Unsupervised Dependency Parsing without Gold Part-of-Speech Tags

V.I. Spitkovsky, H. Alshawi, A.X. Chang and D. Jurafsky

Key Finding

Unsupervised word clusters can surpass the performance of gold part-of-speech tags in dependency grammar induction.

A Question

Why are gold part-of-speech tags so useful in parsing?

Two Potential Reasons:

- Grouping: pooling the statistics of words that play similar syntactic roles improves generalization by reducing sparsity;
- Disambiguation: for words that can take on multiple parts of speech, knowing gold tags limits the parsing search space.

Methodology

We test both hypotheses using two types of tag-sets.

- Tagless Lexicalized Models:

  – *full*: each word gets its own class;
  – *partial*: high frequency words get their own classes, with the rest lumped into a single "rare" cluster;
  – *none*: all words lumped into one big "cluster."

- One-Class-per-Word Remappings:

  – *most-frequent class*: uses a word's most common gold tag;
  – *most-frequent pair*: maps each word to the set of up to two of its most common gold tags;
  – *union all*: maps each word to the set of all gold tags associated with it.

word	*most-frequent class*	*most-frequent pair*	*union all*
it	{PRP}	{PRP}	{PRP}
gains	{NNS}	{VBZ, NNS}	{VBZ, NNS}
the	{DT}	{JJ, DT}	{VBP, NNP, NN, JJ, DT, CD}

Example tag reassignments derived from manually annotated categories.

Acknowledgments

Partially funded by Air Force Research Laboratory (AFRL), under prime contract no. FA8750-09-C-0181, and by NSF, via award #IIS-0811974. We thank Omri Abend, Spence Green, David McClosky and anonymous reviewers for many helpful comments on draft versions of the paper.

Experiment #1:

An Ablative Analysis and Induced Tags

Unsupervised Word Clusters

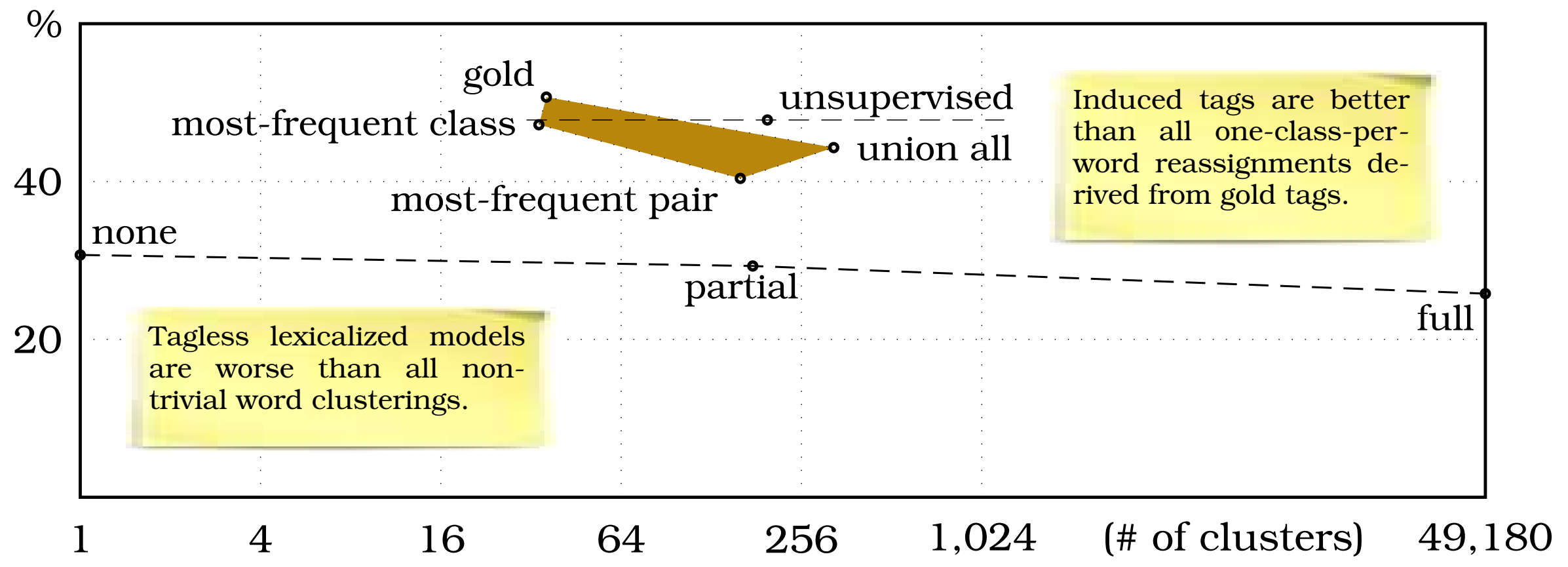
Adjectives, especially ones that take comparative (or other) complements.

