Refining syntactic categories using local contexts

Experiments in unlexicalized PCFG parsing

John Pate and Detmar Meurers

The Ohio State University

TLT '07
December 8, 2007

Overview

- Introduction: Syntactic categories and the independence assumption of PCFGs

Overview

- Introduction: Syntactic categories and the independence assumption of PCFGs
- Explore effect of contextualization on PCFG parsing

Overview

- Introduction: Syntactic categories and the independence assumption of PCFGs
- Explore effect of contextualization on PCFG parsing
  - Encode the local tree context in each LHS category to obtain new category distinctions. (Experiment 1)

Overview

- Introduction: Syntactic categories and the independence assumption of PCFGs
- Explore effect of contextualization on PCFG parsing
  - Encode the local tree context in each LHS category to obtain new category distinctions. (Experiment 1)
  - Keep only the new categories distinctions which differ
    - from each other (Experiment 2)
    - from the original category (Experiment 3)

  by clustering based on their probability distribution.

Introduction

Syntactic Categories

- Categories encode distributional properties
  - Cf. substitution/replacement constituency test

    (1) a. *Bob [$_{VP}$ loves ice cream].*

    b. *Bob [$_{VP}$ gave the book to Mary].*

  - Distinct categories need to be introduced where substitution is not possible.

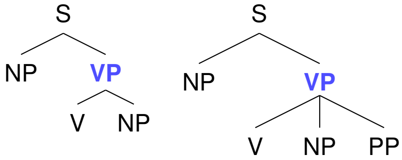
Introduction

Syntactic Categories

- Categories encode distributional properties
  - Cf. substitution/replacement constituency test

    (1) a. *Bob [$_{VP}$ loves ice cream].*

    b. *Bob [$_{VP}$ gave the book to Mary].*
  - Distinct categories need to be introduced where substitution is not possible.
- Two local trees can only “communicate” through the category connecting them. → Locality



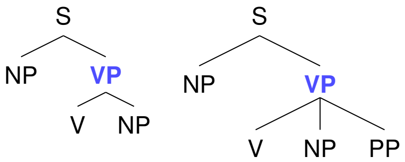
Introduction

Syntactic Categories

- Categories encode distributional properties
  - Cf. substitution/replacement constituency test

    (1) a. *Bob [$_{VP}$ loves ice cream].*

    b. *Bob [$_{VP}$ gave the book to Mary].*
  - Distinct categories need to be introduced where substitution is not possible.
- Two local trees can only "communicate" through the category connecting them. → Locality



- Context in which a category occurs plays no role.
  - e.g., an NP as daughter of S is assumed to be realizable in all the same ways as an NP in a VP.

Introduction

PCFG and its independence assumption

- PCFG rules encode not only possible expansions but a probability distribution over these.

Introduction

PCFG and its independence assumption

- PCFG rules encode not only possible expansions but a probability distribution over these.
- *Expansion vector* for each non-terminal category α:
  - One dimension per right-hand side expansion β_i
  - Encode likelihood of expanding α to each β_i
    - Estimated by counting the number of times a given category takes each of its expansions in a treebank.

Introduction

PCFG and its independence assumption

- PCFG rules encode not only possible expansions but a probability distribution over these.
- *Expansion vector* for each non-terminal category α:
  - One dimension per right-hand side expansion β_i
  - Encode likelihood of expanding α to each β_i
    - Estimated by counting the number of times a given category takes each of its expansions in a treebank.
- Locality strengthened to independence assumption:
  - Wherever a category occurs, it can be realized in the same ways, at the same likelihoods.

Introduction

PCFG and its independence assumption

- PCFG rules encode not only possible expansions but a probability distribution over these.
- *Expansion vector* for each non-terminal category α:
  - One dimension per right-hand side expansion β_i
  - Encode likelihood of expanding α to each β_i
    - Estimated by counting the number of times a given category takes each of its expansions in a treebank.
- Locality strengthened to independence assumption:
  - Wherever a category occurs, it can be realized in the same ways, at the same likelihoods.

$\Rightarrow$ Modify the grammar to better satisfy the independence assumption:

  - Introduce new category distinctions for each context

Contextualization

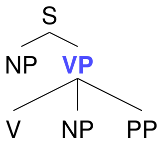
Introduction

- Johnson (1998): Let each category in treebank encode the category label of the mother in its local tree.

