



Comparing Jokes with NLP: How Far Can Joke Vectors Take Us?

Xiaonan Jing, Chinmay Talekar, and Julia Taylor Rayz^(✉)

Purdue University, West Lafayette, IN 47906, USA
{jing, ctalekar, jtaylor1}@purdue.edu

Abstract. This paper describes the results of comparison of jokes based on neural network language models and the General Theory of Verbal Humor knowledge resources. The question to be explored is whether joke similarity can be inferred from joke text similarity. Our results indicate that while similarities in text can be detected by doc2vec and jokes with somewhat similar texts can be consistently clustered, there is very little, if any, correlation between doc2vec cosine similarity and KR-annotated joke similarity.

Keywords: Doc2vec · Computational humor · General theory of verbal humor Clustering

1 Introduction

The usefulness and motivations of computational humor have been intensely discussed [1–10], with applications varying from friendlier computer systems and human-computer interfaces [2, 6, 11] and more effective communication [12] to education [13–15] and “edutainment” [16], to advertising, politics and commerce [7, 17], to information security [7], to detection of unintended humor [7, 8, 18] to embodied agents with personalities [6] to understanding how human brain works [4, 9]. It can also be important to detect and exclude humor from reporting it as accurate information because humor is, in principle, not committed to the literal truth of the statements [19].

Accepting that computational humor is important, the question of its feasibility arises. Within text-based humor, a number of humor generators, computational humor recognizers/detectors, and a few comparisons of jokes have been attempted. We will concentrate on the latter two categories as computational humor detection can be either done by comparing a text with a previously known set of humorous text – and if a new text is close enough to one of the known ones, classify it as humorous – or, attempt to classify it based on some features, that, hopefully, have some (humor) theoretical foundations [20].

Machine learning classification methods have been used to separate texts into humorous and non-humorous subsets, with a possible analysis of the contrastive features [21–27]. These results, while impressive, have not, however, led to any insights into the nature or structure of verbal jokes. Semantic methods, tightly associated with humor theories, tend to be more illuminating in this respect [19, 28], but still emphasize the distinctions between jokes and non-jokes rather than the constitutive features of the

former, including optimality or even elegance. A SemEval 2017 pun interpretation and detection task forced the participants into using semantic features for pun interpretation task, however, the results show that it remains an “extremely challenging problem” [29]. One of the reasons why this task is so hard is that not only does it require a detection of semantic features, an additional component is reasoning with them to detect the salient semantic features.

Joke comparison could work around the problem of reasoning detection/understanding, if the jokes are similar enough. The question that we are interested in exploring for this paper is *whether joke similarity can be inferred from joke text similarity*.

For this paper, we assume that any joke can be described in terms of form, content, and use [30]. The General Theory of Verbal Humor [31] captures the overlap between use and content. The knowledge resources of GTVH are used as a vector for joke description and comparison. Five GTVH features for joke comparison were experimentally confirmed [32] on perception of 534 human objects; the only exception being (Pseudo) Logical Mechanism. Word embedding algorithms, infamous by now within natural language processing, suggest capturing the overlap between form and content.

In the past five years semantic similarity between words or texts started to look more promising. Semantic similarity between two texts can be measured by lexical similarity (surface level) or, more popular, word embedding similarity. Word embeddings, following the distributional hypothesis of [33], are trained on word co-occurrence in large corpora and are reported to capture semantic word similarity. The most popular word embedding models are SkipGram [34], CBOW [35] and GloVe [36]. Word vectors, produced by these models, allow not only a comparison of word similarity, but also semantic similarity of patterns [37]. While original word2vec built one representation per word (not per word sense), there has been work in recent years that took multiple senses of words into account [38–40] as well as phrase embeddings [41, 42].

[43] introduced paragraph2vec, often called doc2vec, that is based on word2vec idea, and [44] confirmed its performance. In paragraph2vec, “every paragraph is mapped to a unique vector” (*ibid*). The paragraph vector, together with the word vectors serve as input for the target word prediction, very similar to the CBOW method. The method can be applicable to texts of various length, such as sentences, paragraphs, longer documents, as represented by the additional input vector and can be used for document classification [45, 46].

The compared jokes are taken from a dataset by [47], containing approximately 11,000 jokes. The jokes are clustered according to their vector similarity, with Elbow method [48] to determine the optimal number of clusters. Based on the results of the comparison, the final version of the paper will discuss the feasibility of the current natural language processing techniques for comparing some of the Knowledge Resources of the General Theory of Verbal Humor.

1.1 Corpus

This paper adopted the web archive joke dataset used in [47]. The dataset includes 10,942 jokes from 13 joke archive websites. It allows the same joke with different

versions (identical or nearly identical) to present multiple times. The lengths of the jokes vary from one-liner jokes to short scripts. We also included the 21 labeled jokes from [32] to the corpus for the training of the joke vectors.

