

Linguistic Ethnography: Identifying Dominant Word Classes in Text

Rada Mihalcea^{1,2}, Stephen Pulman²

¹ Computer Science Department, University of North Texas
rada@cs.unt.edu

² Computational Linguistics Group, Oxford University
sgp@clg.ox.ac.uk

Abstract. In this paper, we propose a method for “linguistic ethnography” – a general mechanism for characterising texts with respect to the dominance of certain classes of words. Using humour as a case study, we explore the automatic learning of salient word classes, including semantic classes (e.g., person, animal), psycholinguistic classes (e.g., tentative, cause), and affective load (e.g., anger, happiness). We measure the reliability of the derived word classes and their associated dominance scores by showing significant correlation across different corpora.

1 Introduction

Text classification is an area in natural language processing that has received a significant amount of interest from both the research and industrial communities, with numerous applications ranging from spam detection and Web directory categorization [4], to sentiment and subjectivity classification [17], emotion recognition [14], gender identification [3] or humour recognition [6]. The task is defined as the automatic identification and labeling of texts that share certain properties, be that a common topic (e.g., “arts”), a common author (e.g., female-authored texts), or a certain feature of the text (e.g., humorous texts).

While there are a number of text classification algorithms that have been proposed to date, there are only a handful of techniques that have been developed to identify the characteristics that are shared by the texts in a given class. Most of the work in this area has focused on the use of weights associated with the words in the text, by using metrics such as tf.idf or information gain, but no attempts have been made to systematically identify broader patterns or word classes that are common in these texts. The relatively small amount of work in this area is understandable since, from a practical perspective, the accurate classification of texts is more important than the identification of general word classes that are specific to the texts in one category.

When the goal however is to *understand the characteristics* of a certain type of text, in order to gain a better understanding of the properties or behaviours modeled by those texts (such as happiness, humour, or gender), then the systematic identification of broad word classes characteristic to the given type of text is considerably more insightful than a mere figure reflecting the accuracy of a text classifier.

Given a collection of texts, characterised by a certain property that is shared by all the texts in the collection, we introduce a method to automatically discover the classes

of words that are dominant in the given type of text. For instance, given a collection of texts that are either humorous, or that reflect the happy mood of the writer, or the specifics of the gender of the author, the method can be used to identify those word classes that are typical to the given texts. For example, the method can find that words that describe *humans* are more often encountered in humorous texts, and thus suggest the human-centeredness of humour. Or, it can find that words that are used to characterize *novelty* are frequently used in texts describing happy moods, and thus indicate a possible connection between novelty and happiness.

In the following, we first introduce the method to automatically assign a dominance score to word classes to indicate their saliency in a type of text. We then describe three lexical resources that define word classes, including Roget’s Thesaurus, Linguistic Inquiry and Word Count, and WordNet Affect. We then illustrate the application of the method to humorous texts, we show the classes that are derived by using the dominance score, and evaluate the consistency of the classes using correlation measures.

2 Identifying Dominant Word Classes in Text

In this section, we describe a method to calculate a score associated with a given class of words, as a measure of saliency for the given word class inside a collection of texts that share a common property.

We define the *foreground* corpus to be the collection of texts for which we want to determine the dominant word classes. All the texts in the foreground corpus are assumed to share a certain property, e.g., humorous texts, female-authored texts, etc.

We define the *background* corpus as a collection of texts that are neutral and do not have the property shared by the texts in the foreground. The background corpus plays the role of a baseline, with respect to which we can determine the saliency of the word classes in the foreground corpus. A good background corpus should consist of a mix of texts balanced with respect to genre and source, all of which lack the property of the foreground texts. The purpose of seeking different sources for the construction of the background dataset is to avoid the bias that could be introduced by a specific source or genre. We want to model the characteristics of the foreground corpus, and thus we do not want to learn features that could be specific to a single-source background collection.

Given a class of words $C = \{W_1, W_2, \dots, W_N\}$, we define the class coverage in the foreground corpus F as the percentage of words from F belonging to the class C :

$$Coverage_F(C) = \frac{\sum_{W_i \in C} Frequency_F(W_i)}{Size_F}$$

where $Frequency_F(W_i)$ represents the total number of occurrences of word W_i inside the corpus F , and $Size_F$ represents the total size (in words) of the corpus F .

