Unsupervised Methods in Deep Processing

LING 571 — Deep Processing in NLP

December 5th, 2018

Ryan Georgi

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Announcements

- Course evaluations are available online until December 14th.
  - Please take the time to fill one out, it's helpful to us for improving the course.
- Remaining grades will be finished ASAP
  - (Including HW#4-EX!)

Degrees of Supervision

Degrees of Supervision

- **Problem**
  - Creating annotated language data is ***expensive***
  - Language research isn't always well-funded
- **Bigger Problem**
  - Newswire English ≠ "Natural Language"

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Degrees of Supervision

- How to get the most "bang for your buck"?
  - What can you do with just raw text?
  - How about raw text and a POS tagger?
  - How about raw text and one or two language experts?

Levels of Supervision

- Supervised
- Unsupervised
- Semi-supervised

Two Example Problems

- **Tasks**
  - Grammar (PCFG) Induction
  - Semantic Role Labeling (SRL)
- **Highlights**
  - Examples of how to merge Shallow Processing Intuitions w/Deep Processing
  - Examples of how to maximize

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Example: Learning a PCFG

- **Supervised**
  - Requires a full treebank with syntactic parses
  - You've implemented the fully supervised case already!
- **Unsupervised**
  - What if we don't have parses available?
  - Can we infer information about constituency from raw text?
- **Semi-Supervised**
  - Maybe we have a few parses available?
  - Maybe we just have some idea what common constituents look like?

Inside-Outside Algorithm

Inside-Outside Algorithm
(Baker, 1979)

- If we have an existing representation of our grammar…
  - Nonterminals
  - Terminals (POS Tags)
  - …maybe even some guesses at rewrite rules
- …can we estimate their probabilities from raw text?

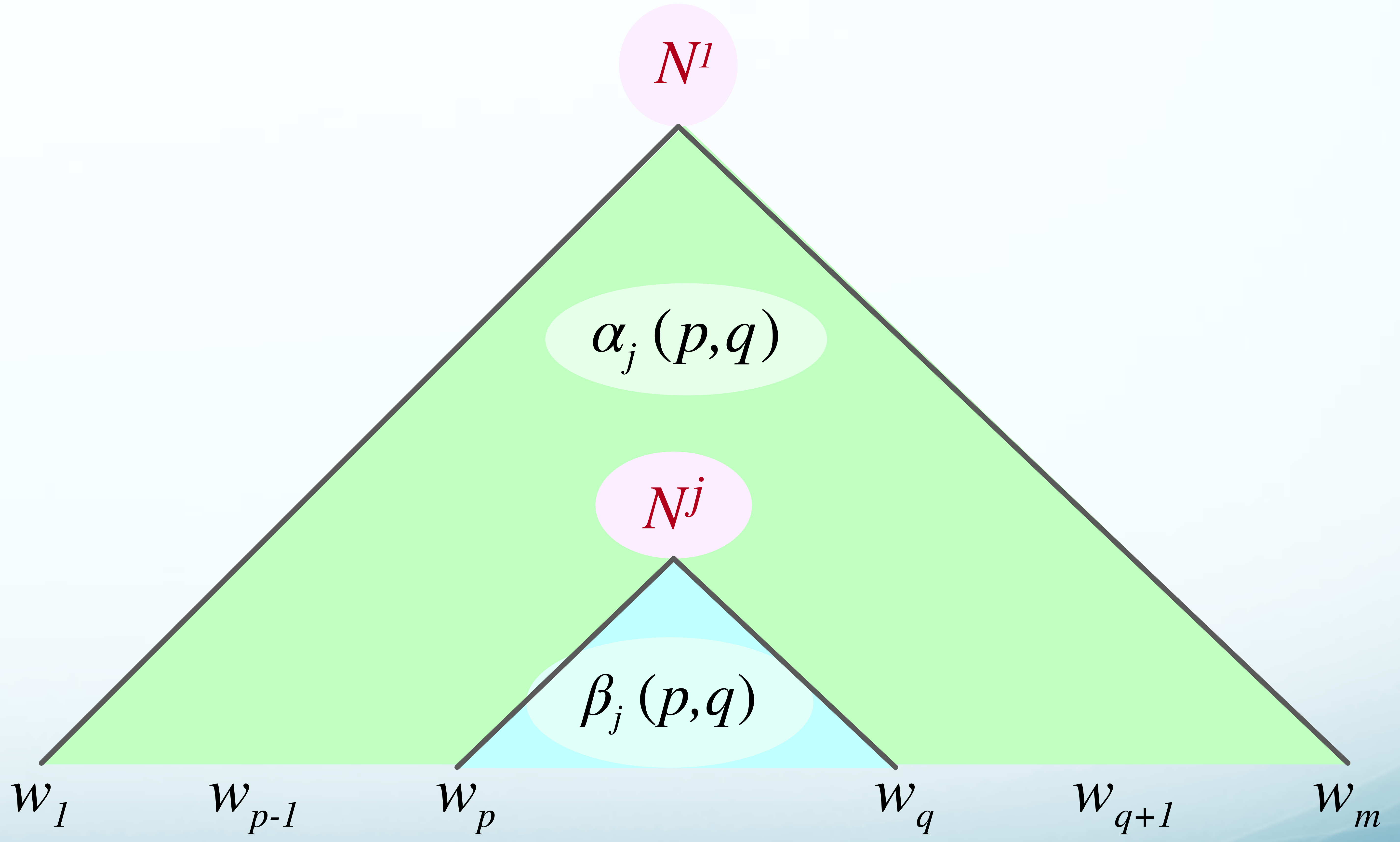
UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Inside-Outside Algorithm