Contextualization

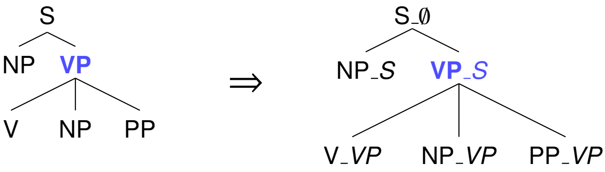
Introduction

- Johnson (1998): Let each category in treebank encode the category label of the mother in its local tree.



Contextualization
Introduction

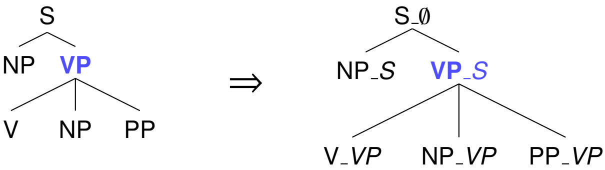
- Johnson (1998): Let each category in treebank encode the category label of the mother in its local tree.



Contextualization

Introduction

- Johnson (1998): Let each category in treebank encode the category label of the mother in its local tree.

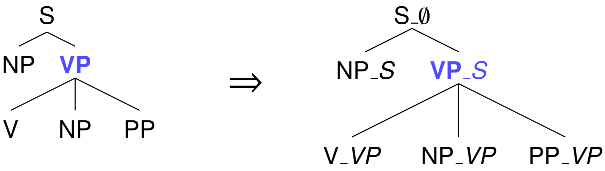


- The new categories differ in their distributional properties, as shown by their probability distributions.
  - Example:

VP $\rightarrow_{0.3}$ V NP PP		VP_*S* $\rightarrow_{0.4}$ V NP_*VP* PP_*VP*	
VP $\rightarrow_{0.6}$ V NP	⇒	VP_*S* $\rightarrow_{0.4}$ V NP_*VP*	...
VP $\rightarrow_{0.1}$ V		VP_*S* $\rightarrow_{0.2}$ V	

Contextualization
Introduction

- Johnson (1998): Let each category in treebank encode the category label of the mother in its local tree.



- The new categories differ in their distributional properties, as shown by their probability distributions.
  - Example:

$\text{VP} \rightarrow_{0.3} \text{V NP PP}$		$\text{VP\_}S \rightarrow_{0.4} \text{V NP\_}VP\ \text{PP\_}VP$	
$\text{VP} \rightarrow_{0.6} \text{V NP}$	$\Rightarrow$	$\text{VP\_}S \rightarrow_{0.4} \text{V NP\_}VP$	. . .
$\text{VP} \rightarrow_{0.1} \text{V}$		$\text{VP\_}S \rightarrow_{0.2} \text{V}$	

- Better estimates for context-enriched categories
  $\rightarrow$ better parsing performance

Contextualization

Experiment 1: Systematically Encoding Local Contexts

- Instead of introducing all mother contexts into the categories, Klein and Manning (2003) introduce specific, linguistically motivated distinctions.

Contextualization

Experiment 1: Systematically Encoding Local Contexts

- Instead of introducing all mother contexts into the categories, Klein and Manning (2003) introduce specific, linguistically motivated distinctions.
- How about automatically introducing relevant distributional distinctions?

Contextualization

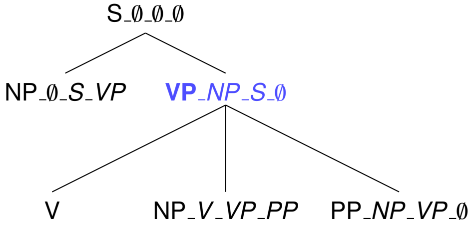
Experiment 1: Systematically Encoding Local Contexts

- Instead of introducing all mother contexts into the categories, Klein and Manning (2003) introduce specific, linguistically motivated distinctions.
- How about automatically introducing relevant distributional distinctions?
  - ⇒ Annotate non-preterminals with local context: mother, left sister, and/or right sister

Contextualization

Experiment 1: Systematically Encoding Local Contexts

- Instead of introducing all mother contexts into the categories, Klein and Manning (2003) introduce specific, linguistically motivated distinctions.
- How about automatically introducing relevant distributional distinctions?
  - ⇒ Annotate non-preterminals with local context: mother, left sister, and/or right sister
    - Example for Cat_*LeftSister*_*Mother*_*RightSister*:



