Information retrieval approach to humorous response generation in dialog systems: a baseline

Vladislav Blinov (cunningplan@yandex.ru), Valeriya Bolotova (lurunchik@gmail.com), and Pavel Braslavski (pbras@yandex.ru)

Ural Federal University, Yekaterinburg, Russia

ABSTRACT

In this paper, we present a baseline IR-based solution to humorous response generation in dialog systems. We describe 1) a corpus consisting of about 48,000 jokes gathered from the VK social network, 2) about 80 test stimuli, 3) BM25 and popularity-based baseline systems, 4) evaluation protocol and results. The study creates a ground for future research in the direction.

Keywords: computational humor, dialog systems, information retrieval approach

Introduction

Proliferation of mobile devices makes mobile virtual assistants, chat bots, dialog and question answering systems very popular. Examples of such systems are Apple Siri, Google Now, Facebook M, and Microsoft Cortana. A mobile virtual assistant should be able not only to respond to the user's questions, control the mobile device and perform simple tasks such as calling a taxi or scheduling a meeting, but also have a sense of humor. This ability helps the application avoid formal or cliched answers to questions that it cannot answer and makes it more human-like, attractive and engaging [1–3].

In this paper we approach the problem of generating a humorous response to the user's stimulus as an information retrieval (IR) task. The study presents two baseline solutions to the problem, one is based on BM25 ranking and the other relies mainly on joke popularity. We collected about 40 thousand unique jokes in Russian from three user communities on the social network VKontakte and evaluated both approaches using about 80 utterances from a mobile assistant log and questions from Humor category of a popular community question answering platform as test stimuli. With the joke collection, test questions, evaluation interface and baseline evaluation results, the study builds a basis for further research in humorous response generation in dialog systems based on re-using available humorous content.

Related Work

The field of computational humor received a significant deal of attention in the 2000s. There are two main directions of research in the field: humor generation [4–6] and humor recognition (see for example [7, 8]). Humor generation studies focus usually on a certain type of jokes. Stock and Strapparava [4] developed HAHAcronym, a system that generates funny deciphers for existing acronyms or produces new ones starting from concepts provided by the user. Ritchie [6] systematized different types of puns and proposed mechanisms for automatic pun generation. Valitutti et al. [5] proposed a method for 'adult' puns made from short text messages by lexical replacement. In the field of information retrieval, Friedland and Allan [9] proposed a domain-specific joke retrieval model based on jokes structure and interchangeable word classes. Surdeanu et al. [10] investigated usefulness of different linguistic features for search in large archives of questions and answers for non-factoid questions. The study does not deal with humorous content, but the approach is still similar to ours, although we employ very straightforward search methods in our current study. Methods for building dialog systems based on human conversation archives, for instance – a large collection of tweets along with responses [11] – are also close to our study, although, again, they use much more sophisticated methods compared to our current techniques.

Data

We gathered a collection of jokes from three popular humor-related user communities on VKontakte, the largest Russian social network. We collected only posts without images or video that gained more than 500 Likes. Table 1 summarizes the sources of the initial corpus.

Group	URL	# of jokes
F*** normality	https://vk.com/trahninormalnost1	64,276
Evil incorporated	https://vk.com/evil_incorparate	63,163
Witty	https://vk.com/ostroym	49,447
	Total	176,886

Table 1. Initial collection of jokes by source.

The three groups differ in number of subscribers and post frequency; moreover, these parameters are unstable over time. To make the popularity of jokes comparable across different sources and time, we normalized Like scores. Firstly, we excluded 637 outliers based on three-sigma rule, even those with very high scores that appeared to be viral posts. Secondly, we smoothed Like scores within groups using sliding window of size 10. Finally, we mapped scores to $(0, 1)$ interval within each group.

Further, we retained only one-liners and two-turn dialog jokes (see Examples 1 and 2, respectively), which resulted in 159,278 jokes. Then, we removed duplicates based on similarity of lemmatized bag-of-words representations. This step reduced drastically the collection size – down to 48,813 jokes. We also labeled jokes without obscene words using a dedicated library[1]; the 'decent' collection contains 41,848 texts.

