“Statistical Identification of Language” –Ted Dunning

Kristinn
Reykjavík University

Languages

- Halló
- Hello
- Hallo
- Hola
- Bonjour
- 안녕하세요
- こんにちは
- 你好
- 你好

Languages

- Halló
  - Íslenska
- Hello
  - English
- Hallo
  - German
- Hola
  - Spanish
- Bonjour
  - French
- 안녕하세요
  - Korean
- こんにちは
  - Japanese
- 你好
  - Chinese (traditional)
- 你好
  - Chinese (simplified)

Introduction

- Statistical based program has been written which learns to distinguish between languages, e.g. Spanish, English, French
  - 100 words of code
  - Only needs a few thousand words of sample text in order to learn the language
  - Works very well with 92%+ accuracy and more accurate with a larger “learning text”.
    - Learning text implies a sample of text which the computer program can “tokenize”

Bayesian Method with Markov Probablity

- Bayesian logic probablity, i.e. deciding which event is causing the observation by observing
- Markov probability is analyzing past events to predict future events, i.e. weather systems.

Previous Work: Unique Letter Combinations

- Enumerating a number of short sequences from text which are unique to a particular language

- Drawback: Languages sometimes adobt words from other cultures, e.g. Geography, Movies, Names, etc..

Language	String
Dutch	"vnd"
English	"ery "
French	"eux "
Gaelic	"mh"
German	" der "
Italian	"cchi "
Portuguese	" seu "
Serbo-croat	"lj"
Spanish	" ir "

Previous Work: Common Words

- Devise a list of commonly used words in a language.
  - English: the, of, to, and, a, in, is, it, you, “etc..”
  - German: der/die/das, und, sein, in, ein, zu, “etc..”
  - Spanish: el/la, de, que, y, a, en, un, ser, se, “etc..”

- Drawback: not all langauge phrases contain these words. Difficult to tokenize a language such as Chinese and therefore impossible to implement this method.

Previous Work: N-gram counting with rank order

- Ad hoc rank ordering of tokenized text. Or, comparing tokenized text to a large library of text from a source such as network news groups.

- Drawback: Input had to be tokenized and the statistical rank order of text was dependant on longer text sizes, i.e. 4K or 700 words

Markov Method

- The Markov model defines a random variables whose values are strings from an alphabet X, and where the probability of a particular string S is:

$$p(S) = p(s_1 \ldots s_n) = p(s_1) \prod_{i=2}^{n} p(s_i \mid s_{i-1})$$

- We are loooking at the sequence of characters in a learning text, but not considering language structure.

```
0  hm 1 imuandno˜doc ni leotLs Aiqe1pdt6cf tlc.teontctrrdsxo˜es loo oil3s
1   ˜ a meston s  oflas n,˜ 2 nikexihiomanotrmo s,˜125 0 3 1 35  fo there
2  s ist  des anat p sup sures Alihows raiial on terliketicany of prelly
3  approduction where.  If the lineral wate probability the or likelihood
4 sumed normal of the normal distribution.  Church, Gale, Willings. This
5 ˜k sub 1} sup {n-k}  .EN  where than roughly 5.  This agreemented by th
6 these mean is not words can be said to specify appear.  McDonald.  1989
```

Bayesian Method

- If we are choosing between A and B given an observation X, where we feel that we know how A or B might affect the distribution of X, we can use Bayes' theorem.

$$p(A, X) = p(A \mid X)p(X) = p(X \mid A)\, p(A)$$

- *looking for what happened before this current character. What is most porbable since this event already occured.*

Summarised

- This method reads from a learning text of a relatively small size.
  - Test results
    - Language: English and Spanish
    - Learning text: 10 training texts of size: 1000, 2000, 5000, 10,000, and 50,000 bytes length
    - Tests Texts: 100 different tests: 10, 20, 50, 100, and 500 bytes in length

