Some of my Best Friends are Linguists

(LREC 2004)

Frederick Jelinek

Johns Hopkins University

THANKS TO: E. Brill, L. Burzio, W. Byrne, C. Cieri, J. Eisner, R. Frank, L. Guthrie, S. Khudanpur, G. Leech, M. Liberman, M. Marcus, M. Palmer, P. Smolensky, and D. Yarowsky

The Quote

"Whenever I fire a linguist our system performance improves"

From my talk entitled:

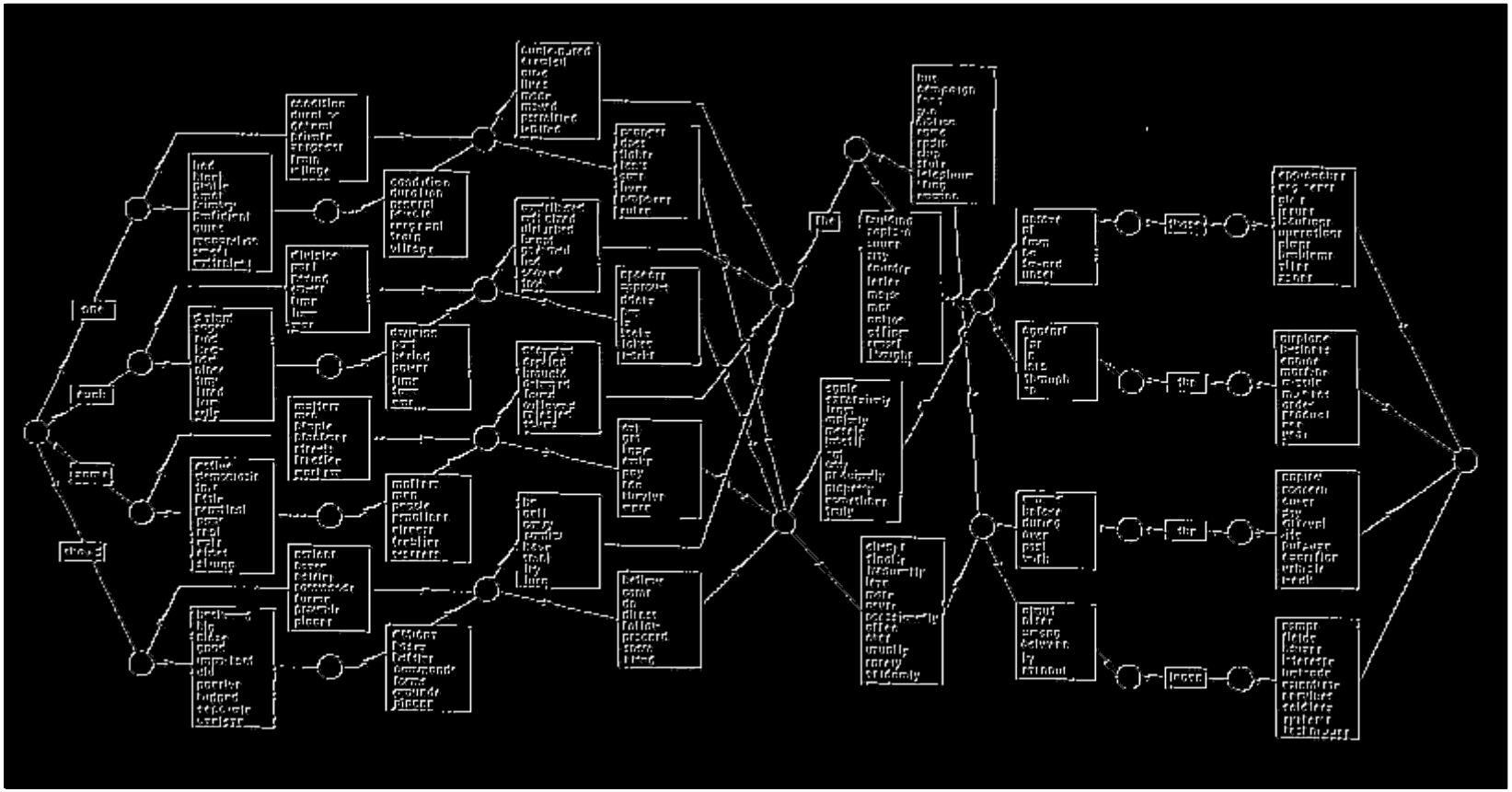
Applying Information Theoretic Methods: Evaluation of Grammar Quality