- A type of ***Expectation Maximization (EM)*** Algorithm
- **Expectation**
  - Given input grammar rules and probabilities…
  - Calculate ***expected*** likelihood of observed input using current rule probabilities
  - *Partial counts* = sum of probabilities for any nonterminal expansion covering ("explaining") the observed span
- **Maximization**
  - Use *partial counts* as if these were true counts in a PCFG induction step
  - Recalculate probabilities based on these new counts

Inside-Outside Algorithm

- With a start symbol N^1
- And some nonterm N^j
- β_j is the "inside" probability
  - …that N^j is a node covering $w_p \ldots w_q$
- And α_j is the "outside" probability
  - Of the rest of the tree being expanded



UNIVERSITY of WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Inside-Outside Algorithm

Inside Probabilities

- Total probability of generating words $w_p \ldots w_q$ from non-terminal N^j.

$$\beta_j(p,q) = \boxed{P\left(w_{pq} \mid N^j_{pq}\right)}$$

- This is the probability of ***all possible expansions*** of any nonterm covering that word sequence.

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Inside-Outside Algorithm

Outside Probabilities

- Total probability of beginning with start symbol N^1 and generating N^j_{pq} and all the words ***outside*** $w_p \ldots w_q$

$$\alpha_j(p,q) = P\left(w_{1(p-1)} \mid N^j_{pq}, w_{(q+1)m}\right)$$

- Zero out impossible (out-of-order) spans

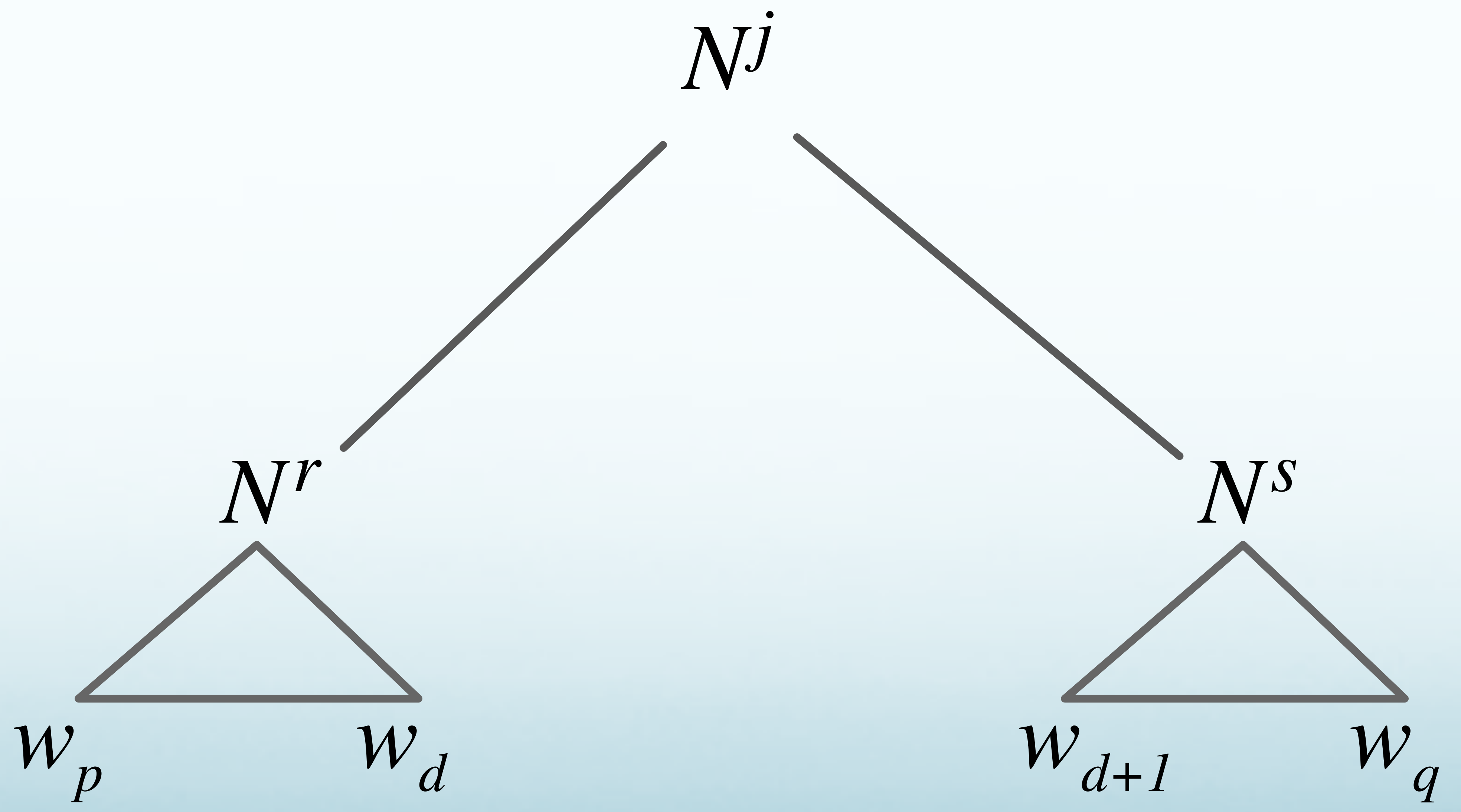
$$\text{when } p > q \qquad \alpha_j(p,q) = \beta_j(p,q) = 0$$

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Calculating Inside Probability

- If a pre-terminal: $\beta_j(k,k) = P\left(N^j \to w_k\right)$
- Otherwise: $\beta_j(p,q) = \sum_{r,s} \sum_{d=p}^{q-1} P\left(N^j \to N^r N^s\right) \cdot \beta_r(p,d) \cdot \beta_s(d+1,q)$

