Humor Recognition and Humor Anchor Extraction

Diyi Yang, Alon Lavie, Chris Dyer, Eduard Hovy
Language Technologies Institute, School of Computer Science
Carnegie Mellon University. Pittsburgh, PA, 15213, USA
{diyiy, alavie, cdyer}@cs.cmu.edu, hovy@cmu.edu

Abstract

Humor is an essential component in personal communication. How to create computational models to discover the structures behind humor, recognize humor and even extract humor anchors remains a challenge. In this work, we first identify several semantic structures behind humor and design sets of features for each structure, and next employ a computational approach to recognize humor. Furthermore, we develop a simple and effective method to extract anchors that enable humor in a sentence. Experiments conducted on two datasets demonstrate that our humor recognizer is effective in automatically distinguishing between humorous and non-humorous texts and our extracted humor anchors correlate quite well with human annotations.

1 Introduction

Humor is one of the most interesting and puzzling research areas in the field of natural language understanding. Recently, computers have changed their roles from automatons that can only perform assigned tasks to intelligent agents that dynamically interact with people and learn to understand their users. When a computer converses with a human being, if it can figure out the humor in human's language, it can better understand the true meaning of human language, and thereby make better decisions that improve the user experience. Developing techniques that enable computers to understand humor in human conversations and adapt behavior accordingly deserves particular attention.

The task of *Humor Recognition* refers to determining whether a sentence in a given context expresses a certain degree of humor. Humor recognition is a challenging natural language problem (Attardo, 1994). First, a universal definition of humor is hard to achieve, because different people hold different understandings of even the same sentence. Second, humor is always situated in a broader context that sometimes requires a lot of external knowledge to fully understand it. For example, consider the sentence, "*The one who invented the door knocker got a No Bell prize*" and "*Veni, Vidi, Visa: I came, I saw, I did a little shopping*". One needs a larger cultural context to figure out the subtle humorous meaning expressed in these two sentences. Last but not least, there are different types of humor (Raz, 2012), such as wordplay, irony and sarcasm, but there exist few formal taxonomies of humor characteristics. Thus it is almost impossible to design a general algorithm that can classify all the different types of humor, since even human cannot perfectly classify all of them.

Although it is impossible to understand universal humor characteristics, one can still capture the possible latent structures behind humor (Bucaria, 2004; Binsted and Ritchie, 1997). In this work, we uncover several latent semantic structures behind humor, in terms of meaning incongruity, ambiguity, phonetic style and personal affect. In addition to humor recognition, identifying anchors, or which words prompt humor in a sentence, is essential in understanding the phenomenon of humor in language. Here, *Anchor Extraction* refers to extracting the semantic units (keywords or phrases) that enable the humor in a given sentence. The presence of such anchors plays an important role in generating humor within a sentence or phrase.

In this work, we formulate humor recognition as a classification task in which we distinguish between humorous and non-humorous instances. Then we explore the semantic structure behind humor from four perspectives: incongruity, am-

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2367–2376,
Lisbon, Portugal, 17-21 September 2015.

biguity, interpersonal effect and phonetic style. For each latent structure, we design a set of features to capture the potential indicators of humor. With high classification accuracy, we then extract humor anchors in sentences via a simple and effective method. Both quantitative and qualitative experimental results are provided to validate the classification and anchor extraction performance.

2 Related Work

Most existing studies on humor recognition are formulated as a binary classification problem and try to recognize jokes via a set of linguistic features (Purandare and Litman, 2006; Kiddon and Brun, 2011). For example, Mihalcea and Strapparava (2005) defined three types of humor-specific stylistic features: Alliteration, Antonym and Adult Slang, and trained a classifier based on these feature representations. Similarly, Zhang and Liu (2014) designed several categories of humor-related features, derived from influential humor theories, linguistic norms, and affective dimensions, and input around fifty features into the Gradient Boosting Regression Tree model for humor recognition. Taylor and Mazlack (2004) recognized wordplay jokes based on statistical language recognition techniques, where they learned statistical patterns of text in N-grams and provided a heuristic focus for a location of where wordplay may or may not occur. Similar work can also be found in (Taylor, 2009), which described humor detection process through Ontological Semantics by automatically transposing the text into the formatted text-meaning representation to detect humor. In addition to language features, some other studies also utilize spoken or multimodal signals. For example, Purandare and Litman (2006) analyzed acoustic-prosodic and linguistic features to automatically recognize humor during spoken conversations. However, the humor related features in most of those works are not systematically derived or explained.