1.2 Joke Vectors Clustering

We trained a Doc2vec [49] model with the open source module Gensim v3.1.0 (with parameters: window size 10, dimensionality 100, and minimum count 1). For the learned joke vectors, we applied the k-means++ clustering algorithm with a cosine

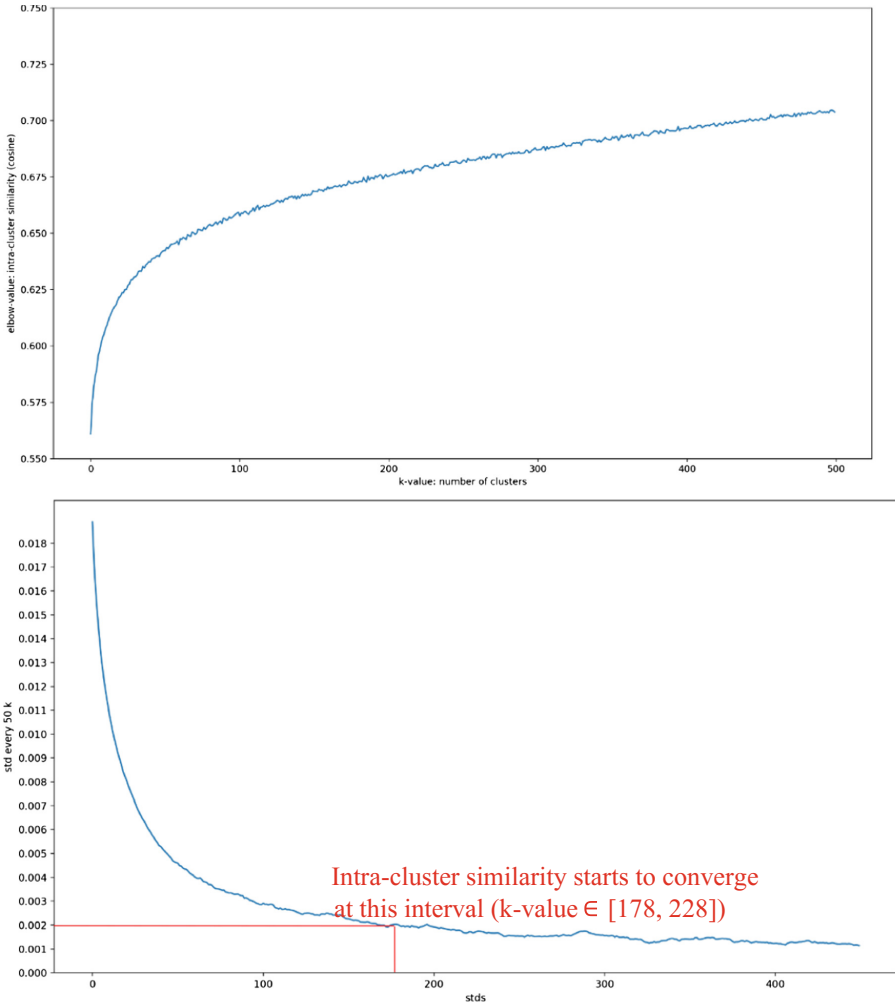


Fig. 1. (1-a top) number of cluster k vs. intra-cluster cosine similarity. (1-b bottom) standard deviations for every 50 consecutive intra-cluster cosine similarities, the standard deviation starts to converge at the interval $k = [178, 228]$, select the k in this interval where the intra-cluster similarity takes minimum between k and $k + 1$.

similarity metric to group them into 181 clusters. K-means++ uses an optimized center initialization process on top of the general k-means algorithm. The number of cluster was determined through the Elbow method used in [50] by testing the k-value equals from 1 to 500 (Fig. 1-a). For every 50 consecutive intra-cluster similarity values, the standard deviation was computed and the interval [178, 228], for the proper k-value, was selected where these standard deviations start to converge (Fig. 1-b). The k-value was selected to be the proper number of clusters for the corpus where the absolute differences between the intra-cluster similarity takes a minimum – select such k when $\min (|I_k - I_{k+1}|)$ for k in [178, 228].

We conducted the k-means++ clustering experiments 10 times using the same initial random seeding on the joke vectors with k-value equals to 181 to study the stability of the joke clusters as well as to compare the computational results with the manually labeled GTVH-based KRs. As k-means++ uses an optimized probability-based random seeding method to select the initial cluster centroids from the dataset, it provides more stable and uniformly distributed centers across multiple experiments.

2 Script Oppositeness and Logical Mechanism

The General Theory of Verbal Humor [31] describes jokes in terms of six knowledge resources: Script Overlap/Oppositeness (SO) – originally introduced in Script-based Semantic Theory of Humor [51], – Logical Mechanism (LM), Situation (SI), Target (TA), Narrative Strategy (NS), and Language (LA). In order to compare jokes, using knowledge resources, it helps to take an inventory of what is available as features for each of the resources.

For a reader unfamiliar with humor theories, a text is considered to be joke-carrying, if it is compatible, fully or in part, with two scripts that overall and oppose [51]. Raskin introduces several high-level types of SOs:

- Real/unreal category
 - Actual vs. non-actual
 - Normal, expected state of affairs vs. abnormal, unexpected state of affairs
 - Good vs. bad
 - Life vs. death
 - Possible, plausible situation vs. fully or partially impossible or much less plausible situation
- Sex/non-sex opposition
 - Overt, unspecified
 - Overt, specified
 - Non-sexual opposition in explicitly sexual humor
 - Specific sexual opposition in explicitly sexual humor
- Ethnic category
 - Language distortion
 - Dumb vs. smart
 - Stingy vs. generous

- Cunning vs. honest/straight forward
- (non-standard scripts, based on ethnicity)
- Political category
 - Denigration jokes
 - A political figure
 - A group or institution
 - A political idea
 - Exposure jokes
 - National traits
 - Political repression
 - Shortages
 - Specific political situation

It should be noted that a joke does not have to be specific to just one sub-category. For example, Raskin notes that political jokes that expose national traits (SSTH: Chapter 7, Sect. 4) are close to ethnic humor; and the script of sexuality can be used in ethnic humor (SSTH: Chapter 6, Sect. 5).

