Learning Annotations

Slav Petrov et al.
Jonathan Clark
Presented by Jonathan Clark

Advanced Machine Translation Seminar

Wednesday, February 17, 2010

Outline

- Why do we care about labels in MT?
- Background
- Learning
- Inference
- Results

Isn't this a parsing paper?

- Yes, but…
  - We use parsers
  - Hypergraph decoders act like parsers
  - Grammar induction and nonterminal granularity is also an issue in SCFG MT

The Parsing Task

- (Over)fit to Penn Treebank by maximizing likelihood of trees that linguists made up to annotate strange WSJ language

Splitting non-terminals

- Lexicalize grammar:
  - (S-did (NP-he (N-he he)) (VP-did) (V-did did))
- Markovize grammar:
  - (S (NP^S (N^NP he))
- Cluster grammar (this work):
  - (S-2 (NP-13 (N-9 he))

Learning: Initialization

- Fix structure
- Label with PTB symbols
  - But we wouldn't have to!

Learning: Splitting

- Annotations are latent
  - One tree becomes many fuzzy trees
- E: P(annotated rule in context)
  - Inside-Outside is O(n) -- fixed structure
- M: Re-estimate preference of annotated RHS's for this LHS

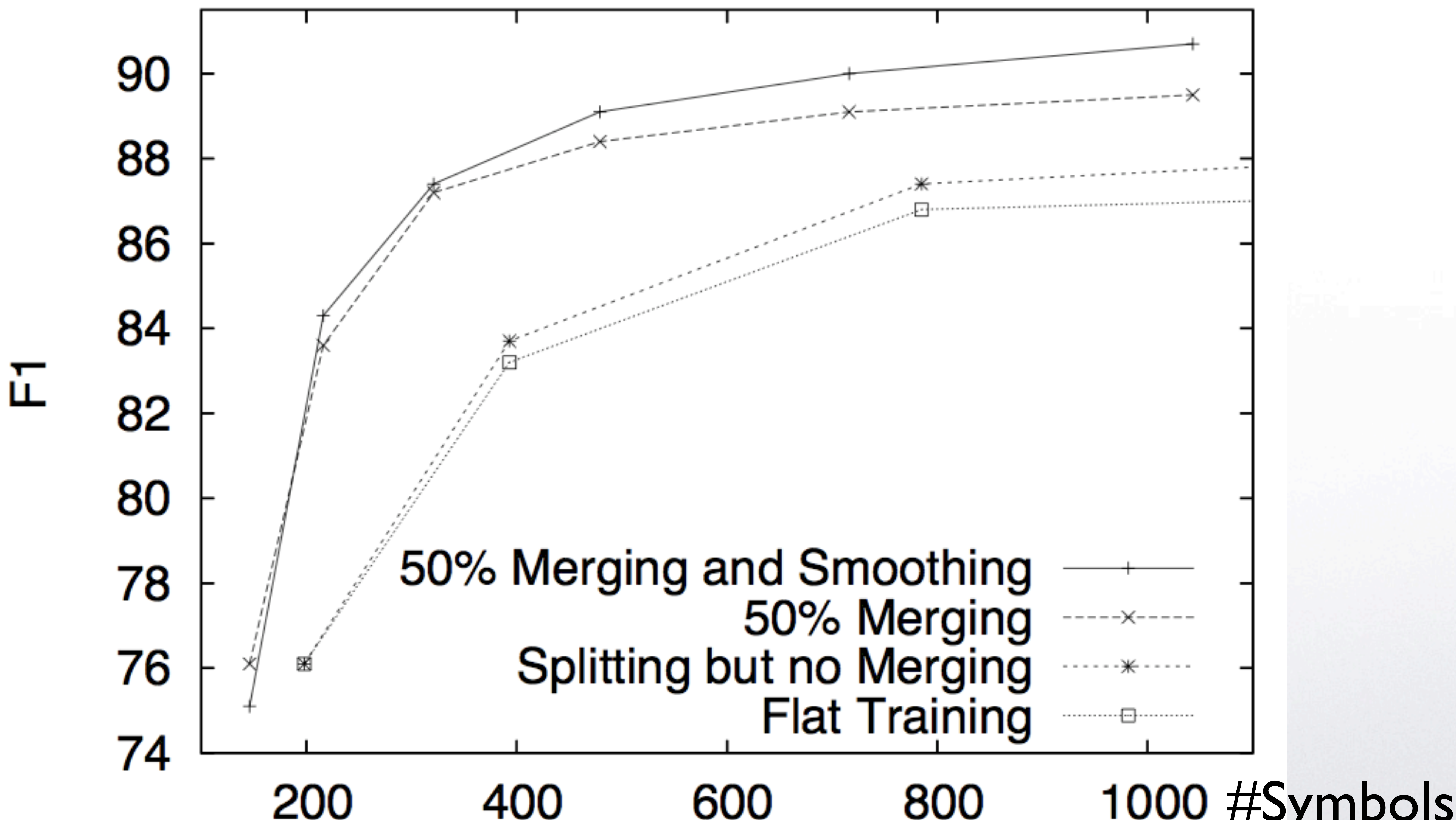
$$\frac{\#\{Ax \rightarrow By\ Cz\}}{\sum_{y',z'} \#\{Ax \rightarrow By'\ Cz'\}}$$