One essential component in humor recognition is the construction of negative data instances. Classifiers based on negative samples that lie in a different domain than humor positive instances will have high classification performance, but are not necessarily good classifiers. There are few existing benchmark datasets for humor recognition and most studies select negative instances specifically. For example, Mihalcea and Strapparava (2005) constructed the set of negative examples by using news title from Reuters news, proverbs and British National Corpus. (Zhang, el. al 2014) randomly sampled 1500 tweets and then asked annotators to filter out humorous tweets.

Compared to humor recognition, humor generation has received quite a lot attention in the past decades(Stock and Strapparava, 2005; Ritchie, 2005; Hong and Ong, 2009). Most generation work draws on humor theories to account for humor factors, such as the Script-based Semantic Theory of Humor (Raskin, 1985; Labutov and Lipson, 2012) and employs templates to generate jokes. For example, Ozbal and Strapparava (2012) created humorous neologism using WordNet and ConceptNet. In detail, their system combined several linguistic resources to generate creative names, more specifically neologisms based on homophonic puns and metaphors. Stock and Strapparava (2005) introduced HAHACRONYM, a system (an acronym ironic re-analyzer and generator) devoted to produce humorous acronyms mainly by exploiting incongruity theories (Stock and Strapparava, 2003).

In contrast to research on humor recognition and generation, there are few studies that identify the humor anchors that trigger humorous effects in general sentences. A certain type of jokes might have specific structures or characteristics that provide pointers to humor anchors. For example, in the problem of "That's what she said" (Kiddon and Brun, 2011), characteristics that involves the using of nouns that are euphemisms for sexually explicit nouns or structures common in the erotic domain might probably give clues to potential humor anchors. Similarly, in the Knock Knock jokes (Taylor and Mazlack, 2004), wordplay is what leads to the humor. However, the wordplay by itself is not enough to trigger the comic effect, thus not equivalent to the humor anchors for a joke. To address these issues, we introduce a formal definition of humor anchors and design an effective method to extract such anchors in this work. To the best of our knowledge, this is the first study on extracting humor anchors that trigger humor in general sentences.

3 Data Preparation

To perform automatic recognition of humor and humor anchor extraction, a data set consisting of both humorous (positive) and non-humorous (negative) examples is needed. The dataset we

use to conduct our humor recognition experiments includes two parts: Pun of the Day [1] and the 16000 One-Liner dataset (Mihalcea and Strapparava, 2005). The two data sets only contain humorous text. In order to acquire negative samples for the humor classification task, we sample negative samples from four resources, including AP News[2], New York Times, Yahoo! Answer[3] and Proverb[4]. Such datasets not only enable us to automatically learn computational models for humor recognition, but also provide us with the chances to evaluate the performance of our model.

However, directly applying sentences extracted from those four resources and simply treating them as negative instances of humor recognition could result in deceptively high performance of classification, due to the domain differences between positive and negative datasets. For example, the humor sentences in our positive datasets often relate to daily lives, such as "*My wife tells me I'm a skeptic, but I don't believe a word she says.*". Meanwhile, sentences in news websites sometimes describe scenes related to wars or politics, such as "*Judge Thomas P. Griesa of Federal District Court in Manhattan stopped short of issuing sanctions*". Such domain differences between descriptive words might make a naive bag of words model perform quite well, without taking into account the deeper semantic structures behind humor. To deal with this issue, we extract our negative instances in a way that tries to minimize such domain differences by (1) selecting negative instances whose words are all contained in our positive instance word dictionary and (2) forcing the text length of non-humorous instances to follow the similar length restriction as humorous examples, i.e. one sentence with an average length of 10-30 words. Here, we assume sentences come from the aforementioned four resources are all non-humorous in nature. Table 1 provides a detailed statistical description to our datasets.