Contextualization
Experimental Setup

- *Treebank used*: WSJ from the Penn Treebank
  - Training set: sections 2–21
  - Development set: section 22
  - Test set: section 23

Contextualization

Experimental Setup

- *Treebank used*: WSJ from the Penn Treebank
  - Training set: sections 2–21
  - Development set: section 22
  - Test set: section 23
- *Prepare treebank* in standard way (Johnson 1998)
  - Remove grammatical function and coindexation tags
  - Remove nodes dominating empty string only

Contextualization
Experimental Setup

- *Treebank used*: WSJ from the Penn Treebank
  - Training set: sections 2–21
  - Development set: section 22
  - Test set: section 23
- *Prepare treebank* in standard way (Johnson 1998)
  - Remove grammatical function and coindexation tags
  - Remove nodes dominating empty string only
- *Parse* using plain PCFG parser (BitPar, Schmid 2004) with gold-standard part of speech as input.

Contextualization
Experimental Setup

- *Treebank used*: WSJ from the Penn Treebank
  - Training set: sections 2–21
  - Development set: section 22
  - Test set: section 23
- *Prepare treebank* in standard way (Johnson 1998)
  - Remove grammatical function and coindexation tags
  - Remove nodes dominating empty string only
- *Parse* using plain PCFG parser (BitPar, Schmid 2004) with gold-standard part of speech as input.
- *Evaluate* against standard WSJ treebank
  - after mapping parse back to original categories (simple stripping off of added contextualization)
  - EVALB parseval package (Sekine and Collins 1997)

Contextualization
Experiment 1: Results

Contextualization	Prec.	Recall	F	Cats	Nonce	Rules
Baseline	74.79	69.94	72.28	28	1	14,974
Mother (M)	81.10	79.64	80.36	300	62	22,696
Left Sister (L)	79.75	78.05	78.89	648	182	32,304
Right Sister (R)	80.10	77.02	78.53	523	170	26,327
L & R	80.49	80.79	80.64	3,004	1,263	47,677
M & L	81.44	80.86	81.15	1,905	723	38,350
M & R	82.37	82.07	82.22	1,592	640	34,390
M & L & R	80.99	81.34	81.16	5,177	2,627	52,756

Contextualization

Experiment 1: Results

Contextualization	Prec.	Recall	F	Cats	Nonce	Rules
Baseline	74.79	69.94	72.28	28	1	14,974
Mother (M)	81.10	79.64	**80.36**	300	62	22,696
Left Sister (L)	79.75	78.05	**78.89**	648	182	32,304
Right Sister (R)	80.10	77.02	**78.53**	523	170	26,327
L & R	80.49	80.79	**80.64**	3,004	1,263	47,677
M & L	81.44	80.86	**81.15**	1,905	723	38,350
M & R	82.37	82.07	**82.22**	1,592	640	34,390
M & L & R	80.99	81.34	**81.16**	5,177	2,627	52,756

- Contextualization improves F-score significantly

Contextualization

Experiment 1: Results

Contextualization	Prec.	Recall	F	Cats	Nonce	Rules
Baseline	74.79	69.94	72.28	28	1	14,974
Mother (M)	81.10	79.64	80.36	300	62	22,696
Left Sister (L)	79.75	78.05	78.89	648	182	32,304
Right Sister (R)	80.10	77.02	78.53	523	170	26,327
L & R	80.49	80.79	80.64	3,004	1,263	47,677
M & L	81.44	80.86	81.15	1,905	723	38,350
M & R	82.37	82.07	**82.22**	1,592	640	34,390
M & L & R	80.99	81.34	**81.16**	5,177	2,627	52,756

- Contextualization improves F-score significantly
- Mother-both-sisters contextualization is *not* the best.

