A Pinch of Humor for Short-Text Conversation: an Information Retrieval Approach

Vladislav Blinov, Kirill Mishchenko, Valeria Bolotova, and Pavel Braslavski

Ural Federal University
vladislav.blinov@urfu.ru, ki.mishchenko@gmail.com
lurunchik@gmail.com, pbras@yandex.ru

Abstract. The paper describes a work in progress on humorous response generation for short-text conversation using information retrieval approach. We gathered a large collection of funny tweets and implemented three baseline retrieval models: BM25, the query term reweighting model based on syntactic parsing and named entity recognition, and the *doc2vec* similarity model. We evaluated these models in two ways: *in situ* on a popular community question answering platform and in laboratory settings. The approach proved to be promising: even simple search techniques demonstrated satisfactory performance. The collection, test questions, evaluation protocol, and assessors' judgments create a ground for future research towards more sophisticated models.

Keywords: natural language interfaces, humor generation, short-text conversation, information retrieval, evaluation, community question answering

1 Introduction

Humor is an essential aspect of human communication. Therefore, sense of humor is a desirable trait of chatbots and conversational agents aspiring to act like humans. Injection of humor makes human-computer conversations more engaging, contributes to the agent's personality, and enhances the user experience with the system [18, 10]. Moreover, a humorous response is a good option for out-of-domain requests [3] and can soften the negative impact of inadequacies in the system's performance [4]. However, if we look at existing mobile personal assistants (for example, Apple Siri[1]), it can be noticed that their humorous answers work on a limited set of stimuli and are far from being diverse.

In this study, we approached the problem of generating a humorous response to the user's utterance as an information retrieval (IR) task over a large collection of presumably funny content. Our approach is exploratory and is not based on a certain theory of humor or a concrete type of jokes. The aim of our study is to implement several retrieval baselines and experiment with different methods of evaluation. IR is a promising approach in conversational systems [29, 28] that can significantly improve quality and diversity of responses.

[1] https://www.apple.com/ios/siri/

First, we gathered about 300,000 funny tweets. After that, we implemented three baselines for tweet retrieval: 1) BM25 – a classical IR model based on term statistics; 2) a query term reweighting model based on syntactic parsing and NER; and 3) a retrieval model based on document embeddings. Finally, we collected user questions and evaluated three baselines *in situ* on a community question answering (CQA) platform and in laboratory settings.

To the best of our knowledge, IR has not been applied to humorous response generation in short-text conversation scenario and no formal evaluation has been conducted on the task before. We have made the tweet collection (as a list of tweet IDs), test questions and assessors' judgments freely available[2] for research. The data creates a solid ground for future research in the field.

2 Related Work

There are two main directions in computational humor research: humor recognition and humor generation.

Humor recognition is usually formulated as a classification task with a wide variety of features – syntactic parsing, alliteration and rhyme, antonymy and other WordNet relations, dictionaries of slang and sexually explicit words, polarity and subjectivity lexicons, distances between words in terms of *word2vec* representations, etc. [24, 16, 11, 30, 31]. A cognate task is detection of other forms of figurative language such as irony and sarcasm [20, 19, 25]. Several recent studies dealing with humor and irony detection are focused on the analysis of tweets, see [31, 19, 20].

Most humor generation approaches focus on puns, as puns have relatively simple surface structure [21, 26, 6]. Stock and Strapparava [23] developed *HAHAcronym*, a system that generates funny deciphers for existing acronyms or produces new ones starting from concepts provided by the user. Valitutti et al. [26] proposed a method for 'adult' puns made from short text messages by lexical replacement. A related study [6] addresses the task of automatic template extraction for pun generation.

Mihalcea and Strapparava [17] proposed a method for adding a joke to an email message or a lecture note from a collection of 16,000 one-liners using latent semantic analysis (LSA). A small-scale user study showed a good reception of the proposed solution. This study is the closest to ours; however, we use an order of magnitude larger collection, implement several retrieval models and place emphasis on evaluation methodology.

Wen et al. [27] explore a scenario, when a system suggests the user funny images to be added to a chat. The work also employs an IR technique among others: candidate images are partly retrieved through Bing search API using query "funny *keywords*", where *keywords* are *tf-idf* weighted terms from the last three utterances.

