

CUBeD

Creating and Using a Benchmark for German Distributional Semantic Models Maximilian Bacher, Isabell Wolter

Ruprecht-Karls-Universität, Institut für Computerlinguistik, Softwareprojekt Wintersemester 2013/2014

TASK

motivation: get a resource to evaluate Distributional Semantic Models that is able to identify semantically related words and to distinguish between different semantic relations (already has been done for English: BLESS dataset by Baroni & Lenci)

goals:

1. create a list of 200 unambiguous target-words and extract several words for each target belonging to certain relations (5 relations + 3 non-relations: needed to make sure that not every word is seen as related to target → “counterexample”)

2. evaluate the ability of DSMs to assign relations between words with the created gold-standard-list

RELATIONS

target	noun	Apfel
HYPER	noun	Obst
COORD	noun	Birne
MERO	noun	Kern
ATTR	adjective	sauer
EVENT	verb	essen
NONRN	noun	Baustelle
NONRA	adjective	laut
NONRV	verb	rennen

hypernym (hyperonym):
“superordinate” term of target

coordinate (cohyponym):
words sharing the same hypernym

meronym: part of target

attribute: characteristic of target

event: prototypical verbs associated with the target word

nonr[n|a|v]: nonrelata with same amount per part of speech as relata

RESOURCES + RELATA EXTRACTION

BLESS

- revision for German 200-words target-list
- translation of meronyms

GermaNet

- hypernyms
- coordinates
- non-relata nouns

Schulte im Walde

- attributes
- events

SdeWaC

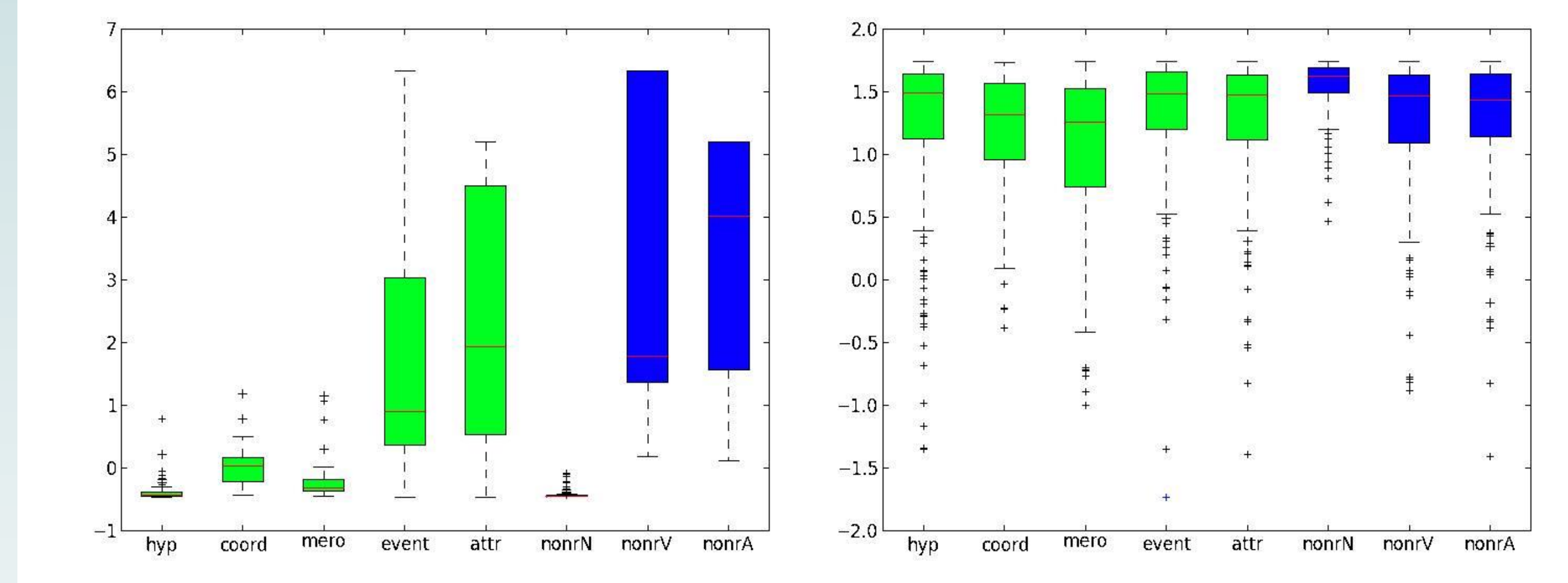
- frequency analysis
- extraction of non-relata via Pointwise Mutual Information

