Latent Semantic Analysis
a tutorial at CogSci'2005

Benoît Lemaire

Laboratoire Leibniz-IMAG (CNRS UMR 5522)
University of Grenoble
France
Benoit.Lemaire@imag.fr
Outline

- Intuitive introduction
  - representing the meaning of words from the analysis of huge corpora
- LSA technique
- Tests
- Cognitive models
- Limits
- Competing models
Intuitive introduction
The pilot parks the plane
From words to texts

The pilot parks the plane
From words to texts

The pilot parks the plane
From words to texts

The pilot parks the plane

The diagram illustrates the relationships between words and concepts, showing how words like 'pilot', 'plane', 'park', 'flew', 'flying', etc., are connected in a semantic space. This helps to understand how meaning is represented in texts.
Usages of LSA
LSA as a tool

- Compare texts
- Control experiment material
LSA as a cognitive model of semantic memory

- Build lots of cognitive models
  - Text comprehension
  - Interface navigation
  - ...

Corpus

Introduction
LSA technique
Tests
Cognitive models
Limits
Competing models
Humans also construct the meaning of words from written material

Does LSA account for this process?
Outline

- Intuitive introduction
- LSA technique
- Tests
- Cognitive models
- Limits
- Competing models
Semantic information is drawn from raw texts

- Take a huge corpus
- Split it into paragraphs
- Determine the statistical context in which each word occurs
A first approach based on co-occurrence

- Define the meaning of a word from the co-occurring words
- Two words are similar if they occur in the same paragraphs
  - `plane = (1, 0, 0, 1, 0, 0, 0, 0, 2, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1)`
  - `airport = (1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 2, 1, 0, 0, 0, 0, 0, 0, 1)`
- Does not work well (French, Perfetti)
A better approach

- Two words are similar if they occur in the same paragraphs
A better approach

- Two words are similar if they occur in the same paragraphs

- Two words are similar if they occur in similar paragraphs
A better approach

- Two words are similar if they occur in the same paragraph

- Two words are similar if they occur in similar paragraphs

- Two paragraphs are similar if they contain common words
A better approach

- Two words are similar if they occur in the same paragraph
- Two words are similar if they occur in similar paragraphs
- Two paragraphs are similar if they contain common words
- Two paragraphs are similar if they contain similar words
A mutual recursion

- Mathematicians know how to solve such circular systems
- Suppose $P$ paragraphs, $W$ words
- Build $M$, a $P \times W$ matrix
  - $M[i,j] = \text{nb of occurrence of word } i \text{ in paragraph } j$
- Each word is a $P$-dimension vector
- Reduce the matrix to its best 300 dimensions
Example (Landauer & Dumais 1997)

§1: Human machine *interface* for ABC *computer* applications
§2: A *survey* of *user* opinion of *computer system response time*
§3: The *EPS user interface* management *system*
§4: *System* and *human system* engineering testing of *EPS*
§5: Relation of *user* perceived *response time* to error measurement
§6: The generation of random, binary, ordered *trees*
§7: The intersection *graph* of paths in *trees*
§8: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
§9: *Graph minors*: A *survey*

| 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 |
Matrix reduction

$$M = W \times 0 \times P$$

words = paragraphs
Matrix reduction

\[ M \text{ paragraphs} = W \times 0 \times P \]

\[ M' \text{ words} = W' \times 0 \times P' \]
Example (Landauer & Dumais 1997)

§1: Human machine interface for ABC computer applications
§2: A survey of user opinion of computer system response time
§3: The EPS user interface management system
§4: System and human system engineering testing of EPS
§5: Relation of user perceived response time to error measurement
§6: The generation of random, binary, ordered trees
§7: The intersection graph of paths in trees
§8: Graph minors IV: Widths of trees and well-quasi-ordering
§9: Graph minors: A survey

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

\[ r(\text{human}, \text{user}) = -0.38 \]
\[ r(\text{human}, \text{user}) = 0.94 \]
Meaning as vectors

