Computational Intelligence, Volume 22, Number 2, 2006

LEARNING TO LAUGH (AUTOMATICALLY): COMPUTATIONAL MODELS FOR HUMOR RECOGNITION

RADA MIHALCEA

Department of Computer Science, University of North Texas, Denton, TX 76203

CARLO STRAPPARAVA

ITC – irst, Istituto per la Ricerca Scientifica e Tecnologica, I-38050, Povo, Trento, Italy

Humor is one of the most interesting and puzzling aspects of human behavior. Despite the attention it has received in fields such as philosophy, linguistics, and psychology, there have been only few attempts to create computational models for humor recognition or generation. In this article, we bring empirical evidence that computational approaches can be successfully applied to the task of humor recognition. Through experiments performed on very large data sets, we show that automatic classification techniques can be effectively used to distinguish between humorous and non-humorous texts, with significant improvements observed over a priori known baselines.

Key words: computational humor, humor recognition, sentiment analysis, one-liners.

1. INTRODUCTION

. . . pleasure has probably been the main goal all along. But I hesitate to admit it, because computer scientists want to maintain their image as hard-working individuals who deserve high salaries. Sooner or later society will realize that certain kinds of hard work are in fact admirable even though they are more fun than just about anything else. (Knuth 1993)

Humor is an essential element in personal communication. While it is merely considered a way to induce amusement, humor also has a positive effect on the mental state of those using it and has the ability to improve their activity. Therefore computational humor deserves particular attention, as it has the potential of changing computers into a creative and motivational tool for human activity (Stock, Strapparava, and Nijholt 2002; Nijholt et al. 2003).

Previous work in computational humor has focused mainly on the task of humor generation (Binsted and Ritchie 1997; Stock and Strapparava 2003), and very few attempts have been made to develop systems for automatic humor recognition (Taylor and Mazlack 2004; Mihalcea and Strapparava 2005). This is not surprising, since, from a computational perspective, humor recognition appears to be significantly more subtle and difficult than humor generation.

In this article, we explore the applicability of computational approaches to the recognition of verbally expressed humor. In particular, we investigate whether automatic classification techniques represent a viable approach to distinguish between humorous and non-humorous text, and we bring empirical evidence in support of this hypothesis through experiments performed on very large data sets.

Because a deep comprehension of humor in all of its aspects is probably too ambitious and beyond the existing computational capabilities, we chose to restrict our investigation to the type of humor found in *one-liners*. A one-liner is a short sentence with comic effects and an interesting linguistic structure: simple syntax, deliberate use of rhetoric devices (e.g., alliteration, rhyme), and frequent use of creative language constructions meant to attract the readers' attention. While longer jokes can have a relatively complex narrative structure, a one-liner must produce the humorous effect "in one shot," with very few words. These characteristics make this type of humor particularly suitable for use in an automatic learning

setting, as the humor-producing features are guaranteed to be present in the first (and only) sentence.

We attempt to formulate the humor-recognition problem as a traditional classification task, and feed positive (humorous) and negative (non-humorous) examples to an automatic classifier. The humorous data set consists of one-liners collected from the Web using an automatic bootstrapping process. The non-humorous examples are selected such that they are structurally and stylistically similar to the one-liners. Specifically, we use four different negative data sets: (1) Reuters news titles; (2) proverbs; (3) sentences from the British National Corpus (BNC); (4) commonsense statements from the Open Mind Common Sense (OMCS) corpus. The classification results are encouraging, with accuracy figures ranging from 79.15% (One-liners/BNC) to 96.95% (One-liners/Reuters). Regardless of the non-humorous data set playing the role of negative examples, the performance of the automatically learned humor-recognizer is always significantly better than a priori known baselines.

The experimental results prove that computational approaches can be successfully used for the task of humor recognition. An analysis of the results shows that the humorous effect can be identified in a large fraction of the jokes in our data set using surface features such as alliteration, word-based antonymy, or specific vocabulary. Moreover, we also identify cases where our current automatic methods fail, which require more sophisticated techniques such as recognition of irony, detection of incongruity that goes beyond word antonymy, or commonsense knowledge. Finally, an analysis of the most discriminative content-based features identified during the process of automatic classification helps us point out some of the most predominant semantic classes specific to humorous text, which could be turned into useful features for future studies of humor generation.

The remainder of the article is organized as follows. We first describe the humorous and non-humorous data sets. We then show experimental results obtained on these data sets using several heuristics and two different text classifiers. Finally, we conclude with a discussion, a detailed analysis of the results, and directions for future work.

2. HUMOROUS AND NON-HUMOROUS DATA SETS

To test our hypothesis that automatic classification techniques represent a viable approach to humor recognition, we needed in the first place a data set consisting of humorous (positive) and non-humorous (negative) examples. Such data sets can be used to automatically *learn* computational models for humor recognition, and at the same time *evaluate* the performance of such models.

Humorous data: While there is plenty of non-humorous data that can play the role of negative examples, it is significantly harder to build a very large and at the same time sufficiently "clean" data set of humorous examples. We use a dually constrained Web-based bootstrapping process to collect a very large set of one-liners. Starting with a short *seed* set consisting of a few one-liners manually identified, the algorithm automatically identifies a list of Web pages that include at least one of the seed one-liners, via a simple search performed with a Web search engine. Next, the Web pages found in this way are HTML parsed, and additional one-liners are automatically identified and added to the seed set. The process is repeated several times, until enough one-liners are collected. As with any other bootstrapping algorithm, an important aspect is represented by the set of constraints used to steer the process and prevent as much as possible the addition of noisy entries. Our algorithm uses (1) a *thematic* constraint applied to the theme of each webpage, via a list of keywords that have to appear in the URL of the Web page; and (2) a *structural* constraint, exploiting HTML annotations indicating text of similar genre (e.g., lists, adjacent paragraphs, and others).

