Fine-grained Sentiment Analysis on German Twitter

Wladimir Sidorenko
uladzimir.sidarenka@uni-potsdam.de

University of Potsdam

July 10, 2014

Table of Contents

Unknown words are really a problem for existing NLP-tools:

Example

Leg_NN den_ART Karl_NE weg_PTKVZ ,_$, denn_KON kannste_VVFIN immer_ADV noch_ADV hauen_VVINF ,_$, der_ART heutige_ADJA @Tatort_NN ist_VAFIN mal_ADV wieder_ADV richtig_ADJD gut_ADJD :_: -D_ADJA

Possible solutions:

- adjust the tools (domain adaptation);

Possible solutions:

- adjust the tools (domain adaptation);
- adjust the text (text normalization).

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
- What should be normalized?
- How should we normalize?
- How can we measure the quality of text normalization?

10,000 randomly selected tweets from a corpus of 24,179,871 Twitter messages that were gathered in April 2013. After sentence splitting and tokenization we got a list of **129,146** tokens (**32,538** token types). These tokens were successively processed with open source tools `hunspell` and `TreeTagger`.

Rate of OOV Tokens

From the previously obtained token list, 26,018 tokens were considered as OOV by `hunspell`, and 28,389 were regarded as OOV by `TreeTagger`.

Table: OOV rate in analyzed tweets

	hunspell		TreeTagger	
	% of OOV tokens	% of OOV types	% of OOV tokens	% of OOV types
OOV rate	20.15	46.96	21.98	58.24

Classes of OOV-Tokens

In which classes can OOV-tokens be divided?

Classes of OOV-Tokens

In which classes can OOV-tokens be divided?

- Limitedness of machine-readable lexicons;

Classes of OOV-Tokens

In which classes can OOV-tokens be divided?

- Limitedness of machine-readable lexicons;
- Stylistic specifics of text genre;

Classes of OOV-Tokens

In which classes can OOV-tokens be divided?

- Limitedness of machine-readable lexicons;
- Stylistic specifics of text genre;
- Sloppiness of user input.

In order to measure how OOV-tokens were distributed among these classes, we selected and analyzed all OOV-tokens with frequency higher than 1 and 1,000 randomly selected Hapax Legomena.

TABLE: Distribution of OOV-tokens among three major classes

OOV class	hunspell		TreeTagger	
	% of OOV tokens	% of OOV types	% of OOV tokens	% of OOV types
Limitedness of lexicons	45.87	54.62	40.46	43.36
Stylistic specifics of text genre	41.65	33.64	48.02	44.59
Deviating spelling	11.87	10.75	9.09	8.23

TABLE: Distribution of OOV-tokens in the class "Limitedness of machine-readable lexicons"

OOV subclass	hunspell		TreeTagger	
	% of OOV tokens	% of OOV types	% of OOV tokens	% of OOV types
Common German words	7.27	8.66	2.74	3.46
Compounds	1.27	2.65	2.5	4.54
Abbreviations	3.96	4.8	3.26	3.43
Interjections	5.99	4.6	5.56	4.27
Person names	4.77	6.49	2.31	3.46
Geographical names	1.53	2.6	1.16	1.87
Company names	2.28	2.87	4.34	3
Product names	2.16	2.65	2.45	3.22
Neologisms	1.37	1.35	3.32	2.38
Loan words	3.7	4.06	3.28	2.86
Foreign words	11.57	13.89	9.54	10.87
Total	45.87	54.62	40.46	43.36

TABLE: Distribution of OOV-tokens in the class "Stylistic specifics of text genre"

OOV-Unterklasse	hunspell		TreeTagger	
	% of OOV tokens	% of OOV types	% of OOV tokens	% of OOV types
@-mentions	13.12	20.49	16.14	21.84
hashtags	7.41	6.26	13.02	10.56
hyperlinks	2.45	0.4	4.88	6.05
emoticons	2.01	0.74	6.86	1.2
slang words	16.66	5.75	7.12	4.94
Total	41.65	33.64	48.02	44.59

