	Language Processing for Different Domains and Genres	
introduction		
approach	Detection of non-generalising rules	
material		
methods	Dickinson (2008) Dickinson & Foster (2009)	
experiments		
conclusion		
	by Fai Greeve	

	The general approach
introduction	
approach	Treebank
material	Rule extraction
methods	Filtering rules
experiments	Equivalence Frequency Similarity
conclusion	Cirinanty
10 December 2009	



Treebanks

introduction

approach

material

methods

experiments

conclusion

Penn Tree Bank •Wall Street Journal •The Brown Corpus •Switchboard •ATIS

British National Corpus

	Rule extraction
introduction	"John loves Mary"
approach	(S (NP (N John))
material	(VP (V loves)
	(NP (N Mary)))
methods	
	S> NP VP
experiments	NP> N
conclusion	VP> V NP
10 December 2009	

Ad-Hoc Rules
Rules used for specific constructions in one data set and unlikely to be used again.
For example:
•erroneous rules
 ungeneralizable rules rules for ungrammatical texts rules inconsistent with the rest of the annotation scheme.



Filtering by Frequency

introduction

approach

material

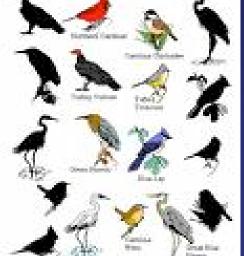
methods

experiments

conclusion











A less general approach

introduction

approach

material

methods

experiments

conclusion



Equivalence



similarity



	(non) equivalence
introduction approach	1) Remove daughter categories that are always non-predictive to phrase categorisation
material	2) Group head-equivalent lexical
methods	categories
experiments	
conclusion	

	Examples equivalence
introduction	Behind Woolfs modern lighthouse
	From the Hatters tea party
approach	[D Robind [Dro Moolfe [AD] modern [N] lighthouse1]]]
motorial	[P Behind [Pro Woolfs [ADJ modern [N lighthouse]]]] [P From [Det the [Pro Hatters [N tea [N party]]]]]]
material	
methods	P Pro AdJ N
	P Det Pro N N
experiments	
	P Pro N
conclusion	P Pro N N

	Levenshtein distance		
introduction	Measures the amount of difference between two sequences, a distance of 1		
approach	is highly similar	25, a Uistai	
material	start:	train	
methods			4
	Deletion	rain	+1
experiments	Insertion	rains	+2
conclusion	Substitution	gains	+3

	Modified Levenshtein distance
introduction	Deletion
approach	The cat died naturally (VP> V Adv) By deletion comparable to VP> V by 1 step
material	Insertion <i>The cat died</i> (VP> V)
methods	By insertion comparable to VP> V Adv by 1 step
experiments	Substitution * <i>The cat naturally</i> (VP> Adv)
conclusion	By substitution comparable to VP> V by only 1. Really?!

	Whole Daughters Scoring
introduction	1: Map a rule to its equivalence class
approach	2: For every rule token within the
material	equivalence class, add a score of 1
methods	3: For every rule token within a high similar equivalence class, add a score of 1/2.
experiments	
conclusion	
10 December 2000	

	Examples Whole Daughters Scoring	
introduction	For the equivalence class PP	
approach material	Compare: On the Wizards path P Pro N	
methods	To: Behind Woolfs modern lighthouse P Pro N +1	
experiments conclusion	From the Hatters tea party P Pro N N +1/2	

	Whole Daughters Scoring: Corpora Independent
introduction	1: For every identical rule token,
approach	add 1
material	2: For every highly identical rule token,
	add 1/2
methods	
experiments	
conclusion	
10 December 2009	

	Examples Whole Daughters Scoring Corpus Independent	
introduction	Compare:	
	On the Wizards path	
approach	P Det Pro N	
material	То:	
matchai	Behind Woolfs modern lighthouse	
methods	P Pro Adj N	+0
experiments	From the Hatters tea party	
	P Det Pro N N	+1/2
conclusion		

Advantages Whole Daughter S Corpus Independent	Scoring
introduction most western air fleets	
approach	N V
material WDS (old) score: 1,547	
methods because reduced NP> Adj N V is	
experiments to NP> Adj N	
conclusion WDSCI (new) score: 7	

	Disadvantages Whole Daughter Scoring Corpus Independent
introduction	"Quest for Fire" was the first time
approach	S> "NP" VP
material	WDS (old) score: 159,444
methods	because similar to
	reduced rule
experiments	S> NP VP
Conclusion	WDSCI (new) score: 0

	With and	witho	ut equiva	lence classes
introduction	Whole Daug	phter Sco	oring (old)	
approach	Threshold 1 25	Rules 311 2.683	Ungeneralizabilty 100% 97.50%	
material	50 100	3.548 4.596	<mark>96.93%</mark> 96.15%	ickinson et al. (2008) p. 366
methods	Whole Daugh	nter Scor		ependent (new)
	Threshold	Rules	, , ,	
	1	1625		
experiments	2	2.801 3.515		
	4	4.011		
conclusion	5	4.412		
				Dickinson et al. (2009) p. 6

	With and without equivalence classes Corpus dependent and corpus independent				
introduction	BNC 1000 training and evaluation				
approach	Old New	Threshold 35 3		Ungeneralizability 88.59% 94.14%	
material	Old New	50 5		88.51% 92.52%	
methods	WSJ train	ning and Bl	NC 100	0 evaluation	Dickinson et al. (2009) p. 6
		Threshold	Rules		
	Old		1600	98.92%	
experiments	New	1	1600	99.25%	
	Old	81	4300	96.84%	
	New	5	4300	98.66%	
conclusion				· · ·	Dickinson et al. (2009) p. 7
10 December 2009					

	Whole Daughters Scoring Corpora Independent
introduction	1) For every identical rule token,
approach	add 1
material	>frequency score
methods	2) For every highly identical rule token,
experiments	add 1/2 >similarity score
conclusion	
10 December 2000	

	Only Frequency: score results				
introduction					
approach		Threshold 1	Rules 8776	Ungeneralizabilty 98.30%	
material		2	10.741	97.52%	
		3 4	11.601 12.131	97.00% 96.64%	
methods				Dicl	kinson et al. (2009) p. 7
experiments					
conclusion					

	Only similarity: score results				
introduction					
	Threshold	Rules	Ungeneralizabilty		
approach	0	1851	98.27%		
	1	2.622	98.05%		
material	2	3.147	97.87%		
	4	3.865	97.52%	ļ	
methods			Dick	kinson et al. (2009) p. 8	
experiments					
conclusion					
10 December 2009					



conclusion

introduction

approach

material

methods

experiments

conclusion

Treebank Rule extraction Filtering rules Equivalence Frequency Similarity

Whole Daughter Scoring Corpus Independent

Complementary function of frequency and similarity



Bibliography

Main article:

Markus Dickinson and Jennifer Foster. 2009. Similarity Rules! Exploring Methods for Ad-Hoc Rule Detection. *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT-7 2009)*. Groningen, The Netherlands.

Background reading:

Markus Dickinson. 2008. Ad Hoc Treebank structures. *The 46th Annual Meeting* of the Association for Computational Linguistics (ACL) with the Human Language Technology Conference (HLT) (ACL-08). Columbus, OH.