TABLE 1. Sample Examples of One-Liners, Reuters Titles, BNC Sentences, Proverbs, and OMCS Sentences

One-liners
Take my advice; I don't use it anyway.
I get enough exercise just pushing my luck.
Beauty is in the eye of the beer holder.
Reuters titles
Trocadero expects tripling of revenues.
Silver fixes at two-month high, but gold lags.
Oil prices slip as refiners shop for bargains.
BNC sentences
They were like spirits, and I loved them.
I wonder if there is some contradiction here.
The train arrives three minutes early.
Proverbs
Creativity is more important than knowledge.
Beauty is in the eye of the beholder.
I believe no tales from an enemy's tongue.
OMCS sentences
Humans generally want to eat at least once a day.
A file is used for keeping documents.
A present is a gift, something you give to someone.

Two iterations of the bootstrapping process, started with a small seed set of 10 one-liners, resulted in a large set of about 24,000 one-liners. After removing the duplicates using a measure of string similarity based on the longest common subsequence, we are left with a final set of 16,000 one-liners, which are used in the humor-recognition experiments. A more detailed description of the Web-based bootstrapping process is available in Mihalcea and Strapparava (2005). The one-liners humor style is illustrated in Table 1, which shows three examples of such one-sentence jokes.

Non-humorous data: To construct the set of negative examples required by the humor-recognition models, we tried to identify collections of sentences that were non-humorous, but similar in structure and composition to the one-liners. We did not want the automatic classifiers to learn to distinguish between humorous and non-humorous examples based simply on text length or obvious vocabulary differences. Instead, we seek to enforce the classifiers to identify humor-specific features, by supplying them with negative examples similar in most of their aspects to the positive examples, but different in their comic effect.

We tested four different sets of negative examples, with three examples from each data set as illustrated in Table 1. All non-humorous examples are enforced to follow the same length restriction as the one-liners, that is, one sentence with an average length of 10–15 words.

1. *Reuters* titles, extracted from news articles published in the Reuters newswire over a period of one year (from August 20, 1996, to August 19, 1997) (Lewis et al. 2004). The titles consist of short sentences with simple syntax, and are often phrased to catch the readers' attention (an effect similar to the one rendered by the one-liners).

2. *Proverbs* extracted from an online proverb collection. Proverbs are sayings that transmit, usually in one short sentence, important facts or experiences that are considered true by many people. Their property of being condensed, but memorable sayings make them very similar to the one-liners. In fact, some one-liners attempt to reproduce proverbs, with a comic effect, as in the example, *"Beauty is in the eye of the beer holder*," derived from *"Beauty is in the eye of the beholder."*
3. *British National Corpus (BNC)* sentences, extracted from BNC—a balanced corpus covering different styles, genres, and domains. The sentences were selected such that they were similar in content with the one-liners: we used an information retrieval system implementing a vectorial model to identify the BNC sentence most similar to each of the 16,000 one-liners.[1] Unlike the Reuters titles or the proverbs, the BNC sentences have typically no added creativity. However, we decided to add this set of negative examples to our experimental setting to observe the level of difficulty of a humor-recognition task when performed with respect to simple text.
4. *Open Mind Common Sense (OMCS)* sentences. OMCS is a collection of about 800,000 commonsense assertions in English as contributed by volunteers over the Web. It consists mostly of simple single sentences, which tend to be explanations and assertions similar to glosses of a dictionary, but phrased in a more common language. For example, the collection includes such assertions as "keys are used to unlock doors," and "pressing a typewriter key makes a letter." Because the comic effect of jokes is often based on statements that break our commonsensical understanding of the world, we believe that such commonsense sentences can make an interesting collection of "negative" examples for humor recognition. For details on the OMCS data and how it has been collected, see Singh (2002). From this repository we use the first 16,000 sentences.[2]

To summarize, the humor-recognition experiments rely on data sets consisting of humorous (positive) and non-humorous (negative) examples. The positive examples consist of 16,000 one-liners automatically collected using a Web-based bootstrapping process. The negative examples are drawn from (1) Reuters titles, (2) proverbs, (3) BNC sentences, and (4) OMCS sentences.

3. AUTOMATIC HUMOR RECOGNITION

We experiment with automatic classification techniques using (a) heuristics based on humor-specific stylistic features (alliteration, antonymy, slang); (b) content-based features, within a learning framework formulated as a typical text classification task; and (c) combined stylistic and content-based features, integrated in a stacked machine learning framework.

3.1. Humor-Specific Stylistic Features

Linguistic theories of humor (Attardo 1994) have suggested many *stylistic features* that characterize humorous texts. We tried to identify a set of features that were both significant and feasible to implement using existing machine-readable resources. Specifically, we focus

[1] The sentence most similar to a one-liner is identified by running the one-liner against an index built for all BNC sentences with a length of 10–15 words. We use a *tf.idf* weighting scheme and a cosine similarity measure, as implemented in the Smart system (ftp.cs.cornell.edu/pub/smart).

[2] The first sentences in this corpus are considered to be "cleaner," as they were contributed by trusted users (Push Singh, personal communication, July, 2005).

on alliteration, antonymy, and adult slang, which were previously suggested as potentially good indicators of humor (Ruch 2002; Bucaria 2004).

Alliteration: Some studies on humor appreciation (Ruch 2002) show that structural and phonetic properties of jokes are at least as important as their content. In fact one-liners often rely on the readers' awareness of attention-catching sounds, through linguistic phenomena such as alliteration, word repetition, and rhyme, which produce a comic effect even if the jokes are not necessarily meant to be read aloud. Note that similar rhetorical devices play an important role in wordplay jokes, and are often used in newspaper headlines and in advertisements. The following one-liners are examples of jokes that include one or more alliteration chains:

Veni, Vidi, Visa: I came, I saw, I did a little shopping.
Infants don't enjoy infancy like adults do adultery.