Shahaf et al. [22] investigate the task of ranking cartoon captions provided by the readers of New Yorker magazine. They employ a wide range of linguistic

[2] `https://github.com/micyril/humor`

features as well as features from manually crafted textual descriptions of the cartoons. Jokes comparison/ranking task is close to ours, however, the settings and data are quite different.

Augello et al. [2] described a chatbot nicknamed *Humorist Bot*. The emphasis was made on humor recognition in the humans' utterances following the approach proposed in [16]; the bot reacted to jokes with appropriate responses and emoticons. Humorous response generation was restricted to a limited collection of jokes that was triggered when the user asked the bot to tell one.

The information retrieval approach to short-text conversations became popular recently [8, 29, 28]. The method benefits from the availability of massive conversational data, uses a rich set of features and learning-to-rank methods. Our approach follows the same general idea; however our exploratory study employs simpler retrieval models with a weak supervision.

3 Data

3.1 Joke Collection

To gather a collection of humorous tweets, we started with several "top funny Twitter accounts" lists that can be easily searched online[3]. We filtered out accounts with less than 20,000 followers, which resulted in 103 accounts. Table 1 lists top 10 most popular accounts in the collection. Then, we downloaded all available text-only tweets (i.e. without images, video, and URLs) and retained those with at least 30 likes or retweets (366,969 total). After that, we removed duplicates with a Jaccard similarity threshold of 0.45 using a Minhash index implementation[4] and ended up with a collection of 300,876 tweets. Here is an example from our collection (359 likes, 864 retweets)[5]:

Life is a weekend when you're unemployed.

To validate the proposed data harvesting approach, we implemented a humor recognizer based on a dataset consisting of 16,000 one-liners and 16,000 non-humorous sentences from news titles, proverbs, British National Corpus, and Open Mind Common Sense collection [16]. We employed a concatenation of *tf-idf* weights of unigrams and bigrams with document frequency above 2 and 300-dimensional *doc2vec* representations (see Section 4.3 for details) as a feature vector. A logistic regression classifier achieved 10-fold cross-validation accuracy of 0.887 (which exceeds previously reported results [16, 30]). We applied this classifier to our collection as well as to a sample of tweets from popular media accounts, see Table 2. The results confirm that the approach is sound; however, we did not filter the collection based on the classification results since the training data is quite different from the tweets. For example, many emotional and sentiment rich tweets from the *realDonaldTrump* account are considered to be funny by our classifier.

[3] See for example `http://www.hongkiat.com/blog/funny-twitter-accounts/`

[4] `https://github.com/ekzhu/datasketch`

[5] `https://twitter.com/MensHumor/status/360113491937472513`

Table 1: Accounts with highest numbers of followers in the collection

Account	# of followers
ConanOBrien	21,983,968
StephenAtHome	11,954,015
TheOnion	9,128,284
SteveMartinToGo	7,828,907
Lmao	5,261,116
AlYankovic	4,355,144
MensHumor	3,543,398
TheTweetOfGod	2,285,307
Lord_Voldemort7	2,026,292
michaelianblack	2,006,624

Table 2: Humor recognition in tweets

Collection/account	Classified as humorous	# of tweets
Funny Accounts	258,466/85.9%	300,876
The Wall Street Journal (wsj)	142/9.7%	1,464
The Washington Post (washingtonpost)	195/21.5%	907
The New York Times (nytimes)	240/19.8%	1,210
Donald J. Trump (realDonaldTrump)	7,653/59.1%	12,939

