'Speech' Recognition: Using NLP to Attribute Quotes to Speakers

LING 227 Final Project

Luke Benz
Will Langhorne
Kevin Truong

December 15, 2017

Introduction

History remembers great speakers for their idiosyncracies. From JFK's authenticity and passion to Donald Trump's divisive rhetoric, U.S. presidents are particularly identifiable by their speeches. While nearly any citizen can name the presidents who said "Nothing to fear but fear itself" (Franklin Delano Roosevelt), "Ask not what your country can do for you–ask what you can do for your country" (John F. Kennedy), and "Yes We Can!" (Barack Obama), it remains to be seen whether a language model can do the same. Our project aims to build a language model that determines the most likely speaker of a given input quote. Morover, our project seeks to investigate how the model discernability changes when trained on three types of data: presidential speeches, part of speech tags for presidental speechs, and politicans' tweets.

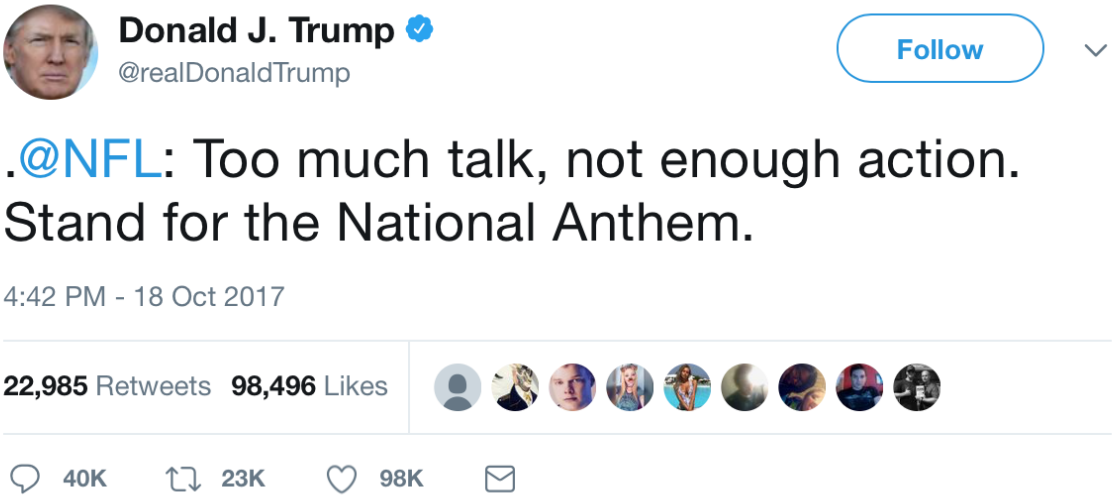
Methodogy

Data Sources

Speeches from each of the 44 presidents and 2016 Democratic Nominee Hillary Clinton were obtained from on open source online corpus compiled by The Grammar Lab. Tweets from Presidents Donald Trump and Barack Obama and various Senators were collected from data made public by FiveThirtyEight.

Data Cleaning

Tweets were parsed and tokenized before using the model. All words were lowercased so that various capitalization wouldn't affect the results. All punctuation (including commas) was removed from tweets and was replaced with beginning and end of sentence speech tags (<s> and </s>). Given that tweets can be quoted (in the form of a hyperlink to said tweet), everything after (and including) the text pattern "http" was removed when such a pattern appeared in a tweet. Additionally, Twitter offers users the option to share or "Retweet" posts that other users have made. Since retweets, indicated in our corpus by RT, aren't necessarily direct words from a given user, all retweets were removed from the training corpus. When all is said and done, this tweet



is represented as follows:

['<s>', u'@nfl', u':', u'too', u'much', u'talk', u'not', u'enough', u'action', '</s>', '<s>', u'stand', u'for', u'the', u'national', u'anthem', '</s>']

Presidential speechs were cleaned in a manner that matched the output of the tweet parser. That is, all punction marks were removed and sentences were marked with appropriate tags. The first two lines of each document was removed as it provided the name and date of each speech. A few speechs were transcripts of interviews, and thus had reporter questions that were not presidential text. Given that these interjections were infrequent and could not be easily removed with regular expressions, we decided to leave them in with the assumption that they would not change any bi-gram or tri-gram probabilities in a significant way.

