


Data-Driven Semantic Language Modeling

Jerome R. Bellegarda
Spoken Language Group
Apple Computer


Outline

- Motivation
- Latent Semantic Analysis (LSA)
- Application to Spoken Language
- Semantic Language Modeling
- Perspectives



Outline

- Motivation
 - local / global language constraints
 - inherent n-gram limitations
 - beyond n-grams
- Latent Semantic Analysis (LSA)
- Application to Spoken Language
- Semantic Language Modeling
- Perspectives




Language

- Local Constraints
 - short-span relationships
 - syntactic constructs (e.g., "to get rid of")
 - phrasal entries (e.g., "New York City")
- Global Constraints
 - large-span effects
 - gender/number/tense agreements
 - underlying semantic fabric




n-Gram Modeling

- Trade-Off
 - effective in capturing short-span effects
 - parameter estimation less reliable as $n \uparrow$
 - local horizon resulting from low value of n
- Example
 - stocks fell sharply in afternoon trading as a result of this announcement
 - stocks, as a result of this announcement, sharply fell in afternoon trading



Possible Directions

- Robust Estimation
 - train on more/larger corpora
 - more sophisticated smoothing
- Information Aggregation
 - class-based models
- Span Extension
 - structured language models
 - word triggers, semantic analysis



Structured LMs

- **Syntactic Information**
 - sentence parse sub-trees define **headwords**
 - n-grams on headwords rather than words
 - \implies equivalence classes on n-gram history
 - caveat: reliance on parser!
- **Example**
 - **stocks** = NP headword; **fell** = VP headword
 - \implies structured bigram OK in both cases

Word Triggers

- **Semantic Information**
 - (**stock**, **fell**) = trigger pair
 - seeing **stocks** boosts probability of **fell**
- **Drawbacks**
 - trigger pair selection (combinatorial issue)
 - low frequency triggers typically eliminated
 - \implies break transitivity property
 - proven successful only for self-triggers

Outline

- Motivation
- **Latent Semantic Analysis (LSA)**
 - word-document matrix
 - dimensionality reduction
 - semantic vector space
- Application to Spoken Language
- Semantic Language Modeling
- Perspectives

Latent Semantic Analysis

- **Overview**
 - originally used for information retrieval
 - dual concepts of **word** and **document**
 - document implies semantic consistency
- **Trigger Extension**
 - word "trigger pairs" appear in similar docs
 - doc "trigger pairs" contain similar words
 - \implies analyze word-document co-occurrences

Co-Occurrence Matrix

- w_{ij} = weighted count (w_i, d_j)

Expression for w_{ij}

- c_{ij} = count of w_i in document d_j
- $w_{ij} = G_i L_{ij} c_{ij}$
- L_{ij} : local weight
 - importance of w_i in current document d_j
- G_i : global weight
 - overall importance of w_i in entire corpus

Weighting

- **Definitions**
 - n_j : number of words in document d_j
 - e_i : normalized entropy of w_i in corpus
- **Weights**
 - $L_{ij} = 1 / n_j$: length normalization
 - $G_i = 1 - e_i$: measure of indexing power

Word-Document Matrix

- **Two Representations**
 - words in space of dimension N
 - documents in space of dimension M

Dimensionality Reduction

- **SVD Analysis**
 - R = number of singular values

$$W_{(M \times N)} = U_{(M \times R)} S_{(R \times R)} V^T_{(R \times N)}$$

words
documents

Benefits

- **Vector Representation**
 - single space for words and documents
 - closeness = semantic similarity
 - parsimonious dimension ($100 < R < 300$)
- **Consequences**
 - discrete entities \rightarrow continuous space (S)
 - amenable to usual clustering techniques
 - \rightarrow uncover high-level semantic regions

