Latent Analysis of Syntactic-Semantic Information

Dharmendra Kanejiya

kanejiya@hotmail.com

Department of Electrical Engineering

Indian Institute of Technology Delhi

Outline

- Latent Semantic Analysis
- Syntactically Enhanced LSA
- Applications
    - Intelligent tutoring system ‘AutoTutor’
    - Statistical language modeling

Natural Language Processing

- phonetic - relation between sounds and phonemes
- morphological - components of a word
- syntactic - structure of words, phrases and sentences
- semantic - meaning of and relationships among words
- pragmatic - domain world knowledge

Syntactic Analysis

- Hyrarchical structure in a sentence
- Part-of-speech tags, Phrases, Parse trees
- Parsing
  - Context Free Grammar
  - Probabilistic CFG
  - Dependency grammar

Semantic Analysis

- Compositional semantics
  - Knowledge representation : predicate logic, frames, conceptual dependency
  - Meaning of a sentence from meaning of its parts
  - Uses parse-tree and rules to derive meaning
- Classification based approach
  - Bayesian networks, Speech acts
- Statistical-algebraic approach : LSA
  - Words and documents as vectors

Latent Semantic Analysis

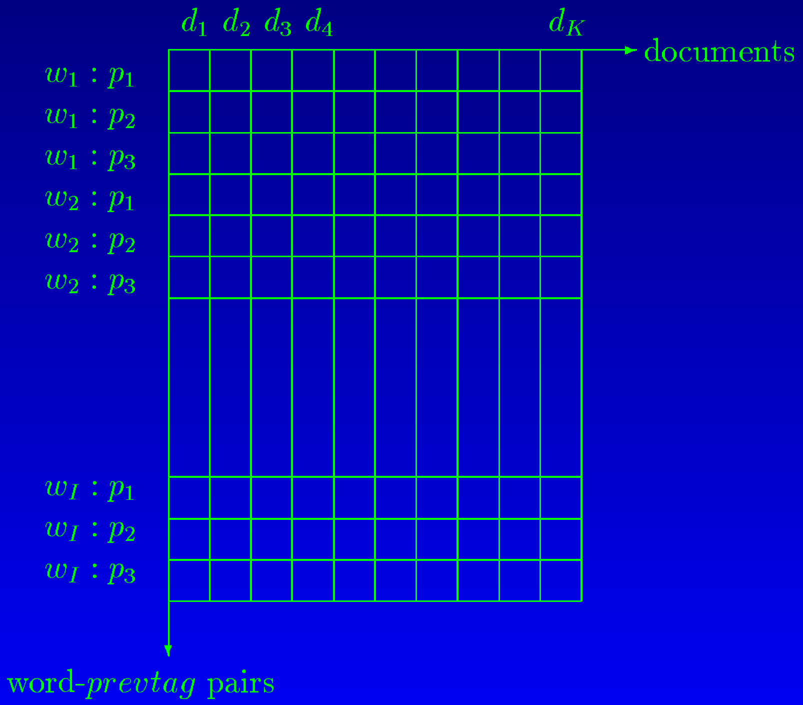
- Extracts salient semantic relationships between words and documents in a corpus
- Generate a word-document co-occurrence statistics matrix $\mathbf{W}$
- Apply entropic scaling and document length normalization
- Perform SVD : $\mathbf{W} \approx \mathbf{USV}^T$
- Results in a *latent semantic* space, in which words and documents are projected as vectors
- Applied to information retrieval, natural language understanding, cognitive modeling, statistical language modeling

Syntactically Enhanced LSA

- LSA is a ‘bag-of-words’ approach
- Syntax is important for resolving semantic ambiguity
- For accurate knowledge representation, capture the semantic behaviour of a word in each possible syntactic neighborhood
- This gives large-span syntactic-semantic information
- Example : augment words with the POS tag of previous word or the phrase level information

Syntactically Enhanced LSA

- word-*prevtag*-doc structure
- $x_{i\_j,k} = (1 - \varepsilon_{i\_j})\frac{c_{i\_j,k}}{n_k}$



d_1 d_2 d_3 d_4 d_K

documents

$w_1 : p_1$

$w_1 : p_2$

$w_1 : p_3$

$w_2 : p_1$

$w_2 : p_2$

$w_2 : p_3$

$w_I : p_1$

$w_I : p_2$

$w_I : p_3$

word-*prevtag* pairs

SELSA: Projection & Similarity

- Projection of a document $w_{i_1}, w_{i_2}, \ldots, w_{i_n}$

$$\bar{\mathbf{v}}_{sel} = \tilde{\mathbf{v}}\mathbf{S} = \mathbf{d}^{\mathbf{T}}\mathbf{U} = \frac{1}{n}\sum_{p=1}^{n}(1 - \varepsilon_{i_p\_j_{p-1}})\mathbf{u}_{i_p\_j_{p-1}}$$

