SYMBOLIC SYSTEMS 100: Introduction to Cognitive Science

Dan Jurafsky and Daniel Richardson
Stanford University
Spring 2005

May 26, 2005: Empiricism versus Rationalism in Language Learning

IP Notice: Lots of text from the LSA website (Landauer, Dumais, Laham, etc)
And from Peter Chew http://www.sandia.gov/ACG/focusareas/cud/lsa.ppt

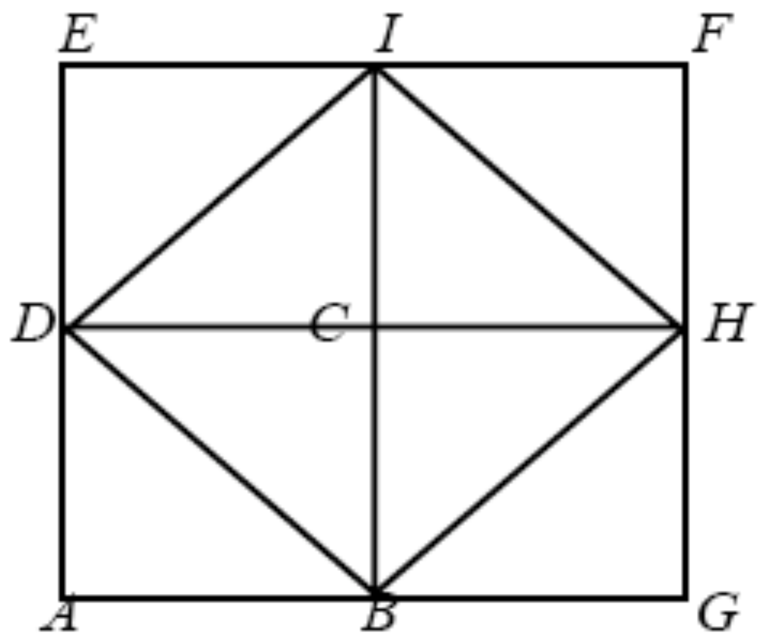
Outline

- Plato's problem
- "Innatist" solution
- Landauer and Dumais alternative
- The LSA Model itself
- Applications of LSA

Plato’s Problem

- **How are humans able to learn to so much given the fragmentary, noisy, and incomplete nature of our interaction with the world?**

BHID is twice as big as ABCD



Plato's answer

- **Knowledge of virtue (or geometry) is innate is some way**
- **His version: "remembered" from a previous life.**

Chomsky's argument from Plato

- The Poverty of the Stimulus argument
  - We don't get enough input from the world to learn the structure of human language
  - Therefore the structure of human language is innate
  - I.e. biologically determined in our genes

Landauer and Dumais response

- Something is certainly innate about language (humans have language and rocks don't)
- But maybe it's not the structure of language.
- Maybe it's something about how we learn?
- We get lots of input!
- So it's possible that much of human knowledge of language is acquired from experience
- But what exactly would such a theory be like?
- How could meaning be induced purely from experience?

Rationalism versus Empiricism

- **Rationalists:**
  - The key factor in human behavior is innate predisposition
- **Empiricists:**
  - The key factor in human behavior is induction from experience

- **(Of course the right answer is that both are true but it's not as shocking to say so)**

LSA: a proof of concept

- L+D make a simplifying assumption:
  - Much of meaning can be learned purely from language input
- L+D propose to build a device for extracting meaning from language data:
  - A 'theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text' (http://lsa.colorado.edu/whatis.html)
- Again, the simplifying assumption:
  - None of its 'knowledge' comes from perceptual information about the physical world
- What is the learning method?

Assumptions underlying LSA

- The meaning of a word is defined distributionally
- Firth (1957) "You shall know a word by the company it keeps"
- Idea: the meaning of a word can be determined by looking at its distribution, I.e. neighboring words
- "distribution" = the range or occurrence of a word among other words

Distributional induction of meaning

- **Two words are similar if they tend to occur in similar contexts**
- **How do we define "similar context"?**
- **Appearing near similar "word neighbors"**

Word similarity and vectors

	stem	pit	debug	subroutine
mango				
apple				
software				
hardware				

How can LSA work if it doesn't know about word order?

