Evaluation and Extension of a Polarity Lexicon for German

Simon Clematide & Manfred Klenner
{simon.clematide, klenner}@cl.uzh.ch

Institute of Computational Linguistics
University of Zurich

WASSA 2010

Background and Goals

PolArt project: http://kitt.cl.uzh.ch/kitt/polart

Multi-lingual compositional sentiment analysis (en, fr, de)

Automatic extension of a prior polarity lexicon of adjectives

- Corpus-based lexicon extension: Which strategy? (Semi-)Automatic?
- Classification experiment: To what degree can we predict polarity orientation and its strength automatically?
- Reliability experiment: How reliable are intellectual polarity decisions?

Why adjectives?

- In general: Recognition of evaluative adjectives is crucial for sentiment detection [Bruce and Wiebe, 1999]
- In particular: Following the results of an application-based evaluation of PolArt

Approaches for (Semi-)Automatic Lexicon Extension

- Coocurrence in the Web [Baroni and Vegnaduzzo, 2004]:
  High *Mutual Information* $\approx$ polarity agreement
- Relational lexical semantics (WordNet) [Kamps et al., 2004]:
  Synonymy $\approx$ same orientation
  Antonymy $\approx$ opposed orientation
- Interesting combinations [Baccianella et al., 2010]: Coocurrence in WordNet glosses (SentiWordNet)
- Translation of sentiment lexica [Waltinger, 2010]
- Occurrencies of coordinated adjectives...

Our Initial Adjective Lexicon

%	Freq	Pol	Examples (randomly selected)
27.1	785	–h	sadistisch (*sadistic*) idiotisch (*idiotic*)
19.5	566	–m	arglos (*unsuspecting*) ablehnend (*refusing*)
19.5	565	+h	schwärmerisch (*enthusiastic*) fachkundig (*expert*)
18.4	533	+m	kühn (*bold*) fruchtbar (*seminal*)
8.8	255	–l	stiefmütterlich (*stepmotherly*) arm (*poor*)
6.7	195	+l	real (*real*) wuchtig (*bulky*)
Total	2899		

Table: Distribution of the polarity classes in our lexicon: Pol(arity): h=high, m=medium, l=low

Negative adjectives are in the majority with 55.4%.
For the classification experiment 2850 adjectives were selected.

Automatic Polarity classification (+/–)

Approach of [Hatzivassiloglou and McKeown, 1997]

"[. . .] conjunctions between adjectives provide indirect information about orientation."

Coordination hypothesis

Coordinated subjective adjectives do have a statistically significant bias towards same orientation polarity.

Example (p-value of [Hatzivassiloglou and McKeown, 1997])

78% of 2748 types of coordinated adjectives have same orientation. Assuming equal distribution of adjectives, the probability of getting 78% or more is lower than 10^{-16}.

Preparation of a German Corpus

Use of `http://wortschatz.uni-leipzig.de` by the way of the PERL SOAP client `wsws.pl`

Application flow

1. For each lexicon entry generate all inflected variants
   `$ wsws.pl Wordforms hilflos` → hilflos hilflose hilflosen hilfloser hilfloses hilflosem hilflosesten hilfloseren hilfloseste hilflosere (*helpless*)
2. Request example sentences (max. 256 per inflected variant):
   `$ wsws.pl Sentences hilfloseren`
3. Chunk sentences by `chunkie`
4. Lemmatize by morphological analyser `GERTWOL`
5. Extract coordinated adjective pairs

Extraction of Coordinated Pairs: An Example

Sentence

Es ist ein veritables Labyrinth mit idyllischen, romantischen und gruseligen Zutaten. (*It's a real maze with idyllic, romantic and scary ingredients.*)

Chunking output with tripartite coordinated adjective phrase

(PPER Es) (VAFIN ist) (NP (ART ein) (ADJA veritables) (NN Labyrinth)) (PP (APPR mit) **(CAP (ADJA idyllischen) ($, ,) (ADJA romantischen) (KON und) (ADJA gruseligen))** (NN Zutaten)) ($. .)

Extracted adjacent pairs, alphabetically ordered

1. "idyllisch/romantisch" (*idyllic/romantic*)
2. "gruselig/romantisch" (*scary/romantic*)

The results of our chunker are quite faulty. For reasons of precision, we did without transitive pairs as "gruselig/idyllisch".

