Formal properties of word pair correlations

March 25, 2015

1 Definition

The "word pair correlation function" can be defined for a single wordtype, or across all words.

Across all words, let's define it as the probability a pair of tokens, spaced r distance apart, are the same word. So if you picked a position t at random, what's the chance the same word is at $t + r$? (Restriction: you sample t only from positions $\{1..N - r\}$ where N is the number of tokens in the corpus.) This is:

$$g(r) = p(w_{t+r} = w_t)$$

For a specific word v, there are two possible ways to define it. We should standardize on one.

$$g_{cond}(r, v) = p(w_{t+r} = v \mid w_t = v)$$

versus

$$g_{joint}(r, v) = p(w_{t+r} = v,\ w_t = v)$$

the joint version (the latter) has much smaller numerical values than the first, and they're especially smaller for rare words. Perhaps that is an argument the conditional form is easier to interpret.

Finally, this could be defined for a set of words — for example, to analyze a single paircorr function curve for all names, or all verbs, etc. Let the set of words be A. Here's a proposal:

$$g(r, A) = p(w_{t+r} \in A \mid w_t \in A)$$

this is equivalent to rewriting all words in the set A to a special symbol, and computing the word paircorr for that symbol (specifically g_{cond}).

2 Pair correlation under Markov models

Consider the case $r = 1$. $g(1)$ can be rewritten in a history conditional form as follows.

$$\begin{aligned} g(1) &= p(w_{t+1} = w_t) && (1)\\ &= \sum_v p(w_{t+1} = v,\ w_t = v) && (2)\\ &= \sum_v p(w_{t+1} = v \mid w_t = v)\, p(w_t = v) && (3) \end{aligned}$$

Step 2 decomposes the probability into the sum over all words that are possible. Step 3 is definition of conditional probability.

A zeroth order (independent unigrams) assumption yields the following equivalency:

$$p(w_{t+1} = v \mid w_t = v) = p(w_{t+1} = v) = p(w_t = v)$$

With this, we can simplify step 3 and fully derive $g(r)$ under the zeroth order assumption.

$$\sum_v p(w_{t+1} = v \mid w_t = v)\, p(w_t = v) \tag{4}$$

$$= \sum_v p(w_{t+1} = v)\, p(w_t = v) \tag{5}$$

$$= \sum_v [p(w_t = v)]^2 \tag{6}$$

$$= g(r) \tag{7}$$

Now let's consider a first-order (bigram) model. For $r = 1$, $g(r)$ is already defined for each word type. So:

$$p(w_{t+1} = w_t) = \sum_v p(w_{t+1} = v \mid w_t = v)$$

$$p(w_{t+2} = w_t) = \sum_v p(w_{t+2} = v, w_t = v) * p(w_t = v)$$

$$p(w_{t+2} = v, w_t = v) \tag{8}$$

$$= p(w_{t+2} = v \mid w_t = v) \tag{9}$$

$$= \sum_u p(w_{t+2} = v, p(w_{t+1} = u \mid w_t = v) \tag{10}$$

$$= \sum_u p(w_{t+2} = v \mid p(w_{t+1} = u, w_t = v) * p(w_{t+1} = u \mid w_t = v) \tag{11}$$

$$= \sum_u p(w_{t+2} = v \mid p(w_{t+1} = u) * p(w_{t+1} = u \mid w_t = v) \tag{12}$$

$$\sum_v p(w_{t+2} = v \mid w_t = v)$$

$$\sum_v p(w_{t+2} = v, p(w_{t+1} = v \mid w_t = v)$$

To do:

1. Derive $g(r)$ under a first-order model.
2. Derive $g(r)$ under an arbitrary order model .. or maybe just second order, that sounds easier.