Language Processing for Different Domains and Genres

Detection of non-generalising rules

Dickinson (2008)
Dickinson & Foster (2009)

by Fai Greeve

10 December 2009

The general approach

Treebank

Rule extraction

Filtering rules

Equivalence
Frequency
Similarity

Treebanks

Penn Tree Bank

- **Wall Street Journal**
- The Brown Corpus
- Switchboard
- ATIS

British National Corpus

Rule extraction

introduction

approach

material

methods

experiments

conclusion

“John loves Mary”

```
(S (NP (N John))
        (VP (V loves)
                (NP (N Mary)))
```

S --> NP VP

NP --> N

VP --> V NP

Ad-Hoc Rules

introduction

approach

material

methods

experiments

conclusion

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

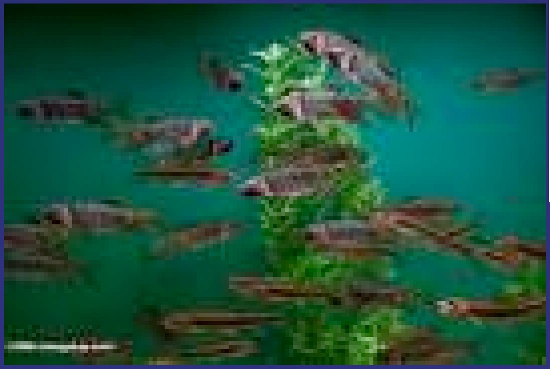
approach

material

methods

experiments

conclusion



A less general approach

Equivalence

similarity

(non) equivalence

introduction

approach

material

methods

experiments

conclusion

1) Remove daughter categories that are always non-predictive to phrase categorisation

2) Group head-equivalent lexical categories

Examples equivalence

Behind Woolfs modern lighthouse

From the Hatters tea party

[P Behind [Pro Woolfs [ADJ modern [N lighthouse]]]]

[P From [Det the [Pro Hatters [N tea [N party]]]]]]

P Pro AdJ N

P Det Pro N N

P Pro N

P Pro N N

introduction

approach

material

methods

experiments

conclusion

Levenshtein distance

Measures the amount of difference between two sequences, a distance of 1 is highly similar

start:	train	
Deletion	rain	+1
Insertion	rains	+2
Substitution	gains	+3

Modified Levenshtein distance

introduction

approach

material

methods

experiments

conclusion

Deletion

The cat died naturally (VP --> V Adv)

By deletion comparable to VP --> V by 1 step

Insertion

The cat died (VP --> V)

By insertion comparable to VP --> V Adv by 1 step

Substitution

**The cat naturally* (VP --> Adv)

By substitution comparable to VP --> V by only 1.
Really?!

Whole Daughters Scoring

1: Map a rule to its equivalence class

2: For every rule token within the equivalence class, add a score of 1

3: For every rule token within a high similar equivalence class, add a score of 1/2.

Examples Whole Daughters Scoring

introduction

approach

material

methods

experiments

conclusion

For the equivalence class PP

Compare:

On the Wizards path

P Pro N

To:

Behind Woolfs modern lighthouse

P Pro N +1

From the Hatters tea party

P Pro N N +1/2

Whole Daughters Scoring: Corpora Independent

1: For every identical rule token, add 1

2: For every highly identical rule token, add ½

Examples Whole Daughters Scoring Corpus Independent

introduction

approach

material

methods

experiments

conclusion

Compare:

On the Wizards path

P Det Pro N

To:

Behind Woolfs modern lighthouse

P Pro Adj N +0

From the Hatters tea party

P Det Pro N N +1/2

Advantages Whole Daughter Scoring Corpus Independent

introduction

approach

material

methods

experiments

conclusion

most western air fleets

NP--> AdjS Adj N V

WDS (old) score: 1,547

because reduced rule NP--> Adj N V is similar to NP --> Adj N

WDSCI (new) score: 7

Disadvantages Whole Daughter Scoring Corpus Independent

introduction

approach

material

methods

experiments

conclusion

“Quest for Fire” was the first time

S --> “NP” VP

WDS (old) score: 159,444

because similar to reduced rule

S --> NP VP

WDSCI (new) score: 0

With and without equivalence classes

Whole Daughter Scoring (old)

