An Impact Analysis of Features in a Classification Approach to Irony Detection in Product Reviews

Konstantin Buschmeier, Philipp Cimiano and Roman Klinger
Semantic Computing Group
Cognitive Interaction Technology – Center of Excellence (CIT-EC)
Bielefeld University
33615 Bielefeld, Germany
kbuschme@techfak.uni-bielefeld.de
{rklinger,cimiano}@cit-ec.uni-bielefeld.de

Abstract

Irony is an important device in human communication, both in everyday spoken conversations as well as in written texts including books, websites, chats, reviews, and Twitter messages among others. Specific cases of irony and sarcasm have been studied in different contexts but, to the best of our knowledge, only recently the first publicly available corpus including annotations about whether a text is ironic or not has been published by Filatova (2012). However, no baseline for classification of ironic or sarcastic reviews has been provided.
With this paper, we aim at closing this gap. We formulate the problem as a supervised classification task and evaluate different classifiers, reaching an F_1-measure of up to 74 % using logistic regression. We analyze the impact of a number of features which have been proposed in previous research as well as combinations of them.

1 Introduction

Irony is often understood as "the use of words that mean the opposite of what you really think especially in order to be funny" or "a situation that is strange or funny because things happen in a way that seems to be the opposite" of what is expected.[1] Many dictionaries make this difference between verbal irony and situational irony (British Dictionary, 2014; New Oxford American Dictionary, 2014; Merriam Webster Dictionary, 2014). The German Duden (2014) mentions sarcasm as synonym to irony, while the comprehension of sarcasm as a special case of irony might be more common. For instance, the Merriam Webster Dictionary (2014) defines sarcasm as "a sharp and often satirical or ironic utterance designed to cut or give pain".[2]

Irony is a frequent phenomenon within human communication, occurring both in spoken and written discourse including books, websites, fora, chats, Twitter messages, Facebook posts, news articles and product reviews. Even for humans it is sometimes difficult to recognize irony. Irony markers are thus often used in human communication, supporting the correct interpretation (Attardo, 2000). The automatic identification of ironic formulations in written text is a very challenging as well as important task as shown by the comment[3]

> "Read the book!"

which in the context of a movie review could be regarded as ironic and as conveying the fact that the film was far worse compared to the book. Another example is taken from a review for the book "Great Expectations" by Charles Dickens:[4]

> "i would recomend this book to friends who have insomnia or those who i absolutely despise."

The standard approach of recommending X implies that X is worthwhile is clearly not valid in the given context as the author is stating that she disliked the book.

[1] as defined in the Merriam Webster Dictionary (2014), http://www.merriam-webster.com/dictionary/irony

[2] http://www.merriam-webster.com/dictionary/sarcasm, accessed April 28, 2014

[3] Example from Lee (2009).

[4] http://www.amazon.com/review/R86RAMEBZSB11, access date March 10, 2014

Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 42–49,
Baltimore, Maryland, USA. June 27, 2014.

In real world applications of sentiment analysis, large data sets are automatically classified into positive statements or negative statements and such output is used to generate summaries of the sentiment about a product. In order to increase the accurateness of such systems, ironic or sarcastic statements need to be identified in order to infer the actual communicative intention of the author.

In this paper, we are concerned with approaches for the automatic detection of irony in texts, which is an important task in a variety of applications, including the automatic interpretation of text-based chats, computer interaction or sentiment analysis and opinion mining. In the latter case, the detection is of outmost importance in order to correctly assign a polarity score to an aspect of a reviewed product or a person mentioned in a Twitter message. In addition, the automatic detection of irony or sarcasm in text requires an operational definition and has therefore the potential to contribute to a deeper understanding of the linguistic properties of irony and sarcasm as linguistic phenomena and their corpus based evaluation and verification.

The rest of this paper is structured as follows: We introduce the background and theories on irony in Section 1.1 and discuss previous work in the area of automatically recognizing irony in Section 1.2. In the methods part in Section 2, we present our set of features (Section 2.1) and the classifiers we take into account (Section 2.2). In Section 3, we discuss the data set used in this work in more detail (Section 3.1), present our experimental setting (Section 3.2) and show the evaluation of our approach (Section 3.3). We conclude with a discussion and summary (Section 4) and with an outlook on possible future work (Section 5).

