

Evaluating Humorous Features: Towards a Humour Taxonomy^{*}

Antonio Reyes, Paolo Rosso, and Davide Buscaldi

Natural Language Engineering Lab - ELiRF
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Spain
areyes,proso,dbuscaldi@dsic.upv.es

Abstract. The importance of the analysis of processes related to cognitive phenomena through Natural Language Processing techniques is acquiring a greater relevance every day. Opinion Mining, Sentiment Analysis or Automatic Humour Recognition are a sample about how this kind of research works grows. In this paper we focus on the study of how the features that define a corpus of humorous data (one-liners) may be used for obtaining a set of parameters that allow us to build a primitive taxonomy of humour. We analyse, through several experiments, a set of well-known features defined in the literature, besides a set of new ones, in order to determine the importance of each one for a humour taxonomy. An evaluation of all the features was performed by means of an automatic classification task over a collection of humorous blogs. The results obtained show that some of the features may represent elemental information for the purpose of creating a humour taxonomy.

1 Introduction

The analysis of phenomena related to cognitive processes is a very important trend in Natural Language Processing research. The study of characteristics linked to the human behaviour such as emotions or mood. [3] is a sample about the importance of this kind of research that leads to the exploration of more abstract spheres that acquire a representation at linguistic level. On that, the investigations in areas such as Opinion Mining [12], Sentiment Analysis [21] or Computational Humour [15], have shown how to address the challenge that these tasks suppose through the use of machine learning or pattern recognition techniques, besides the use of linguistic resources.

In this framework, we present a research work which focuses on the analysis of a set of features that define a corpus of humorous one-liners [17, 18] in order to obtain parameters that allow us to build a primary taxonomy of humour. On this subject, we aim at investigating how the features that have been considered as

^{*} The TEXT-ENTERPRISE 2.0 (TIN2009-13391-C04-03) research project has partially funded this work. The National Council for Science and Technology (CONACyT - Mexico) has funded the research work of Antonio Reyes.

descriptors of these one-liners, besides a set of new features, may be employed for depicting the concepts that underly the phenomenon of humour. We evaluated the hypothesis that this objective implies going through a classification task over a collection of blogs whose main topic is humour.

The outline of this paper is organized as follows. Section 2 describes the research works on Computational Humour, focusing on Automatic Humour Recognition. Section 3 underlines the initial assumptions and the aim of our research. Section 4 explains all the experiments we carried out. Section 5 presents the evaluation and the discussion of the results obtained. Finally, in Section 6 we draw some conclusions and address the further work.

2 Computational Humour

Humour is one of the most amazing and fuzzy aspects of the human behaviour that, despite its common practice, it is still not clearly defined [25, 2]. Cognitive features as well as cultural knowledge, for instance, are some of the variables that must be analysed in order to obtain some answers about how humour works. Factors like these turn humour in a subjective and fuzzy entity that changes according to cultures, societies, persons or mood. Nonetheless, its automatic processing seems promising. For instance, as part of the Affective Natural Processing tasks, the Computational Humour area has demonstrated that this characteristic may be automatically handled from two angles: generation and recognition. The first one builds models from recurrent templates taken into account linguistic patterns. For instance, the research work in [6] and [7] showed the importance of phonetic and semantic patterns as features for automatically generating punch lines. Likewise, the European project “HaHacronym” [27] demonstrated how the incongruity and the opposite senses are relevant triggers for generating humorous meanings.

The recognition task, which is defined as the identification and extraction of humour descriptors from the analysis of internal and external features in textual information, has shown that, given a collection of humorous samples, it is possible to learn the discriminating features that turn these samples into humorous data. The research works in [17–19, 26, 8] have contributed to provide a set of different features that define their data as humorous¹. Some of them are ambiguity, irony, adult slang, antonymy, human centric vocabulary, negative orientation, bag of words, n-grams or professional communities. Although in [19] the authors have experimented with humorous news articles, the others works have focused in the analysis of one-liners, which are short humorous structures that produce their comic effect with few words. The results reported by these authors are encouraging, despite the one-liners whose characteristics suppose to learn more complex features in order to recognise whether an input is humorous or not. For instance, let us consider the example (a):

¹ All these works have focused on verbal humour, which refers to the humour that is expressed linguistically [2].

- (a) Children in the back seats of cars cause accidents, but accidents in the back seats of cars cause children.

The humorous effect in this sentence is caused by the interrelation of opposed concepts given the focused elements in the syntax, i.e., the subjects *children* and *accidents*, respectively. This information is not at surface level and it is necessary to find strategies and methodologies that extract and represent the knowledge that is not given a priori and that determines the relations that turn a sentence into humorous or serious. Thus, in order to obtain the knowledge for identifying what are the features that best describe the patterns that produce humour, Automatic Humour Recognition relies on models and resources that take advantage of linguistic information for describing features such as antonym, alliteration or ambiguity.