3.2 Yahoo!Answers

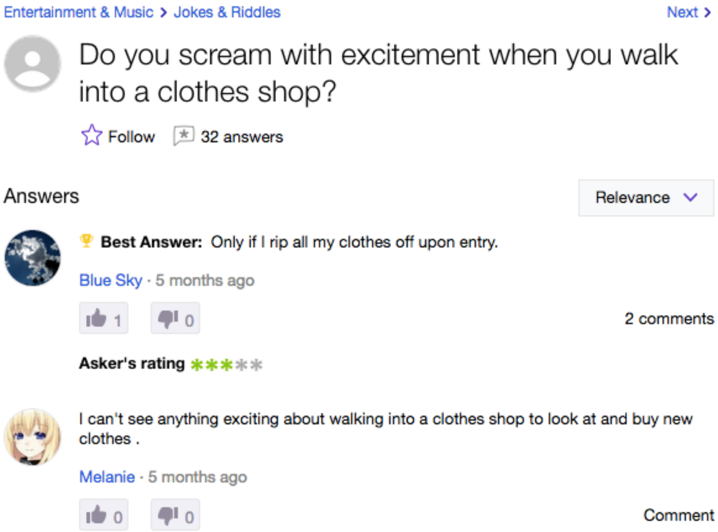
We used *Jokes & Riddles* category of Yahoo!Answers[6] for *in situ* evaluation: as a source of users' questions and measuring reactions of community members to automatically retrieved answers.

Yahoo!Answers is a popular CQA platform where users can ask questions on virtually any subject and vote for answers with 'thumb up' and 'thumb down'; the asker can also nominate the 'best answer' [1]. Fig. 1 represents Yahoo!Answers' user interface with a question and two answers provided by the community in the *Jokes & Riddles* category. Each question has a title and an optional longer description, which we disregarded. Approximately 20 questions are posted in *Jokes & Riddles* daily. The category has an obvious topical bias: there are noticeably many ironic questions on atheism, faith and theory of evolution (see for instance the second question in Table 3). Apart from using Yahoo!Answers to evaluate the retrieved responses, we also gathered historical data to weakly supervise our Query Term Rewighting model (see Section 4.2). We collected 1,371 questions asked during two months in the *Jokes & Riddles* category along with submitted answers; 856 of the threads contain 'best answer' nominations.

4 Retrieval Models

We implemented three joke retrieval baselines: 1) a classical BM25 model based on term statistics; 2) a query term reweighting model based on structural prop-

[6] `https://answers.yahoo.com/dir/index?sid=396546041`



Do you scream with excitement when you walk into a clothes shop?

Best Answer: Only if I rip all my clothes off upon entry.

Blue Sky · 5 months ago

Asker's rating ★★★★★

I can't see anything exciting about walking into a clothes shop to look at and buy new clothes .

Melanie · 5 months ago

Fig. 1: Yahoo!Answers interface

erties of the stimulus (dependency tree and the presence of named entities); and 3) a model based on document embeddings that retrieves semantically similar 'answers' and does not require word overlap with the 'query'. Table 3 shows examples of top-ranked responses by these models.

4.1 BM25

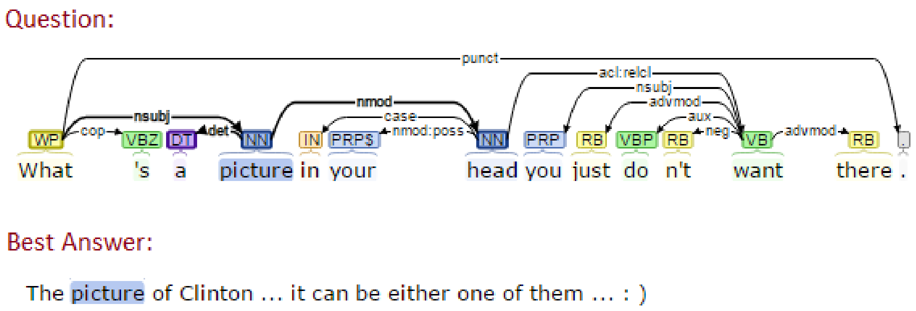
BM25 is a well-known ranking formula [9], a variant of the *tf-idf* approach. It combines term frequency within document (*tf*) and collection-wide frequency (*idf*) to rank documents that match query terms. We did not perform stop-word removal: in has been shown that personal pronouns are important features for humorous content [22, 15]. Since documents (tweets) in our case are rather short, ranking is dominated by the *idf* weights, i.e. rare words. It can potentially be harmful for humor retrieval, since many popular jokes seem to contain mostly common words [22, 15].

