SAND2008-2028C

# *Automated Analysis for Understanding Large Text Corpora*

## Analyzing text to understand individuals and groups

Dan Morrow, PhD
Brett Bader, PhD
Peter Chew, PhD
Ann Speed, PhD

*aespeed@sandia.gov*

NNSA National Nuclear Security Administration

Sandia National Laboratories

# Introduction and Outline

- Analyzing large amounts of text not the only problem in text analysis
  - We can analyze 5MB text in 90 seconds
  - Small amounts of text call validity into question
- Brief primer on text analysis
- Current project

# Text Analysis Primer - 1

- Allows us to develop understanding of people to whom we don't have direct access
- Affords several stand-alone applications
- Two general types
  - "bag of words"
  - natural language processing (NLP)
- This talk focuses on the former

# Text Analysis Primer - 2

- Documents turned into high-dimensional vectors using a universe of possible terms
- Each dimension represents a key word from the term universe.
- For example, the vector

{0 1 3 3 …}

might represent frequencies of the terms

{cat energy political repercussions}

as they appear in a Campaign 2008 news story

- Once we have a vector representation for each text of interest, we can compare the vectors

# Current project

- Leverage research that indicates that peoples' speech becomes more similar as they interact (Cassell & Tversky, 2005; Niederhoffer & Pennebaker, 2002)
- Hypothesis: alliance formation will correlate positively with similarity in ideological expression (i.e., vector similarity)
- Demonstrate the ability to identify these alliances using bag-of-words text analysis techniques

# The Congressional Data

- 83,008 speeches (162 MB) from members of Congress between 1998 & 2000
- Independent special interest group (SIG) ratings
- Experiment:
  - Parametric manipulation of text corpus size, difference in SIG ratings to determine threshold of different text analysis techniques to detect real, but subtle convergences
- Experiment Goals:
  - Classify novel text according to authorship, SIG rating using small amounts of text

