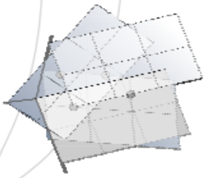
Automated Coding of Qualitative Interviews with Latent Semantic Analysis

ISTA, May 24, 2007, Kharkov



Fridolin Wild
Ingo Feinerer
Vienna University of Economics and Business Administration

Structure of this Talk

- Motivation: Why Automated Coding?
- Latent Semantic Analysis
- Algorithm I: headcount
- Algorithm II: termcount
- Evaluation
- Conclusion & Future Agenda

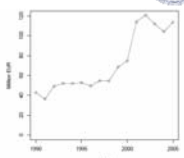
<2>

Motivation

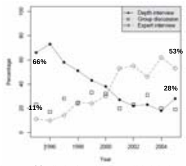
<3>

Motivation

- Increased deployment of qualitative methods in marketing
- But: decrease of in-depth interviews due to high costs
- But: qualitative research has advantages: not feeding analysts expectations so much, open ended, spontaneous associations
- Problem: High Human Resource Costs
- Problem: inherent subjectivity in manual coding:
  - More interviews = more errors
  - More coders = more errors



Members‘ annual turnover 1995-2005 for qual. Res. (Arbeitskre[is] Deutscher Markt- und Sozialforschungsinstitute e.V., 2006)



Development of Qualitative Interview Types (Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V., 2006)

Latent Semantic Analysis

Input (e.g., documents)

term = feature

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

vocabulary = ordered set of features

Only the red terms appear in more than one document, so strip the rest.

{ M } =

	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

TEXTMATRIX

Deerwester, Dumais, Furnas, Landauer, and Harshman (1990): Indexing by Latent Semantic Analysis, In: Journal of the American Society for Information Science, 41(6):391-407

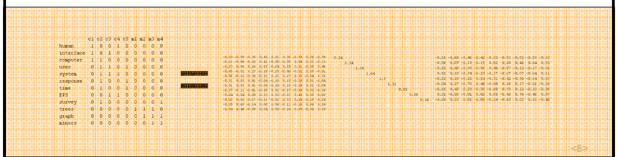
Latent Semantic Analysis

- “Humans learn word meanings and how to combine them into **passage meaning** through experience with ~paragraph unitized verbal environments.”
- “They don’t remember all the separate words of a passage; they remember its **overall gist** or meaning.”
- “LSA learns by ‘reading’ **~paragraph unitized texts** that represent the environment.”
- “It doesn’t remember all the separate words of a text it; it remembers its **overall gist or meaning**.”

(Landauer, 2007) <7>

Singular Value Decomposition

$$M = T\ S\ D^T$$



<8>

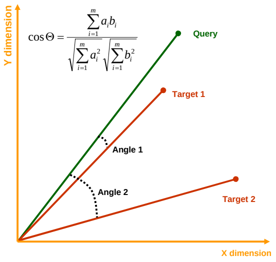
Latent Semantics

- Assumption: language utterances have a semantic structure
- However, this structure is obscured by word usage (noise, synonymy, polysemy, …)
- Proposed LSA Solution: map doc-term matrix using conceptual indices derived statistically (truncated **SVD**) and make similarity comparisons using **angles**

$$T_k\ S_k\ D_k^T = M_k$$

<9>

Similarity in a Latent-Semantic Space



(Landauer, 2007)

Ex Post Updating: Folding-In

- SVD factor stability
  - SVD calculates factors over a given text base
  - Different texts – different factors
  - Challenge: avoid unwanted factor changes (e.g., bad essays)
  - Solution: folding-in of essays instead of recalculating
- SVD is computationally expensive
  - 14 seconds (300 docs textbase, this machine)
  - 10 minutes (3500 docs textbase, this machine)
  - … and rising!

Algorithm I: Headcount

Algorithm I: Headcount

- Calculate latent-semantic space from answers per brand per person
  - e.g. „bad advertisement focused on young target group first net in the market expensive“
- Fold-in concept of interest + synonyms & distinct paraphrases = ‚seed terms‘ defining the concept
  - E.g. „big market share, established, known“
- (Several organised ‚concepts‘ = ‚coding scheme‘)
- Headcount = 100 * number of answers correlating high with the concept / number of answers

$$\mathrm{hc} = \frac{\left|\{d \mid (\mathrm{cor}\,(c,d) > l) \wedge (d \in M)\}\right| \cdot 100}{|M|}$$

Algorithm II: Termcount

Algorithm II: Termcount

- Calculate latent-semantic space from answers
- Fold-in brand name
  - e.g. ‚Mercedes‘
- Fold-in ‚seed-terms‘ for coding construct
  - e.g. ‚secure safe stability‘
- Measure distance between the two vectors = association strength (Pearson‘s product moment correlation coefficient)

$\mathrm{tc}\colon c \times b \mapsto \mathbb{R}$ $\qquad$ $\mathrm{tc} = \mathrm{cor}\,(c,b)$