- Document-document similarity : cosine measure

$$Sim(d_1, d_2) = \frac{\bar{\mathbf{v}}_{1sel}^{T}\bar{\mathbf{v}}_{2sel}}{\| \bar{\mathbf{v}}_{1sel} \| \| \bar{\mathbf{v}}_{2sel} \|}$$

- Word-*prevtag*-document similarity

$$K(w_q, p_{q-1}, H_{q-1}) = \frac{\mathbf{u}_{i_q\_j_{q-1}}\mathbf{S}\tilde{\mathbf{v}}_{q-1}^{T}}{\| \mathbf{u}_{i_q\_j_{q-1}}\mathbf{S}^{\frac{1}{2}} \| \| \tilde{\mathbf{v}}_{q-1}\mathbf{S}^{\frac{1}{2}} \|}$$

Application 1

Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA

AutoTutor

AUTOTUTOR : When you turn on the computer, how is the operating system first activated and loaded into RAM?

STUDENT : When the computer is turned on, the file is automatically booted up.

AUTOTUTOR : Anything else ?

STUDENT : The file contains the operating system.

AUTOTUTOR : What else ?

STUDENT : The operating system is then read into RAM and activated

AUTOTUTOR : Right. The CPU executes instructions in ROM, which help the operating system to start.

STUDENT : After being read to ROM the OS is activated.

AUTOTUTOR : Right. Let’s review what we have just gone over. The operating system must first be stored on the hard disk. When you turn on the computer, the CPU executes instructions stored in ROM. These instructions help boot the operating system. The operating system is then loaded from the hard disk into RAM.

AutoTutor Modules

- Curriculum script
- Language extraction
- Speech act classification
- (Syntactically Enhanced) Latent semantic analysis
- Topic selection
- Dialog move generator
- Talking head with gestures

SELSA and LSA

- Training Corpus
  - Two textbooks on computer literacy
  - Ten papers each from three topics - hardware, operating system and networking
  - Curriculum scripts
- Document unit: a paragraph, a sentence
- Vocabulary size $\sim$ 10000 words
- POS tag-set size = 12
- SVD dimensions : 200 to 400

Evaluation of SELSA and LSA

- 8 questions per topic of HW, OS, NET
- 8 student-answer per question (192 total)
- 4 human raters evaluated these answers based on compatibility score
- 3 were doctorate students, one assistant professor in computer related areas
- SELSA and LSA also evaluated these answers

Compatibility Score

- Each student-answer broken into sentences
- Human score

$$h = \frac{\#\text{sentences that match with any good answer}}{\#\text{sentences in the student-answer}}$$

- SELSA or LSA score

$$l = \frac{\left(\begin{array}{c}\#\text{sentences whose cosine match with any}\\ \text{good answer exceeds a threshold (e.g. 0.5)}\end{array}\right)}{\#\text{sentences in the student-answer}}$$

Performance Measures

- Correlation

$$COR = \frac{(\mathbf{h} - m_h)^T(\mathbf{l} - m_l)}{\| \mathbf{h} - m_h \|\| \mathbf{l} - m_l \|}$$