Test Results

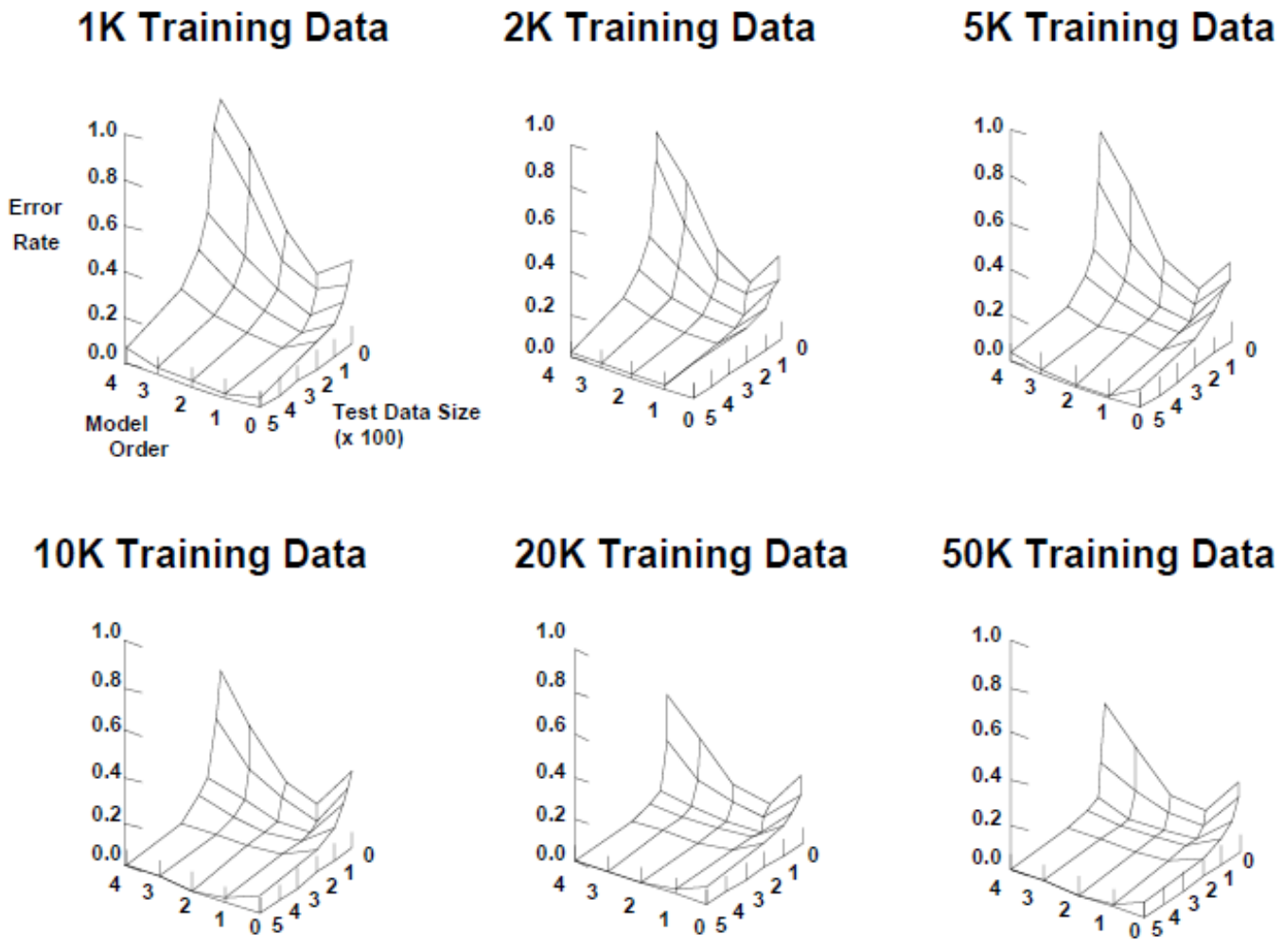

Figure 1

For any given combination of test string size and training set size, there is an optimum order for the language model. In all cases, longer test strings and more training data improve error performance.

Why and Where?

- Genetic sequence analyzers
  - Determining the species which a particular animal or plant, etc..
- Determining the origin of a language.
  - http://whatlanguageisthis.com/

Questions