As slang words we counted:

As slang words we counted:

- colloquial and dialectal expressions, e.g. *nö*, *bissl*

As slang words we counted:

- colloquial and dialectal expressions, e.g. *nö*, *bissl*
- Expressions pertaining to the genre of Internet-based communication, e.g. *LOL*, *ava*

As slang words we counted:

- colloquial and dialectal expressions, e.g. *nö*, *bissl*
- Expressions pertaining to the genre of Internet-based communication, e.g. *LOL*, *ava*
- Spelling deviations that reflected colloquial pronunciation of words, e.g. *Tach*, *nen*

TABLE: Intended (colloquial) vs. unintended (erroneous) “spelling deviations”

OOV-subclass	hunspell % of OOV tokens	hunspell % of OOV types	TreeTagger % of OOV tokens	TreeTagger % of OOV types
Intended deviations	8.06	5.09	5.97	3.7
Unintended deviations	3.81	5.66	3.12	4.54
Total	11.87	10.75	9.09	8.23

TABLE: Distribution of OOV-tokens in the class "Spelling deviations"

OOV-subclass	hunspell		TreeTagger	
	% of OOV tokens	% of OOV types	% of OOV tokens	% of OOV types
Insertions	1	1.66	0.79	1.08
Deletions	8.3	6.28	6.55	5.33
Substitutions	2.57	2.81	1.75	1.82
Total	11.87	10.75	9.09	8.23

TABLE: Distribution of OOV-tokens among three major classes

OOV class	hunspell		TreeTagger	
	% of OOV tokens	% of OOV types	% of OOV tokens	% of OOV types
Limitedness of lexicons	45.87	54.62	40.46	43.36
Stylistic specifics of text genre	41.65	33.64	48.02	44.59
Deviating spelling	11.87	10.75	9.09	8.23

TABLE: Distribution of OOV-tokens among three major classes

OOV class	hunspell		TreeTagger	
	% of OOV tokens	% of OOV types	% of OOV tokens	% of OOV types
Limitedness of lexicons	45.87	54.62	40.46	43.36
Stylistic specifics of text genre	41.65	33.64	48.02	44.59
Deviating spelling	11.87	10.75	9.09	8.23

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{5}$ *of all tokens,* $\approx \frac{1}{2}$ *of all types are unknown*
- What should be normalized?
- How should we normalize?
- How can we measure the quality of text normalization?

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{5}$ *of all tokens,* $\approx \frac{1}{2}$ *of all types are unknown*
- What should be normalized?
  - *Stylistic specifics of text genre;*
- How should we normalize?
- How can we measure the quality of text normalization?

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{5}$ *of all tokens,* $\approx \frac{1}{2}$ *of all types are unknown*
- What should be normalized?
  - *Stylistic specifics of text genre;*
  - *Deviating spellings;*
- How should we normalize?
- How can we measure the quality of text normalization?

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{10}$ *of all tokens,* $\approx \frac{1}{4}$ *of all token types need normalization*
- What should be normalized?
  - *Stylistic specifics of text genre;*
  - *Deviating spellings;*
- How should we normalize?
- How can we measure the quality of text normalization?

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{10}$ *of all tokens,* $\approx \frac{1}{4}$ *of all token types need normalization*
- What should be normalized?
  - *Stylistic specifics of text genre;*
  - *Deviating spellings;*
- How should we normalize?
  - *Stylistic specifics of text genre?*
  - *Intended spelling deviations?*
  - *Unintended spelling deviations?*
- How can we measure the quality of text normalization?

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{10}$ *of all tokens,* $\approx \frac{1}{4}$ *of all token types need normalization*
- What should be normalized?
  - *Stylistic specifics of text genre;*
  - *Deviating spellings;*
- How should we normalize?
  - *Stylistic specifics of text genre? - rule-based*
  - *Intended spelling deviations?*
  - *Unintended spelling deviations?*
- How can we measure the quality of text normalization?