- **Word order is not part of model**
- **How can LSA still produce intuitive results?**

Assume:

- vocabulary of 100,000 (10^5) words
- sentences are 20 words long
- word order important only within sentences

$\Rightarrow$ **Contribution (in bits) to passage 'information':**

– From word order:	$\log_2(20!) \approx$	61.0774	15.53%
– From word choice:	$\log_2((10^5)^{20}) \approx$	332.1928	84.47%
– Total	$\log_2(20! \times 10^{100}) \approx$	393.2702	100.00%

Singular Value Decomposition

- **Singular Value Decomposition (SVD) is a form of factor analysis**
- **Any $m \times n$ matrix A can be written using an SVD of the form**

$A = UDV^T$

where:

U is an $m \times n$ matrix (a 'hanger' matrix)
D is an $n \times n$ diagonal matrix (a 'stretcher' matrix)
V^T is an $n \times n$ matrix (an 'aligner' matrix)

(see http://www.uwlax.edu/faculty/will/svd/index.html)

Application of SVD to LSA

- Assemble a large corpus of natural language
- Parse corpus into meaningful passages
- Form matrix with passages as rows and words as columns
- SVD applied to re-represent the words and passages as vectors in a high-dimensional 'semantic space'

SVD: an example (1)

Titles of Technical Memos

- c1: Human machine interface for ABC computer applications
- c2: A survey of user opinion of computer system response time
- c3: The EPS user interface management system
- c4: System and human system engineering testing of EPS
- c5: Relation of user perceived response time to error measurement

- m1: The generation of random, binary, ordered trees
- m2: The intersection graph of paths in trees
- m3: Graph minors IV: Widths of trees and well-quasi-ordering
- m4: Graph minors: A survey

SVD Example: The original matrix

	c1	c2	c3	c4	c5	m1
human	1			1		
interface	1		1			
computer	1	1				
user		1	1		1	
system		1	1	2		
response		1			1	
time		1			1	
EPS			1	1		
survey		1				

SVD: an example (3)

SVD of matrix on previous slide

[illegible]	[illegible]	[illegible]	1	-0.11	-0.34	0.52	0.06	-0.41
[illegible]	[illegible]	[illegible]	5	0.28	0.50	-0.07	0.01	-0.11
[illegible]	[illegible]	[illegible]	9	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.1	[illegible]	[illegible]	[illegible]	0.00	0.01
0.64	-0.17	0.36	0.3	[illegible]	[illegible]	[illegible]	0.03	0.27
0.27	0.11	-0.43	0.0	[illegible]	[illegible]	[illegible]	0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	0.68	0.23

×

2.35					
	1.64				
		1.50			
			1.31		
				0.85	
					0.56

×

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	[illegible]	-0.76	0.45	-0.07

=

1			1					
1		1						
1	1							
	1	1		1				
	1	1	2					
	1			1				
	1			1				
		1	1					
	1							1
					1	1	1	
						1	1	1

2.35

1.64

1.50

1.31

0.85

0.56

SVD: an example (4)

2-D reconstruction of original matrix

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85

$\underline{r}$ (human.user) = .94

$\underline{r}$ (human.minors) = -.83

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

	Before	After
r (human - user) =	-.38	.94
r (human - minors) =	-.28	-.83

Intuitive evaluation of LSA validity
(Landauer 2002: 12)

thing-things	.61
husband-wife	.87
man-husband	.22
chemistry-physics	.65
blackbird-black	.04
go-went	.71

Applications of LSA

- Information retrieval (search engines)
  - Computation of passage similarity
- Text Assessment
  - Automated grading of essays for quality and quantity of content
- Automatic summariation of text
  - Determine best subset of text to portray same meaning
- As a thesaurus
  - Finding synonyms
  - vocabulary tests
  - subject matter tests

