Room for Improvement?

Upper limits for collocation extraction with statistical association measures

Collocations and statistical association

- **Collocations** (Firth 1957) are pairs of words (such as *day* and *night* or *cow* and *milk*) that show a strong tendency to occur close to each other (i.e. to **co-occur**), in terms of
  - **surface proximity** (e.g. within a distance of five words)
  - **textual segments** (e.g. sentence, paragraph, Web page)
  - a **syntactic relation** (e.g. adjective + noun, verb + direct object)
- Such attraction between words can be quantified by (statistical) **association measures** (**AM**)
  - AMs compare the **observed** co-occurrence **frequency *O*** in a **corpus** with the **expected frequency *E*** under independence assumptions (as if the words were distributed at random)

Simple association measures

$$\text{MI} = \log_2 \frac{O}{E} \qquad \text{MI}^k = \log_2 \frac{O^k}{E} \qquad \text{local-MI} = O \cdot \log_2 \frac{O}{E}$$

$$\text{z-score} = \frac{O-E}{\sqrt{E}} \qquad \text{t-score} = \frac{O-E}{\sqrt{O}} \qquad \text{simple-ll} = 2\left(O \cdot \log \frac{O}{E} - (O-E)\right)$$

- In theoretical linguistics, collocations are treated as an **epiphenomenon** with a variety of underlying causes:
  - idioms (*red herring, kick the bucket*)
  - compounds, terms (*bus stop, support vector machine*)
  - lexical collocations (*commit a crime*)
  - semantic families (*day, night, time, year*)
  - cultural stereotypes & facts of life (*bucket and spade*)

Dossier: Collocates of *bucket* (noun)

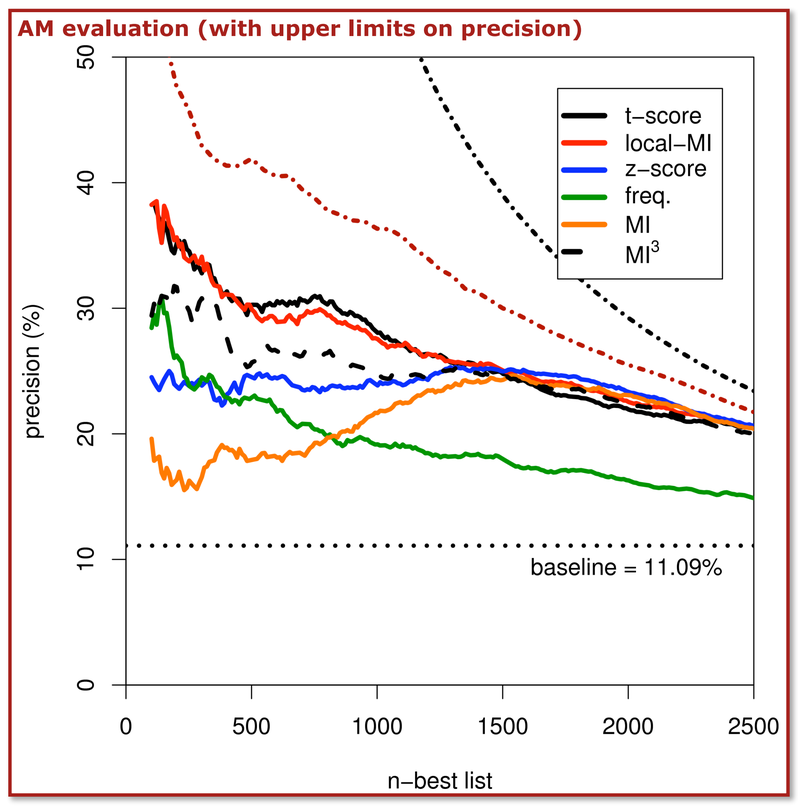
noun	*f*	local MI	verb	*f*	local MI	adjective	*f*	local MI
water	183	1023.77	throw	36	168.87	large	37	114.79
spade	31	288.11	fill	30	139.45	single-record	5	64.53
plastic	36	225.83	empty	14	96.73	full	21	63.23
size	41	195.89	randomize	9	96.11	cold	13	55.52
record	38	163.95	hold	31	78.93	small	21	45.61
slop	14	162.62	put	37	77.96	galvanized	4	43.47
mop	16	155.47	carry	26	71.95	ten-record	3	40.17
ice	22	125.76	tip	10	59.30	empty	9	38.41
bucket	18	125.49	kick	12	59.28	old	20	35.67
seat	21	89.21	chuck	7	44.85	steaming	4	31.89
coal	16	77.25	use	31	42.31	clean	7	27.47
density	11	63.64	weep	7	41.73	leaky	3	25.91
brigade	10	62.31	pour	9	40.73	wooden	6	25.50
sand	12	61.32	take	42	37.57	bottomless	3	25.17
algorithm	9	60.77	fetch	7	35.13	galvanised	3	24.70
shop	17	59.49	get	46	34.73	big	12	23.86
container	10	59.10	douse	4	33.03	iced	3	22.62
champagne	10	56.79	store	7	31.82	warm	6	19.55
shovel	7	56.50	drop	10	31.49	hot	6	17.05
oats	7	54.93	pick	11	28.89	pink	3	11.15