To extract this feature, we identify and count the number of alliteration/rhyme chains in each example in our data set. The chains are automatically extracted using an index created on top of the CMU pronunciation dictionary.[3]

The underlying algorithm is basically a matching device that tries to find the largest and longest string matching chains using the transcriptions obtained from the pronunciation dictionary. For example, in the second sentence above the algorithm finds two alliteration chains of length two: (*infan*-ts, *infan*-cy) and (*adult*-s, *adult*-ery), exploiting, respectively, the phonetic matchings "`ih1 n f ah0 n`" and "`ah0 d ah1 l t`" found using the pronunciation dictionary. The algorithm avoids matching noninteresting chains such as, for example, series of definite/indefinite articles, by using a stopword list of functional words that cannot be part of an alliteration chain.

Antonymy: Humor often relies on some type of incongruity, opposition, or other forms of apparent contradiction. While an accurate identification of all these properties is probably difficult to accomplish, it is relatively easy to identify the presence of *antonyms* in a sentence. For instance, the comic effect produced by the following one-liners is partly due to the presence of antonyms:

A clean desk is a sign of a cluttered desk drawer.
Always try to be modest and be proud of it!

The lexical resource we use to identify antonyms is WORDNET (Miller 1995), and in particular the *antonymy* relation among nouns, verbs, adjectives, and adverbs. For adjectives we also consider an indirect antonymy via the *similar-to* relation among adjective synsets. Despite the relatively large number of *antonymy* relations defined in WORDNET, its coverage is far from complete, and thus the *antonymy* feature cannot always be identified. A deeper semantic analysis of the text, such as word sense disambiguation or domain disambiguation, could probably help detect other types of semantic opposition, and we plan to exploit these techniques in future work.

Adult slang: Humor based on adult slang is very popular. Therefore, a possible feature for humor recognition is the detection of sexual-oriented lexicon in the sentence. The following represent examples of one-liners that include such slang:

The sex was so good that even the neighbors had a cigarette.
Artificial Insemination: procreation without recreation.

[3] Available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

To form a lexicon required for the identification of this feature, we extract from WORDNET DOMAINS[4] all the synsets labeled with the domain SEXUALITY. The list is further processed by removing all words with high polysemy ($\geq$4). Next, we check for the presence of the words in this lexicon in each sentence in the corpus, and annotate them accordingly. Note that, as in the case of antonymy, WORDNET coverage is not complete, and the *adult slang* feature cannot always be identified.

Finally, in some cases, all three features (alliteration, antonymy, adult slang) are present in the same sentence, as for instance the following one-liner:

Behind every great$_{al}$ man$_{ant}$ is a great$_{al}$ woman$_{ant}$, and
behind every great$_{al}$ woman$_{ant}$ is some guy staring at her behind$_{sl}$!

3.2. Content-Based Learning

In addition to stylistic features, we also experimented with *content-based features*, through experiments where the humor-recognition task is formulated as a traditional text-classification problem. Specifically, we compare results obtained with two frequently used text classifiers, Naïve Bayes and Support Vector Machines (SVM), selected based on their performance in previously reported work, and for their diversity of learning methodologies.

Naïve Bayes: The main idea in a Naïve Bayes text classifier is to estimate the probability of a category given a document using joint probabilities of words and documents. Naïve Bayes classifiers assume word independence; however, despite this simplification, these algorithms were shown to perform well on text classification (Yang and Liu 1999). While there are several versions of Naïve Bayes classifiers (variations of multinomial and multivariate Bernoulli), we use the multinomial model, previously shown to be more effective (McCallum and Nigam 1998).

Support Vector Machines: SVM are binary classifiers that seek to find the hyperplane that best separates a set of positive examples from a set of negative examples, with maximum margin (Vapnik 1995). Applications of SVM classifiers to text categorization led to some of the best results reported in the literature (Joachims 1998).

4. EXPERIMENTAL RESULTS

The goal of the studies reported in this article is to find out to what extent automatic classification techniques can be successfully applied to the task of humor recognition. Several experiments were conducted to gain insights into various aspects related to an automatic humor-recognition task: classification accuracy using stylistic and content-based features, learning rates, impact of the type of negative data, impact of the classification methodology.

All evaluations are performed using stratified 10-fold cross-validations, for accurate estimates. The baseline for all the experiments is 50%, which represents the classification accuracy obtained if a label of "humorous" (or "non-humorous") would be assigned by default to all the examples in the data set. Experiments with uneven class distributions were also performed, and are reported in Section 5.

[4] WORDNET DOMAINS assigns each synset in WORDNET with one or more "domain" labels, such as SPORT, MEDICINE, ECONOMY. See http://wndomains.itc.it.

TABLE 2. Humor-Recognition Accuracy Using Alliteration, Antonymy, and Adult Slang

	One-Liners			
Heuristic	Reuters	BNC	Proverbs	OMCS
Alliteration	74.31%	59.34%	53.30%	55.57%
Antonymy	55.65%	51.40%	50.51%	51.84%
Adult slang	52.74%	52.39%	50.74%	51.34%
All	76.73%	60.63%	53.71%	56.16%

TABLE 3. Number of Examples in Each Data Set with a Feature Value Different from 0 (Out of the Total of 16,000 Examples), for Alliteration, Antonymy, and Adult Slang

	Data Set				
Heuristic	One-Liners	Reuters	BNC	Proverbs	OMCS
Alliteration	8,323	555	4,675	6,889	6,539
Antonymy	2,124	319	1,164	1,960	1,535
Adult slang	1,074	177	604	828	645

4.1. Heuristics Using Humor-Specific Features

In a first set of experiments, we evaluated the classification accuracy using stylistic humor-specific features: alliteration, antonymy, and adult slang. These are numerical features that act as heuristics, and the only parameter required for their application is a threshold indicating the minimum value admitted for a statement to be classified as humorous (or non-humorous). These thresholds are learned automatically using a decision tree applied on a small subset of humorous/non-humorous examples (1,000 examples). The evaluation is performed on the remaining 15,000 examples, with results shown in Table 2.[5] We also show, in Table 3, the number of examples in each data set that have a feature value different from 0, for each of the three humor-specific features. A sample decision tree learned for the One-liners/BNC data set is shown in Figure 1, and classification results obtained on all data sets are listed in Table 2.

