Does Latent Semantic Analysis Reflect Human Associations?

Tonio Wandmacher[1], Ekaterina Ovchinnikova[1] and Theodore Alexandrov[2]

[1]Institute of Cognitive Science
University of Osnabrück
{eovchinn@uos.de, twandmac@uos.de}

[2]Center for Industrial Mathematics
University of Bremen
theodore@math.uni-bremen.de

Lexical Semantics Workshop at ESSLLI 2008 – Hamburg

August 7, 2008

Disclaimer

1. We do not want to improve the methodology
2. We do not want to investigate human associative memory
3. We do not claim our results to be optimal (better look at J. Bullinaria's work)

Disclaimer

1. We do not want to improve the methodology
2. We do not want to investigate human associative memory
3. We do not claim our results to be optimal (better look at J. Bullinaria's work)

- But we do want to get a deeper understanding of what a distributional method like LSA can and cannot contribute to human semantic processing

Introduction & Method

LSA: Basics

Original goal: Improving Information Retrieval, nowadays used for nearly everything, e.g.

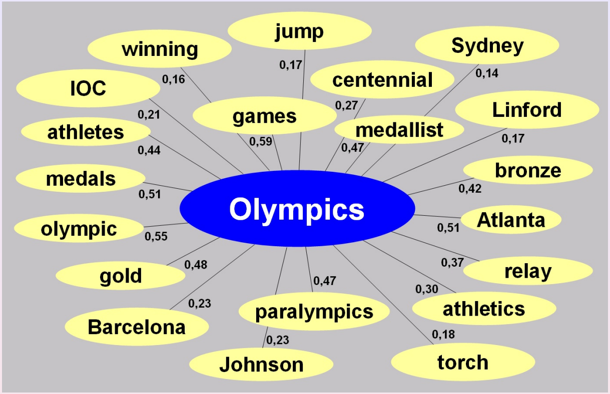
- Text summarization (Wade-Stein&Kintsch 2001)
- Cognitive Modelling (Landauer&Dumais 1997)
- Metaphor comprehension (Kintsch 2000)
- Evaluating student essays (Graesser et al. 2001)

Typical claims

"[. . .] *the similarity estimates derived by LSA are* **not simple contiguity frequencies**, **co-occurrence counts, or correlations in usage**, *but depend on a powerful mathemati- cal analysis that is capable of correctly inferring much* **deeper relations**." (Landauer et al. 1998, p. 263)

LSA: Example

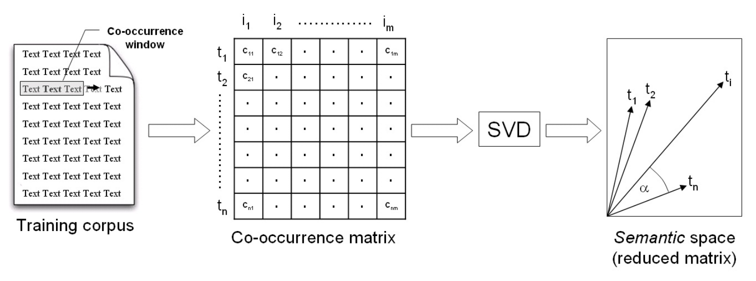
Nearest neighbors for '**Olympics**':



- All kinds of semantic and associative relations represented
- Looks quite like a human being's mind map
- So: **To what extent do LSA-generated similarities correspond to human associations?**

LSA: Method

- Based on a co-occurrence matrix (term×document or term×term)
- Reduction by *Singular Value Decomposition* (noise reduction, revealing "latent" structures)
- Comparing term vectors by cosine (length norm.)



LSA: Method (2)

- Using the **Infomap** toolkit (v. 0.8.6)[1], we construct a term$\times$term-matrix (80.000$\times$3.000)

Advantages of term$\times$term models

- Matrix does not grow with corpus ($\rightarrow$ more training data)
- Fixed co-occurrence window; no problems with defining 'document' (sentence?, paragraph?, book?)

- Training corpus: 108M words from *The Times* and *The Guardian* (1996 - 1998)
- SVD reduction to 300 dims
- Four spaces using co-occurrence windows of $\pm 5, \pm 25, \pm 50, \pm 75$ words

[1] http://infomap-nlp.sourceforge.net

Tasks and Results

Free Association Tasks

Free Association

First word(s) coming to a person's mind after being presented a *cue* word.