4.2 Query Term Reweighting Model (QTR)

This approach is inspired by the notion of *humor anchors* introduced in [30]. According to the authors, a humor anchor is a set of word spans in a sentence that enables humorous effect. To detect humor anchors, they firstly extracted a set of candidates based on syntactic parsing, and then searched for a subset that caused a significant drop in humor score when removed from the original sentence. In our study, we followed the idea of humor anchors to modify term weighting scheme: BM25 scores in a query are adjusted corresponding to the syntactic roles of matched terms.

In order to calculate weight adjustments, we used 573 question-'best answer' pairs that have word overlap (out of 856, see Section 3.2).

We applied dependency parsing and named entity recognition (NER) from Stanford CoreNLP [14] to the questions and counted dependency and NER labels of the overlapping words (see Figure 2 and 3, respectively). Then, we set adjustment coefficients accordingly to the counts. As expected, the *nominal subject* and *main verb* (root) roles and the *person* NE are the most frequent.



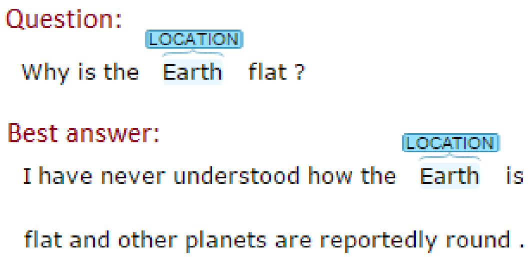
Question:

What 's a picture in your head you just do n't want there .

Best Answer:

The picture of Clinton … it can be either one of them … :)

Fig. 2: Learning term reweighting based on dependency parsing: accounting for syntactic tags of overlapping words (*picture, nsubj*)



Question:

Why is the Earth flat ?

Best answer:

I have never understood how the Earth is flat and other planets are reportedly round .

Fig. 3: Learning term reweighting based on NER: accounting for NE types of overlapping words (*Earth, LOCATION*)

4.3 doc2vec

doc2vec model [13] generates vector representations of sentences and paragraphs, thus extending *word2vec* model from word to document embeddings. In contrast to two previous models, *doc2vec* is capable of finding jokes semantically close to the 'query' even when there is no word overlap between them (recall, about one third of 'best answers' in *Jokes & Riddles* category have no word overlap with the corresponding questions). See for example the *doc2vec* response to the second question in Table 3. We used cosine similarity between the question's and documents' vector representations to find semantically closest jokes.

We followed the same pipeline as described in [12]: tokenizing, lowercasing, and inferring embeddings with the same initial learning rate and number of epochs. In our experiments, we used the Distributed Bag of Words (DBOW) model pre-trained on the English Wikipedia[7].

Table 3: Examples of top-ranked responses by the three models

BM25	QTR	doc2vec
	Is it true Hilary Clinton is secretly Donald Trump's mom?*	
Is it true eminem thanked his mom's spaghetti	At the very least, I'm far less concerned about Hillary Clinton's physical ailments than I am about Donald Trump's mental ones.	I'm not convinced Donald Trump knows what sex is.
	Why do you atheist not apply the same standards of evidence on your own "theories" as you do to challenge the existence of God?	
The existence of conspiracy theories is a myth.	I have a mosquito bite on the inside of the arch of my foot thus disproving the existence of God.	Science is true whether or not you believe it, but religion is true whether or not it's true.

*Original spelling

5 Evaluation

Humor evaluation is challenging, since the perception of humor is highly subjective and is conditioned on the situation and the socio-cultural background. We evaluated joke retrieval in two modes: 1) *in situ* – top-1 ranked responses of each model were presented to the CQA users in a 'humorous' category and 2) top-3 responses for a subset of the same CQA questions were assessed by three judges in lab settings. The former approach allows evaluation in real-life environment, however it scales poorly and is harder to interpret. The latter one

[7] https://github.com/jhlau/doc2vec

is more controllable, but it is not clear how well few judges represent an 'average' user.