Fitting the Model

The model we choose to implement is an n-gram model with $n = 2$. To deal with the issue of unseen counts, we use Good-Turing Smoothing, explained in more detail below. Bigram and Trigram counts are computed from the various training corpuses. We will use three different types of data to train the model in order to see if any type of data works best. They are as follows:

1. Text from presidential speech corpus.
2. Part of speech tags from presidential speech corpus.
3. Tweet text from Obama, Trump and several U.S. senators.

Since the corpus of presidential speeches does not come with POS tags, we use a Hidden Markov Model to compute the most likely tag sequence for each sentence. The HMM is trained on data from the Brown Corpus built into Python's Natural Language Toolkit library.

Good Turing Smoothing

The smoothing method chosen for this project was the simple Good Turing smoothing method. This discounting algorthim uses the frequency of singletons words or N-grams to estimate the frequency of zero count events. The Good Turing algorthim first computes N_c or the number of N-grams that occur c times:

$$N_c = \sum_x count(x) = c$$

The intuition behind the algorithm is then to use the number of N-grams that occur c times to determine the probability of N-grams which occur c+1 times. The updated counts denoted by $c*$ are calculated using the following formula:

$$c* = (c+1)\frac{N_{c+1}}{N_c}$$

These new counts $c*$ are then used to replace the maxium likelihood counts for each N_c. The probability of a zero-count event according the Good Turing algorithm is calculated using the following equation:

$$P(N-gramwithzero-count) = \frac{N_1}{N}$$

In this equation, N_1 is the number of counts seen once and N is the total number of counts seen in the training set of the model.

An issue that arose when implementing this algorithm was that there were holes in the set of raw counts of counts N_c. Since the re-estimated value of a given $c*$ depends on N_{c+1} these holes had to filled in with an estimation in order for the model to function. This was accomplished by using the Simple Good-Turing method proposed by (Gale and Sampson 1995). This method works by smoothing the unseen N_c bins to replace all zeros. A linear regression function which maps the N_c to c in log space was used for this smoothing:

$$\log(N_c) = a + b\log(c)$$

Testing the Model

For each of the three types of training files, we have compiled a list of 10 quotes, each from a different speaker. The 10 possible speakers are then used as training files for the model. The probability of each quote for a given speaker is calculated as follows, using addition of logarithms to account for small probabilities.

$$\mathbb{P}(Q) = \prod_{i=i}^{n} \mathbb{P}(w_i|w_{i-1}) = 2^{\sum_{i=1}^{n} \log_2(\mathbb{P}(w_i|w_{i-1}))}$$

After computing the probability of each sentence for all possibile speakers, we attribute to each quote the speaker that yields the highest probability for the given quote. That is, our model predicts the speaker S of a quote Q as follows:

$$\hat{S} = \underset{s \in \text{ Possible Speakers}}{\arg\max} \mathbb{P}(Q|s)$$

Results

Presidential Speech Corpus

The probilities of each of the following presidential quotes were computed using the relavent speakers speechs as training files for the bigram model.

- Quote 1: <s> yes we can </s> - Barack Obama

- Quote 2: <s> and so my fellow Americans ask not what your country can do for you ask what you can do for your country </s> - JFK
- Quote 3: <s> when the president does it that means that it is not illegal </s> - Richard Nixon
- Quote 4: <s> the only thing we have to fear is fear itself </s> - FDR
- Quote 5: <s> liberty when it begins to take root is a plant of rapid growth </s> - George Washington
- Quote 6: <s> the care of human life and happiness and not their destruction is the first and only object of good government </s> - Thomas Jefferson
- Quote 7: <s> make america great again </s> -Donald Trump
- Quote 8: <s> a house divided against itself cannot stand </s> - Abraham Lincoln
- Quote 9: <s> we must teach our children to resolve their conflicts with words not weapons </s> - Bill Clinton
- Quote 10: <s> i would rather belong to a poor nation that was free than to a rich nation that had ceased to be in love with liberty </s> -Woodrow Wilson

Probability outputs for each speaker on each of the ten quotes are shown below in Figure 1. Overall, the model correctly identified 2/10 (20%) of the people who said these tweets. The two speakers that the model identified correctly were Barack Obama (quote 1) and Donald Trump (quote 7). Given that our model attributed 7 quotes to either Trump or Obama, this may be the result of larger training corupuses for these two presidents. For earlier presidents, our corpus contains fewer speeches than for more recent presidents. We have 50 Obama speeches in our data set compared to only 21 for George Washinton. Using more data for recent presidents likely yields fewer unseen bigrams, and thus higher probabilities. The most likely quote-speaker pair was Donald Trump "Make America Great Again", nearly 1000 times more probable than Obama's famous "Yes we Can!" slogan.

