Using WordNet to Supplement Corpus Statistics

Rose Hoberman and Roni Rosenfeld

November 14, 2002

Data, Statistics, and Sparsity

- Statistical approaches need large amounts of data
- Even with lots of data long tail of infrequent events
  (in 100MW over half of word types occur only once or twice)
- Problem: Poor statistical estimation of rare events
- Proposed Solution: Augment data with linguistic or semantic knowledge
  (e.g. dictionaries, thesauri, knowledge bases...)

WordNet

- Large semantic network, groups words into synonym sets
- Links sets with a variety of linguistic and semantic relations
- Hand-built by linguists (theories of human lexical memory)
- Small sense-tagged corpus

WordNet: Size and Shape

- Size: 110K synsets, lexicalized by 140K lexical entries
  - 70% nouns
  - 17% adjectives
  - 10% verbs
  - 3% adverbs
- Relations: 150K
  - 60% hypernym/hyponym (IS-A)
  - 30% similar to (adjectives), member of, part of, antonym
  - 10% ...

WordNet Example: Paper IS-A ...

- paper $\rightarrow$ material, stuff $\rightarrow$ substance, matter $\rightarrow$ physical object $\rightarrow$ entity
- composition, paper, report, theme $\rightarrow$ essay $\rightarrow$ writing ... abstraction

  $\rightarrow$ assignment ... work ... human act
- newspaper, paper $\rightarrow$ print media ... instrumentality $\rightarrow$ artifact $\rightarrow$ entity
- newspaper, paper, newspaper publisher $\rightarrow$ publisher, publishing house $\rightarrow$ firm, house, business firm $\rightarrow$ business, concern $\rightarrow$ enterprise $\rightarrow$ organization $\rightarrow$ social group $\rightarrow$ group, grouping
- ...

This Talk

- Derive numerical word similarities from WordNet noun taxonomy.
- Examine usefulness of WordNet for two language modelling tasks:

1. Improve perplexity of bigram LM (trained on very little data)
   - Combine bigram data of rare words with similar but more common *proxies*
   - Use WN to find similar words
2. Find words which tend to co-occur within a sentence.
   - Long distance correlations often semantic
   - Use WN to find semantically related words

Measuring Similarity in a Taxonomy

- Structure of taxonomy lends itself to calculating distances (or similarities)
- Simplest distance measure: length of shortest path (in edges)
- Problem: edges often span different semantic distances
- For example:
  plankton IS-A living_thing
  rabbit IS-A leporid ... IS-A mammal IS-A vertebrate IS-A ... animal IS-A living_thing

Measuring Similarity using Information Content

- Resnik's method: use structure *and* corpus statistics
- Counts from corpus $\Rightarrow$ probability of each concept in the taxonomy $\Rightarrow$ "information content" of a concept.
- Similarity between concepts $=$ the information content of their least common ancestor: $sim(c_1, c_2) = -\log(p(lca(c_1, c_2)))$
- Other similarity measures subsequently proposed

Similarity between Words

- Each word has many senses (multiple nodes in taxonomy)
- Resnik's word similarity: max similarity between any of their senses
- Alternative definition: the weighted sum of $sim(c_1, c_2)$, over *all* pairs of senses c_1 of w_1 and c_2 of w_2, where more frequent senses are weighted more heavily.
- For example:
  TURKEY vs. CHICKEN
  TURKEY vs. GREECE

Improving Bigram Perplexity

- Combat sparseness $\longrightarrow$ define equivalence classes and pool data
- Automatic clustering, distributional similarity, ...
- But... for *rare* words not enough info to cluster reliably
- Test whether bigram distributions of semantically similar words (according to WordNet) can be combined to reduce the bigram perplexity of rare words

Combining Bigram Distributions

- Simple linear interpolation
- $p^s(\cdot|t) = (1-\lambda)p_{gt}(\cdot|t) + \lambda p_{ml}(\cdot|s)$
- Optimize lambda using 10-way cross-validation on training set
- Evaluate by comparing the perplexity on a new test set of $p^s(\cdot|t)$ with the baseline model $p_{gt}(\cdot|t)$.

