Defining a Gold Standard for a Swedish Sentiment Lexicon: Towards Higher-Yield Text Mining in the Digital Humanities

Jacobo Rouces, Lars Borin, Nina Tahmasebi, Stian Rødven Eide

University of Gothenburg

UNIVERSITY OF GOTHENBURG

Språk-BANKEN

DHN2018

Introduction

- There is an increasing demand for multilingual sentiment analysis
- There is limited coverage for sentiment lexicons in Swedish
- We build a GS to evaluate different automatic methods
  - Connotations are assumed to be part of the word sense
  - We use SALDO (Swedish lexicon structured in terms of word senses, semantics and morphology)

Methodology (1)

- Fine-grained Direct Annotation (DA) is
  - ✔ Straightforward
  - ✖ But problematic because human annotators don’t have (consciously) available numerical scores, but rather rely on comparisons. It is easy to become inconsistent with other annotators or even with one self over time (“scale drifting”)

Methodology (2)

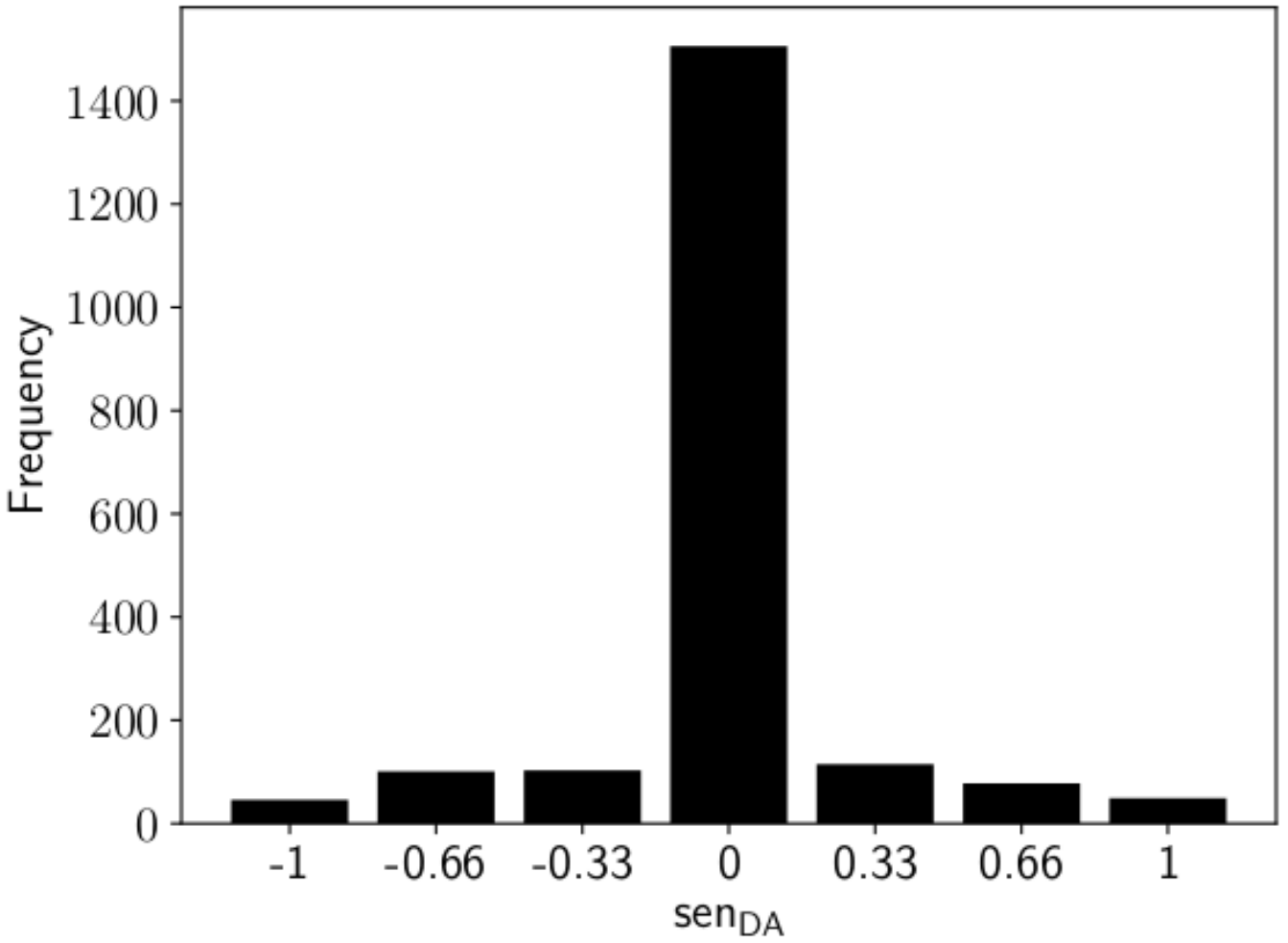
- Best-Worst Scaling (BWS) solves this by presenting annotators with 4-tuples of elements to select the one ranked highest and the one ranked lowest, and builds the numerical scores from the statistics of the choices
  - However, positive and negative words are rare (~10% each) so most 4-tuples do not offer a choice
- Therefore: multi-stage approach
  - Coarse-grained {-1,0,1} DA of 1998 elements by three annotators after joint annotation (synchrozing)
  - BWS scaling over DA with -2>=score>2

