Riddle Generation using Word Associations

Paloma Galván[1], Virginia Francisco[1], Raquel Hervás[1], Gonzalo Méndez[2]
[1]Departamento de Ingeniería del Software e Inteligencia Artificial
[2]Instituto de Tecnología del Conocimiento
Universidad Complutense de Madrid, Spain
palomagalvan@ucm.es, {virginia,raquelhb,gmendez}@fdi.ucm.es

Abstract

In knowledge bases where concepts have associated properties, there is a large amount of comparative information that is implicitly encoded in the values of the properties these concepts share. Although there have been previous approaches to generating riddles, none of them seem to take advantage of structured information stored in knowledge bases such as Thesaurus Rex, which organizes concepts according to the fine grained ad-hoc categories they are placed into by speakers in everyday language, along with associated properties or modifiers. Taking advantage of these shared properties, we have developed a riddle generator that creates riddles about concepts represented as common nouns. The base of these riddles are comparisons between the target concept and other entities that share some of its properties. In this paper, we describe the process we have followed to generate the riddles starting from the target concept and we show the results of the first evaluation we have carried out to test the quality of the resulting riddles.

Keywords: Computational Creativity, Riddle Generation, Word Associations

1. Introduction

In knowledge bases where concepts have associated properties, there is a large amount of comparative information that is implicitly encoded in the values of the properties that these concepts share. This kind of information can be useful in tasks where it is required to automatically establish relations between concepts, such as the generation of comparisons between entities based on shared properties (e.g. *this shirt is as white as snow*). These tasks are included in computational creativity, trying to simulate the natural behaviour of human beings to be creative using computer programs.

In this paper we present a riddle generator that creates riddles about concepts represented as common nouns. The base of these riddles are comparisons between the target concept (i.e. *a shirt*) and other entities that share some of its properties (i.e. *snow*). The resulting riddles are composed as a sequence of comparisons following this template: "What is... as *attribute* as *concept*?", where *attribute* is a property of the target concept which is the answer to the riddle, and *concept* is a different entity that shares the value of the *attribute* with the target concept. For example, "*What is... as hard as concrete and as transparent as hair?*" is a riddle generated for the concept *diamond* by the riddle generator.

In order to gather information about the features that characterize the target concept of the riddle, and to obtain similar concepts according to those features, a word association resource called Thesaurus Rex (Veale and Li, 2013) has been used.

This paper is organized as follows. Section 2 provides a overview of the state of the art of riddle and puzzle generation. Section 3 explains our approach to generate riddles using word associations. In Section 4 we present the evaluation and results obtained from a series of questionnaires where people tried to solve two sets of riddles generated by our system. Finally, Section 5 contains some conclusions and future work.

2. Related Work

Although the generation of riddles may seem a difficult task from a computational point of view, there have been previous attempts to the automatic generation of riddles.

De Palma and Weiner (1992) propose a model of a knowledge representation that contains the data to generate or solve riddles. They develop an algorithm that generates a guess based on homophonous concepts.

JAPE (Binsted and Ritchie, 1997; Ritchie, 2003) is a computer program which generates simple punning riddles using templates with slots where words or phrases are inserted. To determine which words must to be incorporated to the final riddle, the system makes use of predefined schemas (manually built from previously known jokes), which establish relationships between words which must hold to build a joke. The program was tested by 120 children that rated generated riddles, human-generated texts, and non-joke texts for 'jokiness' and 'funniness'. The evaluation confirmed that riddles generated were jokes, and that there is no significant difference in 'funniness' or 'jokiness' between punning riddles generated by their system and published human-generated jokes.

Some of the authors of JAPE (Cunningham et al., 2000) have furtherly developed STANDUP (Waller et al., 2009), a large-scale pun generator to allow children with communication disabilities to improve their linguistic skills. The pun generation followed the same steps used in JAPE, but several improvements had to be introduced in order to adapt the generated puns to the target audience, i.e. children with communication disabilities: speech output, picture support, restricted topics or use of familiar words. The system was evaluated with real users over a short period, and although no positive effects could be observed on the long term, the authors report a change in the attitude of the children towards communication.

Colton (2002a) extended the HR automated theory formation system (Colton, 2002b) to enable it to automatically generate puzzles given background information about a set

of objects of interest. They generate three types of puzzles: odd one out, next in sequence and analogy puzzles. They found that the main problem with puzzle generation was ensuring the uniqueness of the concept supposed to explain the puzzle solution.

Pintér et al. (2012) propose a knowledge-lean method to generate three types of word puzzles (odd one out, choose the related word, and separate the topics) from unstructured and unannotated document collections. The difficulty of the puzzles can be adjusted. The algorithm is based on topic models, semantic similarity, and network capacity.