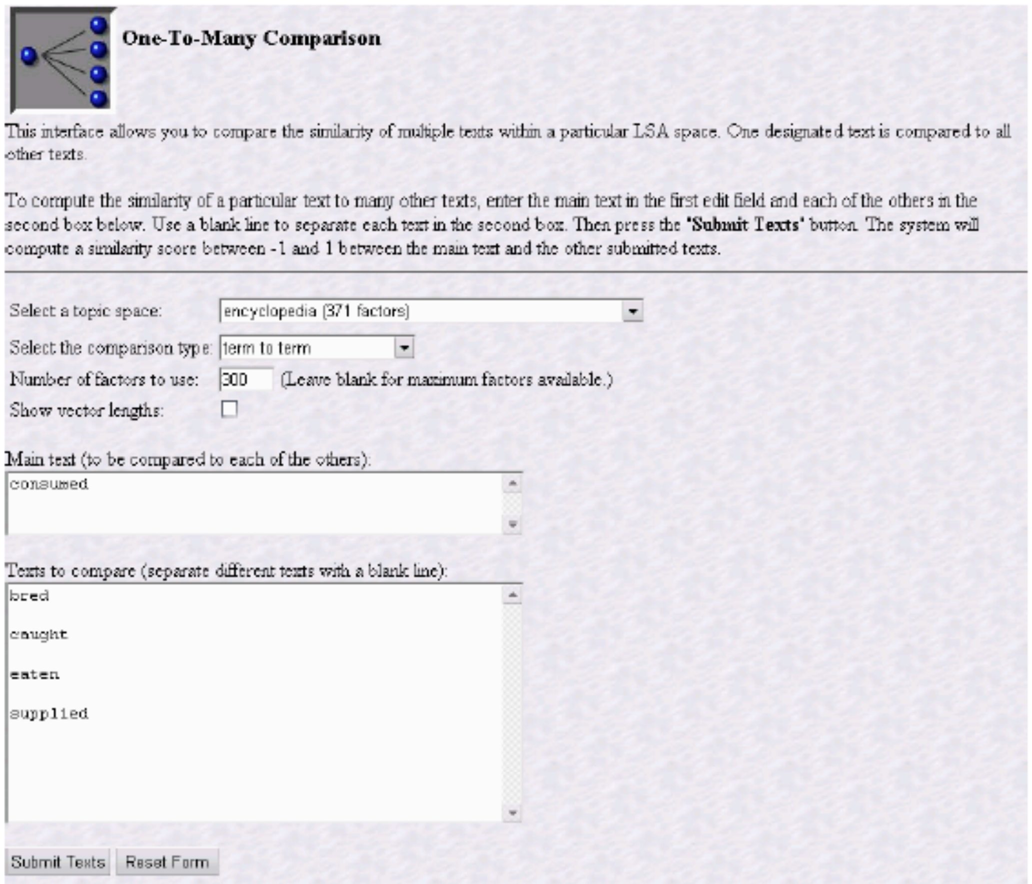
Taking the synonym test

- Test LSA's model of word knowledge
- Can it pass the TOEFL test?
  - 80 item test from ETS (Educational Testing Service)
  - Test of English as a Foreign Language
- Trained on AP newswire and Grollier's Encyclopedia
- 45 million words of text, 60,000 word vocabulary
- Roughly equivalent to what a child would have read by the eighth grade

Taking the synonym test

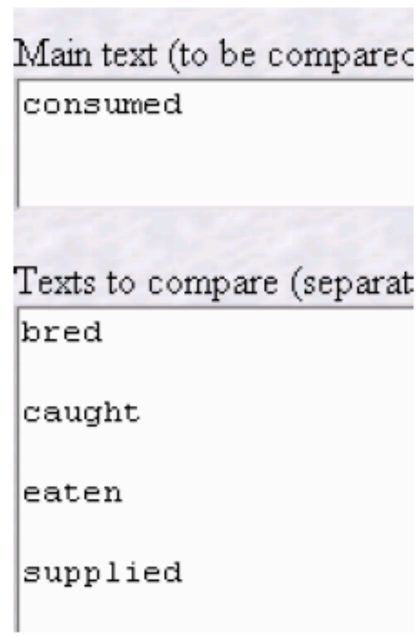
- **Given a test item:**
  - **levied:**
    - **A) imposed**
    - **B) believed**
    - **C) requested**
    - **D) correlated**
- **Use LSA to compute the cosine between the test word and each of the putative synonyms.**
- **Choose the word with the highest cosine.**

Thesaurus (word similarity)



Encyclopedia corpus

300 dimensions



Thesaurus (word similarity)