Each joke should, theoretically, fall under at least one of these SOs, however, it is important to remember that most SOs identified for a particular joke are likely to be finer grain, especially, for comparison within a category. For example, jokes (1) and (2) are both of the Normal/Abnormal SO type, jokes (3) and (4) are of the Possible/Impossible SO type, but each could be analyzed further to the SO below (Chapter 4, Sect. 4):

- (1) Should a person stir his coffee with his right hand or his left hand? Neither. He should use a spoon. [51, originally 53:21].
 - SO: hand with tool vs. bare hand
- (2) When is a joke not a joke? Usually. [51, originally 53:21]
 - SO: good joke vs. bad joke
- (3) Why is a drawn tooth like a thing forgot? Because it is out of one's head. [51, originally 53:22].
 - SO: head vs. mind
- (4) Why does a donkey eat thistles? Because he's an ass. [51, originally 53:23].
 - SO: animal vs. abuse

As can be seen, a result of a joke comparison based on knowledge resources (KRs) is trivially dependent on the grain size that is specified for KRs. Jokes (1) and (2) may or may not have the same SOs, depending on what an annotator specified. For the purposes of this paper, we will stay with the hierarchy introduced above, but will not create finer grain hierarchy.

The next KR that can be described as a hierarchy of features is the Logical Mechanism. A Logical Mechanism is a resource that describes the connection between scripts. [52] present a list of all known LMs (Fig. 2) and a taxonomy of some of them:

role-reversals	role exchanges	potency mappings
vacuous reversal	juxtaposition	Chiasmus
garden-path	figure-ground reversal	faulty reasoning
almost situations	analogy	self-undermining
inferring consequences	reas. from false prem.	missing link
coincidence	parallelism	implicit parall.
proportion	ignoring the obvious	false analogy
exaggeration	field restriction	Cratylism
meta-humor	vicious circle	referential ambiguity

Fig. 2. List of known LMs [52]

- Syntagmatic Relationships
 - Reversals
 - Vacuous
 - Actantial
 - Role
 - Role Exchange
 - Potency Mapping
 - Figure/Ground
 - Chiasmus
 - Direct Spatial Relationships
 - Juxtaposition
 - Sequence
 - Parallelism
 - Proportion
 - Implicit Parallelism
 - Explicit Parallelism
- Reasoning
 - Correct
 - From False Premise
 - Missing Link
 - Almost Situation
 - Coincidence
 - Analogy
 - Faulty
 - Cratylism (puns)
 - Exaggeration
 - Ignoring the Obvious
 - Field Restriction
 - False Analogy

- Meta
 - Metahumor
 - Garden Path
 - Self-reflexive

Using these resources as a starting point, it is possible to annotate jokes that we attempt to analyze. We are interested in comparing 21 jokes, originally used in [32] relative to jokes adopted from [47].

3 GTVH-Based Jokes Comparison: Between Jokes and Noise

The paper uses 3 sets of jokes, listed below, taken from [32].

Set 1:

- Anchor1: What do you call it when a blonde dyes her hair brown? Artificial Intelligence.
- LA1: What's the result of a blonde dyeing her hair brown? Artificial Intelligence.
- NS1: When a blonde dyes her hair brown, it's called Artificial Intelligence.
- TA1: What do you call it when a fair-haired sorority girl dyes her hair brown? Artificial Intelligence.
- SI1: What do you call it when a blonde "lipsyncs" Einstein on the screen? Artificial Intelligence.
- LM1: What do you call it when a blonde dyes her hair brown? Illiteracy: she could not read the label on the bottle.
- SO1: What do you call it when a blonde dyes her hair brown? Serial murderer: her five boyfriends hanged themselves.

Set 2:

- Anchor2: Why did the chicken cross the road? It wanted to get to the other side.
- TA2: Do you know the reason why the chicken decided to cross the road? Because it wanted to get to the other side.
- NS2: The reason the chicken crossed the road is that it wanted to get to the other side.
- TA2: Why did the turtle cross the road? It wanted to get to the other side.
- SI2: Why did the chicken eat an octagonal-headed worm? Because it was hungry.
- LM2: Why did the chicken cross the road? Nothing ventured, nothing gained.
- SO2: Why did the chicken cross the road? He saw a blonde hen on the other side.

Set3:

- Anchor3: How many Poles does it take to screw in a light bulb? Five. One to hold the light bulb and four to turn the table he's standing on.
- TA3: The number of Polack's needed to screw in a light bulb? Five – one to hold the bulb and four turn the table.

- NS3: It takes five Poles to screw in a light bulb: one to hold the light bulb and four to turn the table he's standing on.
- TA3: How many Irishman does it take to screw in a light bulb? Five. One to hold the light bulb and four to turn the table he's standing on.
- SI3: How many Poles does it take to wash a car? Two. One to hold the sponge and one to move the car back and forth.
- LM3: How many Poles does it take to screw in a light bulb? Five. One to hold the light bulb and four to look for the right screwdriver.
- SO3: How many Poles does it take to screw in a light bulb? Five. One to take his shoes off, get on the table, and screw in the light bulb and four to wave the air deodorants to kill his foot odor.