Normalization of stylistic specifics of Twitter genre

Example

@Merkel soll für die nächsten 4 Jahre Kanzlerin bleiben.
%User soll für die nächsten 4 Jahre Kanzlerin bleiben.

Example

@Merkel Steinbrück wirds sicherlich nicht gelingen in die zweite Runde zu kommen.
Steinbrück wirds sicherlich nicht gelingen in die zweite Runde zu kommen.

Example

Wenn ich mir die Wahlnacht so Revue passieren lasse, dann gefiel mir der Kommentar des stellv. Chefredakteurs im **fb.me/34N8K2KTw**

Wenn ich mir die Wahlnacht so Revue passieren lasse, dann gefiel mir der Kommentar des stellv. Chefredakteurs im **%Link**

Example

#Schubs des Tages: Warum habe ich es verdient, glcklich zu sein? Deine Antwort? **url9.de/JLc**

Schubs des Tages: Warum habe ich es verdient, glcklich zu sein? Deine Antwort?

Example

Heute vor 7 Jahren: #Berlin Wuhlheide, blauer Himmel, 25 Grad... erstes #PearlJam Konzert inkl. Present Tense **:-D** legendär! (mit **@Vochlchen**)

Heute vor 7 Jahren: Berlin Wuhlheide, blauer Himmel, 25 Grad... erstes PearlJam Konzert inkl. Present Tense **%PosSmiley** legendär! (mit **%User**)

Normalization of intended spelling deviations

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{10}$ *of all tokens*, $\approx \frac{1}{4}$ *of all types need normalization*
- What should be normalized?
  - *Stylistic specifics of text genre;*
  - *Spelling deviations;*
- How should we normalize?
  - *Stylistic specifics of text genre? - rule-based*
  - *Intended spelling deviations?*
  - *Unintended spelling deviations?*
- How can we measure the quality of text normalization?

Normalization of intended spelling deviations

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{10}$ *of all tokens,* $\approx \frac{1}{4}$ *of all types need normalization*
- What should be normalized?
  - *Stylistic specifics of text genre;*
  - *Spelling deviations;*
- How should we normalize?
  - *Stylistic specifics of text genre? - rule-based*
  - *Intended spelling deviations? - rule-based*
  - *Unintended spelling deviations?*
- How can we measure the quality of text normalization?

- Omissions of ‘e’ in unstressed syllables, e.g. *würd*, *zuguckn* etc.;
- Omissions or replacement of unstressed consonants in final word positions, e.g. *nich* instead of *nicht* or *Tach* instead of *Tag*;
- Multiple repetitions of characters as way of expressing prolongated vowels, e.g. *Hilfeeee*, *süüüüß*;
- Omissions of ‘ei’ in indefinite articles, e.g. *ne* instead of *eine* or *nem* instead of *einem*;
- Omissions of ‘he’ in verb prefixes *herauf-*, *heraus-*, *herum-* etc., e.g. *rauszukriegen*, *rumbasteln*.

$D \leftarrow dictionary$
if w_i !~ /e$/ AND $w_i \notin D$ AND $w_i +$ 'e' $\in D$ **then**
    $w_i \leftarrow w_i +$ 'e'
end if

Example

So. Das Wahlergebnis gestern **hab** ich nur geträumt, oder?
So. Das Wahlergebnis gestern **habe** ich nur geträumt, oder?

Example

Wulff tritt zurück, Georg **Schramm** wird neuer Bundespräsident
Wulff tritt zurück, Georg **Schramme** wird neuer Bundespräsident

$D \leftarrow dictionary$
if w_i !~ /e$/ AND $w_i \notin D$ AND $w_i +$ 'e' $\in D$ AND
$log(P(w_{i-1}, w_i)) + log(P(w_i)) + log(P(w_i, w_{i+1})) <$
$log(P(w_{i-1}, w_i^*)) + log(P(w_i^*)) + log(P(w_i^*, w_{i+1}))$ **then**
$w_i \leftarrow w_i +$ 'e'
end if

Example

So. Das Wahlergebnis gestern **hab** ich nur geträumt, oder?
So. Das Wahlergebnis gestern **habe** ich nur geträumt, oder?

