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):
  - 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)

\[
MI = \log_2 \frac{O}{E} \quad \text{MI}^2 = \log_2 \frac{O^2}{E} \quad \text{local-MI} = O - \log_2 \frac{O}{E}
\]

\[
z\text{-score} = \frac{O - E}{\sqrt{E}} \quad t\text{-score} = \frac{O - E}{\sqrt{O}} \quad \text{simple-S} = 2 \left( O - \frac{O^2}{E} - (O - E) \right)
\]

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

- **A sonic barrier**?
  - most evaluation studies have found many AMs with similar performance
  - best-performing group often includes a simple AM
  - new AM equations give no substantial improvements

**Two key questions**

- **(A)** What might significantly better AMs look like?
- **(B)** How much room for improvement is there?

**The role of machine learning**

- **Multiword extraction as a classification task**
  - AM can be understood as a function \( g(E,O) \) that assigns an association score to each data point = word pair (Evert 2004, Sec. 3.3)
  - after threshold application, this becomes a binary classifier (+/- MWE) on a two-dimensional real-valued feature space
  - **decision boundary** is determined by the implicit equation \( g(E,O) = C \)

- **Idea**: can use general machine learning techniques for this task supervised learning, using manual annotation from evaluation as gold standard
  - for 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**
  - very easy to overtrain the model (poor generalisation)
  - data not separable by simple boundary
  - learned classifier may not match intuitions about collocativity

- **Can give partial answer to (A), but no answer to (B)**

**Soundness and upper limits**

- Evert (2004) suggests two **soundness conditions**
  - If \( O \) is increased, \( g(E,O) \) must also increase (for fixed \( O \) / \( g(E,O) > 0 \))
  - If \( E \) is increased, \( g(E,O) \) must decrease (for fixed \( O \) / \( g(E,O) < 0 \))

- capture basic intuitions about statistical association and collocativity

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

- **Employ soundness conditions as intuitive model bias**, and allow model to overtrain = **theoretical upper limits**

  - compare precision of current AMs to upper limits = answer to question (B)
  - optimal decision boundaries suggest possible new AM equations = partial answer to (A), complementary to ML
  - translation of boundaries into sound AM equation counteracts overtraining!

- **Preliminary results**
  - show some room for improvement
  - current implementation is limited to small data sets (= 1500 candidates)

- **Work in progress**
  - cross-validation (on different subsets of a data set as well as different subsets of a corpus) to assess degree of overtraining
  - test applicability in fine-tuning setting
  - **practical machine learning of sound AMs**: linear classifier using association scores of prototypical sound AMs as features
  - new AMs (= shapes of decision boundaries) suggested by upper-limit experiments will be crucial for successful machine learning

---

References


References

http://www.cogsci.uni-osnabrueck.de/~CL/