Ranking Proxies

- Score each candidate proxy s for target word t
  1. WordNet similarity score: $wsim_{max}(t, s)$
  2. KL Divergence: $D(p_{gt}(\cdot|t) \,||\, p_{ml}(\cdot|s))$
  3. Training set perplexity reduction of word s, i.e. the improvement in perplexity of $p^s(\cdot|t)$ compared to the 10-way cross-validated model.
  4. Random: choose proxy randomly
- Choose highest ranked proxy (ignore actual scales of scores)

Experiments

- 140MW of Broadcast News
  - Test: 40MW reserved for testing
  - Train: 9 random subsets of training data (1MW - 100MW)
- From nouns occurring in WordNet:
  - 150 target words (occurred < 2 times in 1MW)
  - 2000 candidate proxies (occurred > 50 times in 1MW)

Methodology

for *each* size training corpus:

- Find highest scoring proxy for each target word and each ranking method
- Target word: ASPIRATIONS
  best Proxies: SKILLS DREAMS DREAM/DREAMS HILL
- Create interpolated models and calculate perplexity reduction on test set
- Average perplexity reduction: weighted average of the perplexity reduction achieved for each target word, weighted by the frequency of each target word in the test set

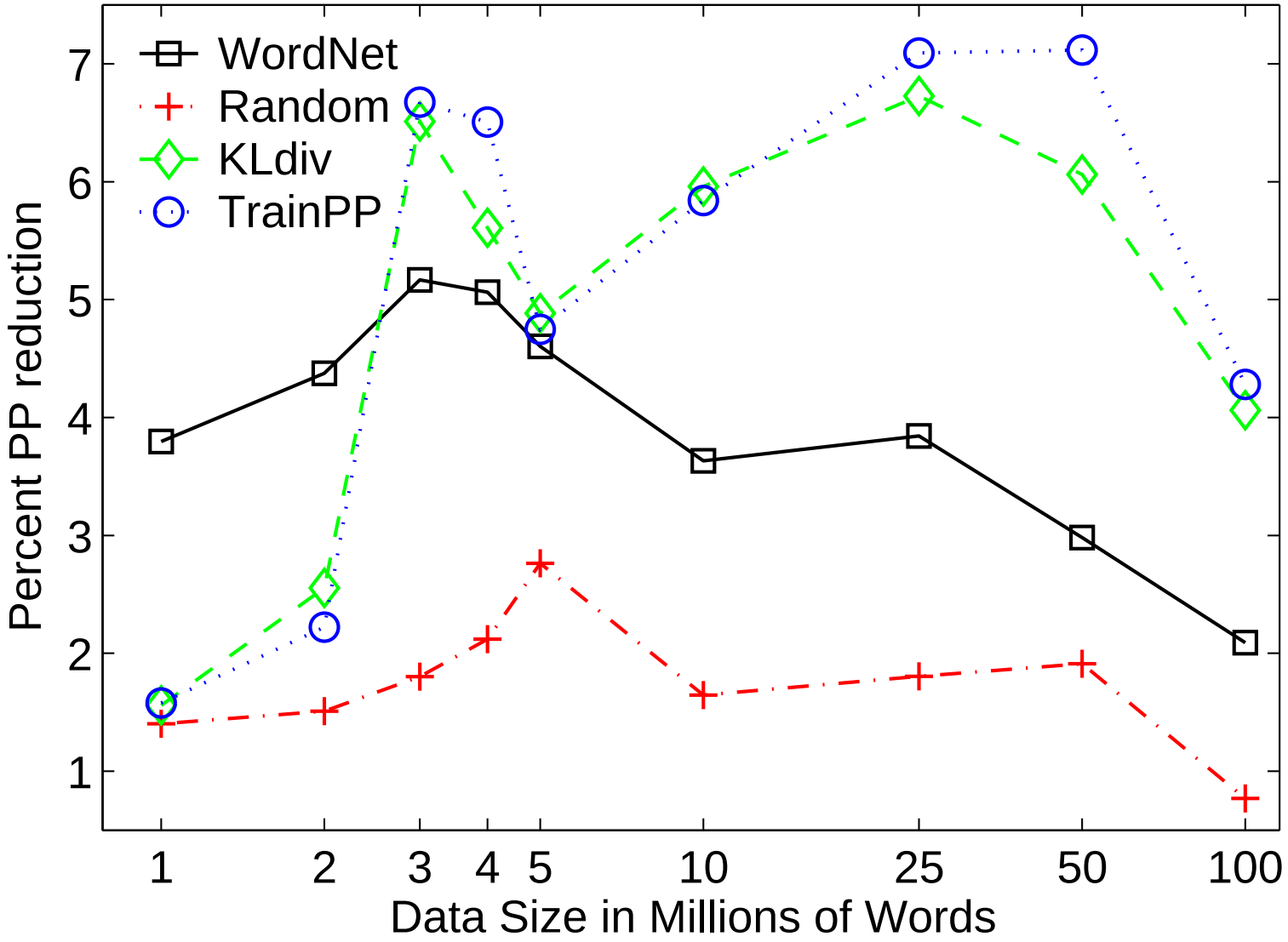


Figure 1: Perplexity reduction as a function of training data size for four similarity measures.