3 Humour Features

As noted in the previous section, the features obtained through the analysis of one-liners have allowed to automatically discriminate humorous from non humorous data with a high percentage of accuracy. That is why we think that these features may be employed for describing another kind of humorous data beyond only one-liners. This means that the underlying concepts that trigger the humorous effect in the one-liners are common to any kind of joke and consequently, to any kind of verbal humour. For instance, a feature such as adult slang is not an one-liner privative feature but it appears in other data such as a punning riddle or a discussion about humour. Therefore, we think that the set of features that have showed its effectiveness for discriminating humorous from serious data could represent elemental concepts that may be used to build a general taxonomy of humour.

In this framework, our objective is to assess some of the most relevant features reported in the literature as general descriptors of humour, specifically one-liners humour, in order to find some hints for conceptualising a humour taxonomy. We expect that such features provide information for classifying any kind of verbal humour.

We addressed this issue through a feature extraction task and an automatic classification process. That is, given the one-liners corpus used by Mihalcea and Strapparava in their experiments, we extract the main features they reported for characterising humour. Besides those features, we performed some experiments over the same corpus in order to retrieve other kind of discriminating characteristics. Afterward, using the whole set of features, we evaluate the importance of each one of them through an automatic classification process over a test set composed by a collection of blogs whose main topic is humour.

The features we considered in this research work, according to the results depicted in [17–19], were:

1. *stylistic features*, focusing on adult slang;
2. *human centric vocabulary*, focusing on personal pronouns;

3. *human centeredness*, focusing on social relationships;
4. *polarity*, focusing on the positive or negative orientation of the data.

Alongside these features, we took into account the following aspects:

1. *wh – phrases*, focusing on interrogative pronouns;
2. *nationalities*, focusing on adjectives of nations;
3. *keyness*, focusing on the extraction of the most representative subjects of the data.
4. *discriminative items*, focusing on the words that belong to a same cluster;
5. *ambiguity*, focusing on the sense dispersion of the words.

These features were tested using the Naïve Bayes and the Multinomial Logistic Regression (MLR) classifiers [28]. The data sets and the experiments performed are described in the following section.

4 Experiments on Features Extraction

The experiments reported in this section are divided in two phases: (i) in the first one we extracted from the one-liners corpus all the features we described in the previous section (Sections 4.2 to 4.10); (ii) in the second one, we automatically labeled each blog according to these features (Section 4.11).

4.1 Data sets

The corpus of one-liners was automatically collected from the web through a bootstrapping process described in [16]. It contains 16,000 one-liners. This corpus, as we have already mentioned, was the main database for extracting the features that define humour. On the other hand, we decided to test the set of features over a collection of blogs because, being a heterogeneous site where humour is not only represented by one-liners but by jokes, gags, punning riddles or even by humorous and serious discussions about humour, they are a good source to study any type of verbal humour. The collection of blogs we used was retrieved from the web through an automatic request to Google search engine. Keywords such as *punch line*, *humour*, *joke*, *funny*, *laughter*, *laugh line*, *gag*, *gag line*, *tag line* and so on, were the seeds for retrieving the results. A total of 200 humorous blogs integrated the collection². It is necessary to mention that, given the automatic process, it was possible that the blogs have information not related to humour. Thus, for minimising the noise, the collection was evaluated according to the measures proposed in [22] for estimating features such as domain broadness, shortness, stylometry and structure on corpora³. In Table 1 we

² Some statistics about the collection are: 23,363 types; 168,100 tokens; tokens/types relation = 7.19. This collection, enhanced with more blogs, will be made soon available.

³ Before measuring these features on the collection, we eliminated the stopwords, enhancing the list with words such login, username, copyright, next, top, etc., in order to delete information not related to the topic of the request.

show the results obtained employing the tool that implements all these measures: the Watermarking Corpora On-line System (WaCOS)⁴.

Table 1. Blogs representative features

Feature	Result
Domain broadness	Wide
Shortness	Short texts
Stylometry	General Language Style
Structure	Complex

The information in this table indicates:

- i) the collection is not restricted only to one topic (broadness), for instance politics, but several ones. This impacts on the fact of having different kinds of discourses expressing humorous information;
- ii) the blogs we will classify are written without following a standard pattern (stylometry) whereby they do not share a surface similarity (structure) that implies a trend in the way that humour is expressed.

According to this information, the collection seems to be wide enough for covering a broad spectrum of humour and ways in which it is linguistically expressed. Therefore, we considered the collection as valid for our purposes⁵.

4.2 Stylistic Features

According to [18], the sexual information from example (b) represents one of the most relevant features for discriminating humour. Therefore, we reproduced their experiment about *adult slang* extracting all the words labeled with the tag “sexuality” in WordNet Domains v. 3.2 [4] for getting our first feature⁶.

(b) Artificial Insemination: procreation without recreation.