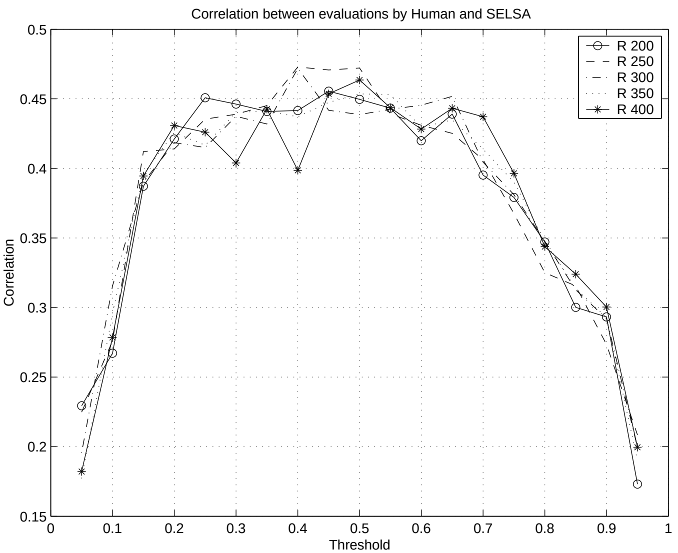
- Mean absolute difference

$$MAD = \frac{1}{192}\sum_{i=1}^{192} |h_i - l_i|$$

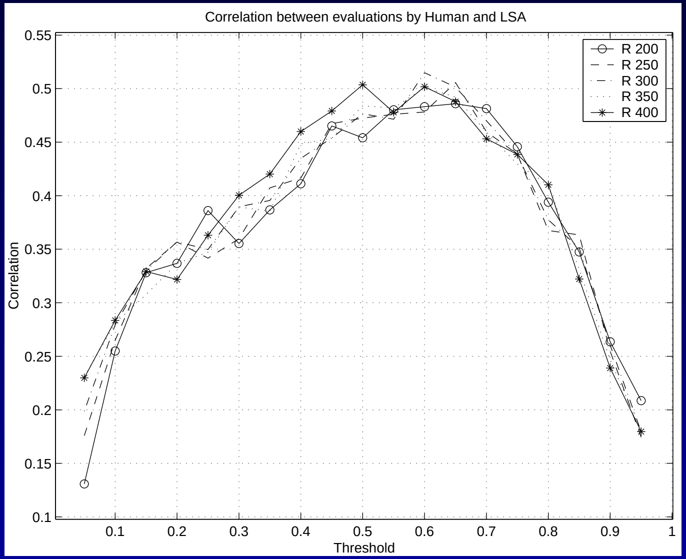
- Correct vs False

$$\begin{aligned} \#CORRECT &= \sum_{i=1}^{192} \mathbf{I}(|h_i - l_i| < 0.05) \\ \#FALSE &= \sum_{i=1}^{192} \mathbf{I}(|h_i - l_i| > 0.95) \end{aligned}$$

Correlation Measure



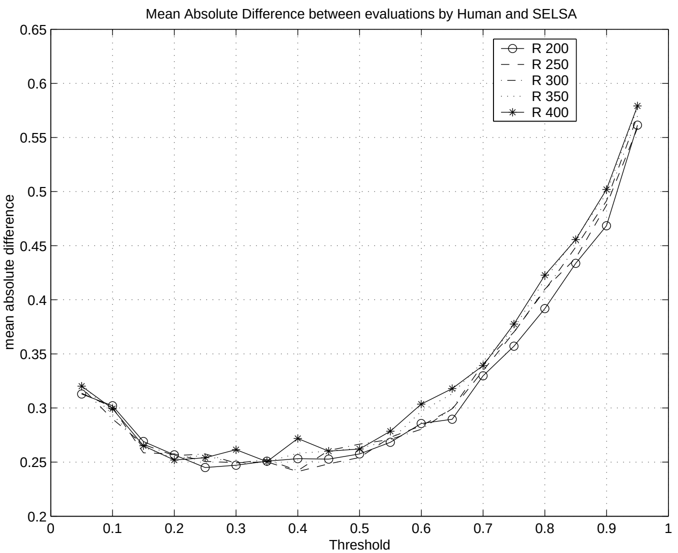
(a) Human-SELSA



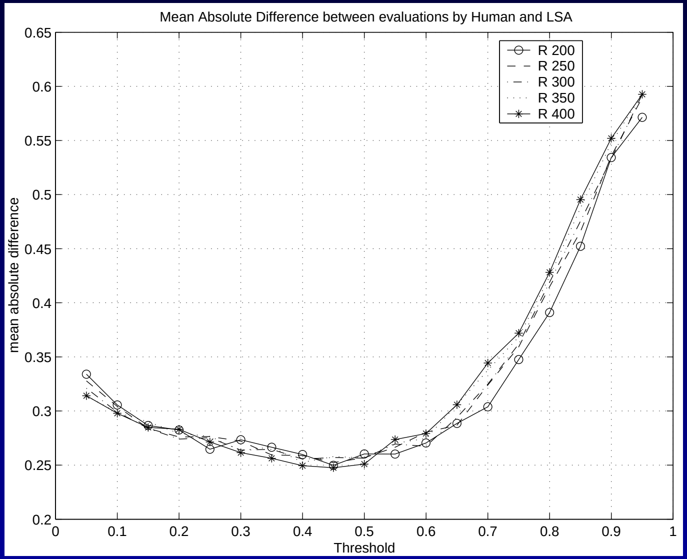
(b) Human-LSA

	Correlation	SVD dim	Threshold	10% TW
Best human-SELSA	0.47	250	0.40	0.42
Best human-LSA	0.51	250	0.65	0.29
Inter-human	0.59			

