Computational Experiments on Detecting Meaning shift in Jokes

Eugeniia Zakovorotnaia,
National Research University
Higher School of Economics
Moscow, Russia
0000-0002-3655-4369

2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) | 978-1-6654-6480-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/SIBIRCON56155.2022.10017009

***Abstract*—The paper describes an experimental approach to detect the meaning shift, one of the most fundamental characteristics of humor, which is studied by many scientists in different interdisciplinary theoretical methodologies. We measured cosine similarity between setups and punchlines and explained these results through the set of objective criteria such as cosine results limitations, punchline length, three groups of words in joke's parts. We also decided to investigate the originality of obtained distribution of cosine similarity to the same distributions calculated for fiction texts. The results demonstrate crucial differences in distributions for all the verification texts. We described an automatic approach for extracting meaning concepts of setup and punchline aided by word embeddings of its top three semantically closest words. We also provide a comparative analysis of distribution of cosine similarities among jokes. The proposed approach allows obtaining embeddings for setup and punchline made on the top of the closest words in semantic space.**

***Keywords* —NLP, humor studies, experiments, BERT, joke**

I. Introduction

In computer linguistics, humor studies are mainly dedicated to classification, generation, corpus creation and literature review. However, there is lack of objective and formal features to define texts as jokes – in the majority of papers researchers accept the prior statement that all texts collected by automatic parsers from social networks and humor forums can be considered as humorous. This deficiency of definitive limitations becomes a primary problem for researchers working with such material because of human subjectivity. This paper describes an experimental approach to detect one of the most fundamental formal features of humor, named *meaning*, *semantic* or *frame shift*, which is the compatibility of two different concepts (*frames*) with one text, in other words.

The meaning shift in humorous texts is a key element of the incongruity theory [2,3]. Philosophical methodology argues that humor is based on a combination of contradictory concepts: good and evil, day and night, light and dark, etc. The incongruity theory separated into two linguistic methodologies – the semantic theory and the general theory of verbal humor [4,6]. The semantic theory claims that humor is based on the context ambiguity of different jokes' parts. As a result, if two frames overlap in a joke, then the ambiguity of interpretations becomes visible and leads to the humorous effect. All plots, ideas, characters, notions, events are presented by a frame (V. Raskin [*] also called it a *script*) that is a certain knowledge structure corresponding to the meaning of a word or a group of words. Formally, each script is a limited domain of the single continuous multidimensional graph which is the lexicon of the language [4]. Frames can be expressed by one or more polysemous words or expressions. In case a word is polysemous, the formal description will display a decomposition of its meanings expressed by more monosemantic words [4]. In jokes it is possible to observe different levels of frames' opposition, from a combination of two negations to alternatives within the same concept, for example, *good manners* vs *bad manners*. If opposite scripts fully or partially overlap on a text, they are compatible with it. In the first case, both frames are perfectly compatible with the joke, and there is nothing in them that could be perceived as odd, redundant, or absent. V. Raskin also introduces here a borderline situation when one script is more compatible with the entire text than the other. In case of the partial overlapping, there are some parts of the joke's text which are incompatible with both scripts. This effectively means that the combination of such scripts is unusual for the given text from the point of view of its correct semantic interpretation. Since Grice maxims are violated in humor, such texts are not perceived as meaningless, unlike in ordinary communication [4]. In addition to oppositional concepts, the semantic shift also involves a trigger; it reminds one of the first script when reading the second, imposes a different, non-obvious interpretation on it in a way that makes a compatibility of two opposite frames and joke's text more plausible and less impossible. The trigger is applied to figure out the mechanism that underlies the joke; the methodology distinguishes two main types: ambiguity and contradiction.