Лекарства так подорожали, что скоро мы будем дарить их друг другу на день рождения, чтобы дожить до следующего.

Example 1. One-liner

– Давай на камень, ножницы, бумага?
– Сереж, я ушла от тебя два года назад, я не вернусь, отстань!

Example 2. Dialog joke

Baseline Models

We address the task of generating a humorous response to a user's stimulus as a ranked retrieval over a large collection of humorous content, using user's input as query. In case of one-liners we return the entire joke, in case of dialog jokes – the second (final) turn. We have implemented

[1] https://github.com/oranmor/russian_obscenity

two baseline approaches: popularity-based ranking (Likes model) and BM25 scoring. The former can be regarded as an analogue of query-independent ranking based on document authority (e.g. PageRank) – a funny joke is potentially still funny, even if it is not quite in the context. The latter is based on query-dependent textual similarity between the query and response (the joke should be on-topic). The Likes model requires minimal overlap between the question and candidate responses (one common noun or verb) and ranks the responses by descending normalized Like scores. BM25 ranking model uses a well-known ranking formula [12].

Evaluation

We evaluated two baseline systems using a dedicated annotation tool, see Figure 1. Top-3 results of each model were presented to an assessor in random order. Each response was judged on a four-point scale (from 0 to 3, with corresponding emoticons in the evaluation interface). We did not use pooling, each model was evaluated by three assessors (authors of the paper) independently. Therefore, some stimulus-response pairs were evaluated more than once by an assessor. As test stimuli, we selected 34 speech recognized utterances from Everfriends mobile application[2] log и 47 questions from Humor category of a popular Russian community question answering platform[3]. Table 2 cites some test stimuli.

Source	Test stimulus
Otvety@Mail.Ru	Почему Винни-Пух просто не отпустил шарик?
	Что может вас насторожить?
	А зачем вы спящую жену солью посыпаете?
	Вы сколько дней можете без первого, второго и третьего?
Everfriends	спой мне песенку
	где можно спрятать труп
	как завоевать сердце девушки
	как отпраздновать свой день рождения

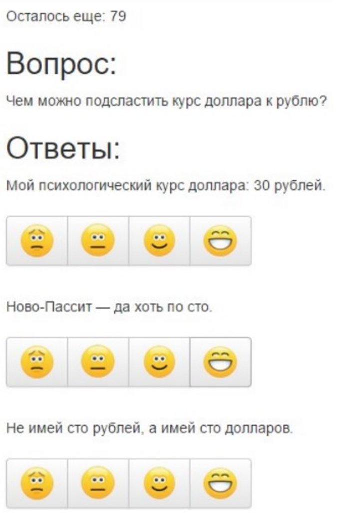
Table 2. Examples of test stimuli.

Relevance score for a query–document pair is an average over all assessors' labels. Table 3 shows exemplary stimuli and the responses of the systems along with averaged assessors' judgments. We employed Discounted Cumulative Gain (DCG) [13] as quality metrics. Table 4 shows DCG@1 (in a nutshell, averaged score of the top-ranked results for 81 queries) and DCG@3 scores for both baseline systems; for the whole collection and the subcollection without obscene words.

We also calculated Cohen's kappa [14] as a measure of inter-annotator agreement. We used weighted variant (weights are absolute differences between labels) for pairwise agreement. Averaged pairwise kappa statistics for three assessors in our experiments is 0.39. Averaged self-agreement for three assessors is 0.71 (recall, there are multiple judgments by the same assessor for a subset of query–document pairs).

[2] http://www.everfriends.com/, now discontinued.

[3] https://otvet.mail.ru/humor



Осталось еще: 79

Вопрос:

Чем можно подсластить курс доллара к рублю?

Ответы:

Мой психологический курс доллара: 30 рублей.

😟 😐 🙂 😁

Ново-Пассит — да хоть по сто.

😟 😐 🙂 😁

Не имей сто рублей, а имей сто долларов.

