문학석사 학위논문

Using Figurative Language and Other Co-textual Markers for the Automatic Classification of Irony

비유언어와 문맥 표지를 이용한 반어법 자동 분류 연구

2014 년 7 월

서울대학교 대학원
언어학과 언어학전공
Andrew Cattle

Using Figurative Language and Other Co-textual Markers for the Automatic Classification of Irony

지도 교수 신효필

이 논문을 문학석사 학위논문으로 제출함
2014 년 6 월

서울대학교 대학원
언어학과 언어학전공
Andrew Cattle

Andrew Cattle 의 문학석사 학위논문을 인준함
2014 년 6 월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

Abstract

Using Figurative Language and Other Co-textual Markers for the Automatic Classification of Irony

Cattle, Andrew

Department of Linguistics

The Graduate School

Seoul National University

This thesis proposes a linguistic-based irony detection method which uses these frequently co-occurring figurative languages to identify areas where irony is likely to occur. The detection and proper interpretation of irony and other figurative languages represents an important area of research for Computational Linguistics. Since figurative languages typically convey meanings which differ from their literal interpretations, interpreting such utterances at face value is likely to give incorrect results. Irony in particular represents a special challenge as, unlike some figurative languages like hyperbole or understatement which express sentiments which are more-or-less in line

with their literal interpretation, differing only in intensity, ironic utterances convey intended meanings incongruent with – or even the exact opposite of – their literal interpretation. Compounding the need for effective irony detection is irony's near ubiquitous use in online writings and computer-mediated communications, both of which are commonly used in Computational Linguistics experiments.

While irony in spoken contexts tends to be denoted using prosody, irony in written contexts is much harder to detect. One of the major difficulties is that irony typically does not present with any explicit clues such as punctuation marks or verbal inflections. Instead, irony tends to be denoted using paralinguistic, contextual, or pragmatic cues. Among these are the co-occurrence of figurative languages such as hyperbole, understatement, rhetorical questions, tag questions, or other ironic utterances which alert the listener that the speaker does not expect to be interpreted literally.

This thesis introduces a divide-and-conquer approach to irony detection where co-occurring figurative languages are identified independently and then fed into an overall irony detector. Experiments on both short-form Twitter tweets and longer-form Amazon product reviews show not only that co-textual figurative languages are useful in the automatic classification of irony but that identifying these co-occurring figurative languages separately yields better overall irony detection by resolving conflicts between conflicting features, such as those for hyperbole and understatement.

This thesis also introduces detection methods for hyperbole and understatement in general contexts by adapting existing approaches to irony detection. Before this point hyperbole detection was focused only on specialized contexts while understatement detection had been largely ignored. Experiments show that these proposed automated hyperbole and understatement detection methods outperformed methods which rely on fixed vocabularies.

Keywords: Figurative Language, Irony, Sarcasm, Hyperbole, Understatement
Student number: 2011-24258

Table of Contents

1 Introduction

Figurative languages typically convey meanings which differ from their literal interpretations. While some figurative languages, such as hyperbole or understatement, express meanings that are more-or-less in line with their literal interpretations, differing only in intensity, irony presents a special challenge as intended meanings may be incongruent with – or even the opposite of – their literal interpretations. Compounding the problem is the wide-spread usage of irony in English and other languages, especially in online discourse and computer-mediated communications. Irony detection is a non-trivial task as irony typically does not present with any explicit clues such as punctuation marks[1] or verbal inflections. Instead, irony tends to be marked using paralinguistic, contextual, or pragmatic cues.

Irony presents a significant problem for automated sentiment analysis and opinion mining. A sentiment analysis or opinion mining system which is unable to correctly identify irony and extract the intended meaning cannot be expected to return accurate results. Consider a company which wishes to gauge customer satisfaction by using data mining techniques on utterances gathered from social media. A naïve solution which doesn't consider irony may misinterpret ironic statements as legitimately positive statements. This may lead to the company overestimating their customer satisfaction and thus potentially costing them significant revenue. This is supported by Carvalho et al.

[1] See http://en.wikipedia.org/wiki/Irony_punctuation for an overview of how irony can be marked using punctuation. Note that in English such markings are optional and not used in the vast majority of cases.

(2009) which found that 35% of their errors identifying positive sentiments were due to the misinterpretation of verbal irony.

This thesis proposes a linguistic-based irony detection method which uses frequently co-occurring figurative languages to identify areas where irony is likely to occur. Specifically, this thesis will examine the effects of hyperbole, understatement, rhetorical questions, and tag questions on the automatic classification of irony. This is the first work to use understatement in automated irony detection. Although previous works have employed simplistic hyperbole and question-based features, this thesis represents the most sophisticated use of these features in irony detection. Finally, this thesis is the first work to employ machine-learning based methods for the automatic classifications of hyperbole and understatement.

1.1 What is Irony?

Before one can begin the task of automatically detecting irony, one must first examine irony from a theoretical perspective. Irony is a complex phenomenon with multiple competing definitions and formations. What is generally agreed upon is that irony can be split into two main types. S*ituational irony* is irony arising from physical or conceptual juxtapositions. V*erbal irony*, also called *sarcasm*, is irony arising from a discrepancy between the literal and intended interpretations of an utterance (Colston, 1997). As this thesis focuses on the detection of verbal irony in texts, except where otherwise noted "irony" may be taken to refer to verbal irony and may be used interchangeably with "sarcasm".

Traditional pragmatic theory identifies irony as a willful violation of conversational maxims such as *utterances should be relevant to the topic at hand* (Grice, 1975's Maxim of Quality) or *utterances should contain all sufficient relevant details* (Grice, 1975's Maxim of Quantity). According to Grice (1975), the violation of these maxims is what signals to listeners that an utterance may have a second, non-literal meaning. Kruez and Glucksberg (1989) introduced the Echoic Reminder Theory of irony, noting that previous models of irony failed to account for the fact that positive statements are more easily identified as irony than negative ones, such as statements (1) and (2). Under this theory, irony is an allusion to shared expectations for the purposes of highlighting a discrepancy between the expectation and the reality.

(1) A fine friend you are. (reproduced from Kruez and Glucksberg, 1989)

(2) You're a terrible friend. (reproduced from Kruez and Glucksberg, 1989)

Regardless of which theory of irony one subscribes to, it should be noted that irony detection by humans is not perfect. Listeners often have to resort to questions like "are you joking?" to confirm whether and utterance is ironic (Kreuz et al., 1999). Kruez and Caucci (2007) notes that "speakers will only employ sarcasm if they are reasonably certain that their hearers will interpret it correctly." This naturally raises the question of how speakers ensure their ironic intent is understood. Kreuz et al. (1999) finds that the amount of "common ground" between the speaker and the listener has a large effect on the listener's readiness and ability to identify irony. Additionally, spoken discourses allow speakers to use laughter or ironic tone of voice (Kreuz and Roberts, 1995; Tepperman et al., 2006) to denote ironic intent while face-to-face discourses further

permit behavioural cues such as winking, eye rolling, smirking, nodding, or even so-called "air quotes" (Kruez et al., 1999).

Written discourse does not allow such cues, making irony identification in written discourses a significantly more difficult task. This was demonstrated in González-Ibáñez et al. (2011) which asked human judges to classify *tweets* collected from the social networking site Twitter as ironic or non-ironic. Some of these utterances had been explicitly marked by their author as sarcastic using Twitter's *hashtag* feature. When these explicit annotations where removed, the human participants were only able to achieve and accuracy of 63%. This was reinforced by the results of Riloff et al. (2013) which found a human-baseline recall of only 45% given a similar experimental set-up.

1.2 Irony and Co-textual Markers

Studies have identified several irony support strategies using co-textual markers which speakers use to covertly signal their ironic intent (see Burgers et al., 2013 and Whalen et al., 2013 for reviews). Kreuz and Caucci (2007) identified several lexical factors which aid in the perception of irony; namely the presence of adjectives and adverbs, the presence of interjections, and the usage of either exclamation points or question marks. Other typological clues include so-called "ironic quotes", emoticons, and laughter onomatopoeias (Burgers et al., 2013).

(3) Kentrell is soooo smart OMG, seriously the modern day Einstein !!! Oh Jeez ! he is the alpha and omega oh gosh ! #sarcasm -_-

The tweet in (3) displays another common indicator of irony; hyperbole (Kreuz and Roberts, 1995; Burgers et al., 2013; Whalen et al., 2013). Other types of figurative language can also be used to signal ironic intent. These include understatement, such as in (4), metaphor, and even other ironic statements (Burgers et al., 2013; Whalen et al., 2013).

(4) Only 50 more problems! Yay! #sarcasm

(5) But don't you just love hearing you might have torn a ligament? I know I sure do #sarcasm #nothanks

(6) Saturday and Sunday classes next week. Great, isn't it? #sarcasm

Rhetorical devices such as rhetorical questions or tag questions can also be used as part of an irony support strategy (Kreuz and Roberts, 1995; Kreuz and Caucci, 2007; Burgers et al., 2013; Whalen et al., 2013). Tweet (5) includes an example of a rhetorical question while (6) includes a tag question.

Finally, Burgers et al. (2013) identifies several stylistic factors that help denote ironic intent. These include the use of cynicism or humour as well as abrupt changes in register. Repetition is another way speakers signal irony. Consider (7), an excerpt from an ironic Amazon book review. The review's author uses repetition not only in their use of "Yes, the author…" but also in the excerpt's call-and-response structure.

(7) Yes, the author has read all the other books. Me, too. Yes, the author knows that Stephanie is torn between two hotties. I got that, too. Yes, the author

knows to include wacky characters and purportedly amusing scenarios. (reproduced from Filatova, 2012)

As stated above, this thesis focuses on the use of hyperbole, understatement, rhetorical questions, and tag questions for irony detection. This thesis is also the first work to offer generalized classifiers for hyperbole and understatement. Thus it is necessary to provide a theoretical background for each of these language devices.