Threshold	Rules	Ungeneralizabilty
1	311	100%
25	2.683	97.50%
50	3.548	96.93%
100	4.596	96.15%

Dickinson et al. (2008) p. 366

Whole Daughter Scoring Corpus Independent (new)

Threshold	Rules	Ungeneralizabilty
1	1625	99.51%
2	2.801	99.43%
3	3.515	98.97%
4	4.011	98.85%
5	4.412	98.75%

Dickinson et al. (2009) p. 6

With and without equivalence classes
Corpus dependent and corpus independent

introduction

approach

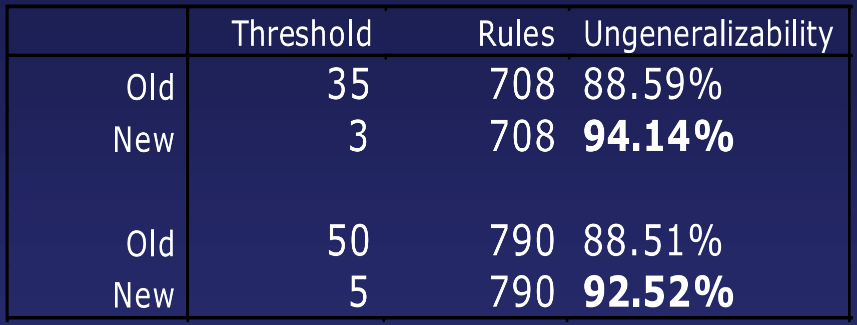
material

methods

experiments

conclusion

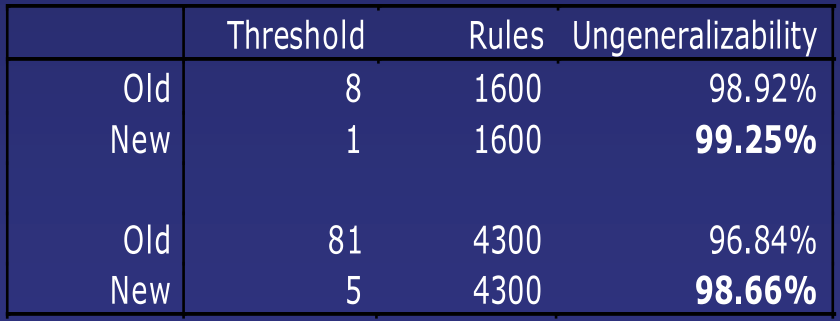
BNC 1000 training and evaluation



	Threshold	Rules	Ungeneralizability
Old	35	708	88.59%
New	3	708	**94.14%**
Old	50	790	88.51%
New	5	790	**92.52%**

Dickinson et al. (2009) p. 6

WSJ training and BNC 1000 evaluation



	Threshold	Rules	Ungeneralizability
Old	8	1600	98.92%
New	1	1600	**99.25%**
Old	81	4300	96.84%
New	5	4300	**98.66%**

Dickinson et al. (2009) p. 7

Whole Daughters Scoring
Corpora Independent

introduction

approach

material

methods

experiments

conclusion

1) For every identical rule token, add 1

>frequency score

2) For every highly identical rule token, add ½

>similarity score

Only Frequency: score results

Threshold	Rules	Ungeneralizabilty
1	8776	98.30%
2	10.741	97.52%
3	11.601	97.00%
4	12.131	96.64%

Dickinson et al. (2009) p. 7

Only similarity: score results

Threshold	Rules	Ungeneralizabilty
0	1851	98.27%
1	2.622	98.05%
2	3.147	97.87%
4	3.865	97.52%

Dickinson et al. (2009) p. 8

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.