Dataset	#Positive	#Negative
Pun of the Day	2423	2403
16000 One Liners	16000	16002

Table 1: Statistics on Two Datasets

4 Latent Structures behind Humor

In this section, we explore the latent semantic structures behind humor in four aspects: (a) Incongruity; (b) Ambiguity; (c) Interpersonal Effect and (d) Phonetic Style. For each latent structure, a set of features is designed to capture the corresponding indicators of humor.

4.1 Incongruity Structure

"Laughter arises from the view of two or more inconsistent, unsuitable, or incongruous parts or circumstances, considered as united in complex object or assemblage, or as acquiring a sort of mutual relation from the peculiar manner in which the mind takes notice of them" (Lefcourt, 2001). The essence of the laughable is the incongruous, the disconnecting of one idea from another (Paulos, 2008). Humor sometimes relies on a certain type of incongruity, such as opposition or contradiction. For example, the following 'clean desk' and 'cluttered desk drawer' example (Mihalcea and Strapparava, 2005) presents an incongruous/contrast structure, resulting in a comic effect.

A clean desk is a sign of a cluttered desk drawer.

Direct identification of incongruity is hard to achieve, however, it is relatively easier to measure the semantic disconnection in a sentence. Taking advantage of Word2Vec[5], we extract two types of features to evaluate the meaning distance[6] between content word pairs in a sentence (Mikolov et al., 2013):

- Disconnection: the maximum meaning distance of word pairs in a sentence.
- Repetition: the minimum meaning distance of word pairs in a sentence.

4.2 Ambiguity Theory

Ambiguity (Bucaria, 2004), the disambiguation of words with multiple meanings (Bekinschtein et al., 2011), is a crucial component of many humor jokes (Miller and Gurevych, 2015). Humor and ambiguity often come together when a listener expects one meaning, but is forced to use another

[1]Pun of the Day: `http://www.punoftheday.com/` This constructed dataset will be made public.

[2]http://hosted.ap.org/dynamic/fronts/HOME?SITE=AP

[3]https://answers.yahoo.com/

[4]Manually extracted 654 proverbs from Proverb websites

[5]https://code.google.com/p/word2vec/

[6]We take the generic Word2Vec vectors without training new vectors for our specific domain. In addition, vectors associated with senses (Kumar Jauhar et al., 2015) might be alternative advantageous in this task.

meaning. Ambiguity occurs when the words of the surface sentence structure can be grouped in more than one way, thus yielding more than one associated deep structures, as shown in the example below.

Did you hear about the guy whose whole left side was cut off? He's all right now.

The multiple possible meanings of words provide readers with different understandings. To capture the ambiguity contained in a sentence, we utilize the lexical resource WordNet (Fellbaum, 1998) and capture the ambiguity as follows:

- Sense Combination: the sense combination in a sentence computed as follows: we first use a POS tagger (Toutanova et al., 2003) to identify Noun, Verb, Adj, Adv. Then we consider the possible meanings of such words $\{w_1, w_2 \cdots w_k\}$ via WordNet and calculate the sense combinations as $log(\prod_{i=1}^{k} n_{w_i})$. n_{w_i} is the total number of senses of word w_i.
- Sense Farmost: the largest Path Similarity[7] of any word senses in a sentence.
- Sense Closest: the smallest Path Similarity of any word senses in a sentence.

4.3 Interpersonal Effect

Besides humor theories and linguistic style modeling, one important theory behind humor is its social/hostility focus, especially regarding its interpersonal effect on receivers. That is, humor is essentially associated with sentiment (Zhang and Liu, 2014) and subjectivity (Wiebe and Mihalcea, 2006). For example, a sentence is likely to be humorous if it contains some words carrying strong sentiment, such as 'idiot' as follows.

Your village called. They want their Idiot back.

Each word is associated with positive or negative sentiments and such measurements reflect the emotion expressed by the writer. To identify the word-associated sentiment, we use the word association resource in the work by (Wilson et al., 2005), which provides annotations and clues to measure the subjectivity and sentiment associated with words. This enables us to design the following features.

- Negative (Positive) Polarity: the number of occurrences of all Negative (Positive) words.
- Weak (Strong) Subjectivity: the number of occurrences of all Weak (Strong) Subjectivity oriented words in a sentence. It is the linguistic expression of people's opinions, evaluations, beliefs or speculations.