Contextualization

Experiment 1: Results

Contextualization	Prec.	Recall	F	Cats	Nonce	Rules
Baseline	74.79	69.94	72.28	28	1	14,974
Mother (M)	81.10	79.64	80.36	300	62	22,696
Left Sister (L)	79.75	78.05	78.89	648	182	32,304
Right Sister (R)	80.10	77.02	78.53	523	170	26,327
L & R	80.49	80.79	80.64	3,004	1,263	47,677
M & L	81.44	80.86	81.15	1,905	723	38,350
M & R	82.37	82.07	82.22	1,592	640	34,390
M & L & R	80.99	81.34	81.16	**5,177**	**2,627**	**52,756**

- Contextualization improves F-score significantly
- Mother-both-sisters contextualization is *not* the best.
  - high number of categories and rules, with over half of the categories occurring only once → data sparsity

Contextualization

Experiment 1: Results

Contextualization	Prec.	Recall	F	Cats	Nonce	Rules
Baseline	74.79	69.94	72.28	28	1	14,974
Mother (M)	81.10	79.64	80.36	300	62	22,696
Left Sister (L)	79.75	78.05	78.89	648	182	32,304
Right Sister (R)	80.10	77.02	78.53	523	170	26,327
L & R	80.49	80.79	80.64	3,004	1,263	47,677
M & L	81.44	80.86	81.15	1,905	723	38,350
M & R	82.37	82.07	82.22	1,592	640	34,390
M & L & R	80.99	81.34	81.16	5,177	2,627	52,756

- Contextualization improves F-score significantly
- Mother-both-sisters contextualization is *not* the best.
  - high number of categories and rules, with over half of the categories occurring only once → data sparsity

⇒ Idea: Keep only distributionally distinct new categories, i.e., new distinctions differing in probability distribution.

Clustering: Hierarchical Clustering

- Basic idea: For each original category, compare the expansion vectors of all its contextualized categories.
  - Example grammar (mother-contextualized categories):

NP_$S \rightarrow_{0.2}$ Det N NP_$S \rightarrow_{0.8}$ Pron	NP_S: $[0.2, 0.8]$
NP_$VP \rightarrow_{0.8}$ Det N NP_$VP \rightarrow_{0.2}$ Pron	NP_VP: $[0.8, 0.2]$
NP_$RelS \rightarrow_{0.3}$ Det N NP_$RelS \rightarrow_{0.7}$ Pron	NP_$RelS$: $[0.3, 0.7]$

Clustering: Hierarchical Clustering

- Basic idea: For each original category, compare the expansion vectors of all its contextualized categories.
  - Example grammar (mother-contextualized categories):

NP_$S \rightarrow_{0.2}$ Det N NP_$S \rightarrow_{0.8}$ Pron	NP_S: $[0.2, 0.8]$
NP_$VP \rightarrow_{0.8}$ Det N NP_$VP \rightarrow_{0.2}$ Pron	NP_VP: $[0.8, 0.2]$
NP_$RelS \rightarrow_{0.3}$ Det N NP_$RelS \rightarrow_{0.7}$ Pron	NP_$RelS$: $[0.3, 0.7]$

- Hierarchical Clustering assesses distributional similarity throughout a vector space.
  - Recursive, bottom-up algorithm that
    - begins with every vector in its own cluster and
    - merges the least distant clusters.

Clustering: Hierarchical Clustering

- Basic idea: For each original category, compare the expansion vectors of all its contextualized categories.
  - Example grammar (mother-contextualized categories):

NP_*S* $\rightarrow_{0.2}$ Det N NP_*S* $\rightarrow_{0.8}$ Pron	NP_*S*: $[0.2, 0.8]$
NP_*VP* $\rightarrow_{0.8}$ Det N NP_*VP* $\rightarrow_{0.2}$ Pron	NP_*VP*: $[0.8, 0.2]$
NP_*RelS* $\rightarrow_{0.3}$ Det N NP_*RelS* $\rightarrow_{0.7}$ Pron	NP_*RelS*: $[0.3, 0.7]$

- Hierarchical Clustering assesses distributional similarity throughout a vector space.
  - Recursive, bottom-up algorithm that
    - begins with every vector in its own cluster and
    - merges the least distant clusters.
  - We use Manhattan Distance as distance metric:

$$\text{Manhattan}(p, q) = \sum_{i=0}^{n} |p_i - q_i|$$

Clustering: Hierarchical Clustering

- Basic idea: For each original category, compare the expansion vectors of all its contextualized categories.
  - Example grammar (mother-contextualized categories):