Word Clustering

- **Closeness Measure**
 - word w_i mapped to $\bar{u}_i = u_i S$
 - metric:

$$K(\bar{u}_i, \bar{u}_m) = \frac{u_i S^2 u_m^T}{\|u_i S\| \|u_m S\|}$$
- **Outcome**
 - a set of word clusters $\{C_k\}$ in S , $1 \leq k \leq K$
 - see: Bellegarda, Trans. SAP, Sept. 98

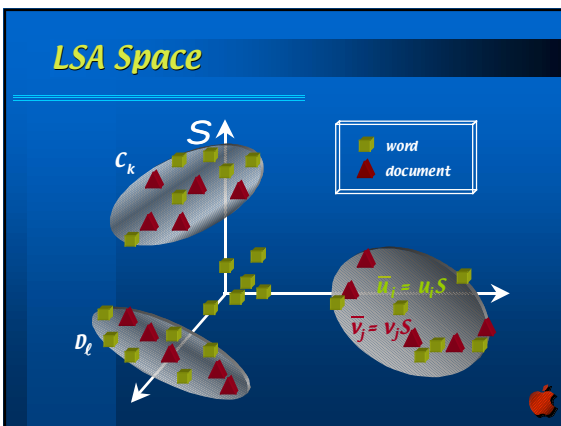
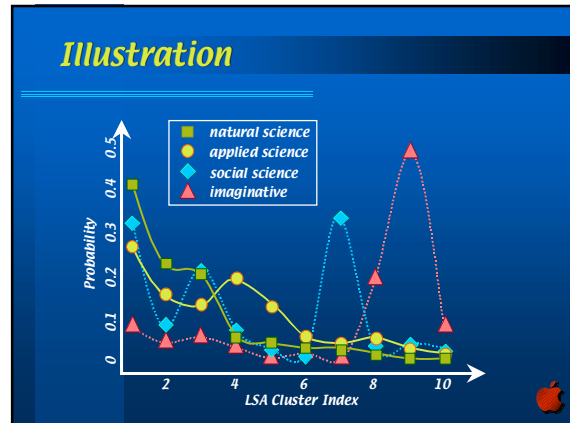
Two Examples

- Andy, antique, antiques, art, artist, artist's, artists, artworks, auctioneers, Christie's, collector, **drawings**, gallery, Gogh, fetched, **hysteria**, masterpiece, museums, painter, painting, paintings, Picasso, Pollock, reproduction, Sotheby's, van, Vincent, Warhol
- appeal, appeals, attorney, attorney's, counts, court, court's, courts, condemned, convictions, criminal, decision, defend, defendant, dismisses, dismissed, hearing, **here**, indicted, indictment, indictments, judge, judicial, judiciary, jury, juries, lawsuit, leniency, overturned, plaintiffs, prosecute, prosecution, prosecutions, prosecutors, **ruled**, ruling, sentenced, sentencing, suing, suit, suits, witness

Document Clustering

- Closeness Measure**
 - document d_j mapped to $\bar{v}_j = v_j S$
 - metric:

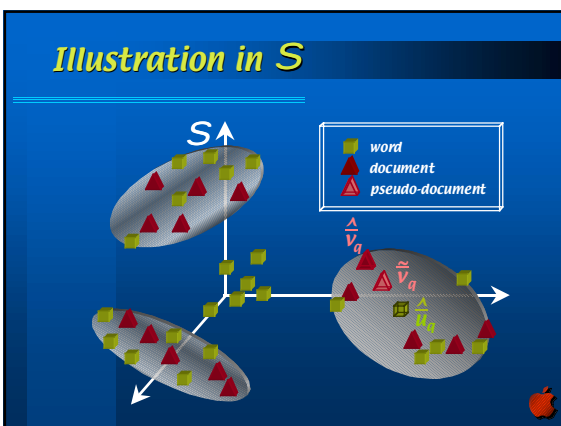
$$K(\bar{v}_j, \bar{v}_m) = \frac{v_j S^2 v_m^T}{\|v_j S\| \|v_m S\|}$$
- Outcome**
 - a set of document clusters $\{D_\ell\}$ in S , $1 \leq \ell \leq L$
 - see: Gotoh & Renals, EuroSpeech, Sept. 97



