Automated Grammar Induction

Experimental Results

Linas Vepštas

AGI 2022

19 August 2022

Experimental Results

Word-pair Mutual Information

Basic definitions:

- Word Pair: (u, w)
- Count: $N(u, w)$
- Frequentist probability: $p(u, w) = N(u, w) / N(*, *)$
- Star == wildcard sum over all entries in that location
- Lexical Attraction (MI):

$$MI(u, w) = \log_2 \frac{p(u, w)}{p(u, *)\, p(*, w)}$$

- Not symmetric: $(u, w) \neq (w, u)$

Experimental Results

Characterizing Word–Pair Data Sets

Sparse matrix with global properties

- Log width and height: $\log_2 N_L$ and $\log_2 N_R$
- Log total number of nonzero entries: $\log_2 D_{\mathrm{Tot}}$
- Log total number of observations: $\log_2 N_{\mathrm{Tot}}$
- Sparsity: $-\log_2 D_{\mathrm{Tot}}/N_L \times N_R$
- Rarity: $\log_2 D_{\mathrm{Tot}}/\sqrt{N_L \times N_R}$ is independent of dataset size!
- Entropy: $H_{\mathrm{Tot}} = \sum_{w,v} p(w,v) \log_2 p(w,v)$
- Marginal Entropy: $H_{\mathrm{Left}} = \sum_{w} p(w,*) \log_2 p(w,*)$
- Total MI:

$$MI = H_{\mathrm{Tot}} - H_{\mathrm{Left}} - H_{\mathrm{Right}} = \sum_{w,v} p(w,v) \log_2 \frac{p(w,v)}{p(w,*)\, p(*,v)}$$

Experimental Results

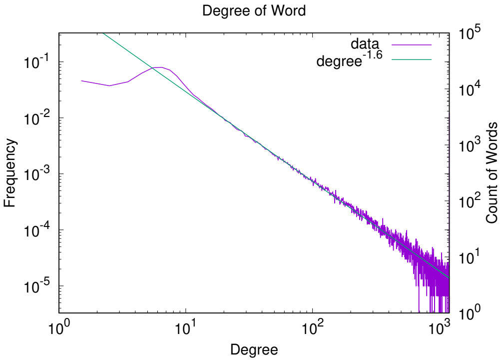
Example Word-Pair Data Sets

Corpus	1	2	3	4	5
$\log_2 N_L$	16.678	17.097	18.214	18.600	19.019
$\log_2 N_R$	16.690	17.117	18.228	18.620	19.039
$\log_2 D_{\mathrm{Tot}}$	23.224	23.797	24.748	25.180	25.627
Sparsity	10.144	10.416	11.693	12.040	12.431
Rarity	6.540	6.690	6.527	6.570	6.598
$\log_2 N_{\mathrm{Tot}}/D_{\mathrm{Tot}}$	4.779	5.079	5.128	5.235	5.335
Total Entropy	17.827	17.889	18.378	18.503	18.631
Left Entropy	9.7963	9.8102	10.069	10.109	10.148
Right Entropy	9.5884	9.5463	9.8321	9.8801	9.9265
MI	1.5572	1.4677	1.5227	1.4863	1.4431

Experimental Results

Sample Size Effects

Vertex degree: For word w, how many pairs (u, w) is it in?

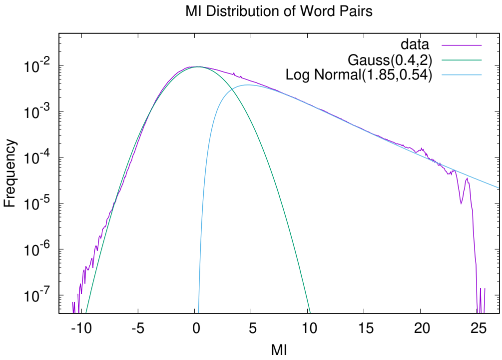


- Zipfian, with exponent $\gamma \approx 1.6$.
- Left side: 2/3rds of the data-set contains junk: bad punctuation, typos, bad quote segmentation, stray markup.

Experimental Results

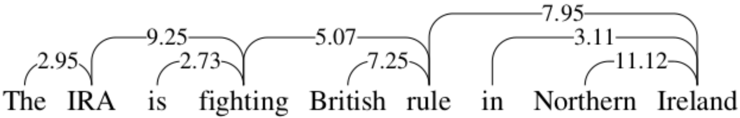
MI Distribution

28 Million word-pairs



- Sum of two curves: Gaussian and Log–Normal
- Theory: ??? Gaussian is presumably “common-mode noise”
- Uniform random under-sampling of pairs -> Gaussian
- Same for Mandarin Chinese

Maximum Spanning Tree Parse of English.