1.2.1 Hyperbole

Hyperbole is the purposeful exaggeration of a statement for rhetorical effect. Cano Mora (2009) notes that hyperbole is "a long neglected trope despite its ubiquity in everyday conversation." Given this ubiquity it comes as a surprise that very little work has been done on the explicit automatic detection of hyperbole. Perhaps this is because, unlike irony, the use of hyperbole does not create significant discrepancy between an utterance's literal and intended interpretations (barring other social factors). Contrasting the hyperbolic (8) with its literal counterpart in (9), one can see that both present fairly similar sentiments with only a slight difference in intensity. While hyperbole can be used as part of a face management strategy (Whalen et al., 2009), it may also be used to purposefully increase the intensity of a statement. Such usages would further minimize the need for dedicated hyperbole detection in sentiment analysis tasks.

(8) That was the best sandwich I've ever eaten in my life!

(9) That was a very good sandwich.

That being said, there are situations in which it may be useful to detect hyperbole. Cano Mora (2009) notes that "exaggeration [is] by far the figure that most often [interacts] with other non-literal forms… [interacting] with every other type of non-literal language with the exception of its logical opposite, understatement." One example of such interactions is how hyperbole can signal ironic intent, as discussed in Section 1. Naturally, the proper detection and interpretation of hyperbole would be a perquisite for exploring these interactions as well as having possible applications in sentiment analysis, politeness profiling, and the automated analysis of face management.

1.2.2 Understatement

Understatement is the purposeful down playing of a statement for rhetorical effect. Like its logical opposite, hyperbole, understatement has been underrepresented in linguistic analysis considering its ubiquity in daily speech. Most works discussing understatement only do so as a means of comparison against other forms of figurative language such as hyperbole or irony (Berntsen and Kennedy, 1996; Colston, 1997).

Also like hyperbole, it is possible that this lack of interest in understatement is because it has a rather mild effect on sentiment analysis. Comparing the understated (10) against its literal counterpart in (11), one can see that both utterances exclude the possibility that the speaker disliked the sandwich in question. Where they differ is that while (11) expresses an unambiguously positive opinion, (10), if interpreted literally, fails to exclude to possibility that the speaker is indifferent. However, in natural speech such a

reading is unlikely and thus “not bad” can generally be assumed to express a positive opinion.

(10) That sandwich was not bad.

(11) That sandwich was very good.

By contrast to sentiment analysis, understatement is incredibly important in face management (Whalen et al., 2009), in addition to its usefulness in signaling ironic intent as discussed in Section 1.2. Although hyperbole and understatement have different effects, speakers often employ them in similar contexts and for similar purposes. As such, understatement detection has the same potential applications as those discussed for hyperbole detection in Section 1.2.1. Namely, sentiment analysis, politeness profiling, and face management analysis.

1.2.3 Rhetorical Questions

Rhetorical questions, which are intended by speakers to lead, persuade, or impress listeners, differ from genuine questions in that they are not a sincere attempt to illicit information (Schmidt-Radefeldt, 1977). Interestingly, rhetorical questions remain an effective persuasion strategy despite the fact that listeners are well versed with, and can readily identify, this tactic (Frank, 1990). As such, rhetorical questions are of great interest in conversation or discourse analysis. Rhetorical questions may also have use in opinion mining as they may signal a speaker’s private state. For example, even though (12) does not contain any explicit opinions it is reasonable to assume the speaker has a positive opinion towards chocolate.

(12) Is there anything better than chocolate?

One of the major difficulties when attempting to differentiate rhetorical questions from genuine questions is that, like irony, rhetorical questions tend not to present any explicit clues. Given that rhetorical questions and genuine questions have very different intents, the only sure way to distinguish between the two is through the non-linguistic context of the utterance. (Schmidt-Radefeldt, 1977) Moreover, Frank (1990) challenges the traditionally held view that unlike genuine questions, rhetorical questions do not illicit responses from listeners, stating "there are also instances where taking into account the hears' responses may only complicate, rather than facilitate, analysis and interpretation." This can be seen in the hypothetical exchange in (13) where a parent, A, is chastising their teenage child, B, regarding peer pressure. Even though the question asked by A is contextually clearly meant to be rhetorical, B still offers a response.

(13) A: If all your friends jumped off a bridge, would you jump too?

B: Of course not!

1.2.4 Tag Questions

Tag questions have a wide range of usages from requesting confirmation, providing emphasis, or simply allowing the speaker to confirm the listener is still engaged in the conversation. Although common in spoken discourses, tag questions rarely appear in formal written utterances. This may explain why their detection in text has not received any attention. However, the increase of informal written discourses such as chat logs and social media interactions means that tag question identification may have applications in discourse analysis.

2 Previous Works

2.1 Irony Detection

Utsumi (1995) and Utsumi (1996) represent some of the first attempts to develop a computational model of irony. These papers lay out an algorithm for detecting irony from a formal pragmatic view point. Unfortunately, this approach requires knowledge of the speaker's and listeners' private states – expectations, desires, etc. This means implementing these algorithms is impractical both in accurately identifying such private states given a finite amount of context but also modeling these states such that useful comparisons between the private state and uttered opinion can be made.

Due to these limitations, later sarcasm detection works tended to focus on lexical or paralinguistic cues. Tepperman et al. (2006) used prosody and laughter to identify sarcasm in spoken language systems. Inspired by the work of Kruez and Caucci (2007), several studies including Carvalho et al. (2009), González-Ibáñez et al. (2011), and Vanin et al. (2013) used such features as exclamation marks, quotation marks, ellipsis, emoticons, and laughter onomatopoeias to aid in the identification of irony in text.

Another common approach was to look for specific phrases or patterns which tend to denote irony. Tepperman et al. (2006) looked for the spoken phrase "yeah right". Carvalho et al. (2009) employed several fixed phrases common to the expression of irony in Brazilian Portuguese.

The main disadvantage of these approaches is a lack of coverage as they not only rely on manually compiled lists of phrases and structures, but they also fail to detect variations

of these phrases or structures which may appear in real word data. Davidov et al. (2010) and Tsur et al. (2010) attempt to rectify this by utilizing the automated pattern extraction techniques developed in Davidov and Rappoport (2006) to automatically extract phrases and structures from ironic texts, resulting in a greater coverage of patterns. Additionally, their solution was capable of identifying near and partial matches making it a much more flexible solution. Go and Bhayani (2010) and González-Ibáñez et al. (2011) implement a somewhat simpler approach to automatic pattern extraction, using surface n-grams and POS tag n-grams.

Several studies attempted to simplify the irony detection task by limiting themselves specific forms of irony expression. Veale and Hao (2010) examined ironic similes by using web search APIs to judge the semantic appropriateness of the simile/comparison. Riloff et al. (2013) tackled the identification of ironic statements in the form of a positive sentiment combined with a generally negative situation, such as in (14), using a bootstrapping approach.

(14) I love having to the work on my day off.

While the lexical and typological aspects of irony have been thoroughly explored and exploited in the various studies discussed so far, surprisingly little attention has been paid to the figurative languages, rhetorical devices, and stylistic features discussed in Section 1. Go and Bhayani (2010) looks for exaggeration words as well as for other stylistic features such as profanity and alliteration. González-Ibáñez et al. (2011) takes a

more psycholinguistic approach, making use of LIWC+[2] and WordNet Affect[3] lexical categories.

However, the most in-depth examination of these co-textual markers has been Reyes and Rosso (2011) which uses humour-detection related features as well as politeness profiling, polarity, and affect to identify irony in Amazon product reviews. This approach is continued in Reyes et al. (2012) which models semantic and syntactic ambiguity in additional to using polarity, emotional scenarios, and unexpectedness (i.e. semantic un-relatedness) to differentiate ironic tweets from humourous tweets, political tweets, or technology-related tweets. Pérez (2012) offers a more in-depth analysis of this approach.

2.2 Detection of Co-textual Markers

With the exception of the automated identification and interpretation of metaphor, which has been an active area of research (see Shutova, 2010 for a review), very little work has been done on the automatic detection of any of the co-textual irony markers discussed in Section 1.2. The detection of these markers was generally treated as a subtask or offshoot of a larger natural language processing task.

Automated hyperbole detection, for example, is treated as a subtask of irony detection in Go and Bhayani (2010). The irony detection system developed in Go and Bhayani (2010) included an "Exaggeration" feature which they defined as "words like 'so', 'very', 'absolutely' which are extremely polar in nature". Wu and Kao (2012)

[2] Linguistic Inquiry and Word Counting, http://www.liwc.net/
[3] http://wndomains.fbk.eu/wnaffect.html

presents to first look at hyperbole detection as an independent problem, proposing a detection method for number-based hyperboles such as the one in (15). Their approach takes a poll of real-world expected values and compares them against the uttered value. Any sufficiently large discrepancy is classified as hyperbole.

(15) These tickets must have cost you like $1000000!

The main shortcoming of these hyperbole detection approaches is that they lack coverage. According to the results of Cano Mora (2009), hyperbole of the type covered in Go and Bhayani (2010) account for only a third all hyperbole while number-based hyperboles such as those tackled in Wu and Kao (2012) account for only 14%. This leaves a large room for improvement.

Given the lack of interest in hyperbole detection, it is no surprise that automated understatement detection has been completely unexplored. However, due to the similarities between hyperbole and understatement and their relationship as logical opposites, it stands to reason that the hyperbole approach of Wu and Kao (2012) could be adapted to look for number-based understatements, such as the one in (16), with minimal effort.

(16) It's not a big deal. It only took me like 2 minutes.

Although there has been some work on the automatic classification of questions, these works have not specifically addressed rhetorical questions or tag questions. For example, Li et al. (2011) tackles the task of identifying tweets which attempt to illicit information. While it is tempting to assume that any question not inviting information is in fact a

rhetorical tweet, this is not necessarily the case as this would also include such categories as advertisements, titles, and trivia question/answer pairs. The application of these techniques to rhetorical question detection would be a matter for further examination.

3 Data Collection

3.1 Twitter Data

Twitter[4] is a microblogging site which allows users to submit short messages, called *tweets*, of up to 140 characters in length. Due to the site's popularity and the short, relatively self-contained nature of tweets, Twitter has been a popular source of data in sentiment analysis and opinion mining tasks; Pak and Paroubek (2010) and Davidov et al. (2010) being notable early examples.