Example

Wulff tritt zurück, Georg **Schramm** wird neuer Bundespräsident
Wulff tritt zurück, Georg **Schramm** wird neuer Bundespräsident

Evaluation

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{10}$ *of all tokens,* $\approx \frac{1}{4}$ *of all types need normalization*
- What should be normalized?
  - *Stylistic specifics of text genre;*
  - *Spelling deviations;*
- How should we normalize?
  - *Stylistic specifics of text genre? - rule-based*
  - *Intended spelling deviations? - rule-based*
  - *Unintended spelling deviations?- statistically/ML*
- How can we measure the quality of text normalization?

Evaluation

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{10}$ *of all tokens,* $\approx \frac{1}{4}$ *of all types need normalization*
- What should be normalized?
  - *Stylistic specifics of text genre;*
  - *Spelling deviations;*
- How should we normalize?
  - *Stylistic specifics of text genre? - rule-based*
  - *Intended spelling deviations? - rule-based*
  - *Unintended spelling deviations?- statistically/ML*
- How can we measure the quality of text normalization?
  - *intrinsically (OOV-rate, precision, recall, F-measure)*

Evaluation

Questions regarding text normalization:

- How relevant is text normalization for German tweets?
  $\approx \frac{1}{10}$ *of all tokens,* $\approx \frac{1}{4}$ *of all types need normalization*
- What should be normalized?
  - *Stylistic specifics of text genre;*
  - *Spelling deviations;*
- How should we normalize?
  - *Stylistic specifics of text genre? - rule-based*
  - *Intended spelling deviations? - rule-based*
  - *Unintended spelling deviations?- statistically/ML*
- How can we measure the quality of text normalization?
  - *intrinsically (OOV-rate, precision, recall, F-measure)*
  - *extrinsically (Performance and quality of succeeding analysis modules)*

Intrinsic evaluation (OOV Rate)

The OOV-rate for tokens decreased by 5.6 % to 14.55 % for `hunspell` and by 8.9 % to 13.08 % for `TreeTagger`.

Intrinsic evaluation (Restoration of spelling deviations)

Measured on 1,492 tweets with 1,480 spellingdeviations

Table: Evaluation results

Input text	BLEU	NIST	Precision	Recall	F-Score
Without normalization	0.7929	12.55	–	–	–
With normalization	0.8766	13.2638	0.8793	0.4584	0.6027

Extrinsic evaluation (Tagging)

After normalization, PoS-Tagging accuracy improved by 6.41 % from 80.56 % to 86.97 %.[1]

[1] Measured on 200 randomly selected tweets.

Sentiment Corpus

For developing and testing our sentiment analysis system, we have created a corpus of 3996 Twitter messages. This corpus consists of four major topic parts (two political and two non-political ones) each of which was sampled using three disjoint selection criteria.

The covered topics are:

1. Political:
   - Tweets containing political terms (March 27 – May 25, 2013);
   - Tweets pertaining to the federal election 2013 (June 15 – September 30, 2013);
2. Non-political:
   - General tweets with no particular topic (March 31 – April 30, 2013);
   - Tweets pertaining to the pope election 2013 (March 13 – March 14, 2013).

The selection criteria for each of the topics are:

1. Presence of polar terms (SentiWS [4]);
2. Presence of smileys and exclamation marks;
3. Others.

For each of the above criteria, we sampled 333 messages for each topic. All messages were sampled disjointly so that tweets which fell into one of the preceding categories were excluded from the next ones.