Three tasks:

1. **Discriminating** between three classes of association strengths
2. Measuring **correlation** between human association strengths and LSA similarity
3. **Predicting** the most frequent human response

- Datasets are based on the *Edinburgh Associative Thesaurus* (EAT)

Free Association Tasks

Task 1

Discrimination

Discrimination Task

Task

Discriminating between three classes of association strengths:

- FIRST: strongly associated pairs (>50% of the responses)
- HAPAX: association pairs produced by a single subject
- RANDOM: random cue-target combinations

Discrimination Task

Task

Discriminating between three classes of association strengths:
- FIRST: strongly associated pairs (>50% of the responses)
- HAPAX: association pairs produced by a single subject
- RANDOM: random cue-target combinations

- Results for 299 of the 300 suggested pairs:

	Right	**Wrong**	**Accuracy**
FIRST (th=0.23)	50	50	50%
HAPAX (th=0.02)	63	32	68%
RANDOM	68	17	78.2%
Total (F/H/R)	181	119	60.33%
HAPAX or RANDOM	189	11	94.5%
FIRST/HAPAX or RANDOM	239	61	**79.66%**

Free Association Tasks

Task 2

Correlation

Correlation Task

Task

Measuring correlation between human association strengths and LSA similarity.

- Cue-target pairs were selected by *stratified sampling* $\rightarrow$ human association strength [0,1] is uniformly distributed.

Correlation Task

Task

Measuring correlation between human association strengths and LSA similarity.

- Cue-target pairs were selected by *stratified sampling* $\rightarrow$ human association strength [0,1] is uniformly distributed.

Results for 239 of the 240 suggested pairs:

- *Pearson* correlation of **0.353**
- *Kendall* correlation coefficient of **0.263**
- Both are significant with a *p*-value<0.01

Correlation Task (2)

Contrasting results:

- In (Cramer, Waltinger & Wandmacher, *to appear*) we measured LSA correlation with *semantic relatedness* judgements by human subjects
- 35 subjects to rated 320 word pairs on a 5-level scale ("*Is X related to Y?*")
- LSA space was trained on 101M words from German newspaper text (*Süddeutsche Zeitung*)
- Here, a *Pearson* correlation of **0.62** was measured

Why are the results so different ?

Correlation Task (2)

Contrasting results:

- In (Cramer, Waltinger & Wandmacher, *to appear*) we measured LSA correlation with *semantic relatedness* judgements by human subjects
- 35 subjects to rated 320 word pairs on a 5-level scale ("*Is X related to Y?*")
- LSA space was trained on 101M words from German newspaper text (*Süddeutsche Zeitung*)
- Here, a *Pearson* correlation of **0.62** was measured

Why are the results so different ?

Well, we don't know yet ...

Free Association Tasks

Task 3
Response Prediction

Response Prediction

Task

Predicting the most frequent human response.

- Cue condition: Assoc. strength for 1^{st} response at least 3 times higher than for 2^{nd} response.
- Evaluation by calculating the *average LSA rank* of the correct response ($Rank_{min} = 100$).

Response Prediction

Task

Predicting the most frequent human response.

- Cue condition: Assoc. strength for 1^{st} response at least 3 times higher than for 2^{nd} response.
- Evaluation by calculating the *average LSA rank* of the correct response ($Rank_{min} = 100$).

- Results for 199 of the 200 suggested pairs
- The resulting average rank of the correct response is **51.89**

Target rank	1	2	3	4	5	6	7-99	100
Frequency	31	10	7	5	6	7	43	89

Co-occurrence Window

- Size of the co-occurrence window is a crucial factor for establishing semantic relatedness
- Previous works used rather "small" windows:

Lund & Burgess, 1996:	± 8 words
Rapp, 2002:	± 2 words
Cederberg & Widdows, 2003:	± 15 words
Peirsman, Heylen & Geeraerts, 2008:	± 1 to ± 10 words

Co-occurrence Window

- Size of the co-occurrence window is a crucial factor for establishing semantic relatedness
- Previous works used rather "small" windows:

Lund & Burgess, 1996:	± 8 words
Rapp, 2002:	± 2 words
Cederberg & Widdows, 2003:	± 15 words
Peirsman, Heylen & Geeraerts, 2008:	± 1 to ± 10 words

- We created 4 spaces using $\pm 5, \pm 25, \pm 50, \pm 75$ words

	± 5	± 25	± 50	± 75
Correlation (r)	0.254	0.326	0.347	**0.354**
Disc. (Acc.)	54.67	55.67	58.67	**60.33**
Pred. (Av. Rank)	62.61	54.11	52.69	**51.89**

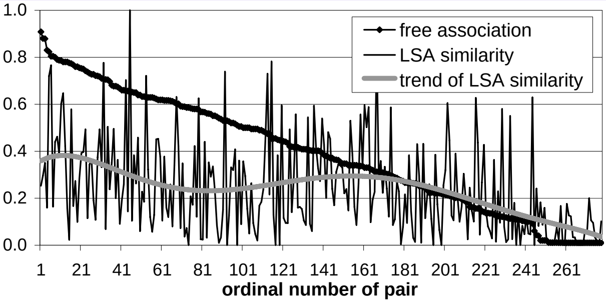
- Larger windows seem to work better in all 3 tasks
- Similar results for word prediction
- Closer investigation needed ...