# Early Results

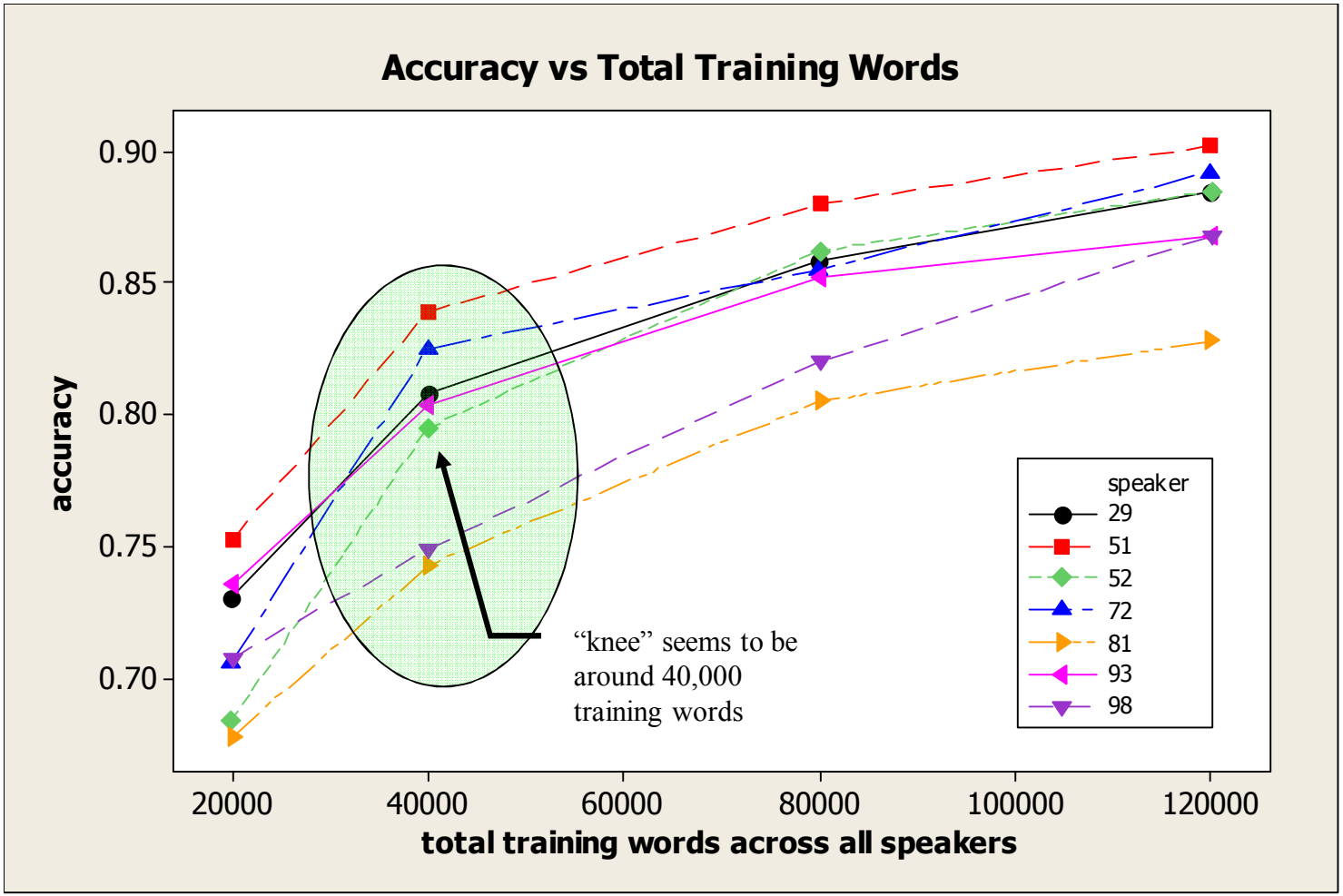


Figure 1: Effect of number of training words on classification accuracy – classification by speaker

# Early Results

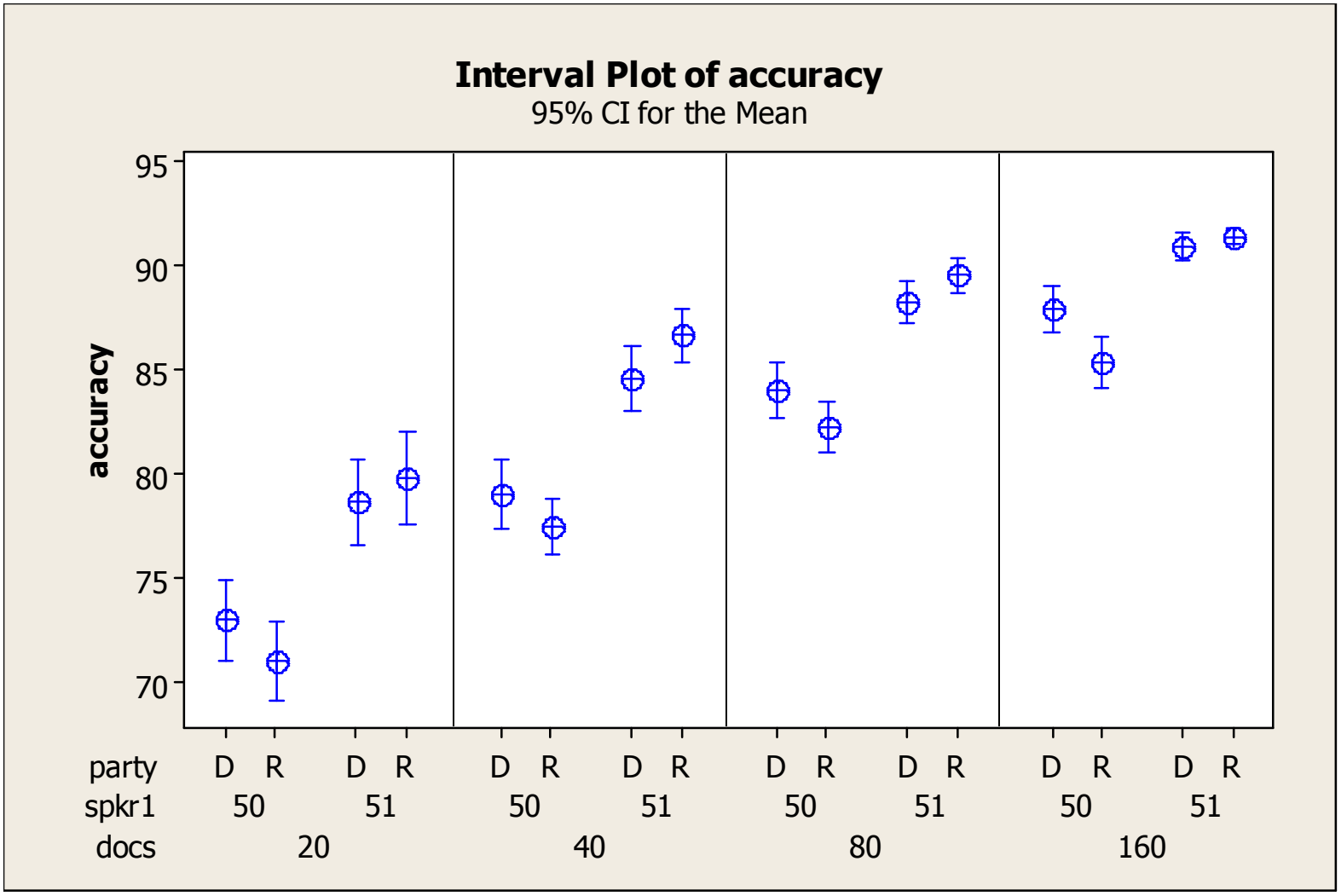


Figure 2: Classification of text along party lines as a function of the number of 500 word documents