- Collocations are fundamental to **lexical priming** theories of language (Hoey 2005). From a psychological point of view, they represent **cognitively salient** patterns in the linguistic experience of a learner (Lund & Burgess 1996).

Which association measure?

- A large number of AMs have been proposed
  - see **www.collocations.de/AM** for a comprehensive listing
  - standard mathematical arguments are fruitless, and often not valid for linguistic data (cf. Dunning 1993; Evert 2004, Ch. 4)
- Typical application of AMs: **multiword extraction**
  - candidate word pairs with sufficiently high association scores are identified as potential (lexicalised) **multiword expressions** (**MWE**)
  - cutoff threshold often determined implicitly to give **n-best set**
  - **evaluation** of AMs: extracted MWEs are validated manually, resulting in precision and recall values for each AM and data set
  - evaluation example: PP-verb combinations from *Frankfurter Rundschau corpus*, extraction based on chunk-parsed data (Evert 2004, Ch. 5)
  - 5102 candidate pairs with f ≥ 30
  - **true positives** (TP) are FVG (*in Frage stellen*) and figurative expressions (*über die Bühne gehen*)
  - manual annotation by Brigitte Krenn

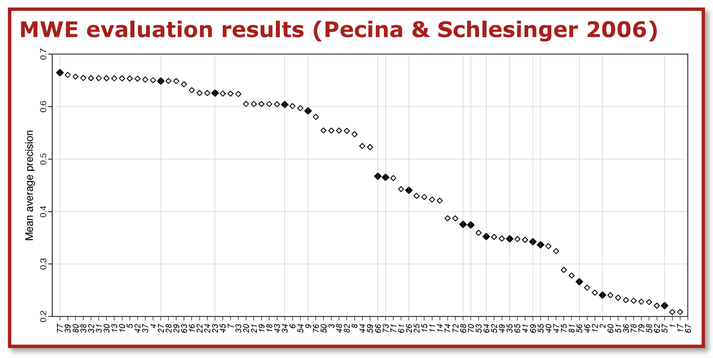
AM evaluation (with upper limits on precision)



A sonic barrier?

- Most evaluation studies have found many AMs with similar performance
- Best-performing group often with simple AM
- New AM equations have given no substantial improvement

MWE evaluation results (Pecina & Schlesinger 2006)



Two key questions

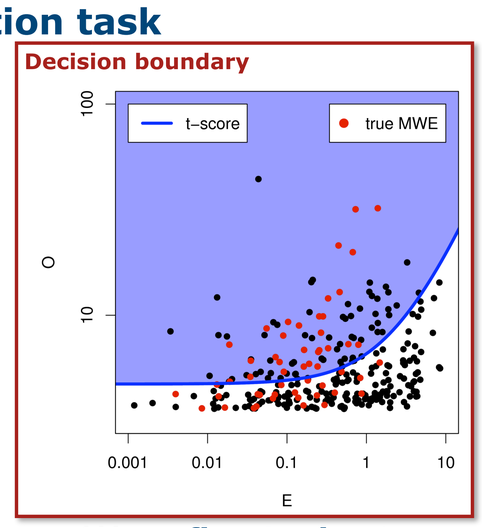
(A) What might significantly better AMs look like?

(B) How much room for improvement is there?