Tweets typically contain certain features typical of online speech such as hyperlinks, slang, abbreviations, and emoticons. Additionally, Twitter users may refer to each other using the format @<*username*> or explicitly mark the topic or theme of the tweet using the format #<*tag*>. These so-called *hashtags* commonly refer to specific events, such as using *#Sochi2014* to refer to the 2014 Winter Olympics in Sochi, Russia, or to emotions or private states, such as *#happy*, *#upset*, or *#tired*. These hashtags are informal and are created by the users themselves.

Davidov et al. (2010) notes that since hashtags are added by the author of a tweet, the inclusion of *#sarcasm* or a similar hashtag in a tweet represents a reliable indication that it was intended to be interpreted sarcastically and thus can serve as a gold standard for sarcastic texts. This approach has been continued in such sarcasm and irony detection works as Go and Bhayani (2010), González-Ibáñez et al. (2011), Reyes et al. (2012), Vanin et al. (2013), and Riloff et al. (2013).

[4] http://twitter.com/

It should be noted that this hashtag-based data collection approach can be expanded to collect different types of figurative languages. For example, Reyes et al. (2012) used *#humour* to identify humourous texts. Remembering that the creation and usage of Twitter hashtags is entirely at the discretion of Twitter users, it is not unreasonable to assume that other types of figurative language are also explicitly marked using hashtags in much the same way *#sarcasm* is used to explicitly mark sarcastic intent.

One major disadvantage of using Twitter as a corpus is that Section I Article 4.A of Twitter's Developer Rules of the Road clearly states that users may not "sell, rent, lease, sublicense, redistribute, or syndicate … Twitter Content to any third party without prior written approval from Twitter." (2013) This makes the sharing of Twitter-based corpora extremely difficult and means it is easier for researchers to compile their own individual Twitter-based corpora than to create and distribute a standardized corpus. The fact that all Twitter-based Irony Detection works use different datasets makes it impossible to compare their results directly.

In line with previous Twitter-based Irony Detection experiments (Davidov et al., 2010; Tsur et al., 2010; Go and Bhayani, 2010; González-Ibáñez et al., 2011; Reyes et al., 2010; Vanin et al., 2013; Riloff et al., 2013), this thesis compiled its own Twitter figurative language corpora. These corpora consisted of real world tweets collected using Tweepy[5], a Python implementation of Twitter's streaming API[6]. Tweets were

[5] http://github.com/tweepy
[6] http://dev.twitter.com/docs/streaming-apis

collected between August 10th, 2013 and October 21st, 2013. Tweets were assigned labels based on their hashtags, as will be described in Sections 3.1.1, 3.1.2, and 3.1.3.

Several heuristic measures were implemented to further refine the data. Retweets, when a user republishes another user's tweet, were filtered out by using common retweet patterns. Non-English tweets were identified and removed using a stop word-based approach. Tweets consisting of only usernames, hashtags, and hyperlinks were removed as they were deemed a poor fit for linguistic-based analyses. Finally, usernames and hyperlinks found in tweets were replaced with generic placeholders and all hashtags used in the annotation process were removed.

Tweets not containing any of the target hashtags were also collected for use as negative examples in experiments. These *general tweets* were collected using Twitter Streaming API's random sampling method, ensuring the collected tweets were representative of the type of language used on Twitter. These tweets were then subjected to the same labeling and sanitization processes detailed above resulting in a total of 422284 unique tweets.

It should be noted that because hashtags are completely optional, not all genuine examples of a specific phenomenon are labeled. Thus, there exists a possibility that false-negative examples may appear in the data. Although this study assumes that such false-negative examples represent an insignificant portion of the data, an assumption implicitly shared in all studies using similar hashtag-based annotated data, this may be a topic for future discussion.

3.1.1 Twitter Irony Corpus

Tweets containing the hashtags *#sarcasm* or *#sarcastic* were assumed to represent true examples of irony. The hashtags *#irony* and *#ironic* were purposefully avoided to prevent the collection of examples of situational irony. A total of 612401 ironic tweets were collected and sanitized as detailed above. For computational reasons, 10000 tweets were randomly selected from the full set of ironic tweets to be used as the test data. A further 10000 tweets were randomly selected from the full set of general tweets, for a total of 20000 tweets.

3.1.2 Twitter Hyperbole Corpus

Tweets containing one or more of the hashtags *#hyperbole, #exaggeration, #exaggerating,* or *#overstatement* were taken to represent true examples of hyperbole. This resulted in 3708 hyperbolic tweets. An equal number of tweets were randomly selected from the full set of general tweets and used as non-hyperbolic examples for total of 7416 tweets.

3.1.3 Twitter Understatement Corpus

Tweets containing the hashtag *#understatement* were taken to represent true examples of understatement. This resulted in 7255 understated tweets. An equal number of tweets were randomly selected from the full set of general tweets and used as non-understated examples for total of 14510 tweets.

3.2 Amazon Data

While much attention has been paid to irony on Twitter due to the ease of collecting author-annotated data, Amazon product reviews are another common area of interest for irony detection. The most obvious difference between Twitter data and Amazon product reviews is that while tweets are limited to 140 characters, reviews can be much longer. Additionally, since, even in reviews written with ironic intent, not every utterance is itself ironic, Amazon product reviews contain a greater amount of context than tweets which are effectively context-free. This presents a very different challenge for irony detection compared to Twitter-based approaches. (Davidov et al., 2010; Filatova, 2012) Without Twitter's length restrictions authors are also better able to structure their ideas and provide co-textual irony markers to signal irony in advance. As such, it is the conjecture of this thesis that the detection of contextual irony markers will be even more beneficial for Amazon product reviews than for tweets.

Amazon product reviews lack Twitter's hashtag feature and thus there is no easily way to identify ironic reviews. Luckily, the Sarcasm Corpus[7] introduced in Filatova (2012) consists of annotated ironic and non-ironic Amazon product reviews. It is important to note that Filatova (2012)'s Sarcasm Corpus annotated irony at a macro level. That is, while the reviews themselves are annotated as ironic or not but the individual utterances in each review are not. Although Filatova (2012) asked annotators to identify the specific utterances which make a review ironic, these are not explicitly marked in the Sarcasm Corpus. Additionally, although Sarcasm Corpus reviews contain metadata,

[7] http://storm.cis.fordham.edu/~filatova/SarcasmCorpus.html

such as the product being reviewed and the review's star rating, this thesis is focused on the linguistic aspects of irony and thus only the review's title and body were considered.

4 Experimental Set-up

Machine learning algorithms have been increasing in popularity for use in Natural Language Processing tasks due to their ability to automatically extract non-obvious patterns from feature sets. Figure 1 shows the basic structure of a machine learning classifier.

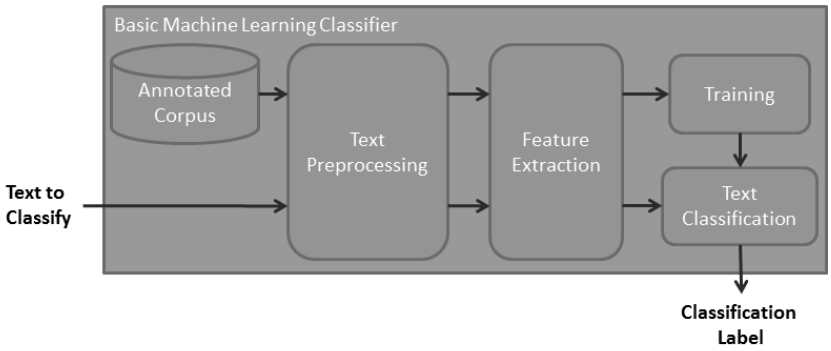

Figure 1 Normal Machine Learning Architecture

Inspired by irony detection irony classifiers of Go and Bhayani (2010) and González-Ibáñez et al. (2011), this thesis also adopted an n-gram based machine learning approach to irony detection. Moreover, given the interrelationships between irony and other forms of figurative language described in Section 1.2, this thesis posits that this same approach can be used for hyperbole and understatement detection tasks as well. In their simplest form, classifiers were trained on surface n-gram and POS tag n-gram frequencies. While the majority of the classifiers in this thesis followed this structure, variations and alternative approaches will be described when required.

All n-grams and POS tag n-grams used in this thesis were generated by tokenizing and POS tagging each document in a corpus using the Punkt tokenizer and Penn Treebank Maxent POS tagger implementations included with NLTK[8] 3.0. From these tokens and POS tags, n-grams were generated for all values of n such that $1 \leq n \leq 4$. Separate classifiers were trained for each individual set of n-grams, $1 \leq n \leq 4$, as well as for select combinations thereof. Classifiers were trained using the Linear Support Vector Machine (SVM) implementation found in SKLearn[9]. All experiments were conducted using 90% of the data for training and the remaining 10% for testing. All results were subjected to a 10-fold cross validation.

4.1 Hyperbole Detection

A series of fixed word list-based feature sets were created and used to establish a baseline performance for hyperbole detection. The first list, the Hyperbole Word List (HWL), was created by manually selecting keywords from the sample hyperbole words and phrases included in Cano Mora (2009) as well as through native-speaker intuition. The HWL contains 185 unique words which can be found in Appendix 1. Since the sample hyperbole words and phrases in Cano Mora (2009) cover a wide range of hyperbole categories, the HWL is expected to offer greater coverage of real-world hyperbole phrases than the word list used in Go and Bhayani (2010), which appears to be limited to intensifiers. Three other lists were generated using the HWL as a seed. The Hyperbole Stem List (HSL) consists of 149 word stems generated by removing

[8] Natural Language Toolkit, a popular Python library for processing text. http://nltk.org/
[9] SciKit Learn, a popular Python library for machine learning. http://scikit-learn.org/

inflections from HWL words using the Porter Stemmer[10] implementation included with NLTK. The Thesaurus-Expanded Hyperbole Word List (TEHWL) consists of 1389 words which were generated by collecting all synonyms found in all WordNet 3.1[11] synsets for each HWL word. Each synset represents only one sense (or meaning) of a word. Most words have multiple synsets and not all of them are necessarily themselves hyperbolic. As such, the TEHWL is expected to generate more false-positives than the HWL or HSL. Finally, the Thesaurus-Expanded Hyperbole Stem List (TEHSL) consists of 1273 word stems generated by removing inflections from TEHWL words, again using the Porter Stemmer. For each tweet the frequency of each HWL and TEHWL word was computed along with the total number of matches for each list. Frequencies for HSL and TEHSL words were computed in a similar manner but with the extra step of first stemming each word in the tweet.