The general theory of verbal humor by S. Attardo and V. Raskin combines not only elements of semantics, but also textual linguistics, narrative theory and pragmatics. The methodology not only considers the opposition of frames, but also introduces additional criteria for analyzing the joke, named *Knowledge Resources*, which consist of a logical mechanism, a purpose, a narrative strategy, a language, and a situation. The trigger of semantic theory is replaced here by the logical mechanism, which explains how two opposite meanings, named *scripts*, are compatible with a text, and how the meaning shift appears during reading jokes. This notion includes not only ambiguity and contradiction, but also false analogies, such as Garden-Path phenomena, or textual figure-ground reversals [6].

The frame or semantic shift in humorous texts has been studied using qualitative [7-9] and quantitative approaches [10-11] in interdisciplinary research. The relevance of this work is to develop a methodology of objective and quantitative criteria for finding and describing meaning shift in jokes, as well as using a pre-trained neural network with Encoder architecture, BERT. The discovered results will be further developed for automatic search of triggers or logical mechanisms by which the semantic shift occurs. We worked with a corpus which consists of 1100 short humorous texts written in the English language, which were taken from the book "1001 Jokes" [5] by psychologist Richard Weissman. The experimental approach on the dataset contains jokes with variable syntactic structures (jokes-riddles/question-answer, one-shot liner, puns, etc.) and linguistic characteristics (presence/absence of language play, presence/absence extralinguistic data), as well as with different lengths, topics, sentiment let to analyze the effectiveness of objective criteria in the formal description of such a complicated, abstract, subjective phenomenon as humor.

II. Detection of Meaning Shift

Most researchers divide jokes into two parts: setup, where a first frame is introduced, and punchline, where the first frame is replaced with a second. In this paper, a punchline is always represented by the last sentence of a joke or the final part of this sequence; the rest of the joke is considered as the setup. The idea of the meaning shift detection experiments is to calculate the cosine similarity between two parts of jokes, which would show the semantic difference between two frames compatible with the same text.

The first experiment concerns the setup and punchline of a joke. Preprocessed and marked up humorous texts were used as an input of the pre-trained BERT neural network, the 'bert-base-uncased' model [1] which creates word embeddings based on its context. This approach has advantages over the frame ontologies, as it does not require the construction of complex graph dependencies between semantic meanings of polysemous words. In addition, in this research, context-dependent vector representations of individual parts of a joke can be considered as analogous to frames. The amount of data used for BERT training allows the model to take into account different semantic meanings of words and their transformation in the contexts of joke parts, therefore, their full semantic concepts are preserved.

We calculated embeddings for setup and punchline by averaging its words embeddings and measured cosine similarity between them. While analyzing the results, there is a set of the following criteria.

- Three conditional intervals of the cosine similarity: [0; 0.5], (0.5; 0.7), [0.7; 1]. The introduction of this criterion relies on the theoretical proposition of semantic theory about two scripts overlapping, fully or partially, on the text of a joke. Thus, jokes belong to the third interval since corresponding texts are compatible with both scripts despite their opposition. Jokes belong to the first interval since two opposite frames are compatible partially on the text of a joke. Finally, there is a borderline case, when one of the scripts is more compatible with the text than the second one. This distribution is explained by the presence or absence of certain lexical data.
- Three groups of words able producing effect on cosine similarity results, if they occur both in setup and punchline:
    - similar word forms;
    - antecedents and their anaphoras, in other words, co-referential links;
    - synonyms or words closely located in the semantic space. Those were extracted from the setup for punchline's words by measuring the cosine distance.

  This criterion explains the meaning shift in jokes at the computational and structural levels: the cosine similarity increases in case of semantically similar words in different scripts of one joke. If two opposite frames are described by semantically similar words, the joke text is more compatible with each of them.

- Length of punchline – the shorter it is, the more unambiguous its context.

The results showed that 91 jokes have cosine similarity < 0.5; the majority (688) belongs to (0.5; 0.7) interval; the rest of jokes (320) have scores > 0.7.

The distribution of the results of cosine similarity between setup and punchline context-dependent embeddings is shown in Fig. 1.