Generalization

- New Document**
 - at time q : construct vector for $\{w_1 \dots w_q\}$
 - pseudo-document \tilde{d}_q
 - need associated vector in S : \tilde{v}_q

The diagram shows a word vector w_i being multiplied by a matrix U to produce a pseudo-document vector \tilde{d}_q . This vector is then multiplied by the matrix S to produce the associated vector \tilde{v}_q . The equation $\tilde{v}_q = \tilde{v}_q S = \tilde{d}_q^T U$ is shown below.

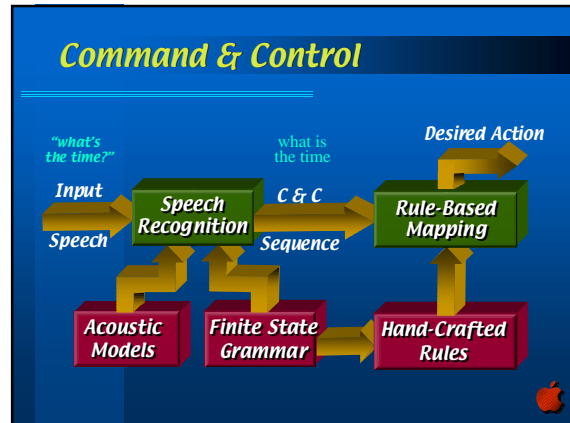


Usage

- Closest Document**
 - assign new document to most relevant topic
 - semantic classification problem
 - application to command & control
- Closest Word**
 - select word based on new document history
 - semantic prediction problem
 - application to language modeling

Outline

- Motivation
- Latent Semantic Analysis (LSA)
- Application to Spoken Language
 - command & control
 - language modeling
- Semantic Language Modeling
- Perspectives

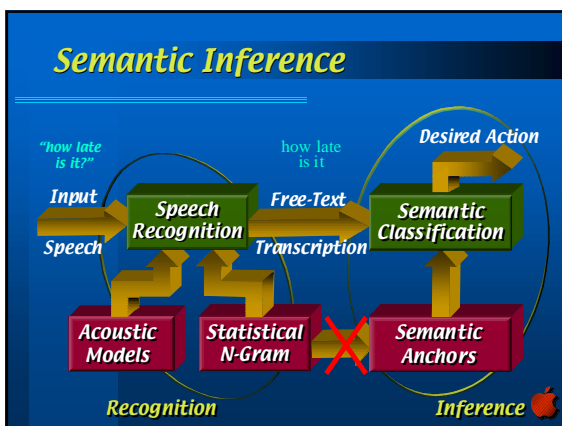


Problem

- Lack of Flexibility
 - FSG forced to play two **contradictory** roles
 - Language constraints: for recognition
 - Semantic constraints: for action mapping
- FSG Too Impoverished
 - hard to exhaustively encapsulate domain
 - many semantically correct paths missed
- FSG Too Rich
 - many paths not worth extra complexity
 - users can't remember what is/isn't OK

Solution

- Divide & Conquer
 - Recognition: what did the user say?
 - Action mapping: what does it mean?
- Recognition
 - No longer any "right" and "wrong" syntax
 - User's choice simply transcribed as is
- Action Mapping
 - Infer action from meaning of transcription