4.4 Phonetic Style

Many humorous texts play with sounds, creating incongruous sounds or words. Some studies (Mihalcea and Strapparava, 2005) have shown that the phonetic properties of humorous sentences are at least as important as their content. Many one-liner jokes contain linguistic phenomena such as alliteration, word repetition and rhyme that produce a comic effect even if the jokes are not necessarily meant to be humorous in content.

What is the difference between a nicely dressed man on a tricycle and a poorly dressed man on a bicycle? A tire.

An alliteration chain refers to two or more words beginning with the same phones. A rhyme chain is defined as the relationship that words end with the same syllable. To extract this phonetic feature, we take advantage of the CMU Pronouncing Dictionary [8] and design four features as follows:

- Alliteration: the number of alliteration chains in a sentence, and the maximum length of alliteration chains.
- Rhyme: the number of rhyme chains and the maximum length of rhyme chains.

5 Humor Anchor Extraction

In addition to humor recognition, identifying anchors, or which words prompt humor in a sentence, is also essential in understanding humor language phenomena. In this section, we first define what humor anchors are and then describe how to extract such semantic units that enable humor in a given sentence.

5.1 Humor Anchor Definition

The semantic units or humor anchors enable humor in a given sentence, and are reflected in the form of sentence words. However, not every single word can be a humor anchor. For example, *I am glad that I know sign language; it is pretty handy*. In this one-liner, words such as 'am' and 'is' are not able to enable humor

[7] Path Similarity: `http://www.nltk.org/howto/wordnet.html`

[8] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

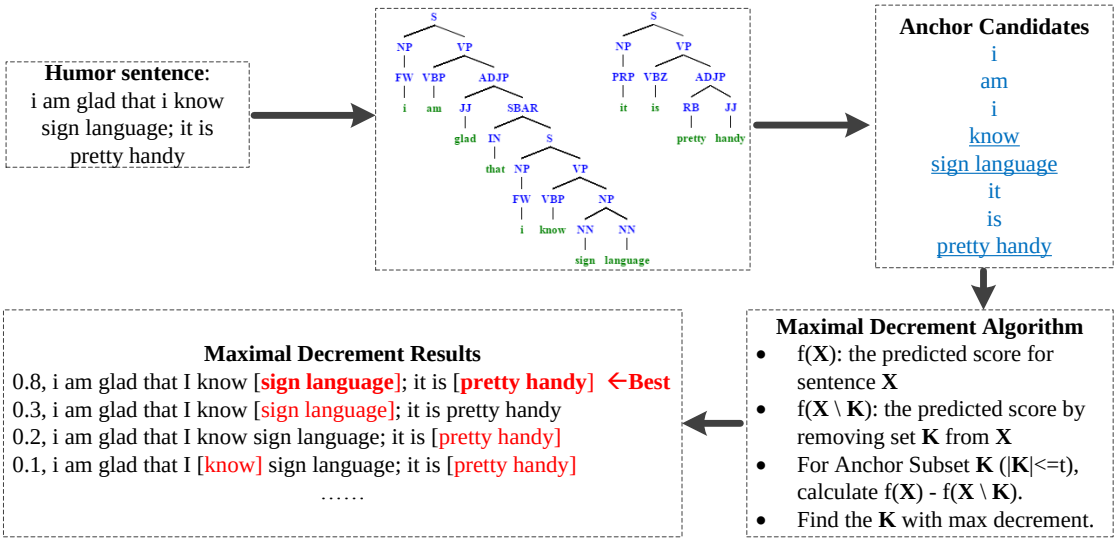

Figure 1: Humor Anchor Extraction Overview. Based on the parsing output of each sentence, we generate its humor anchor candidates. We then apply the Maximal Decrement algorithm to these candidates. The humor anchor subset that gives the maximal decrement is the extracted humor anchors for that sentence.

via themselves. Similarly, 'sign' or 'language' itself are not capable to prompt comic effect. The possible anchors in this example should contain both 'sign language' and 'handy'; it is the combination of these two spans that triggers humor. Therefore, formally defined, a humor anchor is a meaningful, complete, minimal set of word spans in a sentence that potentially enable one of the latent structures of Section 4 to occur. (1) *Meaningful* means humor anchors are meaningful word spans, not meaningless stop words in a sentence; (2) *Completeness* shows that all possible humor anchors should be covered by this anchor set and no individual span in this anchor set is capable enough to enable humor; (3) *Minimal* emphasizes that it is the combination of these anchors together that prompts comic effect; discarding any anchors from this candidate set destroys the humorous effect.