4.3 Human Centric Vocabulary

One of the most important features reported in the literature is the presence of words that make reference to “human-related scenarios”. For instance, the pronoun *You* appears with a frequency greater than 25% in the one-liners whereas the pronoun *I* occurs 15% [18]. That is why we selected personal pronouns, specifically, first, second and third (masculine and feminine) singular, for integrating the set of elements in this feature. Besides them, we included their correspondent reflexive pronouns for getting a broader coverage on this feature.

⁴ This tool is available at: <http://users.dsic.upv.es/grupos/nle/demos.html>.

⁵ In Appendix A we show a sample about the information that appears in the blogs.

⁶ At the end of the Section 4.10 we include a graphic that represents the statistics of all the experiments.

4.4 Human Centeredness

As reported in [19], human centeredness tends to find out what are the most discriminating features in humorous data given four a priori semantic classes: persons, social groups, social relations and personal pronouns. We only selected the most salient one for representing this feature: *social relations*. The items that integrated this class were chosen as the authors reported, i.e., retrieving all the nouns that deal the synsets *relation*, *relationship* and *relative* in [20].

4.5 Polarity

According to [19], the negative orientation is a very important discriminating feature when speaking of humour. Therefore, in order to verify this assertion, we automatically labeled the corpus with a public tool for Sentiment Analysis: Java Associative Nervous Engine (Jane16)⁷. The underlying algorithm of this tool creates a model of positive and negative words and sentences which are crawled in Internet. Depending on their occurrence, they are ranked and a weight is assigned to each of them. In this way, the positive and negative data sets are retrieved. The labeling process matches the information provided with that one in the database and computes the occurrence and weights for assigning the correspondent label.

Furthermore we used [11]⁸ over a set of elements that we found in the experiments reported in the clustering section in order to assess the role that a small set of words could play in the overall objective of this work.

4.6 WH-phrases

A common humorous structure handled in Computational Humour is the punning riddle [5]. This kind of jokes takes advantage of syntactic recurrent templates: the *wh - phrases*. These structures are syntactic constituents that are characterised by question words or entire phrases. A humorous example of this structures appears in (c).

(c) *What* are the 3 words you never want to hear while making love? Honey, I'm home!

The quantity of jokes that relies on this template is substantial, at least in the corpus of one-liners. That is why we used the interrogative pronouns for representing a humorous feature that may provide elements about the most profiled topics in the jokes.

⁷ This tool is available at <http://opusintelligence.com/download.jsp>.

⁸ This tool is available at <http://sentiwordnet.isti.cnr.it>.

4.7 Nationalities

The professional communities are a set of elements that have been associated to humour [17, 18]. For instance, the one-liner that appears in (d) is a sample about this assumption.

(d) *Parliament* fighting inflation is like the *Mafia* fighting crime.

Instead of using this category, we employed a wordlist with adjectives of nationalities for noticing whether or not the information about toponyms is as relevant as the professional communities in the definition of a humour taxonomy.

4.8 Keyness

We extracted the most representative items from the one-liners corpus according to their *keyness* value. This measure estimates the keyness through the frequency comparison of each word in a corpus against the frequency of the same word in a reference one. The values are computed taking into account the Log Likelihood test [10].

For retrieving the items with greater keyness value, we generated a list with all the words from the one-liners corpus, except the stopwords. Likewise, in order to obtain a reference corpus, we used the 3-gram section of [9]. Given both corpora, we computed the keyness.

Furthermore, we added the items that Jane16's service *scope* identified as keywords in the one-liners corpus. Some of the representative items according to the keyness value and Jane16 are: *mad*, *paranoid*, *sick*, *hell* and *mistake*.

4.9 Discriminative Items

In order to identify how much similar the items in the one-liners are and be able to determine a set of discriminative features, we carried out five different clustering experiments. We employed two tools: Cluto and SenseClusters. Cluto⁹ is a set of algorithms that “operate either directly in the objects feature space or in the objects similarity space” [13], maximising or minimising a criterion function over the solution. On the other hand, SenseClusters¹⁰ is a package that integrates Cluto's algorithms besides a set of tools for identifying similar contexts. SenseClusters works seeking “lexical features to build first and second order representations of contexts” [14].

In the first experiment we worked with Cluto selecting vector space, direct clustering method, H2 criterion function and cosine similarity function as discriminative features¹¹. The number of requested clusters was 20. Figure 1 shows how the most discriminating features are distributed in each one of the 20 clusters.

⁹ Available at <http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview>.

¹⁰ Available at <http://www.d.umn.edu/~tpederse/senseclusters.html>.

¹¹ For a detailed explanation about the meanings of these parameters, see [13].