Considering the fact that these features represent *stylistic* indicators, the style of Reuters titles turns out to be the most different with respect to one-liners, while the style of proverbs is the most similar. Note that for all data sets the alliteration feature appears to be the most useful indicator of humor, which is in agreement with previous linguistic findings.

4.2. Text Classification with Content Features

The second set of experiments was concerned with the evaluation of content-based features for humor recognition. Table 4 shows results obtained using the four different sets

[5] We also experimented with decision trees learned from a larger number of examples, but the results were similar, which confirms our hypothesis that these features are heuristics, rather than learnable, properties that improve their accuracy with additional training data.

```
alliteration = 0
|   adult slang = 0
|   |   antonymy <= 1 : no
|   |   antonymy > 1 : yes
|   adult slang > 0 : yes
alliteration > 0 : yes
```

FIGURE 1. Sample decision tree for the application of the three heuristics for humor recognition.

TABLE 4. Humor-Recognition Accuracy Using Naïve Bayes and SVM Text Classifiers

	One-Liners			
Classifier	Reuters	BNC	Proverbs	OMCS
Naïve Bayes	96.67%	73.22%	84.81%	82.39%
SVM	96.09%	77.51%	84.48%	81.86%

of negative examples, with the Naïve Bayes and SVM text classifiers. Learning curves are plotted in Figure 2.

Once again, the content of Reuters titles appears to be the most different with respect to one-liners, while the BNC sentences represent the most similar data set. This suggests that joke content tends to be very similar to regular text, although a reasonably accurate distinction can still be made using text-classification techniques. Interestingly, proverbs can be distinguished from one-liners using content-based features, which indicates that despite their stylistic similarity (see Table 2), proverbs and one-liners deal with different topics.

4.3. Combining Stylistic and Content Features

Encouraged by the results obtained in the first two experiments, we designed a third experiment that attempts to jointly exploit stylistic and content features for humor recognition. The feature combination is performed using a stacked learner, which takes the output of the text classifier, joins it with the three humor-specific features (alliteration, antonymy, adult slang), and feeds the newly created feature vectors to a machine learning tool. Given the relatively large gap between the performance achieved with content-based features (text classification) and stylistic features (humor-specific heuristics), we decided to implement the second learning stage in the stacked learner using a memory-based learning system, so that low-performance features are not eliminated in the favor of the more accurate ones.[6] We use the Timbl memory-based learner (Daelemans et al. 2001), and evaluate the classification using a stratified 10-fold cross-validation. Table 5 shows the results obtained in this experiment, for the four different data sets.

Combining classifiers results in a statistically significant improvement ($p < 0.0005$, paired t-test) with respect to the best individual classifier for the One-liners/Reuters and One-liners/BNC data sets, with relative error rate reductions of 8.9% and 7.3%, respectively.

[6]Using a decision tree learner in a similar stacked learning experiment resulted into a flat tree that takes a classification decision based exclusively on the content feature, ignoring completely the remaining stylistic features.

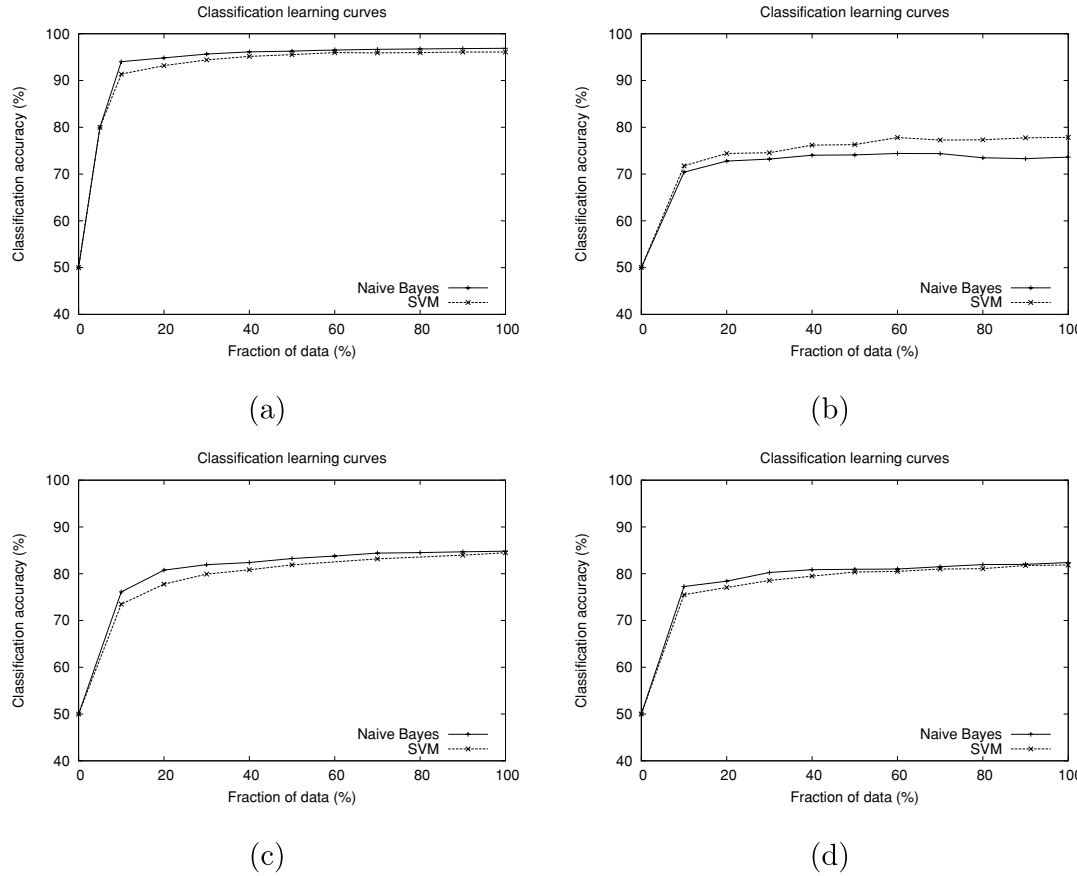

FIGURE 2. Learning curves for humor recognition using text-classification techniques, with respect to four different sets of negative examples: (a) Reuters, (b) BNC, (c) Proverbs, (d) OMCS.