Cluster #173		*Cluster #188*	
1.	open	1.	get
2.	free	2.	make
3.	further	3.	take
4.	higher	4.	find
5.	lower	5.	give
6.	similar	6.	keep
7.	leading	7.	pay
8.	present	8.	buy
9.	growing	9.	win
⋮		⋮	
37.	**cool**	42.	improve
⋮		⋮	
1,688.	up-wind	2,105.	zero-out

Bare-stem verbs (infinitive stems).

Representative members for two of Clark's (2000) flat word groupings.

Results



Induced tags are better than all one-class-per-word reassignments derived from gold tags.

Tagless lexicalized models are worse than all non-trivial word clusterings.

Parsing performance (directed dependency accuracy on WSJ15) versus the number of syntactic categories, for grammar inducers using different word clustering schemes.

Our Answer

- Grouping: appears to be vital to grammar induction;
- Disambiguation: not as crucial as grouping, but quite helpful — makes the difference between manual annotation effort and induced tags, for one-class-per-word assignments.

Conjecture:

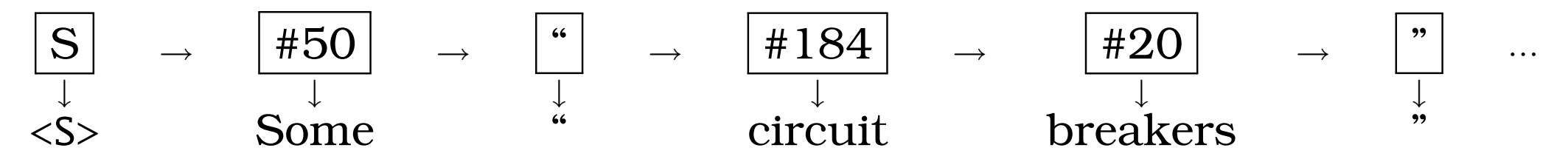
Context-sensitive unsupervised clusters should, analogously, perform better than one-class-per-word induced tags.

Experiment #2:

Context-Sensitive Unsupervised Clustering

Training the unTagger

1. Start with unsupervised cluster assignments for words in your text, and record the left- and right-context distributions of tags — $\mathbb{P}_{\mathtt{R}}(t_i \mid t_{i-1})$ and $\mathbb{P}_{\mathtt{L}}(t_i \mid t_{i+1})$ — from, e.g.:



2. Replicate the text 100-fold and inject context-colored noise to break the initial deterministic assignment of tags:

$$t'_i := \begin{cases} l, \text{ w.p. } 0.1 \cdot \mathbb{P}_{\mathtt{L}}(l \mid t_{i+1}); \\ r, \text{ w.p. } 0.1 \cdot \mathbb{P}_{\mathtt{R}}(r \mid t_{i-1}); \\ t_i \text{ otherwise (w.p. } 0.8). \end{cases}$$

3. Finally, use these perturbed sequences $\{t'_i\}$ to initialize Viterbi training of a bitag HMM, and run to convergence.

(Available at **http://nlp.stanford.edu/pubs/goldtags-data.tar.bz2**.)

Results

Some "circuit breakers" installed after the October 1987 crash failed their first test, traders say, unable to **cool** *the selling panic in both stocks and futures.*

$$\overset{\uparrow}{\boxed{\#188}}$$

word clustering scheme	*accuracy*	
gold tags	58.4	
one-class-per-word induced tags	58.2	(-0.2)
context-sensitive induced tags	59.1	(+0.7)

Directed dependency accuracies on Section 23 of WSJ (all sentences) for experiments with our recent state-of-the-art system, from CoNLL-2011.

Summary

- Word Clustering: classic unsupervised word clustering techniques of Clark (2000) and Brown et al. (1992) are well-suited to dependency parsing and grammar induction
  — *should we stop using gold tags?*
- Sequence Modeling: even a bitag HMM can relax classic one-class-per-word clustering schemes, resulting in context-sensitive cluster assignments that outperform gold tags
  — *should we start using soft clustering?*