1.1 Background

Irony is an important and frequent device in human communication that is used to convey an attitude or evaluation towards the propositional content of a message, typically in a humorous fashion (Abrams, 1957, p. 165–168). Between the age of six (Nakassis and Snedeker, 2002) and eight years (Creusere, 2007), children are able to recognize ironic utterances or at least notice that something in the situation is not common (Glenwright and Pexman, 2007). The principle of inferability (Kreuz, 1996) states that figurative language is used if the speaker is confident that the addressee will interpret the utterance and infer the communicative intention of the speaker/author correctly. It has been shown that irony is ubiquitous, with 8 % of the utterances exchanged between interlocutors that are familiar with each other being ironic (Gibbs, 2007).

Utsumi (1996) claim that an ironic utterance can only occur in an *ironic environment*, whose presence the utterance implicitly communicates. Given the formal definition it is possible to computationally resolve if an utterance is ironic using first-order predicate logic and situation calculus. Different theories such as the *echoic account* (Wilson and Sperber, 1992), the *Pretense Theory* (Clark and Gerrig, 1984) or the *Allusional Pretense Theory* (Kumon-Nakamura et al., 1995) have challenged the understanding that an ironic utterance typically conveys the opposite of its literal propositional content. However, in spite of the fact that the attributive nature of irony is widely accepted (see Wilson and Sperber (2012)), no formal or operational definition of irony is available as of today.

1.2 Previous Work

Corpora providing annotations as to whether expressions are ironic or not are scarce. Kreuz and Caucci (2007) have automatically generated such a corpus exploiting Google Book search[5]. They collected excerpts containing the phrase "said sarcastically", removed that phrase and performed a regression analysis on the remaining text, exploiting the number of words as well as the occurrence of adjectives, adverbs, interjections, exclamation and question marks as features.

Tsur et al. (2010) present a system to identify sarcasm in Amazon product reviews exploiting features such as sentence length, punctuation marks, the total number of completely capitalized words and automatically generated patterns which are based on the occurrence frequency of different terms (following the approach by Davidov and Rappoport (2006)). Unfortunately, their corpus is not publicly available. Carvalho et al. (2009) use eight patterns to identify ironic utterances in comments on articles from a Portuguese online newspaper. These patterns contain positive predicates and utilize punctuation, interjections, positive words, emoticons, or onomatopoeia and acronyms for laughing as well as some Portuguese-specific patterns considering the verb-morphology. González-Ibáñez et al. (2011) differentiate between sarcastic and positive or negative Twitter messages. They

[5] http://books.google.de/

exploit lexical features like unigrams, punctuation, interjections and dictionary-based as well as pragmatic features including references to other users in addition to emoticons. Reyes et al. (2012) distinguish ironic and non-ironic Twitter messages based on features at different levels of linguistic analysis including quantifiers of sentence complexity, structural, morphosyntactic and semantic ambiguity, polarity, unexpectedness, and emotional activation, imagery, and pleasantness of words. Tepperman et al. (2006) performed experiments to recognize sarcasm in spoken language, specifically in the expression "yeah right", using spectral, contextual and prosodic cues. On the one hand, their results show that it is possible to identify sarcasm based on spectral and contextual features and, on the other hand, they confirm that prosody is insufficient to reliably detect sarcasm (Rockwell, 2005, p. 118).

Very recently, Filatova (2012) published a product review corpus from Amazon, being annotated with Amazon Mechanical Turk. It contains 437 ironic and 817 non-ironic reviews. A more detailed description of this resource can be found in Section 3.1. To our knowledge, no automatic classification approach has been evaluated on this corpus. We therefore contribute a text classification system including the previously mentioned features. Our results serve as a strong baseline on this corpus as well as an "executable review" of previous work.[6]

2 Methods

We model the task of irony detection as a supervised classification problem in which a review is categorized as being ironic or non-ironic. We investigate different classifiers and focus on the impact analysis of different features by investigating what effect their elimination has on the performance of the approach. In the following, we describe the features used and the set of classifiers compared.