- The meaning of a each word is represented as a vector

computer = (1,4,5,-1,0,2,-2,5,-6,8,11.....8,0,0,-5,7,1)
disk = (2,4,6,-1,0,3,-4,8,-5,8,12......9,0,1,-9,4,1)

Illustrative purpose only:
does not work in 2 dimensions!
Comparison with symbolic approaches

Introduction
LSA technique
Tests
Cognitive models
Limits
Competing models
Comparison with symbolic approaches

- Representation is absolute
- Good for drawing inferences
- Not so good for comparisons
- Hard to compound nodes

**Diagram:**
- Vehicle
  - Human-propelled vehicle
    - Bike
      - Guide
      - Function: ride
      - HAS: wheels
    - Bicycle
      - Synonym: bicycle
      - HAS: wheels
Comparison with symbolic approaches

- Representation is relative
- Good for semantic comparisons
- Easy to compound words
Semantic comparison

- Semantic distance = cosine between vectors
- A value between –1 and 1
- Other measures:
  - Euclidian distance
  - Hellinger distance
  - ...
Outline

- Intuitive introduction
- LSA technique
- Tests
- Cognitive models
- Limits
- Competing models
Word comparisons

- A semantic space built from a 4.6 million word encyclopedia (lsa.colorado.edu)
  - Closest words to bicycle:
    - pedals: .84
    - handlebars .79
    - bicycles .79
    - pedaling .75
    - bike .75
  - fly – insect: .26
  - fly -plane: .48
  - insect – plane: .02
Existing semantic spaces

- **English**
  - encyclopedia, literature, psychology textbooks, scientific textbooks (available at lsa.colorado.edu)

- **French**
  - newspaper, literature, children tales, children production (some of them are available at lsa.colorado.edu)
What is the right number of dimensions?

- Too few dimensions would not capture the latent semantic structure.
- Too much dimensions would describe idiosyncratic structures.
- An open issue.
- 300 dimensions seems a good value for the whole language (Dumais found a maximum of performance at 90 dim. for a specific domain).
From words to texts

- Sum up word vectors to get text vectors

\[ S = The \ computer \ disk \ was \ broken \ by \ the \ kid \]

\[
\begin{align*}
\text{computer} &= (1, 4, 5, -1, 0, 2, -2, 5, \ldots, 8, 0, -5, 7, 1) \\
\text{disk} &= (2, 4, 6, -1, 0, 3, -4, 8, \ldots, 9, 0, -9, 4, 1) \\
\text{broken} &= (5, -3, 8, 6, 5, -2, 2, 0, \ldots, 6, 1, 7, 1, 2) \\
\text{kid} &= (1, 0, 0, -9, 4, 5, -1, 2, \ldots, 0, 2, 3, 4, 5) \\
S &= (9, 5, 19, -5, 9, 8, -5, 15, \ldots, 23, 3, -4, 16, 9)
\end{align*}
\]
Text comparisons

- A semantic space built from a 4.6 million word encyclopedia (lsa.colorado.edu)
  - The cat was lost in a forest / My little feline disappeared in the trees: .66
  - The radius of spheres / a circle's diameter: .55 (Landauer)
  - The radius of spheres / the music of spheres: .01 (Landauer)
No syntax processing

- Paragraphs are bags of words
- Semantic is determined from the lexical level
We learned that this is about birds, nests and a building process

The analysis unit is the paragraph
Outline

- Intuitive introduction
- LSA technique
- Tests
- Cognitive models
- Limits
- Competing models
Models based on LSA

- Semantic representation
- Vocabulary acquisition
- Text comprehension
- Free text assessment

Introduction
LSA technique
Tests
Cognitive models
Limits
Competing models
Semantic representation

- Comparison of words
  - TOEFL test
  - Vocabulary test for children
- Comparison of sentences
  - Measure of coherence
- Comparison of texts
Comparison of words