The results of this word-list based approach were compared against a surface n-gram and POS n-ram based machine learning classifier. The classifier was trained on n-grams generated from the Twitter Hyperbole Corpus described in Section 3.1.2, both following the method detailed at the beginning of Section 4.

4.2 Understatement Detection

Given the effect of hedges to weaken an assertion or to create distance between an assertion and a speaker, it comes as no surprise that hedges have a strong relationship with understatement (Hübler, 1983). Following the example set by the hyperbole

[10] http://tartarus.org/martin/PorterStemmer/
[11] http://wordnet.princeton.edu/

detection experiments described in Section 4.1, a series of fixed word list-based feature sets were created and used to establish a baseline performance for understatement detection. The first list, the Hedge Word List (HedgeWL), was created by manually selecting keywords from example hedging phrases found across several popular grammar websites as well as through native-speaker intuition. The HedgeWL contains 45 unique words which are reproduced in Appendix 2. As before, three other lists were generated using the HedgeWL as a seed. The Hedge Stem List (HedgeSL) consists of 40 word stems generated by removing inflections from HedgeWL. The Thesaurus-Expanded Hedge Word List (TEHedgeWL) consists of 341 words which were generated by collecting all synonyms found in all WordNet synsets for each HWL word. Again, since most words have multiple senses, and thus multiple synsets, and not all of them are necessarily themselves hedges, the TEHedgeWL is expected to generate some false-positive matches. Finally, the Thesaurus-Expanded Hedge Stem List (TEHedgeSL) consists of 321 word stems generated by removing inflections from TEHedgeWL words. Once again, HedgeWL and TEHedgeWL word frequencies and total counts were computed for each tweet. Frequencies for HedgeSL and TEHedgeSL words were computed in a similar manner but with the extra step of first stemming each word in the tweet.

The results of this word-list based approach were compared against a surface n-gram and POS n-ram based machine learning classifier. The classifier was trained on n-grams generated from the Twitter Understatement Corpus described in Section 3.1.3, both following the method detailed at the beginning of Section 4.

4.3 Rhetorical Question Detection

Although, as discussed in Section 0, rhetorical questions often appear without any explicit clues, Schmidt-Radefeldt (1977) does note several structures which do tend to indicate rhetorical intent. First and foremost is the "question and direct answer" structure seen in (17). Here a speaker asks a question and then immediately supplies an answer. Another extremely common strategy is the embedding of the *wh*-question into matrix sentences, such as in (18). Finally, "Auto-responsive Rhetorical Questions" (ARQs) are questions where the speaker sets up a context in which no answer expect the one intended by the speaker can be considered acceptable. Such questions take two forms. The first is questions utilizing "Expressions of Exclusive Absoluteness" (EEAs), like (19). The second is questions utilizing 'summing up' phrases, like (20).

(17) And what do I have to show for it? **Nothing.**

(18) Do you know **how much that costs**?

(19) Who would burn a cheque **other than** a fool? (reproduced from Schmidt-Radefeldt, 1977)

(20) It had to be John. **After all,** who else had the motive and opportunity?

Although theoretically a syntactic parser, such as the Stanford Parser[12], should be able to reliably identify such syntactic structures as embedded *wh*-questions, early experimentation found that these tools had trouble returning consistent results given mild variations of the same sentence. "Question and direct answer" structures also

[12] http://nlp.stanford.edu/downloads/lex-parser.shtml

proved difficult to automatically detect in all but the simplest yes/no questions. As such, these avenues were eventually dropped.

Unlike other forms of rhetorical questions, ARQs have a lexical component, whether it be an EEA, such as those listed in Table 1, or one of the aforementioned 'summing up' phrases, like *after all* or *in the end*. As such, these forms were relatively straight forward to detect. Also unlike other forms or rhetorical questions, embedded wh-phrases leave a syntactic footprint. While no reliable parsing-based detection method was discovered, such structures could still be identified by looking at the sequence of POS tags generated by a sentence. Using these observations, ad hoc methods were developed for detecting these types of rhetorical questions using regular expressions.

Table 1 Examples of EEAs

Expressions of Exclusive Absoluteness
apart from
aside from
Barring
But
Except
Excluding
if not
other than
save for
short of

4.4 Tag Question Detection

Tag questions, unlike rhetorical questions, are unmistakable from a syntactic perspective; they should be easily identifiable using existing parsing methods. Unfortunately, early experimentation revealed that distinguishing a tag question from any other added clause was problematic using the parser's output alone.

(21) <modal or aux><optional negative contraction> <pronoun>?

English tag questions tend to follow the structure in (21). Given that English contains a relatively small number of modal and auxiliary verbs as well as a small number of pronouns, this was deemed to be one situation were a fixed list seemed to be an acceptable solution. Several idiomatic tag question forms such as "yes?", "right?", and "eh?" were also identified by manually examining utterances in the Switchboard Dialog Act Corpus[13] which had been hand annotated as a tag question. Twenty four regular expressions were then created, collectively capable of matching all of the compiled tag questions.

Three variations on these regular expressions were created. Context 1 looked for any match, no matter where it came in the sentence. The major disadvantage of this particular set of regular expressions was that they could not differentiate between a tag question and regular subject-verb inversion. Context 2 looked for matches which occurred only immediately before question marks, immediately before the end of an

[13] ftp://ftp.ldc.upenn.edu/pub/ldc/public_data/swb1_dialogact_annot.tar.gz

utterance, or as an interjection. Context 3 was the same as Context 2 but added matches immediately before periods and exclamation points.

Experimentation showed that Context 3 was the most strongly correlated with irony; followed closely by Context 2. This result was somewhat expected since Twitter users do not always conform to standard grammar or punctuation rules. Context 1 actually proved to be correlated with non-ironic utterances. Manual examination of the Context 1 matches confirmed that most of the matches were indeed false positives, making Context 1 a better indicator of genuine questions as opposed to rhetorical ones. Given these results, Context 3 was chosen for use in the irony detection experiments described below.

4.5 Irony Detection

The results of Go and Bhayani (2010) and González-Ibáñez et al. (2011) show that a machine learning classifier trained on surface n-gram and POS n-gram frequencies is an effective method for detecting irony. As such, a baseline irony classifier was trained following the structure of Figure 1.

A variation on this structure added features representing the co-textual markers hyperbole, understatement, rhetorical questions, and tag questions. Hyperbole and understatement classifiers were created following the methods outlined in Sections 4.1 and 4.2. Although previous experiments used a 90/10 split for training/test data, the hyperbole classifier was trained using the entire Twitter Hyperbole Corpus and the understatement classifier was trained using the entire Twitter Understatement Corpus. It

should be noted that there was no overlap between the Twitter Irony Corpus and either the Twitter Hyperbole Corpus or the Twitter Understatement Corpus. It is also important to note that the hyperbole and understatement classifiers were trained completely independently from the irony classifier and, for the purposes of this experiment, they were considered as black boxes. A document's hyperbole feature consisted of the hyperbole classifier's output for that document mapped to a binary value such that 0 indicated an absence of hyperbole while 1 indicated that a document was hyperbolic. Similarly, the understatement feature consisted of the output of the understatement classifier. The rhetorical questions feature and tag questions feature were simply the number of rhetorical questions and tag questions, respectively, detected using the ad hoc patterns defined in Sections 4.3 and 4.4. A fifth and final new feature called "Total Marker Count" was the sum of the hyperbole, understatement, rhetorical questions, and tag questions features. These features were then combined with the regular n-gram feature, resulting in the structure seen in Figure 2.

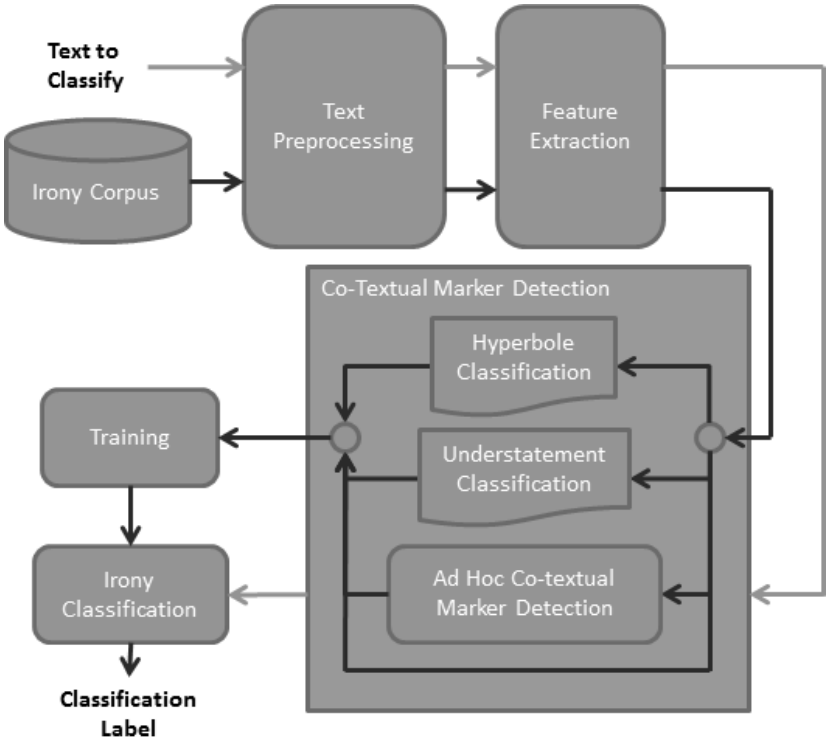

Figure 2 Irony Detection Algorithm Architecture

4.5.1 Twitter Data

In addition to a co-textual marker based classifier which was trained following the method described in the previous section, an additional classifier was trained on surface n-gram and POS n-grams alone to serve as a baseline.

4.5.2 Amazon Product Review Data

Until this point we have been implicitly assuming each document is a single utterance. While this seems to be a reasonable assumption for Twitter data given their short length, Amazon product reviews can be several paragraphs long and cover numerous topics.