2.1 Features

To estimate if a review is ironic or not, we measure a set of features. Following the idea that irony is expressing the opposite of its literal content, we take into account the imbalance between the overall (prior) polarity of words in the review and the star-rating (as proposed by Davidov et al. (2010)). We assume the imbalance to hold if the star-rating is positive (*i. e.*, 4 or 5 stars) but the majority of words is negative, and, vice versa, if the star-rating is negative (*i. e.*, 1 or 2 stars) but occurs with a majority of positive words. We refer to this feature as *Imbalance*. The polarity of words is determined based on a dictionary consisting of about 6,800 words with their polarity (Hu and Liu, 2004).[7]

The feature *Hyperbole* (Gibbs, 2007) indicates the occurrence of a sequence of three positive or negative words in a row. Similarly, the feature *Quotes* indicates that up to two consecutive adjectives or nouns in quotation marks have a positive or negative polarity.

The feature *Pos/Neg&Punctuation* indicates that a span of up to four words contains at least one positive (negative) but no negative (positive) word and ends with at least two exclamation marks or a sequence of a question mark and an exclamation mark (Carvalho et al., 2009). Analogously, the feature *Pos/Neg&Ellipsis* indicates that such a positive or negative span ends with an ellipsis ("…"). *Ellipsis and Punctuation* indicates that an ellipsis is followed by multiple exclamation marks or a combination of an exclamation and a question mark. The *Punctuation* feature conveys the presence of an ellipses as well as multiple question or exclamation marks or a combination of the latter two. The *Interjection* feature indicates the occurrence of terms like "wow" and "huh", and *Laughter* measures onomatopoeia ("haha") as well as acronyms for grin or laughter ("*g*", "lol"). In addition, the feature *Emoticon* indicates the occurrence of an emoticon. In order to capture a range of emotions, it combines a variety of emoticons such as happy, laughing, winking, surprised, dissatisfied, sad, crying, and sticking tongue out. In addition, we use each occurring word as a feature (*bag-of-words*).

All together, we have 21,773 features. The number of specific features (*i. e.*, without bag-of-words) alone is 29.

2.2 Classifiers

In order to perform the classification based on the features mentioned above, we explore a set of standard classifiers typically used in text classification research. We employ the open source machine learning library *scikit-learn* (Pedregosa et al., 2011) for Python.

[6]The system as implemented to perform the described experiments is made available at `https://github.com/kbuschme/irony-detection/`

[7]Note that examples can show that this is not always the case. Funny or odd products ironically receive a positive star-rating. However, this feature may be a strong indicator for irony.

We use a support vector machine (SVM, Cortes and Vapnik (1995)) with a linear kernel in the implementation provided by libSVM (Fan et al., 2005; Chang and Lin, 2011). The naïve Bayes classifier is employed with a multinomial prior (Zhang, 2004; Manning et al., 2008). This classifier might suffer from the issue of over-counting correlated features, such that we compare it to the logistic regression classifier as well (Yu et al., 2011).

Finally, we use a decision tree (Breiman et al., 1984; Hastie et al., 2009) and a random forest classifier (Breiman, 2001).

3 Experiments and Results

3.1 Data Set

The data set by Filatova (2012) consists of 1,254 Amazon reviews, of which 437 are ironic, *i. e.*, contain situational irony or verbal irony, and 817 are non-ironic. It has been acquired using the crowd sourcing platform Amazon Mechanical Turk[8]. Note that Filatova (2012) interprets sarcasm as being verbal irony.

In a first step, the workers were asked to find pairs of reviews on the same product so that one of the reviews is ironic while the other one is not. They were then asked to submit the ID of both reviews, and, in the case of an ironic review, to provide the fragment conveying the irony.