Emo-expressions (*expressive subjective elements* [7]) - lexical items with polar evaluative sense, e.g. *gut, schrecklich, kritisieren, zum Besten halten etc.*;

Diminishers (*down-toners* [5]) - words or phrases which decrease the intensity of an emo-expression term, e.g. *weniger, bisschen, kaum etc.*

Intensifiers - lexical elements which strengthen the polar evaluative sense of an emo-expression, e.g. *recht, super, außerordentlich etc.*

Negations - language elements which reverse the polarity of subjective meaning expressed by an ESE, e.g. *nicht, kein, etc.*

Sentiment - minimal complete coherent syntactic or discourse-level unit that expresses a polar evaluative opinion of a person or organization about some particular subject, topic, or event, e.g. *Ich hasse diese Reform*, *ein ausgezeichneter Film*, *Meine Mutter ruft mich heulend an. Man hat einen Argentinier zum Papst gewählt.*;

Source - the immediate originator of a polar evaluative opinion who either directly expresses her opinion or whose opinion is being cited;

Target - subject or event which is being evaluated in a sentiment.

Example

[[Ich]$_{\textbf{source}}$[hasse]$_{\textbf{emo-expression}}$[Merkel]$_{\textbf{target}}$]$_{\textbf{sentiment}}$.

Table: Distribution of emotional expressions across topics and selection criteria in corpus.

Selection Criterion	Politics		Non-politics	
	General Politics	Federal Election	General Discussions	Pope Election
Polar Terms	225	199	270	163
Emoticons	426	415	457	364
Other	76	75	82	54

TABLE: Distribution of sentiments across topics and selection criteria in corpus.

Selection Criterion	Politics		Non-politics	
	General Politics	Federal Election	General Discussions	Pope Election
Polar Terms	90	105	79	83
Emoticons	68	71	35	50
Other	54	46	17	30

TABLE: Inter-annotator agreement for the sentiment corpus across topics. POL = political topics; FE = federal election 2013; PE = Pope election 2013; GEN = general tweets; TOT = total

Markable Type	Annotator 1					Annotator 2				
	POL	FE	PE	GEN	**TOT**	POL	FE	PE	GEN	**TOT**
Sentiment	0.35	0.35	0.45	0.41	**0.39**	0.27	0.29	0.36	0.34	**0.32**
Source	0.39	0.27	0.41	0.41	**0.37**	0.38	0.28	0.4	0.4	**0.36**
Target	0.32	0.38	0.4	0.39	**0.38**	0.26	0.28	0.31	0.32	**0.3**
Emo-Expression	0.64	0.57	0.68	0.66	**0.64**	0.6	0.54	0.65	0.63	**0.61**
Intensifier	0.46	0.48	0.21	0.62	**0.52**	0.46	0.48	0.21	0.6	**0.51**
Diminisher	0.67	0.44	0.0	0.4	**0.37**	0.67	0.44	0.0	0.4	**0.37**
Negation	0.44	0.1	0.36	0.21	**0.28**	0.44	0.1	0.36	0.21	**0.28**

Sentiment Analysis

Table: Classification results for automatic sentiment analysis (token-based).

ML-System	Sentiment	Source	Target	Other
MLN	na	na	na	na
SVM	3.4	10.7	0	94.5
Bayes Net	15.7	9.4	5.8	89
NB	15.9	7.5	8.9	78.4
Multinomial NB	17.5	9.8	11	85.6
CRF	16.53	17.65	7.89	94.47

Features:

- Formal:
  - Initial three characters of word form;
  - Final three characters of word form;
  - Character class of word (title, upper, lower, alphabetic mixed, alnum, digit, punct, mixed);
- Morphological:
  - Case;
  - Gender;
  - Degree of Comparison;
  - Mood;
  - Tense;
  - Person;
- Lexical:
  - Word Form;
  - Polarity Score (SentiWS* [4] and GermanPolarityClues [6]);
  - Class of modal verb (lexical or true modal);
- Syntactical:
  - Dependency relation of preceding and current word;
  - Dependency relation of current word;
  - Dependency relation of current and next word;
  - Lemma of parent;
  - PoS-Tag of grandmother;
  - Form of grandmother;
  - Polarity class of grandmother;
  - Child Lemma + Dependency Relation;
  - Child Lemma + Dependency Relation + Lemma;
  - Child PoS-Tag + Dependency Relation + PoS-Tag;
  - Cummulative polarity class for children (polarity class of the sum of children's scores);