Disagreement Analysis

Quantitative Analysis

Correlation Analysis

- Plotting human and LSA similarity values for correlation data (ordered by human values)

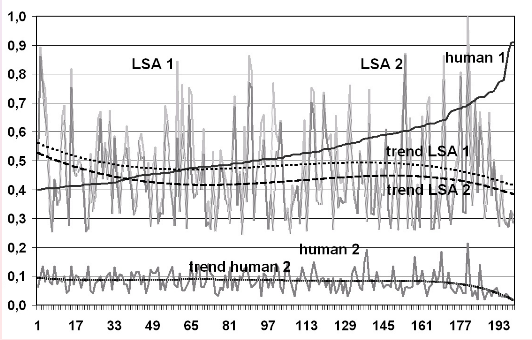


- LSA easily establishes low association
- No correlation for higher half of human values
  $\rightarrow$ Correct estimation of high association strength is complicated

Prediction Analysis

- **Human**: Strength of 1st response at least 3 times higher than strength of 2nd
  LSA: No large difference between 1st and 2nd values
- **Human**: When the 1st response strengths ($\geq$0.65) increase, the strengths of the 2nd responses decrease
  LSA: no such effect observed

Human response strengths vs. LSA values for 1st and 2nd response



Relation Analysis

- Word pairs of the prediction data were manually categorized wrt. their semantic relation; average rank was determined for each sub-group

relation	av. rank	# pairs	example
opposition	24.42	31	*female, male*
co-hyponymy	40.50	6	*july, august*
near-synonymy	46.98	47	*incorrect, wrong*
pred.-arg.	49	13	*eating, food*
hypo-/hyperonymy	53.32	22	*finch, bird*
mero-/holonymy	58.43	21	*deck, ship*
topical rel.	62.65	31	*prefect, school*
collocation	77.59	17	*wizard, oz*
attribute-class rel.	85.86	7	*sugar, sweet*

- LSA is good at predicting oppositions and co-hyponyms

Disagreement Analysis

Qualitative Analysis

Pairs with the highest LSA values

cue	**target**	**human**	**LSA**
ha	ha	0.66	1.00
inland	revenue	0.31	0.84
four	five	0.45	0.78
question	**answer**	0.71	0.78
good	**bad**	0.80	0.77
grammar	school	0.53	0.74
below	**above**	0.47	0.73
daughter	**son**	0.63	0.72
vehicle	car	0.82	0.72
parish	church	0.66	0.70
boy	**girl**	0.78	0.65

- LSA strongly associates oppositions
  (15 from 19 oppositions found in the correlation task data sets have got LSA values $>$0.22)

Typical disagreements

- LSA similarity is symmetric, human associations are not

cue	**target**	**human**	**LSA**
wrong	*right*	0.72	0.493
right	*wrong*	0.42	0.493

- LSA is corpus-dependent (here: newspaper ads)

cue	**target**	**human**	**LSA**	**NN for cue**
fresh	*lobster*	0.01	0.2	*flavour* (0.39) *soup* (0.37) *vegetables* (0.37) *potato* (0.36) *chicken* (0.36)

Best assoc. from EAT for '*fresh*': *air, fish, new, stale, fresher*

Typical disagreements (2)

- LSA underestimates associations between concepts and their salient properties

cue	target	human	LSA	NN for cue
snow	*white*	0.408	0.09	*snowfalls* (0.65) *winds* (0.624) *weather* (0.612) *slopes* (0.61) *temperature* (0.608)

Typical disagreements (3)

- LSA generates neighbors of the prominent meaning only and suppresses other domains
  (cf. also Wandmacher, 2005)

cue	target	human	LSA	NN for cue
nurse	*hospital*	0.156	0.627	*nurses* (0.64) *hospital* (0.627) *patient* (0.597) *doctors* (0.554) *patients* (0.525)

cue	target	human	LSA	NN for cue
eve	*adam*	0.567	0.024	*christmas* (0.657) *festive* (0.535) *yuletide* (0.456) *festivities* (0.453) *presents* (0.408)

Typical disagreements (4)

- LSA underestimates collocations that consist of elements without meaning overlaps

cue	target	human	LSA	NN for cue
shotgun	*wedding*	0.402	0.06	*gun* (0.54) *pistol* (0.50) *shooting* (0.46) *shotguns* (0.45) *firearms* (0.44)

Conclusion

- LSA is able to predict human associations to some extent, but it does not account for all aspects of associative memory
- LSA estimates for weakly associated terms are much closer to those of humans than for strongly associated terms
- Larger co-occurrence windows (of around ± 75 words) provide better results in all tasks
- Oppositions and co-hyponyms are rather well predicted, however collocations and attribute-class relations are not
- Most disagreements between LSA and human evaluations seem to be corpus-related
- *Polysemous terms*: LSA generates neighbors of the prominent meaning only and tends to suppress other domains

Future work

- Solving the question why the results for the 'semantic relatedness' data were so different
- Trying out different (and larger) corpora
- Performing these tasks using more techniques (dimension reduction, order of similarity etc.)

The end

The end

Thank you for your attention!

The end

Thank you for your attention!

Questions?

Our results can be found on:
www.ikw.uos.de/∼twandmac/FA-Results-WOA.zip