5.2 Anchor Extraction Method

Based on the humor anchor requirements listed above, we scoped humor anchor candidates to words or phrases that belong to the syntactic categories of Noun, Verb, Noun Phrase, Verb Phrase, ADVP or ADJP. Those properties are acquired via a sentence parse tree. To generate anchor candidates, we parsed each sentence and selected words or phrases that satisfy one or more of the latent structure criteria by first extracting the minimal parse subtrees of NP, VP, ADVP and ADJP and then adding remaining Nouns and Verbs into candidate sets.

The above anchor generation process provides us with all possible anchors that might enable humor. It satisfies the *Meaningful* and *Completeness* requirements. To extract a *Minimal* set of anchors, we proposed a simple and effective method of Maximal Decrement. Its basic idea is summarized as follows: Each complete sentence has a predicted humor score, which is computed via a humor recognition classifier trained on all data points. This humor recognizer is not limited to any specific classifiers or features as long as it provides good classification accuracy, which guarantees the generalization ability of our anchor extraction method. We next enumerate a subset of anchors from all potential anchors for this sentence. Then, we recompute the predicted humor score by providing the classifier with features associated with the current sentence, after removing that subset of anchors. Note that our designed humor structural features are all word order free, thereby not distinguishing between complete and incomplete sentences. The subset of humor anchor candidates that provides the maximum decrement of humor predicted scores is then returned as the extracted humor anchor set.

Mathematically, X_i is the word set of sentence i. Let f denote the trained classifier on all data instances. $f(X_i)$ is the predicted humor score

for sentence i before performing any operations. Denote $K_i (K_i \subset X_i)$ as the subset of words that we need to remove from sentence i. The size of K_i should be smaller than a threshold t, $|K_i| \leq t$. $f(X_i/K_i)$ is the recomputed humor score for sentence i after removing K_i. Our Maximal Decrement method tries to maximize the following objective by enumerating all possible K_is. The subset K_i that gives the maximal decrement is returned as our extracted humor anchors for sentence i. The system overview is shown in Figure 1.

$$\arg \min_{|K_i \leq t|} f(X_i) - f(X_i/K_i) \quad (1)$$

6 Experiment

In this section, we validate the performance of different semantic structures we extracted on humor recognition and how the combination of the structures contributes to classification. In addition, both qualitative and quantitative results regarding humor anchor extraction performance are explored.

6.1 Humor Recognition

We formulate humor recognition as a traditional text classification problem, and apply Random Forest to perform 10 fold cross validation on two datasets. Random Forest is an ensemble of decision trees [9] for classification (regression) that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes output by individual trees. Unlike single decision trees, which are likely to suffer from high variance or high bias, random forests use averaging to find a natural balance between the two extremes.

In addition to the four latent structures behind humor, we also design a set of K Nearest Neighbor (KNN) features that uses the humor classes of the K sentences ($K = 5$) that are the closest to this sentence in terms of meaning distance in the training data. We use several methods to act as baselines for comparison with our classier. **Bag of Words** baseline is used to capture a multiset of words in a sentence that might differentiate humor and non-humor. **Language Model** baseline assigns a humor/nonhumor probability to words in a sentence via probability distributions. **Word2Vec** baseline represents the meaning of sentences via Word2Vec (Mikolov et al., 2013) distributional semantic meaning representation. We implemented an earlier work (Mihalcea and Strapparava, 2005) that exploits stylistic features including alliteration, autonomy and adult slang and ensembles with bag of words representations, denoted as **SaC Ensemble**. It is worth mentioning that our datasets are balanced in terms of positive and negative instances, giving a random classification accuracy of 50%.