- Synonymy part of the TOEFL test (Landauer & Dumais, 97)
  - 80 items: 1 word/4 alternative words
  - Guess which one is the closest
  - Example: levied
    - imposed
    - believed
    - requested
    - correlated
  - LSA was trained on a 4.6 million word corpus
  - LSA score: 64.4%
  - Foreign applicants to US colleges: 64.5%
Vocabulary test: LSA vs children
(Denhière & Lemaire, 2004)

- 115 items
- 4 definitions: correct, close, distant, unrelated
- Subjects were in grade 2 to 4
- Example (translated from French):

**slope**
- rising road
- tilted surface which goes up or down
- small piece of ground
- face of a rock or a mountain
A children's semantic space

- Stories and tales for children: ~ 1,600,000 words
- Children productions: ~ 800,000 words
- Textbooks: ~ 400,000 words
- Encyclopedia: ~ 400,000 words
- Dictionary: ~ 50,000 words

TOTAL: 3.2 million words
Vocabulary test: results

Children data

Correct  Close  Far  Unrelated

2nd grade  3rd grade  4th grade
Vocabulary test: results

Children+Model data

Correct Close Far Unrelated

2nd grade
3rd grade
4th grade
LSA
Children+Model data

- 2nd grade
- 3rd grade
- 4th grade
- 115 words
- 99 words

Correct | Close | Far | Unrelated
Vocabulary test: word lemmatization

Children+Model data

2nd grade
3rd grade
4th grade
115 words
99 words
lemm.
Vocabulary test: verb lemmatization

Children+Model data

Correct  Close  Far  Unrelated
2nd grade
3rd grade
4th grade
115 words
99 words
verb lemm.

0  5  10  15  20  25  30  35  40  45  50  55  60  65  70

Children+Model data
Comparison of sentences

- Use LSA to measure the coherence of a text
- Compute the semantic similarities between adjacent sentences

\[
\text{coherence (T)} = \frac{\sum_{i=1}^{n-1} \text{cosine(sentence } i, \text{ sentence } i+1)}{n-1}
\]
Foltz et al. (1998) experiment

- McNamara et al. (1996) data
- 4 versions of text about the heart
  - Low local coherence, low macro coherence (lm)
  - Low local coherence, high macro coherence (IM)
  - High local coherence, low macro coherence (Lm)
  - High local coherence, high macro coherence (LM)
- Corpus: 21 texts, 100 dimensions
Foltz et al. (1998) experiment

Results:
- Coherence(lm) = 0.177
- Coherence(lM) = 0.205
- Coherence(Lm) = 0.210
- Coherence(LM) = 0.242
Can LSA assess text recall?

- Measure M1: cosine between source text and texts recalled
- Measure M2: number of propositions recalled

<table>
<thead>
<tr>
<th>Text</th>
<th>Recall</th>
<th>Nb subjects</th>
<th>Correlation M1/M2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poule</td>
<td>Immediate</td>
<td>n = 48 (8,3 years old)</td>
<td>0.49</td>
</tr>
<tr>
<td>Dragon</td>
<td>Immediate</td>
<td>n = 48 (8,3 years old)</td>
<td>0.64</td>
</tr>
<tr>
<td>Dragon</td>
<td>Delayed</td>
<td>n = 48 (8,3 years old)</td>
<td>0.7</td>
</tr>
<tr>
<td>Araignée</td>
<td>Immediate</td>
<td>n = 57 (16-20 years old)</td>
<td>0.81</td>
</tr>
<tr>
<td>Clown</td>
<td>Immediate</td>
<td>n = 57 (16-20 years old)</td>
<td>0.65</td>
</tr>
<tr>
<td>Géant</td>
<td>Immediate</td>
<td>n = 130 (16-20 years old)</td>
<td>0.6</td>
</tr>
<tr>
<td>Ourson</td>
<td>Immediate</td>
<td>n = 44 (16-20 years old)</td>
<td>0.48</td>
</tr>
<tr>
<td>Clown</td>
<td>Summary</td>
<td>n = 44 (16-20 years old)</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Comparison of texts
Foltz et al. (1996)

