"Let Everything Turn Well in Your Wife": Generation of Adult Humor Using Lexical Constraints

Alessandro Valitutti
Department of Computer Science
and HIIT
University of Helsinki, Finland

Hannu Toivonen
Department of Computer Science
and HIIT
University of Helsinki, Finland

Antoine Doucet
Normandy University – UNICAEN
GREYC, CNRS UMR–6072
Caen, France

Jukka M. Toivanen
Department of Computer Science
and HIIT
University of Helsinki, Finland

Abstract

We propose a method for automated generation of adult humor by lexical replacement and present empirical evaluation results of the obtained humor. We propose three types of lexical constraints as building blocks of humorous word substitution: constraints concerning the similarity of sounds or spellings of the original word and the substitute, a constraint requiring the substitute to be a taboo word, and constraints concerning the position and context of the replacement. Empirical evidence from extensive user studies indicates that these constraints can increase the effectiveness of humor generation significantly.

1 Introduction

Incongruity and taboo meanings are typical ingredients of humor. When used in the proper context, the expression of contrasting or odd meanings can induce surprise, confusion or embarrassment and, thus, make people laugh. While methods from computational linguistics can be used to estimate the capability of words and phrases to induce incongruity or to evoke taboo meanings, computational generation of humorous texts has remained a great challenge.

In this paper we propose a method for automated generation of adult humor by lexical replacement. We consider a setting where a short text is provided to the system, such as an instant message, and the task is to make the text funny by replacing one word in it. Our approach is based on careful introduction of incongruity and taboo words to induce humor.

We propose three types of lexical constraints as building blocks of humorous word substitution. (1) The *form constraints* turn the text into a pun. The constraints thus concern the similarity of sounds or spellings of the original word and the substitute. (2) The *taboo constraint* requires the substitute to be a taboo word. This is a well-known feature in some jokes. We hypothesize that the effectiveness of humorous lexical replacement can be increased with the introduction of taboo constraints. (3) Finally, the *context constraints* concern the position and context of the replacement.

Our assumption is that a suitably positioned substitution propagates the *tabooness* (defined here as the capability to evoke taboo meanings) to phrase level and amplifies the semantic contrast with the original text. Our second concrete hypothesis is that the context constraints further boost the funniness.

We evaluated the above hypotheses empirically by generating 300 modified versions of SMS messages and having each of them evaluated by 90 subjects using a crowdsourcing platform. The results show a statistically highly significant increase of funniness and agreement with the use of the humorous lexical constraints.

The rest of this paper is structured as follows. In Section 2, we give a short overview of theoretical background and related work on humor generation. In Section 3, we present the three types of constraints for lexical replacement to induce humor. The empirical evaluation is presented in Section 4. Section 5 contains concluding remarks.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 243–248,
Sofia, Bulgaria, August 4-9 2013.

2 Background

Humor, Incongruity and Tabooness A set of theories known as *incongruity theory* is probably the most influential approach to the study of humor and laughter. The concept of incongruity, first described by Beattie (1971), is related to the perception of incoherence, semantic contrast, or inappropriateness, even though there is no precise and agreed definition. Raskin (1985) formulated the incongruity concept in terms of *script opposition*. This has been developed further, into the *General Theory of Verbal Humor* (Attardo and Raskin, 1991). A cognitive treatment of incongruity in humor is described by Summerfelt et al. (2010).

One specific form of jokes frequently discussed in the literature consists of the so called *forced reinterpretation jokes*. E.g.:

> *Alcohol isn't a problem, it's a solution...*
> *Just ask any chemist.*

In his analysis of forced reinterpretation jokes, Ritchie (2002) emphasises the distinction between three different elements of the joke processing: CONFLICT is the initial perception of incompatibility between punchline and setup according to the initial obvious interpretation; CONTRAST denotes the perception of the contrastive connection between the two interpretations; while INAPPROPRIATENESS refers to the intrinsic oddness or tabooness characterising the funny interpretation. All three concepts are often connected to the notion of incongruity.