In a second step, each collected review was annotated by five additional workers and remained in the corpus if three of the five new annotators concurred with the initial category, *i. e.*, ironic or non-ironic. The corpus contains 21,744 distinct tokens[9], of which 5,336 occur exclusively in ironic reviews, 9,468 exclusively in non-ironic reviews, and the remaining 6,940 tokens occur in both ironic and non-ironic reviews. Thus, all ironic reviews comprise a total of 12,276 distinct tokens, whereas a total of 16,408 distinct tokens constitute all non-ironic reviews. On average, a single review consists of 271.9 tokens, a single ironic review of an average of 261.4 and a single non-ironic review of an average of 277.5 tokens. The distribution of ironic and non-ironic reviews for the different star-ratings is shown in Table 2. Note that this might be a result of the specific annotation procedure applied by the annotators to search for ironic reviews. Nevertheless, this motivates a simple baseline system which just takes one feature into account: the numbers of stars assigned to the respective review ("Star-rating only").

[8] `https://www.mturk.com/mturk/`, accessed on March 10, 2014

[9] Using the TreeBankWordTokenizer as implemented in the Natural Language Toolkit (NLTK) (`http://www.nltk.org/`)

3.2 Experimental Settings

We run experiments for three baselines: The *star-rating* baseline relies only on the number of stars assigned in the review as a feature. The *bag-of-words* baseline exploits only the unigrams in the text as features. The *sentiment word count* only uses the information whether the number of positive words in the text is larger than the number of negative words.

We emphasize that the first baseline is only of limited applicability as it requires the explicit availability of a star-rating. The second baseline relies on standard text classification features that are not specific for the task. The third baseline relies on a classical feature used in sentiment analysis, but is not specific for irony detection.

We refer to the feature set "All" encompassing all features described in Section 2.1, including bag-of-words and the set "Specific Features".

In order to understand the impact of a specific feature A, we run three sets of experiments:

- Using all features with the exception of A.
- Using all specific features with the exception of A.
- Using A as the only feature.

In addition to evaluating each single feature as described above, we evaluate the set of positive and negative instantiations of features when using the sentiment dictionary. The "Positive set" and "Negative set" take into account the respective subsets of all specific features.

Each experiment is performed in a 10-fold cross-validation setting on document level. We report recall, precision and F_1-measure for each of the classifiers.

3.3 Evaluation

Table 1 shows the results for the three baselines and different feature set combinations, all for the different classifiers. The star-rating as a feature alone is a very strong indicator for irony. However, this result is of limited usefulness as it only regards reviews of a specific rating as ironic, namely results with