NP_$S \rightarrow_{0.2}$ Det N NP_$S \rightarrow_{0.8}$ Pron	NP_S: $[0.2, 0.8]$
NP_$VP \rightarrow_{0.8}$ Det N NP_$VP \rightarrow_{0.2}$ Pron	NP_VP: $[0.8, 0.2]$
NP_$RelS \rightarrow_{0.3}$ Det N NP_$RelS \rightarrow_{0.7}$ Pron	NP_$RelS$: $[0.3, 0.7]$

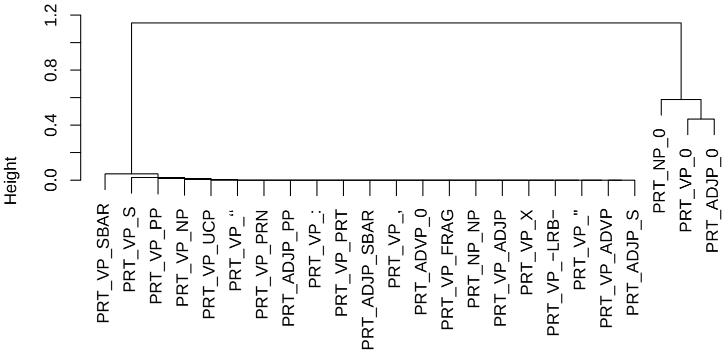
- Hierarchical Clustering assesses distributional similarity throughout a vector space.
  - Recursive, bottom-up algorithm that
    - begins with every vector in its own cluster and
    - merges the least distant clusters.
  - We use Manhattan Distance as distance metric:

$$\text{Manhattan}\,(p,q) = \sum_{i=0}^{n} |p_i - q_i|$$

    - similar results with Jensen-Shannon Divergence

Clustering: Hierarchical Clustering
Example for resulting dendrogram

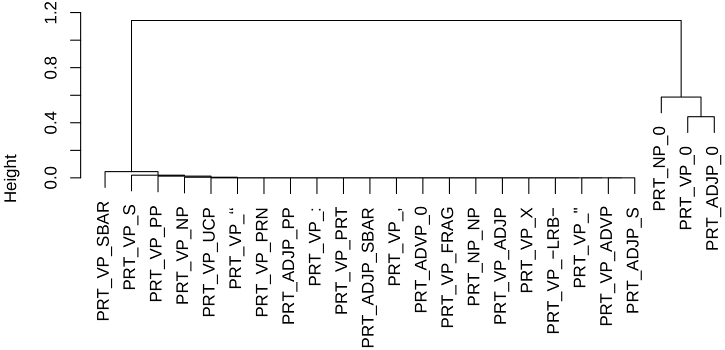
Particle (PRT) contextualized with Mother-Right-Sister:



Clustering: Hierarchical Clustering

Example for resulting dendrogram

Particle (PRT) contextualized with Mother-Right-Sister:

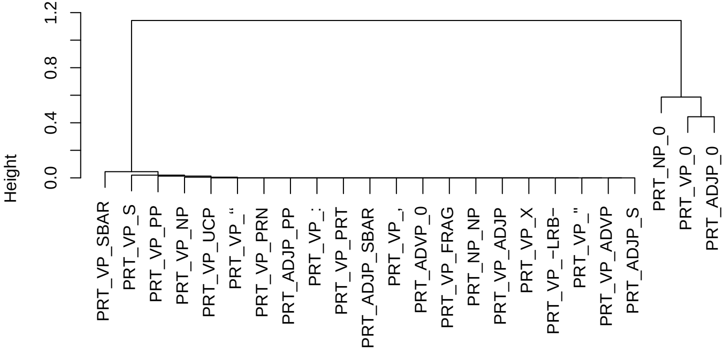


- Distance between *clusters* is the greatest distance found in pairwise comparisons of cluster members (“complete-link clustering”).

Clustering: Hierarchical Clustering

Example for resulting dendrogram

Particle (PRT) contextualized with Mother-Right-Sister:



- Distance between *clusters* is the greatest distance found in pairwise comparisons of cluster members (“complete-link clustering”).
  - Comparable results for vector averages and observed cluster member count.

