

Improved Unsupervised POS Induction Using Intrinsic Clustering Quality and a Zipfian Constraint

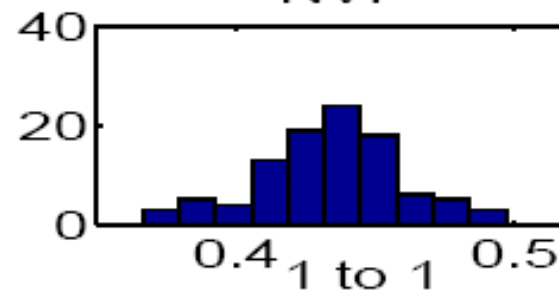
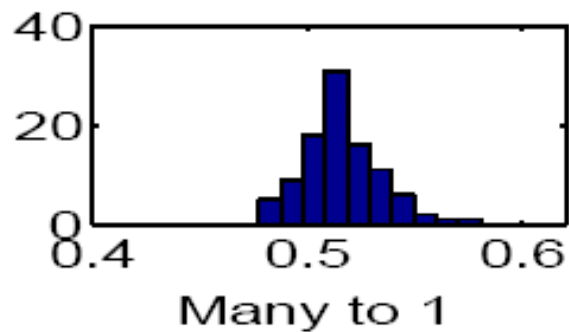
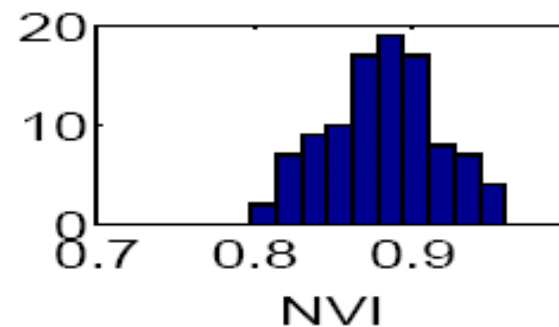
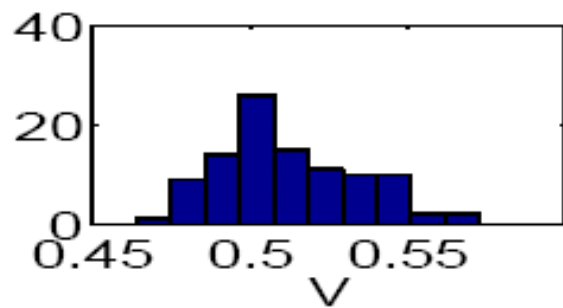
Roi Reichart, Raanan Fattal and
Ari Rappoport



Problem Definition

- Most unsupervised POS induction algorithms try to optimize a non-convex function.
- Therefore, they converge to a local maximum.
- Consequently, many papers report average results for these algorithms.
- A single run of the algorithm is not guaranteed to induce clustering of the quality reported for the algorithm.

Example – The Clark Tagger



Background – The Clark Tagger

- Optimizing a probability function consisting of a likelihood term multiplied by a prior.
- The likelihood is the traditional HMM model.
- The prior biases words with similar morphology to share the same cluster and cluster size distribution to be skewed.
- All instances of the same word type are assigned the same induced cluster.

Background – The Clark Tagger

■ Optimization:

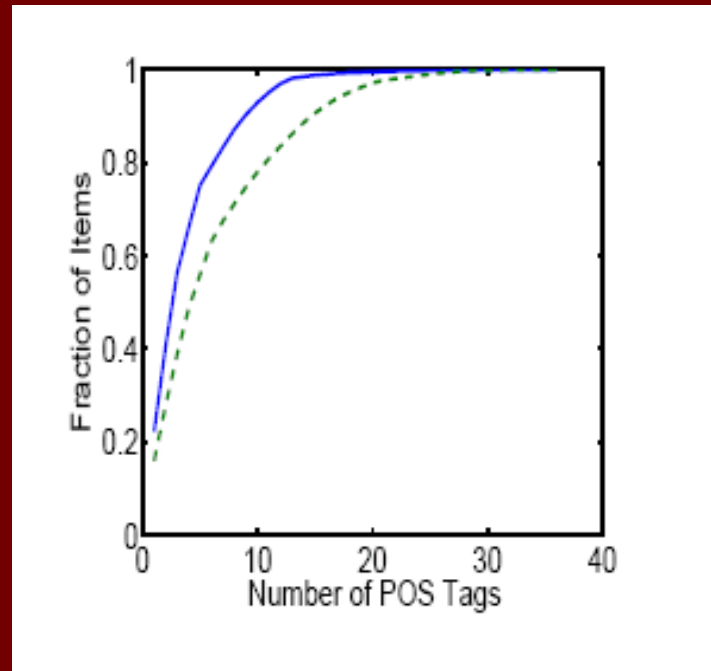
1. Start with initial assignment of word types to clusters.