	Linear SVM			Logistic Regression			Decision Tree			Random Forest			Naive Bayes		
Feature set	R.	P.	F_1	R.	P.	F_1	R.	P.	F_1	R.	P.	F_1	R.	P.	F_1
Star-rating only	66.7	78.4	71.7	66.7	78.4	71.7	66.7	78.4	71.7	66.7	78.4	71.7	66.7	78.4	71.7
BOW only	61.8	67.2	64.1	63.3	76.0	68.8	53.8	53.4	53.4	21.7	70.4	32.9	48.1	77.4	59.1
Sentiment Word Count	57.3	59.4	58.1	57.3	59.4	58.1	57.3	59.4	58.1	57.3	59.4	58.1	0.0	100.0	0.0
All + Star-rating	69.0	74.4	71.3	68.9	81.7	74.4	71.7	73.2	72.2	34.0	85.0	48.2	55.3	79.7	65.0
All (= Sp. Features + BOW)	61.3	68.0	64.3	62.2	75.2	67.8	55.0	59.8	56.9	24.1	73.2	35.3	50.9	77.3	61.2
All − Imbalance	62.4	67.1	64.4	62.5	75.0	67.9	53.0	54.3	53.3	22.3	75.9	33.8	47.8	75.8	58.4
All − Hyperbole	61.3	68.0	64.3	62.2	75.2	67.8	57.1	61.5	58.9	22.3	79.6	34.4	50.9	77.3	61.2
All − Quotes	61.3	68.0	64.3	62.8	75.1	68.2	57.2	61.7	59.1	25.9	76.8	38.5	50.6	77.0	60.9
All − Pos/Neg&Punctuation	61.5	67.9	64.4	62.4	75.2	68.0	56.7	60.1	58.0	21.8	77.8	33.5	50.9	77.3	61.2
All − Pos/Neg&Ellipsis	61.0	67.4	63.8	63.0	75.1	68.3	57.6	60.5	58.8	29.0	79.2	42.2	50.4	76.6	60.7
All − Ellipsis and Punctuation	61.3	68.0	64.3	62.4	75.2	68.0	55.1	59.7	56.9	24.6	73.6	36.2	50.9	77.3	61.2
All − Punctuation	61.8	67.9	64.5	62.5	74.9	67.8	56.1	61.2	58.3	28.6	78.1	41.5	50.2	76.7	60.6
All − Injections	61.3	68.0	64.3	62.2	75.0	67.8	56.1	61.8	58.5	24.1	75.2	35.6	50.9	77.3	61.2
All − Laughter	61.3	68.2	64.4	62.4	75.3	68.0	56.6	60.9	58.2	24.0	79.3	36.5	50.9	77.3	61.2
All − Emoticons	61.3	68.2	64.4	62.6	75.3	68.1	57.7	60.2	58.6	24.3	76.5	36.7	50.9	77.3	61.2
All − Negative set	61.0	68.0	64.1	62.3	74.7	67.7	59.0	61.1	59.7	25.4	76.8	37.6	50.2	76.6	60.5
All − Positive set	62.6	67.3	64.6	62.5	75.7	68.2	53.7	55.1	54.2	20.5	67.7	31.1	47.8	75.8	58.4
Sp. Features	37.5	77.2	50.2	38.2	77.5	50.8	38.3	76.0	50.6	38.3	74.8	50.2	34.3	80.5	47.7
Sp. Features − Imbalance	9.3	50.4	15.4	11.0	54.1	18.1	11.3	48.5	18.1	12.9	47.4	20.0	5.9	55.8	10.3
Sp. Features − Hyperbole	37.5	77.4	50.3	38.2	77.5	50.8	38.3	76.7	50.7	38.8	76.4	51.2	34.3	80.9	47.8
Sp. Features − Quotes	37.7	76.9	50.3	38.0	78.1	50.7	37.8	75.6	50.1	38.3	73.6	50.0	34.3	80.5	47.7
Sp. Features − Pos/Neg&Punctuation	37.7	77.9	50.5	37.8	77.6	50.5	37.1	74.5	49.2	38.2	73.8	49.9	33.3	80.2	46.7
Sp. Features − Pos/Neg&Ellipsis	37.7	77.3	50.4	38.1	78.2	50.9	37.9	76.2	50.4	39.1	72.3	50.3	34.5	79.7	47.8
Sp. Features − Ellipsis and Punctuation	37.8	76.9	50.3	37.8	76.9	50.3	38.3	75.8	50.6	39.0	72.5	50.5	34.5	80.2	47.9
Sp. Features − Punctuation	37.1	79.7	50.3	37.6	78.7	50.6	37.0	76.7	49.6	38.4	75.4	50.5	32.6	78.9	45.6
Sp. Features − Interjections	37.7	76.9	50.3	37.9	77.5	50.6	38.1	76.1	50.4	38.7	75.2	50.7	34.3	80.5	47.7
Sp. Features − Laughter	37.8	77.3	50.5	38.0	77.7	50.7	37.3	75.5	49.6	37.5	73.4	49.4	34.5	81.2	48.0
Sp. Features − Emoticons	37.3	78.2	50.2	38.2	77.5	50.8	38.0	75.4	50.2	38.7	75.0	50.7	33.4	80.7	46.8
Sp. Features − Positive set	10.5	48.7	17.1	11.0	56.3	18.1	9.9	49.3	16.3	12.3	50.8	19.5	6.3	64.8	11.0
Sp. Features − Negative set	37.7	78.2	50.6	38.0	78.7	50.9	38.2	75.1	50.3	37.6	72.0	48.9	34.9	79.8	48.3
Imbalance only	36.9	81.4	50.4	36.9	81.4	50.4	36.9	81.4	50.4	36.9	81.4	50.4	0.0	100.0	0.0
Hyperbole only	0.0	80.0	0.0	0.0	90.0	0.0	0.0	80.0	0.0	0.2	55.0	0.4	0.0	100.0	0.0
Quotes only	3.9	45.5	7.0	0.9	67.0	1.7	4.0	43.8	7.0	2.5	52.2	4.5	0.0	100.0	0.0
Pos/Neg&Punctuation only	0.9	90.0	1.8	0.5	90.0	0.9	0.0	90.0	0.0	0.4	90.0	0.8	0.9	90.0	1.8
Pos/Neg&Ellipsis only	6.8	59.0	12.1	6.8	59.0	12.1	6.8	59.0	12.1	6.8	59.0	12.1	0.0	100.0	0.0
Ellipsis and Punctuation only	0.9	90.0	1.7	0.4	90.0	0.8	0.9	90.0	1.7	0.9	90.0	1.7	0.0	100.0	0.0
Punctuation only	5.4	64.6	9.8	5.4	64.6	9.8	3.3	60.8	6.2	4.0	60.8	7.5	4.7	64.6	8.6
Interjections only	0.5	75.8	0.9	0.3	82.5	0.5	0.5	75.8	0.9	1.4	74.2	2.7	0.0	100.0	0.0
Laughter only	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	80.0	0.0	0.0	100.0	0.0
Emoticons only	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	80.0	0.0
Positive set only	36.9	81.4	50.4	36.9	81.1	50.4	37.1	80.5	50.5	37.3	79.3	50.5	32.4	80.7	45.6
Negative set only	8.2	54.5	14.1	7.3	48.8	12.5	8.8	49.4	14.8	9.0	49.9	15.2	0.0	80.0	0.0

