Evolution of Language from an Information Theoretic Point of View

Kata Naszádi Hannah Seitz Madhumita

Formal vs. Informal Hindi - Evolution of Language from an Information Theoretic Point of View

Question

Can the difference between formal and informal Hindi be explained by information theoretic principles?

Background

Formal Hindi: Uses words directly derived from Sanskrit.
Informal Hindi: Uses evolved forms of those words.
Example: 'chandrama' (formal) vs. 'chaand' (informal)

Predictions

Formal Hindi is a less optimal code for communication than informal Hindi.

Predicting word length

Assuming that optimal code approximates Huffman-coding, in informal Hindi information content is a better predictor of word length

Training and Test Corpora

Ideal corpus:

- **Informal Hindi:** Modern literature text written in colloquial language.
- **Formal Hindi:** Modern literature text written in formal(pure) Hindi.

Feasible corpus:

- **Informal Hindi:** Crawled corpus of news articles. A small corpus can be obtained from corpus for WMT '14 translation task.
- **Formal Hindi:** Crawled corpus of Hindi Wikipedia and editorials from newspapers.

Using news articles and editorials from the same set of newspapers controls for domain differences in terms of topics and time lapse.

Procedure

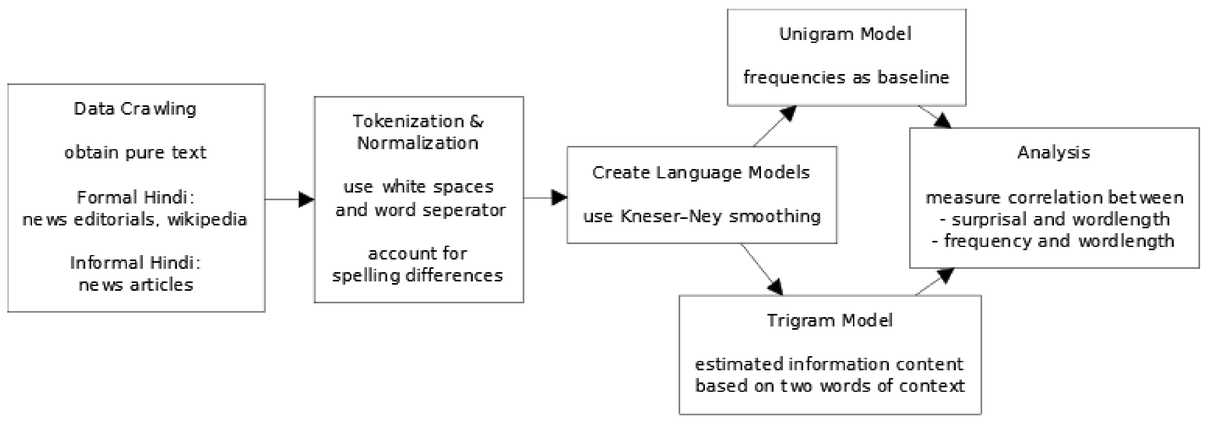

Fig. 1: pipeline of the experiment

Normalization of words

Hindi text consists of some spelling variations on account of usage of consonants instead of certain nasal vowels like the 'anuswar'. In order to evaluate word length correctly, word spellings need to be normalized according to common standards.

Control for domain differences

Obtain data from the same domains to do the same analysis for a language, which does not have a formal and an informal variant, e.g. English.
This shows which part of the results can be explained by domain differences.

Expected Results

Predicting word length from average information content

$$\sum_c P(C = c)|W = w) * logP(W = w|C = c)$$

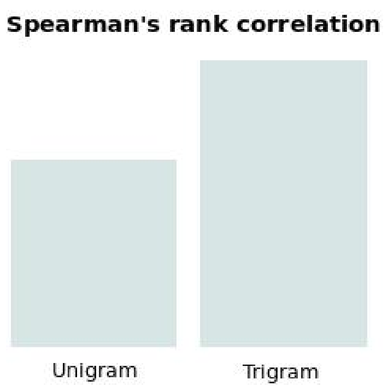

Fig. 2: within language

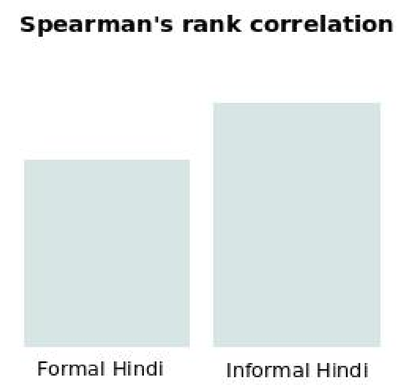

Fig. 3: between languages

Alternative hypothesis for wordform changes

The changed forms usually have shorter word lengths and have easier consonant structures, making it phonologically easier to produce. If it is only **pronunciation** that governs changes in wordforms, we would expect that **frequency** explains which words shortened the most.

Future Prospects: Similar Behavior for Other Languages Expected

Modern Standard Arabic: derived from Quran, mostly used in written/ formal spoken context. Spoken language is usually less conservative. Additionally needs to consider that there are different variants of Arabic spoken in different countries.

Syntax

In language variants, where syntax has also been changed, it would be interesting to measure how average **dependency length** changed.

Persian: Several differences between formal and informal Persian, some of them being differences in pronunciations and verb endings. E.g., "mi:ravad" (formal - meaning "someone is going") changes to "mi:reh" (informal). Meaning remains the same.