Aleksandra Piwowarek
15 October 2012

COLLOCATIONS PART 2

ANOTHER METHOD FOR DISCOVERING COLLOCATIONS:

Pointwise mutual information in NLP -

for events x' and y':

(1) $I(x',y') = \log_2 \frac{P(x'y')}{P(x')P(y')}$

(2) $= \log_2 \frac{P(x'|y')}{P(x')}$

(3) $= \log_2 \frac{P(y'|x')}{P(y')}$

roughly a measure of how much words tell us about each other or:

> "The amount of information provided by the occurrence of the event represented by [y'] about the occurrence of the event represented by [x']" (Book, #).

more precisely, this measures the reduction of uncertainty about the occurrence of one word when we are told abou the occurrence of another, or how certain we can be that x' will occur given what we know about y'

Example:

$$I(\text{Ayatollah, Ruhollah}) = \log 2 \frac{(20/14307668)}{(42/14307668) \times (20/14307668)} = 18.38$$

P*roblems* with this method include:

- not a good measure of what an interesting correspondance b/w events is
- even with large corpuses, innacurate maximum likelihood estimates and artificially inflated mutual information scores can occur
- is not particularly good with low frequency events
- refers to something else in information theory - the expectation of the quanity:

$$I(X;Y) = \mathfrak{S}_{P(x,y)} \log_2 \frac{P(X,Y)}{P(X)P(Y)}$$

MORE ON COLLOCATIONS

Collocations:

various **definitions** including:

> "A sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components" (Choueka, 1988).

- need not necessarily be a consecutive phrase (ex: knock... door)

- typical **criteria** for a collocation include:

> - **Non-compositionality:** the meaning of a collocation cannot be derived from the meaning of its parts. The meaning may be either completely different than the sum of its parts (ex: kick the bucket) or may have an added connotation that cannot be predicted from the parts (ex: "white wine", "white hair", "white woman", for which each "white" has a slightly different meaning.)
>
> - **Non-substitutability**: it is impossible to other words for the components of a collocation, even if the meaning is the same in context. (ex: the "white" in "white wine" cannot be substituted with "yellow", yielding "yellow wine", even though the description is just as accurate, given that white wine is yellowish white in colour).
>
> - **Non-modifiability**: cannot be freely modified with additional lexical material or through grammatical transformations (ex: "a frog in one's throat" cannot be modified to produce "an ugly frog in wone's throat" even though nouns like frog can usually be modified by adjectives like ugly).

- can often test for a collocation by translating it into another language word by word.

> Ex: English "make a decision" into French word by word is "faire une décision", while the correct combination would be "prendre une décision". The phrase is most likely a collocation.

- Some authors suggest that **collocation** be used to include words that are strongly associated with one another, but do not necessarily occur a common grammatical unit with a certain order (ex: nurse, doctor). The book instead suggests the vocabulary **association** and **co-occurence** for words likely to be used in the same context, recommending the earlier, more narrow definition of collocation.

Subclasses of Collocation

- **light verbs** – verbs with little semantic context in collocations
    ex: “make”, “take” or “do” in collocations like “make a decision”

- **verb particle constructions** or **phrasal verbs –** combination of a main verb and a particle, and often correspond to a single lexeme in other languages
    ex: “to tell off” in English (compare to “réprimander” in French)
- **proper nouns** or **proper names** – often included in the category of collocations in computational work, despite being different from lexical collocations; useful in approaches that look for fixed phrases that appear in the same form throughout a text

- **terminological expressions or phrases** - refer to concepts and objects in technical domains, and are often fairly compositional; it is useful to treat them as collocations to ensure their being treated consistently throughout a text
    ex: hydraulic oil filter

All information taken more or less directly from: “Foundations of statistical natural language processing” by Christopher Manning and Hinrich Schuetze.