Table 1: Comparison of different classification methods using different feature sets. “All” refers to the features described in Section 2 including bag-of-words (“BOW”). “Sp. Features” are “All” without “BOW”.

a positive rating by the author, as explained by Table 2, which shows the more real-world compatible result of a rich feature set in addition. Obviously, the depicted distribution is very similar to the distribution of the manually annotated data set, which can obviously not be achieved by the star-rating feature alone.

The best result is achieved by using the star-rating together with bag-of-words and specific features with a logistic regression approach (leading to an F_1-measure of 74 %). The SVM and decision tree have a comparable performance on the task, which is albeit lower compared to the performance of the logistic regression approach.

Using the task-agnostic pure bag-of-words approach leads to a performance of 68.8 % for logistic regression; this classifier has the property of dealing well with correlated features and the additional specific features cannot contribute positively to the result. Similarly, the F_1-measure of 64.1 % produced by the SVM cannot be increased by including additional features. In contrast, a positive impact of additional features can be observed for the decision tree in the case that specific features are combined with bag-of-word-based features, reaching close to 59 % F_1 in comparison to 53.4 % F_1 for bag-of-words alone.

It would be desirable to have a model only or mainly based on the problem-specific features, as this leads to a much more compact and therefore ef-

ficient representation than taking all words into account. In addition, the model would be easier to understand. By exploiting task-specific features alone, the performance reaches at most an F_1-measure of 50.9 %, which shows that task-agnostic features such as unigram features are needed. A significant drop in performance when leaving out a feature or feature set can be observed for the *Imbalance* feature and the *Positive set*. Both these feature sets take into account the star-rating.

The task-specific features alone yield high precision results at the expense of a very low recall. This clearly shows that task-specific features should be used with standard, task-independent features (the bag-of-words). The most helpful task-specific features are: Imbalance, Positive set, Quotes and Pos/Neg&Ellipses.

4 Discussion and Summary

The best performance is achieved with very corpus-specific features taking into account meta-data from Amazon, namely the product rating of the reviewer. This leads to an F_1-measure of 74 %. However, we could not show a competitive performance with more problem-specific features (leading to 51 % F_1) or in combination with bag-of-word-based features (leading to 68 % F_1).