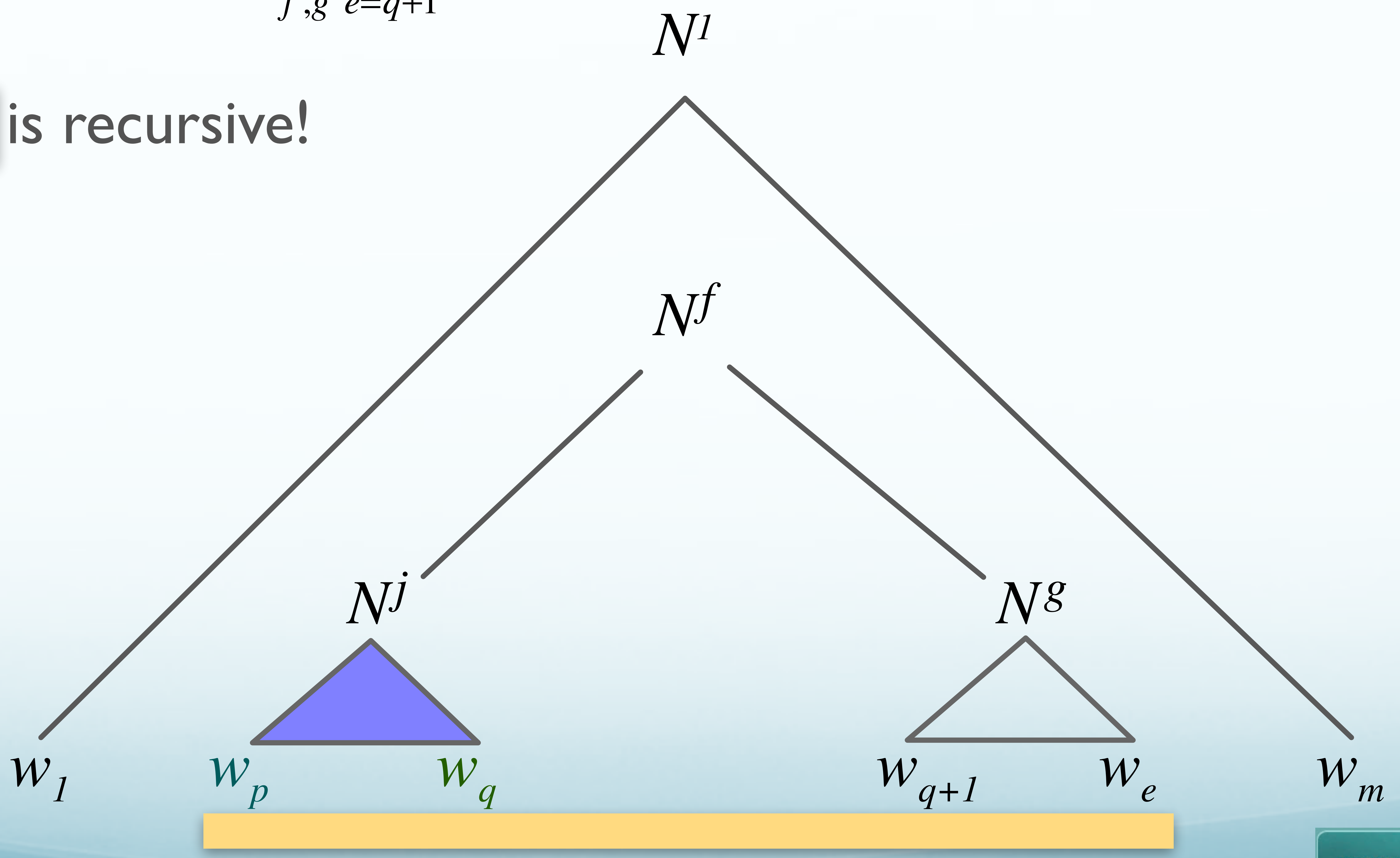


UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Calculating Outside Probability

$$\alpha_j(p,q) = \sum_{f,g} \sum_{e=q+1}^{m} \alpha_f(p,e) \cdot P\left(N^f \rightarrow N^j N^g\right) \cdot \beta_g(q+1,e)$$

Note that this part is recursive!



UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Inside-Outside Algorithm

Fully Unsupervised Setting

- **Setup**
  - Choose set of nonterminals
  - Initialize all possible (CNF-Compatible) rules with random weights
- **Problems**
  - Massive parameter space
  - Meaning of nonterminals is random
  - Might do okay inducing constituency
    - …but internal nodes are going to be somewhat meaningless

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Inside-Outside Algorithm

Semi-supervised Setting

- **Setup**
  - Choose set of nonterminals
  - Initialize some set of learned rules, usually from small treebank
- **Improvements**
  - Bootstraps nonterminals to some linguistic knowledge
  - Rules out many impossible constituents
- **Problems**
  - Algorithm prefers grammars concentrating probability on a few rules *(de Marcken, 1995)*
  - Still many local optima

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Semi-Supervised Grammar Induction

Haghighi & Klein (2006)

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Prototype-Driven Grammar Induction

Haghighi & Klein (2006)

- What if:
  - You still don't have syntactically parsed corpora
  - …but you have some good ideas of what some constituents look like?

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Prototype-Driven Grammar Induction

[Haghighi & Klein (2006)](#)

- Provide some "prototypical" constituent structures:

	Prototypes
NP	DT NN
	JJ NNS
	NNP NNP
VP	VBN IN NN
	VBD DT NN
	MD VB CD
…	…

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Prototype-Driven Grammar Induction

Haghighi & Klein (2006)

- **Hypothesis**
  - If a prototype is seen as a constituent, it must receive the prototype's entry label
  - This will provide "pressure" to allocate probability mass to the correct nonterminals
- **Implementation**
  - Using Inside-Outside algorithm, if N^j is dominating a span of POS in prototype list…
  - Zero out partial counts for any rule where LHS does not match that of prototype

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Prototype-Driven Grammar Induction

Other "Tricks"

- **Expand The Prototype List**
  - In addition to manual prototypes,
  - Use context vectors to expand to sequences found in similar settings
- **Constrain what might be a constituent**
  - Use Constituent-Context Model (CCM) *(Klein & Manning, 2002)*
  - Use unparsed data and contextual modeling to form distributional clusters
    - Clusters represent what is frequently a ***constituent*** vs. ***distituent***
  - Add to inside-outside by multiplying bracket scores with inside-outside scores

Prototype-Driven Grammar Induction

Results

- Include both Labeled/Unlabeled Bracketing
- Pure Inside-Outside is terrible
- Just adding prototypes is a huge improvement
- Using prototypes with induced brackets produces the best (non-oracle) result

	Labeled			Unlabeled		
Setting	Prec.	Rec.	F_1	Prec.	Rec.	F_1
No Brackets						
PCFG× NONE	23.9	29.1	26.3	40.7	52.1	45.7
PROTO× NONE	51.8	62.9	56.8	59.6	76.2	66.9
Gold Brackets						
PCFG× GOLD	47.0	57.2	51.6	78.8	100.0	88.1
PROTO× GOLD	64.8	78.7	71.1	78.8	100.0	88.1
CCM Brackets						
CCM	-	-	-	64.2	81.6	71.9
PCFG× CCM	32.3	38.9	35.3	64.1	81.4	71.8
PROTO× CCM	56.9	68.5	62.2	68.4	86.9	76.5
BEST	59.4	72.1	65.1	69.7	89.1	78.2
UBOUND	78.8	94.7	86.0	78.8	100.0	88.1

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Prototype-Driven Grammar Induction

Conclusions

- Using fairly basic speaker intuitions…
- Combined with shallow processing techniques
- Doesn't reach state-of-the-art, but might allow for reasonable performance on a previously unseen language/domain!