Similarly, we define the class C coverage for the background corpus B :

$$Coverage_B(C) = \frac{\sum_{W_i \in C} Frequency_B(W_i)}{Size_B}$$

The *dominance score* of the class C in the foreground corpus F is then defined as the ratio between the coverage of the class in the corpus F with respect to the coverage of the same class in the background corpus B :

$$Dominance_F(C) = \frac{Coverage_F(C)}{Coverage_B(C)} \quad (1)$$

A dominance score close to 1 indicates a similar distribution of the words in the class C in both the foreground and the background corpus. Instead, a score significantly higher than 1 indicates a class that is dominant in the foreground corpus, and thus likely to be a characteristic of the texts in this corpus. Finally, a score significantly lower than 1 indicates a class that is unlikely to appear in the foreground corpus. Note that if the background corpus is compiled so that it is balanced and mixed across different genres and sources, a score lower than 1 does not indicate a class that is characteristic to the background corpus, but a class that is *not characteristic* to the foreground corpus.

3 Word Classes

We use classes of words as defined in three large lexical resources: Roget’s Thesaurus, Linguistic Inquiry and Word Count, and the six main emotions from WordNet Affect. For each lexical resource, we only keep the words and their corresponding class. Note that some resources include the lemmatised form of the words (e.g., Roget), while others include an inflected form (e.g., LIWC); we keep the words as they originally appear in each resource. Any other information such as morphological or semantic annotations are removed for consistency purposes, since not all the resources have such annotations available.

3.1 Roget

Roget is a thesaurus of the English language, with words and phrases grouped into hierarchical classes. A word class usually includes synonyms, as well as other words that are semantically related. Classes are typically divided into sections, subsections, heads and paragraphs, allowing for various granularities of the semantic relations used in a word class. We only use one of the broader groupings, namely the heads. The most recent version of Roget (1987) includes about 100,000 words grouped into close to 1,000 head classes. Table 1 shows three classes, together with a sample of the words included in these classes.

3.2 Linguistic Inquiry and Word Count (LIWC)

LIWC was developed as a resource for psycholinguistic analysis, by Pennebaker and colleagues [10, 11]. The 2001 version of LIWC includes about 2,200 words and word stems grouped into about 70 broad categories relevant to psychological processes (e.g.,

emotion, cognition). The LIWC lexicon has been validated by showing significant correlation between human ratings of a large number of written texts and the rating obtained through LIWC-based analyses of the same texts. Table 1 shows three LIWC classes along with a set of sample words included in these classes.

3.3 WordNet Affect

WordNet Affect [15] is a resource that was created starting with WordNet [8], by annotating synsets with several emotions. It uses several resources for affective information, including the emotion classification of Ortony [9]. WordNet Affect was constructed in two stages. First, a core resource was built based on a number of heuristics and semi-automatic processing, followed by a second stage where the core synsets were automatically expanded using the semantic relations available in WordNet.

We extracted the words corresponding to the six basic emotions defined by [9], namely anger, disgust, fear, joy, sadness, surprise. We show three of these classes and a few sample words in Table 1.

Class	Words
Roget	
PERFECTION	perfection, faultlessness, lawlessness, impeccability, purity, integrity, chastity
MEDIOCRITY	mediocrity, dullness, indifference, normality, commonness, inferiority
SAFETY	safety, security, surety, assurance, immunity, safeguard, protect, insured
LIWC	
OPTIM(ISM)	accept, best, bold, certain, confidence, daring, determined, glorious, hope
TENTAT(IVE)	any, anyhow, anytime, bet, betting, depending, doubt, fuzzy, guess, hesitant
SOCIAL	adult, advice, affair, anyone, army, babies, band, boy, buddies, calling, comrade
WordNet Affect	
ANGER	wrath, umbrage, offense, temper, irritation, lividity, irascibility, fury, rage
JOY	worship, adoration, sympathy, tenderness, regard, respect, pride, preference, love
SURPRISE	wonder, awe, amazement, astounding, stupefying, dazed, stunned, amazingly

Table 1. Three word classes from each lexical resource, along with sample words.