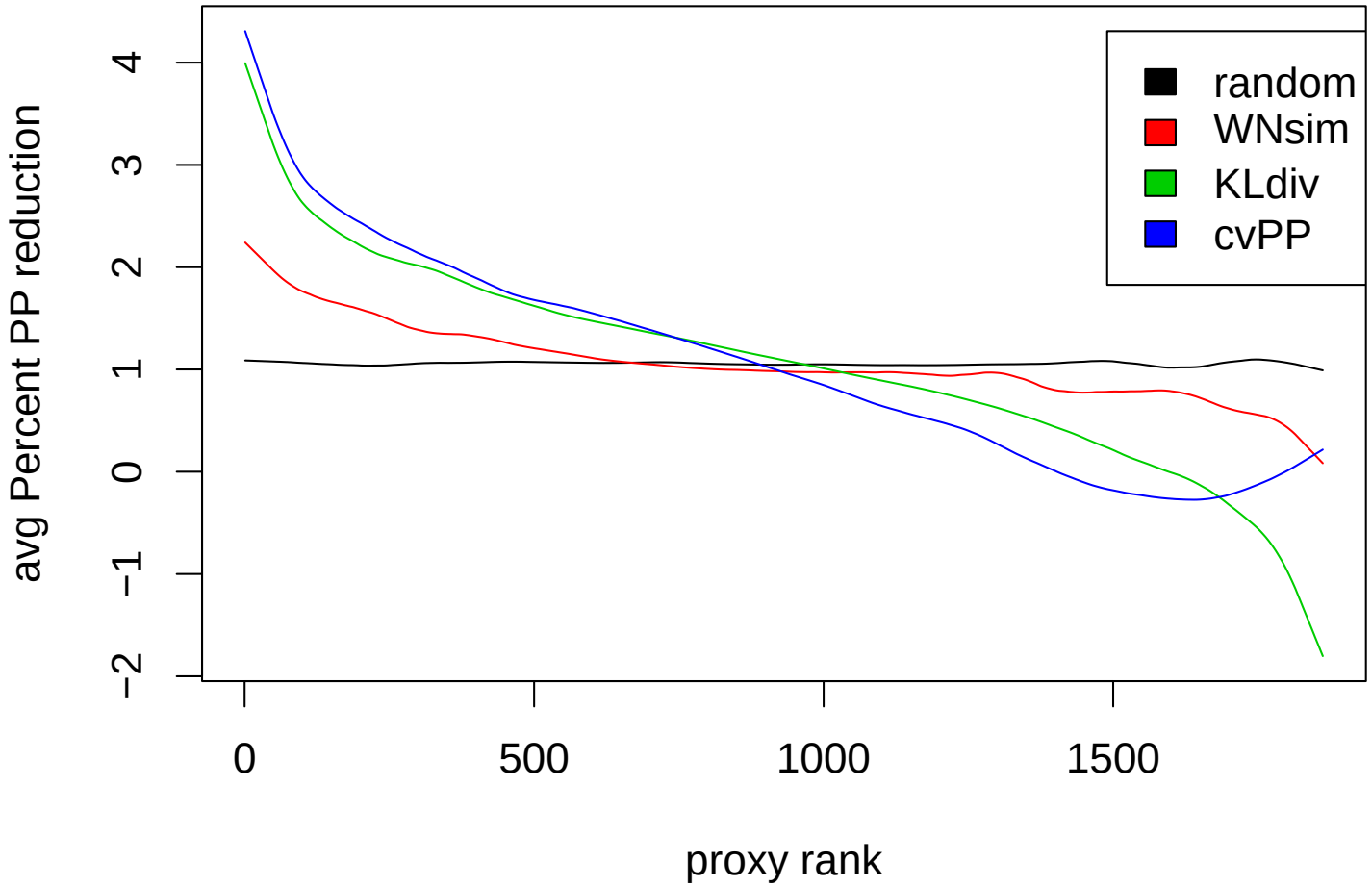


Figure 2: Perplexity reduction as a function of proxy rank for four similarity measures.

Error Analysis

%	Type of Relation	Examples
45	Not an IS-A relation	rug-arm, glove-scene
40	Missing or weak in WN	aluminum-steel, bomb-shell
15	Present in WN	blizzard-storm

Table 1: Classification of best proxies for 150 target words.

- Each target word$\Rightarrow$ proxy with largest test PP reduction $\Rightarrow$ categorized relation

- Also a few topical relations (TESTAMENT-RELIGION) and domain specific relations (BEARD-MAN)

Modelling Semantic Coherence

- N-grams only model short distances
- In real sentences content words come from same semantic domain
- Want to find long-distance correlations
- Incorporate semantic similarity constraint into exponential LM

Modelling Semantic Coherence II

- Find words that co-occur within a sentence.
- Association statistics from data only reliable for high frequency words
- Long-distance associations are semantic
- Use WN ?

Experiments

- "Cheating experiment" to evaluate usefulness of WN
- Derive similarities from WN for only *frequent words*
- Compare to measure of association calculated from *large* amounts of data. (ground truth)
- Question: are these two measures correlated?

"Ground Truth"

- 500,000 noun pairs
- Expected number of chance co-occurrences > 5
- Word pair association: (Yule's statistic) $Q = \frac{C_{11} \cdot C_{22} - C_{12} \cdot C_{21}}{C_{11} \cdot C_{22} + C_{12} \cdot C_{21}}$