- Compare texts and their summaries
- Britt et al. (1994) data
  - 24 participants read 21 texts
  - Participants write an essay
- Corpus: 21 texts + 8 encyclopedia texts + 2 book excerpts
- Assessing the essays:
  - \textit{Score 1} = average similarity between each sentence and the closest text sentence
  - \textit{Score 2} = same but compare with the main 10 text sentences
Comparison of texts
Foltz et al. (1996)

<table>
<thead>
<tr>
<th></th>
<th>Judge1</th>
<th>Judge2</th>
<th>Judge3</th>
<th>Judge4</th>
<th>Score1</th>
<th>Score2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge2</td>
<td>.575**</td>
<td></td>
<td></td>
<td></td>
<td>.418*</td>
<td>.589*</td>
</tr>
<tr>
<td>Judge3</td>
<td>.582**</td>
<td>.412*</td>
<td></td>
<td></td>
<td>.552**</td>
<td>.626**</td>
</tr>
<tr>
<td>Judge4</td>
<td>.367</td>
<td>.317</td>
<td>.384+</td>
<td></td>
<td>.117</td>
<td>.240</td>
</tr>
</tbody>
</table>

** p<.01   * p<.05    + p<.06
Models based on LSA

- Semantic representation
- Vocabulary acquisition
- Text comprehension
- Free text assessment
A model of learning

- LSA learns the meaning of words from texts
- Does it mimic the human rate of learning?
- Landauer & Dumais (1997) experiment:
  - Most of the lexicon is acquired from text
  - Children data:
    - 3500 words read / day
    - 50 new words encountered / day
    - 10 words learned / day
    - Controlled experiments lead to 2.5 words learned / day
Direct and indirect effects

- How do children acquire the remaining 7.5 words/day?
- The poverty of stimulus problem
  - how do children acquire as much knowledge on the basis of little information?
- Hypothesis: a word is learned...
  - From texts containing it (direct effect)
  - From texts *not* containing it (indirect effect)
LSA simulation

- TOEFL test score is function of:
  - T: total number of paragraphs
  - S: number of paragraphs containing the stem word

\[ \text{score} = 0.128 \times \log(0.076 \times T) \times \log(31.910 \times S) \]

- Correlation with observed score = .98

- After “reading” 25,000 paragraphs, what is the effect of the 25,001th?
  - When the stem word belongs to it?
  - When the stem word does not belong to it?
Example

- Suppose a word that appeared twice in the 25,000 paragraphs
  \[ \text{Score} = 0.128 \times (\log 0.076 \times 25000) \times (\log 31.91 \times 2) = 0.75809 \]

- Direct effect: the word belongs to the 25,001th paragraph
  \[ \text{Score} = 0.128 \times (\log 0.076 \times 25001) \times (\log 31.91 \times 3) = 0.83202 \]

- Indirect effect: the word does not belong to the 25,001th paragraph
  \[ \text{Score} = 0.128 \times (\log 0.076 \times 25001) \times (\log 31.91 \times 2) = 0.75810 \]
Effects are cumulated over all word-frequency band

Direct effect:
- .0007 words gained per word encountered
- .0007 * 3500 = 2.5 words acquired / day

Indirect effect:
- .15 words gained per paragraph encountered
- .15 * 50 = 7.5 words acquired / day

The meaning of a word would be acquired
- for 25% by reading text containing it
- for 75% by reading texts not containing it
Direct and indirect effects
(Lemaire et al., to appear)

acheter (to buy)/magasin (store)
Cosine acheter/magasin

- Total gain: 0.58
- Gain due to the occurrence of acheter: -0.10
- Gain due to the occurrence of magasin: -0.19
- Gain due to the co-occurrence: 0.73
- Gain due to 2\textsuperscript{nd} order co-occurrences: 0.11
- Gain due to 3\textsuperscript{rd} and more co-occurrences: 0.03
Same measures on 28 words