The KRrs for the third set, LightBulb, were previously labeled by [31], as shown in Table 1. The KRrs for the Blonde set and CrossingTheRoad set are demonstrated in Table 2.

Table 1. Knowledge resources of jokes in Set3 [31]

Joke/KR	LA	NS	TA	SI	LM	SO
1	LA1	Riddle	Poles	Light bulb	Figure-ground	Dumbness
2	LA2	Riddle	Poles	Light bulb	Figure-ground	Dumbness
3	LA1*	Expository	Poles	Light bulb	Figure-ground	Dumbness
4	LA1*	Riddle	Irish	Light bulb	Figure-ground	Dumbness
5	LA1*	Riddle	Poles	Car wash	Figure-ground	Dumbness
6	LA1*	Riddle	Poles	Light bulb	False analogy	Dumbness
7	LA1*	Riddle	Poles	Light bulb	Figure-ground	Dirtyness

Table 2. Knowledge resources of jokes in Set1 and Set2.

Joke/KR	LA	NS	TA	SI	LM	SO
Anchor1	Std	Riddle	Blonde	Dyeing hair	Faulty reasoning	Dumbness
LA1	Change	Riddle	Blonde	Dyeing hair	Faulty reasoning	Dumbness
NS1	Std	Expository	Blonde	Dyeing hair	Faulty reasoning	Dumbness
TA1	Std	Riddle	Sorority girl	Dyeing hair	Faulty reasoning	Dumbness
SI1	Std	Riddle	Blonde	Lipsyncing	Faulty reasoning	Dumbness
LM1	Std	Riddle	Blonde	Dyeing hair	Missing link	Dumbness
SO1	Std	Riddle	Blonde	Dyeing hair	Faulty reasoning	Life/death
Anchor2	Std	Riddle	n/a	Crossing road	Faulty reasoning	Implausible
LA2	Change	Riddle	n/a	Crossing road	Faulty reasoning	Implausible
NS2	Std	Expository	n/a	Crossing road	Faulty reasoning	Implausible
TA2	Std	Riddle	n/a	Crossing road	Faulty reasoning	Implausible
SI2	Std	Riddle	n/a	Eating worm	Faulty reasoning	Implausible
LM2	Std	Riddle	n/a	Crossing road	Metahumor	Implausible
SO2	Std	Riddle	n/a	Crossing road	Faulty reasoning	Overt sex

There are several observations that can be made by comparing Tables 1, 2 and the actual texts of the jokes in the three sets. The first one is that even if we stay at the high level description of the jokes, the difference between jokes in X1/Y1 and X2/Y2 and X3/Y3 can be predicted based on the distance of values in the resources. In other words, if $x = \text{LM}$ and $y = \text{SO}$, the jokes in Set 3 vary in LM (false analogy vs. figure ground), which differ at the highest level of the LM hierarchy, and SO (dumbness vs dirtiness), which, arguably, both belong to the ethnic category. The jokes in Set 2 vary in LM (metahumor vs. faulty reasoning), which differ at the reasoning level (but not the highest level), and the SO (implausible vs. overt sex) differ at the highest level. The jokes in Set 1 vary in LM (missing link vs. faulty reasoning), which differ in the reasoning level again, and the SO (dumbness vs. life/death) vary at the highest level again. Thus, these jokes vary at the highest SO level and second highest LM level in Sets 1 and 2, and at, at most, the second highest SO level and highest LM level in Set 3.

According to the GTVH, jokes that only vary in SO should be less similar than jokes that vary only in LM. It is reasonable to assume that jokes that only vary in SO at the highest level of hierarchy are less similar than jokes that only vary in SO at a finer grain level. Similarly, jokes that only vary in LM at the highest level of the hierarchy (the coarsest level) should be less similar than jokes that only vary in LM in at the finer grain size. It should be noted that LM placement in resource placement hierarchy was not confirmed experimentally [32]. However, if one is to follow theoretical predictions [31] and common sense, LM and SO jokes in Sets 1 and 2 (SO1/LM1 and SO2/LM2) should be more different than jokes in Set 3 (SO3/LM3) as set 1 and 2 come with the highest difference in SOs and SO should carry more weight than LM (Set 3 comes with the highest level of difference in LM, but not SO). However, cosine similarity of joke vectors tells a different story: $\text{cosine}(\text{SO1}, \text{LM1}) = 0.55$, $\text{cosine}(\text{SO2}, \text{LM2}) = 0.42$, $\text{cosine}(\text{SO3}, \text{LM3}) = 0.18$.

There are several reasons why the theoretical prediction with common sense flavor is not met. The simplest one is that the knowledge resources are annotated at a very high level, and only one of the KR is chosen to be annotated (remember that there may be more than one SOs in each joke). On the opposite end of the spectrum, but perhaps, a more reasonable one is that joke vectors may not contain enough information to infer SOs and LMs as our hypothesis was that they can only measure the overlap between form and content, which corresponds to LA, NS, and possibly TA and SI.

In order to explore what joke vectors can actually capture, and whether it correlates with any of the knowledge resources, we compare the three sets of jokes to over 10,000 others, collected in [47]. We are interested in jokes that are most similar to the set of 21 jokes, again, according to joke vectors. The results are shown in Table 3.