	Word 1 Yes	Word 1 No
Word 2 Yes	C_{11}	C_{12}
Word 2 No	C_{21}	C_{22}

- Q ranges from -1 to 1

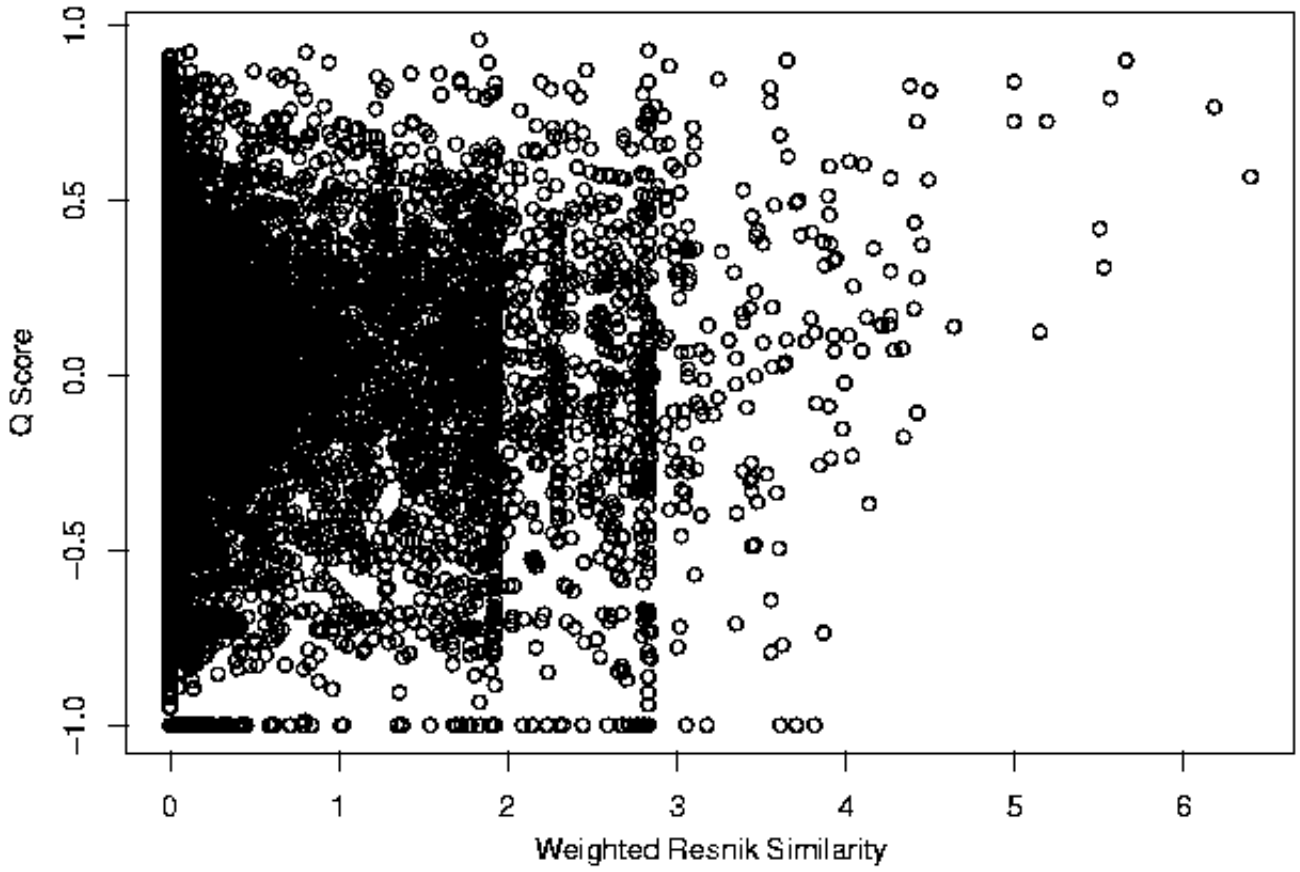
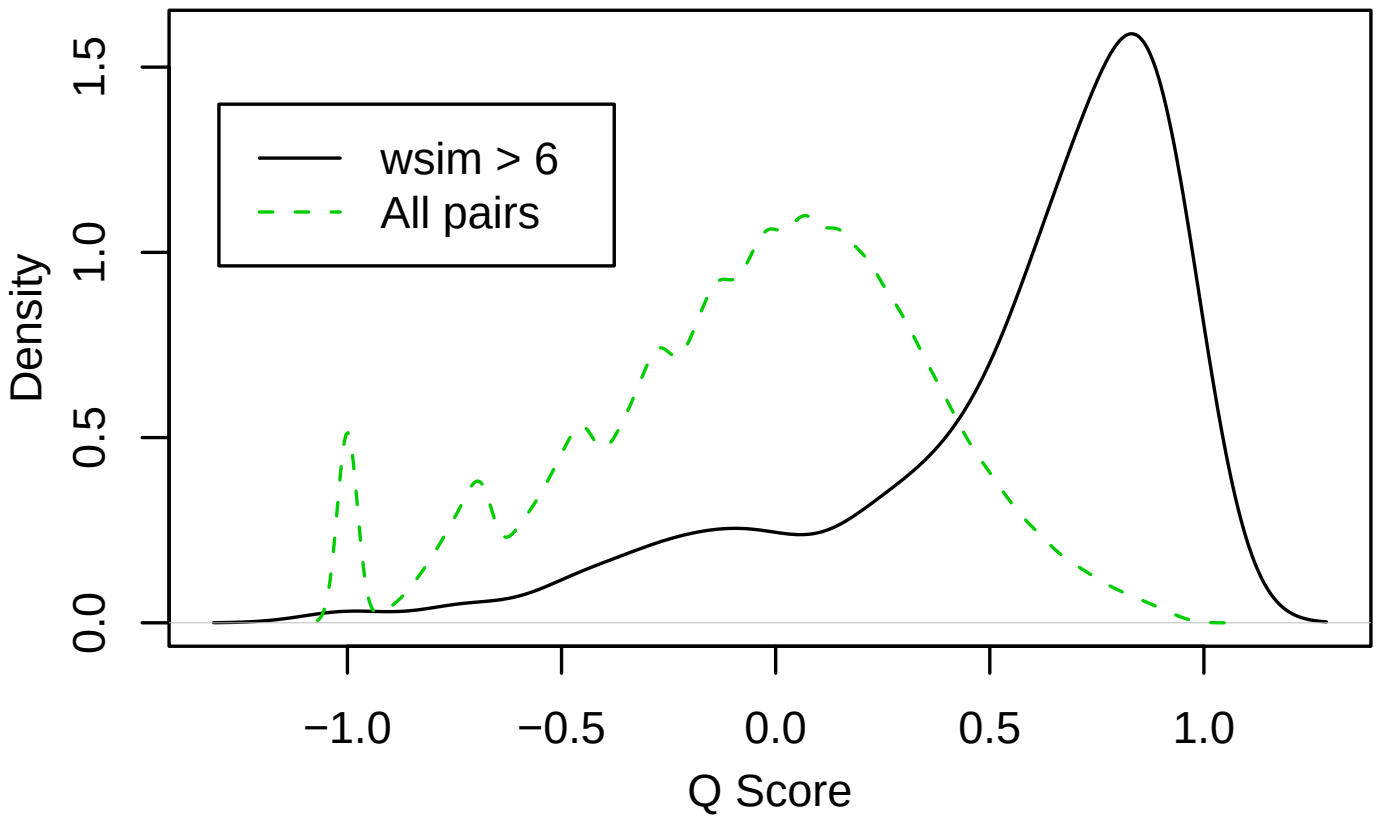