5.1 Yahoo!Answers

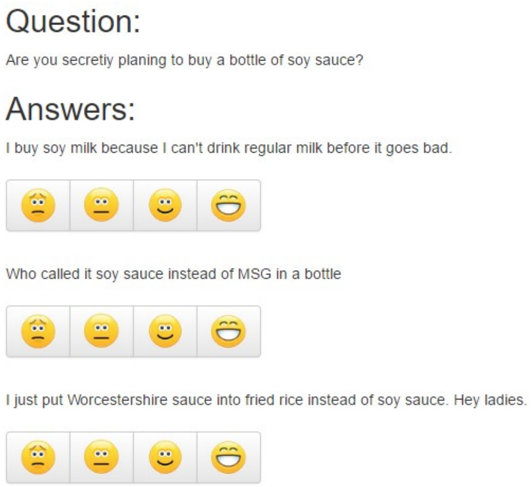
For six consecutive days, we manually posted top-1 ranked responses by the models to questions asked during the day. We have submitted responses to 101 questions in total; in five threads there was no voting activity (i.e. neither a 'best answer' selected, nor any votes submitted), so we excluded them from the analysis. Each question received 22 answers on average (including three from our models). The CQA users' votes were collected two weeks later. Evaluation on Yahoo!Answers allows potentially for evaluation models against each other, as well as comparison of automatic responses with those by the users. Table 6 summarizes the users' reaction to retrieved answers by model: upvotes (+), downvotes (−), 'best answer' nominations (BA), and number of times the model outperformed the other two ('best model'). If we rank all answers by their votes (the 'best answer' on the first position if present), there are 5.72 rank positions on average; mean position of an oracle (the best variant out of three models' answers) is 3.92. The last column in Table 6 presents an average percentage of the users, whose answers had lower rank positions than the model's responses. Table 4 shows similar statistics for seven most active answerers in the same 96 question threads. Obviously, users are much more selective, they do not answer all questions in a row and thus have higher average scores per answer. Nevertheless, automatically retrieved answers do not look completely hopeless when compared to these users' performance (except for several good scorers, e.g. User6).

Table 4: The most active CQA users (96 questions)

User	*# answers*	*+*	*–*	*BA*	*Users below*
User1	23	25	0	1	32.1%
User2	20	33	1	0	33.1%
User3	20	13	1	2	15.9%
User4	17	20	1	0	16.6%
User5	15	24	0	0	45.6%
User6	13	28	0	7	71.1%
User7	13	8	0	0	19.5%

5.2 Lab Evaluation

For lab evaluation we sampled 50 questions from the ones we answered on Yahoo!Answers. Top-3 results for each model were collected for evaluation, yielding 433 unique question-answer pairs. The question-response pairs were presented to three assessors in a dedicated evaluation interface in a random order, three at a time. The assessors were asked to judge responses with the context in mind, i.e. an out-of-context joke, even when it is funny by itself, is expected to be scored low. The responses were judged on a four-point scale (from 0 to 3), with corresponding emoticons in the evaluation interface (see Fig. 4).



Question:

Are you secretiy planing to buy a bottle of soy sauce?

Answers:

I buy soy milk because I can't drink regular milk before it goes bad.

😟 😐 🙂 😁

Who called it soy sauce instead of MSG in a bottle

😟 😐 🙂 😁

I just put Worcestershire sauce into fried rice instead of soy sauce. Hey ladies.

😟 😐 🙂 😁

Fig. 4: The annotation tool for laboratory evaluation

The relevance score for a question–response pair is an average over three assessors' labels (see Table 5 for some examples). Table 7 shows the averaged scores of the top-ranked responses and DCG@3 scores [7] for the three models and the oracle that composes the best output from the nine pooled results. The averaged pairwise weighted Cohen's kappa [5] is 0.13, which indicates a low agreement among assessors. Here is an example of a question–answer pair that received three different labels (☺/😃/😐) from three assessors:

Q: Do you scream with excitement when you walk into a clothes shop?
A: Do hipsters in the Middle East shop at Turban Outfitters?