TABLE 5. Humor-Recognition Accuracy for Combined Learning Based on Stylistic and Content Features

One-Liners			
Reuters	BNC	Proverbs	OMCS
96.95%	79.15%	84.82%	82.37%

No improvement is observed for the One-liners/Proverbs and One-liners/OMCS data sets, which is not surprising since, as shown in Table 2, proverbs and commonsense statements cannot be clearly differentiated from the one-liners using stylistic features, and thus the addition of these features to content-based features is not likely to result in an improvement.

5. DISCUSSION

The results obtained in the automatic classification experiments prove that computational approaches represent a viable solution for the task of humor recognition, and good

performance can be achieved using classification techniques based on stylistic and content features.

Despite our initial intuition that one-liners are most similar to other creative texts (e.g., Reuters titles, or the sometimes almost identical proverbs), and thus the learning task would be more difficult in relation to these data sets, comparative experimental results show that in fact it is more difficult to distinguish humor with respect to regular text (e.g., BNC sentences). Note, however, that even in this case the combined classifier leads to a classification accuracy that improves significantly over the a priori known baseline.

In addition to the four negative data sets, we also performed an experiment using a corpus of arbitrary sentences randomly drawn from all the negative sets. The humor recognition with respect to this negative mixed data set resulted in 63.76% accuracy for stylistic features, 77.82% for content-based features using Naïve Bayes, and 79.23% using SVM. These figures are comparable to those reported in Tables 2 and 4 for One-liners/BNC, which suggests that the experimental results reported in the previous sections do not reflect a bias introduced by the negative data sets, since similar results are obtained when the humor recognition is performed with respect to arbitrary negative examples.

As indicated in Section 2, the negative examples were selected structurally and stylistically similar to the one-liners, making the humor-recognition task more difficult than in a real setting. Nonetheless, we also performed a set of experiments where we made the task even harder, using uneven class distributions. For each of the four types of negative examples, we constructed a data set using 75% non-humorous examples and 25% humorous examples. Although the baseline in this case is higher (75%), the automatic classification techniques for humor recognition still improve over this baseline. The stylistic features lead to a classification accuracy of 87.49% (One-liners/Reuters), 77.62% (One-liners/BNC), 76.20% (One-liners/Proverbs), and 77.12% (One-liners/OMCS), and the content-based features used in a Naïve Bayes classifier result in accuracy figures of 96.19% (One-liners/Reuters), 81.56% (One-liners/BNC), 87.86% (One-liners/Proverbs), and 88.91% (One-liners/OMCS).

Finally, in addition to classification accuracy, we were also interested in the variation of classification performance with respect to data size, which is an aspect particularly relevant for directing future research. Depending on the shape of the learning curves, one could decide to concentrate future work either on the acquisition of larger data sets, or toward the identification of more sophisticated features. Figure 2 shows that regardless of the type of negative data, there is significant learning only until about 60% of the data (i.e., about 10,000 positive examples, and the same number of negative examples). The rather steep ascent of the curve, especially in the first part of the learning, suggests that humorous and non-humorous texts represent well-distinguishable types of data. An interesting effect can be noticed toward the end of the learning, where for both classifiers the curve becomes completely flat (One-liners/Reuters, One-liners/Proverbs, One-liners/OMCS), or it even has a slight drop (One-liners/BNC). This is probably due to the presence of noise in the data set, which starts to become visible for very large data sets.[7] This plateau is also suggesting that more data are not likely to help improve the quality of an automatic humor-recognizer, and more sophisticated features are probably required.

5.1. Word Similarities in Different Semantic Spaces

As an additional assessment of the differences/similarities between our data sets, we conducted an analysis of the semantic spaces that can be derived for certain themes, starting

[7] We also like to think of this behavior as if the computer is losing its sense of humor after an overwhelming number of jokes, in a way similar to humans when they get bored and stop appreciating humor after hearing too many jokes.

TABLE 6. Words Semantically Related to "Beauty" Mined from the Different Data Sets

Data Set	Semantic Space for *"Beauty"*
One-liners	blinded squint skin beholder quarrel sharpened . . .
Proverbs	scent cosmetics fading mourning skin-deep gladden . . .
OMCS	comb accessory salon hair gel mousse curling . . .
BNC	beautiful delight exquisite elegance charm perfect . . .
Reuters	ambitious glittering feature plans mercedes palace . . .

with the various data sets used in our experiments. Specifically, given a thematic word such as, for example, "beauty," we attempt to automatically create lists of semantically similar words using corpus-based measures of word semantic similarity, relying on one data set at a time. The trust is that these semantic spaces centered around given themes/words will highlight the similarities or differences between the specifics of the various corpora used to generate the lists of semantically similar words.