Direct Annotation

- Direct annotation
  - POS filtering SALDO entries
    - Out: Multi-words expressions, single-letter lemmas (names of letters, musical notes, units of measurement, etc.)
    - In: Single-word adjectives, interjections, nouns, and verbs having a lemma two letters or longer.
  - Sampling according to frequencies in Gigaword corpus.
  - Average of 3 annotators assigning {+1,0,-1} labels to 1998 SALDO entries.

Direct Annotation

Results



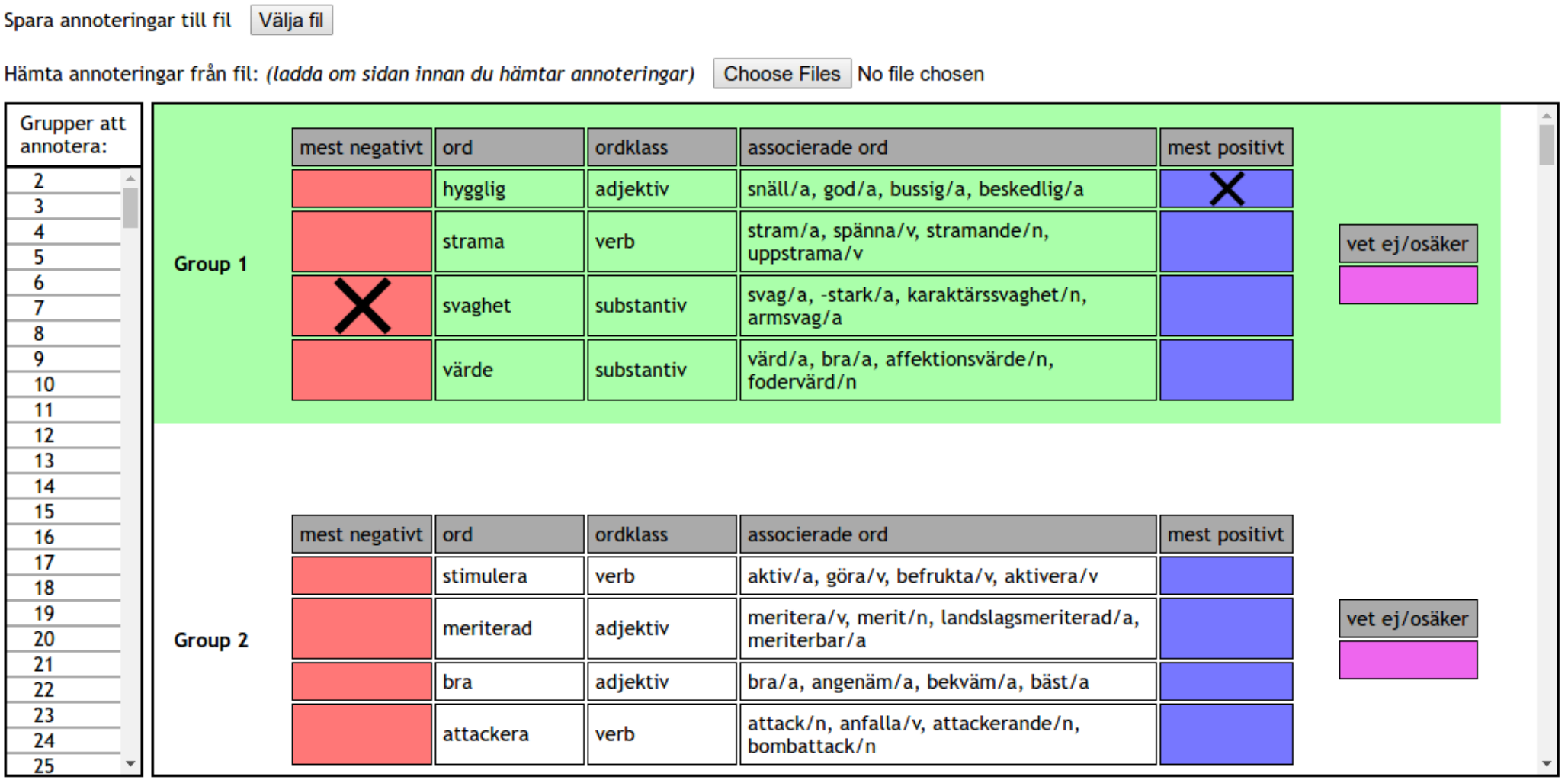
Best-Worst Scaling

- Best-Worst Scaling annotation
  - Users choose most positive and negative from each 4-tuple of SALDO entries
  - ~90% neutrals means that many tuples have no most positive or negative

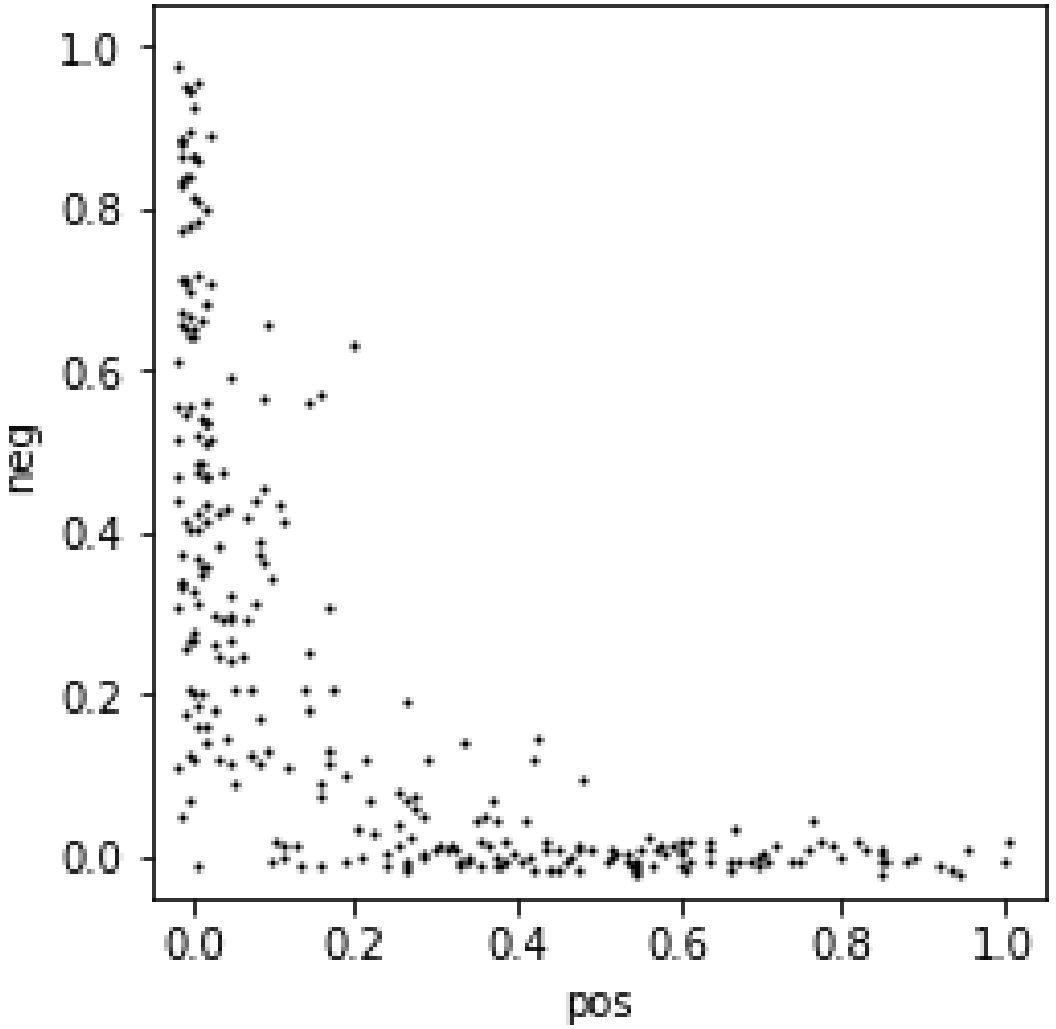
    → From direct annotation results, choose only those with |x|>0.5 (278 elements)
  - score = normalize(positives-negatives)