Table 5: Example question–response pairs and their averaged relevance scores

Score	*Question*	*Response*
3.00	Does evolution being a theory make it subjective?	There is no theory of evolution, just a list of creatures Chuck Norris allows to live.
2.67	Can you find oil by digging holes in your backyard?	Things to do today: 1.Dig a hole 2. Name it love 3. Watch people fall in it.
1.33	Why don't they put zippers on car doors?	Sick of doors that aren't trap doors.
0.67	What if you're just allergic to working hard?	You're not allergic to gluten.
0.33	What test do all mosquitoes pass?	My internal monologue doesn't pass the Bechdel test. :(

Table 6: CQA users' reaction (96 questions)

Model	*+*	*-*	*BA*	*Best model*	*Users below*
BM25	19	3	0	3	16.5%
QTR	14	1	2	6	16.1%
doc2vec	15	2	2	3	14.3%
Oracle	23	1	4	–	19.4%

Table 7: Lab evaluation results (50 questions)

Model	*Avg. score @1*	*DCG@3*
BM25	1.34	2.78
QTR	1.15	2.38
doc2vec	1.25	2.63
Oracle	1.91	3.61

6 Conclusion and Future Work

The most important outcome of the conducted experiment is that a combination of a simple approach to harvesting a joke collection and uncomplicated retrieval models delivers satisfactory performance for humorous response generation task. On the one hand, we may hypothesize that the size of the collection does matter – even simple methods can yield reasonable results when they have a lot of variants to choose from. On the other hand, it seems that when a situation implies a whimsical response, an unexpected, illogical or even inconsistent answer can still be considered funny.

The evaluation on the CQA platform showed that automatic methods for humorous response generation have at least some promise compared to humans. At the same time this evaluation does not reveal an absolute winner among three models. Keeping in mind short-text conversation scenario, best answer nominations seem to be the most appropriate quality measure that proves the advantage of the QTR and *doc2vec* models. However, best answer selection is very competitive in contrast to one-to-one conversation scenario (the asker receives about 20 answers on average); 'thumb up' and 'thumb down' scores from community members seem to be less subjective and biased. In terms of these two scores, BM25 slightly outperforms two other models.

If we look at CQA users' up– and downvotes only, lab evaluation confirms the advantage of BM25 over the other two models to some extent. What seems to be more important in case of top-3 results evaluation is that the models deliver quite diverse responses – the oracle's scores are significantly higher. The average score of the oracle's top response is close to *funny* ☺, which is promising. The results suggest that a deeper question analysis, humor-specific features and advanced ranking methods can potentially deliver higher-quality responses.

Although a low agreement among assessors in laboratory settings is expected, it constitutes a serious obstacle for future work. Lab evaluation, successfully used in various information retrieval tasks, in our case proves that humor is a highly subjective and contextualized area. Additional efforts must be undertaken to ensure a higher inter-annotator agreement and reliability of judgments. We will explore the opportunity to account for assessors' personality traits (such as Big Five[8]), socio-demographic characteristics, language proficiency, and humor-

[8] `https://en.wikipedia.org/wiki/Big_Five_personality_traits`

specific profiling (cf. Jester project[9]) that can potentially help interpret and reconcile divergent assessments. We will also consider crowdsourcing humor evaluation, as several recent studies suggest. In addition, we plan to conduct user studies to better understand the perception and role of humor in short-text conversations.

We also plan to build a sizable collection of dialog jokes, which will allow us to harness advanced features already explored in humor recognition and combine them using learning-to-rank methods. State-of-the-art humor recognition methods can also be applied to improve the quality of the joke corpus.

To sum up, the study demonstrates that the information retrieval approach to humorous response generation is a promising direction of research. The current collection of tweets, test questions, evaluation protocol and assessors' judgments create a solid ground for further investigations of the IR-based humorous response generation.