Therefore, it is unreasonable to treat entire Amazon product reviews as a single utterance. To address this issue, reviews were split into individual sentences using the Punkt tokenizer included with NLTK 3.0. Each sentence was then processed as a single, independent utterance using the method defined Section 4.5. For each review, these sentence-level features, including co-textual irony marker features, were summed to create a single document-level set of features which was then supplied to the machine learning algorithm.

The discourse-like nature of Amazon reviews also allowed for an additional co-textual marker; sarcasm. While the Twitter-based experiment described in Section 4.5.1 is concerned with utterance-level irony, this Amazon product review-based experiment is concerned with document-level irony. Inspired by results of Burgers et al. (2013) which showed that ironic utterances may be used to signal further ironic utterances, the presence of a large number of ironic sentences in a document may be a strong indication that the overall document is also ironic. A sentence-level sarcasm classifier was created following the method described in Section 4.5.1. Like hyperbole and understatement, each sentence in a review was supplied to the sarcasm classifier separately and the output was mapped to a binary value where 1 indicated sentence-level ironic intent and 0 indicated no ironic-intent. These values were then summed and supplied to the document-level irony classifier described in this section. Unlike the hyperbole and understatement classifiers which were trained on a combination of unigrams and bigrams, based on the results of the Twitter data experiment in Section 5.3.1, the

sarcasm classifier was trained on unigrams, bigrams, as well as the co-textual markers hyperbole, understatement, rhetorical questions, tag questions, and total marker count.

5 Results and Discussion

5.1 Hyperbole

Table 2 Hyperbole Classification Results. Bold values represent the highest result achieved.

	Precision	Recall	F-Score
Unigrams	0.768651	0.74036	0.754051
Bigrams	0.786633	0.730769	0.75743
Trigrams	0.766116	0.670792	0.715
4-grams	0.758651	0.529155	0.623274
Unigram + Bigram	**0.797017**	**0.754365**	**0.7748**
Unigram + Bigram + Trigram	0.794845	0.753877	0.773633
Unigram + Bigram + Trigram + 4-gram	0.789575	0.745999	0.767012
HWL	0.683648	0.426078	0.524627
HSL	0.684175	0.464935	0.553343
TEHWL	0.695185	0.494452	0.486627
TEHSL	0.70094	0.517834	0.595011

The results of Table 2 highlight the advantage of using an n-gram based approach over a fixed word list. While all the fixed word lists showed precisions higher than chance, their real weak point was there lack of coverage which resulted in poor recall scores. The n-gram based approaches overcome this problem by allowing the machine-learning algorithm extract patterns from all bigrams instead of using only the subset of words used in the fixed lists. This resulted in better coverage.

Encouraged by these initial results, a second experiment was conducted to test the effectiveness of adding word list count features to the n-gram based classifier. Using the

combined unigram and bigram features as a base, a new feature was added representing the number word list members matched by each tweet. The words themselves were not added since the inclusion of unigrams makes the individual word counts redundant. Unfortunately, this was shown to have no statistically significant effect.

The most informative features for combined unigrams and bigrams, seen in Table 3, show some interesting patterns. The majority of these hyperbole words fit into the hyperbole categories described in Cano Mora (2009), such as "Idea of sorrow or pain" (*painful*), "Idea of non-existence" (*'m_never*), or "Idea of violence, destruction" (*attacked*). It is important to note that these are not the type of hyperboles which would have been captured using the existing methods.

Table 3 Most Informative Features for Unigram/Bigram Linear SVM Classifier

Hyperbole	Not Hyperbole
bald	*#teamfollowback*
space	*true_!*
#stupid	*mins*
48	*the_question*
not_true	*pandora*
btw	*like_not*
voice	*#np*
correct	*getting_into*
you_mean	*is_amazing*
it_a	*._@USER*
mate	*a_new*
million	*yes_i*
#butreally	*away_.*
famous	*absolute*
'm_never	*lucky*
painful	*pics*
so_good	*@USER_what*
worse	*thing_.*
attacked	*my_favorite*
die	*would_like*

(22) Grapefruit juice is amazing👌😊

(23) Ham is now my favorite food.

(24) Huell is my favorite Breaking Bad character.

(25) I'm so lucky to have this amazing guy

With the exception of *is_amazing* and *my_favorite*, the non-hyperbolic words are not particularly interesting as there are no easy to extract patterns. What is worrying, however, is the inclusion of *is_amazing* and *my_favorite* which intuition would dictate should be more likely to be hyperbolic instead of strongly non-hyperbolic. Examining the data, there are some examples, such as (22) and (23), which definitely appear to be hyperbolic. However, in the case of *my_favorite*, the majority of tweets appeared to be literal, for example (24). *is_amazing* is slightly trickier to explain. Many of the tweets, like (25), were aimed at loved ones and thus explicit marking of hyperbole may have been seen as a sarcastic utterance. There is also an argument to be made that the speakers are using an informal definition of "amazing", something more akin to "very very good" than "causing great wonder", and thus these are not genuine hyperbole. Regardless, this perfectly showcases the shortcomings of using Twitter hashtags for annotation. As discussed in Section 3.1, hashtags are entirely at the author's discretion and thus many examples of figurative languages may go unmarked. Despite this, the hyperbole classifier was able to identify several intuitively hyperbolic patterns in the data such as *famous*, *million*, or *painful*, proving the viability of this annotation method.

Even though the n-gram based approach proved superior to the fixed word list approach, it is still worthwhile to examine the fixed list results. Specifically, the results in Table 2

clearly show that both stemming and thesaurus-based list expansion result not only large gains in recall, but also small gains in precision. This comes as no surprise given that fixed lists look only at a finite subset of words or stems and that If a tweet does not contain any of these words/stems then the classifier has no information with which to judge the tweet as hyperbolic or not. By either adding more words to the list, as with thesaurus expansion, or by making the list items easier to match, as with stemming, the number of tweets containing the relevant words/stems is expected to go up, giving the classifier more information to work with and thus providing better results. Additionally, the improved results of stem lists compared to their respective word lists also provide evidence that inflections are not important to the expression of hyperbole.

As discussed in Section 4.1, the TEHWL was generated by naively selecting all synonyms for all senses of each HWL word in WordNet, not all of which are necessarily hyperbolic. While the inclusion of non-hyperbolic words would be expected to introduce false positives, this was mitigated by the use of a machine learning algorithm to automatically weight the individual items in the list. This means that any non-hyperbolic words were implicitly identified by the machine learning algorithm and given a low or negative weight. Examining the most informative TEHWL features shown in Table 4, one can see that typically non-hyperbolic words such as *circumstances*, *bundle*, or *peck* received non-hyperbolic weights while words such as *starving*, *eternally*, and *ruined* were correctly identified as being typically hyperbolic.

Table 4 Most Informative Features for TEHWL Linear SVM Classifier

Hyperbole	Not Hyperbole
chip	*bonkers*
curve	*trouble*
beard	*circumstances*
severe	*waste*
hr	*burst*
age	*wonderful*
starving	*bundle*
inch	*madly*
eternally	*heavy*
cypher	*sum*
highly	*cracked*
pass	*beyond*
humanity	*vivid*
ruined	*shameful*
massive	*cracking*
correctly	*amazed*
million	*keen*
pile	*c*
load	*peck*
starvation	*outstanding*

However, as with the most informative unigram/bigram features in Table 3, Table 4 also seems to contain some false negatives such as *bonkers* or *madly*; both of which would be expected to be hyperbolic. Again, this seems to be due to a limitation of the hashtag annotation method used for all Twitter-based experiments in this thesis. The use of

bonkers in (26) is clearly hyperbolic but the author has simply not explicitly annotated their tweet as such.

(26) i'm going bonkers just waiting for eyes to ship. it hasn't even been 14 business days for them. this 3-4 month doll wait is gonna kill me XD

5.2 Understatement

Table 5 Understatement Classification Results. Bold values represent the highest result achieved.

	Precision	Recall	F-Score
Unigrams	0.767925	0.758162	0.762719
Bigrams	0.750017	0.732069	0.740847
Trigrams	0.691339	0.71884	0.704723
4-grams	0.632247	0.7117	0.669453
Unigrams + Bigrams	**0.785221**	**0.77757**	**0.781272**
Unigrams + Bigrams + Trigrams	0.783329	0.774353	0.778708
Unigrams + Bigrams + Trigrams + 4-grams	0.777833	0.77118	0.774385
HedgeWL	0.759194	0.163325	0.268672
HedgeSL	0.722719	0.173018	0.278728
TEHedgeWL	0.750602	0.293452	0.42177
TEHedgeSL	0.73219	0.320452	0.445467

As with the hyperbole word lists results in Section 5.1, Table 5 shows that while the hedge word lists show very good precision, rivalling that of the n-gram based classifier, they fail when it comes to coverage. Understatement is simply too complex a phenomenon and natural language is simply too diverse to be encapsulated in a fixed word list.

As with hyperbole, a second experiment was conducted to test the effectiveness of adding word list count features to the n-gram based classifier. Using the combined unigram and bigram features as a base, a new feature was added representing the number word list members matched by each tweet. The words themselves were not added since the inclusion of unigrams makes the individual word counts redundant. Again, like with hyperbole, this was shown to have no statistically significant effect.

The most informative features for understatement, seen in Table 6, once again offer some interesting patterns. The non-understated features include typically hyperbolic phrases such as *the_worst* and *slightly_obsessed*. Given that hyperbole is overstatement, the opposite of understatement, this is not surprising.

Table 6 Most Informative Features for Understatement

Understatement	Not Understatement
#excited	*slightly_obsessed*
#pissedoff	*@USER _amazing*
#tired	*moves*
lawn	*is_love*
the_back	*!_ @USER*
#exhausted	*list*
slightly	*the_worst*
not_feeling	*a_cold*
#bored	*pathetic*
of_8	*homie*
exhaustion	*(POS TAG) VBZ_CC*
finding	*is_it*
#gutted	*vs.*
in_tomorrow	*an_ugly*
smashed	*is_this*
rough	*what's*
this_morning	*yes_you*
getting	*trending*
's_hot	*haha_i*
walt	*got_ta*

What is surprising is the pro-understatement features' inclusion of high intensity language like *#excited* which one would not expect to see in an understated utterance. (27) provides a perfect example of why this may be. In such cases, users are not including *#understatement* in their tweets to signal genuine understatement, but rather

creating an ironic context in which even hyperbole fails to capture the level of their excitement. In fact, if the purpose of understatement is to downplay the importance or intensity of a statement then explicitly marking it with a *#understatement* tag may be counterproductive as it calls extra attention that the statement you wished to downplay.