Statistics on Types of Coordinated Pairs I

# Adj	570	1140	1710	2280	2850
Sent	852.8	796.6	753.6	736.8	715.5
AA	50.3	45.6	41.4	38.2	35.6
$\overline{AA}$	29.4	30.6	30.3	29.8	29.2
$\bar{A}\bar{A}$	2.4	4.9	7.4	9.8	12.3
$_{\pm}\vec{A}\vec{A}$	1.8	3.7	5.7	7.5	9.5
$_{\pm 3}\vec{A}\vec{A}$	0.8	1.7	2.6	3.4	4.4

Adj:
Number of used lexicon entries
Sent:
Mean number of sentences per lexicon entry containing at least one adjective: decreasing (one sentence may contain more than one adjective)
AA:

Statistics on Types of Coordinated Pairs II

Mean number of types of coordinated adjective pairs per lexicon entry: decreasing (new ones get more rare)

$\overline{AA}$:

Mean number of types of coordinated adjective pairs with at least one adjective from our lexicon: Constant

$\bar{A}\bar{A}$:

Mean number of types of coordinated adjective pairs with both adjectives from our lexicon: Increasing proportionally

$_{\pm}\vec{A}\vec{A}$:

Mean number of types of coordinated pairs with same-orientation adjectives (only +/–) from our lexicon: Increasing proportionally

$_{\pm 3}\vec{A}\vec{A}$:

Mean number of types of coordinated pairs with same-orientation adjectives (+/–h, +/–m, +/–l) from our lexicon: Increasing proportionally

Sparse data problem

Statistics on Types of Coordinated Pairs III

249 adjectives never show up in a coordinated pair in combination with a known adjective partner.
150 only with a single partner.
140 only with 2 partners.

Testing the Coordination Hypothesis for German (+/–)

Occurrences of coordinated adjective pairs using the sentences from the whole test lexicon (2850 lemmas)

- Frequency of the types of category $\bar{A}\bar{A}$: 35156
- Distribution of the polarity: +: 54% –: 46%

Chi-Square-Test by R

	++	+–	––
Expected Frequency	0.30	0.50	0.20
Empirical Frequency	0.43	0.23	0.34

```
X-squared = 10326.55, df = 2, p-value < 2.2e-16
```

Coordination Hypothesis w.r.t. Polarity Strength: Winners

Pair	Expected	Empirical	Difference
-h-h	5.2	11.1	+5.9
+h+m	11.5	16.6	+5.1
+h+h	6.9	11.0	+4.1
-h-m	7.3	10.3	+3.0
+m+m	4.8	7.1	+2.3
-m-m	2.5	4.6	+2.1
-m-l	2.1	3.5	+1.4
+m+l	2.9	3.8	+1.0
+h+l	3.4	4.1	+0.7
-h-l	3.0	3.7	+0.7
-l-l	0.4	0.7	+0.3
+l+l	0.4	0.7	+0.3

Observation: Strong polarity with same orientation profits most!

Coordination Hypothesis w.r.t. Polarity Strength: Losers

Pair	Expected	Empirical	Difference
+h-h	12.1	4.4	-7.7
+m-h	10.0	3.7	-6.3
+h-m	8.3	3.6	-4.7
+m-m	6.9	3.6	-3.3
+h-l	3.4	1.8	-1.6
+l-h	3.0	1.5	-1.5
+m-l	2.9	1.8	-1.1
+l-m	2.1	1.4	-0.6
+l-l	0.9	0.8	-0.1

Observation: Weak oppositions distribute randomly!

Automatic Classification: "Baseline"

Decision rule for an adjective x

1. Count all occurrences of all known subjective adjectives which appear combined with x in a coordinated pair.
2. Set the orientation of x to the orientation of adjective z which co-occurs most often with x.

