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

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

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

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

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
  - 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
  - equivalence classes on n-gram history
  - caveat: reliance on parser!
- **Example**
  - stocks = NP headword; fell = VP headword
  - 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
  - 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
  - analyze word-document co-occurrences

**Co-Occurrence Matrix**

\[ w_{ij} = \text{weighted count} \left( w_i, \, d_j \right) \]

- \( c_{ij} \): count of \( w_i \) in document \( d_j \)
- \( w_{ij} = G_i \cdot L_{ij} \cdot c_{ij} \)
- \( L_{ij} \): 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**
  - $R = \text{number of singular values}$
  - $W_{(MxN)} = U_{(MxR)} S_{(RxR)} V_{(RxN)}^{T}$
  - $d_j$: documents
  - $w_i$: words

Benefits

- **Vector Representation**
  - single space for words and documents
  - closeness = semantic similarity
  - parsimonious dimension ($100 < R < 300$)

- **Consequences**
  - discrete entities $\rightarrow$ continuous space
  - amenable to usual clustering techniques
  - uncover high-level semantic regions

Word Clustering

- **Closeness Measure**
  - word $w_i$ mapped to $\overline{u}_i = u_i S$
  - metric:
  $$K(\overline{u}_i, \overline{u}_m) = \frac{u_i S^T u_m}{||u_i S|| ||u_m S||}$$

- **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

© Apple Computer, Inc.
Document Clustering

- **Closeness Measure**
  - document $d_j$ mapped to $\mathbf{v}_j = v_j S$
  - metric:
    $$ K(\mathbf{v}_j, \mathbf{v}_m) = \frac{\mathbf{v}_j S^2 v_m^T}{||v_j S|| ||v_m S||} $$

- **Outcome**
  - a set of document clusters $\{D_i\}$ in $S$, $1 \leq i \leq L$
  - see: Gotoh & Renals, EuroSpeech, Sept. 97

---

Illustration

- **LSA Cluster Index**
  - Illustration in LSA Space
  - S
  - $c_k$
  - $d_j$
  - $\mathbf{v}_j = v_j S$

- **Generalization**
  - **New Document**
    - at time $q$: construct vector for $\{w_1, \ldots, w_q\}$
    - pseudo-document $\mathbf{d}_q$
    - need associated vector in $S$: $\mathbf{v}_q$
  - $\mathbf{v}_q^T \mathbf{d}_q = \mathbf{v}_j^T \mathbf{d}_q$

- **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

Semantic Inference

• Closeness
  • command $d_q$ (or cluster $D_q$) mapped to $\bar{v}_l$
  • new variant $\tilde{d}_q$ mapped to $\tilde{v}_l$
  \[ K(\tilde{v}_q, \bar{v}_l) = \frac{\bar{v}_l^T S \tilde{v}_q}{||\bar{v}_l|| \cdot ||\tilde{v}_q||} \]

• Classification
  • $\Pr(\tilde{d}_q / D_q) = \text{suitably normalized } K(\tilde{v}_q, \bar{v}_l)$
Institute for Mathematics and Its Applications Workshop -- Mathematical Foundations of Natural Language Modeling

October 30-November 3, 2000

**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?

**Illustration**

- **Predictive Power**
  - time co-occur in two commands

**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
  - 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**

**Context Representation**

- **LSA Component**
  - **Rationale**
    - predict word on basis of entire history \( \tilde{a}_{q-1} \)
    - “relevance” to current document
    - especially useful for content words
    - complementary to n-gram information
  - **Direct Model**
    \[
    Pr ( w_q | \tilde{a}_{q-1} ) = Pr ( \tilde{u}_q | \tilde{v}_{q-1} )
    \]
Implementation

- Closeness
  - new metric needed:
  \[ K(\tilde{\mathbf{u}}_q, \tilde{\mathbf{v}}_{q-1}) = \frac{\mathbf{u}_q S^{1/2} \mathbf{v}_{q-1}^T S^{1/2}}{\|\mathbf{u}_q S^{1/2}\| \|\mathbf{v}_{q-1} S^{1/2}\|} \]
  - similar to classification with scaling by \( S^{1/2} \)
- Probability
  - \( \Pr(w_q | \tilde{\mathbf{a}}_{q-1}) = \text{suitably normalized } K(u_q, v_{q-1}) \)

Integration

- General Formulation
  - \( H_{q-1} : \text{integrated history for word } w_q \)
  \[ \Pr(w_q | H_{q-1}) = \Pr(w_q | w_{a_1} \ldots w_{a_{n-2}}, \tilde{\mathbf{a}}_{q-1}) \]
  - \( n \)-gram
  - LSA
- After Manipulation
  \[ \Pr(w_q | H_{q-1}) = \frac{\Pr(w_q | w_{a_1} \ldots w_{a_{n-2}}, \Pr(\tilde{\mathbf{a}}_{q-1} | w_{l-1}))}{\sum \Pr(w_q | w_{a_1} \ldots w_{a_{n-2}}, \Pr(\tilde{\mathbf{a}}_{q-1} | w_{l-1}))} \]

Clustering

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

Smoothing

- Word
  \[ \Pr(w_q | \tilde{\mathbf{a}}_{q-1}) = \sum_{k=1}^{\infty} \Pr(w_q | C_k) \Pr(C_k | \tilde{\mathbf{a}}_{q-1}) \]
- Document
  \[ \Pr(w_q | \tilde{\mathbf{a}}_{q-1}) = \sum_{l=1}^{\infty} \Pr(w_q | D_l) \Pr(D_l | \tilde{\mathbf{a}}_{q-1}) \]
- Joint
  \[ \Pr(w_q | \tilde{\mathbf{a}}_{q-1}) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \Pr(w_q | C_k) \Pr(C_k | D_l) \Pr(D_l | \tilde{\mathbf{a}}_{q-1}) \]

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

<table>
<thead>
<tr>
<th></th>
<th>PO + SA component</th>
<th>n-gram + LSA with various smoothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>bigram</td>
<td>215 - 147</td>
<td>116 - 106 - 102</td>
</tr>
<tr>
<td></td>
<td>32%</td>
<td>46% - 51%</td>
</tr>
<tr>
<td>trigram</td>
<td>142 - 115</td>
<td>103 - 98</td>
</tr>
<tr>
<td></td>
<td>19% - 28%</td>
<td>31% - 33%</td>
</tr>
</tbody>
</table>
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 portfolio remains tiny.

- **N-gram Alone**
  - 0.22: Wells as recognized in error
  - 0.06: Wells’s correctly recognized

- **W/ Semantic Component**
  - 0.31: Wells’s
  - 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
  - untie 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?