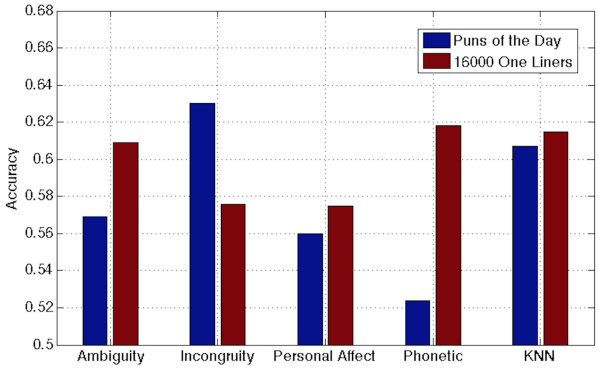

Figure 2: Different Latent Structures' Contribution to Humor Recognition

We first explored how different latent semantic structures affect humor recognition performance and summarize the results in Figure 2. It is evident that *Incongruity* performs the best among all latent semantic structures in the context of Pun of the Day and both *Ambiguity* and *Phonetic* substantially contribute to recognition performance on the 16000 One Liners dataset. The reason behind the differences in performance with *Incongruity* and with *Phonetic* lies in the different nature of the corpus. Most puns are well structured and play with contrasting or incongruous meaning. However, humor sentences in the 16000 One Liners often rely on the reader's awareness of attention-catching sounds (Mihalcea and Strapparava, 2005). This demonstrates that humor characteristics are expressed differently in different contexts and datasets.

We also investigated how the combination of such semantic structures performs compared with our proposed baselines, as shown in Table 2. Here, we denote the combination of four latent structures and KNN features as Human Centric Features (**HCF**). From Table 2, we found that (1) HCF (21 features in total) has a bigger contribution to humor recognition, compared with Bag of Words and Language Model (LM). The

[9] https://www.kaggle.com/wiki/RandomForests

	Pun of the Day				16000 One Liners			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
HCF	0.705	0.696	0.736	0.715	0.701	0.685	0.746	0.714
Bag of Words	0.632	0.623	0.686	0.650	0.673	0.708	0.662	0.684
Language Model	0.627	0.602	0.762	0.673	0.635	0.645	0.596	0.620
Word2Vec	0.833	0.804	0.880	0.841	0.781	0.767	0.809	0.787
SaC Ensemble	0.763	**0.838**	0.655	0.735	0.662	0.628	0.796	0.701
Word2Vec+HCF	**0.854**	0.834	**0.888**	**0.859**	**0.797**	**0.776**	**0.836**	**0.805**

Table 2: Comparison of Different Methods of Humor Recognition

inadequacy of LM also indicates that we can alleviate the domain differences and capture the real humor. (2) SaC Ensemble is inferior to the combination of Word2Vec and HCF because it does not involve enough latent structures such as Interpersonal Effect and distributional semantics. (3) The combination of Word2Vec and HCF (Word2Vec+HCF) gives the best classification performance because it takes into account both latent structures and semantic word meanings. Such a conclusion is consistent across two datasets. This indicates that our extracted latent semantic structures are effective in capturing humorous meaning.

6.2 Anchor Extraction

Qualitative Evaluation

The above humor recognition classifier provides us with decent accuracy in identifying humor in the text. To better understand which words or semantic units enable humor in sentences, we performed humor anchor extraction as described in Section 5.2. We set the size of the humor anchor set as 3, i.e. $t = 3$. The classifier that is used to predict the humor score is trained on all data instances. Then all predicted humorous instances are collected and input into the humor anchor extraction component. Based on the Maximal Decrement method, a set of humor anchors is extracted for each instance.

Table 3 presents selected extracted humor anchor results, including both successful and unsatisfying extractions. As we can see, extracted humor anchors are quite reasonable in explaining the humor causes or focuses. For example, in the sentence "*I used to be a watchmaker; it is a great job and I made my own hours*", our method selected 'watchmaker', 'made' and 'hours' as humor anchors. It makes sense because each word is necessary and essential to enable humor. Deleting 'watchmaker' will make the combination of 'made' and 'hours' helpless to the comic effect. To sum up, our extracted anchor extraction works fairly well in identifying the focus and meaning of humor language.