Learning Rates (F-Measure) for Strength Classification

Pol	E	S1	S2	S3	S4	S5
+h	F	$32_{\pm 16}$	$39_{\pm 8}$	$45_{\pm 8}$	$50_{\pm 4}$	$52_{\pm 4}$
+m	F	$23_{\pm 8}$	$30_{\pm 8}$	$25_{\pm 9}$	$27_{\pm 5}$	$30_{\pm 8}$
+l	F	$0_{\pm 0}$	$12_{\pm 13}$	$13_{\pm 8}$	$6_{\pm 10}$	$9_{\pm 6}$
-l	F	$3_{\pm 11}$	$0_{\pm 0}$	$8_{\pm 7}$	$8_{\pm 4}$	$5_{\pm 4}$
-m	F	$17_{\pm 16}$	$30_{\pm 10}$	$33_{\pm 6}$	$33_{\pm 4}$	$33_{\pm 5}$
-h	F	$35_{\pm 10}$	$50_{\pm 10}$	$54_{\pm 6}$	$57_{\pm 7}$	$60_{\pm 4}$

Table: Learning rates of F-measure for baseline algorithm with ten-fold cross-validation: $Sn = 2850 \times n/5$.

Binary Orientation Classification with Baseline

Pol	E	S1	S2	S3	S4	S5
+	P	$75_{\pm 11}$	$81_{\pm 5}$	$84_{\pm 3}$	$83_{\pm 3}$	$84_{\pm 4}$
+	R	$63_{\pm 11}$	$72_{\pm 6}$	$74_{\pm 3}$	$77_{\pm 5}$	$81_{\pm 4}$
+	F	$67_{\pm 7}$	$76_{\pm 4}$	$79_{\pm 3}$	$80_{\pm 4}$	$82_{\pm 3}$
-	P	$82_{\pm 9}$	$87_{\pm 5}$	$90_{\pm 5}$	$91_{\pm 4}$	$93_{\pm 3}$
-	R	$42_{\pm 5}$	$63_{\pm 6}$	$72_{\pm 5}$	$74_{\pm 5}$	$76_{\pm 3}$
-	F	$55_{\pm 4}$	$73_{\pm 5}$	$80_{\pm 5}$	$81_{\pm 4}$	$83_{\pm 2}$

Table: Learning rates for baseline algorithm with ten-fold cross-validation: S$n = 2850 \times n/5$.
E(valuation measure): P=precision, R=recall, F=F-Measure

Binary Classification with Maximum Entropy Approach

Training idea

- For each subjective adjective, compute the set of all other subjective adjectives that co-occur in an extracted coordination pair (so-called *coordination fellows*).
- For each positive adjective each positive coordination fellow acts as a feature. In the same way, for each negative adjective each negative coordination fellow acts as a feature.
- To account for pure frequency effects which proved to be powerful in the baseline algorithm, several features based on raw counts were defined: For example, whether at least 60, 70, or 80 percent of all occurrences of coordination fellows of an adjective are positive or negative.

We used the `megam` tool.

Binary Classification with Maximum Entropy Approach

Pol	E	S1	S2	S3	S4	S5
+	P	$77_{\pm 9}$	$84_{\pm 5}$	$87_{\pm 4}$	$87_{\pm 3}$	$87_{\pm 4}$
+	R	$61_{\pm 10}$	$71_{\pm 6}$	$75_{\pm 3}$	$78_{\pm 4}$	$80_{\pm 4}$
+	F	$68_{\pm 6}$	$77_{\pm 5}$	$81_{\pm 3}$	$82_{\pm 3}$	$84_{\pm 2}$
-	P	$78_{\pm 10}$	$84_{\pm 6}$	$89_{\pm 4}$	$90_{\pm 4}$	$90_{\pm 4}$
-	R	$51_{\pm 6}$	$69_{\pm 6}$	$78_{\pm 4}$	$80_{\pm 3}$	$82_{\pm 2}$
-	F	$61_{\pm 5}$	$76_{\pm 5}$	$83_{\pm 4}$	$85_{\pm 3}$	$86_{\pm 2}$

Table: Learning rates with ten-fold cross-validation: S$n = 2850 \times n/5$

About 3 percent better than our base line. Behaves as expected: More training data = better results!

Problems of Strength Classification

Problem

Learning rates for strength classification did not converge with more training data using machine learning.

Possible reasons

- Wrong classification approach
- Noisy training data

Next step

Determine the inter-annotator agreement of our sentiment classifications in the lexicon.

