took the 200 most common words/types for each occupation and manually tagged their corresponding parts of speech. Although there is a high likelihood of human error present in the annotations, we tried to control for this by opting for rather general categories (i.e. noun, conjunction, adjective). Also, to account for word with multiple part-of-speech tags, we tagged all instances of such words with their most prominent POS tag. We grouped the appropriate POS tags into their respective class (i.e. open, closed). Following annotation, we compared the proportion of open to closed class words for each occupation (see Appendix A, Tables 1 - 2).

Ultimately, with the exception of the athlete category, it was apparent that closed class words occur more frequently in terms of occupation. This was to be expected in the sense that language needs to be articulated grammatically, which requires more function words. That being said, they

don’t actually differ that much from the frequency of open class words. This leads to the interpretation that word class proportion is roughly equal among the eight chosen occupations and possibly to all occupations.

5.2 Word Saliency

To analyze the specific word choice a little more in-depth, we identified and compared the most salient words/types for each occupation (see Appendix A, Table 3). It’s important to note that many of the salient words were not the most frequent. In fact, the salient words for each category were the 31+ most frequent, following predominantly closed class words, such as punctuation, determiners, and pronouns. That being said, the salient words, when compared across occupations, weren’t as distinct as previously speculated. For example, although “novel” appeared only for the author category and is one of the few words unique to this category, most words overlapped with multiple categories. For instance, the word “movie” appeared in both the actor and director categories. This outcome is indicative of the performance of the text classifier in the sense that directors were frequently misclassified as actors. Given that this occurred with several of the other categories, we inferred that, at least based on the analysis conducted on spoken interviews across a range of occupations, word choice is not a distinguishable characteristic.

References

“ASAP Sport.” *ASAP Sport*, 2019, www.asapsports.com.
“Interview Directory.” *The Talks*, 2019, www.the-talks.com/interviews/.
“National Public Radio.” *NPR*, 2019, www.npr.org/.

Appendix A. Word Analysis Tables

	POS Tag	Description	Example
Open Class	V	Verb	have, started, need
	N	Noun	film, book, show
	Adv	Adverb	always, never
	Adj	Adjective	big, great
	Int	Interjection	yeah, No, Well
Closed Class	Sym	Symbol (either punctuation, unintelligible "words", or unseparated contractions)	?, ca, 's, don’t
	Det	Determiner	the, a, one
	Conj	Conjunction	and, but
	Prep	Preposition	in, at, from
	Part	Particle (strictly grammatical)	as, about
	Ex	Existential "there"	there
	Modal	Modal Auxiliary Verbs	will, should, must
	Pro	Pronoun	I, his, myself

Table 1: Annotation Key

	Actor	Athlete	Author	Business Leader	Director	Musician	Politician	Talk Show Host
Open (Percentage)	46%	50%	45.5%	44.5%	42%	48.5%	47%	43.5%
Closed (Percentage)	54%	50%	54.5%	55.5%	58%	51.5%	53%	56.5%

Table 2: Open vs. Closed Class Percentages

Actor	Athlete	Author	Business Leader	Director	Musician	Politician	Talk-Show Host
people	time	book	people	people	people	people	people
time	play	writing	make	film	music	country	think
movie	people	people	years	movie	time	president	time
work	game	write	company	movies	song	America	years
good	years	books	world	time	love	years	show
years	guys	story	work	films	songs	work	money
film	year	years	money	work	feel	Trump	company
character	hard	writer	business	years	life	tax	school
show	playing	read	life	story	years	system	dollars
school	team	novel	year	world	work	believe	love
films	players	characters	country	trying	album	world	
play	work	idea	hundred	years	write	American	
	happy	stories	point	making	play	House	
	played	point	Apple	first	year	United	
	trying	wrote	working	great	world	campaign	
		written	book	try		problem	
				started		States	
				big		decisions	

Table 3: Salient Words Across Occupations