In his integrative approach to humor theories, Martin (2007) discusses the connection between tabooness and incongruity resolution. In particular, he discusses the *salience hypothesis* (Goldstein et al., 1972; Attardo and Raskin, 1991), according to which "the purpose of aggressive and sexual elements in jokes is to make salient the information needed to resolve the incongruity".

Humor Generation In previous research on computational humor generation, puns are often used as the core of more complex humorous texts, for example as punchlines of simple jokes (Raskin and Attardo, 1994; Levison and Lessard, 1992; Venour, 1999; McKay, 2002). This differs from our setting, where we transform an existing short text into a punning statement.

Only few humor generation systems have been empirically evaluated. The JAPE program (Binsted et al., 1997) produces specific types of punning riddles. HAHAcronym (Stock and Strapparava, 2002) automatically generates humorous versions of existing acronyms, or produces a new funny acronym, starting with concepts provided by the user. The evaluations indicate statistical significance, but the test settings are relatively specific. Below, we will present an approach to evaluation that allows comparison of different systems in the same generation task.

3 Lexical Constraints for Humorous Word Substitution

The procedure gets as input a segment of English text (e.g.: *"Let everything turn well in your life!"*). Then it performs a single word substitution (e.g: *'life'* $\rightarrow$ *'wife'*), and returns the resulting text. To make it funny, the word replacement is performed according to a number of lexical constraints, to be described below. Additionally, the text can be appended with a phrase such as "*I mean 'life' not 'wife'.*" The task of humor generation is thus reduced to a task of lexical selection. The adopted task for humor generation is an extension of the one described by Valitutti (2011).

We define three types of lexical constraints for this task, which will be described next.

3.1 Form Constraints

Form constraints (FORM) *require that the original word and its substitute are similar in form.* This turns the text given as input into a kind of *pun*, "text which relies crucially on phonetic similarity for its humorous effect" (Ritchie, 2005).

Obviously, simply replacing a word potentially results in a text that induces "conflict" (and confusion) in the audience. Using a phonetically similar word as a replacement, however, makes the statement pseudo-ambiguous, since the original intended meaning can also be recovered. There then are two "conflicting" and "contrasting" interpretations — the literal one and the original one — increasing the likelihood of humorous incongruity.

Requiring the substitute to share part-of-speech with the original word works in this direction too, and additionally increases the likelihood that the resulting text is a valid English statement.

Implementation We adopt an extended definition of punning and also consider orthographically similar or rhyming words as possible substitutes.

Two words are considered *orthographically similar* if one word is obtained with a single character deletion, addition, or replacement from the other one.

We call two words *phonetically similar* if their phonetic transcription is orthographically similar according to the above definition.

Two words *rhyme* if they have same positions of tonic accent, and if they are phonetically identical from the most stressed syllable to the end of the word.

Our implementation of these constraints uses the WordNet lexical database (Fellbaum, 1998) and CMU pronunciation dictionary[1]. The latter also provides a collection of words not normally contained in standard English dictionaries, but commonly used in informal language. This increases the space of potential replacements. We use the TreeTagger[2] POS tagger in order to consider only words with the same part-of-speech of the word to be replaced.

3.2 Taboo Constraint

Taboo constraint (TABOO) *requires that the substitute word is a taboo word or frequently used in taboo expressions, insults, or vulgar expressions.* Taboo words "represent a class of emotionally arousing references with respect to body products, body parts, sexual acts, ethnic or racial insults, profanity, vulgarity, slang, and scatology" (Jay et al., 2008), and they directly introduce "inappropriateness" to the text.

Implementation We collected a list of 700 taboo words. A first subset contains words manually selected from the domain SEXUALITY of WordNet-Domains (Magnini and Cavaglià, 2000). A second subset was collected from the Web, and contains words commonly used as insults. Finally, a third subset was collected from a website posting examples of funny autocorrection mistakes[3] and includes words that are not directly referring to taboos (e.g.: 'stimulation') or often retrieved in jokes evoking taboo meanings (e.g.: 'wife').