- Total gain: \(0.13\)
- Gain due to the occurrence of \(W_1\): \(-0.16\)
- Gain due to the occurrence of \(W_2\): \(-0.19\)
- Gain due to the co-occurrence: \(0.34\)
- Gain due to 2\(^{nd}\) order co-occurrences: \(0.05\)
- Gain due to 3\(^{rd}\) and more co-occurrences: \(0.09\)
Models based on LSA

- Semantic representation
- Vocabulary acquisition
- **Text comprehension**
- Free text assessment
Text comprehension

- Cognitive models can be simulated on computers, using LSA as a model of semantic memory:
  - the construction-integration model (Kintsch, 1998)
  - A predication algorithm
  - Metaphor comprehension
A Computational model of text comprehension (Lemaire et al, to appear)
Example (translated from French)

- A woodcutter was walking in the forest

- Select associates:
  - walk: stroll, meet, pick
  - woodcutter: ax, forest, cottage
  - forest: glade, oak, wood
Example (translated from French)

- A woodcutter was walking in the forest

walk/woodcutter/forest
walk
pick
meet
stroll

woodcutter
ax
cottage

forest
oak
wood
ax

glade
Example (translated from French)

- A woodcutter was walking in the forest
Example (translated from French)

- *A woodcutter was walking in the forest*
Example (translated from French)

- *A woodcutter was walking in the forest*
The construction-integration model

- Before LSA, the construction step was performed by hand
- Link weights were set approximately
- All nodes being vectors, the weight of a link is the cosine of the corresponding nodes
- A question remains: what is the vector of $P(A_1, \ldots, A_n)$?
The predication algorithm
(Kintsch, 2001)

- The centroid rule:
  \[ \overrightarrow{P(A)} \rightarrow \overrightarrow{P+A} \]
- ...does not work well
  - The meaning of the predicate depends on the meaning of the arguments
  - *The horse ran / The color ran*
  - *shark(lawyer) ≠ shark + lawyer*

- A better rule:
  \[ \overrightarrow{P(A)} \rightarrow \overrightarrow{P+A+M_1+\ldots+M_n} \]
  - \( M_i \) are vectors that are close to both \( P \) and \( A \)
Example 1: **construction step** (Kintsch, 2001)

The horse ran

```
ran
  \__________\                      \_____________
  |          |                      |            |
  | stopped  |                      | horse     |
  \__________\                      \_____________
  |          |                      |            |
  | ran      |                      | down      |
  \__________\                      \_____________
  |          |                      |            |
  | hopped   |                      |            |
```

ran(horse)
Example 1: integration step

The horse ran ran(horse)

The horse ran ➔ horse+ran+stopped
Example 2: construction step

The color ran ran(color)

![Diagram showing the construction step with probabilities: ran -> down (.11), ran -> stopped (.06), hopped -> color (.02).]
Example 2: integration step

The horse ran ran(horse)

The color ran color+ran+down
### Predication algorithm: test
(Kintsch, 2001)

<table>
<thead>
<tr>
<th>Landmarks</th>
<th>ran</th>
<th>horse ran</th>
<th>color ran</th>
</tr>
</thead>
<tbody>
<tr>
<td>gallop</td>
<td>.33</td>
<td>.75</td>
<td>.29</td>
</tr>
<tr>
<td>dissolve</td>
<td>.01</td>
<td>.01</td>
<td>.11</td>
</tr>
</tbody>
</table>
## Inferences (Kintsch, 2001)