- Cutting each edge in half yields jigsaws ("disjuncts")
- Count these – Count word-jigsaw pairs (w, d)
- Repeat the matrix game.
- Matrix is (very) rectangular

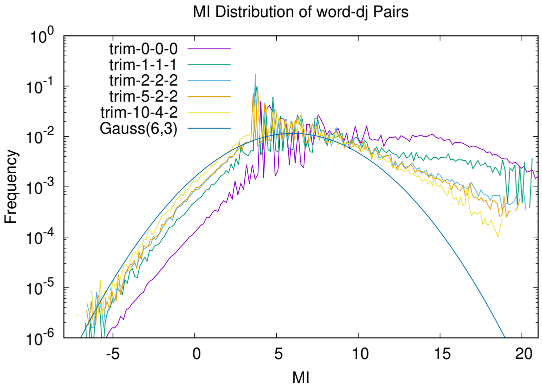
Experimental Results

Jigsaw Data Sets Characterization.

Trim cuts	full set	1-1-1	2-2-2	5-2-2	10-4-2
$\log_2 N_{\mathrm{words}}$	18.526	15.542	13.644	12.889	12.249
$\log_2 N_{\mathrm{disjuncts}}$	24.615	20.599	18.662	18.447	17.369
$\log_2 D_{\mathrm{Tot}}$	24.761	20.967	19.247	19.086	18.443
Sparsity	18.380	15.174	13.058	12.251	11.175
Rarity	3.191	2.896	3.095	3.418	3.634
$\log_2 N_{\mathrm{Tot}}/D_{\mathrm{Tot}}$	0.356	2.248	3.384	3.461	3.889
Total Entropy	24.100	19.486	17.711	17.508	16.875
Left Entropy	23.494	18.346	16.417	16.163	15.379
Right Entropy	10.157	7.937	7.280	7.268	7.258
MI	9.550	6.796	5.987	5.923	5.763

Experimental Results

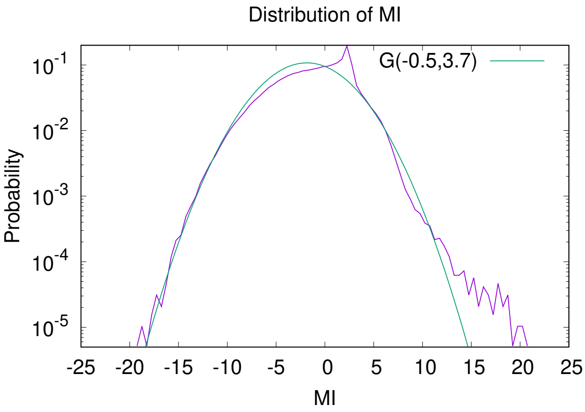
Distribution of Jigsaw (Disjunct) MI



- This is $MI(w, d)$ for word w and jigsaw d
- Unclean. Obscure meaning.

Experimental Results

Distribution of Similarity



- Wow! Gaussian!

- Inner product: $i\left(w,v\right)=\sum_d p\left(w,d\right)p\left(v,d\right)$
- MI of inner product:

$$MI\left(w,v\right)=\log_2\frac{i\left(w,v\right)i\left(*,*\right)}{i\left(w,*\right)i\left(v,*\right)}$$

- Variation of Information (VI):

$$VI\left(w,v\right)=\log_2\frac{i\left(w,v\right)}{\sqrt{i\left(w,*\right)i\left(v,*\right)}}$$

- Various Jacquard distances...
- *Not the cosine distance!!! Its terrible!*

Experimental Results

Spin Glasses

Gaussian Orthogonal Ensemble

- A high–dimensional sphere.
- With a uniform random distribution on it.
- Dimension of space == size of vocabulary.
- A vector for word w has direction $MI(w, u)$.
- Each vector corresponds to the syntactic usage of that word.
- Syntax is maximally leveraged by English speakers!
- Probably holds in other languages, too.
- This is about the effectiveness of grammar in communications.

Experimental Results

Similarity and Clustering

Clustering generalizes from specifics

Top–ranked Clusters		
+ — “ ” _	? . !	must would
, ;	He It I There	he she
was is	of in to from	are were
but and that as	has was is had could	might should will may

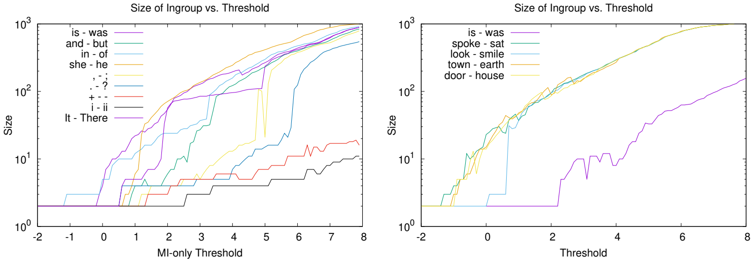
Not “just” similar words, but also:

- Similar grammatical behavior.
- Similar structure.
- Similar semantics.

Experimental Results

Word-sense Disambiguation

Each word–vector is a linear sum of multiple word-senses



- Exclusive club, Common interests
- How exclusive?
  - There's a natural threshold to nearest neighbors.
- Common interests?
  - Disjuncts not shared by majority are different word senses

Experimental Results

Conclusion

We've learned:

- Information–theoretic foundations are central.
- Experimental confirmation is central.
- Structure can be extracted from undifferentiated samples.
- Structure is a synonym for grammar.
- Recursion: structure defines a new random, uniform sampling.
- So sample again, to find differences and structure at the next level.