Evaluation schemes:

- Binary Overlap [1]:
  Precision $= \frac{|\{p|p \in P \wedge \exists c \in C \text{ s.t. } f(c,p)\}|}{|P|}$; Recall $= \frac{|\{c|c \in C \wedge \exists p \in P \text{ s.t. } f(c,p)\}|}{|C|}$;
  where C is the set of correct spans, P is the set of predicted spans, and $f(c, p)$ is a function which yields "true" if the spans overlap and "false" otherwise;

Evaluation schemes:

- Binary Overlap [1]:
  Precision $= \frac{|\{p | p \in P \wedge \exists c \in C \text{ s.t. } f(c,p)\}|}{|P|}$; Recall $= \frac{|\{c | c \in C \wedge \exists p \in P \text{ s.t. } f(c,p)\}|}{|C|}$;
  where C is the set of correct spans, P is the set of predicted spans, and $f(c,p)$ is a function which yields "true" if the spans overlap and "false" otherwise;
- Proportional Overlap [3]:
  Precision $= \frac{\text{Score}(C,P)}{|P|}$; Recall $= \frac{\text{Score}(P,C)}{|C|}$;
  where $\text{Score}(S,S') = \sum_{s \in S} \sum_{s' \in S'} f(s,s')$ and $f(s,s') = \frac{|s \cap s'|}{|s'|}$;

Evaluation schemes:

- Binary Overlap [1]:
  Precision $= \frac{|\{p|p \in P \wedge \exists c \in C \text{ s.t. } f(c,p)\}|}{|P|}$; Recall $= \frac{|\{c|c \in C \wedge \exists p \in P \text{ s.t. } f(c,p)\}|}{|C|}$;
  where C is the set of correct spans, P is the set of predicted spans, and $f(c,p)$ is a function which yields "true" if the spans overlap and "false" otherwise;
- Proportional Overlap [3]:
  Precision $= \frac{\text{Score}(C,P)}{|P|}$; Recall $= \frac{\text{Score}(P,C)}{|C|}$;
  where $\text{Score}(S,S') = \sum_{s \in S} \sum_{s' \in S'} f(s,s')$ and $f(s,s') = \frac{|s \cap s'|}{|s'|}$;
- Exact Match [1]:
  the same as binary overlap except that $f(c,p)$ yields "true" iff the compared spans completely agree on their boundaries.

TABLE: Classification results for automatic sentiment analysis (binary overlap).

Classification Element	Precision	Recall	F-Measure
Training Set			
Sentiment	99.23	86.27	92.29
Source	91.56	75.55	82.78
Target	95.99	75.69	84.64
Test Set			
Sentiment	25	16.04	19.55
Source	47.06	25	32.65
Target	31.51	18.11	23

TABLE: Classification results for automatic sentiment analysis (binary overlap).
Sentiment is emo-expression

Classification Element	Precision	Recall	F-Measure
Training Set			
Sentiment	94.38	81.43	87.43
Source	92.31	48.54	63.62
Target	96.95	56.83	71.66
Test Set			
Sentiment	76.54	68.5	72.29
Source	25	18.75	21.43
Target	15.46	11.81	13.39

TABLE: Classification results for automatic sentiment analysis (proportional overlap).

Classification Element	Precision	Recall	F-Measure
Training Set			
Sentiment	97.62	84.94	90.84
Source	90.4	73.71	81.21
Target	93.55	74.02	82.65
Test Set			
Sentiment	21.31	14.53	17.28
Source	40	25	30.77
Target	26.06	13.75	18

TABLE: Classification results for automatic sentiment analysis (proportional overlap). Sentiment is emo-expression

Classification Element	Precision	Recall	F-Measure
Training Set			
Sentiment	93.62	80.5	86.57
Source	92.07	48.26	63.33
Target	94.39	55.58	69.96
Test Set			
Sentiment	74.38	67.27	70.65
Source	22.22	18.75	20.34
Target	12.16	10.56	11.3