Unsupervised Semantic Role Labeling

[Lang & Lapata (2010)]

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Unsupervised Semantic Role Labeling

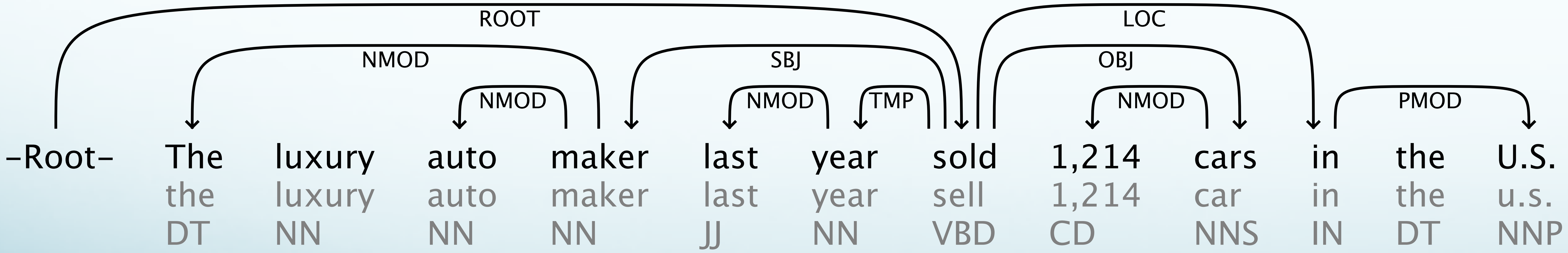
- Available Resources
  - Dependency parser (with syntactic functions)
  - POS tags
- Unavailable Resources
  - Role-annotated corpora

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Unsupervised Semantic Role Labeling

- Helpful Insight
  - Syntactic functions of dependencies correlate strongly to semantic roles
  - For instance, OBJ is almost always ARG1 (*PROTO-PATIENT*)
  - Can use this as cue for ***canonical*** argument form

	A0	A1	TMP	MNR
SBJ	54514	19684	15	7
OBJ	3359	51730	93	54
ADV	162	3506	976	2308
TMP	5	60	15167	22
PMOD	2466	4860	142	62
OPRD	37	5554	1	36
LOC	17	145	43	157
DIR	0	178	15	6
MNR	5	48	13	3312
PRP	9	50	11	6
LGS	2168	36	2	2
PRD	413	830	31	38
NMOD	422	388	25	59
EXT	0	20	2	12
DEP	18	150	25	65
SUB	3	84	4	2
CONJ	198	331	22	8
ROOT	62	147	84	2
	64517	88616	16803	6404



UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Unsupervised Semantic Role Labeling

- Problem formulation:
  - Treat induction of roles as a clustering problem
  - Clusters represent a predicate and an argument relating in a specific way
  - Predicates will have ***canonical*** theta frames, and alternations
    - …how to avoid only labeling everything as canonical?

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Unsupervised Semantic Role Labeling

