SentiMerge: Combining Sentiment Lexicons in a Bayesian Framework

Guy Emerson
Thierry Declerck

LG-LP, 2014

Motivation

Sie zeichnet sich durch wunderbare Melodien

Motivation

Sie zeichnet sich durch wunderbare Melodien

Motivation

Sie zeichnet sich durch wunderbare Melodien

Motivation

Lemma	Sentiment
wunderbar	+
falsch	-
Angst	-
Frieden	+
unantastbar	+

Motivation

Lemma	Sentiment
wunderbar	1.0
falsch	-0.9
Angst	-0.8
Frieden	0.9
unantastbar	0.9

Motivation

Lemma	Sent. 1	Sent. 2	Sent. 3
wunderbar	1.0	0.4	0.8
falsch	-0.9	-0.3	-0.9
Angst	-0.8	-0.4	-1.0
Frieden	0.9	0.5	1.0
unantastbar	0.9	0.3	-0.8

Outline

- Data Sources
- Normalising Scores
- Combining Scores
- Evaluation

Data Sources

- Clematide and Klenner (C&K)
- SentimentWortschatz
- GermanSentiSpin
- GermanPolarityClues

Data Sources

- Clematide and Klenner (C&K)
- SentimentWortschatz
- GermanSentiSpin
- GermanPolarityClues
- *MLSA corpus (for evaluation)*

Data Sources

	vergöttern, V
C&K	1.000
PolarityClues	0.333
SentiWS	0.004
SentiSpin	0.245

Normalising Scores

- Intuitively: if a source has small scores, increase them; if large scores, decrease them
- Formally: minimise the square difference between sources (under a suitable constraint)

Normalising Scores

- Consider words in the overlap between two sources
- For each source, divide the scores by the root mean square

$$\sqrt{\frac{1}{n}\sum_i x_i^2}$$

Normalising Scores

- For each pair of sources, calculate the root mean squares in the overlap
- Average for each source

Normalising Scores

Lexicon	Root mean square
C&K	0.845
PolarityClues	0.608
SentiWS	0.267
SentiSpin	0.560

Normalised Data Sources

	vergöttern, V
C&K	1.183
PolarityClues	0.547
SentiWS	0.015
SentiSpin	0.438

Combining Scores

- Assume true polarity values distributed normally
- Assume each source independently introduces a linear error term, also normal

Combining Scores

	Variance
Prior	0.528
C&K	0.328
PolarityClues	0.317
SentiWS	0.446
SentiSpin	0.609

Combining Scores

$$\hat{x} = \frac{\sum \sigma_a^{-2} x_a}{\sigma^{-2} + \sum \sigma_a^{-2}}$$

Combined Data Sources

	vergöttern, V
C&K	1.183
PolarityClues	0.547
SentiWS	0.015
SentiSpin	0.438
Sentimerge	0.509

Evaluation Data

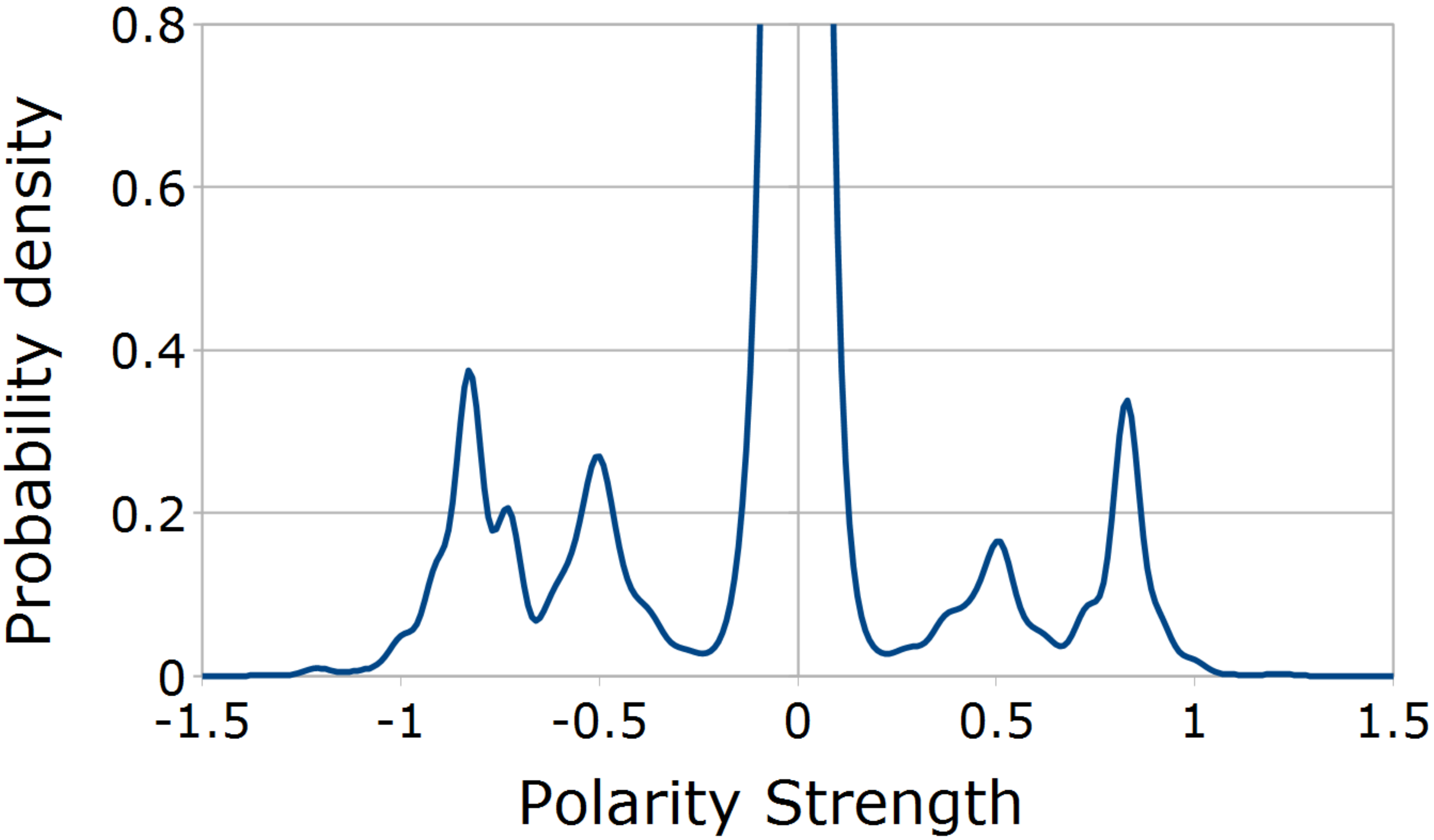
- Content words in MLSA
  - Sentiment from 2nd layer
  - Lemma and PoS from 3rd layer (with manual correction)
- 1001 types, 1424 tokens
- 378 positive, 399 negative

Discretisation

- Test data polarities are simply positive or negative

- Need a threshold to convert from numerical scores

Discretisation



Evaluation

Lexicon	Precision	Recall	F-score
C&K	0.754	0.733	0.743
PolarityClues	0.705	0.564	0.626
SentiWS	**0.803**	0.513	0.621
SentiSpin	0.557	0.668	0.607
Majority vote	0.548	**0.898**	0.679
SentiMerge	0.708	0.815	**0.757**

Conclusion

- Bayesian combination outperformed original sources, as well as a baseline method
- Merged resource being published as part of LLOD