RELATA STATISTICS

overall amount of (non-)relata: **18,000**

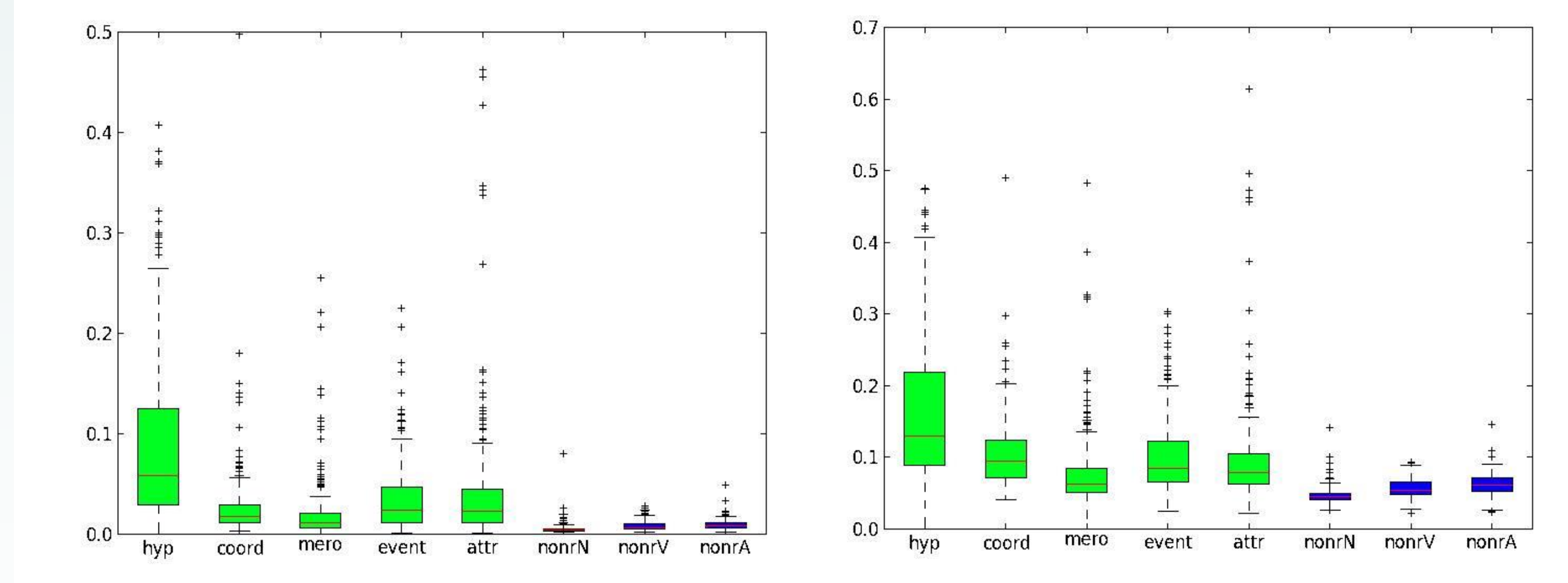
relation	relata count	average relata per target
HYPER	1263	6.315
COORD	2830	14.15
MERO	998	4.99
ATTR	1945	9.725
EVENT	1964	9.82
NONRN	5091	25.455
NONRA	1945	9.725
NONRV	1964	9.82

EVALUATION



frequency baseline

random cardinality baseline



2 content words window

20 content words window

DSMS

- **baselines:**
 - frequency: each relatum is assigned its frequency (normalized with z-score)
 - random cardinality: each relatum is assigned a random number between 0 and 100,000 (normalized with z-score)
- **cooccurrences:** counts of cooccurrences of each relatum with its target in a window of 2, 5 and 20 content words

AVERAGE PRECISION

frequency baseline:

HYPER	COORD	MERO	ATTR	EVENT	NONRN	NONRA	NONRV
0.24	0.16	0.17	0.08	0.06	0.56	0.08	0.06

random cardinality baseline:

HYPER	COORD	MERO	ATTR	EVENT	NONRN	NONRA	NONRV
0.07	0.15	0.05	0.11	0.11	0.29	0.11	0.11

2 content words window:

HYPER	COORD	MERO	ATTR	EVENT	NONRN	NONRA	NONRV
0.08	0.32	0.06	0.18	0.15	0.18	0.11	0.11

5 content words window:

HYPER	COORD	MERO	ATTR	EVENT	NONRN	NONRA	NONRV
0.09	0.32	0.07	0.16	0.14	0.18	0.11	0.10

20 content words window:

HYPER	COORD	MERO	ATTR	EVENT	NONRN	NONRA	NONRV
0.10	0.31	0.06	0.16	0.14	0.18	0.12	0.11