<table>
<thead>
<tr>
<th>Inference 1</th>
<th>Inference 2</th>
<th>Inference 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>The student washed the table</td>
<td>The student was clean</td>
<td>The table was clean</td>
</tr>
<tr>
<td>Predication</td>
<td>Centroid</td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.71</td>
</tr>
<tr>
<td>The student dropped the glass</td>
<td>The student was broken</td>
<td>The glass was broken</td>
</tr>
<tr>
<td>Predication</td>
<td>Centroid</td>
<td>.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.83</td>
</tr>
<tr>
<td>The hunter shot the elk</td>
<td>The hunter was dead</td>
<td>The elk was dead</td>
</tr>
<tr>
<td>Predication</td>
<td>Centroid</td>
<td>.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.91</td>
</tr>
<tr>
<td></td>
<td>Predication</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td>Predication</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.54</td>
</tr>
<tr>
<td></td>
<td>Predication</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.70</td>
</tr>
</tbody>
</table>
Metaphors

- **My lawyer is a shark** \( \text{shark(lawyer)} \)

The meaning of *my lawyer is a shark* is not a simple combination of the meaning of *lawyer* and the meaning of *shark*
Metaphors

- Look for words that are similar to *shark* but also close to *lawyer*
Metaphors

My lawyer is a shark

My lawyer is a shark

lawyer

shark

vicious

aggressive

Cognitive models

Competing models
Metaphors

- Much more neighbors of the predicate need to be considered
- According to Kintsch, predication fails if number of neighbors is less than 100!
An incremental algorithm
(Lemaire & Bianco, 2003)

- Does not work on a predefined set of neighbors but look for them progressively

*This kid is intelligent*
1... 2... 3 *clever* 4... 5 *brilliant* 6... 7... 8... 9 *sage*

*This kid is a scientist*
1... 2... 3 ...................................................................................................
 .......................................................................................
 ........................................................................97 *invention* ..........................................................
 .........................................................................................128 *knowledge*
 .................................................................................138 *wonderful*
Experiment

- Does the model account for the processing time of various kind of metaphors?
- 12 texts, 4 conditions:
  - Synonym (child / kid)
  - Conventional metaphor (child / scientist)
  - Original metaphor (child / sage)
  - Bad metaphor (child / colonel)
Subjects (N=49)

Subjects

Additionnal time (ms)

Types of metaphors

- Synonym
- Conv. met.
- Unconv. Met.
- Bad met.
Subjects and model

- Correlation: .62
Models based on LSA

- Semantic representation
- Vocabulary acquisition
- Text comprehension
- Free text assessment
Free text assessment

- Assessing the content of a student essay
- Rely on a domain-specific corpus
- Compare the student essay with a reference text:
  - Pre-graded essays
  - Text to be summarized
  - Portions of a text course

Intelligent Essay Assessor

- Foltz et al. (1999)
- **Holistic score:**
  - Compare student essay with pre-graded essays
  - Give the grade of the closest one
- **Gold standard score**
  - Compare student essay with a teacher essay
  - Give the grade according to the semantic similarity
  - Correlation with teacher's grades around .8
Summary Street

- Helps the student to summarize a text
- Indicate the part of the text that are well or not well summarized
- An experiment with 60 6th grade students has shown a positive effect
Summary Street

**Guest,** these bars show how well your summary covered the sections of the text you read. If the bar passes the black line, then you’ve written enough information about that section. When your summary contains enough information about every section, you can advance to the next level and will receive more advice on how to improve your summary.

The blue dashed lines show how much information your previous summary contained, so you can see if you are improving or not.