Learning: Merging

- Oops, we overfitted… and ran out of memory
- We don't need 16 types of commas
- Merging allows us to:
  - Consider dependencies among splits
  - Split more
  - Approximate likelihood loss efficiently
    - Ignore interactions in same tree, same symbol

Learning: Smoothing

- Interpolate with average of annotations
  - 0.01 goes to other annotations
- Gives significant gain in results



F1
90
88
86
84
82
80
78
76
74
200
400
600
800
1000
#Symbols
50% Merging and Smoothing
50% Merging
Splitting but no Merging
Flat Training

≤ 40 words	LP	LR	CB	0CB
Klein and Manning (2003)	86.9	85.7	1.10	60.3
Matsuzaki et al. (2005)	86.6	86.7	1.19	61.1
Collins (1999)	88.7	88.5	0.92	66.7
Charniak and Johnson (2005)	90.1	**90.1**	**0.74**	**70.1**
This Paper	**90.3**	90.0	0.78	68.5

all sentences	LP	LR	CB	0CB
Klein and Manning (2003)	86.3	85.1	1.31	57.2
Matsuzaki et al. (2005)	86.1	86.0	1.39	58.3
Collins (1999)	88.3	88.1	1.06	64.0
Charniak and Johnson (2005)	89.5	**89.6**	**0.88**	**67.6**
This Paper	**89.8**	**89.6**	0.92	66.3

Inference: Parsing

- Extra annotations are nuisance variable
- Options:
  - Max Derivation
  - Variational Inference
    - Maximum rules expected correct (Again, may feel a bit like MBR)

Inference: Pruning

- Coarse-to-fine pruning
- Threshold pruning of low probability symbols
  - 16X speedup, little effect on quality

VBZ

VBZ-0	gives	sells	takes
VBZ-1	comes	goes	works
VBZ-2	includes	owns	is
VBZ-3	puts	provides	takes
VBZ-4	says	adds	Says
VBZ-5	believes	means	thinks
VBZ-6	expects	makes	calls
VBZ-7	plans	expects	wants
VBZ-8	is	’s	gets
VBZ-9	’s	is	remains
VBZ-10	has	’s	is
VBZ-11	does	Is	Does

DT

DT-0	the	The	a
DT-1	A	An	Another
DT-2	The	No	This
DT-3	The	Some	These
DT-4	all	those	some
DT-5	some	these	both
DT-6	That	This	each
DT-7	this	that	each
DT-8	the	The	a
DT-9	no	any	some
DT-10	an	a	the
DT-11	a	this	the

ADVP

ADVP-0	RB-13 NP-2	RB-13 PP-3	IN-15 NP-2
ADVP-1	NP-3 RB-10	NP-3 RBR-2	NP-3 IN-14
ADVP-2	IN-5 JJS-1	RB-8 RB-6	RB-6 RBR-1
ADVP-3	RBR-0	RB-12 PP-0	RP-0
ADVP-4	RB-3 RB-6	ADVP-2 SBAR-8	ADVP-2 PP-5
ADVP-5	RB-5	NP-3 RB-10	RB-0
ADVP-6	RB-4	RB-0	RB-3 RB-6
ADVP-7	RB-7	IN-5 JJS-1	RB-6
ADVP-8	RB-0	RBS-0	RBR-1 IN-14
ADVP-9	RB-1	IN-15	RBR-0

SINV

<table>
<tr><td>SINV-0</td><td>VP-14 NP-7</td><td>VP-14</td><td>VP-15 NP-7 NP-9</td></tr>
<tr><td>SINV-1</td><td colspan="3">VP-14 NP-7 .-0
S-6 ,-0 VP-14 NP-7 .-0
S-11 VP-14 NP-7 .-0</td></tr>
</table>