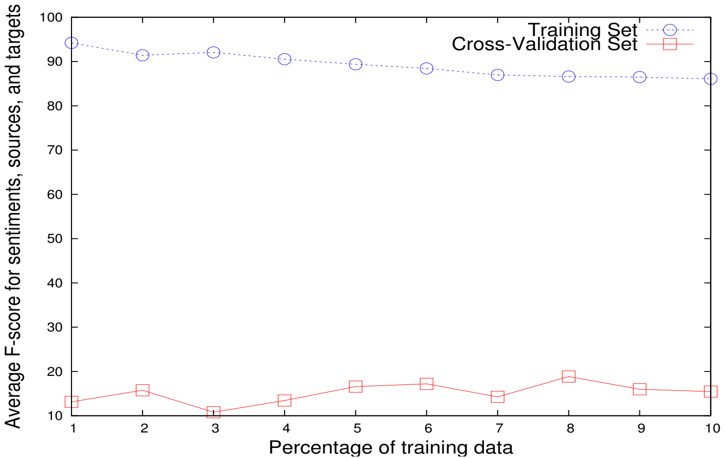
TABLE: Classification results for automatic sentiment analysis (exact match).

Classification Element	Precision	Recall	F-Measure
Training Set			
Sentiment	87.37	72.7	79.36
Source	88.24	71.17	78.79
Target	85.54	66.44	74.79
Test Set			
Sentiment	13.95	9.09	11.01
Source	40	25	30.77
Target	14.67	8.66	10.89

TABLE: Classification results for automatic sentiment analysis (exact match).
Sentiment is emo-expression

Classification Element	Precision	Recall	F-Measure
Training Set			
Sentiment	90.9	78.39	84.18
Source	89.51	46.72	61.39
Target	80.08	45.6	58.11
Test Set			
Sentiment	70.84	63.21	66.81
Source	20.83	15.62	17.86
Target	8.25	6.3	7.14

Learning Curve



Problems and Open Questions

- Bad overfitting;
- Inconsistent tagging sequences;
- Flat tagging scheme;
- Relation linking.

TABLE: Classification results for automatic sentiment analysis (binary overlap; linear chain CRFs).

Classification Element	Precision	Recall	F-Measure
Training Set			
Sentiment	99.23	86.27	92.29
Source	91.56	75.55	82.78
Target	95.99	75.69	84.64
Test Set			
Sentiment	25	16.04	19.55
Source	47.06	25	32.65
Target	31.51	18.11	23

Preliminary Conclusions and Perspectives

Conclusions:

- Preprocessing matters (w/25.067 vs. wo/18.277);
- Quality of polarity dictionaries is important (sentiws/25.067 vs. gpc/23.903);

Perspectives:

- Different classifiers (higher order CRFs, structural SVMs, etc.);
- Experiments with polarity dictionaries and ontologies;

Eric Breck, Yejin Choi, and Claire Cardie.
Identifying expressions of opinion in context.
In Manuela M. Veloso, editor, *IJCAI*, pages 2683–2688, 2007.

Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors.
Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta. European Language Resources Association, 2010.

Richard Johansson and Alessandro Moschitti.
Reranking models in fine-grained opinion analysis.
In Chu-Ren Huang and Dan Jurafsky, editors, *COLING*, pages 519–527. Tsinghua University Press, 2010.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer.
Sentiws - a publicly available german-language resource for sentiment analysis.
In Calzolari et al. [2].

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede.
Lexicon-based methods for sentiment analysis.
Computational Linguistics, 37(2):267–307, 2011.

Ulli Waltinger.
Germanpolarityclues: A lexical resource for german sentiment analysis.
In Calzolari et al. [2].

Janyce Wiebe, Theresa Wilson, and Claire Cardie.
Annotating expressions of opinions and emotions in language.

Language Resources and Evaluation, 39(2-3):165–210, 2005.