The baseline only predicting based on the star-rating itself is highly competitive, however, not applicable to texts without meta-data and of limited use due to its naturally highly biased outcome towards positive reviews being non-ironic and negative reviews being ironic. Our results show that the best results are achieved via meta-data and it remains an open research task to develop comparably good approaches only based on text features.

It should be noted that the corpus used in this work is not a random sample from all reviews available in a specific group of products. We actually assume ironic reviews to be much more sparse when sampling equally distributed. The evaluation should be seen from the angle of the application scenario: For instance, in a discovery setting in which the task is to retrieve examples for ironic reviews, a highly precise system would be desirable. In a setting in which only a small number of reviews should be used for opinion mining, the polarity of a text would be discovered taking the classifier's result into account – therefore a system with high precision and high recall would be needed.

	Distribution			
	Corpus		Predicted	
Rating	ironic	non-ironic	ironic	non-ironic
5	114	605	126	593
4	14	96	17	93
3	20	35	14	41
2	27	17	17	27
1	262	64	192	134
1–5	437	817	366	888

Table 2: Frequencies for the different star-ratings of a review, as annotated, and according to the logistic regression classifier with the feature set "All – Imbalance".

5 Future Work

As discussed at the end of the last section, a study on the distribution of irony in the entirety of available reviews is needed to better shape the structure and characteristics of an irony or sarcasm detection system. This could be approached by performing a random sample from reviews and annotation, though this would lead to a substantial amount of annotation work in comparison to the directed selection procedure used in the corpus by Filatova (2012).

Future research should focus on the development of approaches analyzing the vocabulary used in the review in a deeper fashion. Our impression is that many sarcastic and ironic reviews use words and phrases which are non-typical for the specific domain or product class. Such out-of-domain vocabulary can be detected with text similarity approaches. Preliminary experiments taking into account the average cosine similarity of a review to be classified to a large set of reviews from the same product class have been of limited success. We propose that future research should focus on analyzing the specific vocabulary and develop semantic similarity measures which we assume to be more promising than approaches taking into account lexical approaches only.

Most work has been performed on text sets from one source like Twitter, books, reviews, etc. Some of the proposed features mentioned in this paper or previous publications are probably transferable between text sources. However, this still needs to be proven and further development might be necessary to actually provide automated domain adaption for the area of irony and sarcasm detection. We assume that not only the vocabulary changes

(as known in other domain adaptation tasks) but actually the linguistic structure might change.

Finally, it should be noted that the corpus is actually a mixture of ironic and sarcastic reviews. Irony and sarcasm are not fully exchangeable and can be assumed to have different properties. Further investigations and analyses regarding the characteristics that can be transferred are necessary.

Acknowledgements

Roman Klinger has been funded by the "It's OWL" project ("Intelligent Technical Systems Ostwestfalen-Lippe", http://www.its-owl.de/), a leading-edge cluster of the German Ministry of Education and Research. We thank the reviewers for their valuable comments. We thank Christina Unger for proof-reading the manuscript and helpful comments.

References

Meyer Howard Abrams. 1957. *A Glossary of Literary Terms*. Cengage Learning Emea, 9th edition.

Salvatore Attardo. 2000. Irony markers and functions: Towards a goal-oriented theory of irony and its processing. *Rask: Internationalt Tidsskrift for Sprog og Kommunikation*, 12:3–20.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, California.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

British Dictionary. 2014. MacMillan Publishers. Online: http://www.macmillandictionary.com/dictionary/british/irony. accessed April 28, 2014.

Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, TSA '09, pages 53–56, New York, NY, USA. ACM.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3).

Herbert H. Clark and Richard J. Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General*, 113(1):121–126.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Marlena A. Creusere. 2007. A developmental test of theoretical perspective on the understanding of verbal irony: Children's recognition of allusion and pragmatic insincerity. In Raymond W. Jr. Gibbs and Herbert L. Colston, editors, *Irony in Language and Thought: A Cognitive Science Reader*, chapter 18, pages 409–424. Lawrence Erlbaum Associates, 1st edition.