Evaluation

<16>

Methodology

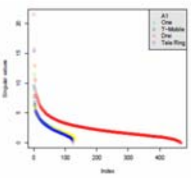
- Pseudo Experiment to evaluate validity
- External validation: machine findings against human analysis results
- Two real-life data sets:
  - Set 1: Austrian Mobile Phone Market (Marketmind, Soja Ehrenberger, Wolfgang Rejzlik)
  - Set 2: German & US Automobile Sector (for Mercedes, Andreas Strebinger)

<17>

Data-Set 1: Mobile Phone Market

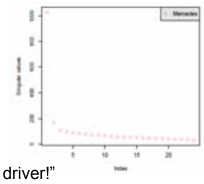
- 969 Interviews conducted by MarketMind
- Open questions to activate brand associations:
  - "Which image do you perceive if you consider brand X?"
  - "Please imagine brand Z. What do you associate?"
  - "What are your impressions and feelings you relate to brand Y?"
- Up to 10 short answers per interview
- Questions and answers in German
- Short answers (Ø: 103 chars, std. dev.: 61 chars, Ø: 14 words)

Brand	n	μ	σ
A1	134	94	62
One	120	106	62
T-Mobile	125	96	53
Drei	468	112	68
Tele Ring	122	105	59



<18>

Data-Set 2: Automobile Sector



- 24 German interviews about brand ‚Mercedes' in USA and Germany
- Each interview had ~ 64 questions
  - "If I buy a Mercedes, I have a good feeling because . . . "
  - "Please characterise a typical Mercedes driver!"
  - "Please tell me three things you directly associate with Mercedes!"
- length: long answers (each interview 3500 to 11.500 words, Ø: ~ 3500 words)
- 1624 answers (for 1624 questions)

Results for Algorithm I

Brand	$\mathcal{D}$	$\mathcal{T}$	ρ	p-value
A1	0.5	0.6	0.52	0.000
One	0.5	0.6	0.51	0.000
T-Mobile	0.5	0.6	0.40	0.004
Drei	0.5	0.6	0.37	0.008
Tele Ring	0.5	0.6	0.36	0.012
A1	0.5	0.5	0.61	0.000
One	0.5	0.5	0.59	0.000
T-Mobile	0.5	0.5	0.35	0.014
Drei	0.5	0.5	0.46	0.001
Tele Ring	0.5	0.5	0.45	0.001
A1	0.3	0.5	0.55	0.000
One	0.3	0.5	0.52	0.000
T-Mobile	0.3	0.5	0.31	0.030
Drei	0.3	0.5	0.44	0.001
Tele Ring	0.3	0.5	0.37	0.009

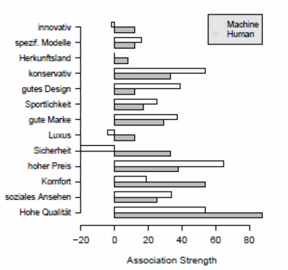
- D: share of cumulative singular values
- T: Threshold
- ρ: Spearman's rho

=> highly significant

=> correlation with human judgement in a range slightly less than human-human interrater correlation

=> Expl: TeleRing was very small data-set!

Results for Algorithm II



- Spearman's rho = .51
- p-value = .07
- Pearson can have negative values: outlier at ‚security': seed terms very different from human coders interpretation

Conclusion

Conclusion & Future Work

- Acceptable Validity: near human results
- Eliminates coding subjectivity: High Reliability
- Proposal: headcount for large corpora, termcount for smaller and more lengthy ones
- Future work:
  - fine tuning
  - Test with more data-sets
  - Ease applicability through provision of a software package
  - Ease Coding Construct Exploration: interpretable similarity value! (association strength?)

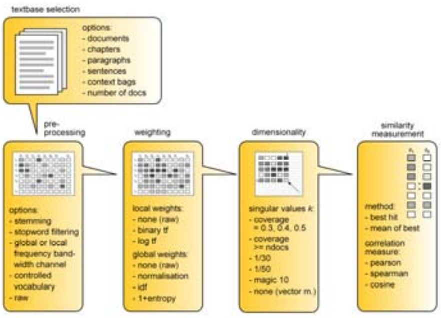
#eof.

Word Order Neglection?

- Educated adult understands ~100,000 word forms
- An average sentence contains 20 tokens.
- Thus $100{,}000^{20}$ possible combinations of words in a sentence
- $\therefore$ maximum of $\log_2 100{,}000^{20}$
  = **332 bits in word choice alone.**
- $20! = 2.4 \times 10^{18}$ possible orders of 20 words
  = maximum of **61 bits from order of the words.**
- 332/(61+ 332) = **84% word choice**

(Landauer, 2007)

LSA Process & Driving Parameters



Parameter Settings

- Stopwords filtered
- Minimum word length = 2
- Share of .5/.4/.3 of the cumulative singular values
- No background corpus
- Pearson Correlation as similarity measure
- No weighting