Computation

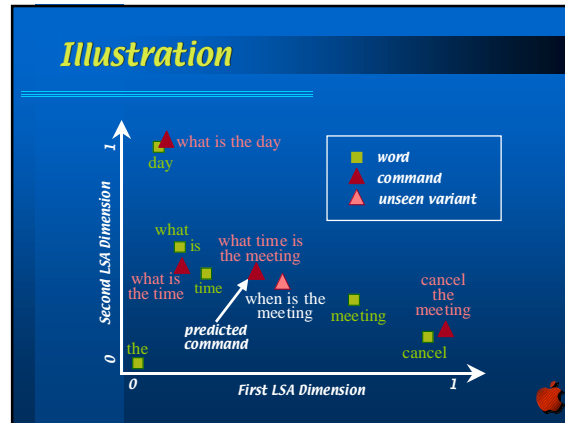
- Closeness
 - command d_ℓ (or cluster D_ℓ) mapped to \bar{v}_ℓ
 - new variant \tilde{d}_a mapped to \tilde{v}_a
$$\rightarrow K(\tilde{v}_a, \bar{v}_\ell) = \frac{\tilde{v}_a^T S^2 v_\ell^T}{\|\tilde{v}_a S\| \|v_\ell S\|}$$
- Classification
 - $\Pr(\tilde{d}_a | D_\ell) =$ suitably normalized $K(\tilde{v}_a, \bar{v}_\ell)$

Example

- **4 Commands**
 - what *and is* always co-occur
 - what is the time?
 - what is the day?
 - what time is the meeting?
 - cancel the meeting.
- **New Wording**
 - when is the meeting?

Annotations:

- predictive power: day, cancel vs the
- what is, the, time co-occur in two commands
- meeting does not help disambiguate
- when has never been seen

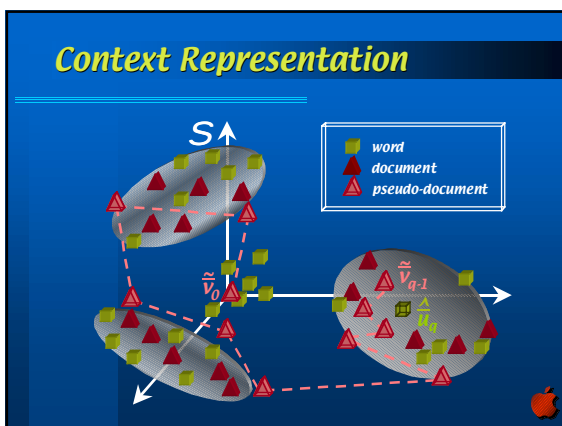


Actual Usage

- **Semantic Anchor**
 - add this to favorites folder
- **Actual Variant**
 - "set it aside where I keep the things I like best"
- **Top 5**
 - 0.96: add this to favorites folder **predicted command**
 - 0.24: add this to startup items
 - 0.14: add this to the Apple menu
 - 0.07: open the speakable items folder
 - 0.01: make this application speakable

Outline

- Motivation
- Latent Semantic Analysis (LSA)
- Application to Spoken Language
- **Semantic Language Modeling**
 - LSA component
 - integration with n-grams
 - perplexity / word error rate reduction
- Perspectives



LSA Component

- **Rationale**
 - predict word on basis of entire history \tilde{d}_{q-1}
 - "relevance" to current document
 - especially useful for content words
 - complementary to n-gram information
- **Direct Model**

$$Pr(w_q | \tilde{d}_{q-1}) = Pr(\bar{u}_q | \tilde{v}_{q-1})$$

Implementation

- Closeness**
 - new metric needed:

$$K(\bar{u}_q, \tilde{v}_{q-1}) = \frac{u_q S \tilde{v}_{q-1}^T}{\|u_q S^{1/2}\| \|\tilde{v}_{q-1} S^{1/2}\|}$$

- similar to classification with scaling by $S^{1/2}$
- Probability**
 - $\Pr(w_q | \tilde{a}_{q-1}) =$ suitably normalized $K(u_q, \tilde{v}_{q-1})$

Integration

- General Formulation**
 - H_{q-1} : integrated history for word w_q

$$\Pr(w_q | H_{q-1}) = \Pr(w_q | \underbrace{w_{q-1} \dots w_{q-n+1}}_{n\text{-gram}}, \underbrace{\tilde{a}_{q-1}}_{LSA})$$