(27) #excited is an #understatement

Also of note in (27) is that the hashtags used do not appear outside of the discourse but in fact play a role in the sentence. This too can lead to false-positive annotations as utterances *about* understatement are not necessarily themselves understated. González-Ibáñez et al. (2011) addresses a similar issue in their sarcasm detection experiments by manually reviewing and removing any discourse-intensive #sarcasm tags.

Despite these issues, the understatement classifier was still able to capture some useful patterns, such as the hedge word *slightly* or *not_feeling*, which uses negation to avoid directly expressing a negative sentiment. Compare the tweet in (28) against its literal interpretation in (29).

(28) fair to say not feeling 100% today #understatement

(29) I feel sick

Table 7 Most Informative Features for TEHedgeWL

Understatement	Not Understatement
slight	*conceive*
tad	*twin*
prick	*smell*
slightly	*image*
possibly	*spirit*
reasonably	*pinch*
somewhat	*refer*
span	*tone*
bit	*venture*
kinda	*reason*
fairly	*debate*
quite	*sting*
reckon	*upright*
concern	*scrap*
impact	*speculation*
rather	*approximately*
considered	*relate*
so-called	*evidently*
classify	*reach*
screen	*insignificant*

The fixed word list results confirm the expectation expressed in Section 4.2 that hedges are used to express understatement. This is backed up by the list of TEHedgeWL's Most Informative Features, found in Table 7, which weighted such hedges as *slight*, *tad*, and *somewhat* as being highly understated while weighting more clinical or formal terms

such as *approximately* or *insignificant* as being non-understated. As discussed for hyperbole in Section 5.1, these results also show how a machine learning-based approach can help mitigate the false positives introduced by naïve thesaurus-based expansion methods such as those used to create the THedgeWL, as described in Section 4.2.

5.3 Irony

5.3.1 Twitter

Table 8 shows the results of classifiers trained only on co-textual markers. The purpose of this test was to provide a proof-of-concept that computers, like humans, can make use of co-textual irony features to identify ironic intent in Twitter data.

Table 8 Proof-of-concept Results for Co-Textual Irony Markers on Twitter Data

	Precision	Recall	F-Score
Hyperbole Only	0.663323	0.449554	0.535802
Understatement Only	0.686012	0.507104	0.583098
Rhetorical Questions Only	0.570177	0.05017	0.092087
Tag Questions Only	0.492236	0.520927	0.413831
Total Count of Co-textual Markers Only	0.631946	0.658939	0.64504
All Co-textual Markers plus Total Count	0.655151	0.628186	0.641277

These results show that hyperbole alone can provide a better-than-chance indicator of whether a document is ironic or not. Similarly, understatement alone can also provide a better-than-chance indicator of whether a document is ironic or not. Furthermore, Table

9 shows that both hyperbole and understatement serve as positive indicators or irony, in line with this thesis' hypothesis that the presence of co-textual markers signals ironic intent by the author.

Table 9 Proof-of-Concept Irony Marker Significance for Twitter Data

Ironic	Not Ironic
HYPERBOLE *UNDERSTATEMENT* *TAG_QUESTIONS* *TOTAL_MARKER_COUNT*	*RHETORICAL_QUESTIONS*

Conversely, tag questions seem to be no better than chance at predicting irony. This is unexpected since tag questions typically do not appear in written discourse and one may reasonably suspect that the inclusion of tag questions in written discourse would be a deliberate act by the author, presumably to reach a specific stylistic goal. However, such an expectation would implicitly assume a formal writing style and that there is no real-time interaction between the author and the reader. Tweets tend to be written in a more conversational style and users are free to interact with each other. In this respect, Twitter-based discourse seems more akin to spoken discourse than formal written discourse and appears to employ tag questions accordingly. If this analysis is correct then the use of tag questions to denote irony in formal written discourse may in fact be a special case of Burgers et al. (2013)'s "change of register" co-textual irony marker.

Finally, Table 1 shows that the presence of rhetorical question patterns in a document is a stronger indicator of literal intent than ironic intent, against expectation. Examination

of the test data showed that the rhetorical patterns described in Section 4.3 matched only two tweets out of the 20000 tweet test set. As such, no conclusions regarding the use of rhetorical questions in irony detection can be drawn from this experiment.

The following tables show the results for classifiers trained on unigram features (Table 10), bigram features (Table 11), and combined unigram and bigram features (Table 12). Statistical significance was computed using a one-tailed t-test on paired data, using the results of the corresponding n-gram only classifier as a baseline.

Table 10 Unigram Sarcasm Detection Results for Twitter Data. Bold values represent the highest result achieved. *Italicized values* represent $p \leq 0.05$ that mean is greater than

	Precision	Recall	F-Score
Unigrams	0.755327	0.737505	0.746141
w/ Hyperbole	0.757254	0.738838	0.747732
w/ Understatement	0.757528	***0.740711***	*0.7488*
w/ Rhetorical Questions	0.755604	0.737297	0.746165
w/ Tag Questions	0.755818	*0.737597*	0.746431
w/ Total Marker Count	*0.758995*	*0.740329*	***0.749318***
w/ All Markers and Total Marker Count	***0.759582***	*0.738813*	*0.748863*

Table 11 Bigram Sarcasm Detection Results for Twitter Data. Bold values represent the highest result achieved. *Italicized values* represent $p \leq 0.05$ that mean is greater than

	Precision	Recall	F-Score
Bigrams	0.774233	0.725391	0.748876
w/ Hyperbole	0.773591	0.725814	0.748817
w/ Understatement	0.773313	0.727863	0.749788
w/ Rhetorical Questions	0.774465	0.725701	0.749157
w/ Tag Questions	**0.774572**	0.725497	0.749094
w/ Total Marker Count	0.773099	0.726206	0.748799
w/ All Markers and Total Marker Count	0.774177	**0.729195**	**0.750906**

Table 12 Unigram and Bigram Combined Sarcasm Detection Results for Twitter Data. Bold values represent the highest result achieved. *Italicized values* represent $p \leq 0.05$ that mean is greater than

	Precision	Recall	F-Score
Unigrams + Bigrams	0.787215	0.761914	0.774193
w/ Hyperbole	0.784953	0.760437	0.772322
w/ Understatement	**0.787705**	***0.764552***	***0.775798***
w/ Rhetorical Questions	0.786786	0.761513	0.773786
w/ Tag Questions	0.787277	0.761923	0.774244
w/ Total Marker Count	0.786922	0.762631	0.774436
w/ All Markers and Total Marker Count	0.786148	0.76125	0.773301

Table 13 Most Informative Irony Features for Twitter Data

Ironic	Not Ironic
(POS TAG) CD_WRB	*(POS TAG) NNP_VBG*
far_from	*october*
at_once	*happy_birthday*
#school	*,_here*
online_.	*dis*
#bestfriends	*driving_me*
❤_URL	*be_very*
not_like	*not_gon*
just_love	*a_hypocrite*
clearly	*._going*
woohoo	*today_i*
classy	*tub*
#funny	*defense_.*
the_twins	*slept*
problem_,	*pregnant*
baseball	*._love*
buzzing	*you_thank*
friday_night	*this_great*
#stupidity	*your_parents*
spectacular	*oh_well*

Although the proof-of-concept suggest co-textual markers are useful in detecting irony, they resulted in only small, although still statistically significant, real-world performance gains over the baseline system. The proof-of-concept results prove that hyperbole and understatement are more likely to occur in ironic tweets than non-ironic

ones. As such, it stands to reason that the baseline irony classifier must be covertly learning hyperbole and understatement patterns. This can be seen in the Most Informative Features shown in Table 13 which captures such seeming hyperbolic patterns as *just_love* and *spectacular*. Such redundancy may help explain the small real-world performance gains.

However, despite their small magnitude, the statistical significance of these gains provide further evidence that co-textual markers are useful in the automatic detection of irony and suggest that a divide-and-conquer approach to irony detection where individual co-textual markers are identified by using highly specialized classifiers may still capture specific patterns that an all-in-one classifier may miss. This should not be surprising as co-textual markers like hyperbole and understatement may have conflicting features which would be difficult to model using a single, generalized classifier.

Finally, it is important to examine how well humans perform at this task. González-Ibáñez et al. (2011) found that humans were only able to achieve an accuracy of 63% at identifying irony in tweets. Like the experiments described in this thesis, tweets containing *#sarcasm* hashtags were taken to be true examples of sarcasm. Also like the experiments described in this thesis, these author-supplied irony annotations were removed prior to presenting the tweets to the participants. Riloff et al. (2013) performs and almost identical experiment and found that their human participants had a recall of only 45%. González-Ibáñez et al. (2011) and Riloff et al. (2013) use different datasets than the ones used in this study and thus the results are not directly comparable. However, it is interesting to note that irony detection is a difficult task, even for humans.

Davidov et al. (2010) suggests that may be because *#sarcasm* hashtags are biased towards the most difficult examples when authors fear their irony would otherwise go undetected.

5.3.2 Amazon Product Reviews

Similar to Table 8, the results shown in Table 14 represent a proof-of-concept experiment attempting to show that sentence-level co-textual irony markers can use used by detect document-level ironic intent.

Table 14 Co-textual Marker Proof-of-Concept Results for Amazon Data

	Precision	**Recall**	**F-Score**
Hyperbole Only	0.023003	0.006796	0.010493
Understatement Only	0.049841	0.008668	0.014768
Sarcasm	0	0	0
Rhetorical Questions Only	0	0	0
Tag Questions Only	0.569179	0.078266	0.135593
Total Count of Co-textual Markers Only	0.036172	0.064593	0.046375
All Co-textual Markers plus Total Count	0.571079	0.150123	0.186555

Unlike the proof-of-concept results seen for Twitter data in Section 5.3.1, co-textual markers alone seem useless for identifying ironic reviews. However, the real-world results show a different story. The following tables show the results for classifiers trained on unigram features (Table 15), bigram features (Table 16), and combined unigram and bigram features (Table 17). As before, statistical significance was

computed using a one-tailed t-test on paired data, using the results of the corresponding n-gram only classifier as a baseline.