3.3 Contextual Constraints

Contextual constraints (CONT) *require that the substitution takes place at the end of the text, and in a locally coherent manner.*

By local coherence we mean that the substitute word forms a feasible phrase with its immediate predecessor. If this is *not* the case, then the text is likely to make little sense. On the other hand, if this *is* the case, then the taboo meaning is potentially expanded to the phrase level. This introduces a stronger semantic "contrast" and thus probably contributes to making the text funnier. The semantic contrast is potentially even stronger if the taboo word comes as a surprise in the end of a seemingly innocent text. The humorous effect then is similar to the one of the forced reinterpretation jokes.

Implementation Local coherence is implemented using n-grams. In the case of languages that are read from left to right, such as English, expectations will be built by the left-context of the expected word. To estimate the level of expectation triggered by a left-context, we rely on a vast collection of n-grams, the 2012 Google Books n-grams collection[4] (Michel et al., 2011) and compute the cohesion of each n-gram, by comparing their expected frequency (assuming word indepence), to their observed number of occurrences. A subsequent Student t-test allows to assign a measure of cohesion to each n-gram (Doucet and Ahonen-Myka, 2006). We use a substitute word only if its cohesion with the previous word is high.

In order to use consistent natural language and avoid time or location-based variations, we focused on contemporary American English. Thus we only used the subsection of Google bigrams for American English, and ignored all the statistics stemming from books published before 1990.

4 Evaluation

We evaluated the method empirically using CrowdFlower[5], a crowdsourcing service. The aim of the evaluation is to measure the potential effect of the three types of constraints on funniness of texts. In particular, we test the potential effect of

[1] available at `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

[2] available at `http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger`

[3] `http://www.damnyouautocorrect.com`

[4] available at `http://books.google.com/ngrams`

[5] available at `http://www.crowdflower.com`

adding the tabooness constraint to the form constraints, and the potential effect of further adding contextual constraints. I.e., we consider three increasingly constrained conditions: (1) substitution according only to the form constraints (FORM), (2) substitution according to both form and taboo constraints (FORM+TABOO), and (3) substitution according to form, taboo and context constraints (FORM+TABOO+CONT).

One of the reasons for the choice of taboo words as lexical constraint is that they allows the system to generate humorous text potentially appreciated by young adults, which are the majority of crowdsourcing users (Ross et al., 2010). We applied the humor generation method on the first 5000 messages of *NUS SMS Corpus*[6], a corpus of real SMS messages (Chen and Kan, 2012).

We carried out every possible lexical replacement under each of the three conditions mentioned above, one at a time, so that the resulting messages have exactly one word substituted. We then randomly picked 100 such modified messages for each of the conditions. Table 1 shows two example outputs of the humor generator under each of the three experimental conditions. These two examples are the least funny and the funniest message according to the empirical evaluation (see below).

For evaluation, this dataset of 300 messages was randomly divided into groups of 20 messages each. We recruited 208 evaluators using the crowdsourcing service, asking each subject to evaluate one such group of 20 messages. Each message in each group was judged by 90 different participants.

We asked subjects to assess individual messages for their funniness on a scale from 1 to 5. For the analysis of the results, we then measured the effectiveness of the constraints using two derived variables: the *Collective Funniness* (CF) of a message is its mean funniness, while its *Upper Agreement* (UA(t)) is the fraction of funniness scores greater than or equal to a given threshold t. To rank the generated messages, we take the product of Collective Funniness and Upper Agreement UA(3) and call it the overall *Humor Effectiveness* (HE).

In order to identify and remove potential scammers in the crowdsourcing system, we simply asked subjects to select the last word in the message. If a subject failed to answer correctly more than three times all her judgements were removed. As a result, 2% of judgments were discarded as untrusted. From the experiment, we then have a total of 26 534 trusted assessments of messages, 8 400 under FORM condition, 8 551 under FORM+TABOO condition, and 8 633 under FORM+TABOO+CONT condition.

The Collective Funniness of messages increases, on average, from 2.29 under condition FORM to 2.98 when the taboo constraint is added (FORM+TABOO), and further to 3.20 when the contextual constraints are added (FORM+TABOO+CONT) (Table 2). The Upper Agreement UA(4) increases from 0.18 to 0.36 and to 0.43, respectively.