Dmitry Davidov and Ari Rappoport. 2006. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 297–304, Sydney, Australia, July. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Duden. 2014. Duden Verlag. Online: http://www.duden.de/rechtschreibung/Ironie. accessed April 28, 2014.

Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. 2005. Working set selection using second order information for training support vector machines. *Jounral of Machine Learning Reasearch*, 6:1889–1918.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 392–398, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Raymond W. Jr. Gibbs. 2007. Irony in talk among friends. In Raymond W. Jr. Gibbs and Herbert L. Colston, editors, *Irony in Language and Thought: A Cognitive Science Reader*, chapter 15, pages 339–360. Lawrence Erlbaum Associates, 1st edition, May.

Melanie Harris Glenwright and Penny M. Pexman. 2007. Children's perceptions of the social functions of verbal irony. In Raymond W. Jr. Gibbs and Herbert L. Colston, editors, *Irony in Language and Thought: A Cognitive Science Reader*, chapter 20, pages 447–464. Lawrence Erlbaum Associates, 1st edition.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual*

Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, HLT '11, pages 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. 2009. *Elements of Statistical Learning*. Springer, 2nd edition.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Roger J. Kreuz and Gina M. Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, New York, April. Association for Computational Linguistics.

Roger J. Kreuz. 1996. The use of verbal irony: Cues and constraints. In Jeffery S. Mio and Albert N. Katz, editors, *Metaphor: Implications and Applications*, pages 23–38, Mahwah, NJ, October. Lawrence Erlbaum Associates.

Sachi Kumon-Nakamura, Sam Glucksberg, and Mary Brown. 1995. How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General*, 124(1):3–21, Mar 01. Last updated - 2013-02-23.

Lillian Lee. 2009. A tempest or, on the flood of interest in: sentiment analysis, opinion mining, and the computational treatment of subjective language. Tutorial at ICWSM, May.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Merriam Webster Dictionary. 2014. Merriam Webster Inc. Online: `www.merriam-webster.com/dictionary/irony`. accessed April 28, 2014.

Constantine Nakassis and Jesse Snedeker. 2002. Beyond sarcasm: Intonation and context as relational cues in children's recognition of irony. In A. Greenhill, M. Hughs, H. Littlefield, and H. Walsh, editors, *Proceedings of the Twenty-sixth Boston University Conference on Language Development*, Somerville, MA, July. Cascadilla Press.

New Oxford American Dictionary. 2014. Oxford University Press. Online: `http://www.oxforddictionaries.com/us/definition/american_english/ironic`. accessed April 28, 2014.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Patricia Rockwell. 2005. Sarcasm on television talk shows: Determining speaker intent through verbal and nonverbal cues. In Anita V. Clark, editor, *Psychology of Moods*, chapter 6, pages 109–122. Nova Science Pubishers Inc.

Joseph Tepperman, David Traum, and Shrikanth S. Narayanan. 2006. "yeah right": Sarcasm recognition for spoken dialogue systems. In *Proceedings of InterSpeech*, pages 1838–1841, Pittsburgh, PA, September.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM – A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 162–169. The AAAI Press.

Akira Utsumi. 1996. A unified theory of irony and its computational formalization. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 962–967, Stroudsburg, PA, USA. Association for Computational Linguistics.

Deirdre Wilson and Dan Sperber. 1992. On verbal irony. *Lingua*, 87:53–76.

Deirdre Wilson and Dan Sperber, 2012. *Explaining Irony*, chapter 6, pages 123–145. Cambridge University Press, 1st edition, April.

Hsiang-Fu. Yu, Fang-Lan Huang, and Chih-Jen Lin. 2011. Dual coordinate descent methods for logistic regression and maximum entropy. *Machine Learning*, 85(1–2):41–75, October.

Harry Zhang. 2004. The optimality of naive bayes. In Valerie Barr and Zdravko Markov, editors, *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, pages 3–9. AAAI Press.