Table 15 Unigram Co-textual Marker Results for Amazon Data. Bold values represent the highest result achieved. *Italicized values* represent $p \leq 0.05$ that mean is greater than

	Precision	**Recall**	**F-Score**
Unigrams	0.685323	0.65087	0.659335
w/ Hyperbole	0.694108	0.65727	0.665775
w/ Understatement	*0.692661*	0.646349	*0.659611*
w/ Sarcasm	0.687282	0.655814	0.661426
w/ Rhetorical Questions	0.688692	0.653302	0.66091
w/ Tag Questions	0.693763	0.655661	*0.664535*
w/ Total Marker Count	0.682622	0.638749	0.652095
w/ All Markers and Total Marker Count	**0.697794**	***0.66722***	***0.673338***

Table 16 Bigram Co-textual Marker Results for Amazon Data. Bold values represent the highest result achieved. *Italicized values* represent $p \leq 0.05$ that mean is greater than

	Precision	**Recall**	**F-Score**
Bigrams	0.660086	0.535333	0.585346
w/ Hyperbole	0.651406	0.527463	0.576111
w/ Understatement	0.656411	0.530875	0.58121
w/ Sarcasm	0.658256	0.537808	0.58594
w/ Rhetorical Questions	0.654466	0.530875	0.579982
w/ Tag Questions	0.671317	0.55122	0.601389
w/ Total Marker Count	0.663614	0.533809	0.584319
w/ All Markers and Total Marker Count	**0.673687**	***0.579089***	***0.618499***

Table 17 Combined Unigram and Bigram Co-textual Marker Results for Amazon Data. Bold values represent the highest result achieved. *Italicized values* represent $p \leq 0.05$ that mean is greater than

	Precision	**Recall**	**F-Score**
Unigram + Bigram	0.695864	0.640436	0.66033
w/ Hyperbole	0.685844	0.640646	0.656143
w/ Understatement	0.695776	0.640436	0.66032
w/ Sarcasm	0.693689	0.629727	0.653399
w/ Rhetorical Questions	0.689342	0.638568	0.656617
w/ Tag Questions	0.69696	0.638574	0.660605
w/ Total Marker Count	0.690478	0.641195	0.657504
w/ All Markers and Total Marker Count	**0.699898**	**0.652922**	**0.668677**

While the Twitter-based experiments showed strong proof-of-concept results but small real-world performance gains, the experiments using Amazon product reviews seem to show the opposite story with very poor proof-of-concept results but comparatively larger real-world performance gains. This may be partially due to the fact that the hyperbole and understatement classifiers were trained using Twitter data, which is a very different style of writing than Amazon reviews (Filatova, 2012). This may be resulting in an increased coverage where the hyperbole and understatement classifiers identify patterns which are common in Twitter data that may not be easily identified by a classifier trained on Amazon product reviews, mitigating the partial redundancy observed in the Twitter-based experiments.

Another factor is the long form nature of Amazon reviews. Since authors are not limited to 140 characters, they are free to supply greater amounts of context, including these co-textual markers. Tweets are normally only a single sentence long and thus co-textual

markers such as hyperbole and understatement need to be part of the ironic utterance itself. Amazon reviews, on the other hand, allow authors to use these co-textual markers to signal irony well in advance of the actual ironic utterance.

Table 18 Most Informative Irony Features for Amazon Data

Ironic	Not Ironic
using	*?*
(POS TAG) CC_IN	*nothing*
best	*no*
(POS TAG) RB_IN	*by*
books	*(POS TAG) NN_MD*
feel	*toothbrush*
.	*10*
read	*could*
fun	*(POS TAG) RB_PRP*
(POS TAG) IN_,	*(POS TAG) RB_PRP$*
(POS TAG) :_NN	*thing*
(POS TAG) VBG_.	*(POS TAG) CD_NNS*
item	*(POS TAG) CC_JJR*
very	*people*
loved	*dick*
itchy	*(POS TAG) VBD_NNS*
works	*pay*
nice	*(POS TAG) NN_PRP$*
(POS TAG) VBN_IN	*off*
(POS TAG) NN_CC	*if_you*

As with the Twitter-based experiment in Section 5.3.1, there appears to be some overlap between the n-gram features and the individual co-textual markers. The Most Informative Features found in Table 18 show stereotypically hyperbolic patterns such as *best* and *loved*.

Table 19 Proof-of-Concept Irony Marker Significance for Amazon Data

Ironic	Not Ironic
UNDERSTATEMENT	*RHETORICAL_QUESTIONS*
SENT_SARCASM	*TOTAL_MARKER_COUNT*
HYPERBOLE	
TAG_QUESTIONS	

Table 19 shows that, similar to Twitter data, the presence of sentence-level hyperbole or understatement is indicative of document-level irony. Additionally, it shows that sentence-level irony also indicates document-level irony, in line with expectations. As with Twitter data, the presence of the rhetorical question patterns discussed in Section 4.3 indicates that a document is not ironic, against expectation. Again, this due to the low hit rate of these patterns and as such no conclusions can be drawn about the effectiveness of rhetorical questions in general in the automatic detection of irony.

What is more difficult to explain is why the total marker count was classified as non-ironic while the individual markers were classified as ironic. It is important to know that the total marker count is just that, a simple sum of all of the co-textual marker values. Since hyperbole, understatement, and sarcasm are all binary values, each feature has a maximum value of 1. By contrast, a sentence can have multiple tag questions or

rhetorical formations. As such, tag questions and rhetorical questions may be artificially inflating the total marker count. Another difficulty is, as described above and in Section 1, irony/sarcasm have a large overlap with hyperbole and understatement. This means that utterances which express sarcasm through the use of hyperbole or understatement are more heavily weighted in the total marker count than utterances which express sarcasm without the use of these co-textual markers.

This also hints at a further problem: what types of ganging effects exist between co-textual irony markers? Kreuz and Roberts (1995) showed that hyperbole and veridicality have an additive effect on humans' perceptions of irony while Kreuz et al. (1999) showed that tag questions and the amount "common ground" shared between the speaker and listener did not have any additive effect. While the results of the total marker count experiments and all marker experiments on Twitter data shown in Table 8 seems to suggest some kind of additive or ganging effect does exist between hyperbole and understatement, such an effect cannot be seen in the Amazon experiments shown in Table 14. Thus, further research into the interactions between co-textual markers in irony detection is needed.

Another interesting aspect of these results is that they suggest understatement may be more useful than hyperbole for irony detection. This seems more pronounced with Amazon data than with Twitter data. This may be because tweets and Amazon reviews both create contexts where authors are more likely to invoke hyperbole in all utterances, not only ironic ones. Writing a legitimate tweet or review is time consuming. Authors may only make the effort if they feel very strongly about the topic/product. This is

exemplified by the Amazon review in (30), which was marked as legitimate by Filatova (2012).

(30) This book is **simply amazing**. It is the first one that I read from Oscar Wilde and I have to admit I was **amazed by how good** he was in choosing the right words to form sentences. Dorian Gray is an **amazing** character and I think that it **appeals to everyone** in a certain way. It is a **must-read**

This may also explain why this trend is more pronounced in Amazon data. Compared to a tweet, which has a maximum length of 140 characters, writing an Amazon product review requires a much greater amount of effort, if only due to its longer length. Additionally, while Twitter can be used to discuss any topic a user wishes to discuss, the topic of an Amazon review is limited to a specific product or category of products. Furthermore, the Twitter user experience is designed around encouraging users to tweet; from the "Compose new tweet" box found on Twitter's main page[14] to official apps for mobile devices. Compare this to Amazon which requires users to navigate to the appropriate product page and then scroll almost all the way to the bottom of the page when they wish to write a review. Such barriers to entry may further skew the active user base towards even more polarized opinions. The very fact that 605 out of 817 legitimate Amazon reviews in Filatova (2012) have a rating of five stars further supports this hypothesis.

[14] For logged-in users only

6 Conclusions and Future Work

This thesis has documented the creation of generalized detection methods for both hyperbole and understatement. To the author's best knowledge, this is the first attempt at such systems. The results of Sections 5.1 and 5.2 show that the n-gram based classification method introduced in this thesis outperform word list based approaches with the best results being observed in both cases using a combination of unigrams and bigrams.

As discussed in Sections 3.1, 5.1, and 5.2, Twitter data tends to be noisy with a high rate of false negatives and, especially in the case of understatement, a large potential for false positives. While it would require much more effort, the creation of manually annotated hyperbole and understatement corpora may further improve the reliability of these automatic detection techniques. It should be noted that despite these issues with the quality of the training data, the most informative features shown in Table 3 and Table 6 still show the n-gram based detection method was able to extract meaningful patterns.

This thesis also evaluated the effectiveness of including co-textual irony markers in the automatic detection of irony for both short documents, such as Twitter messages, and medium length documents, such as Amazon product reviews. The results of Table 8 indicate that the simply presence of hyperbole or understatement offers a better-than-chance indication that a document is ironic while the results discussed in Sections 5.3 suggest that such features can indeed lead to real-world performance advantages. Furthermore, the use of automated hyperbole and understatement detection methods

show that this approach to irony detection is indeed feasible for real-world applications as figurative languages do not need to be explicitly marked.

While these initial results are encouraging, the lack of research into how humans detect and interpret the individual co-textual irony markers (both inside and outside of ironic contexts) presents a major obstacle not only to developing improved detection techniques for the individual co-textual markers but also for irony detection as a whole. The discovery of any unique features helpful in detecting specific co-textual markers would be expected to increase irony detection performance not only by increasing the accuracy of the individual co-textual marker classifier but also by helping to mitigate the redundancies of n-gram patterns between co-textual marker classifiers and the baseline irony classifier described in Section 5.3.1.

More accurate detection of co-textual markers would be expected to result in more accurate detection of irony by allowing automatic detection methods to more readily identify ironic contexts. As such, it may be worthwhile to explore more advanced co-textual marker detection systems, such as employing the pattern extraction techniques of Tsur et al. (2010) and Davidov et al. (2010). Psycholinguistic features, such as LWIC's word categories, have shown promise in the automatic classification of irony (González-Ibáñez et al., 2011). Given the use of hyperbole and understatement in signalling irony (as described in Section 1.2), this approach may also prove fruitful for hyperbole and understatement. Finally, combining the irony markers described in this thesis with the figurative language features of Reyes and Rosso (2011), Reyes et al. (2012), and Pérez

(2012) would offer an even more complete view of the use of co-textual features to detect irony.

