Words and Echoes: Assessing and Mitigating the Non-Randomness Problem in Word Frequency Distribution Modeling

Marco Baroni\textsuperscript{1}  Stefan Evert\textsuperscript{2}

\textsuperscript{1}Center for Mind/Brain Sciences
University of Trento
\textsuperscript{2}Cognitive Science Institute
University of Osnabrück

ACL Conference
Prague, 27 June 2007

Outline

Introduction
LNRE models
Evaluation of LNRE models
Results 1
Non-randomness and echoes
Results 2
Conclusion

What is word frequency distribution modelling?

- We are interested in analyzing type-token statistics . . .
  - such as vocabulary size, type-token ratio,
    or the proportion of hapax legomena
  - ... in (random) samples . . .
    - more about (non-)randomness later
  - ... from type-rich populations . . .
    - words, n-grams and phrases are just the obvious examples
    - also subcategorisation patterns, named entities, treebank
      grammar rules, collocations, insect species, etc.
  - ... with a skewed, “Zipfian” distribution
    - in fact, our models are all based on Zipf’s law

Type-token statistics

Given a sample of $N_0$ tokens, we are interested in these observations:

- vocabulary size $V$ (= number of different types)
- number $V_1$ of hapaxes (= types occurring just once)
- frequency spectrum $V_m$ for $m \in \mathbb{N}$ (= types occurring exactly $m$ times)
- development of $V(N)$ and $V_m(N)$ for increasing samples of $0 \leq N \leq N_0$ tokens (→ vocabulary growth)
- not in frequencies of specific types
  - focus on low-frequency data
LNRE models & applications

- Statistical models for such distributions are known as LNRE models (Baayen 2001) and allow us to:
  - estimate population vocabulary size $S$
  - model distribution of type probabilities in population
  - extrapolate vocabulary growth
  - predict frequency spectrum of unseen data

- Some applications of LNRE models
  - measuring morphological productivity
  - vocabulary richness (stilometry, child language acquisition)
  - quantifying data sparseness
  - empirically justified Bayesian priors
  - Good-Turing smoothing
  - reliability of statistical inference from low-frequency data

LNRE population models

- LNRE model describes distribution of type probabilities in a population with a large number of rare events

- One possibility is to specify an equation for Zipf-ranked type probabilities, e.g. the Zipf-Mandelbrot law

$$\pi_k = \frac{C}{(k + b)^a} \quad (a > 1, b > 0)$$

- Better representation as type density function
  - E.g. for Zipf-Mandelbrot:

$$g(\pi) = C\pi^{-\alpha - 1} \quad (\alpha = \frac{1}{a})$$

- LNRE models in zipfR library: ZM, IZM, GiGP

Expectation & variance

- Expected values $E[V(N)]$ and $E[V_m(N)]$ for random sample of $N$ tokens can easily be calculated:

$$E[V] = \int_0^1 (1 - e^{-N\pi})g(\pi) \, d\pi$$

$$E[V_m] = \int_0^1 \frac{(N\pi^m)^m}{m!} e^{-N\pi}g(\pi) \, d\pi$$

- Variances $\text{Var}[V(N)]$ and $\text{Var}[V_m(N)]$ are slightly uglier, but also easy to calculate (same for covariances)
LNRE parameter estimation

- Estimate LNRE model parameters by comparison of observed and expected frequency spectrum
- Nonlinear minimization of cost function (e.g. MSE)
- Measure goodness-of-fit by multivariate chi-squared test (Baayen 2001)
- General observation: GIGP (and fZM) achieve much better fit than simple ZM model
  - ZM assumes an infinite population vocabulary!

Goodness-of-fit & evaluation

- Goodness-of-fit measures how well model describes training data (df-adjustment corrects for overtraining)
- Evaluation measures we are really interested in:
  - accurate extrapolation of vocabulary growth
  - reliable prediction of unseen data
  - how well model describes true population distribution
- No problem! For a random sample, goodness-of-fit is a reliable predictor of “interesting” evaluation measures
  - overtraining controlled by variance estimates
- Unfortunately… corpora aren’t random samples
  - key problem: not sampled at token level
  - our empirical evaluation will show how seriously LNRE models are affected by the non-randomness of corpus data

Outline

Introduction
LNRE models
Evaluation of LNRE models
Results 1
Non-randomness and echoes
Results 2
Conclusion

Data-set preparation and model training

- Corpora:
  - British National Corpus (English “balanced” corpus)
  - deWaC (German Web data)
  - la Repubblica (Italian newspaper data)
- From each corpus, we take 20 non-overlapping samples of randomly selected documents
- Each of the samples split into
  - 1 million tokens for training
  - 3 million tokens for testing
- Parameters of ZM, fZM and GIGP estimated on each training set
- Models used to predict vocabulary size $V$ and number of hapaxes $V_1$ at sample sizes of 1, 2 and 3 million tokens
rMSE

- Prediction performance measured by relative error:
  \[ e = \frac{E[V(N)] - V(N)}{V(N)} \]

- Square root of mean square relative error (rMSE), across 20 samples:
  \[ \sqrt{\text{rMSE}} = \sqrt{\frac{1}{20} \cdot \sum_{i=1}^{20} (e_i)^2} \]

Outline

Introduction
LNRE models
Evaluation of LNRE models
Results 1
Non-randomness and echoes
Results 2
Conclusion

la Repubblica rMSE (V)

la Repubblica rMSE (V₁)
**Goodness-of-fit on training set and prediction accuracy**