Mean Absolute Difference



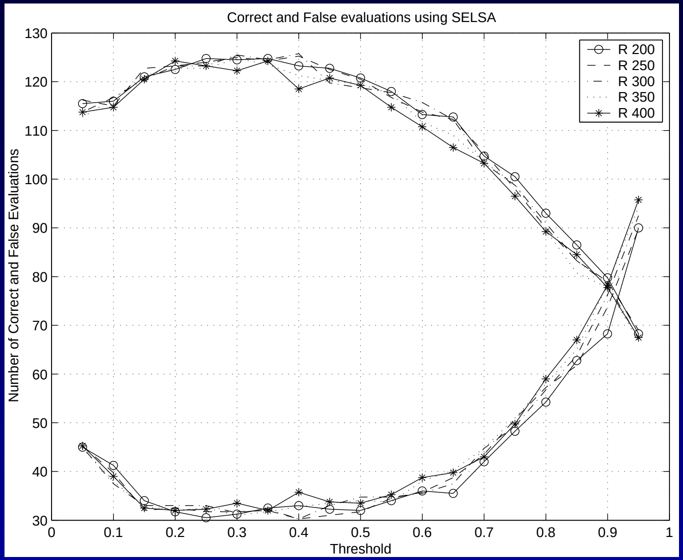
(c) Human-SELSA



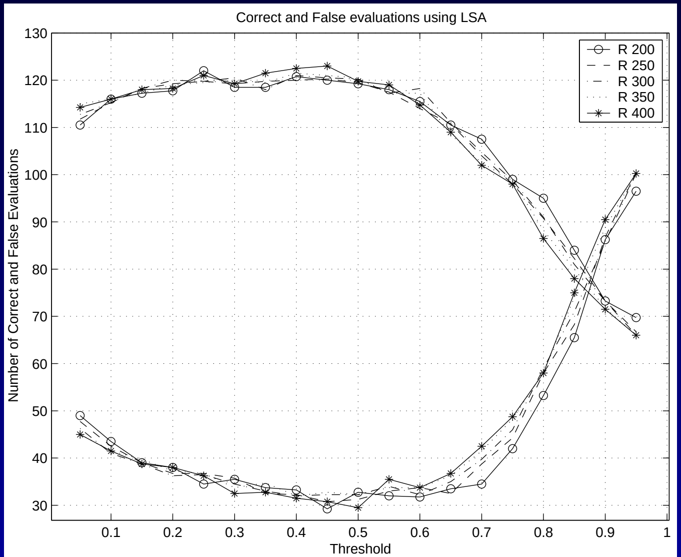
(d) Human-LSA

	minimum MAD	SVD dim
Best human-SELSA	0.2412	250
Best human-LSA	0.2475	400
Inter-human	0.2050	

Correct vs False



(e) Human-SELSA



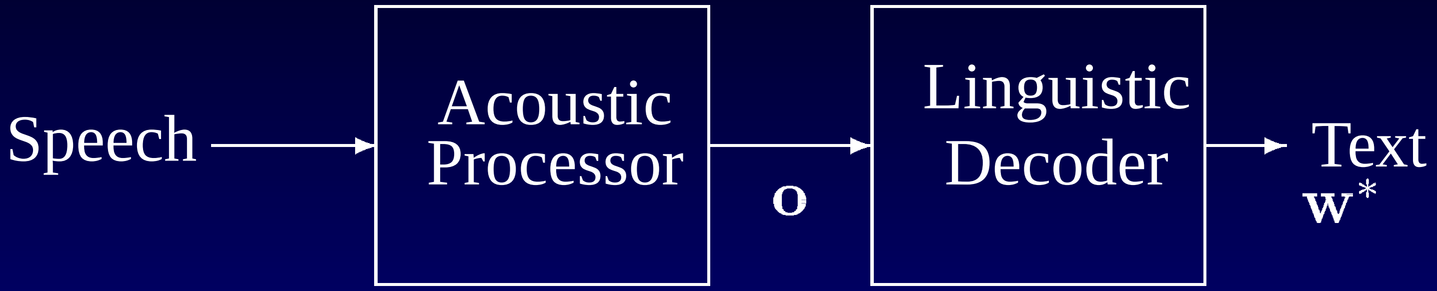
(f) Human-LSA

	max #Correct	min # False	SVD dim
Best human-SELSA	126	30	300
Best human-LSA	123	30	400
Inter-human	132	23	