Best-Worst Scaling

Web interface (open-source, reusable)



Best-Worst Scaling



DHN2018

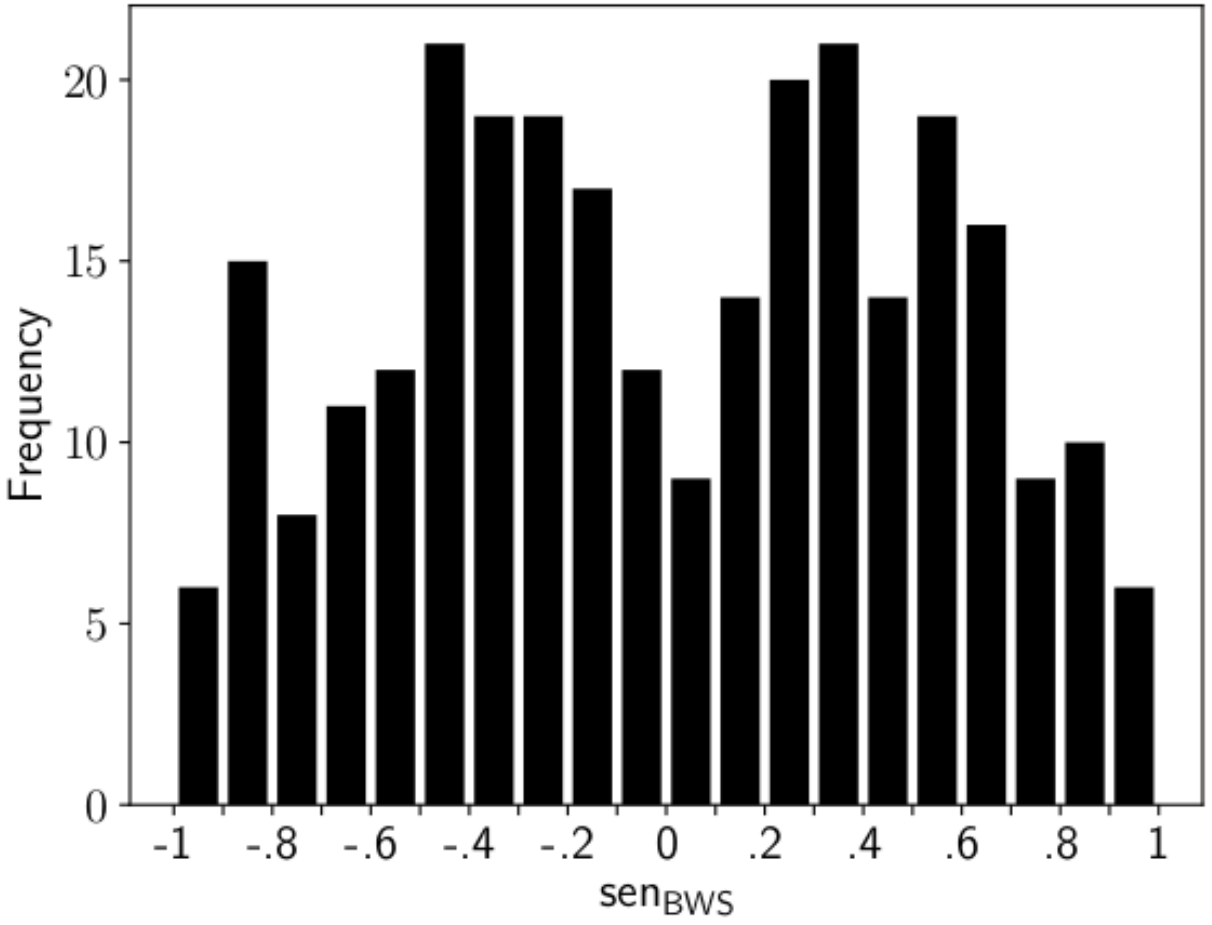
avg(min(pos,neg)) = 0.022
#(min(pos,neg)>0) 86 of 278