Pruning the Dendrogram
Basics

- Clustered categories should consist of maximally intersubstitutable contextualized categories

Pruning the Dendrogram
Basics

- Clustered categories should consist of maximally intersubstitutable contextualized categories
- Prune subdendrogram to define clusters:
  - Pruning at root node returns original categories
  - Pruning at terminal nodes equivalent to full contextualization

Pruning the Dendrogram
Basics

- Clustered categories should consist of maximally intersubstitutable contextualized categories
- Prune subdendrogram to define clusters:
  - Pruning at root node returns original categories
  - Pruning at terminal nodes equivalent to full contextualization

⇒ Determine place to prune the dendrogram based on
  - distance between new categories (Experiment 2)
  - information gain over original category (Experiment 3)

Pruning the Dendrogram

Experiment 2: Pruning by Merge Height

- In hierarchical clustering every merge is performed between clusters at some distance from each other.
  - This distance is encoded in the **height** of the merge.

Pruning the Dendrogram

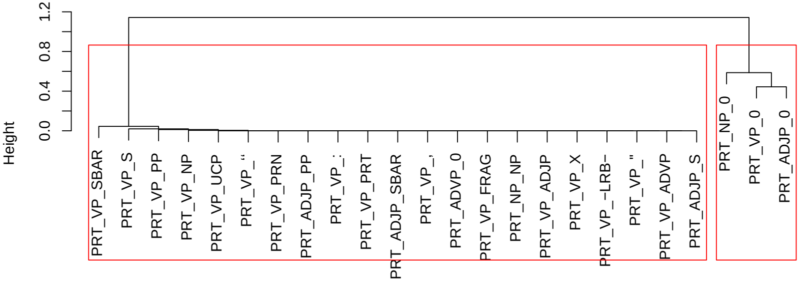
Experiment 2: Pruning by Merge Height

- In hierarchical clustering every merge is performed between clusters at some distance from each other.
  - This distance is encoded in the **height** of the merge.
- Prune dendrograms at nodes whose merge height is greater than some cut-off value.

Pruning the Dendrogram

Experiment 2: Pruning by Merge Height

- In hierarchical clustering every merge is performed between clusters at some distance from each other.
    - This distance is encoded in the **height** of the merge.
- Prune dendrograms at nodes whose merge height is greater than some cut-off value.
- Example: Pruning PRT dendrogram at height 0.7:



THE OHIO STATE UNIVERSITY

Pruning the Dendrogram

Background: Expected Information Gain

- Pruning according to merge height ensures that members of each cluster are similar to each other, but are clusters *different from original category*?

Pruning the Dendrogram

Background: Expected Information Gain

- Pruning according to merge height ensures that members of each cluster are similar to each other, but are clusters *different from original category*?
- The **Kullback-Leibler Divergence** (KLD) expresses the information wasted by using some distribution p to encode the behavior of distribution q:

$$\mathrm{KLD}(p,q) = \sum_{i=0}^{n} p_i \cdot log_2\left(\frac{p_i}{q_i}\right)$$

Pruning the Dendrogram
Background: Expected Information Gain

- Pruning according to merge height ensures that members of each cluster are similar to each other, but are clusters *different from original category*?
- The **Kullback-Leibler Divergence** (KLD) expresses the information wasted by using some distribution p to encode the behavior of distribution q:

$$\text{KLD}(p, q) = \sum_{i=0}^{n} p_i \cdot log_2\left(\frac{p_i}{q_i}\right)$$

- KLD(p, q) with p : Original Category
  q : Clustered Contextualized Category
  - = information wasted by pretending the contextualized categories covered by q behave like p

Pruning the Dendrogram

Background: Expected Information Gain

- Pruning according to merge height ensures that members of each cluster are similar to each other, but are clusters *different from original category*?
- The **Kullback-Leibler Divergence** (KLD) expresses the information wasted by using some distribution p to encode the behavior of distribution q:

$$\mathrm{KLD}(p,q) = \sum_{i=0}^{n} p_i \cdot log_2\left(\frac{p_i}{q_i}\right)$$

- $\mathrm{KLD}(p,q)$ with p : Original Category
  q : Clustered Contextualized Category
  - = information wasted by pretending the contextualized categories covered by q behave like p
  - = cost of ignoring contextual distinctions covered by q