As a measure of semantic similarity, we use the latent semantic analysis (LSA) proposed in Landauer, Foltz, and Laham (1998), considered to be one of the most reliable mechanisms for detecting semantic similarities. Given a word optionally reflecting a certain theme, the LSA process will produce a list of semantically similar words, ranked in reversed order of their similarity. In LSA, term co-occurrences within the documents of a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-document matrix built for the corpus at hand.[8]

As an example, consider the semantic spaces derived for the theme of "beauty," starting with each of the four corpora used in our experiments (one-liners, proverbs, BNC, and Reuters). Table 6 shows the words most closely related to "beauty" derived from each corpus. Interestingly, the semantic space created starting with the one-liners data set includes words that do not seem to recognize the real beauty (e.g., *blinded*, *squint*), or words suggesting lack of objectivity or even disagreement (i.e., *beholder*, *quarrel*). This is in agreement with earlier theories of humor that suggest incongruity and opposition as sources of laughter. Instead, in the proverbs data set, the semantic space of "beauty" seems to suggest that beauty vanishes soon (e.g., *fading*, *mourning*) reflecting the educational purpose of the proverbial sayings. The commonsense data set provides a list of words that are typically related to feminine beauty (e.g., beauty salon accessories such as *comb*, *curling*, etc.). Perhaps not surprising, the Reuters data set suggests that beauty is related to achieving economically important targets (e.g., *Mercedes*, *palace*). Finally, BNC seems to be the most "neutral" in its suggestions, including words such as *elegance*, *delight*, or *perfect*.

For a quantitative analysis, we measured the overlap of the semantic spaces built from each data set, centered in each of the 100 most frequently occurring words from the one-liners data set. For each of these 100 words, we selected the 100 most similar words in the LSA space of each data set. We then measured the overlap of these automatically constructed semantic spaces, obtaining a score of the similarity overlap among the different data sets. Table 7 shows the results, with figures normalized with respect to the largest similarity. According to this

[8]The SVD operation is applied after removing stopwords and considers only the first 100 dimensions.

TABLE 7. Word Similarity: A Quantitative Analysis

One-liners	–				
Proverbs	0.77	–			
OMCS	0.84	0.75	–		
BNC	1.00	0.84	0.97	–	
Reuters	0.44	0.43	0.44	0.49	–
	One-liners	Proverbs	OMCS	BNC	Reuters

analysis, proverbs and Reuters titles are the least similar data sets. Instead, one-liners seem to be very similar to the BNC and commonsense data, and very different from Reuters. This fact is in agreement with our content-based classification results, showing again that humor tends to be very similar to regular text.

5.2. Where Computers Fail

In an attempt to understand the errors made by our current automatic humor-recognition system, as well as the additional sources of knowledge required to identify humorous intent in text, we performed two corpus-based studies.

The first study targets the applicability (*coverage*), the *precision*, and the *recall* of the stylistic and content-based features described in the previous section. Considering the entire corpus of 16,000 one-liners, we found that humor is due, at least in part, to alliteration in 8,323 cases, to word-based antonymy in 2,124 cases, to adult slang in 1,074 cases, and to content-based features in 13,000 cases. Somehow surprising, all these features have similar *precision* figures, ranging from 61% (content-based features) to 65% (alliteration). Recall, however, is considerably different: the adult slang has the smallest recall (6.4%), followed by antonymy (13.2%), alliteration (52.0%), and corpus-based features (81.2%).[9]

The second study was concerned with the additional knowledge required to identify humor in those cases where the stylistic and content-based features implemented in the previous section failed. Starting with a randomly selected sample of 50 BNC sentences, we analyzed the most similar one-liner corresponding to each BNC sentence (as found with the Smart information retrieval system, see Section 2), and tried to identify the source of humor which was present in the humorous one-liner, but not in the BNC sentence, despite their vocabulary similarities. The following additional knowledge sources, potentially useful for humor recognition, were identified:

Irony: The humorous effect in about half of the jokes in our sample data set was due to irony targeted either to the speaker himself (*"Did anyone see my lost carrier?"*), to the dialogue partner (*"Honk if you want to see my finger"*), or to entire professional communities, such as lawyers or programmers (*"Criminal lawyer is a redundancy"*).

Ambiguity: About 20% of the jokes we manually analyzed were based on word ambiguity, and the corresponding potential misinterpretations. For instance, the one-liner *"Change is*

[9]Both precision and recall are determined with respect to the annotations in a mixed corpus of jokes and BNC sentences. From all the statements automatically labeled as jokes, we determined how many were correctly annotated (precision); and from all the jokes in the gold-standard data set, we determined how many were correctly identified as such by the automatic system (recall).

inevitable, except from a vending machine" exploits the ambiguity, and consequently wrong expectations, induced by the word *change*. The statement *change is inevitable* refers to change with the meaning of *action of changing something*, but this meaning is suddenly changed in the second part of the one-liner to that of *balance of money*. This shift of meaning leads to surprise, which then creates the humorous effect.

Incongruity: While we tried to capture the effect of incongruity as modeled through word-based antonymy (Section 4.1), there are however cases (about 10% in our study corpus) where the incongruity stands in opposition among entire phrases—much harder to identify than the opposition among single words. For instance, the comic effect in the one-liner *"A diplomat is someone who can tell you to go to hell in such a way that you will look forward to the trip"* is based on the opposition between *go to hell* and *look forward to the trip*, which cannot be captured with the help of thesauri or semantic networks such as WordNet, but requires different, possibly corpus-based approaches to incongruity detection.

Idiomatic expressions: A relatively large number of one-liners (22%) are based on a reinterpretation of idiomatic expressions. Idioms are typically noncompositional expressions where the semantics of the idiomatic phrase cannot be fully inferred from the semantics of the component words. This potential disjunction of meaning can be exploited with the goal of creating a comic effect: the meaning expectation created by an idiom can be suddenly changed to the (unexpected) meaning inferred by one of the component words, which results in surprise and consequently humor. For example, the one-liner *"Despite the high cost of living, it remains popular"* starts by using the meaning of cost as in *cost of living* (idiomatic expression: *average cost of basic necessities in life*), and then switches to the meaning of cost as in *the monetary value of life*. Similarly, the one-liner *"I used to have an open mind, but my brains kept falling out"* is based on the effect caused by the reinterpretation of the word *open* in the idiom *open mind*, with a switch from the meaning of *open* as *receptive to new ideas* (intended) to that of *open* as *exposed, uncovered* (unintended, but possible).