speaker	quote_1	quote_2	quote_3	quote_4	quote_5	quote_6	quote_7	quote_8	quote_9	quote_10
Barack Obama	2.01E-06	1.08E-45	1.10E-18	1.53E-27	3.18E-27	2.63E-56	1.70E-06	2.54E-06	2.55E-24	2.02E-55
JFK	2.43E-17	2.76E-86	2.56E-80	3.69E-72	1.63E-104	2.40E-119	2.50E-41	1.06E-77	7.81E-98	4.29E-162
Richard Nixon	3.91E-27	1.67E-112	3.49E-59	6.16E-60	1.01E-110	3.50E-123	2.14E-44	1.96E-65	2.47E-108	6.20E-153
FDR	3.17E-14	2.10E-26	5.07E-36	2.58E-21	1.89E-13	1.72E-65	7.16E-15	7.20E-12	1.80E-60	7.90E-48
George Washington	3.13E-42	1.27E-166	1.89E-74	6.46E-89	3.43E-131	1.71E-103	1.00E-50	2.34E-72	7.32E-145	3.47E-188
Thomas Jefferson	2.33E-32	2.38E-166	3.99E-79	3.32E-86	7.55E-116	3.52E-128	2.50E-51	5.46E-72	2.08E-125	3.48E-219
Donald Trump	2.11E-07	1.37E-52	1.83E-34	9.93E-07	9.88E-17	3.29E-45	4.06E-07	0.00121539	2.74E-15	1.36E-30
Abe Lincoln	4.37E-24	6.54E-157	8.58E-62	7.76E-68	2.30E-104	9.34E-169	2.22E-41	9.44E-72	4.39E-123	2.86E-178
Bill Clinton	3.19E-07	4.57E-90	8.13E-48	8.04E-53	8.77E-97	9.03E-118	4.42E-27	6.73E-31	1.92E-82	4.85E-151
Woodrow Wilson	6.56E-38	5.21E-130	2.74E-75	6.22E-54	8.13E-98	2.94E-104	7.69E-42	9.04E-67	7.86E-121	6.29E-150

Key
Correctly Identified
Incorrectly Identified

Probilities in each column are sorted from largest (green) to smallest (red)

Figure 1: Presidential Quote Results

Part of Speech Tags

The probilities of each of the following presidential quotes in POS tag form were computed using the relavent speakers speechs as training files for the bigram model.

- Quote 1: <s> ADV PRON VERB </s> - Barack Obama
- Quote 2: <s> CONJ ADV DET NOUN PRT VERB ADV DET DET NOUN VERB VERB ADP PRON VERB DET PRON VERB VERB ADP DET NOUN </s> - JFK
- Quote 3: <s> ADV DET NOUN VERB PRON ADP NOUN ADP PRON VERB ADV ADJ </s> - Richard Nixon
- Quote 4: <s> DET ADJ NOUN PRON VERB PRT VERB VERB VERB PRON </s> - FDR

- Quote 5: <s> NOUN ADV PRON VERB PRT VERB PRT VERB DET NOUN ADP ADJ NOUN </s> - George Washington
- Quote 6: <s> DET NOUN ADP ADJ NOUN CONJ NOUN CONJ ADV DET NOUN VERB DET ADJ CONJ ADJ NOUN ADP ADJ NOUN </s> - Thomas Jefferson
- Quote 7: <s> VERB NOUN ADJ ADV </s> - Donald Trump
- Quote 8: <s> DET NOUN VERB ADP PRON VERB NOUN </s> - Abe Lincoln
- Quote 9: <s> PRON VERB VERB DET NOUN PRT VERB DET NOUN ADP NOUN ADV NOUN </s> - Bill Clinton
- Quote 10 <s> PRON VERB ADV VERB ADP DET ADJ NOUN PRON VERB ADJ ADP ADP DET ADJ NOUN PRON VERB VERB PRT VERB ADP VERB ADP NOUN </s> - Woodrow Wilson

Probability outputs for each speaker on each of the ten POS tagged quotes are shown below in Figure 2. Overall, the model correctly identified 1/10 (10%) of the people who uttered the POS Tag sequences. Notice that the probabilities of each speaker-tag sequence pair are much larger than each speaker-quote pair. This is likely due to the fewer possible tags, and thus fewer unseen bigrams, as compared to training with words. Our model attributed every single POS tag to Donal Trump. Since there was little variation on each quote and Trump has the most speeches in our corpus, this could be due to Trump having fewer unseen bigrams. In any case, there seems to be very little overall discrenability between speakers on the basis of POS tags.