😟 😐 🙂 😁

Figure 1. Annotation tool.

Score	Stimulus	Response
2.67	Почему ты молчишь?	Да просто не на что жаловаться.
2.50	Как прошел твой день?	Как слон по мандаринам!
2.50	Твое любимое блюдо?	Ну, я неплохо ставлю чайник!
2.29	У вас есть план?	Выжить.
2.25	Ну давай познакомимся поближе.	Нет, я жду кота, он сейчас подойдет.
1.83	У вас есть план?	Пока никаких. А какие есть предложения?
1.50	На кого обычно похожи дети?	На нас.
0.75	Откуда брать виртуальные деньги?	Бери деньги, пойдем покажу!
0.17	А как с буржуазией разговаривать?	Буржуазия!
0.00	Как вы реагируете на хамов?	На свет.

Table 3. System's responses and their evaluation scores.

Discussion and Future Work

Evaluation results show that accounting for context is crucial for humor generation in dialog. The comparison of retrieval results for the whole collection and the 'decent' subcollection (86% of the whole collection) implies that corpus size is crucial for IR-based response generation approach. Assessors' agreement is fair, taking into account the number of judges (three) and four-graded judgments. Thus, we can hope to obtain reliable evaluation results in controlled laboratory settings even for such a seemingly subjective domain as humor.

The main contribution of the study is building a basis for further research consisting of the test collection (jokes, partially marked as obscene, and queries), annotation tool, and evaluation of two baseline models.

In the future we plan to address dialog modeling using large datasets, e.g. now we are investigating the possibility to harness sitcom subtitles archives to this end. Another direction of our future research is plugging in word sense similarity (in the form of word embeddings or using thesauri) into retrieval models. Moreover, we plan to implement simple pun generation heuristics based on ambiguous or similar-sounding words.

	DCG@1		DCG@3	
	w/o obscenity	whole collection	w/o obscenity	whole collection
Likes model	0.76	0.76	1.64	1.69
BM25	1.03	1.19	2.03	2.19

Table 4. Baseline evaluation results.

References

1. Bellegarda, J. R. Spoken language understanding for natural interaction: The siri experience. In Natural Interaction with Robots, Knowbots and Smartphones, 3–14 (2014).
2. 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, 171–191 (2013).
3. Khooshabeh, P., McCall, C., Gandhe, S., Gratch, J. & Blascovich, J. Does it matter if a computer jokes. In CHI, 77–86 (2011).
4. Stock, O. & Strapparava, C. Getting serious about the development of computational humor. In IJCAI, vol. 3, 59–64 (2003).
5. Valitutti, A., Toivonen, H., Doucet, A. & Toivanen, J. M. “let everything turn well in your wife”: Generation of adult humor using lexical constraints. In ACL (2), 243–248 (2013).
6. Ritchie, G. Computational mechanisms for pun generation. In ENLG Wokshop, 125–132 (2005).
7. Mihalcea, R. & Strapparava, C. Learning to laugh (automatically): Computational models for humor recognition. Computational Intelligence 22, 126–142 (2006).
8. Reyes, A., Rosso, P. & Veale, T. A multidimensional approach for detecting irony in twitter. Language resources and evaluation 47, 239–268 (2013).
9. Friedland, L. & Allan, J. Joke retrieval: recognizing the same joke told differently. In CIKM, 883–892 (2008).
10. Surdeanu, M., Ciaramita, M. & Zaragoza, H. Learning to rank answers to non-factoid questions from web collections. Computational Linguistics 37, 351–383 (2011).
11. Ritter, A., Cherry, C. & Dolan, W. B. Data-driven response generation in social media. In EMNLP, 583–593 (2011).
12. Jones, K. S., Walker, S. & Robertson, S. E. A probabilistic model of information retrieval: development and comparative experiments. Information Processing & Management 36, 779–840 (2000).
13. Järvelin, K. & Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. TOIS 20, 422–446 (2002).
14. Carletta, J. Assessing agreement on classification tasks: the kappa statistic. Computational linguistics 22, 249–254 (1996).