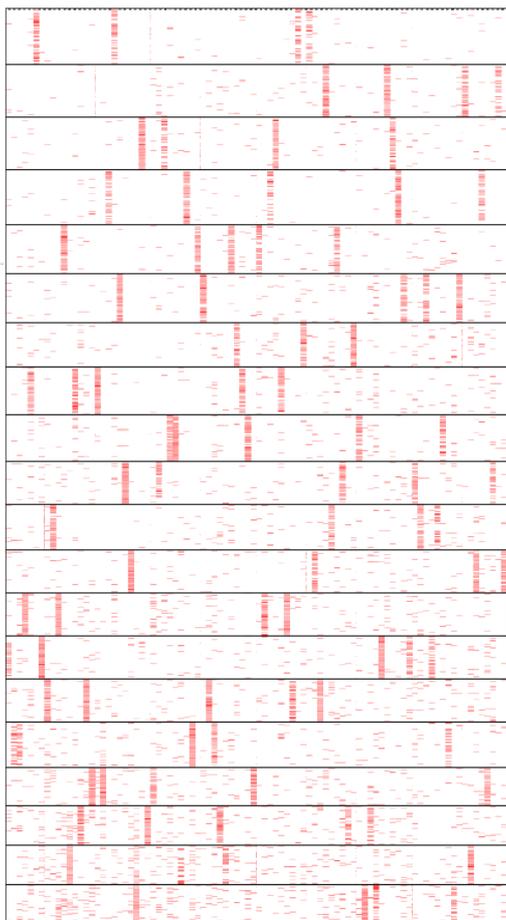


Fig. 1. Items distribution. The rows represent each cluster and the dots indicate how the items are distributed in the cluster

The rest of the clustering experiments were carried out with SenseClusters. In each one of them we varied the parameters and the number of the requested clusters. In Table 2 we summarise the processes¹².

Table 2. SenseCluster parameters per experiment

Exp.	Space	Cl. Method	Cr. Function	LSA	Order	Cluststop	Clusters
1	Vector	RB/Direct	UPGMA	Yes	Bi/Co	All	2
2	Similarity	Agglo/RBR/Graph	H2	Yes	Uni/Bi	Gap	2
3	Vector	Direct	H2	Not	Co	None	20
4	Vector	RB	I2	Not	Bi	Pk	27

The set of all discriminating items generated with both Cluto and SenseCluster tools were recorded in a wordlist for removing all duplicated ones. The remaining items were first labeled with their POS tags using Freeling¹³, described in [1], and then, with their positive, neutral or negative polarity tag according to SentiWordNet¹⁴. This list was used for a second polarity labeling (see Section 4.5) over the collection of blogs.

4.10 Ambiguity

In several research works it has been pointed out that humour takes advantage of linguistic ambiguity for producing its funny effect [17, 18, 26, 23, 24]. That is why we performed an experiment for verifying how much valuable information could provide a humour characterization through the representation of semantic ambiguity. The experiment consisted in measuring the *sense dispersion* for each noun of the one-liners. This measure is based on the hypernym distance between synsets, calculated with respect to the WordNet ontology [20]. This distance was calculated using the formula depicted in [23], which appears in (1):

¹² The abbreviations in the table indicate: Exp., number of experiment; Cl. Method, clustering method employed; RB means repeated bisections; Agglo means agglomerative clustering; RBR means repeated bisections globally optimized; Graph means graph partitioning-based clustering; Cr. Function, criterion function employed; LSA, Latent Semantic Analysis representation; Order, represents contexts; Bi means bigrams; Uni means unigrams; Co means co-occurrences; Cluststop, cluster stopping measure; Gap means adapted gap statistics; Pk means pk measures.

For a detailed explanation consult <http://search.cpan.org/dist/Text-SenseClusters>

¹³ Available at <http://garraf.epsevg.upc.es/freeling>.

¹⁴ In Appendix B we provide a list with the first 50 discriminative items labeled with their POS and polarity tags.

$$\delta(w_s) = \frac{1}{P(|S|, 2)} \sum_{s_i, s_j \in S} d(s_i, s_j) \quad (1)$$

where S is the set of synsets (s_1, \dots, s_n) for the word w ; $P(n, k)$ is the number of permutations of n objects in k slots; and $d(s_i, s_j)$ is the length of the hypernym path between synsets (s_i, s_j) . For instance, the noun *killer* has four synsets¹⁵. Taking only into account the synsets s_i and s_j , we obtain as first common hypernym *physical entity*. The number of nodes to reach this hypernym is 6 and 2, respectively. Thus, the dispersion of *killer* is the sum of those distances divided by 2. Now, considering all its synsets, we obtain six possible combinations whose distance among them and their first common hypernym generates a dispersion of 6,83. The formula in (2) shows how the total dispersion per noun is calculated:

$$\delta_{TOT} = \sum_{w_s \in W} \delta(w_s) \quad (2)$$

where W is the set of nouns in the collection N .

The underlying assumption of this measure is to quantify the difference among the senses of a word. This means, that a word with senses that differ significantly is more likely to be used to create humour than a word with senses that differ slightly. The average sense dispersion of the whole set of nouns in the one-liners corpus (which was calculated as: $\bar{\delta}_W = \frac{\delta_{TOT}}{|W|}$) was 7,63.