We analyzed the distributions of Collective Funniness values of messages, as well as the distributions of their Upper Agreements (for all values from UA(2) to UA(5)) under the three conditions. According to the one-sided Wilcoxon rank-sum test, both Collective Funniness and all Upper Agreements increase from FORM to FORM+TABOO and from FORM+TABOO to FORM+TABOO+CONT statistically significantly (in all cases $p < .002$). Table 3 shows p-values associated with all pairwise comparisons.

5 Conclusions

We have proposed a new approach for the study of computational humor generation by lexical replacement. The generation task is based on a simple form of punning, where a given text is modified by replacing one word with a similar one.

We proved empirically that, in this setting, humor generation is more effective when using a list of taboo words. The other strong empirical result regards the context of substitutions: using bigrams to model people's expectations, and constraining the position of word replacement to the end of the text, increases funniness significantly. This is likely because of the form of surprise they induce. At best of our knowledge, this is the first time that these aspects of humor generation have been successfully evaluated with a crowdsourcing system and, thus, in a relatively quick and economical way.

The statistical significance is particularly high, even though there were several limitations in the experimental setting. For example, as explained in Section 3.2, the employed word list was built

[6]available at `http://wing.comp.nus.edu.sg/SMSCorpus`

Experimental Condition	Text Generated by the System	CF	UA(3)	HE
FORM	Oh oh...Den muz change plat liao...Go back have yan jiu again... Not 'plat'...'plan'.	1.68	0.26	0.43
FORM	Jos ask if u wana melt up? 'meet' not 'melt'!	2.96	0.74	2.19
FORM+TABOO	Got caught in the rain.Waited half n hour in the buss stop. Not 'buss'...'bus'!	2.06	0.31	0.64
BASE+TABOO	Hey pple... $ 700 or $ 900 for 5 nights...Excellent masturbation wif breakfast hamper!!! Sorry I mean 'location'	3.98	0.85	3.39
FORM+TABOO+CONT	Nope...Juz off from berk... Sorry I mean 'work'	2.25	0.39	0.87
FORM+TABOO+CONT	I've sent you my fart.. I mean 'part' not 'fart'...	4.09	0.90	3.66

Table 1: Examples of outputs of the system. CF: Collective Funniness; UA(3): Upper Agreement; HE: Humor Effectiveness.

	Experimental Conditions		
	FORM	FORM+TABOO	FORM+TABOO+CONT
CF	2.29 ± 0.19	2.98 ± 0.43	3.20 ± 0.40
UA(2)	0.58 ± 0.09	0.78 ± 0.11	0.83 ± 0.09
UA(3)	0.41 ± 0.07	0.62 ± 0.13	0.69 ± 0.12
UA(4)	0.18 ± 0.04	0.36 ± 0.13	0.43 ± 0.13
UA(5)	0.12 ± 0.02	0.22 ± 0.09	0.26 ± 0.09

Table 2: Mean Collective Funniness (CF) and Upper Agreements (UA(·)) under the three experimental conditions and their standard deviations.

	Hypotheses	
	FORM → FORM+TABOO	FORM+TABOO → FORM+TABOO+CONT
CF	10^{-15}	9×10^{-5}
UA(2)	10^{-15}	1×10^{-15}
UA(3)	10^{-15}	7×10^{-5}
UA(4)	10^{-15}	2×10^{-4}
UA(5)	10^{-15}	2×10^{-3}

Table 3: P-values resulting from the application of one-sided Wilcoxon rank-sum test.

from different sources and contains words not directly referring to taboo meanings and, thus, not widely recognizable as "taboo words". Furthermore, the possible presence of crowd-working scammers (only partially filtered by the gold standard questions) could have reduced the statistical power of our analysis. Finally, the adopted humor generation task (based on a single word substitution) is extremely simple and the constraints might have not been sufficiently capable to produce a detectable increase of humor appreciation.

The statistically strong results that we obtained can make this evaluation approach attractive for related tasks. In our methodology, we focused attention to the correlation between the parameters of the system (in our case, the constraints used in lexical selection) and the performance of humor generation. We used a multi-dimensional measure of humorous effect (in terms of funniness and agreement) to measure subtly different aspects of the humorous response. We then adopted a comparative setting, where we can measure improvements in the performance across different systems or variants.