Problem and Survey

How reliable are polarity orientation judgements applied to lexicon entries in isolation?

Experiment

- 20 persons
- 7 polarity classes:
  Highly positive (+h),
  Medium positive (+m),
  Neutral (0),
  Medium negative (-m),
  Highly negative (-h),
  Not decidable (na)
- Timeout: 12 seconds at most

60 randomly selected adjectives from our lexicon (including neutral ones)

Distributions of Orientation Classifications

Orientation	+	−	neut	N/A
Relative frequencies (all test persons)	0.43	0.26	0.21	0.10
Relative frequencies (from lexicon)	0.53	0.37	0.10	

[1]

Question

How to compare our PolArt lexicon with our test persons?

[1] Unfortunately, our random sample from the lexicon had a bias towards positive items.

What's the General Polarity? Voting

- Majority decides
- What to do with ties? Choose the most frequent category!
- Is there a measure for the randomness of a decision? (measure of variability)

Relative entropy of a categorial variable: H_{rel}

- $H_{rel} = 1$ if each category is chosen by the same amount of persons
- $H_{rel} = 0$ if everyone chooses the same category
- Formally: $H_{rel} = -\frac{\sum_{i=1}^{n}(p_i \times log(p_i))}{log(n)}$

The Majority Decision: Positives

Some examples ordered by frequency

Adjective	Pol (PolArt)	Freq	H_{rel}
ehrlich (*honest*)	+	20	0.00
blitzschnell (*lightning*)	+	19	0.14
gedankenreich (*rich in ideas*)	+	18	0.23
gradlinig (*straight*)	+	17	0.30
sorgenlos (*carefree*)	+ (–)	17	0.30
energisch (*energetic*)	+ (–)	16	0.46
aufopfernd (*devoted*)	+	15	0.53
leistungsfoerdernd (*efficiency increasing*)	+	14	0.59
folgerichtig (*consequential*)	+	13	0.47
anruehrend (*touching*)	+	12	0.49
schuldlos (*innocent*)	+	11	0.81
meistgespielt (*most often played*)	+	10	0.62
atmosphaerisch (*atmospheric*)	+	10	0.79

Question: Is entropy a good indicator for orientation ambiguity?

The Majority Decisions: Negatives

Some examples ordered by frequency

Adjective	Pol (PolArt)	Freq	H_{rel}
unausstehlich (*insufferable*)	–	20	0.00
desorientiert (*disoriented*)	–	18	0.23
unedel (*ignoble*)	–	17	0.37
unnoetig (*unnecessary*)	–	15	0.50
unchristlich (*unchristian*)	–	14	0.59
unangepasst (*unadapted*)	–	13	0.74
melodramatisch (*melodramatic*)	–	12	0.77
sprachbehindert (*speech impaired*)	–	11	0.70
betaeubt (*stunned*, *dazed*)	–	10	0.74
monarchisch (*monarchic*)	– (0)	9	0.68

Differences between PolArt and majority have typically low frequency and high entropy.

The Majority Decisions: Neutrals

Some examples ordered by frequency

Adjective	Pol (PolArt)	Freq	H_{rel}
zeichnerisch (*graphic*)	0	17	0.42
surreal (*surreal*)	0	14	0.66
schicksalhaft (*fateful*)	0 (–)	11	0.61
taubstumm (*deaf-mute*)	0 (–)	11	0.67
riesenhaft (*gigantic*)	0	11	0.81
angenommen (*assumed*)	0	10	0.72
laeuferisch (*running*)	0 (+)	10	0.82
nichtbehindert (*non-handicapped*)	0 (+)	10	0.87
saturiert (*satisfied*)	0 (+)	8	0.94
kritisch (*critical*)	0 (–)	7	0.96

Question: Where to draw the line between low polarity and neutral sentiment? It's ok for sentiment lexicons to boost polarity.

Correlation: Cohen's Kappa

How strong is the agreement between two vectors of categorizations?