Compare student essay with relevant portions of the course

The course is structured:

#T The first chapter
#N In the first part, we will study the way...
#N Now, I will present a method...
...
#T The second chapter
#N ...
Apex Assessor

Notions of Topic 9 : Le climat de la classe

Définition de la notion de climat
L'écologie
Le milieu
Le système social
La culture

Please write your essay below :

EDMONDS dans ses synthèses des travaux sur l'efficacité des écoles a mis à jour cinq facteurs associés à de meilleures performances des écoles. Ce "modèle des cinq facteurs" d'EDMONDS aura un gros retentissement et sera à l'origine de la mise en place de plusieurs programmes d'amélioration des écoles (projets RISE, SIP et LSDP notamment). Ces cinq facteurs sont les suivants :
- une forte direction ("leadership") ;
- des attentes élevées concernant les performances des élèves ;
- un climat discipliné, sans toutefois être rigide ;
- un fort accent mis sur l'enseignement des savoirs de base (lecture, écriture, mathématiques) ;
- des évaluations et des contrôles fréquents des progrès des élèves.
L'école est conçue comme un système social qui, en tant que tel, a un fonctionnement spécifique, développe un système particulier de

GENERAL GRADE: 8.8/20
The following Notions were covered very poorly:
- l'écologie (0.03)

The following Notions were covered poorly:
- le système social (0.34)

The following Notions were covered well:
- définition de la notion de climat (0.63)
- le milieu (0.58)
- la culture (0.61)
Correlation with teacher's grades: .62
- Not only for grading but for engaging students in a revising process
- Can be used at a distance
Outline

- Intuitive introduction
- LSA technique
- Tests
- Cognitive models
- Limits
- Competing models
Strengths and limits

- **LSA strengths**
  - fully automatic
  - vectorial representation allows:
    - easy representation of sequence of words
    - easy comparisons between words or texts
  - a cognitive model of learning and representation

- **LSA weaknesses**
  - no syntax processing
  - representations are not explicit
  - not incremental
  - paragraphs are bag of words
Outline

- Intuitive introduction
- LSA technique
- Tests
- Cognitive models
- Limits
- Competing models
Cognitive models of semantic memory could mimic:
- the human semantic associations or
- the human *construction* of these associations over a long period of time

Semantic networks are good model of semantic associations BUT... they do not account for the construction of these associations
Features

- **Input** (corpus, word association norms)
- **Knowledge representation** (vector-based, network-based)
- **Addition of a new text** (incrementally or not)
- **Unit of context** (paragraph, sliding window)
- **Use of high-order co-occurrences** (yes/no)
- **Compositionality** (representing the meaning of a text from the meaning of its words)
LSA

- Input: **corpus**
- Knowledge representation: **vector-based**
- Addition of a new text: **not incremental**
- Unit of context: **paragraph**
- Use of high-order co-occurrences: **yes**
- Compositionality: **easy**
Word co-occurrence matrix built from a N-word sliding window

Sample Global Co-occurrence Matrix for the Sentence “the horse raced past the barn fell”

<table>
<thead>
<tr>
<th></th>
<th>barn</th>
<th>horse</th>
<th>past</th>
<th>raced</th>
<th>the</th>
</tr>
</thead>
<tbody>
<tr>
<td>barn</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>fell</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>horse</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>past</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>raced</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

- vector = concatenation of row and column
- barn = <0,2,4,3,6,0,5,0,0,0>
- Measure of similarity = Euclidian distance
HAL
(Burgess, 1998)

Table 2
Example Neighborhoods for *beetles* and *frightened*

<table>
<thead>
<tr>
<th>Word</th>
<th>Neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>beetles</em></td>
<td>original, hand,</td>
</tr>
<tr>
<td></td>
<td>song, movie,</td>
</tr>
<tr>
<td></td>
<td>album, songs</td>
</tr>
<tr>
<td><em>frightened</em></td>
<td>scared, upset,</td>
</tr>
<tr>
<td></td>
<td>shy, embarrassed</td>
</tr>
<tr>
<td></td>
<td>anxious, worried</td>
</tr>
</tbody>
</table>
HAL
(Burgess, 1998)
HAL
(Burgess, 1998)

- Input: corpus
- Knowledge representation: vector-based
- Addition of a new text: incremental
- Unit of context: sliding window
- Use of high-order co-occurrences: no
- Compositionality: easy
Word association space
(Steyvers et al., in press)