It can be seen from the table that there is a curious grouping between jokes within the sets – four most similar jokes to the seven in the first set come from the same set; two most similar jokes to the seven in the second set come from the same set and one joke that does not come from the same set is clearly very similar to them; and three most similar jokes from the third set come from the same set and one joke that does not come from the same set is clearly very similar to them. While the rest of the jokes do not look like they have much similarity (and there is no seeming correlation in a non-technical sense of the word to the KRs), we were curious how the rest of the jokes in the dataset cluster.

Table 3. Most similar jokes to 21 from [32] and their cosine similarity

	Cosine sim	Joke texts
Anch1	0.88	TA1
LA1	0.71	Q: What’s the national bird of Iraq? A: DUCK!
NS1	0.64	Your Mama’s so stupid... she called the 7-11 to see when they closed!
TA1	0.88	Anchor1
SI1	0.81	TA1
LM1	0.70	Q: What do you call a bee that lives in a graveyard? A: A zombee!
SO1	0.73	TA1
Anch2	0.90	TA2
LA2	0.56	Q: How many programmers does it take to change a lightbulb? A: We looked at the light fixture and decided there’s no point trying to maintain it. We’re going to rewrite it from scratch. Could you wait two months?
NS2	0.71	Q: Why did the chicken cross the ‘net’? A: It wanted to get to the other site!
TA2	0.90	Anchor2
SI2	0.66	Q: What’s black and white and red all over? A: A zebra with chicken pox!
LM2	0.80	Borodin nothing to do!! Gone Chopin. Bach in a minuet.
SO2	0.66	How lawyers do it...
Anch3	0.80	Q: How many SAS programmers does it take to change a light bulb? A: How many?
LA3	0.73	NS3
NS3	0.77	Anchor3
TA3	0.73	Anchor3
SI3	0.66	Your Mama’s so stupid... she sold the car for gas money!
LM3	0.65	Q: What did the left eye say to the right eye? A: ‘Between you and me, something smells!’
SO3	0.63	Q: How does Stan Collymore change a lightbulb? A: He holds it in the air, and the world revolves around him

4 Clustering Results

We inspected the locations in the clusters for the labeled 21 jokes and identified the KRs contained in the same cluster for further analysis. Table 4 shows these KRs for the 10 experiments. The three sets of jokes, with Set1 being “blonde”, Set2 being “chicken”, and Set3 being “lightbulb” jokes, are labeled in different sets of colors in Table 4. The KRs appeared frequently together in the same cluster are labeled with the same color. For the “blonde” jokes, the frequently observed groups are (A1, TA1) and (SI1, SO1). Similarly, for the “chicken” and “lightbulb” jokes, the frequently observed groups are (A2, LM2), (LA2, TA2, NS2) and (LA3, NS3), (TA3, SO3) respectively. However, the clusters containing these jokes do not always stay consistent, their

final cluster center for a specific cluster. Cosine similarity between KR and its final center shows the cosine similarity of each KR to its final cluster center.

Three initial centers (joke223, joke1571, joke4869) that appeared multiple times when the KRs in Set1 are clustered together. As the initial centers were selected among the Dov2vec generated joke vectors, the original joke texts for the three centers are listed as following:

- joke223 - Q: What do you call a naughty monkey? A: A badboon!*
- joke1571 - Q: What do you call a blonde with 90% of her intelligence gone? A: Divorced.*
- joke4869 - What do you call a lesbian asian? Minjeeta*

Notice that the three jokes above have the same NS, namely riddles, which matches the NS KR of the majority of Set1 jokes.

Cluster Size. Another observation is that the cluster sizes stay relatively stable in the repeated experiments – the cluster initialized with joke223 contains more than 200 jokes with an exception of the resulting cluster in experiment R8; the cluster initialized with joke1571 contains around 100 jokes on average; the cluster initialized with joke4869 contains approximately 180 jokes. Figure 3 shows that there are approximately 25 clusters with sizes over 100 and among which approximately 5 clusters with size over 200 for each of the 10 experiments. While the cluster size distribution stays consistent across different experiments, it indicates that the KR-based jokes in larger clusters are consistently similar to more jokes in the corpus than those in the smaller clusters. i.e. A1 is similar to more jokes in the corpus than S11.

Initial-final center similarity. The centers shift as more jokes are added to the cluster. It is interesting to see how much the center has shifted from the initialization to its final