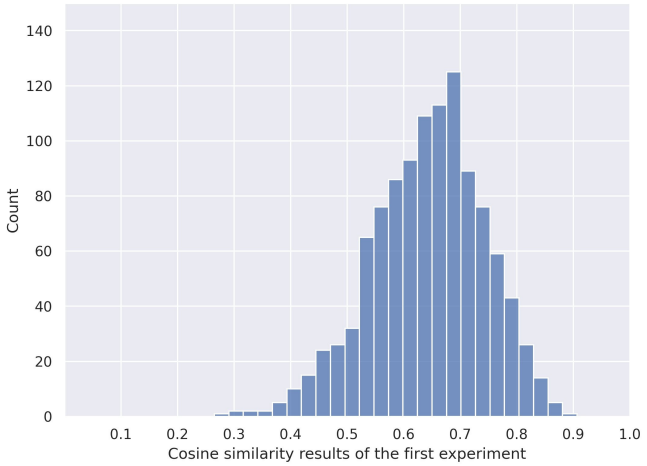

Fig. 1. The results of cosine similarity between setup and punchline context-dependent embeddings

In the first group (91), antecedents and their anaphoras are presented in 71% of humorous texts, while similar word forms are absent both in setup and punchline in 80% and synonyms/semantically related words are absent in ~87%. The average length of punchline is ~ 2 words, the maximal length is 11 words.

In the most numerous group (688), similar word forms occurred in setup and punchline in 61% cases, synonyms/semantically related words and co-referential links occurred in ~47% and ~48%, respectively. 15% of the punchlines have a length of 5 words. The average length of punchline is 5 words, the maximal length is 12 words.

The rest group, which amounts to a third of the entire corpus (320), is described by a high content of three word groups in setup and punchline. The synonyms/semantically related words were mentioned in 94% of humorous texts, while both similar word forms and co-referential links only in 70% and 73%, respectively. The average length of punchline is 7, the max length is 32 words.

Qualitative analysis of the cosine similarity between context-dependent vectors of setup and punchline indicates the compatibility of different parts of the joke with the same text, hence the meaning shift is detected. 81% of the jokes from the corpus have similar vocabulary in setups and punchlines, this explains the obtained scores of cosine similarity for most of the jokes. The prevalence of cosine similarity scores in the range (0.5; 0.7) reflects a compatibility of the joke texts with at least one of the scripts, while in a third of the corpus (cosine similarity > 0.7) opposite frames overlap fully on jokes. This indicates that a large part of the jokes are more characterized by the high compatibility of different scripts within texts by using similar vocabulary in different parts of the joke.

We also decided to investigate the originality of obtained distribution of cosine similarity to the same distributions calculated for fiction texts. For this sake we applied our measurement to the corpus of 3300 pairs of sentences from masterpieces originally written in the English language ("Dracula" and "Dracula's Guest" by B. Stoker, "The Inimitable Jeeves" and "My Man Jeeves" by P.G. Wodehouse, "Good Omens" by T. Pratchett and mixtures, 1100 sentences for each author). These books were chosen because our aim was to oppose different genres such as horror and satire, respectively since they may contain incongruity and contrast them on semantic and emotional levels. The pairs of sentences were selected as follows. In the beginning, we extracted sentences at even positions in the texts. If the length of this sentence belonged to the interval of the average lengths of the punchlines, from 2 to 7 words,

978-1-6654-6480-2/22/$31.00 ©2022 IEEE

then we considered this sentence as a punchline and a previous one as a setup.

We evaluated cosine similarity results in two samples among jokes and 1100 pairs of sentences for each author using the t-test. For P.G. Wodehouse p-value was 2e-52, for T. Pratchett the indicator showed p-value=8e-57, for B. Stoker p-value was 2e-54, for a corpus with pairs of sentences mixed from all books p-value was 3e-55. The results demonstrate crucial differences in distributions for all the verification texts, which means that there are unique semantic links between the investigated parts of jokes.