In the last experiment of the features extraction phase, we calculated the total sense dispersion per isolated noun. The results were recorded in a list for measuring the average sense dispersion in the blogs. In Figure 2 we depict the overall distribution of every feature in terms of number of items.¹⁶

4.11 Feature Representativeness

Once finished the feature extraction task, we investigated how much representative every feature was. To verify such representativeness, we looked for the features in the blogs through a binary distinction: absence/presence. We used the following algorithm:

1. Let (i_1, i_n) be the items that define every feature f_h .
2. Let (b_1, b_m) be the collection of blogs.
3. If any i_n occurs in b_m with frequency ≥ 4 , then f_h was a representative feature for b_m .

Besides searching the representativeness of each feature, we measured the total sense dispersion for all the blogs according to the formula described in

¹⁵ cf. WordNet v. 3.0.

¹⁶ The items with an * in the figure refers to Human Centric Vocabulary and Human Centeredness, respectively; whereas the @ indicates that the items depicted are those ones from SentiWordNet.

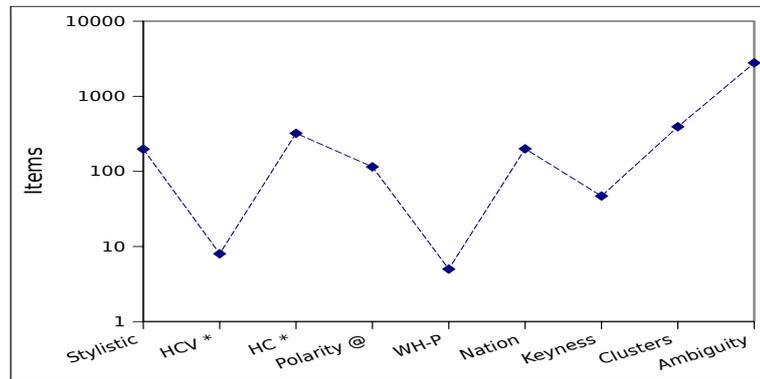


Fig. 2. Items retrieved per feature

Section 4.10. The results obtained are depicted in Figure 3. The graph (a) shows how representative, in terms of presence, every feature was; the plot (b) displays the total sense dispersion per blog.

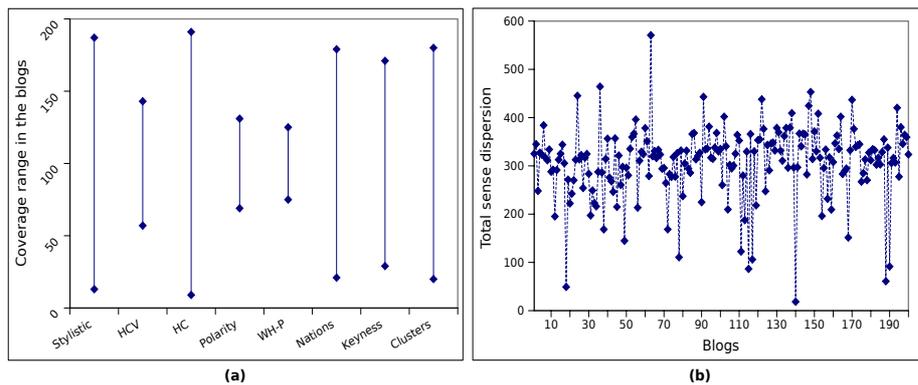


Fig. 3. Feature representativeness results: (a) feature representativeness and (b) sense dispersion per blog

5 Evaluation and Discussion

The classification experiments were carried out in order to understand how difficult is to automatically detect the features and to evaluate whether or not this set may provide hints for building a humour taxonomy. We automatically classified the blogs using the Weka's Bayes and MLR classifiers. We evaluated each classifier employing all the features. The cross validation method was used

as test. It is necessary to mention that, for the polarity feature, besides using the results obtained with Jane16, we incorporated those ones of SentiWordNet, dividing the items retrieved in *Positive* and *Negative*, according to their polarity tag, and removing all of neutral tag (see Appendix B). Also, with respect to sense dispersion, given the differences among ranges, we normalised the values assigning 0 to all the values between 0 and 200; 1 to the values between 201 and 400; and 2 from 401 onwards. The results obtained are displayed in Figure 4. The graph (a) shows the classification accuracy for the state-of-the-art reported features, including SentiWordNet polarity results; whereas the graph (b) shows the classification accuracy for the rest of features.