Workshop on Evaluation of NLP Systems, Wayne PA, December 1988

Hope Springs Eternal

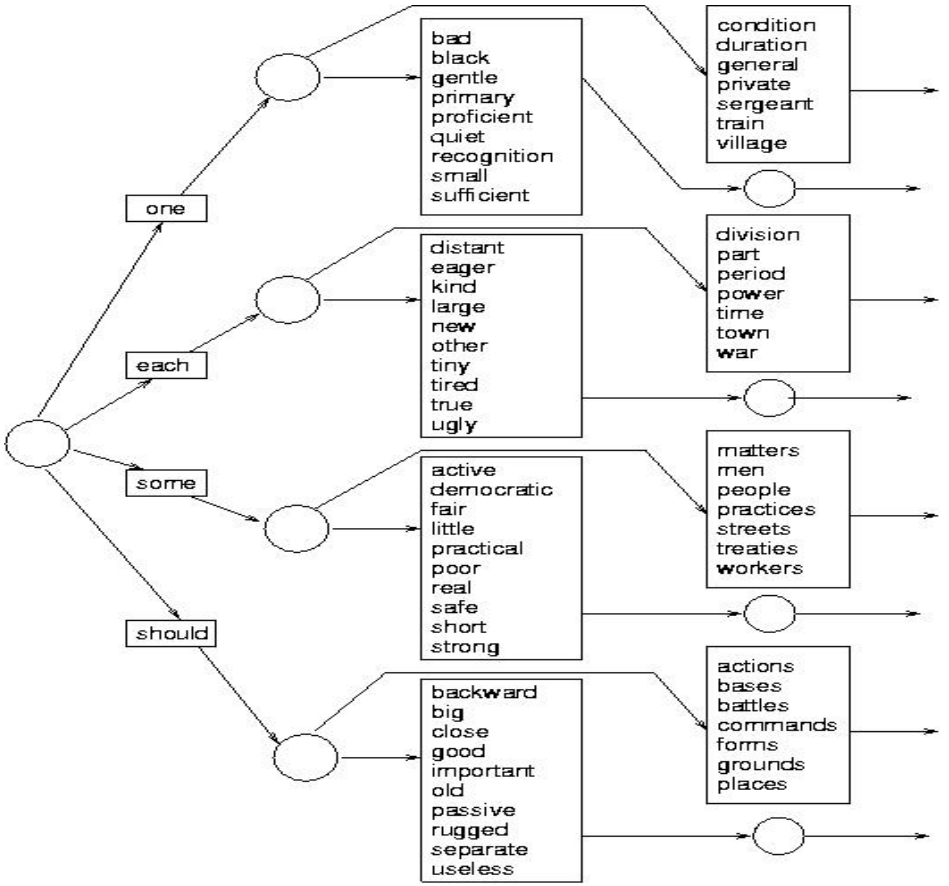
- My colleagues and I always hoped that linguistics will eventually allow us to strike gold
  - HMM tagging (Bahl & Mercer 1976)
- The quote accentuated a certain situation that existed in ASR in the seventies and in NLP in the eighties
- The following is an illustration

The Raleigh Finite State Language



May 28, 2004 Johns Hopkins

Zoom on Raleigh Language



When Linguists Left the Group

Task: New Raleigh Language

- Acoustic model 1:
  - phonetic baseforms: three ←→ ?rí
  - Model statistics estimated by experts (35% accuracy)
- Acoustic model 2:
  - phonetic baseforms: three ←→ ?rí
  - Model statistics estimated automatically from data (75% accuracy)
- Acoustic model 3:
  - orthographic baseforms: three ←→ THREE
  - Model statistics estimated automatically from data (43% accuracy)

Judgment of J.R. Pierce

“ Whither Speech Recognition?” JASA 1969

…ASR is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, or going to the moon.

Most recognizers behave not like scientists, but like mad inventors or untrustworthy engineers.

…performance will continue to be very limited unless the recognizing device understands what is being said with something of the facility of a native speaker (that is, better than a foreigner fluent in the language)

Any application of the foregoing discussion to work in the general area of pattern recognition is left as an exercise for the reader.