Application 2

Statistical Language Modeling using Syntactically Enhanced LSA

Speech Recognition



$$\mathbf{w}^* = \arg\max_{\mathbf{w}} P(\mathbf{w}|\mathbf{o}) = \arg\max_{\mathbf{w}} P(\mathbf{o}|\mathbf{w})P(\mathbf{w})$$

- $P(\mathbf{o}|\mathbf{w})$ - *acoustic model*
- $P(\mathbf{w})$ - *language model*

Language Modeling

- Words in a language are connected through syntactic, semantic and pragmatic dependencies
- Language models capture these dependencies using deterministic or statistical methods
- Applications :
  - speech recognition
  - machine translation
  - optical character recognition
  - spelling correction
  - spoken dialog systems
  - bioinformatics

Statistical Language Modeling

- Probabilitiy of a sequence of l words

$$P(\mathbf{w}) = \prod_{q=1}^{l} P(w_q | w_{q-1}, w_{q-2}, ..., w_1)$$

- Estimation of $P(w_q | w_{q-1}, w_{q-2}, ..., w_1)$
- Equivalence classification of history $H_{q-1} = w_{q-1}, w_{q-2}, ..., w_1$
- Examples : n-gram LM, structured LM, LSA based LM

N-gram LM

- $P(w_q|H_{q-1}) \approx P(w_q|w_{q-1}, w_{q-2}, \ldots, w_{q-n+1})$
- *de facto* LM
- Limitations :
  - Predictive power *vs* unreliable estimate
  - n typically 2 (*bi-gram*) or 3 (*tri-gram*)
  - unable to capture large-span relationships in a language
- Solution :
  - *syntactic large-span* : structured LM
  - *semantic large-span* : LSA based LM
  - *syntactic-semantic large-span* : SELSA

SLM using LSA

- Prediction of a word based on semantic ‘closeness’ to the history
- Projection of history on LS space as a vector $\tilde{\mathbf{v}}_{\mathbf{q-1}}$
- Semantic similarity between a word and the history

$$K(w_q, H_{q-1}) = \frac{\mathbf{u_q S \tilde{v}^T_{q-1}}}{\| \mathbf{u_q S^{\frac{1}{2}}} \| \| \mathbf{\tilde{v}_{q-1} S^{\frac{1}{2}}} \|}$$

$$P^{(lsa)}(w_q | H_{q-1}) = f(K(w_i, H_{q-1}); \forall w_i \in \mathcal{V})$$

- Integrate LSA based *large-span semantic* probability with *local-span* n-gram probability

SLM using SELSA

- Let $p_j \in \mathcal{P}$ be j^{th} POS tag and p_{q-1} be POS tag of w_{q-1}

$$\begin{aligned}
P^{(sel)}(w_q|H_{q-1}) & \\
&= \sum_{p_j \in \mathcal{P}} P(w_q, p_j|H_{q-1}) \\
&= \sum_{p_j \in \mathcal{P}} P(w_q|p_j, H_{q-1}) P(p_j|H_{q-1}) \\
&= \sum_{p_j \in \mathcal{P}} P(w_q|p_j, H_{q-1}) \mathbf{1}(p_j = p_{q-1}) \\
&= P(w_q|p_{q-1}, H_{q-1})
\end{aligned}$$

SLM using SELSA

- Define a syntactic-semantic ‘closeness’ measure

$$K(w_q, p_{q-1}, H_{q-1}) = \frac{\mathbf{u}_{i_q\_j_{q-1}} \mathbf{S} \tilde{\mathbf{v}}^T_{q-1}}{\| \mathbf{u}_{i_q\_j_{q-1}} \mathbf{S}^{\frac{1}{2}} \| \| \tilde{\mathbf{v}}_{q-1} \mathbf{S}^{\frac{1}{2}} \|}$$

- SELSA based probability

$$\begin{aligned} & P^{(sel)}(w_q | p_{q-1}, H_{q-1}) \\ = \; & f(K(w_i, p_{q-1}, H_{q-1}); \forall w_i \in \mathcal{V}) \end{aligned}$$