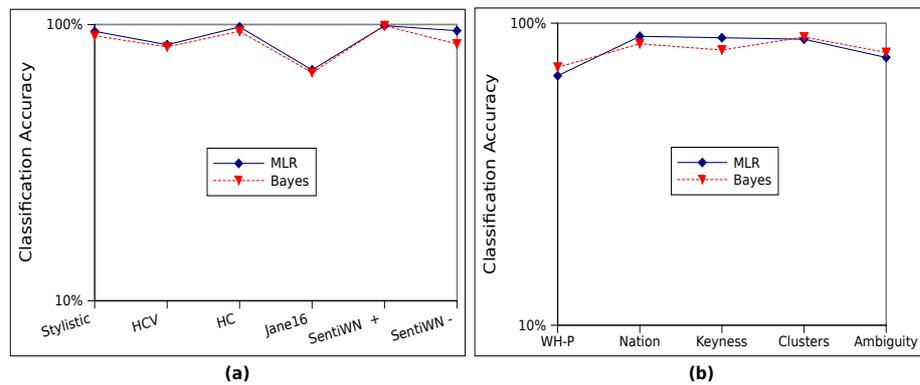


Fig. 4. Classification accuracy for: (a) state-of-the-art and (b) new mined features

According to the information illustrated in these figures, we can notice that there are features that play a more important role for characterising how the humour is expressed by the bloggers. For instance, the WH-phrases seem not to have relevance for representing humour. Likewise, the Jane16's polarity results do not enough reflect the same behaviour reported in [19]. Probably this behaviour is due to the polarity data sets employed. Ambiguity also seems not to show a great impact. However, the experiments must be run over a bigger collection of blogs (or other kind of data) for verifying this behaviour. Moreover, it is evident that the state-of-the-art features have overall better learning curves than the new ones. This fact, besides ratifying the results reported in the literature, establishes that the items of these features may represent basic elements for producing a joke, whereas the items in the new features may help to represent background or adjacent information in a humorous process.

Furthermore, after a manual analysis of the results obtained both with the feature extraction task and the classification process in order to check the viability for having some parameters for our initial purpose, we propose that the features may be divided in two classes: *low level features* and *high level features*. The first one integrates a set of features that represent prototypical information

for characterising humour. This means that there are items that recurrently are used to promote humorous situations and, consequently, they may be identified as common humorous topics. For instance, as has been pointed out in [17,18], jokes about sexuality or self-referential which are represented by elements in the slang, personal pronouns, relationship categories, nationalities or keyness.

Concerning to the second class, we understand as high level features the information that is not clearly related to humorous topics, as the previous ones, but that it is used for producing humour through linguistic strategies. Under this perspective, features such as polarity, discriminative items or ambiguity play as the source that represents this class. For instance, in the example (e), we can behold how humour is generated by information that is not related to any of the state-of-the-art features, not even by a polarity clue. However, its funny effect relies on information beyond the presence of prototypical items but on the use of linguistic ambiguity as a trigger of the humour effect.

(e) Jesus saves, and at today's prices, that's a miracle!

Now, from the two classes above mentioned and with the available features, we think that it is possible to build a general structure that roughly represent the underlying humour's topics. In Figure 5 we illustrate how we conceptualise, from the results obtained in this research work, the humour taxonomy.

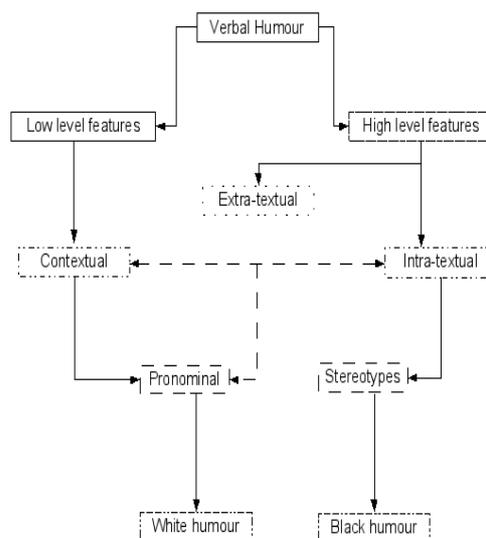


Fig. 5. Towards a primitive humour taxonomy

As noted in this figure, we can identify and extract, according to the items that more recurrently appear, subclasses such as:

1. stereotypes, humour about ethnic groups;
2. pronominal, self-referential humour;
3. white humour, positive polarity orientation;
4. black humour, negative polarity orientation.

And deeper representations such as:

1. contextual, based on items that denote exaggeration, incongruity or absurd;
2. intra-textual, based on linguistic ambiguity;
3. extra-textual, based on pragmatic and cultural information.

6 Conclusions and Further Work

In this paper we evaluated the set of features that have been identified in the main research works on Automatic Humour Recognition as discriminating items between serious and humorous texts (specifically with respect to one-liners), and a set of new ones obtained on the basis of the study of the keyness value, nationalities, discriminating items, and ambiguity, in order to establish some basic parameters for characterising any kind of verbal humour. We aimed at assessing this hypothesis through a classification process over a collection of blogs automatically retrieved from Internet and whose main topic was humour. The results give us some clues about what features have a greater weight for defining humour. Moreover, it seems probable that, through the items that constitute every feature, some of them may be used for conceptualising basic information for building an automatic humour taxonomy.