Commonsense knowledge: Finally, a large fraction of the one-liners (50%) involved understanding of commonsense knowledge, often broken to the effect of creating humor. For instance, *"I like kids, but I don't think I could eat a whole one,"* in addition to the ambiguity of the word *like* (*I like to play with kids* versus *I like to eat kids*), it is also based on the commonsense fact that *one cannot eat kids*.[10] Similarly, many jokes are based on a reinterpretation of the common understanding associated with popular beliefs. For instance, the joke *"Don't drink and drive. You might hit a bump and spill your drink"* is based on the meaning typically inferred by the phrase *don't drink and drive* (because it is unsafe for the driver), to an unexpected reinterpretation *you might spill your drink* (suggesting unsafety for the drink).

5.3. What Makes Us Laugh

Finally, as a final analysis of our experimental results, we tried to identify and classify the content-based humor-specific features learned from our data sets, which could constitute useful features for future studies of humor generation. We examined the most discriminative content-based features learned during the text classification process (Section 4.2), and tried

[10] As a counterexample, consider the similar statement *"I like chicken, but I don't think I could eat a whole one,"* which is not amusing. The lack of humor in this sentence is due to the fact that we typically eat chicken, and therefore this is a regular commonsensical (non-humorous) statement.

to classify them into semantic classes. The following frequently occurring word classes were identified:

Human-centric vocabulary: One-liners seem to constantly make reference to human-related scenarios, through the frequent use of words such as *you*, *I*, *man*, *woman*, *guy*, etc. For instance, the word *you* alone occurs in more than 25% of the one-liners (*"You can always find what you are not looking for"*), while the word *I* occurs in about 15% of the one-liners (*"Of all the things I lost, I miss my mind the most"*). This supports earlier suggestions made by Freud (1905), and later on by Minsky (1980), that laughter is often provoked by feelings of frustration caused by our own, sometime awkward, behavior.

Negation: Humorous texts seem to often include negative word forms, such as *doesn't*, *isn't*, *don't*. For instance, about 3,000 of the 16,000 jokes in our collection contain some form of negation, for example, *"Money can't buy you friends, but you do get a better class of enemy,"* or *"If at first you don't succeed, skydiving is not for you."*

Negative orientation: In addition to negative verb forms, one-liners seem to also contain a large number of words with a negative polarity, such as adjectives with negative connotations like *bad*, *illegal*, *wrong* (*"When everything comes your way, you are in the wrong lane"*), or nouns with a negative load, for example, *error*, *mistake*, *failure* (*"User error: replace user and press any key to continue"*). Both the negative verb forms and the words with negative orientations are potential reflections of the incongruity-based theories of humor.

Professional communities: Many jokes seem to target professional communities that are often associated with amusing situations, such as lawyers, programmers, policemen. About 100 one-liners in our collection fall under this category, for example, *"It was so cold last winter that I saw a lawyer with his hands in his own pockets."*

Human "weakness": Finally, the last significantly large semantic category that we identified refers to events or entities that are often associated with "weak" human moments, including nouns such as *ignorance*, *stupidity*, *trouble* (*"Only adults have trouble with child-proof bottles"*), *beer*, *alcohol* (*"Everybody should believe in something, I believe I'll have another beer"*), or verbs such as *quit*, *steal*, *lie*, *drink* (*"If you can't drink and drive, then why do bars have parking lots?"*). As mentioned before, this kind of vocabulary seems to relate to theories of humor that explain laughter as an effect of frustration or awkward feelings, when we end up laughing "at ourselves" (Minsky 1980).

In addition to these often encountered semantic classes, a more extensive analysis of content-based humor-specific features is likely to reveal other humor-specific content features, which could be turned into useful input features for systems for automatic humor generation.

6. RELATED WORK

While humor is relatively well studied in scientific fields such as linguistics (Attardo 1994) and psychology (Freud 1905; Ruch 2002), to date there is only a limited number of research contributions made toward the construction of computational humor prototypes. Most of the computational approaches to date on style classification have focused on the categorization of more traditional literature genres, such as fiction, scitech, legal, and others (Kessler, Nunberg, and Schuetze 1997), and much less on creative writings such as humor.

One of the first attempts in computational humor is perhaps the work described in Binsted and Ritchie (1997), where a formal model of semantic and syntactic regularities was devised, underlying some of the simplest types of puns (*punning riddles*). The model was then exploited in a system called JAPE that was able to automatically generate amusing puns.

Another humor-generation project was the HAHAcronym project (Stock and Strapparava 2003), whose goal was to develop a system able to automatically generate humorous versions of existing acronyms, or to produce a new amusing acronym constrained to be a valid vocabulary word, starting with concepts provided by the user. The comic effect was achieved mainly by exploiting incongruity theories (e.g., finding a religious variation for a technical acronym).

Another related work, devoted this time to the problem of humor comprehension, is the study reported in Taylor and Mazlack (2004), which focused on a very restricted type of wordplays, namely, the "Knock-Knock" jokes. The goal of the study was to evaluate to what extent wordplay can be automatically identified in "Knock-Knock" jokes, and if such jokes can be reliably recognized from other non-humorous text. The algorithm was based on automatically extracted structural patterns and on heuristics heavily based on the peculiar structure of this particular type of jokes. While the wordplay recognition gave satisfactory results, the identification of jokes containing such wordplays turned out to be significantly more difficult.

Computational humor research can also be found under the more general area of technologies for human-computer interaction, including intelligent interface, personal robots, and other (see Nijholt (2003) for an introduction).

7. CONCLUSION

A conclusion is simply the place where you got tired of thinking. (anonymous one-liner)

The creative genres of natural language have been traditionally considered outside the scope of any computational modeling. In particular humor, because of its puzzling nature, has received little attention from computational linguists. However, given the importance of humor in our everyday life, and the increasing importance of computers in our work and entertainment, we believe that studies related to computational humor will become increasingly important.