References

1. Adamic, L.A., Zhang, J., Bakshy, E., Ackerman, M.S.: Knowledge sharing and yahoo answers: everyone knows something. In: Proc. of WWW. pp. 665–674 (2008)
2. Augello, A., Saccone, G., Gaglio, S., Pilato, G.: Humorist bot: Bringing computational humour in a chat-bot system. In: Proc. of CISIS. pp. 703–708 (2008)
3. Bellegarda, J.R.: Spoken language understanding for natural interaction: The Siri experience. In: Natural Interaction with Robots, Knowbots and Smartphones, pp. 3–14 (2014)
4. Binsted, K.: Using humour to make natural language interfaces more friendly. In: Proc. of the AI, ALife and Entertainment Workshop (1995)
5. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational linguistics 22(2), 249–254 (1996)
6. Hong, B.A., Ong, E.: Automatically extracting word relationships as templates for pun generation. In: Proc. of CALC. pp. 24–31 (2009)
7. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. TOIS 20(4), 422–446 (2002)
8. Ji, Z., Lu, Z., Li, H.: An information retrieval approach to short text conversation. arXiv preprint arXiv:1408.6988 (2014)
9. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. Information Processing & Management 36(6), 779–840 (2000)
10. Khooshabeh, P., McCall, C., Gandhe, S., Gratch, J., Blascovich, J.: Does it matter if a computer jokes? In: Proc. of CHI. pp. 77–86 (2011)
11. Kiddon, C., Brun, Y.: That's what she said: double entendre identification. In: Proc. of ACL-HLT, Vol. 2. pp. 89–94 (2011)
12. Lau, J.H., Baldwin, T.: An empirical evaluation of doc2vec with practical insights into document embedding generation. In: Proc. of the 1st Workshop on Representation Learning for NLP. pp. 78–86 (2016)
13. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proc. of ICML. pp. 1188–1196 (2014)

[9] `http://eigentaste.berkeley.edu/about.html`

14. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: ACL System Demonstrations. pp. 55–60 (2014)
15. Mihalcea, R., Pulman, S.: Characterizing humour: An exploration of features in humorous texts. In: Proc. of CICLing. pp. 337–347 (2007)
16. Mihalcea, R., Strapparava, C.: Learning to laugh (automatically): Computational models for humor recognition. Computational Intelligence 22(2), 126–142 (2006)
17. Mihalcea, R., Strapparava, C.: Technologies that make you smile: adding humor to text-based applications. IEEE Intelligent Systems 21(5), 33–39 (2006)
18. Niculescu, A., van Dijk, B., Nijholt, A., Li, H., See, S.L.: Making social robots more attractive: the effects of voice pitch, humor and empathy. International journal of social robotics 5(2), 171–191 (2013)
19. Rajadesingan, A., Zafarani, R., Liu, H.: Sarcasm detection on Twitter: A behavioral modeling approach. In: Proc. of WSDM. pp. 97–106 (2015)
20. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in Twitter. Language resources and evaluation 47(1), 239–268 (2013)
21. Ritchie, G.: Can computers create humor? AI Magazine 30(3), 71–81 (2009)
22. Shahaf, D., Horvitz, E., Mankoff, R.: Inside jokes: Identifying humorous cartoon captions. In: Proc. of KDD. pp. 1065–1074 (2015)
23. Stock, O., Strapparava, C.: Getting serious about the development of computational humor. In: Proc. of IJCAI. pp. 59–64 (2003)
24. Taylor, J.M., Mazlack, L.J.: Computationally recognizing wordplay in jokes. In: Proc. of CogSci. pp. 1315–1320 (2004)
25. Tsur, O., Davidov, D., Rappoport, A.: ICWSM–A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In: Proc. of ICWSM. pp. 162–169 (2010)
26. Valitutti, A., Toivonen, H., Doucet, A., Toivanen, J.M.: "Let everything turn well in your wife": Generation of adult humor using lexical constraints. In: Proc. of ACL, Vol. 2. pp. 243–248 (2013)
27. Wen, M., Baym, N., Tamuz, O., Teevan, J., Dumais, S., Kalai, A.: OMG UR funny! Computer-Aided Humor with an Application to Chat. In: Proc. of ICCC. pp. 86–93 (2015)
28. Yan, R., Song, Y., Wu, H.: Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In: Proc. of SIGIR. pp. 55–64 (2016)
29. Yan, Z., Duan, N., Bao, J., Chen, P., Zhou, M., Li, Z., Zhou, J.: DocChat: An information retrieval approach for chatbot engines using unstructured documents. In: Proc. of ACL. pp. 516–525 (2016)
30. Yang, D., Lavie, A., Dyer, C., Hovy, E.: Humor recognition and humor anchor extraction. In: Proc. of EMNLP. pp. 2367–2376 (2015)
31. Zhang, R., Liu, N.: Recognizing humor on Twitter. In: Proc. of CIKM. pp. 889–898 (2014)