J.R. Pierce Activities

- Respected communication engineer at ATT Bell Laboratories
  - Colleague of Shannon, Shockley, Bardeen, Darlington, etc
- Head of the 1966 Automatic Language Processing Committee of Defense Dept.
  - Put a stop to government support of MT
  - Believed MT research would neither effect early cost reduction, nor improve performance, nor is meeting an operational need.
  - Is not an intellectually challenging field per se

The Situation in 1970s

- Rules and AI govern NLP and speech research
- No distinction between training and test
- IBM linguists had respect but underestimated ASR problem
- Chomsky thought that statistics were illegitimate
- ARPA project on ASR (1971 – 1976) dominated by AI (except for Jim Baker at CMU)

The View of the IBM Group

- Linguistic intuition combined with ability to extract information will determine the structure of models and their parameterization
- Parameter values will be estimated from (annotated) data
- We will rely on advice of linguists to create resources
  - E.G., *Annotating Noun Argument Structure for NomBank*
- The problem is **not** of direct interest to linguists

Creation of Linguistic Resources

- Brown Corpus (1967)
- Lancaster – Oslo – Bergen corpus (1970)
- Lancaster POS tagging by rule (1982)
- Lancaster treebank (1983 – 1986)
  - Geoff Leech and Geoff Sampson
- IBM commissions 2 – 3 M word treebank at Lancaster (1987)

Founding of LDC

- Meeting with Jack Schwartz at DARPA (1987)
- UPenn willing to host LDC (Austin, Jan 1988)
- DARPA workshop in Mohonk, NY (May 1988)
- UPenn treebank (1992)
- British National Corpus (1991 – 95)

Data Driven Parsing

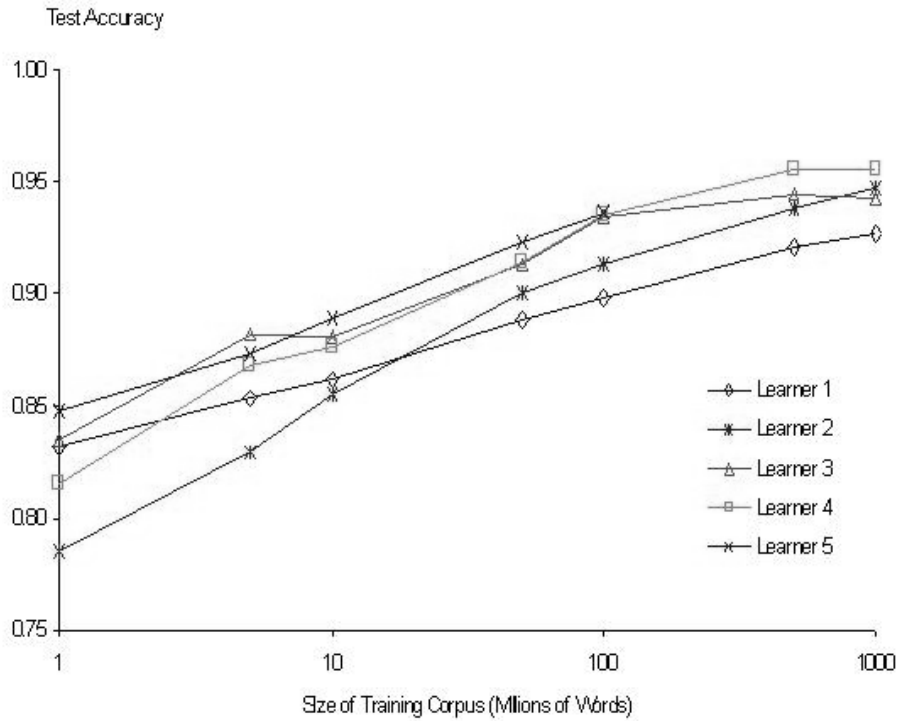
- UPenn – IBM project (NSF) on parser development (1990)
  - Brill transformation based learning
  - Ezra Black headword rules
  - PARSEVAL
  - History-based parsing: Spatter
- ACL 1990: 39 articles, 1 statistical
- ACL 2003: 62 articles, 48 statistical

Data Driven Machine Translation

- French – English statistical translation at IBM (1986)
  - Canadian Hansards data
  - Application of Maximum Entropy Estimation
- DARPA project (1991)
  - Candide (IBM), Pangloss (NMSU/CMU/ISI), Lingstat (Dragon)
- Warren Weaver (Machine Translation of Languages, 1955):
  - *When I look at an article in Russian I say: "This is really written in English but it has been coded in some strange symbols. I will now proceed to decode it."* (letter to Norbert Wiener, March 1947)
  - *…the matter is probably absolutely basic – namely the statistical character of the problem.*