Pruning the Dendrogram

Background: Expected Information Gain

- Pruning according to merge height ensures that members of each cluster are similar to each other, but are clusters *different from original category*?
- The **Kullback-Leibler Divergence** (KLD) expresses the information wasted by using some distribution p to encode the behavior of distribution q:

$$\mathrm{KLD}(p,q) = \sum_{i=0}^{n} p_i \cdot log_2\left(\frac{p_i}{q_i}\right)$$

- $\mathrm{KLD}(p,q)$ with p : Original Category
  q : Clustered Contextualized Category
  - = information wasted by pretending the contextualized categories covered by q behave like p
  - = cost of ignoring contextual distinctions covered by q

(Makes use of $0 \cdot log(0) = 0$ and $0 \cdot log(\frac{0}{0}) = 0$).

Pruning the Dendrogram

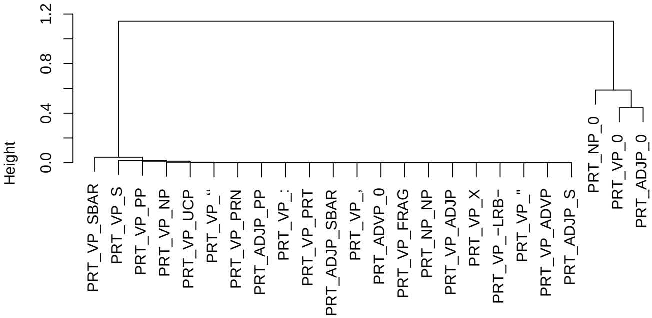
Experiment 3: Pruning by Expected Information Gain

- Descend from root into subdendrograms, assigning sufficiently divergent subdendrograms to new cluster
  - i.e., upon obtaining a KLD suffiently divergent from original category (given some cut-off value).

Pruning the Dendrogram

Experiment 3: Pruning by Expected Information Gain

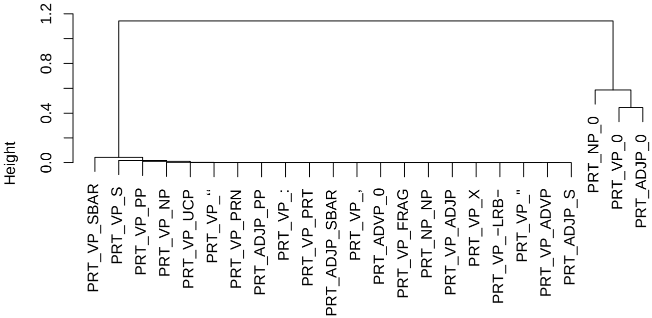
- Descend from root into subdendrograms, assigning sufficiently divergent subdendrograms to new cluster
  - i.e., upon obtaining a KLD suffiently divergent from original category (given some cut-off value).



Pruning the Dendrogram

Experiment 3: Pruning by Expected Information Gain

- Descend from root into subdendrograms, assigning sufficiently divergent subdendrograms to new cluster
  - i.e., upon obtaining a KLD suffiently divergent from original category (given some cut-off value).



- When descending the dendrogram to the leaves without encountering sufficient divergence, assign largest coherent subdendrogram to its own cluster.

Clustering Results

	F	Categ.	Nonce	Rules
Baseline	72.28	28	1	14,974
Mother and Both Sisters				
Unclustered	81.16	5,177	2,627	52,756
Height-Clust.	82.37	1,672	556	33,628
KLD-Clust.	82.28	495	87	28,781

Clustering Results

	F	Categ.	Nonce	Rules
Baseline	72.28	28	1	14,974
Mother and Both Sisters				
Unclustered	81.16	**5,177**	**2,627**	**52,756**
Height-Clust.	82.37	**1,672**	**556**	**33,628**
KLD-Clust.	82.28	**495**	**87**	**28,781**

- Both clustering approaches
  - dramatically reduce the number of categories, rules, and nonce categories

Clustering Results

	F	Categ.	Nonce	Rules
Baseline	72.28	28	1	14,974
Mother and Both Sisters				
Unclustered	81.16	5,177	2,627	52,756
Height-Clust.	**82.37**	1,672	556	33,628
KLD-Clust.	**82.28**	495	87	28,781

- Both clustering approaches
  - dramatically reduce the number of categories, rules, and nonce categories
  - while (marginally) improving the F-Score