One-to-Many Comparison Results

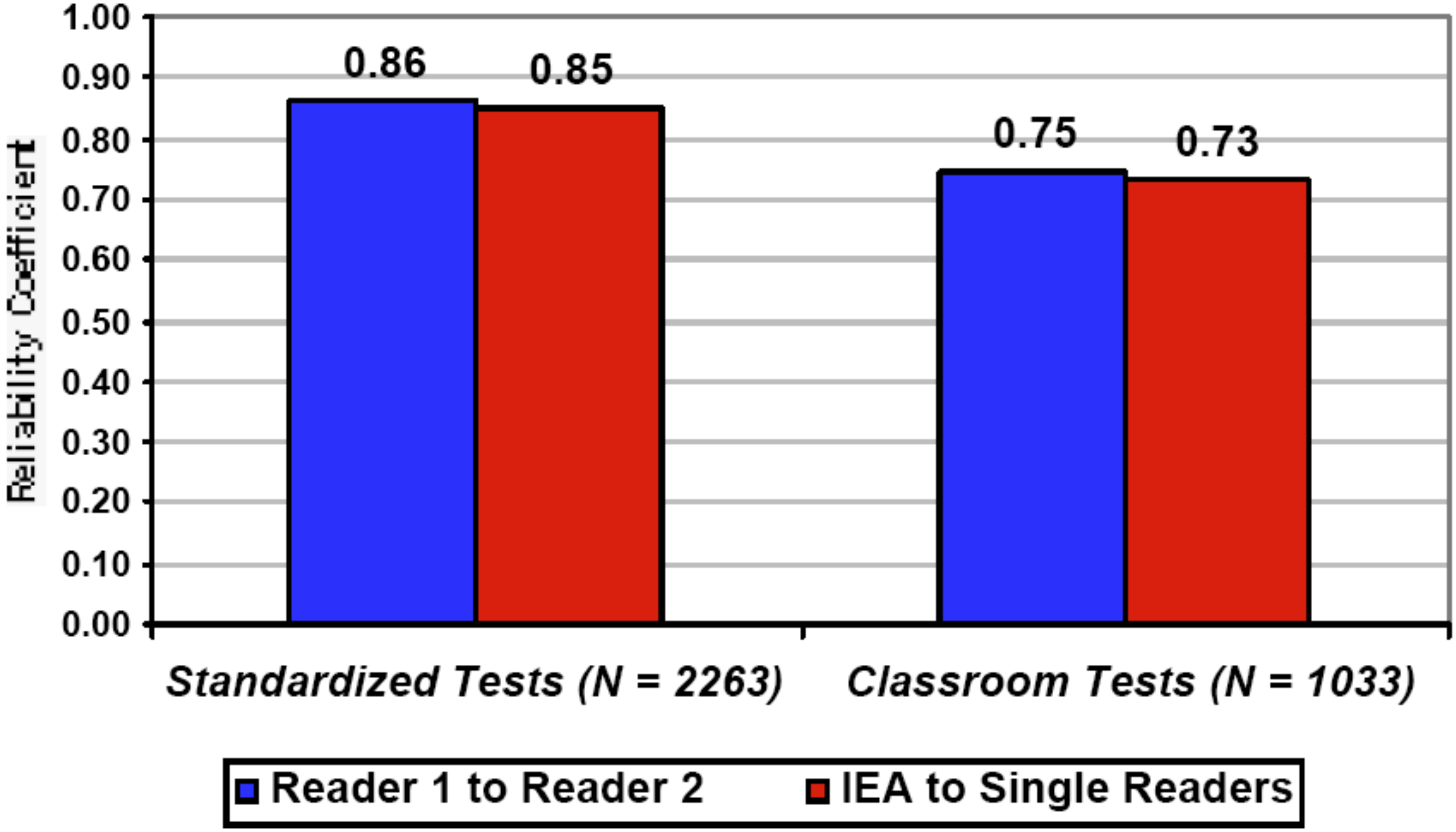
The submitted texts' similarity matrix (in **term to term** space):