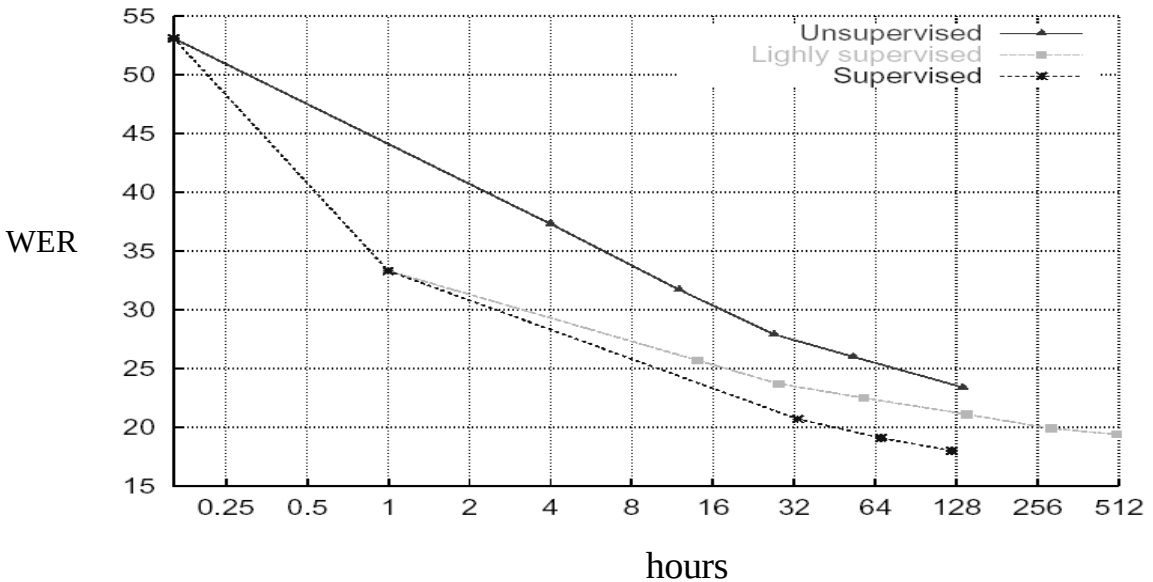
Benefit of Data

Banko & Brill: Mitigating the Paucity-of-Data Problem (HLT 2001)



Benefit of Data

LIMSI: Lamel (2002) – Broadcast News



Supervised:	transcripts
Lightly supervised:	closed captions

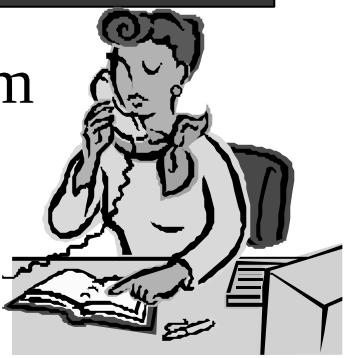
About Data

- *"There is no data like more data"* (Mercer at Arden House, 1985)
- *"More data is more important than better algorithms"* (Brill's opinion)
- M. Lesk:
  - Library of Congress: 20 terabytes
  - Eventual ASCII on Web: 800 terabytes
  - World's writing: 160 terabytes / year
- P. Lyman & H.R. Varian:
  - $5*10^6$ terabytes generated in 2002 in all formats