Figure 3: Looking for Correlation: WordNet similarity scores versus Q scores for 10,000 noun pairs



Only 0.1% of wordpairs have WordNet similarity scores above 5 and only 0.03% are above 6.

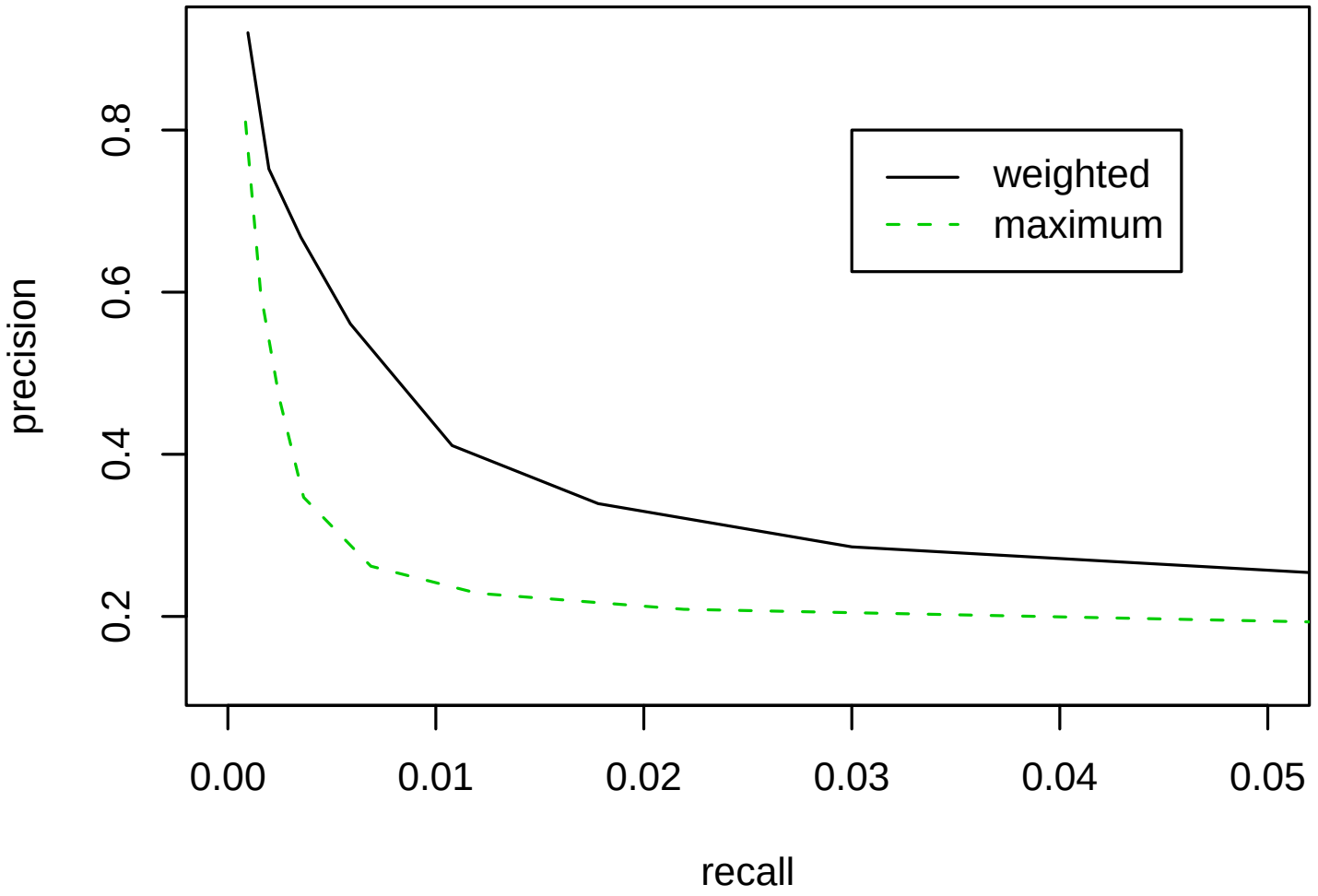


Figure 4: Comparing effectiveness of two WordNet word similarity measures

Relation Type	Num	Examples
WN	**277(163)**	
part/member	87 (15)	finger-hand, student-school
phrase isa	65 (47)	death tax IS-A tax
coordinates	41 (31)	house-senate, gas-oil
morphology	30 (28)	hospital-hospitals
isa	28 (23)	gun-weapon, cancer-disease
antonyms	18 (13)	majority-minority
reciprocal	8 (6)	actor-director, doctor-patient
non-WN	**461**	
topical	336	evidence-guilt, church-saint
news and events	102	iraq-weapons, glove-theory
other	23	END of the SPECTRUM

Table 2: Error Analysis

Conclusions?

- Very small bigram PP improvement when little data available
- Words with very high WN similarity do tend to co-occur within sentences,
- *However* recall is poor because most relations topical (but WN adding topical links)
- Limited types and quantities of relationships in WordNet compared to the spectrum of relationships found in real data
- WN word similarities weak source of knowledge for 2 tasks

Possible Improvements, Other Directions?

- Interpolation weights should depend on ...
  - data AND WordNet score
  - relative frequency of target and proxy word
- Improve WN similarity measure
  - consider frequency of senses but don't dilute strong relations
  - info content misleading for rare but high level concepts
  - learn a function from large amounts of data?
  - learn which parts of taxonomy are more reliable/complete?
- Consider alternative framework
  - class $\rightarrow$ word / word $\rightarrow$ class / class $\leftarrow$ word / word $\leftarrow$ class
  - provide WN with more constraints (from data)