Correlation: $r = -0.89$

**Outline**

- **Introduction**
- **LNRE models**
- **Evaluation of LNRE models**
- **Results 1**
  - Non-randomness and echoes
- **Results 2**
- **Conclusion**

**Term clustering**

- *chondritic* occurs 4 times in the BNC, but all occurrences are in the same (scientific) document
- As famously put by Church (2000):
  - The chance of two Noriegas is closer to $p/2$ than $p^2$
- Term clustering leads to *underestimation* of vocabulary size (because number of hapaxes is reduced)
Baayen's (2001) *partition-adjusted models*

- Only current non-randomness correction method that can be used in the context of LNRE modeling
  - Models of Church and Gale (1995) and Katz (1996) account explicitly for non-random distributions (of the term clustering kind), but there is no tractable mathematical model that would integrate them into LNRE statistics
  - For Baayen's parameter-adjusted models, population distribution depends on $N \rightarrow$ not a proper LNRE model
- Population partitioned into
  - normal types that satisfy random sampling assumption and
  - totally underdispersed types that concentrate all occurrences in a single “burst”
- Standard LNRE model used for normal part of the population; simple linear growth for underdispersed part

*Echo adjustment*

- Tackle non-randomness as a *pre-processing* problem: the issue is with the way we count occurrences of types
- Rare, topic-specific content words occur maximally once in a document
- All other apparent instances of such words are instances of a special “anaphoric” type that has function of “echoing” the content words in a document
- Before:
  ... the result of an impactor of carbonaceous chondritic composition ... A typical strength of a chondritic impactor is ...
- After:
  ... the result of an impactor of carbonaceous chondritic composition ... A typical strength of a ECHO ECHO is ...

*Echo adjustment*

- After echo adjustment, we are effectively counting *document frequencies*, that are not subject to within-document term clustering effects
- However, by replacing repeated words with echo tokens, we can stick to word token sampling model, so that LNRE models can be applied “as is”

**Outline**

- Introduction
- LNRE models
- Evaluation of LNRE models
- Results 1
- Non-randomness and echoes
- Results 2
- Conclusion
Goodness-of-fit on training set and prediction accuracy
Correlation: $r = 0.94$

Outline

- Introduction
- LNRE models
- Evaluation of LNRE models
- Results 1
- Non-randomness and echoes
- Results 2
- Conclusion
Directions for future work

- Echo-adjusted predictions pertain to distributions of document frequencies: what are the implications of this?

- Quality still not fully satisfying, especially at large prediction sizes (we would like to extrapolate $V$ and other quantities to 100 times the training size and more!)

Some references

H. Baayen. 1992. Quantitative aspects of morphological productivity. *Yearbook of Morphology* 1991, 109-150.

H. Baayen. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

K. Church. 2000. Empirical estimates of adaptation: the chance of two Noriegas is closer to $p/2$ than $p^2$. *Proceedings of the 17th Conference on Computational Linguistics*, 180-186.

K. Church and W.A. Gale. 1995. Poisson mixtures. *Journal of Natural Language Engineering* 1, 163-190.

S. Evert. 2004. A simple LNRE model for random character sequences. *Proceedings of JADT 2004*, 411-422.

S. Evert and M. Baroni. 2006. Testing the extrapolation quality of word frequency models. *Proceedings of Corpus Linguistics 2005*.

S. Katz. 1996. Distribution of content words and phrases in text and language modeling. *Natural Language Engineering*, 2(2) 15-59.
Appendix: result details for $\sqrt{rMSE}$

rMSE for $E[V]$ vs. $V$ (DEWAC)

| Method | $N_0$ | $2N_0$ | $3N_0$ |
|--------|-------|--------|--------|
| ZM     |       |        |        |
| IZM    |       |        |        |
| GIGP   |       |        |        |
| echo   |       |        |        |
| echo   |       |        |        |
| partition |    |        |        |

Appendix: bias & variance of predictors

Relative error: $E[V]$ vs. $V$ (REPUBBLICA)

| Method | $N_0$ | $2N_0$ | $3N_0$ |
|--------|-------|--------|--------|
| ZM     |       |        |        |
| IZM    |       |        |        |
| GIGP   |       |        |        |
| echo   |       |        |        |
| echo   |       |        |        |
| partition |    |        |        |

Appendix: bias & variance of predictors

Relative error: $E[V]$ vs. $V$ (BNC)

| Method | $N_0$ | $2N_0$ | $3N_0$ |
|--------|-------|--------|--------|
| ZM     |       |        |        |
| IZM    |       |        |        |
| GIGP   |       |        |        |
| echo   |       |        |        |
| echo   |       |        |        |
| partition |    |        |        |

Appendix: bias & variance of predictors

Relative error: $E[V]$ vs. $V$ (DEWAC)

| Method | $N_0$ | $2N_0$ | $3N_0$ |
|--------|-------|--------|--------|
| ZM     |       |        |        |
| IZM    |       |        |        |
| GIGP   |       |        |        |
| echo   |       |        |        |
| echo   |       |        |        |
| partition |    |        |        |
Appendix: bias & variance for randomized data

Relative error: $E[V]$ vs. $V$ (REPUBBLICA)

Relative error: $E[V]$ vs. $V$ (BNC)

Relative error: $E[V]$ vs. $V$ (DEWAC)

Appendix: prediction vs. extrapolation

rMSE for $E[V]$ vs. $V$ (test set & extrapolation, REPUBBLICA)
Appendix: prediction vs. extrapolation

rMSE for $E[V]$ vs. $V$ (test set & extrapolation, BNC)

rMSE for $E[V]$ vs. $V$ (test set & extrapolation, DEWAC)

Appendix: type & probability density of LNRE models