Great Challenge: Annotating Data

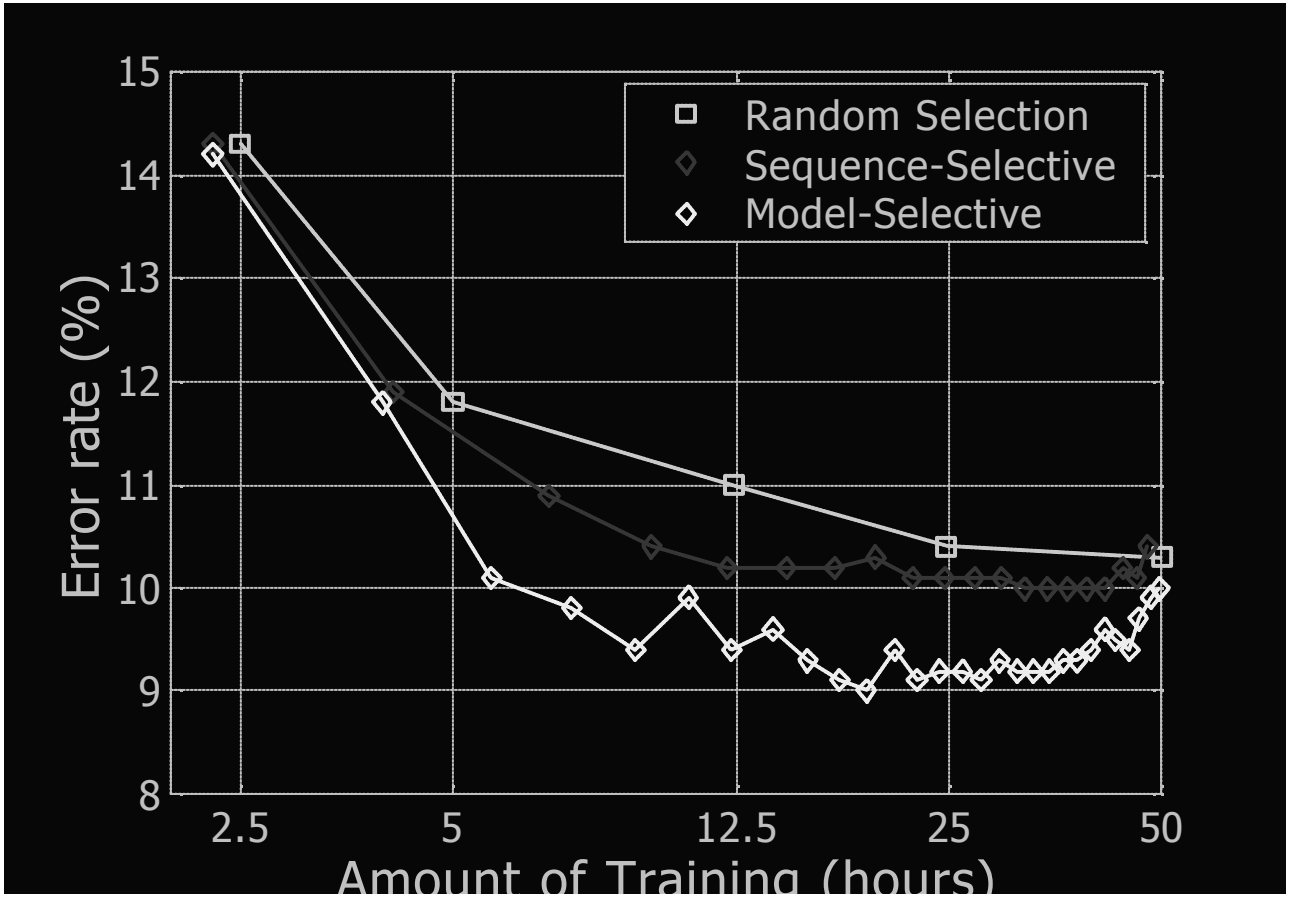
- Produce annotated data with minimal supervision
- Active learning
  - Identify reliable labels
  - Identify best candidates for annotation
- Co-training
- Bootstrap (project) resources from one application to another

Active Training in ASR



OGI Alpha-Digit Corpus, experiment by Terry Kamm
  - English Alphabet and Digits
    - A through Z and 0 through 9
  - Usually six alphanumerics per string
  - Spoken over Telephone
- 3000 speakers, 50 hours training
- Each reads a list of 19 or 29 alphanumeric strings
  - (e.g. “8 h a 8 b h”)
- Baseline of 11.1% WER reported [Hamaker 1998]

Active Training by Terry Kamm



Word Gender Labeling

- Work of Cucerzan and Yarowsky
- Seed process by assigning *natural* gender: `girl, fighter, actress, baby, etc.`
- Iteratively induce gender by considering
  - Context: `sa mère` vs. `his mother`
  - Morphology: in French gender correlated with long suffixes, e.g., -aison, such as in `maison, liaison, raison`
  - Context: In Romanian, only left word context indicates gender, right context does not

How Else Can Linguists Help?

- Help structure systems capable of extracting knowledge under minimal supervision
- MT Example:
  - Diagnostic English sentences to be translated by native informers
    - Designed for the type of language in question
    - Translation elicits basic facts of morphology, system of syntax, inflections, tense structure, etc.
  - New machine learning algorithms extract the rest

Conclusions

- Physicists study physical phenomena

  Linguists study language phenomena
- Engineers learned to take advantage of the insights of physicists
- It is our task to figure out how to make use of the insights of linguists

Point of View of Some Linguists

- Much of what is found in a corpus should be ignored:
  - Like air resistance in physics
- Generative linguists are interested in a set of principles
  - The connection between form and meaning
  - Would rather use informants than corpus
  - Prefer to carry out controlled experiments (like physicists)
- "Relationship of linguists to data is the same as that of physicists to their backyard."

Objective Evaluation Measures

- Needed to be able to optimize systems
- Maximum likelihood in tagging is inadequate (Merialdo, 1984)
- Measures based on trigrams
  - Require human labor
  - BLEU and NIST for MT
  - ROUGE for summarization
- Kulesza – Shieber for MT (SVMachines)
  - Similarity to human-produced