The second experiment aims to extract descriptive plot concepts (frames) from setup and punchline, measure the cosine similarity between them, and compare them with the first results. The idea of this experiment is that the higher values of the cosine similarity calculated between a part of a joke and its word lead to the higher influence of the context-dependent embedding of that word on the formal representation of this part. Consequently, we assume that a set of such context-dependent vectors express the main concept of a part of a joke. The advantage of this approach is deletion of similar words in setup and punchline, which increases the cosine similarity score. In the beginning, the cosine similarity was measured between the joke part embedding and each joke part's word embedding. Based on the results, three closest words were selected for setup and punchline. Articles were removed from the tops of the setups' and punchlines' closest words. The second joke part embedding was calculated with the resulting top of closest word embeddings by the arithmetic mean. The final distribution is shown in Fig. 2.

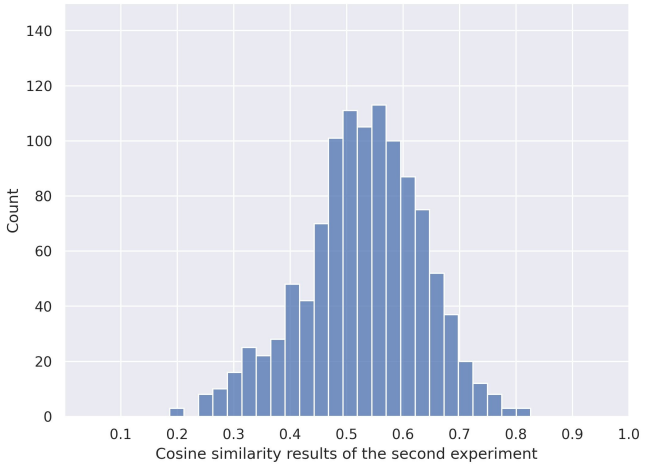

Fig. 2. The results of cosine similarity between setup and punchline context-depended embeddings based on closest words to them

Comparative analysis of the cosine similarity calculations between the context-dependent vectors of the first and second experiments divided the corpus into two parts. In 10% of cases, negative results appeared and the cosine similarity between the second joke part embedding turned out to be higher than between the first one. In 60% of jokes, the difference between the first and second cosine similarity is <= 0.1, while only 8% cases are > 0.2. Nevertheless, the division of jokes on three cosine similarity limitations by the second cosine similarity results has changed: ~36% jokes (400) entered the first group, where is cosine similarity < 0.5; the majority (656) has scores to (0.5; 0.7); the remaining (43) belongs to cosine similarity > 0.7. It's important to mention that all of the jokes in the third group have synonyms/semantically related words. Comparative analysis of the first and second experiments results showed little difference between them, which means that the idea of the second experiment is justified, since the main concepts of the jokes are extracted. However, the vector representations from the second experiment cannot be considered an unstructured script analogy; for more than a third of the jokes, the initial distribution within the cosine similarity intervals changed, and consequently, their characteristics about frame opposition and the level of their compatibility with the text.

III. Conclusions

In this paper we demonstrated that the meaning shift can be detected using averaged embedding vectors of setup and punchline of a joke written in the English language. For such detection, we introduced several formal criteria. We also proved a hypothesis that cosine similarity between setup and punchline of a joke has a distribution that differs from one calculated over neighboring sentences of a fiction text. Finally, we automatically detected top-3 word embeddings from setup and punchline; however, such an approach was not fruitful.

Aiming to reflect the meaning shift, the first experiment offered an explanation of the results of the cosine similarity measurement through a set of formal criteria: cosine results limitations, punchline length, three groups of words in joke's parts. The presence of similar words on orthographic and semantic levels results in high cosine similarity not for the total corpus but only for its major part (81%). In addition, there aren't any explicit distinctions for each of the cosine similarity intervals since all categories of similar words are presented in all three intervals. Nevertheless, the average punchline length is remarkably different for jokes with full and partial overlapping of opposite scripts.