2. Perform till convergence:

Go over the word types in a predefined random order and in each iteration move the addressed word type to the cluster that yields the best improvement of the probability function (if any).

The Main Insight

Imposing a Zipfian constraint on the distribution of word types to clusters may result in a more regularized clustering – i.e. it will be easier to identify the high quality runs of the tagger.



Imposing The Zipfian Constraint

- At initialization: Word types are assigned to clusters whose sizes are distributed according to the Zipfian distribution.
- At optimization: Swapping instead of moving.

Identifying High Quality Runs

- Likelihood.
- Probability function.
- Language model perplexity:

$$\sqrt[N]{\prod_{i=1}^N \frac{1}{p(w_i | w_1 \dots w_{i-1})}}$$

Results

Alg.	Data Probability				Likelihood				Perplexity			
	V		m-to-1		V		m-to-1		V		m-to-1	
	SRC	KT	SRC	KT	SRC	KT	SRC	KT	SRC	KT	SRC	KT
CT	0.2	0.143	0.071	0.045	0.338	0.23	0.22	0.148	0.568	0.397	0.476	0.33
ZCC	0.134	0.094	0.118	0.078	0.517	0.352	0.453	0.321	0.82	0.62	0.659	0.484

Table 2: Correlation of unsupervised quality measures (columns) with clustering quality of two base taggers (CT and ZCC, rows). Correlation is measured by Spearman (SRC) and Kendall Tau (KT) rank correlation coefficients. The quality measures are data probability (left part), likelihood (middle side) and perplexity (right part), and correlation is between these and two of the external evaluation measures, m-to-1 mapping and V (results for the other two clustering evaluation measures, 1-1 mapping and NVI, are very similar). Results for the perplexity quality test used by family Q are superior; data probability and likelihood provide only a mediocre indication for the quality of induced clustering. Note that the correlation values are much higher for ZCC than for CT.

Results

Alg.	V	NVI	1-1	M-1
Q(ZCC)				
no punct.	0.538 (85, 2.6)	0.849 (82, 3.2)	0.521 (100, 4.3)	0.533 (84, 1.7)
with punct.	0.637 (85, 1.8)	0.678 (82, 2.6)	0.58 (100, 3)	0.591 (84, 1.18)
Q(CT)				
no punct.	0.545 (92, 3.3)	0.837 (88, 4.4)	0.492 (99,1.4)	0.526 (75, 1)
with punct.	0.644 (92, 2.5)	0.662 (88, 4.2)	0.555 (99, 0.5)	0.585 (75, 0.58)

Table 1: Quality of the tagging produced by Q(ZCC) and Q(CT). The top (bottom) line for each algorithm presents the results when punctuation is not included (is included) in the evaluation (Section 4). The left number in the parentheses is the fraction of Clark’s (CT) results that scored worse than our models (% from 100 runs). The right number in the parentheses is 100 times the difference between the score of our model and the mean score of 100 runs of Clark’s (CT). Q(ZCC) is better than Q(CT) in the mappings measures, while Q(CT) is better in the IT measures. Both are better than the original Clark tagger CT.

Results

Alg.	V	VI	M-1	1-1
Q(ZCC)	0.637	2.06	0.591	0.58
Q(CT)	0.644	2.01	0.585	0.555
CT	0.619	2.14	0.576	0.543
HK	—	—	—	0.413
J	—	4.23 - 5.74	0.43 - 0.62	0.37 - 0.47
GG	—	2.8	—	—
G-J	—	4.03 - 4.47	—	0.4 - 0.499
VG	0.54 - 0.59	2.49 - 2.91	—	—
GGTP-45	—	—	0.654	0.445
GGTP-17	—	—	0.702	0.495