- It predicts a word using large-span syntactic-semantic dependencies

SELSA + n-gram LM

- Integrating *large-span syntactic-semantic* information with *local-span n-gram* information

$$P^{(selsa+ng)}(w_q|H_{q-1})$$

$$= \frac{\left[P^{(sel)}(w_q|H_{q-1})\right]^{\xi_{i_q}} \left[P^{(ng)}(w_q|w_{q-1},\ldots,w_{q-n+1})\right]^{1-\xi_{i_q}}}{\sum_{w_i\in\mathcal{V}} \left[P^{(sel)}(w_i|H_{q-1})\right]^{\xi_i} \left[P^{(ng)}(w_i|w_{q-1},\ldots,w_{q-n+1})\right]^{1-\xi_i}}$$

where, $\xi_{i_q} = \frac{1-\varepsilon_{i_q\_j_{q-1}}}{2}$ and $\xi_i = \frac{1-\varepsilon_{i\_j_{q-1}}}{2}$

Experimental Setup

- Corpus : Wall Street Journal year 1989 (4.75 Million words)
- Training : 4.45 M words
- Test : 0.3 M words
- Vocabulary size : 10775 most frequent words
- POS tagset size : 12
- SVD dimensions : 125, 200, 300
- Exponential forgetting factor applied to the ‘history of words’ document

Results

- *Perplexity* on the test corpus indicates average vocabulary for speech recognition task

$$PPL(\mathcal{M}) = \exp\left(-\frac{1}{N} \sum_{q=1}^{N} \log P^{(\mathcal{M})}(w_q | H_{q-1}) \right)$$

- Lower the perplexity better the language model

	LSA+Bi-gram			SELSA+Bi-gram		
R	Aprx Err(%)	PPL	% reduction	Aprx Err(%)	PPL	% reduction
125	86.85	106.16	36.33	91.00	113.13	32.15
200	83.01	104.71	37.20	88.31	111.05	33.40
300	78.74	103.98	37.64	85.28	109.57	34.28
250				86.72	110.18	33.92
400				82.59	108.80	34.74

Table 1: Perplexity reduction relative to Bi-gram at different SVD dimensions

Probability assignments

POS tag	Word	Bi-gram	LSA + Bi-gram	SELSA + Bi-gram
DT	The	0.147639		
NN	post	0.000323	0.000971	**0.002089**
IN	of	0.172703		
JJ	chief	0.000409	**0.001865**	0.001636
NN	operating	0.085153	0.115940	**0.144197**
NN	officer	0.169278	**0.377109**	0.329110
AUX	has	0.003224		
AUX	been	0.167456		
JJ	vacant	0.002846	0.013525	**0.016029**
IN	for	0.195119		
JJR	more	0.007073		
IN	than	0.293022		
DT	a	0.060884		
NN	decade	0.001587	0.001498	**0.001689**
,	,	0.218072		
DT	a	0.042599		
NNP	Ball	0.000052	0.000058	**0.000070**
NN	spokesman	0.000136	**0.000458**	0.000367
VBD	said	0.386456	**0.723014**	0.660258
.	.	0.171475		

Phrase-structure based SELSA

- Mainly three types of phrase structures : NP, VP, others
- Consider a word alongwith the phrase it belongs
- It reduces the sparseness of *prevtag*-based SELSA
- Training corpus : 40M, Vocabulary size : 20000
- Baseline perplexity : Bi-gram 162.88 , Tri-gram 103.12

	LSA+Bi-gram			SELSA+Bi-gram		
R	Aprx Err(%)	PPL	% reduction	Aprx Err(%)	PPL	% reduction
125	90.22	103.26	36.62	90.47	97.00	40.45
200	87.75	102.14	37.29	88.07	95.59	41.31

	LSA+Tri-gram			SELSA+Tri-gram		
R	Aprx Err(%)	PPL	% reduction	Aprx Err(%)	PPL	% reduction
125	90.22	68.42	33.65	90.47	64.78	37.18

Table 2: Perplexity reduction relative to Bi-gram and Tri-gram

Speech Recognition

- As a second-pass language model
- N-best/Lattice rescoring method
- First-pass can consist of a simple language model (e.g. bi-gram or tri-gram) and second-pass can benefit from complex syntactic-semantic analysis
- Experiment underway on WSJ lattices from CLSP, JHU
- A reduction in perplexity generally translates to a reduction in *word error rate*

Conclusions

- SELSA generalizes LSA by including various levels of syntactic information
- AutoTutor Task
  - SELSA more robust, discriminatory and correct than LSA but having less correlation with human
- Statistical Language Modeling
  - LSA and SELSA both reduce bi-gram perplexity significantly
  - LSA better than *prevtag*-based SELSA
  - Phrase-structure based SELSA better than LSA : it captures the coarser syntactic information without increasing sparseness much

Thank You !