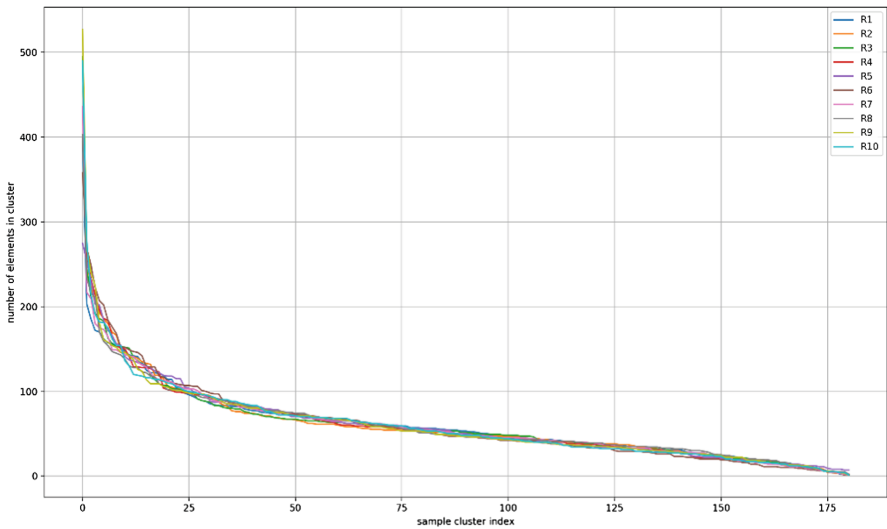


Fig. 3. Number of elements in clusters for experiments R1–R10

result in order to detect how similar the added jokes are thought the steps of addition. The initial-final cosine similarity column in Table 5 indicates how the initial centers moved as elements were added to a cluster. A high similarity value suggests very little movement between the initial and final center. Less movement of the center implies that the elements in that cluster have stronger associations with each other, according to how centers are updated during each iteration in k-means. In the case when the new elements added to the cluster have higher similarity with the rest of the elements that were in the cluster at the beginning of an iteration, the updated cluster center after this iteration should be very similar to the original center at the beginning of this iteration.

It can be observed that the cluster centers initialized with joke4869 moved much less, with an initial-final center similarity staying consistent around 0.942 (obtained by taking the average similarity of the clusters initialized with the same center), throughout the clustering process compared to the cluster centers initialized with joke223 and joke1571, which have the initial-final center similarity around 0.878 and 0.803 respectively. The differences between these three clusters may not seem very significant as the differences between their cosine similarities seem to be very small. However, cosine similarity measures the angles between two vectors in a vector space, a 0.942 cosine similarity suggests that there is approximately a 19° angle between the two vectors being compared. Visualization of hyper plane, in this case 100 dimensional, is not likely possible without information loss/dimension reductions. Nevertheless, a simple way to distinguish such differences could be done by inheriting the two-dimensional vector properties to higher dimensions. Cosine differences of 0.942, 0.878, and 0.803 correspond to, approximately, 19° , 29° , and 37° respectively. One needs to be aware of that n-dimension will also allow the vectors to move in n directions so the degree here to measure the difference between the initial and the final centers is merely a referencing number to compare how much the cluster centers have potentially moved during the clustering process.

Joke with highest similarity value to the final cluster center vector. As mentioned before, the cluster centers were updated after each iteration following the addition of new elements to the cluster. Hence, the final cluster centers do not directly correspond to any joke vectors in the corpus. We mapped each of these final centers (100-dimensional vectors) to its nearest joke vector in the same cluster, namely the joke vector that has a highest cosine similarity with its final center. The sample joke indices and the similarities with their centers for such joke vectors are also shown in Table 5. We could observe some consistency, especially for the clusters initialized with joke223 and joke4869, between the initial and final cluster centers. The clusters initialized with a center of joke223 are frequently associated with the final center represented by joke227; similarly, the clusters initialized with a center of joke4869 are frequently associated with the final center represented by joke4869, with two exceptions being R1 and R8. However, the jokes representing the final centers in these two exceptions, instead of being mapped to some other jokes, were simply switched – joke227 is representing the final center for cluster initialized with joke4869 and joke4869 is representing the final center for cluster initialized with joke223. Further inspection suggested that the clusters in both R1 and R8 contain both jokes; and it can be concluded that the inconsistency observed was due to the small differences caused by the cosine similarity between these

two jokes to their final cluster centers in repeated experiments. In addition, the Doc2vec model suggests that the first most similar joke to joke227 in the corpus is joke4869 with a similarity value of 0.935, and the second most similar joke to joke4869 is joke227, which indicates that in this case similarity is bidirectional in values, but not in rankings. It is possible that joke4869 and joke227, as final cluster center representations, are interchangeable and thus the center movements for clusters initialized with joke223 and joke4869 are consistent, respectively, across repeated experiments. However, the similarity between the two initial centers, joke223 and joke4869, is 0.858, which makes them quite distant and does not rank in top three for either of the most similar jokes for the two.

On the other hand, the clusters initialized with other centers seem to have less stable center movements across repeated experiments. The four clusters initialized with joke1571 in experiments R2, R5, R7, and R9 are presented with various final center representations. Whereas the cosine similarities for these final center representations are relatively low, suggesting the nearest joke vector in that cluster is further away from its final center, compared to the clusters with final center representations of joke227 and joke4869 mentioned above. In addition, although initialized with the same initial center joke1571, the three different final center representations, joke1178, joke1590, and joke1628 are not always together in the same cluster. Therefore, no consistent movements of the cluster centers initialized with joke1571 were observed, which also provides evidence for why the two KR-s, SI1 and SO1, which were frequently clustered together, moved extensively in repeated experiments. The repeated jokes that are closest to the final centers are shown below. It can be seen that the centers share NS with partial LA, and if a TA exists, they share it as well.

joke227 - Q: What do you call a wasp? A: A wanna-bee!

joke4869 - What do you call a lesbian asian? minjeeta

joke1178 - Q. How would a blonde punctuate the following: 'Fun fun fun worry worry worry' A. Fun period fun period fun no period worry worry worry....

joke2150 - What do you call a cheerful flea? A hop-timist!