Learning optimal association measures

- Multiword extraction as a **classification task**
  - AM can be seen as a function ***g(E,O)*** that assigns an association score to each data point = word pair (Evert 2004)
  - after threshold application, this becomes a **binary classifier** (+/− MWE) on a 2-dimensional real-valued feature space
  - **decision boundary** is determined by the implicit equation *g(E,O) = C*
- Use **machine learning** techniques to find optimal classifier
  - supervised learning (with gold standard = manual annotation from evaluation)
  - allows **development** of new general-purpose AMs or **fine-tuning** to a specific task and data set (trained on sample, Evert & Krenn 2005)

Decision boundary



- Problems of the machine-learning approach
  - a **model bias**, i.e. a restriction on the shapes of allowed decision boundaries, is needed to avoid overtraining (➡ poor generalisation)
  - data not separable by simple boundaries ➡ soft margin methods
  - standard models (e.g. SVM with polynomial kernel) are too restricted and learned classifier does not match intuitions about collocativity
- Evert (2004) suggests two **soundness conditions**
  - if *O* is increased, *g(E,O)* must also increase (for fixed *E*), $\partial g/\partial O > 0$
  - if *E* is increased, *g(E,O)* must decrease (for fixed *O*), $\partial g/\partial E \leq 0$

➡ decision boundary is a **simple, monotonically increasing curve**

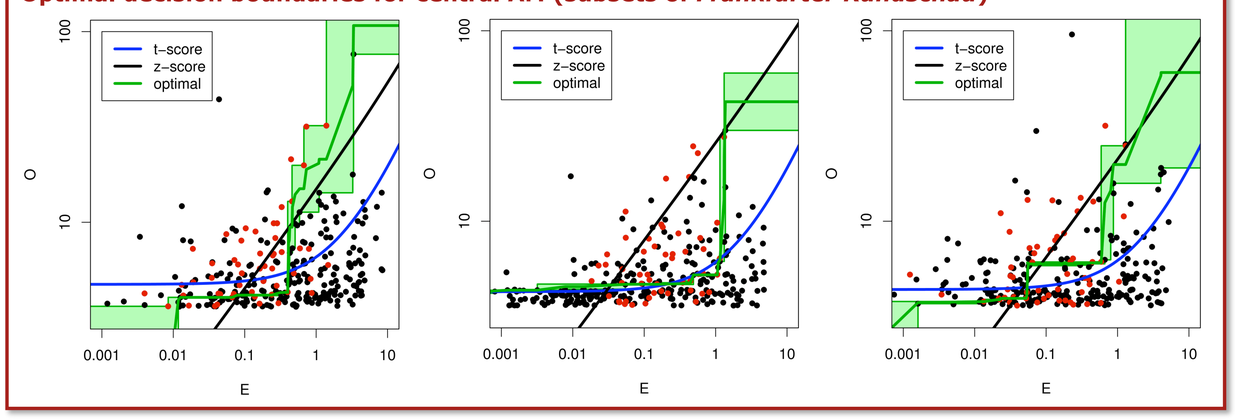
- use soundness as intuitive model bias and allow overtraining

➡ answer to question (B) and suggestions for new AM equations (A)

Preliminary results

- Upper limits on MWE extraction performance
  - shown in evaluation graph on the left ☜
  - broken red line = upper limit for performance of simple AM
  - brocken black line = fundamental limit for perfect separation
- What do the decision boundaries of better AMs look like?
  - optimal decision boundaries for three subsets of the *Frankfurter Rundschau* corpus are shown below (➡ sampling variation)
  - optimal boundary is not unique (shaded area = possible boundaries
  - general pattern: sharp bend close to *E* = 1

Optimal decision boundaries for central AM (subsets of *Frankfurter Rundschau*)



Stefan Evert
Institute of Cognitive Science
University of Osnabrück
http://purl.org/stefan.evert

References

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, IMS, University of Stuttgart. URN urn:nbn:de:bsz:93-opus-23714.

Evert, S. and Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, **19**(4), 450–466.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, pages 1–32. The Philological Society, Oxford.

Hoey, M. (2005). *Lexical Priming: A new theory of words and language*. Routledge, London.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, **28**(2), 203–208.

Pecina, P. and Schlesinger, P. (2006). Combining association measures for collocation extraction. *COLING/ACL 2006 Poster*, pages 651–658, Sydney, Australia.