4 Analysing Humorous Text

As a case study for our method, we analyse the dominant word classes found in humorous text. This follows on previous work on humour recognition using large collections of humorous texts [7], as well as on more recent work including an analysis of the features found in humorous texts [5]. Unlike previous work, where the words found in verbal humour were manually investigated in an attempt to identify more general word classes, the method proposed here is more general and systematic.

4.1 Foreground Corpus: Two Collections of Humorous Texts

There have been only a relatively small number of previous attempts targeting the computational modeling of humour. Among these, most of the studies have relied on small datasets, e.g. 195 jokes used for the recognition of knock-knock jokes [16], or 200 humorous headlines analysed in [2]. Such small collections may not suffice for the robust learning of features of humorous text.

More recently, we proposed a Web-based bootstrapping method that automatically collects humorous sentences starting with a handful of manually selected seeds, which allowed us to collect a large dataset of 16,000 one-liners [6].

In this paper, we use the corpus of one-liners, as well as a second dataset consisting of humorous news articles [5].

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.

Starting with a short seed set consisting of a few one-liners manually identified, the algorithm proposed in [6] automatically identifies a list of webpages that include at least one of the seed one-liners, via a simple search performed with a Web search engine. Next, the webpages 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.

Take my advice; I don't use it anyway.
I get enough exercise just pushing my luck.
I took an IQ test and the results were negative.
A clean desk is a sign of a cluttered desk drawer.
Beauty is in the eye of the beer holder.

Fig. 1. Sample examples of one-liners

Two iterations of the bootstrapping process, started with a small seed set of ten 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, the resulting dataset contains 16,000 one-liners, which are used in the experiments reported in this paper. The one-liners humor style is illustrated in Figure 1, which shows five examples of such one-sentence jokes.

Humorous News Articles. In addition to the one-liners, we also use a second dataset consists of daily stories from the newspaper “The Onion” – a satiric weekly publication with ironic articles about current news, targeting in particular stories from the United States. It is known as “the best satire magazine in the U.S.” (Andrew Hammel, German Joys, <http://andrewhammel.typepad.com>) and “the best source of humour out there” (Jeff Grienfield, CNN senior analyst, <http://www.ojr.org/>).

Canadian Prime Minister Jean Chrétien and Indian President Abdul Kalam held a subdued press conference in the Canadian Capitol building Monday to announce that the two nations have peacefully and sheepishly resolved a dispute over their common border. Embarrassed Chrétien and Kalam restore diplomatic relations. "We are – well, I guess proud isn't the word – relieved, I suppose, to restore friendly relations with India after the regrettable dispute over the exact coordinates of our shared border," said Chrétien, who refused to meet reporters' eyes as he nervously crumpled his prepared statement. "The border that, er... Well, I guess it turns out that we don't share a border after all."

Fig. 2. Sample news article from “The Onion”

All the articles published during August 2005 – March 2006 were collected, which resulted in a dataset of approximately 2,500 news articles. After cleaning the HTML tags, all the news articles that fell outside the 1000–10,000 character length range were removed. This process led to a final dataset of 1,125 news stories with humorous content. Figure 2 shows a sample article from this dataset. This data set was previously used in [5].

4.2 Background corpus

In order to create a background corpus, we compiled a dataset consisting of a mix of non-humorous sentences from four different sources: (1) *Reuters* titles, extracted from news articles published in the Reuters newswire over a period of one year (8/20/1996 – 8/19/1997); (2) *Proverbs* extracted from an online proverb collection; (3) *British National Corpus (BNC)* sentences; and (4) sentences from the *Open Mind Common Sense* collection of commonsense statements.

4.3 Dominant Word Classes in Humorous Text

All the word classes from the resources described in Section 3 were ranked according to the dominance score calculated with formula 1. Those classes that have a high score are the classes that are dominant in humorous text. Table 2 shows the top classes found according to each lexical resource, along with their dominance score and a few sample words.