As further work, besides verifying the behaviour of these features with more data, we aim at investigating what features are more informativeness or whether or not the presence of any of them may change the humorous meaning. Moreover, due to our aim of establishing a verbal humour taxonomy, we plan to verify that this behaviour is similar with data in other languages, (besides other kind of data), in order to take benefit of the insights obtained in this research work for tasks such as machine translation or information filtering.

References

1. Atserias, J., Casas, B., Comelles, E., Gonzalez, M., Padró, L., and Padró, M. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), ELRA. (2006), 48-55.
2. Attardo, S.: Humorous Texts: A semantic and pragmatic analysis. Berlin: Mouton De Gruyter, (2001).
3. Balog, K., Mishne, G., and Rijke, M. Why Are They Excited? Identifying and Explaining Spikes in Blog Mood Levels. In Proceedings 11th Meeting of the European Chapter of the Association for Computational Linguistics. (2006).
4. Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. Revising the Wordnet Domains Hierarchy: semantics, coverage and balancing. In COLING 2004 Multilingual Linguistic Resources. (2004), 94-101.

5. Binsted, K.: Machine humour: An implemented model of puns. PhD thesis. University of Edinburgh, Edinburgh, Scotland, (1996).
6. Binsted, K., Ritchie, G.: Computational rules for punning riddles. In *Humor*. Walter de Gruyter Co. 10: (1997), 25-75.
7. Binsted, K., Ritchie, G.: Towards a model of story puns. In *Humor* 14(3): (2001), 275-292.
8. Buscaldi, D., Rosso, P.: Some experiments in Humour Recognition using the Italian Wikiquote collection. In *Proceedings of the Workshop on Cross Language Information Processing. Int. Conf. WILF-2007, Springer-Verlag, LNAI. 4578: (2007), 464-468.*
9. Brants, T., Franz, A. Web 1T 5-gram corpus version 1. (2006).
10. Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. In *Computational Linguistics*. 19(1): (1993), 61-74.
11. Esuli, A. and Sebastiani, F. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*. (2006), 417-422.
12. Ghose, A., Ipeirotis, P. and Sundararajan, A. Mining using Econometrics: A Case Study on Reputation Systems. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*. (2007).
13. Karypis, G. CLUTO. A Clustering Toolkit. Technical Report 02-017, University of Minnesota, Department of Computer Science.
14. Kulkarni, A., Pedersen, T. SenseClusters: Unsupervised Clustering and Labeling of Similar Contexts. In *Proceedings of the Demonstration and Interactive Poster Session of the 43rd Annual Meeting of the Association for Computational Linguistics*. (2005), 105-108.
15. Mihalcea, R. Multidisciplinary Facets of Research on Humour. In *Proceedings of the Workshop on Cross-Language Information Processing. Int. Conf. WILF-2007, Springer-Verlag, LNAI. 4578: (2007), 412 - 421.*
16. Mihalcea, R., Strapparava, C.: 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)*. 3814: (2005), 84-93.
17. Mihalcea, R., Strapparava, C.: Technologies that make you smile: Adding humour to text-based applications. *IEEE Intelligent Systems*. 21(5): (2006), 33-39.
18. Mihalcea, R., Strapparava, C.: Learning to Laugh (Automatically): Computational Models for Humor Recognition. In *Journal of Computational Intelligence*. 22(2): (2006), 126-142.
19. Mihalcea, R., Pulman, S.: Characterizing Humour: An Exploration of Features in Humorous Texts. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing*. 4394: (2007), 337-347.
20. Miller, G.: Wordnet: A lexical database. In *Communications of the ACM*. 38 (11): (1995), 3941.
21. Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? Sentiment Classification using Machine Learning Techniques, In *Proceedings of EMNLP*. (2002).
22. Pinto, D. On Clustering and Evaluation of Narrow Domain Short-Text Corpora. PhD thesis. Universidad Politcnica de Valencia, Spain, (2008).
23. Reyes, A., Buscaldi, D., Rosso, P.: The Impact of Semantic and Morphosyntactic Ambiguity on Automatic Humour Recognition. In *Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems (NLDB) 2009. Springer-Verlag. Saarbrcken, Germany. (2009).*

24. Reyes, A., Buscaldi, D., Rosso, P.: An Analysis of the Impact of Ambiguity on Automatic Humour Recognition. In Proceedings of the 12th International Conference Text, Speech and Dialogue (TSD) 2009. Springer-Verlag. Plzen, Czech Republic. (2009)
25. Ritchie, G. The Linguistic Analysis of Jokes. Routledge. (2003).
26. Sjöbergh, J., Araki, K.: Recognizing Humor without Recognizing Meaning. In Proceedings of the Workshop on Cross-Language Information Processing. 4578: (2007), 469–476.
27. Stock, O., Strapparava, C.: Hahacronym: A computational humor system. In Demo proc. of the 43rd annual meeting of the Association of Computational Linguistics (ACL05). (2005), 113-116.
28. Witten, I., Frank, E. Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers. Elsevier. (2005).

