**Bootstrapping** ... a walk through example.

**Goal:** company-product pairs  
**Start with:** “Apple” and “iphone”

1. **Count patterns** containing your word pair:

   Apple announces press conference for iPhone  
   Apple said for months they got something dope planned for the iphone  
   Apple iPhone review  
   Apple’s First iPhone Ad  
   Apple Launches iPhone

2. **Sort observed patterns by raw frequency** (we'll do this better later):

   84  X Y review  
   23  X Launches Y  
   8  X ’s First Y Ad  
   4  X announces press conference for Y  
   1  X said for months they got something dope planned for the Y

3. **Search for word pairs with the top n patterns** (often n=1)

   X Y review → (steel, watch), (the, month), (outdoor, green), (prom, dresses)  
   X launches Y → (hammer, new), (signature, new), (greenpeace, a)  
   X 's first Y ad → (GB, title), (year, real), (Asia Pacific, blog)

4. **These word pairs are terrible.**  
The best pattern “X Y review” doesn't match companies and products, but strings like “the year end review” or “top software review”. **Why? **Raw frequency is not a good metric.

5. **Use different patterns**, sorted by pointwise mutual information:

   \[ \text{pmi}("X Y review", \text{apple/iphone}) = \frac{P("X Y review", \text{apple/iphone})}{P("X Y review") \cdot P(\text{apple/iphone})} \]

   Note that “X Y review” is frequent, so the denominator will have a large P(“X Y review”). The pattern “X's first Y ad” will be smaller. PMI thus boosts good patterns that match your relation (numerator) but most importantly, that do not match other relations (denominator).

   9.1  X ’s First Y Ad  
   8.5  X announces press conference for Y  
   8.0  X Launches Y  
   5.4  X Y review
6. Search for new word pairs with the top \textbf{pmi patterns}. Sort by \textbf{PMI again}!

\text{pmi(wordpair, patterns) = } \frac{P(wordpair,patterns)}{P(wordpair)*P(patterns)}

7.6 (microsoft, windows)
6.9 (google, maps)
6.8 (apple, ipad)
… etc.

7. Integrate a “coverage” score as a complementary metric.

\textit{coverage} = “the fraction of patterns a word pair matched”

If you learned 10 patterns, and you extracted (apple, ipad) as a potential pair, how many of the 10 patterns did the pair occur with? If it matched 8 of 10, then coverage = 0.8.

A pair like (prom, dresses) will match many times to a pattern like “X Y review”, but it won’t match the other 9 patterns. This means its PMI score will be very high, due to lots of matches, but the coverage is very low showing lack of diversity.

\textbf{New Metric}: x*PMI + (1-x)*coverage
\textbf{New Filter}: PMI > minPMI && Coverage > minCoverage

8. \textbf{Cautious learning}.

Learn 1 or 2 patterns at a time at most, then learn 1 or 2 word pairs at most. Add them to your learned set, and repeat. This lets you learn high precision terms early, and then cautiously expand your knowledge.

9. \textbf{Noun phrases}.

Match bigrams or longer, not just unigrams. Use a tagger to identify names and organizations. Throw out parts of speech like determiners. Etc….