Highest values

SALDO ID	POS	NEG
slippa..1	0.250	0.194
dödlighet..1	0.188	0.625
protest..1	0.188	0.312
otänkbar..1	0.167	0.194
kapabel..1	0.438	0.156
klaga..1	0.156	0.562
oberäknelig..1	0.156	0.219
överdrift..1	0.179	0.143
lura..2	0.278	0.139
rädd..1	0.139	0.167
erfarenhet..1	0.219	0.125
ihärdig..1	0.156	0.125
kompetent..1	0.333	0.125
stressad..1	0.125	0.562
otillbörlig..1	0.125	0.250
meriterad..1	0.429	0.107
ohyra..1	0.107	0.429
spänning..3	0.464	0.107
svår..2	0.100	0.475
fel..2	0.100	0.425

Results

Results (4 annotators)



Results

Examples:

w	gloss	$\mathbf{pos}_{\mathbf{BWS}}(w)$	$\mathbf{neg}_{\mathbf{BWS}}(w)$	$\mathbf{neu}_{\mathbf{BWS}}(w)$	$\mathbf{sen}_{\mathbf{BWS}}(w)$
svår..1	'difficult'	0.0500	0.3250	0.6250	-0.2750
slippa..1	'be spared'	0.2500	0.1944	0.5556	0.0556
depression..2	'depression	0.0000	0.4688	0.5312	-0.4688
stimulera..1	'stimulate'	0.1250	0.0000	0.8750	0.1250
absurd..1	'absurd'	0.0625	0.4375	0.5000	-0.3750

Results

Inter-annotator agreements:

	nominal	**interval**
$\text{sen}_{\text{DA}}(w)$	0.480	0.529
$\text{pos}_{\text{BWS}}(w)$	0.551	0.889
$\text{neg}_{\text{BWS}}(w)$	0.621	0.893
$\text{neu}_{\text{BWS}}(w)$	0.446	0.744
$\text{sen}_{\text{BWS}}(w)$	0.462	0.927

Table 2: Interannotator agreements (Krippendorff's alpha, nominal and interval) for scores obtained from best-worst scaling (BWS) and direct annotation (DA). Since we used three annotators for sen_{DA}, in order to make the Krippendorff's alpha values comparable, we take the first 3 of the 4 annotators we used for BWS.

Future work: creation of complete lexicon

	DA									BWS
	ρ	τ_p	τ_b	precision	recall	acc.	confusion matrix			τ_b
							GS	SL		
								pos	neu	neg	
graph inheritance	0.39	0.39	0.38	pos: 0.28 neu: 0.91 neg: 0.33	pos: 0.26 neu: 0.90 neg: 0.42	0.82	pos neu neg	10 23 3	28 391 12	1 21 11	0.49
graph inheritance ext	0.33	0.42	0.32	pos: 0.22 neu: 0.90 neg: 0.27	pos: 0.21 neu: 0.89 neg: 0.35	0.81	pos neu neg	8 26 2	30 386 15	1 23 9	0.46
graph random paths	0.30	0.31	0.24	pos: 0.25 neu: 0.90 neg: 0.39	pos: 0.23 neu: 0.90 neg: 0.50	0.82	pos neu neg	9 26 1	29 390 12	1 19 13	0.46
word2vec +logit	0.47	0.21	0.38	pos: 0.37 neu: 0.93 neg: 0.46	pos: 0.54 neu: 0.88 neg: 0.52	0.84	pos neu neg	15 25 1	13 301 11	0 15 13	0.61
word2vec +svc /rbf	**0.55**	**0.15**	**0.45**	pos: 0.65 neu: 0.92 neg: 0.65	pos: 0.46 neu: 0.96 neg: 0.44	**0.89**	pos neu neg	13 7 0	15 328 14	0 6 11	**0.62**
graph random paths	0.32	0.31	0.25	pos: 0.18 neu: 0.90 neg: 0.50	pos: 0.18 neu: 0.89 neg: 0.56	0.82	pos neu neg	5 23 0	23 304 11	0 14 14	0.48

Table 1: Results for evaluating the different methods for constructing the sentiment lexicon in Swedish. Note that the Kendall tau τ_p is a distance, and therefore it is inversely related to the Spearman correlation ρ. GS and SL stand for gold standard and sentiment lexicon respectively.

Questions