Class	Score	Sample words
Roget		
ANONYMITY	3.48	you, person, cover, anonymous, unknown, unidentified, unspecified
ODOR	3.36	nose, smell, strong, breath, inhale, stink, pong, perfume, flavor
SECRECY	2.96	close, wall, secret, meeting, apart, ourselves, security, censorship
WRONG	2.83	wrong, illegal, evil, terrible, shame, beam, incorrect, pity, horror
UNORTHODOXY	2.52	error, non, err, wander, pagan, fallacy, atheism, erroneous, fallacious
PEACE	2.51	law, rest, order, peace, quiet, meek, forgiveness, soft, calm, spirit
OVERESTIMATION	2.45	think, exaggerate, overestimated, overestimate, exaggerated,
INTUITION INSTINCT	2.45	drive, feel, idea, sense, blind, feeling, knowledge, natural, tact
INTELLECTUAL	2.41	woman, brain, student, genius, amateur, intellect, pointy, clerk
DISARRANGEMENT	2.18	trouble, throw, ball, bug, insanity, confused, upset, mess, confuse
LIWC		
YOU	3.17	you, thou, thy, thee, thin
I	2.84	myself, mine
SWEAR	2.81	hell, ass, butt, suck, dick, arse, bastard, sucked, sucks, boobs
SELF	2.23	our, myself, mine, lets, ourselves, ours
SEXUAL	2.07	love, loves, loved, naked, butt, gay, dick, boobs, cock, horny, fairy
GROOM	2.06	soap, shower, perfume, makeup
CAUSE	1.99	why, how, because, found, since, product, depends, thus, cos
SLEEP	1.96	bed, wake, asleep, woke, nap, wakes, napping, waking
PRONOUN	1.84	you, they, his, them, she, her, him, nothing, our, its, themselves
HUMANS	1.79	man, men, person, children, human, child, kids, baby, girl, boy
WordNet Affect		
SURPRISE	3.31	stupid, wonder, wonderful, beat, surprised, surprise, amazing, terrific

Table 2. Dominant word classes from each lexical resource, along with sample words.

5 Evaluation

To evaluate the dominance scores obtained for the word classes, we measure the correlation between the scores derived by using different humorous data sets. Since we are interested in a consistent ranking for the dominance scores when derived from different corpora, we use the Spearman correlation metric to measure ranking consistency.

We evaluate the correlation for three data pairs. First, the one-liners data set is randomly split into two non-intersecting data sets consisting of 8,000 one-liners each. In Table 3, this data set pair is labeled *one-liners vs. one-liners*. Second, the humorous news articles set is split into two separate data sets of approximately 550 news articles each (*news articles vs. news articles*). Finally, the last data set pair measures correlation across corpora: dominance scores derived from the entire corpus of 16,000 one-liners compared to the scores obtained for the entire corpus of 1,125 news articles (*one-liners vs. news articles*).

Table 3 shows the Spearman correlation measured for the three data set pairs, for the dominance scores obtained for the Roget and LIWC word classes. Not surprisingly, the correlation within the same genre (e.g., one-liners vs. one-liners or news articles vs. news articles) is higher than across genres. However, despite the genre and source

	Roget LIWC	
one-liners vs. one-liners	0.95	0.96
news articles vs. news articles	0.84	0.88
one-liners vs. news articles	0.63	0.42

Table 3. Spearman correlation between word class dominance scores derived for different humorous corpora.

differences between the one-liners and the news articles corpora, the correlation is still strong, significant at $p < 0.01$ level using a two-tailed t-test.

For WordNet Affect, because it includes only six classes, we could not calculate the Spearman correlation, since at least 12 points are required for a reliable correlation metric. Instead, the dominance scores obtained for the six emotion classes are listed in Table 4. As seen in the table, the dominance score rankings obtained for the two different data sets (one-liners and humorous news articles) are similar, with *surprise* being by far the most dominant emotion, with a score of 3.31 obtained for the one-liners and 1.91 for the humorous news articles. The *disgust* emotion has also a score larger than 1, but not as significant as the surprise emotion.

Emotion	One-liners	News articles
ANGER	0.81	0.73
DISGUST	1.33	1.16
FEAR	1.12	0.97
JOY	1.13	0.83
SADNESS	0.97	0.85
SURPRISE	3.31	1.91

Table 4. Dominance scores for the six emotions in WordNet Affect.