In this article, we showed that automatic classification techniques can be successfully applied to the task of humor recognition. Experimental results obtained on very large data sets showed that computational approaches can be efficiently used to distinguish between humorous and non-humorous texts. Significant improvements were observed over a priori known baselines, with accuracy figures ranging from 79.15% (One-liners/BNC) to 96.95% (One-liners/Reuters), which compare favorably with the baseline of 50%. To our knowledge, this is the first result of this kind reported in the literature, as we are not aware of any previous work investigating the interaction between humor and techniques for automatic classification.

Through the analysis of learning curves plotting the classification performance with respect to data size, we showed that the accuracy of the automatic humor-recognizer stops improving after a certain number of examples. Given that automatic humor recognition is a rather understudied problem, we believe that this is an important result, as it provides insights into potentially productive directions for future work. The flattened shape of the curves toward the end of the learning process suggests that rather than focusing on gathering more data,

future work should concentrate on identifying more sophisticated humor-specific features, for example, semantic opposition, ambiguity, and others.

Finally, we performed an in-depth analysis of the experimental results, which revealed interesting aspects concerning the specifics of texts with humorous intent. Specifically, using a corpus-based measure of word semantic similarity, we built semantic spaces centered around given themes (e.g., "beauty"), and highlighted in this way the differences between the various data sets used in our experiments. We also investigated the most discriminative content-based features resulting from the text classification experiments, which pointed out some of the semantic characteristics of the words that make us laugh: human centeredness, negation, negative polarity, certain professional communities, human weaknesses. Finally, an analysis of the errors observed during our experiments helped us identify additional sources of knowledge that could be potentially useful for the task of automatic humor recognition: irony, word ambiguity, phrase-based incongruity, idiomatic expressions, commonsense knowledge. We plan to address these aspects in future work.

REFERENCES

ATTARDO, S. 1994. Linguistic Theory of Humor. Mouton de Gruyter, Berlin.

BINSTED, K., and G. RITCHIE. 1997. Computational rules for punning riddles. Humor, **10**(1):25–76.

BUCARIA, C. 2004. Lexical and syntactic ambiguity as a source of humor. Humor, **17**(3):279–309.

DAELEMANS, W., J. ZAVREL, K. VAN DER SLOOT, and A. VAN DEN BOSCH. 2001. Timbl: Tilburg memory based learner, version 4.0, reference guide. Technical report, University of Antwerp, Antwerp, Belgium.

FREUD, S. 1905. Der Witz und Seine Beziehung zum Unbewussten. Deutike, Vienna.

JOACHIMS, T. 1998. Text categorization with support vector machines: Learning with many relevant features. *In* Proceedings of the European Conference on Machine Learning, pp. 137–142, Chemnitz, Germany.

KESSLER, B., G. NUNBERG, and H. SCHUETZE. 1997. Automatic detection of text genre. *In* Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL97), pp. 32–38, Madrid, July.

KNUTH, D. E. 1993. The Stanford Graph Base: A Platform for Combinatorial Computing. ACM Press, New York.

LANDAUER, T. K., P. FOLTZ, and D. LAHAM. 1998. Introduction to latent semantic analysis. Discourse Processes, **25**: 259–284.

LEWIS, D., Y. YANG, T. ROSE, and F. LI. 2004. RCV1: A new benchmark collection for text categorization research. The Journal of Machine Learning Research, **5**:361–397.

MCCALLUM, A., and K. NIGAM. 1998. A comparison of event models for Naive Bayes text classification. *In* Proceedings of AAAI-98 Workshop on Learning for Text Categorization, pp. 44–48, Madison, Wisconsin.

MIHALCEA, R., and C. STRAPPARAVA. 2005. Bootstrapping for fun: Web-based construction of large data sets for humor recognition. *In* Proceedings of the Workshop on Negotiation, Behaviour and Language (FINEXIN 2005), pp. 23–30, Ottawa, Canada.

MILLER, G. 1995. Wordnet: A lexical database. Communications of the ACM, **38**(11):39–41.

MINSKY, M. 1980. Jokes and the logic of the cognitive unconscious. Technical report, MIT Artificial Intelligence Laboratory, Boston, Massachusetts.

NIJHOLT, A. 2003. Disappearing computers, social actors and embodied agents. *In* Proceedings 2003 International Conference on CYBERWORLDS. *Edited by* T. L. Kunii, S. Hock Soon, and A. Sourin, pp. 128–134, Singapore, December 2003.

NIJHOLT, A., O. STOCK, A. DIX, and J. MORKES. (Editors). 2003. *In* Proceedings of CHI-2003 Workshop: Humor Modeling in the Interface, Fort Lauderdale, Florida.

RUCH, W. 2002. Computers with a personality? Lessons to be learned from studies of the psychology of humor. *In* Proceedings of the The April Fools Day Workshop on Computational Humour, pp. 57–70, University of Twente, the Netherlands.

SINGH, P. 2002. The public acquisition of commonsense knowledge. *In* Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access, Palo Alto, California.

STOCK, O., and C. STRAPPARAVA. 2003. Getting serious about the development of computational humour. *In* Proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI-03), pp. 59–64, Acapulco, Mexico.

STOCK, O., C. STRAPPARAVA, and A. NIJHOLT. (Editors). 2002. Proceedings of the The April Fools Day Workshop on Computational Humour, Trento.

TAYLOR, J., and L. MAZLACK. 2004. Computationally recognizing wordplay in jokes. *In* Proceedings of CogSci 2004, pp. 1315–1320, Chicago.

VAPNIK, V. 1995. The Nature of Statistical Learning Theory. Springer, New York.

YANG, Y., and X. LIU. 1999. A reexamination of text categorization methods. *In* Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49, Berkeley, California.