- Based on association norms for 5,000 words
- Word/word matrix scaled to 200-500 dimensions
Figure 2. Correlations of different measures of semantic similarity for different dimensionalities. Data are taken from recognition memory, cued recall, free recall. See text for details.
Word association space
(Steyvers et al., in press)

- Input: association norms
- Knowledge representation: vector-based
- Addition of a new text: not incremental
- Unit of context: N/A
- Use of high-order co-occurrences: no
- Compositionality: easy
Based on Mutual Information (Church & Hanks, 89)
- compares the probability of two words occurring together with the probability of the words occurring separately
  \[ \text{MI}(A,B) = \log_2 \left( \frac{p(A,B)}{p(A).p(B)} \right) \]
- TOEFL test: 1 problem word/4 choices
  \[ \text{score}_1(\text{choice}) = \frac{\text{hits}(\text{problem AND choice})}{\text{hits}(\text{choice})} \]
  \[ \text{score}_2(\text{choice}) = \frac{\text{hits}(\text{problem NEAR choice})}{\text{hits}(\text{choice})} \]
  ...

PMI-IR
(Turney, 2001)
**PMI-IR**  
*(Turney, 2001)*

### Table 3. Results of the TOEFL experiments, including LSA results from [6].

<table>
<thead>
<tr>
<th>Interpretation of ( p(\text{problem} \mid \text{choice}) )</th>
<th>Description of Interpretation</th>
<th>Number of Correct Test Answers</th>
<th>Percentage of Correct Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>score(_1)</td>
<td>co-occurrence using AND operator</td>
<td>50/80</td>
<td>62.5%</td>
</tr>
<tr>
<td>score(_2)</td>
<td>co-occurrence using NEAR</td>
<td>58/80</td>
<td>72.5%</td>
</tr>
<tr>
<td>score(_3)</td>
<td>co-occurrence using NEAR and NOT</td>
<td>59/80</td>
<td>73.75%</td>
</tr>
<tr>
<td>Latent Semantic Analysis</td>
<td></td>
<td>51.5/80</td>
<td>64.4%</td>
</tr>
<tr>
<td>Average Non-English US College Applicant</td>
<td></td>
<td>51.6/80</td>
<td>64.5%</td>
</tr>
</tbody>
</table>
PMI-IR
(Turney, 2001)

- Input: **web**
- Knowledge representation: **N/A**
- Addition of a new text: **incremental**
- Unit of context: **web page**
- Use of high-order co-occurrences: **no**
- Compositionality: **hard**
ICAN
(Lemaire & Denhière, 2004)

- Network-based representation
  - better for quickly retrieving neighbors
  - better for representing asymmetrical relations

![Diagram showing connections between words like electric, plug, cable, and network with associated values]
... if you have such a [device, connect the cable to the network connector] then switch ...

Create or reinforce co-occurrence links
... if you have such a [device, connect the cable to the network connector] then switch ...

- Create or reinforce co-occurrence links
- Create or slightly reinforce 2<sup>nd</sup>-order co-occurrence links
... if you have such a [device, connect the cable to the network connector] then switch ...

- Create or reinforce co-occurrence links
- Create or slightly reinforce 2\textsuperscript{nd}-order co-occurrence links
- Slightly decrease other links
ICAN

- 3.2 million word corpus of texts for children
- Test on 1200 pairs of words (from association norms)
- Average weights:
  - stem/1\(^{st}\) associate: .415
  - stem/2\(^{nd}\) associate: .269
  - stem/3\(^{rd}\) associate: .236
  - stem/last associates: .098
- Correlation with human data: .50
ICAN

- **Input:** corpus
- **Knowledge representation:** network-based
- **Addition of a new text:** incremental
- **Unit of context:** sliding window
- **Use of high-order co-occurrences:** yes
- **Compositionality:** hard
Web pages

- Tests
  - lsa.colorado.edu

- References
  - www-leibniz.imag.fr/~blemaire/lsa.html