7 References

Berntsen, Dorthe, and John M. Kennedy. "Unresolved contradictions specifying attitudes—in metaphor, irony, understatement and tautology." *Poetics* 24.1 (1996): 13-29.

Burgers, Christian, Margot van Mulken, and Peter Jan Schellens. "The use of co-textual irony markers in written discourse." *Humor* 26.1 (2013): 45-68.

Cano Mora, Laura. "All or nothing: a semantic analysis of hyperbole." (2009).

Cano Mora, Laura. "" How to make a mountain out of a molehill": a corpus-based pragmatic and conversational analysis study of hyperbole in interation." (2006).

Carvalho, Paula, et al. "Clues for detecting irony in user-generated contents: oh...!! it's so easy;-)." Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion. ACM, 2009.

Colston, Herbert L. "" I've Never Seen Anything Like It": Overstatement, Understatement, and Irony." *Metaphor and symbol* 12.1 (1997): 43-58.

Davidov, Dmitry, and Ari Rappoport. "Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words." *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006.

Davidov, Dmitry, Oren Tsur, and Ari Rappoport. "Semi-supervised recognition of sarcastic sentences in twitter and amazon." *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2010.

Developer Rules of the Road. (2013, July 2). Retrieved January 18, 2014, from Twitter, http://dev.twitter.com/terms/api-terms

Farkas, Richárd, et al. "The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text." *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*. Association for Computational Linguistics, 2010.

Filatova, Elena. "Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing." *LREC*. 2012.

Frank, Jane. "You call that a rhetorical question?: Forms and functions of rhetorical questions in conversation." *Journal of Pragmatics* 14.5 (1990): 723-738.

Ganter, Viola, and Michael Strube. "Finding hedges by chasing weasels: Hedge detection using Wikipedia tags and shallow linguistic features." *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, 2009.

Gedigian, Matt, et al. "Catching metaphors." *Proceedings of the Third Workshop on Scalable Natural Language Understanding*. Association for Computational Linguistics, 2006.

Go, Alec, and Richa Bhayani. *Exploiting the unique characteristics of tweets for sentiment analysis*. Technical report, Technical Report, Stanford University, 2010.

González-Ibáñez, Roberto, Smaranda Muresan, and Nina Wacholder. "Identifying Sarcasm in Twitter: A Closer Look." ACL (Short Papers). 2011.

Grice, H. Paul. "Logic and conversation." 1975 (1975): 41-58.

Hübler, Axel. *Understatements and hedges in English*. John Benjamins Publishing, 1983.

Kreuz, Roger J., and Gina M. Caucci. "Lexical influences on the perception of sarcasm." *Proceedings of the Workshop on computational approaches to Figurative Language*. Association for Computational Linguistics, 2007.

Kreuz, Roger J., and Richard M. Roberts. "Two cues for verbal irony: Hyperbole and the ironic tone of voice." *Metaphor and symbol* 10.1 (1995): 21-31.

Kreuz, Roger J., and Sam Glucksberg. "How to be sarcastic: The echoic reminder theory of verbal irony." *Journal of Experimental Psychology: General* 118.4 (1989): 374.

Kreuz, Roger J., et al. "Tag questions and common ground effects in the perception of verbal irony." *Journal of Pragmatics* 31.12 (1999): 1685-1700.

Li, Baichuan, et al. "Question identification on twitter." *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011.

Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREC*. 2010.

Pérez, Antonio Reyes. *Linguistic-based patterns for figurative language processing: the case of humor recognition and irony detection*. Diss. Universitat Politècnica de València, 2012.

Riloff, Ellen, et al. "Sarcasm as Contrast between a Positive Sentiment and Negative Situation." *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013.

Reyes, Antonio, and Paolo Rosso. "Mining subjective knowledge from customer reviews: a specific case of irony detection." *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. Association for Computational Linguistics, 2011.

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

Schmidt-Radefeldt, Jürgen. "On so-called 'rhetorical' questions." *Journal of Pragmatics* 1.4 (1977): 375-392.

Shutova, Ekaterina. "Models of metaphor in NLP." *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010.

Slugoski, Ben R., and William Turnbull. "Cruel to be kind and kind to be cruel: Sarcasm, banter and social relations." *Journal of Language and Social Psychology* 7.2 (1988): 101-121.

Tepperman, Joseph, David R. Traum, and Shrikanth Narayanan. "" yeah right": sarcasm recognition for spoken dialogue systems." *INTERSPEECH*. 2006.

Tsur, Oren, Dmitry Davidov, and Ari Rappoport. "ICWSM-A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews." *ICWSM*. 2010.

Utsumi, Akira. "A unified theory of irony and its computational formalization." *Proceedings of the 16th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1996.

Utsumi, Akira. "How to interpret irony by computer: A comprehensive framework for irony." *Proceedings of RANLP*. 1995.

Vanin, Aline A., et al. "Some clues on irony detection in tweets." *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013.

Veale, Tony, and Yanfen Hao. "Detecting Ironic Intent in Creative Comparisons." *ECAI*. Vol. 215. 2010.

Whalen, Juanita M., Penny M. Pexman, and Alastair J. Gill. "“Should Be Fun—Not!” Incidence and Marking of Nonliteral Language in E-Mail." *Journal of Language and Social Psychology* 28.3 (2009): 263-280.

Whalen, Juanita M., et al. "Verbal irony use in personal blogs." *Behaviour & Information Technology* 32.6 (2013): 560-569.

Wu, Jean, and Justine T. Kao. “A Hyperbole is Worth a Thousand Words: A Bayesian Model of Hyperbole Interpretation.” 2012.

Appendix 1 Hyperbole Word List

absolute
absolutely
age
ages
all
always
amazed
amazing
any
anyone
anything
asleep
astonish
astonishing
awful
beautiful
beyond
big
blasted
blasting
blew
blown
brilliant
centuries
century
chaos
completely
crack
crazy
day
days
dead
deadly
definitely
desperate
desperately
devastated
dinosaur
dinosaurs
disaster
disgrace
drained
drugs
enormous
entirely
erupted
every
everybody
everything
everywhere
evil
exact
exactly
excellent
extraordinary
extreme
extremely
feet
flea
foot
forever
frantic
frantically
freezing
frozen
full

fully
gorgeous
great
great
half
haywire
headache
heap
heaven
hell
horrible
horse
hour
hours
huge
hundred
hundreds
ideal
illegible
immense
immensely
impressive
inch
inches
incredible
incredibly
infinitely
insane
instantly
irresistible
killed
killing
lifelong
lifetime
limbo
literally
little
load
loads
lots
lovely
mammoth
massive
mental
mess
mile
miles
millenium
million
millions
minuscule
minute
minutes
month
months
most
mushroom
mushrooming
never
no
nobody
not
nothing
obnoxious
pain
paradise
pathetic
pile
precious

pure

relentless

remotely

revived

reviving

right

rolling

ruin

ruined

scrap

scream

season

second

seconds

sheer

shock

shocked

shocking

sickening

small

smashing

splendid

squeal

starve

starving

terrible

terribly

thousand

thousands

thrilled

thrilling

times

tiny

total

totally

tremendous

unbelievable

utmost

vast

vital

week

weekend

weeks

whole

wicked

wonderful

world

worst

year

year

Appendix 2 Hedge Word List

about

alleged

allegedly

apparent

apparently

appear

appears

arguably

argue

beleive

beleived

bit

consider

considered

couple

experts

fancy

feel

few

guess

insignificant

just

kind

kinda

little

many

often

people

perhaps

probably

quite

reportedly

seem

seemingly

seems

smidge

smidgen

some

somewhat

sort

suppose

tad

think

tiny

touch

국문초록

비유언어와 문맥 표지를 이용한 반어법 자동 분류 연구

Andrew Cattle

언어학과 언어학전공

서울대학교 대학원

본 논문은 고빈도 비유언어(figurative language)를 이용한 반어법 자동 인식 방법을 제안한다. 반어법과 비유언어들(직유법, 은유법, 의인법, 과장법)을 인식하는 문제는 컴퓨터 언어학에서 매우 중요한 분야이다. 이런 비유 언어들은 표면적인 의미와 다른 의미를 내포하기 때문에 그 문장의 의미를 파악하는데 필요한 연구이다. 과장법이나 과소 법 같은 비유언어와 달리 특별히 반어법은 그 표현적 의미와 정 반대 또는 부합하지 않는 의미를 내포하기 때문에 더욱 문제가 된다.

구어에서 반어법이 사용될 때는 운율이라는 요소가 인식에 중요한 역할을 하는 반면, 문어에서 반어법은 운율 정보가 없기 때문에 더 인식이 어렵다. 또한,

반어법은 대부분의 경우 표면적으로 나타나는 명확한 단서를 포함하지 않고, 단지 준언어적, 문맥적 화용적인 단서만을 갖기 때문에 인식에 더 어려움이 크다. 반어법의 단서가 되는 예로는 청자에게 문자 그대로 이해되기를 바라지 않음을 암시하는 과장법, 과소법, 수사적 질문법, 부가 의문문 같은 것들이 존재한다.

본고는 동시에 나타나는 비유언어들을 각각 인식하여 그 결과를 반어법 검출기에 제공하는 방식의 분할-정복법을 소개한다. 짧은 길이의 트위터와 상대적으로 긴 아마존 상품평에 대해 실행한 실험은 이러한 비유언어들을 개별적으로 인식하여 반어법의 자동 인식에 사용하는 것이 비유언어들을 한번에 인식하는 방법 보다 반어법 인식에 효과적이라는 사실을 밝혔다.

또한, 지금까지 개별적으로 제한된 문맥만을 고려한 과장법, 과소법 연구와 달리 본 연구는 반어법 인식에 사용되는 기존의 연구 방법을 과장법과 과소법 인식에도 적용할 수 있는 가능성을 제시하였다는 의의가 있다.

Keywords: Figurative Language, Irony, Sarcasm, Hyperbole, Understatement
Student number: 2011-24258