Cosine similarity between KR-based jokes and their final cluster centers. The cosine similarities between the KR-based jokes and their final cluster centers are shown in the right half of Table 5. We focused our study on the five KR-based jokes from Set1 that appeared frequently together. These cosine similarities provided more evidence for explaining the movements of the jokes between different clusters across repeated experiments. The most stable joke observed is A1, which maintains approximately 0.878 cosine similarity with all of its final cluster centers. Furthermore, A1 only appeared along with the interchangeable final cluster center representations, joke227 and joke4869, as mentioned in the previous section. It is likely not accidental for A1 to be grouped to a cluster with similar final center in the repeated experiments. On the other hand, TA1, which was frequently grouped together with A1, showed slightly less similarities to each of its final cluster centers. However, it appeared to be in the same cluster with A1 seven out of nine times in 10 experiments, with an exception of R3 where A1 did not appear in a cluster along with other KR-based jokes. Although more between-cluster movements were observed for TA1 compared to A1, its stability is predictable - TA1 will be frequently grouped together with A1 in a large number of

such repeated experiments. The cosine similarities for another group of frequently observed jokes, SI1 and SO1, suggest that they are both less similar to their final cluster centers compared to A1 and TA1. Additionally, more between-cluster movement can be seen for SO1 than SI1, in respect to their lower cosine similarities to the final cluster centers. We do not provide an analysis for Set2 and Set3 due to space limitations, but they follow the same general pattern.

5 Epilogue

We were interested in analyzing whether text similarity correlates with GTVH-based jokes similarity. We compared 21 jokes annotated with GTVH-based knowledge resources to a set of over 10,000 jokes. Our results indicate that while similarities in text can be detected by doc2vec, and jokes with somewhat similar texts can be consistently clustered, there is very little, if any, correlation between doc2vec cosine similarity and KR-annotated joke similarity. One of the reasons may be that doc2vec detects form and content similarity, which could be mapped to LA and NS knowledge resources, TA if it exists, and, possibly SI when it is clearly stated. LM and SO are much more difficult to detect, especially at the courser grain level. While from the theoretical perspective it does not come as a particularly surprising result, it means that comparison of two texts by itself cannot classify a joke as a joke or a non-joke as a non-jokes as the presence of SO is a necessary component for a text to be a joke.

References

1. Raskin, V., Attardo, S.: Non-literalness and non-bona-fide in language: An approach to formal and computational treatments of humor. *Pragmatics Cogn.* **2**(1), 31–69 (1994)
2. Binsted, K.: Using humor to make natural language interfaces more friendly. In: *Proceedings of the AI, ALife and Entertainment Workshop, IJCAI (1995)*
3. Raskin, V.: Computer implementation of the general theory of verbal humor. In: Hulstijn, J., Nijholt, A. (eds.) *The International Workshop on Computational Humour*, pp. 9–19. UT Service Centrum, Enschede (1996)
4. Ritchie, G.: Current directions in computational humour. *Artif. Intell. Rev.* **16**(2), 119–135 (2001)
5. Mulder, M.P., Nijholt, A.: *Humour Research: State of the Art*. University of Twente, Enschede, Netherlands (2002)
6. Nijholt, A.: Embodied agents: A new impetus to humor research. In: Stock, O., Strapparava, C., Nijholt, A. (eds.) *The April Fool's Day Workshop on Computational Humor, ITC-irst, Trento, Italy*, pp. 101–111 (2002)
7. Raskin, V.: Quo vadis computational humor? In: Stock, O., Strapparava, C., Nijholt, A. (eds.) *The April Fool's Day Workshop on Computational Humor, ITC-irst, Trento, Italy*, pp. 31–46 (2002)
8. Stock, O., Strapparava, C.: HAHAcronym: Humorous agents for humorous acronyms. In: Stock, O., Strapparava, C., Nijholt, A. (eds.) *The April Fools' Day Workshop on Computational Humor*, pp. 125–135 (2002)
9. Binsted, K., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manarung, R., Pain, H., Waller, A., O'Mara, D.: Computational humor. *IEEE Intell. Syst.* **21**(2), 59–69 (2006)

