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** (AMs) that 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)

- **In theoretical linguistics, collocations are treated as an epiphenomenon with a variety of underlying causes:**
  - idioms (red herring, kick the bucket)
  - compound 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)

- **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](http://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

AM evaluation of AMs: extracted MWEs are validated manually, resulting in precision and recall values for each AM and data set

- **decision 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 PVC (in Frage stellen) and figurative expressions (über die Bühne gehen)
- manual annotation by Brigitte Krenn

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

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 $g(E,O) = O\cdot E$
- **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)

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
- 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 $E$)
  - If $E$ is increased, $g(E,O)$ must decrease (for fixed $O$)
- 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
  - broken 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 $O = 1$

References


Postscript: pages 405-413.

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