In the future, it would be interesting to use a similar setting to empirically investigate more subtle ways to generate humor, potentially with weaker effects but still recognizable in this setting. For instance, we would like to investigate the use of other word lists besides taboo domains and the extent to which the semantic relatedness itself could contribute to the humorous effect.

The current techniques can be improved, too, in various ways. In particular, we plan to extend the use of n-grams to larger contexts and consider more fine-grained tuning of other constraints, too. One goal is to apply the proposed methodology to isolate, on one hand, parameters for inducing incongruity and, on the other hand, parameters for making the incongruity funny.

Finally, we are interested in estimating the probability to induce a humor response by using different constraints. This would offer a novel way to intentionally control the humorous effect.

References

S. Attardo and V. Raskin. 1991. Script theory revis(it)ed: joke similarity and joke representation model. *Humour*, 4(3):293–347.

J. Beattie. 1971. An essay on laughter, and ludicrous composition. In *Essays. William Creech, Edinburgh, 1776*. Reprinted by Garland, New York.

K. Binsted, H. Pain, and G. Ritchie. 1997. Children's evaluation of computer-generated punning riddles. *Pragmatics and Cognition*, 2(5):305–354.

T. Chen and M.-Y. Kan. 2012. Creating a live, public short message service corpus: The nus sms corpus. *Language Resources and Evaluation*, August. published online.

A. Doucet and H. Ahonen-Myka. 2006. Probability and expected document frequency of discontinued word sequences, an efficient method for their exact computation. *Traitement Automatique des Langues (TAL)*, 46(2):13–37.

C. Fellbaum. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

J. H. Goldstein, J. M. Suls, and S.Anthony. 1972. Enjoyment of specific types of humor content: Motivation or salience? In J. H. Goldstein and P. E. McGhee, editors, *The psychology of humor: Theoretical perspectives and empirical issues*, pages 159–171. Academic Press, New York.

T. Jay, C. Caldwell-Harris, and K. King. 2008. Recalling taboo and nontaboo words. *American Journal of Psychology*, 121(1):83–103, Spring.

M. Levison and G. Lessard. 1992. A system for natural language generation. *Computers and the Humanities*, 26:43–58.

B. Magnini and G. Cavaglià. 2000. Integrating subject field codes into wordnet. In *Proc. of the 2^{nd} International Conference on Language Resources and Evaluation (LREC2000)*, Athens, Greece.

R. A. Martin. 2007. *The Psychology of Humor: An Integrative Approach*. Elsevier.

J. McKay. 2002. Generation of idiom-based witticisms to aid second language learning. In *(Stock et al., 2002)*.

J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

V. Raskin and S. Attardo. 1994. Non-literalness and non-bona-fide in language: approaches to formal and computational treatments of humor. *Pragmatics and Cognition*, 2(1):31–69.

V. Raskin. 1985. *Semantic Mechanisms of Humor*. Dordrecht/Boston/Lancaster.

G. Ritchie. 2002. The structure of forced interpretation jokes. In *(Stock et al., 2002)*.

G. Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the 10th European Natural Language Generation Workshop*, Aberdeen, August.

J. Ross, I. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. 2010. Who are the crowdworkers?: Shifting demographics in amazon mechanical turk. In *Proc. of the ACM CHI Conference*.

O. Stock and C. Strapparava. 2002. HAHAcronym: Humorous agents for humorous acronyms. In *(Stock et al., 2002)*.

O. Stock, C. Strapparava, and A. Nijholt, editors. 2002. *Proceedings of the The April Fools Day Workshop on Computational Humour (TWLT20)*, Trento.

H. Summerfelt, L. Lippman, and I. E. Hyman Jr. 2010. The effect of humor on memory: Constrained by the pun. *The Journal of General Psychology*, 137(4):376–394.

A. Valitutti. 2011. How many jokes are really funny? towards a new approach to the evaluation of computational humour generators. In *Proc. of 8th International Workshop on Natural Language Processing and Cognitive Science*, Copenhagen.

C. Venour. 1999. The computational generation of a class of puns. Master's thesis, Queen's University, Kingston, Ontario.