speaker	quote_1	quote_2	quote_3	quote_4	quote_5	quote_6	quote_7	quote_8	quote_9	quote_10
Barack Obama	1.58E-05	2.50E-28	2.50E-16	6.29E-14	1.58E-17	6.27E-26	9.99E-07	2.51E-10	1.58E-17	6.26E-32
JFK	5.61E-06	6.41E-31	8.60E-18	3.63E-15	4.18E-19	2.71E-28	2.73E-07	3.15E-11	4.18E-19	7.39E-35
Richard Nixon	9.15E-06	1.07E-29	4.21E-17	1.39E-14	2.32E-18	3.53E-27	5.03E-07	8.37E-11	2.32E-18	1.77E-33
FDR	1.33E-05	9.33E-29	1.44E-16	3.93E-14	8.67E-18	2.55E-26	8.06E-07	1.78E-10	8.67E-18	2.06E-32
George Washington	4.58E-06	2.01E-31	4.46E-18	2.08E-15	2.06E-19	9.37E-29	2.12E-07	2.10E-11	2.06E-19	1.99E-35
Thomas Jefferson	2.28E-06	3.66E-33	4.64E-19	3.07E-16	1.80E-20	2.42E-30	8.88E-08	5.22E-12	1.80E-20	2.15E-37
Donald Trump	4.29E-05	7.74E-26	6.41E-15	9.78E-13	5.19E-16	1.18E-23	3.48E-06	1.84E-09	5.19E-16	4.10E-29
Abe Lincoln	6.68E-06	1.75E-30	1.51E-17	5.86E-15	7.70E-19	6.76E-28	3.40E-07	4.46E-11	7.70E-19	2.29E-34
Bill Clinton	2.04E-05	1.06E-27	5.67E-16	1.26E-13	3.81E-17	2.35E-25	1.37E-06	4.15E-10	3.81E-17	3.22E-31
Woodrow Wilson	5.97E-06	9.12E-31	1.05E-17	4.29E-15	5.18E-19	3.73E-28	2.95E-07	3.56E-11	5.18E-19	1.10E-34

Key	Probilities in each column are sorted from largest (green) to smallest (red)
Correctly Identified	
Incorrectly Identified	

Figure 2: POS Tags Results

Tweet Data

The probilities of each of the following tweets were computed using the relavent speakers' twitter history as training files for the bigram model.

- Quote 1: <s> yes we can </s> - Barack Obama
- Quote 2: <s> make america great again </s> - Donald Trump
- Quote 3 <s> stark reminder of the threat radical islamic terror still poses to our homeland </s> -Ted Cruz
- Quote 4: <s> it's not by giving massive tax breaks to your billionaire friends </s> - Bernie Sanders
- Quote 5: <s> tax bill equals swamp creature </s> - Tim Kaine
- Quote 6: <s> @SenBernieSanders & I are talking about things democrats are fighting for in this spending bill that would make a big difference to working families </s> - Elizabeth Warren
- Quote 7: <s> whether you are in the media politics or anywhere else abuse of power is unacceptable & shouldn't be tolerated at any place at any level </s> - Lisa Murkowski

- Quote 8: <s> in a foreign relations committee hearing today i discussed how u.S. nuclear forces protect our nation and allies in the 21st century </s> - Marco Rubio
- Quote 9: <s> president trump did the people of utah a great favor today by rolling back harmful land use restrictions in southern utah #utpol </s> - Mike Lee
- Quote 10: <s> better #GetCovered it's the last weekend to sign up for 2018 health coverage before the 12/15 deadline </s> - Chuck Schumer

Probability outputs for each speaker on each of the ten quotes are shown below in Figure 3. Overall, the model correctly identified 5/10 (50%) of the people who said these tweets. The most likely tweet-speaker pair was Donald Trump "Make America Great Again". Interestingly, this combination was over 10,000,000,000 times more likely any other combination! While some of this discrepancy can be expained by the fact that it is a shorter tweet, it also speaks to the unprescedented manner in which President Trump uses Twitter compared to other politicians.