Degree of agreement

- kappa = 0.81-1.00: almost perfect
- kappa = 0.61-0.80: strong
- kappa = 0.41-0.60: moderate
- kappa = 0.10-0.40: weak
- kappa = 0: random

Accuracy (Acc)

Number of correctly classified items

Agreement with majority (orientation classification)

Person	Kappa	Acc
4	0.82	88.33
3/5	0.79	86.67
14	0.74	83.33
9	0.72	81.67
11	0.71	81.67
13/19	0.69	80.00
1	0.64	75.00
12	0.64	76.67
...		
17	0.43	61.67

Agreement between Majority Decision and PolArt

Orientation classification

Agreement	Kappa	Acc
Persons with PolArt (Mean)	0.5	70%
Persons with majority (Mean)	0.6	76%
Majority with PolArt	0.7	82%

Strong correlation of majority with PolArt.

Polarity strength classification

Agreement	Kappa	Acc
Persons with PolArt (Mean)	0.3	46%
Persons with majority (Mean)	0.5	62%
Majority with PolArt	0.5	60%

Moderate correlation of majority with PolArt.

Lexicon Extension

We find a lot of adjectives in coordinated pairs which are not (yet) part of our lexicon.

Generation of candidates

How can we reliably identify adjectives with polarity orientation?

1. Trial: Unknown adjectives that share the most coordination fellows with a known subjective adjective. Computationally complex and unsatisfactory.
2. Trial: Unknown adjectives that have the highest proportion of subjective adjective fellows and occur beyond a certain threshold. Simple and effective.

Semi-automatic Lexicon Extension

- No fully-automatic extension – fine-grained strength classification was needed, but didn't work well enough
- Strategy: Generate high quality candidates for human decision

- Round 0: 2893 adjectives with subjective orientation
- Round 1: 668 candidates (43 completely neutral)
- Round 2: 250 candidates (30 completely neutral)
- Round 3: ...

Automatic Classification

Feasible for the binary polarity orientation task

Conclusion

Automatic orientation classification

- using coordinated adjectives performs comparable to human classification. 10-best human annotators have the following mean F-measures (+: 86%; –: 80%) with respect to PolArt.
- is feasible (if enough data is available)

Polarity strength classification

is hard, for humans as well as for machine learning

Conclusion

- We had a sparse-data problem with our approach. Web-based approaches might help.
- Rare orientation strength classifications as +/–low are hard to learn. We got rid of them.
- We treat polarity as a property of lexicon entries. However, it's a property of word senses. We should allow more than 1 orientation classification per lexicon entry.
- Polarity orientations that are specific to word senses may be expressed by context restrictions of typical collocations, e.g. (sorgloser Umgang (*thoughtless handling*) vs. sorgloses Alter (*carefree age*))
- Differences between "factual" (gehbehindert *mobility impaired*) vs. "subjective" Valuation (spitzfindig *oversubtle*) need clarification

Thank you for the attention[2]!

A demo of compositional sentiment detection and our German lexicon is available under `http://kitt.cl.uzh.ch/kitt/polart`

[2]We would like to thank all students and group members for taking part in the experiment. And last, but not least, Michael Wiegand and Ronny Peter for manual lexicon entry curation.

References I

- Baccianella, S., Esuli, A., and Sebastiani, F. (2010).

  SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.

  In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC'10), Valletta, MT*, pages 2200–2204.

- Baroni, M. and Vegnaduzzo, S. (2004).

  Identifying subjective adjectives through web-based mutual information.

  In *In Proceedings of the 7th Konferenz zur Verarbeitung Natürlicher Sprache (German Conference on Natural Language Processing – KONVENS'04*, pages 613–619.

- Bruce, R. F. and Wiebe, J. M. (1999).

  Recognizing subjectivity: a case study in manual tagging.

  Natural Language Engineering, 5(02):187–205.

References II

- Hatzivassiloglou, V. and McKeown, K. R. (1997).

  Predicting the semantic orientation of adjectives.

  In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Morristown, NJ, USA. Association for Computational Linguistics.

- Kamps, J., Marx, M., Mokken, R. J., and de Rijke, M. (2004).

  Using WordNet to measure semantic orientation of adjectives.

  In *The fourth international conference on Language Resources and Evaluation (LREC)*, volume 4, pages 1115–1118.

- Waltinger, U. (2010).

  GermanPolarityClues: A lexical resource for German sentiment analysis.

  In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC'10), Valletta, MT*, pages 1638–1642. European Language Resources Association (ELRA).