We also found that the details of the criteria proposed in the first experiment need to be refined; our research lacks detailed qualitative analysis of jokes with the high cosine similarity. However, it is possible to detect the meaning shift using its description by the objective and formal criteria. The application of the introduced method, based on the theoretical provisions of the semantic theory and the general theory of verbal humor, has shown positive results. So, the hypothesis of the first experiment was proved.

We calculated cosine similarity to the same distributions for fiction texts in order to demonstrate its distinction from the distribution for short jokes. Statistical pairwise verification of samples of jokes and pairs of sentences from fiction texts demonstrates the results of p-value < 0.05 (very close to 0), which means the presence of unique semantic links between parts of jokes. The experiment with fiction texts shows the importance of conducting an extended comparative analysis and description of jokes with short texts from other genres (fiction books, scientific articles, and news wire) to characterize the semantic shift as one of the basic criteria of jokes and humor in general.

In the second experiment we applied the automatic approach to extract setups' and punchlines' plot concepts aided by word embeddings of its top three semantically closest words. The idea of this approach inspired by the hypothesis that the closest contextualized vector representations of such words exerted a strong impact on the setup's and punchline's main context, thus, it allows obtaining embeddings for setup and punchline made on the top of the closest words in semantic space. Though a comparative analysis of the cosine similarity measurements of the first and second experiments showed non-significant quantitative differences for most of the corpus examples in the interval (-0.033; 0.3764), the initial distribution of the cosine similarity measurements changed for the third of the jokes as well as their characterization about frame opposition and the level of their compatibility with the text. The second experiment is also considered as a successful one, because the main concept of a joke is indeed extracted by this automatic approach according to the results of comparative analysis. However, the averaged vectors of the top of the closest words to the parts of jokes in semantic space should not be considered as their analogues of full context embeddings.

978-1-6654-6480-2/22/$31.00 ©2022 IEEE

The future work includes reconsideration of the automatic approach for scripts extraction, as well as investigation of details of above-mentioned objective criteria. The list of formal characteristics of jokes should be expanded with the reliance on both semantic theory and other theoretical methodologies.

IV. References

[1] J. Devlin, M.-W.Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp.

[2] J. Morreall, "Philosophy of Humor", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.), 2020

[3] V. Raskin, S. Attardo, "Script theory revis(it)ed: joke similarity and joke representation model", Humor - International Journal of Humor Research, 4(3–4), 1991, pp.293-348

[4] V. Raskin, "Semantic Mechanisms of Humor", Springer Netherlands, Dordrecht, 1984 (*references*)

[5] R. Wiseman, "1001 joke", [Online]. Available: https://www.studocu.com/row/document/thammasat-university/new-media-studies/quirkology-1001-jokes-by-richard-wiseman/20822728

[6] A. Salvatore, "Linguistic Theories of Humor", Mouton de Gruyter, New York (1994).

[7] Zh. X. Yong, P. D. Watson, T. T. Torrent, Ol. Czulo, C. F. Baker, "Frame Shift Prediction", 2022, [Online]. Available: https://www.researchgate.net/publication/357645907_Frame_Shift_Prediction

[8] C. S. Q. Siew, T. Engelthaler., T. T. Hills, "Nymph piss and gravy orgies: Local and global contrast effects in relational humor", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(7), 2022, pp. 1047–1063, https://doi.org/10.1037/xlm0001120

[9] S.Frenda, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, P. Rosso, "The unbearable hurtfulness of sarcasm", Expert Systems with Applications, Vol. 193, 2022, https://doi.org/10.1016/j.eswa.2021.116398

[10] K. Binsted *et al*., "Computational humor," in *IEEE Intelligent Systems*, vol. 21, 2006, no. 2, pp. 59-69, doi: 10.1109/MIS.2006.22

[11] L. Gabora, K. Kitto, "Towards a Quantum Theory of Humour", Frontiers in Physics (section: Interdisciplinary Physics), 4. 10.3389/fphy.2016.00053