speaker	quote_1	quote_2	quote_3	quote_4	quote_5	quote_6	quote_7	quote_8	quote_9	quote_10
Barack Obama	1.66E-32	1.58E-33	3.30E-106	8.80E-108	1.00E-60	7.47E-161	9.53E-230	1.02E-169	3.80E-219	1.20E-130
Donald Trump	1.19E-25	5.56E-05	1.08E-97	9.22E-93	3.23E-64	7.89E-171	5.72E-231	4.72E-164	3.31E-184	1.93E-134
Ted Cruz	1.13E-33	1.03E-43	6.65E-88	1.09E-107	1.03E-63	3.52E-214	3.40E-240	1.07E-178	3.94E-207	2.56E-185
Bernie Sanders	1.98E-16	3.34E-48	4.15E-117	2.19E-73	6.27E-54	3.37E-172	5.39E-226	2.95E-182	2.19E-196	1.13E-154
Tim Kaine	1.12E-27	5.72E-27	1.38E-127	1.56E-98	5.42E-64	7.88E-162	8.07E-247	7.67E-167	1.37E-206	7.70E-147
Elizabeth Warren	6.18E-36	2.93E-34	1.10E-126	2.31E-108	1.00E-60	5.75E-178	1.08E-235	3.98E-156	3.93E-200	7.82E-139
Lisa Murkowski	1.27E-23	3.06E-44	8.16E-126	4.75E-113	6.12E-64	6.30E-168	7.97E-222	1.79E-163	5.74E-229	1.22E-140
Marco Rubio	4.39E-33	1.04E-44	9.96E-107	2.84E-118	1.00E-60	8.55E-249	1.29E-239	4.61E-118	1.60E-198	5.11E-167
Mike Lee	1.37E-27	6.20E-35	1.30E-124	2.12E-94	5.58E-54	1.84E-179	1.32E-203	2.19E-158	2.80E-183	3.76E-151
Chuck Schumer	9.07E-27	3.62E-37	1.53E-117	3.13E-101	7.82E-54	2.19E-176	8.03E-238	2.23E-184	9.58E-190	8.92E-163

Key
Correctly Identified
Incorrectly Identified

Probilities in each column are sorted from largest (green) to smallest (red)

Figure 3: Tweet Results

Discussion

Our results have shown that distinguising between speakers using tweets seems to be easier than using direct quotes from speechs. Both of these methods perform much better than using POS tags. These results suggest that people may have the tendency to be more formal/scripted in speeches than on twitter. Moreover, our results seem to indicate that while the types of words presidents use over time may change, the general distribution of part of speech tags since the founding of the United States has remained relatively constant.

Few (published) attempts have been made at speaker identification using n-gram models. (Japi, 2015) using a trigram distribution with Markov Chains to simulate novel text in the style of a given author. (Zheng, Li, Chen, and Huang, 2006) studied the problem of author attribution on English/Chinese online messaging data, and obtained a 70-95 % accuracy rate using advanced machine learning techniques like neural networks and support vector machines. In a monumental 1964 work, Mosteller and Wallace used Bayesian Inference to attribute 12 of the Federalist Papers to John Madison.

The fact that our accuracy is relatively low is not surpirsing given the simplicity of the bigram model. Certain machine learning classification algorithms that have been developed in recent years are likely to outperform our basic counting method. Still, there are a few ways we could improve our methodology to obtain better accuracy. The easiest way to improve accuracy would be to obtain more data, especially for earlier presidents. This would lead to fewer unseen bigrams, and would place all potential speakers on a level playing field. Additionally, with more data, we could attempt to expand our model to a trigram model. Having so many

unseen counts made trigram modeling impossible for this task, but perhaps with enough data, using trigrams would be more powerful. Another extension for future work would be to examine other types of smoothing. We choose to use Good Turing smoothing in this model, but perhaps add-λ smoothing or Katz's Backoff may perform better. Overall, are results are about was to be expected given the limits of an n-gram language model.

Works Cited

- Gale, W. and G. Sampson. *Good Turing Estimation Without Tears.* Journal of Quantitative Linguistics, vol. 2, 217-237, 1995.
- Japi, A. *Mimicking Writing Style With Markov Chains.* The Sopranos, Silicon Valley, and Summer Afternoons. http://aakashjapi.com/mimicking-writing-style-with-markov-chains/
- Mosteller, F. and D. L. Wallace. *Inference and Disputed Authorship: The Federalist.* Reading, MA., 1964 -Zheng, R., Li, J., Chen, H. and Huang, Z. *A framework for authorship identification of online messages: Writing-style features and classification techniques.* J. Am. Soc. Inf. Sci., 57: 378–393, 2006.

Apendix

Code for this project can be found on GitHub.