Features

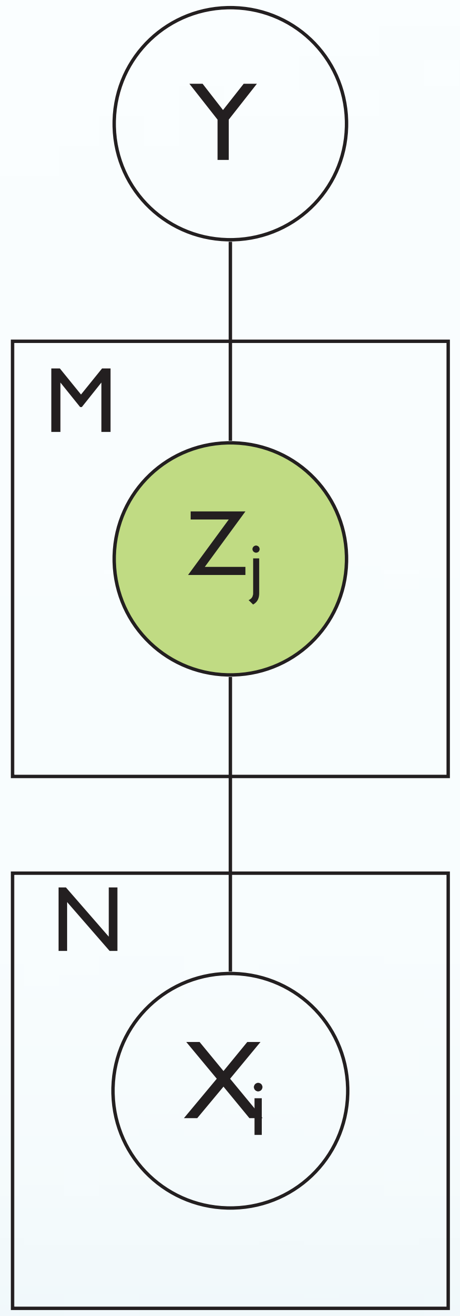
- Clusters? So, what were the features?
  - Predicate lemma
  - Argument lemma
  - Argument POS
  - Preposition between predicate and argument (if one exists)