Quantitative Evaluation

In addition to the above qualitative exploration, we also conducted quantitative evaluations. For each dataset, we randomly sampled 200 sentences. Then for each sentence, 3 annotators are asked to annotate and label the possible humor anchors. To assess the consistency of the labeling in this context, we introduced an Annotation Agreement Ratio (AAR) measurement as follows:

$$AAR(A, B) = \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{|A_i \cap B_i|}{|A_i \cup B_i|}$$

Here, N_s is the total number of sentences. A_i and B_i are the humor anchor sets of sentence i provided by annotator A and B respectively. The AARs on Pun of the Day and 16000 One Liners datasets are 0.618 and 0.433 respectively, computed by averaging the AAR scores between any two different annotators, which indicate relatively reasonable agreement.

As a further step to validate the effectiveness of our anchor extraction method, we also introduced two baselines. The **Random Extraction** baseline selects humor anchors by sampling words in a sentence randomly. Similarly, **POS Extraction** baseline generates anchors by narrowing down all the words in a sentence to a set of certain POS, e.g. Noun, Verb, Noun Phrase, Verb Phrase, ADVP and ADJP and then sampling words from this set.

To evaluate whether our extracted anchors are consistent with human annotation, we used each annotator's extracted anchor list as the ground truth, and compared with anchor list provided by our method. To identify whether two anchors

Result Category	Representative Sentences
Good	Did you hear about the guy who got **hit** in the head with a can of **soda**? He was lucky it was a **soft drink**.
	I was struggling to figure out how **lightning works** then it **struck** me.
	The one who **invented** the **door knocker** got a **No-bell prize**.
	I used to be a **watchmaker**; it is a great job and I **made** my own **hours**.
Bad	I wanted to lose weight, so I **went** to the **paint store**. I **heard** I could get thinner there.
	I **used** to be a **banker** but I **lost** interest

Table 3: Representative Extracted Humor Anchors. Highlighted parts are the extracted humor anchors in a sentence.

are the same, we introduce two measurements: Exact (EX) Matching and At-Least-One (ALO) Matching. Exact Matching requires the two anchors to be exactly the same. For ALO, two anchors are considered the same if they have at least one word in common. Recall, Precision and F1 Score are act as evaluation metrics. We then average the three annotators' individual scores to get the final extraction performance.

Metrics	Recall	Precision	F1
Pun of the Day Dataset			
MDE EX	0.444	0.446	0.438
POS EX	0.166	0.170	0.165
Random EX	0.121	0.116	0.116
MDE ALO	0.782	0.784	0.756
POS ALO	0.364	0.371	0.360
Random ALO	0.297	0.287	0.285
16000 One Liners Dataset			
MDE EX	0.314	0.281	0.288
POS EX	0.104	0.110	0.104
Random EX	0.087	0.075	0.079
MDE ALO	0.675	0.638	0.616
POS ALO	0.386	0.363	0.356
Random ALO	0.341	0.334	0.319

Table 4: Quantitative Result Comparison of Humor Anchor Extraction

The quantitative evaluation results are summarized in Table 4. Maximal Decrement Extraction is denoted as MDE; POS Extraction is denoted as POS, and Random Extraction is denoted as Random. We report both ALO and EX results for MED, POS and Random. From Table 4, we found that MDE performs quite well under the measurement of human annotation in terms of both ALO and EX settings. This again validates our assumption towards humor anchors and the effectiveness of our anchor extraction method.

6.3 Discussion

The above two subsections described the performance of both humor recognition and humor anchor extraction tasks. In terms of humor recognition, incongruity, ambiguity, personal affect and phonetic style are taken into consideration to assist the identification of humorous language. We focus on discovering generalized structures behind humor, and did not take into account sexual oriented words such as adult slang in modeling humorous language. Based on our results, these four latent structures are effective in capturing humor characteristics and such characteristics are expressed to different extents in different contexts. Note that we can apply any classification methods with our humor latent structures. Once such structures help us acquire high recognition accuracy, we can perform the generalized Maximal Decrement extraction method to identify anchors in humorous text.

