CSCI 5832: Natural Language Processing - Fall'18

Assignment 1: How many words?

Keval Dhruveshkumar Shah

(Dated: September 4, 2018)

I have lived in Gujarat (India) all my life and Gujarati has been my primary mode of communication. I believe that I have a good command over the language, with a decent vocabulary in comparison to the speakers born in the 90s. I have observed my grandparents use some words that are still unknown to my parents and the same happens to me when I hear my parents talk. With the increasing influence of the West and English language, the number of words in the 'dying words' category has been increasing with every passing generation. On the other hand, it is a rare occasion when a new Gujarati word is added to the word pool. This implies that the share of English/Hindi words in a Gujarati speaker's vocabulary has been increasing, the total words being the same. This, in turn, leads to a native Gujarati speaker of a newer generation, knowing less Gujarati words. I have used a few facts along with my assumptions to estimate the number of words that I know. This is done by accounting for the words known to my ancestors, that have been unknown to me. This rough estimate results in me knowing approximately **5740** Gujarati words.

I. Facts

- Gujarati is one of the Indian languages (Indo-Aryan descendant). It has more than 50 million speakers and is said to have originated 700 years ago [1].
- The popular Gujarati encyclopedia and dictionary, Bhagavadgomandal has 281,377 words [2].
- The English dictionary Merriam-Webster has approximately 470,000 words [3].
- I was born in 1996, my father in 1970, my grandfather in 1943 and my great-grandfather in 1921. Hence, the average generation gap is of 25 years.

II. Assumptions

- All the words in the vocabulary of a given speaker has to be a subset of the words in the dictionary/encyclopedia. This way the words remain unique and we avoid redundancy in the word pool.
- Shakespeare used 35000 words in his literature. It is safe to assume that he knew 60,000 words [4]. Hence, around the time of origin, A prominent Gujarati language scholar would have a vocabulary of 36000 words (In proportion to the number of words in English and Gujarati dictionaries). I am taking the origin year as my reference because around this time, the language has been uninfluenced by any external factors.
- For a native Gujarati speaker, the vocabulary is half of that of a scholar: 18000 words. Let me assume that my ancestors that lived 700 years ago, had the vocabulary of this size and were not language specialists.
- I will assume that from one generation to the other, there is a 4% decrease in the number of words that a native speaker knows. The main reason is due to the adoption of English/Hindi words over Gujarati words. For example, I have stopped using the word '*phaḷ*' for fruits, I use the English word 'fruit' at all times. In fact, even my parents use 'fruit', it is just my grandparents that use '*phaḷ*'. Similarly, the word 'Guṇākār' is used for multiplication by my parents and not me.

III. Estimating my vocabulary size

- Since the origin of Gujarati language that was 700 years ago, $700/25 = 28$ generations have passed for my family until mine.
- For each generation, the vocabulary decreased by 4% of the previous generation. Hence, this is a Geometric Progression $g(n)$ with a common ratio of 0.96 and an initial value of 18000. We need to find $g(28)$.

$$g(28) = 18000 \times (0.96)^{28} = 5739.4$$

- I know **5740** words of Gujarati.
- **Note:** According to this technique for estimation, the number of words known to a native Gujarati speaker of future generations will reach zero at some point. Since a fluent speaker requires to know a minimum number of words, say n, this technique is not viable when it gives results less than n. A value that is less than n implies that the speaker cannot converse smoothly in the language.

---

[1] T. S. Times, "Gujarati: The language spoken by more than 55 million people." (2017).
[2] B. Sahib, *Bhagwadgomandal* (2017).
[3] Merriam-Webster, "How many words are there in english?" .
[4] B. Efron and R. Thisted, Biometrika **63**, 435 (1976).