Texts	consumed
bred	0.12
caught	0.04
eaten	0.37
supplied	0.17

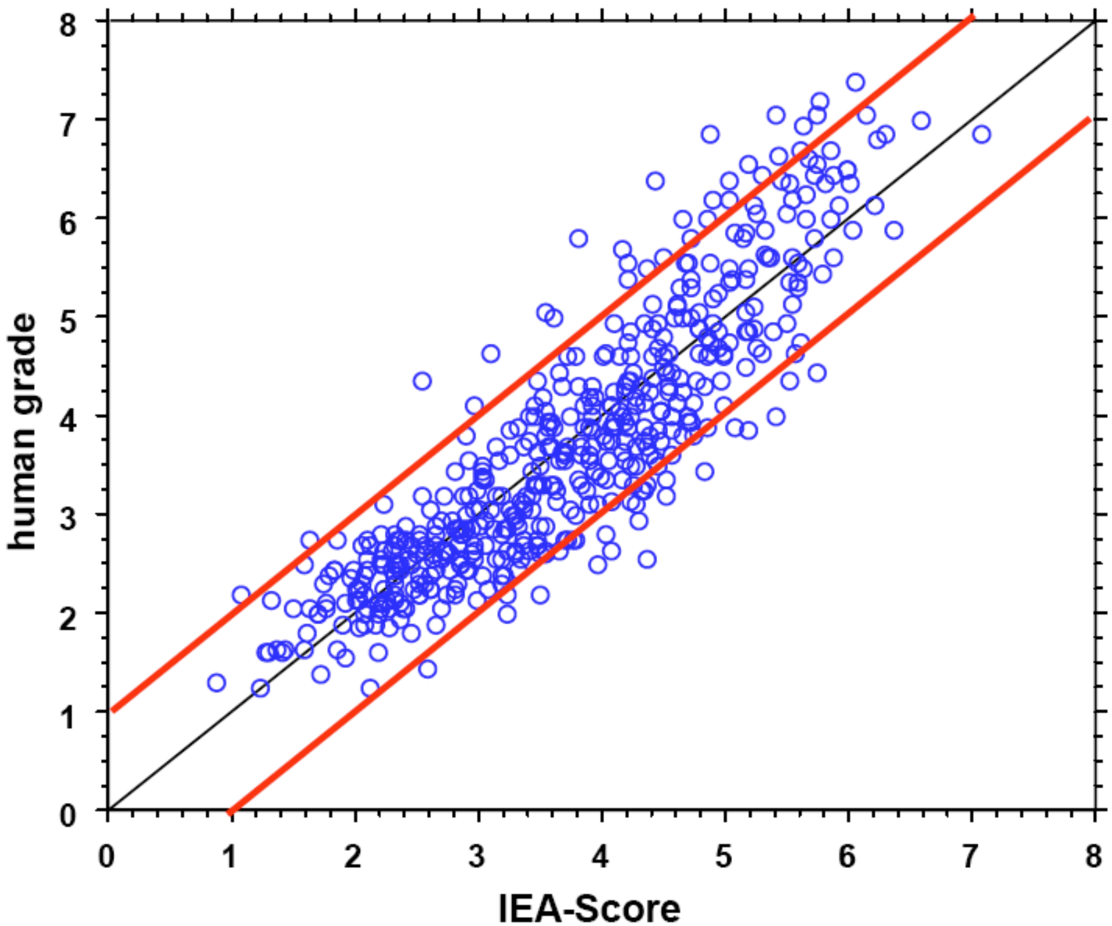
Using LSA to grade essays

- **Middle school, high school, college**
- **Intro psychology, biology, etc**
- **Algorithm:**
  - **Train LSA on texts in the domain**
  - **Pre-grade some "training set" essays**
  - **Use LSA to compare each "test set" essay to the training set graded essays**
  - **Essays which are more similar to "A" essays get an A, more similar to "B" essays get a B, etc**

Inter-rater reliability for standardized and classroom tests



Scattergram for narrative essays



Plagiarism Detection

- LSA-based Intelligent Essay Assessor
- An example of actual plagiarism detected at a university
- 520 student essays total
- For a reader to detect the plagiarism would require 134.940 essay-to-essay comparisons
- In this case, both essays scored by the same reader, and plagiarism went undetected

An example of plagiarism

MAINFRAMES

Mainframes are primarily referred to large computers with rapid, advanced processing capabilities that can execute and perform tasks equivalent to many Personal Computers (PCs) machines networked together. It is characterized with high quantity Random Access Memory (RAM), very large secondary storage devices, and high-speed processors to cater for the needs of the computers under its service.

Consisting of advanced components, mainframes have the capability of running multiple large applications required by many and most enterprises and organizations. This is one of its advantages. Mainframes are also suitable to cater for those applications (programs) or files that are of very high demand by its users (clients). Examples of such organizations and enterprises using mainframes are online shopping websites such as Ebay, Amazon, and computing-giant

MAINFRAMES

Mainframes usually are referred those computers with fast, advanced processing capabilities that could perform by itself tasks that may require a lot of Personal Computers (PC) Machines. Usually mainframes would have lots of RAMs, very large secondary storage devices, and very fast processors to cater for the needs of those computers under its service.

Due to the advanced components mainframes have, these computers have the capability of running multiple large applications required by most enterprises, which is one of its advantage. Mainframes are also suitable to cater for those applications or files that are of very large demand by its users (clients). Examples of these include the large online shopping websites -i.e. : Ebay, Amazon, Microsoft, etc.

Conclusions

- Learning
  - Rationalists: focus on role of innate knowledge (Plato, Chomsky, Pinker)
  - Empiricists: focus on role of experience
- Language learning
  - Lots of information in the input
  - Important: how to generalize beyond just textual information in learning!