10. Hempelmann, C.F.: Computational humor: Going beyond the pun. In: Raskin, V. (ed.) *The Primer of Humor Research*, pp. 333–360. Berlin-New York, Mouton de Gruyter (2008)
11. Morkes, J., Kernal, H.K., Nass, C.: Humor in task-oriented computer-mediated communication and human-computer interaction. In: *Proceedings of CHI*. ACM, New York (1998)
12. Lyttle, J.: The effectiveness of humor in persuasion: The case of business ethics training. *J. Gen. Psychol.* **128**(2), 206–216 (2001)
13. McKay, J.: Generation of idiom-based witticisms to aid second language learning. In: Stock, O., Strapparava, C., Nijholt, A. (eds.) *The April Fools' Day Workshop on Computational Humor*, Trento, Italy, ITC-irst, pp. 77–87 (2002)
14. O'Mara, D., Waller, A.: What do you get when you cross a communication aid with a riddle? *Psychologist* **16**(2), 78–80 (2003)
15. Waller, A., O'Mara, D., Manurung, R., Pain, H., Ritchie, G.: Facilitating user feedback in the design of a novel joke generation system for people with severe communication impairment. In: *International Conference on Human-Computer Interaction*, Las Vegas, NE (2005)
16. Stock, O.: Password swordfish: verbal humor in the interface. In: Hulstijn, J., Nijholt, A. (eds.) *Proceedings of the International Workshop on Computational Humour (TWLT 12)*. University of Twente, Enschede, Netherlands (1996)
17. Stock, O., Strapparava, C.: Automatic production of humorous expressions for catching the attention and remembering. *IEEE Intell. Syst.* 64–67 (2006)
18. Taylor, J.M., Mazlack, L.J.: Toward computational recognition of humorous intent. In: *Cognitive Science Conference 2005* (2005)
19. Taylor, J.M.: *Towards Informal Computer Human Communication: Detecting Humor in Restricted Domain*, Ph.d. thesis, Department of Electrical and Computer Engineering, University of Cincinnati (2008)
20. Taylor, J.M.: Computational treatments of humor. In: Attardo, S. (ed.) *The Routledge Handbook of Language and Humor*. Routledge Vadehra, New York (2017)
21. Mihalcea, R., Strapparava, C.: Computational laughing: automatic recognition of humorous one-liners. In: *Proceedings of CogSci Conference*, Stresa, Italy (2005)
22. Mihalcea, R., Strapparava, C.: Learning to laugh (automatically): computational models for humor recognition. *Comput. Intell.* **22**(2), 126–142 (2006)
23. Mihalcea, R., Pulman, S.: Characterizing humour: an exploration of features in humorous texts. In: Gelbukh, A. (ed.) *CICLing 2007*. LNCS, vol. 4394, pp. 337–347. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-70939-8_30
24. Sjöbergh, J., Araki, K.: Recognizing humor without recognizing meaning. In: Masulli, F., Mitra, S., Pasi, G. (eds.) *WILF 2007*. LNCS (LNAI), vol. 4578, pp. 469–476. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73400-0_59
25. Kiddon, C., Brun, Y.: That's what she said: Double entendre identification. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, vol. 2, pp. 89–94 (2011)
26. Radev, D., Stent, A., Tetreault, J., Pappu, A., Iakopoulou, A., Chanfreau, A., de Juan, P., Vallmitjana, J., Jaimes, A., Iha, R.: Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest (2015)
27. Shahaf, D., Horvitz, E., Mankoff, R.: Inside jokes: identifying humorous cartoon captions. In: *Proceedings of SIGKDD* (2015)
28. Chandrasekaran, A., Kalyan, A., Antol, S., Bansal, M., Batra, D., Zitnick, C.L., Parikh, D.: We are humor beings: Understanding and predicting visual humor (2015)
29. Miller, T., Hempelmann, C.F., Gurevych, I.: SemEval 2017 Task 7: detection and interpretation of English puns. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (2017)

30. Bloom, L., Lahey, M.: *Language Development and Language Disorders*. Columbia University Academic Commons (1978)
31. Attardo, S., Raskin, V.: Script theory revis(it)ed: Joke similarity and joke representation model. *Humor Int. J. Humor Res.* **4**(3), 293–347 (1991)
32. Ruch, W., Attardo, S., Raskin, V.: Toward an empirical verification of the general theory of verbal humor. *Humor Int. J. Humor Res.* **6**(2), 123–136 (1993)
33. Harris, Z.: Distributional structure. *Word* **10**(2/3), 146–162 (1954)
34. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)
35. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781 (2013)
36. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014)
37. Lin, D., Pantel, P.: Discovery of inference rules for question-answering. *Nat. Lang. Eng.* **7**(4), 343–360 (2001)
38. Neelakantan, A., Shankar, J., Passos, A., McCallum, A.: Efficient non-parametric estimation of multiple embeddings per word in vector space. In: *Proceedings of EMNLP 2014, Doha, Qatar*, pp. 1059–1069 (2014)
39. Iacobacci, I., Pilehvar, M.T., Navigli, R.: SENSEMBED: learning sense embeddings for word and relational similarity. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics* (2015)
40. Pilehvar, M., Collier, N.: De-conflated semantic representations. In: *EMNLP 2016* (2016)
41. Yu, M., Dredze, M.: Improving lexical embeddings with semantic knowledge. In: *Association for Computational Linguistics (ACL)*, pp. 545–550 (2014)
42. Hashimoto, K., Tsuruoka, Y.: Adaptive joint learning of compositional and non-compositional phrase embeddings. In: *Proceedings of the 54th Annual Meeting of the ACL* (2016)
43. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning* (2014)
44. Mesnil, G., Ranzato, M., Mikolov, T., Bengio, Y.: Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews (2014). arXiv preprint [arXiv:1412.5335](https://arxiv.org/abs/1412.5335)
45. Carrier, P.L., Cho, K.: LSTM networks for sentiment analysis. *Deep Learning Tutorials* (2014)
46. Kim, Y.: Convolutional neural networks for sentence classification (2014). arXiv preprint [arXiv:1408.5882](https://arxiv.org/abs/1408.5882)
47. Friedland, L., Allan, J.: Joke retrieval: recognizing the same joke told differently. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 883–892. ACM, October 2008
48. Thorndike, R.L.: Who belongs in the family. *Psychometrika* **18**(4), 267–276 (1953)
49. Lau, J.H., Baldwin, T.: An empirical evaluation of doc2vec with practical insights into document embedding generation (2016). arXiv preprint [arXiv:1607.05368](https://arxiv.org/abs/1607.05368)
50. Jing, X.: *Concept-Level Sentiment Analysis of Online Hotel Reviews*, Unpublished Master’s Thesis, Department of Computer and Information Technology, Purdue University (2017)
51. Raskin, V.: *Semantic Mechanisms of Humor*. Reidel, Dordrecht (1985)
52. Attardo, S., Hempelmann, C., Di Maio, S.: Script oppositions and logical mechanisms: Modeling incongruities and their resolutions. *Humor Int. J. Humor Res.* **15**(1), 3–46 (2002)
53. Esar, E.: *The Humor of Humor*. Horizon, New York (1952)