# Early Results

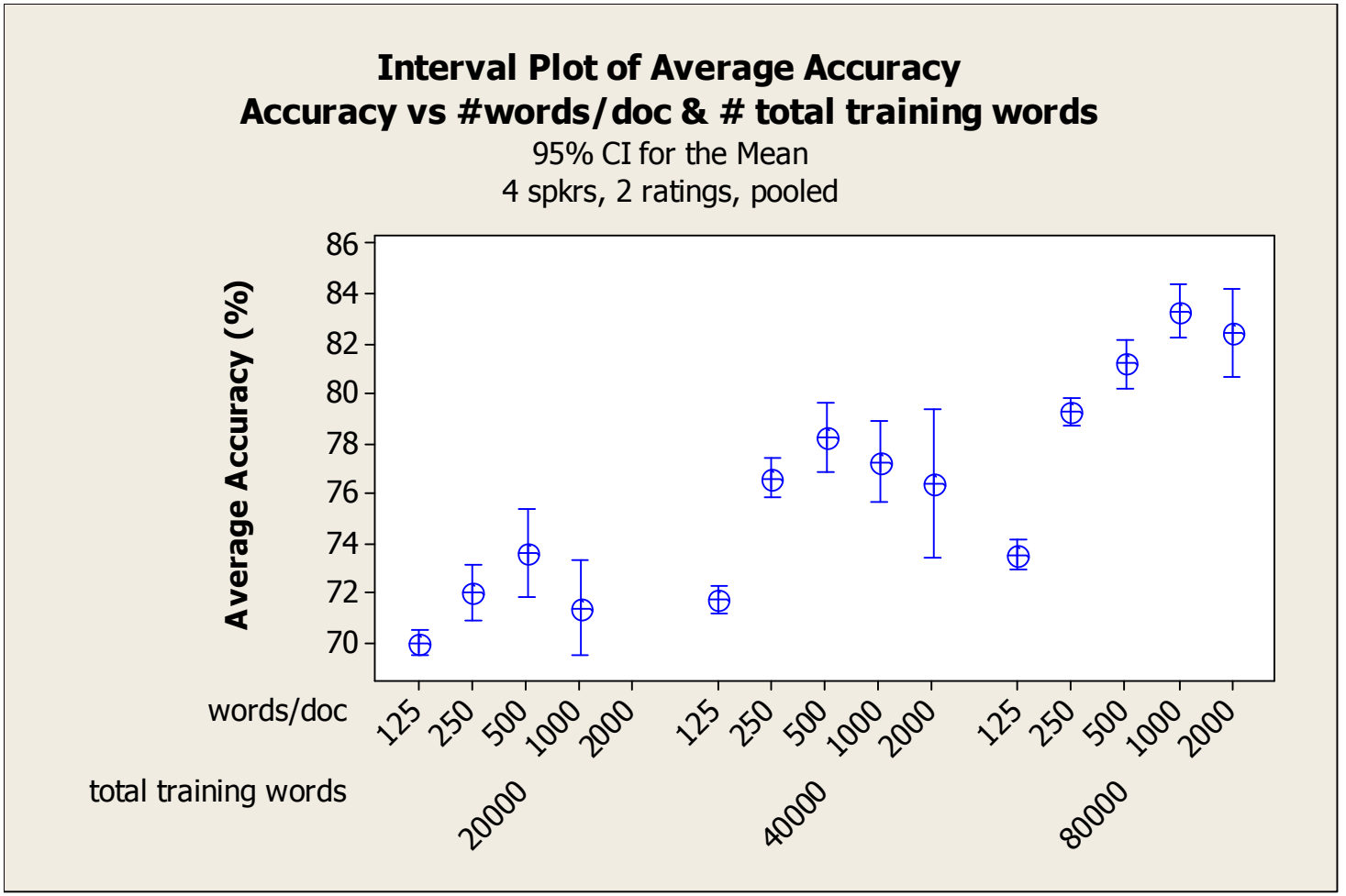


Figure 3: Classification of text according to rating, pooling across speakers

# Discussion

- Even with small amounts of text, classifier accuracy reasonable
- Problems –
  - classifiers are supervised - they require labeled data
  - not clear how to measure ideology using bag-of-words techniques
  - determining ideology from text means texts of interest can't be labeled *a priori*
- Future work
  - unsupervised techniques
  - regression
  - additional work on what "ideology" means for bag-of-words techniques

# References and Acknowledgement


- Cassell, J., & Tversky, D. (2005). The language of online intercultural community formation. Journal of Computer-Mediated Communication, 10, article 2.
- Neiderhoffer, K.G., & Pennebaker, J.W. (2002). Linguistic style matching in social interaction. Journal of Language and Social Psychology, 21, 337-360.



- Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.