For a second evaluation, we also compare the high dominance classes obtained with our method with the observations made in previous work concerning the features of humorous text. For instance, [7] observed that sexual vocabulary was frequently used in humour. This matches the SEXUAL class that we also identified as dominant. Similarly, [5] found human-centered vocabulary and negative polarity as important characteristics of humorous texts. These features correspond to several dominant classes that we automatically identified: YOU, I, SELF, HUMANS (human-centered vocabulary), and WRONG, UNORTHODOXY, DISARRANGEMENT (negative polarity). Swearing vocabulary (among our classes: SWEAR) was also found useful for humour recognition [13]. Finally, surprise [12, 1] was previously identified as one of the elements most frequently encountered in humour. We also found this class as having a high dominance score in humorous texts.

Those observations however were mostly empirical, based on a manual analysis of the words frequently encountered in humour. Instead, our method allows us to systematically identify the word classes that are dominant in humorous texts, which implies

increased coverage (a larger number of word classes can be identified), robustness (the same method can be applied to corpora of different sizes), and portability (besides humour, the method can be used to characterize any other types of texts).

6 Conclusions

In this paper, we proposed a method for “linguistic ethnography,” which automatically identifies the most dominant word classes in text. By using this method, we can take a step further toward the systematic characterization of texts sharing a common property, such as humorous texts or texts authored by same gender writers.

Using humour as a case study, we showed that the automatically learned word classes are reliable, and they correlate well across different corpora sharing the same humorous property. Moreover, we showed that several of the classes automatically identified correspond to previous empirical observations that were based on manual analysis of humorous texts.

Despite its simplicity, the method proposed is systematic, robust, and portable, and can be used to automatically characterize any types of texts. In future work, we plan to integrate the automatically derived dominant word classes into a classifier for humour recognition. We also plan to test the applicability of the method to other types of texts.

References

1. ATTARDO, S., AND RASKIN, V. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research* 4, 3-4 (1991).
2. BUCARIA, C. Lexical and syntactic ambiguity as a source of humor. *Humor* 17, 3 (2004).
3. LIU, H., AND MIHALCEA, R. Of men, women, and computers: Data-driven gender modeling for improved user interfaces. In *International Conference on Weblogs and Social Media* (2007).
4. MCCALLUM, A., AND NIGAM, K. A comparison of event models for Naive Bayes text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization* (1998).
5. MIHALCEA, R., AND PULMAN, S. Characterizing humour: An exploration of features in humorous texts. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics* (Mexico City, 2007).
6. MIHALCEA, R., AND STRAPPARAVA, C. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Human Language Technology / Empirical Methods in Natural Language Processing conference* (Vancouver, 2005).
7. MIHALCEA, R., AND STRAPPARAVA, C. Technologies that make you smile: Adding humor to text-based applications. *IEEE Intelligent Systems* 21, 5 (2006).
8. MILLER, G., LEACOCK, C., RANDEE, T., AND BUNKER, R. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology* (Plainsboro, New Jersey, 1993).
9. ORTONY, A., CLORE, G. L., AND FOSS, M. A. The referential structure of the affective lexicon. *Cognitive Science*, 11 (1987).
10. PENNEBAKER, J., AND FRANCIS, M. Linguistic inquiry and word count: LIWC. Erlbaum Publishers.

11. PENNEBAKER, J., AND KING, L. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77 (1999), 1296–1312.
12. RASKIN, V. *Semantic Mechanisms of Humor*. Kluwer Academic Publications, 1985.
13. SJOBERGH, J., AND ARAKI, K. Recognizing humor without recognizing meaning. In *Proceedings of the Workshop on Cross-Language Information Processing (2007)*.
14. STRAPPARAVA, C., AND MIHALCEA, R. Learning to identify emotions in text. In *Proceedings of the ACM Conference on Applied Computing ACM-SAC 2008 (Fortaleza, Brazil, 2008)*.
15. STRAPPARAVA, C., AND VALITUTTI, A. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (Lisbon, 2004)*.
16. TAYLOR, J., AND MAZLACK, L. Computationally recognizing wordplay in jokes. In *Proceedings of CogSci 2004 (Chicago, August 2004)*.
17. WIEBE, J., AND RILOFF, E. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005) (invited paper) (Mexico City, Mexico, 2005)*.