Both humor recognition and humor anchor extraction suffer from several common issues. (1) **Phrase Meaning**: For example, a humorous sentence "*How does the earth get clean? It takes a meteor shower*" is predicted as non-humorous, because the recognizer does not fully understand the meaning of 'meteor shower', let alone the comic effect caused by 'earth', 'clean' and 'meteor shower'. For the unsatisfying example in Table 3 "*I used to be a banker but I lost interest*", anchor extraction would work better if it recognizes 'lost interest' correctly as a basic semantic unit. (2) **External Knowledge**: For jokes that involve idioms or social phenomena, or need some external knowledge such as "*Veni, Vidi, Visa: I came, I saw, I did a little shopping*", both humor recognition and anchor extraction fail because a broader and implicit comparison of this sentence and its origin ("*Veni, Vidi, Vici: I came, I*

saw, I conquered... ") is hard to be captured from a sentence. (3) **Humor Categorization**: Moreover, a fine granularity categorization of humor might aid in understanding humorous language, because humor has different types of manifestations, such as irony, sarcasm, creativity, insult and wordplay. Therefore, more sophisticated techniques in modeling phrase meaning, external knowledge, humor types, etc., are needed to better expose and define humor for automatic recognition and extraction.

7 Conclusion

In this work, we focus on understanding humorous language through two subtasks: humor recognition and humor anchor extraction. For this purpose, we first designed four semantic structures behind humor. Based on the designed sets of features associated with each structure, we constructed different computational classifiers to recognize humor. Then we proposed a simple and effective Maximal Decrement method to automatically extract anchors that enable humor in a sentence. Experimental results conducted on two datasets demonstrate the effectiveness of our proposed latent structures. The performances of humor recognition and anchor extraction are superior compared to several baselines. In the future, we would like to step further into the discovery of humor characteristics and apply our findings to the process of humor generation.

Acknowledgement

The authors would like to thank Li Zhou, Anna Kasunic, the anonymous reviewers, our annotators and all colleagues who have contributed their valuable comments and suggestions.

References

Salvatore Attardo. 1994. *Linguistic theories of humor*, volume 1. Walter de Gruyter.

Tristan A Bekinschtein, Matthew H Davis, Jennifer M Rodd, and Adrian M Owen. 2011. Why clowns taste funny: the relationship between humor and semantic ambiguity. *The Journal of Neuroscience*, 31(26):9665–9671.

Kim Binsted and Graeme Ritchie. 1997. Computational rules for generating punning riddles. *Humor: International Journal of Humor Research*.

Chiara Bucaria. 2004. Lexical and syntactic ambiguity as a source of humor: The case of newspaper headlines. *Humor*, 17(3):279–310.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Bryan Anthony Hong and Ethel Ong. 2009. Automatically extracting word relationships as templates for pun generation. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, CALC '09, pages 24–31, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chloe Kiddon and Yuriy Brun. 2011. That's what she said: double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 89–94. Association for Computational Linguistics.

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics*.

Igor Labutov and Hod Lipson. 2012. Humor as circuits in semantic networks. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 150–155. Association for Computational Linguistics.

Herbert M Lefcourt. 2001. *Humor: The psychology of living buoyantly*. Springer Science & Business Media.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 531–538. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tristan Miller and Iryna Gurevych. 2015. Automatic disambiguation of english puns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 719–729, Beijing, China, July. Association for Computational Linguistics.

Gözde Ozbal and Carlo Strapparava. 2012. Computational humour for creative naming. *Computational Humor 2012*, page 15.

John Allen Paulos. 2008. *Mathematics and humor: A study of the logic of humor*. University of Chicago Press.

Amruta Purandare and Diane Litman. 2006. Humor: Prosody analysis and automatic recognition for f*r*i*e*n*d*s*. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 208–215, Stroudsburg, PA, USA. Association for Computational Linguistics.

Victor Raskin. 1985. *Semantic mechanisms of humor*, volume 24. Springer.

Yishay Raz. 2012. Automatic humor classification on twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 66–70. Association for Computational Linguistics.

Graeme Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the 10th European Natural Language Generation Workshop*, pages 125–132. Citeseer.

Oliviero Stock and Carlo Strapparava. 2003. Getting serious about the development of computational humor. In *IJCAI*, volume 3, pages 59–64.

Oliviero Stock and Carlo Strapparava. 2005. Hahacronym: A computational humor system. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 113–116. Association for Computational Linguistics.

J Taylor and L Mazlack. 2004. Computationally recognizing wordplay in jokes. *Proceedings of CogSci 2004*.

Julia M Taylor. 2009. Computational detection of humor: A dream or a nightmare? the ontological semantics approach. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03*, pages 429–432. IEEE Computer Society.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1065–1072. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 889–898. ACM.