  - Lemma of left-/rightmost child of argument
  - All syntactic functions of argument's children

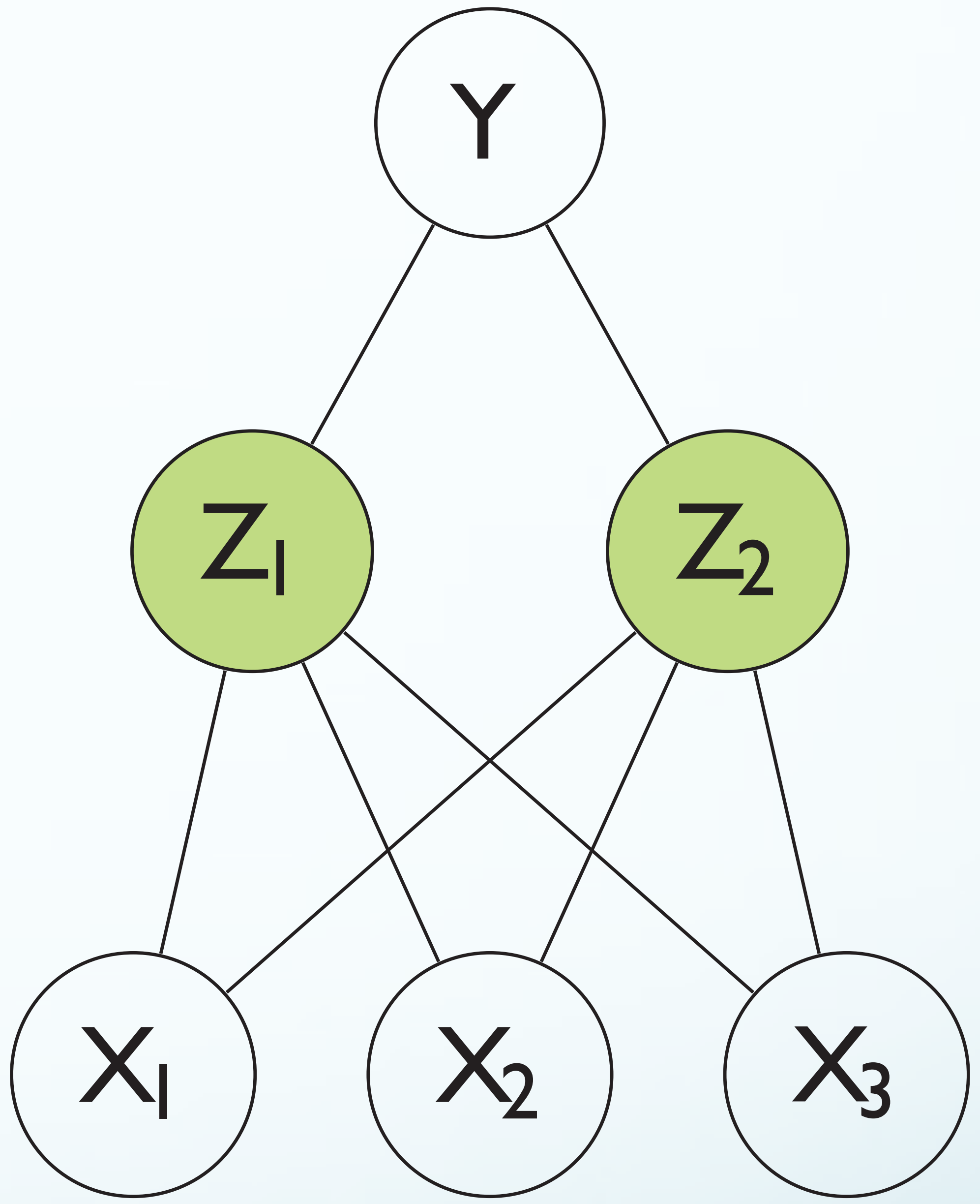
UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Unsupervised Semantic Role Labeling