- After Manipulation**

$$\Pr(w_q | H_{q-1}) = \frac{\Pr(w_q | w_{q-1} \dots w_{q-n+1}) \Pr(\tilde{a}_{q-1} | w_q)}{\sum_w \Pr(w | w_{q-1} \dots w_{q-n+1}) \Pr(\tilde{a}_{q-1} | w)}$$

Clustering

- Rationale**
 - leverage semantic partitions of LSA space
 - individual words \rightarrow semantic events
 - individual documents \rightarrow topics
 - better characterization of $\Pr(w_q | \tilde{a}_{q-1})$
- Clustered LSA Model**
 - exploit available knowledge layer(s) in S
 - mixture modeling \rightarrow act as smoothing
 - use either $\{C_k\}$, $\{D_\ell\}$, or both

Smoothing

- Word**

$$\Pr(w_q | \tilde{a}_{q-1}) = \sum_{k=1}^K \Pr(w_q | C_k) \Pr(C_k | \tilde{a}_{q-1})$$
- Document**

$$\Pr(w_q | \tilde{a}_{q-1}) = \sum_{\ell=1}^L \Pr(w_q | D_\ell) \Pr(D_\ell | \tilde{a}_{q-1})$$
- Joint**

$$\Pr(w_q | \tilde{a}_{q-1}) = \sum_{k=1}^K \sum_{\ell=1}^L \Pr(w_q | C_k) \Pr(C_k | D_\ell) \Pr(D_\ell | \tilde{a}_{q-1})$$

Experiments

- LM Training**
 - financial news (WSJ0), 42 Mwords
 - corpus size: $N=87,000$ documents
 - vocabulary size: $M=20,000$ words
 - n-gram: standard ARPA bigram/trigram
 - SVD: single vector Lanczos method, $R=125$
- Acoustic Setup**
 - training: 7,200 sentences (WSJ0, SI-84)
 - testing: 496 sentences from 12 speakers

Perplexity

Perplexity Reduction	no LSA component	n-gram+LSA with various smoothing			
		none	doc	word	joint
bigram	215	147	116	106	102
	-	32%	46%	51%	53%
trigram	142	115	103	98	95
	-	19%	28%	31%	33%

Word Error Rate

Error Rate Reduction	no LSA component	n-gram+LSA with various smoothing			
		none	doc	word	joint
bigram	16.7%	14.4%	13.4%	12.9%	13.0%
	-	13.7%	19.6%	22.5%	22.0%
trigram	11.8%	10.7%	10.4%	9.9%	9.9%
	-	9.3%	11.7%	15.8%	15.8%

Example

This approach is practiced by Wells Fargo Investment Advisors. Its bond funds diversify as much as possible, usually owning more than 500 different securities. If one company's bonds are clobbered by a recapitalization, the overall impact on Wells's portfolios remains tiny.

- **N-gram Alone**
 - 0.22: Wells as recognized in error
 - 0.06: Wells's
- **W/ Semantic Component**
 - 0.31: Wells's correctly recognized
 - 0.26: Wells as

- ### Outline
- Motivation
 - Latent Semantic Analysis (LSA)
 - Application to Spoken Language
 - Semantic Language Modeling
 - Perspectives
 - summary
 - future directions

- ### Summary
- **Latent Semantic Analysis**
 - large-span, data-driven, vector-based (S)
 - **Semantic Inference**
 - untag recognition / classification constraints
 - **Semantic Language Modeling**
 - n-gram: local info, frequency-dependent
 - LSA: global info, relevance-dependent
 - performance: > 20% reduction in WER

- ### Perspectives
- **Benefits**
 - framework for dimensionality reduction
 - rigorous extension of trigger pair concept
 - efficient integration of semantic structure
 - **Issues**
 - potential polysemy problem
 - sensitive to domain and/or style mismatch
 - not effective against function word errors

- ### Future Directions
- **LSA Component**
 - built-in sense disambiguation?
 - rapid update of LSA space S ?
 - **Further Integration**
 - leverage syntactic knowledge as well
 - use in conjunction with structured LMs?