Appendix A: Sample of Blogs

The following fragments represent the kind of information we found in the blogs.

You don't have to read all the way through. If you just skim read it then you get the general joist of it and it is mediocally funny. The point of the joke (i think) is that it is long and slightly boring (THATS THE POINT!!!!) and this is one joke on this website that i actually felt was slightly funny. If they made the joke shorter then there wouldn't be a joke at all!!!!

An Englishman, an American and an Italian are having a conversation, praising their respective countries.

The Englishman says:

- During the last war we had a ship so large, but so large that for docking maneuvers we needed 24 hours.

The American reply:

- We had a ship so big that to move on it, there was a bus service.

And the Italian:

- This is nothing. We had a ship so large that when at bow the war was over, stern even knew that was started.

A man and his wife were spending the day at the zoo.

She was wearing a loose fitting, pink dress, sleeveless with straps. He was wearing his usual jeans and

T-shirt. As they walked

through the ape exhibit, they passed in front of a large, silverback gorilla. Noticing the wife, the gorilla went crazy. He jumped on the bars, and holding on with one hand and 2 feet he grunted and pounded his chest with his free hand. He was

obviously excited at the pretty lady in the pink dress. The husband, noticing the excitement, thought this was funny. He suggested that his wife tease the poor fellow some more by puckering her lips and wiggling her bottom. She played along and the gorilla got even more excited, making noises that would wake the dead. Then the husband suggested that she let one of her straps fall to show a little more skin. She did and the gorilla was about to tear the bars down. Now show your thighs and sort of fan your dress at him, he said. This drove the gorilla absolutely crazy, and he started doing flips. Then the husband grabbed his wife, ripped open the door to the cage, flung her in with the gorilla and slammed the cage door shut. Now. Tell him you have a headache.

10. The final piece of advice is writing humor takes time. To excel in humor is a lifetime job, and is not something that you can learn in a day or two. Don't think you can read a joke book and start writing funny stuff an hour later. You will have to teach yourself how to be funny. The process is mostly by trial and error, observing other people's comical situations, mistakes, laughing and applying it on yourself, etc. No one can teach you exactly how to write something funny, but the possibilities of creating humor on anything and everything are limitless.

Many companies hold information meetings in the office is not practicing humor, because they do not want to have one of the workers who will be offended. However, at the time the company can cross boundaries on what is acceptable and not acceptable.

Part of the problem with people telling funny jokes or humor is not acceptable is that if someone can not enjoy the job itself in the workplace will be a drab and unhappy workers.

Appendix B: Discriminating Items

Table 3 shows the 50 most discriminating items, according to their POS and polarity tags¹⁷.

¹⁷ When an item belongs to different synsets, it was assigned to the first synset polarity tag according to its POS tag.

Table 3. The 50 most discriminative items in the Mihalcea and Strapparava's one-liners corpus

Positive		Neutral		Negative	
Item	POS	Item	POS	Item	POS
damned	J	circular	J	bad	J
easy	J	foolish	J	common	J
funny	J	front	J	dark	J
high	J	future	J	dead	J
hot	J	green	J	dull	J
meeek	J	homosexual	J	free	J
nice	J	indecisive	J	futile	J
perfect	J	irish	J	hilarious	J
positive	J	lethal	J	impossible	J
real	J	married	J	inverse	J
close	J	middle	J	mad	J
good	J	own	J	negative	J
weak	J	personal	J	old	J
fine	J	photographic	J	paranoid	J
wise	J	proportional	J	sick	J
art	N	remote	J	silent	J
bag	N	suitable	J	stupid	J
care	N	unanimous	J	wrong	J
chance	N	more	J	animal	N
education	N	hard	J	bomb	N
energy	N	little	J	bumper	N
eye	N	usual	J	code	N
fault	N	action	N	difference	N
freedom	N	advance	N	dream	N
fun	N	age	N	fiction	N
genius	N	air	N	habit	N
home	N	alcohol	N	hell	N
ignorance	N	amount	N	hurry	N
important	N	application	N	hydrogen	N
law	N	arrest	N	matter	N
license	N	ass	N	mistake	N
line	N	bar	N	reason	N
mind	N	basket	N	season	N
sharewar	N	bathroom	N	shake	N
strength	N	bed	N	stupidity	N
word	N	beer	N	system	N
die	V	being	N	lie	V
lose	V	bite	N	telekinesis	N
raise	V	blood	N	tourist	N
speed	V	body	N	trouble	N
teach	V	boss	N	worth	N
create	V	box	N	clean	V
see	V	brain	N	forget	V
call	V	bread	N	keep	V
feel	V	bulb	N	succeed	V
learn	V	butter	N	hurt	V
think	V	car	N	kill	V
understand	V	cat	N	missquote	V
		cancel	N	shoot	V
		change	N	suspect	V