Avoiding Overfitting to Canoncial Form

- Proposed Solution:
  - Introduce latent variable into logistic classifier
  - Influence the classifier to learn more abstract relations than just syntactic order or functions





Results

	PU		CA		CP		CR		CF1	
	Mic	Mac	Mic	Mac	Mic	Mac	Mic	Mac	Mic	Mac
SyntFunc	73.2	75.8	82.0	80.9	67.6	65.3	55.7	50.1	61.1	56.7
LogLV	72.5	74.0	81.1	79.4	64.3	60.6	59.7	56.3	61.9	58.4
UpperBndS	94.7	96.1	96.9	97.0	97.4	97.6	90.4	100	93.7	93.8
UpperBndG	98.8	99.4	99.9	99.9	99.7	99.9	100	100	99.8	100

- Key:
  - **SyntFunc** is rule-based baseline mapping syntactic function to semantic role
  - Metrics:
    - PU = Cluster **Purity**
    - CA = **Cluster Accuracy**
    - P/R/F_1
  - Mic/Mac = Micro vs. Macro average

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Results

	PU		CA		CP		CR		CF1	
	Mic	Mac	Mic	Mac	Mic	Mac	Mic	Mac	Mic	Mac
SyntFunc	73.2	75.8	82.0	80.9	67.6	65.3	55.7	50.1	61.1	56.7
LogLV	72.5	74.0	81.1	79.4	64.3	60.6	59.7	56.3	61.9	58.4
UpperBndS	94.7	96.1	96.9	97.0	97.4	97.6	90.4	100	93.7	93.8
UpperBndG	98.8	99.4	99.9	99.9	99.7	99.9	100	100	99.8	100

- Author's system (LogLV) looks very similar to baseline (SyntFunc)
  - …so is there really any improvement?

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Results

	PU		CA		CP		CR		CF1	
	Mic	Mac	Mic	Mac	Mic	Mac	Mic	Mac	Mic	Mac
SyntFunct	73.9	77.8	82.1	81.3	68.0	66.5	55.9	50.3	61.4	57.3
LogLV	82.6	83.7	87.4	85.5	79.1	74.5	73.3	68.5	76.1	71.4

- What about non-canonical forms?
  - Canonical forms are rarer, but this system does a much better job at finding them

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Unsupervised Semantic Role Labeling

Conclusions

- Just because you don't have one type of annotation
  - Look for others!
  - Syntax, word order, POS tags… all can help make decisions about other tasks

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Conclusions

- How useful is your system in making predictions if it basically just chooses the most common thing?
- Deep Processing looks at one set of tasks
  - But make sure to use information from shallow processing
  - …as well as your own intuitions!

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Thank You!