Clustering Results

	F	Categ.	Nonce	Rules
Baseline	72.28	28	1	14,974
Mother and Both Sisters				
Unclustered	81.16	5,177	2,627	52,756
Height-Clust.	**82.37**	1,672	556	33,628
KLD-Clust.	**82.28**	495	87	28,781

- Both clustering approaches
  - dramatically reduce the number of categories, rules, and nonce categories
  - while (marginally) improving the F-Score
- Comparing the two clustering methods:
  - Virtually identical F-score

Clustering Results

	F	Categ.	Nonce	Rules
Baseline	72.28	28	1	14,974
Mother and Both Sisters				
Unclustered	81.16	5,177	2,627	52,756
Height-Clust.	82.37	**1,672**	**556**	**33,628**
KLD-Clust.	82.28	**495**	**87**	**28,781**

- Both clustering approaches
  - dramatically reduce the number of categories, rules, and nonce categories
  - while (marginally) improving the F-Score
- Comparing the two clustering methods:
  - Virtually identical F-score
  - KLD-clustering results in significantly fewer categories, nonce categories, and rules.

Clustering Results
A qualitative observation

- The globally optimized, single cut-off value used may be suboptimal for some (all?) individual categories.

Clustering Results

A qualitative observation

- The globally optimized, single cut-off value used may be suboptimal for some (all?) individual categories.
  - Example: PRT for mother-both-sisters demonstrates three linguistically-motivated groupings in dendrogram:
    1. Particle with no complement
    2. Non-displaced particle
    3. Displaced particle

Clustering Results
A qualitative observation

- The globally optimized, single cut-off value used may be suboptimal for some (all?) individual categories.
  - Example: PRT for mother-both-sisters demonstrates three linguistically-motivated groupings in dendrogram:
    1. Particle with no complement
    2. Non-displaced particle
    3. Displaced particle

    Neither clustering method 'correctly' distinguishes all and only these groupings.

Summary

- Stochastic framework allows for finer distributional distinctions within traditional categories.

Summary

- Stochastic framework allows for finer distributional distinctions within traditional categories.
- Encoding all local context distinctions improves parsing performance but exacerbates data sparsity.

Summary

- Stochastic framework allows for finer distributional distinctions within traditional categories.
- Encoding all local context distinctions improves parsing performance but exacerbates data sparsity.
- Clustering by distance or information gain of the contextualized categories results in smaller grammars
  - containing only the distributionally relevant contextualized categories, thus
  - generalizing better to unseen data and
  - improving parsing performance.

Outlook

- Optimize cut-off for each original category.

Outlook

- Optimize cut-off for each original category.
- Integrate back-off for rare contextualized categories to less specific, better attested ones:
  - e.g. from Mother-Both-Sisters to Mother-Left-Sister

Outlook

- Optimize cut-off for each original category.
- Integrate back-off for rare contextualized categories to less specific, better attested ones:
  - e.g. from Mother-Both-Sisters to Mother-Left-Sister
- Extend contextualization to terminals, using supertagger to pre-process input.
  - Preliminary results with gold contextualized part-of-speech tags obtained F-scores in low 90%

Outlook

- Optimize cut-off for each original category.
- Integrate back-off for rare contextualized categories to less specific, better attested ones:
  - e.g. from Mother-Both-Sisters to Mother-Left-Sister
- Extend contextualization to terminals, using supertagger to pre-process input.
  - Preliminary results with gold contextualized part-of-speech tags obtained F-scores in low 90%
- Explore category contextualization beyond syntactic information (→ prosodic information)

References

Gregory, M., M. Johnson and E. Charniak (2004). Sentence-Internal Prosody Does not Help Parsing the Way Punctuation Does.

Johnson, M. (1998). PCFG Models of Linguistic Tree Representations. *Computational Linguistics* 24(4), 613–632.

Klein, D. and C. Manning (2003). Accurate Unlexicalized Parsing. In *Proceedings of ACL-03*. Sapporo, Japan.

Petrov, S., L. Barrett, R. Thibaux and D. Klein (2006). Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of COLING-ACL-06*. Sydney, Australia.

Schmid, H. (2004). Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of COLING-04*. Geneva, Switzerland.

Sekine, S. and M. J. Collins (1997). EVALB bracket scoring program. web page, http://nlp.cs.nyu.edu/evalb/.