Guerrero et al. (2015) present a Twitter bot that generates riddles about celebrities. The model selects a celebrity, retrieves relevant traits to describe him, generates analogies between his attributes and converts such descriptions into utterances, and, finally, tweets the generated riddle and interact with users by evaluating their answers. To evaluate the riddle generation they asked 86 people to evaluate five riddles. They first asked the participants to guess the answer to the riddle. Then, they presented the correct answer and asked if they knew the person in question. The participants indicated whether they considered the quality of the riddle satisfactory and, if negative, gave the reason why it was not good. The percentage of known celebrities once the answer was presented (54.19%) indicates that the process for the selection of celebrities should be improved. The low number of correct answers (15.58%) suggests that the complexity of the generated riddles was high.

3. Riddle Generation Using Word Associations

The proposed riddle generator receives a common noun as an input, which is the target concept for the riddle. Using Thesaurus Rex, a database of word associations extracted from the web, the system unfolds a series of comparisons between the target concept and other concepts with similar properties in order to create the final riddle.

3.1. Thesaurus Rex

Thesaurus Rex (Veale and Li, 2013) organizes concepts according to the fine grained ad-hoc categories they are placed into by speakers in everyday language (*food*, *drink*, *beverage*...). These categories have an associated weight that represents their relative importance for the given concept. Thesaurus Rex can show different categories for each concept and allows in turn to consult the concepts in each category. If we take as an example the concept *coffee*, some of its categories with more weight are *beverage* or *drink* and some with less weight are *seed* or *poison*. Table 1 shows some categories for coffee and their corresponding weights. Concepts in Thesaurus Rex have also associated properties or modifiers which are also accompanied by a weight indicating how strong its relation to the concept is. For example, for *coffee* some of the modifiers with more weight are *hot*, *acidic* or *stimulating*, and modifiers with less weight are *granulated* or *digestive*.

Category	Weight	Attribute	Weight
drink	4983	hot	3900
smell	185	granulated	10
beverage	7056	acidic	2909
seed	3	dark	1144
intoxicant	14	stimulating	1267
liquid	2541	noncarbonated	24
food	3322	colored	696
poison	5	digestive	7
...	...	...	...

Table 1: Examples of categories and attributes (including weights) for the concept *coffee* on Thesaurus Rex

3.2. Riddle Generation

Table 2 shows a few examples of target concepts and how Thesaurus Rex is used to generate riddles. Taking the first concept, *sun*, as an example, the detailed process to generate a riddle is the following:

1. **Target concept categories.** To obtain the filtered categories to which the target concept belongs, we first extract a list of all the general categories of the concept using a Thesaurus Rex query. From this list, only the $N\%$ of categories with the highest weights are considered as candidates. The value of N is configurable. If a high N value is set, we will have in the list categories with lower weights, which are less relevant to the target concept. In the same way, we can set N to a low value, facing the risk of shortening the list to a single element. In the *sun* example, the categories with higher weights in Thesaurus Rex are *body* and *object*.

2. **Modifier extraction.** In addition to the categories, we also need a list of modifiers associated to the target concept, which is returned by a new query to Thesaurus Rex. From this list, the $N\%$ of attributes with the highest weights are considered as candidates. For example, if our target concept is the noun *sun*, some of the most important properties extracted are: *stellar*, *hot*, *natural* and *yellow*.

3. **Modifier selection.** One of the modifiers previously obtained is randomly selected. This random selection makes the system less repetitive, as the riddles obtained for the same target concept are not always the same as if only the modifier with the highest weight were selected. For the current example, we suppose that the system has chosen the modifier *hot*.

4. **New categories selection.** Using the modifier chosen in the previous step, a new query to Thesaurus Rex is performed in order to obtain new categories that also present this modifier as a highlighted property. In order to obtain comparisons between different kinds of concepts, the new categories that match the categories obtained in step 1 are discarded. In this way, we are avoiding the comparison of the target concept with other concepts in the same category in order to obtain more creative results. In the *sun* example, the new categories selected could be *food* and *beverage*, which are categories that present the *hot* property in Thesaurus Rex.

Target concept	*Categories*	*Modifiers*	*New categories for the selected modifier*	*New query*	*Obtained concepts new query*	*Comparison*
sun	body, object, star...	stellar, **hot**, natural, yellow...	**food**, beverage ...	hot food	chili, **soup**, garlic...	as hot as soup?
whale	animal, mammal, predator...	large, **migratory**, marine, aquatic...	**bird**, fish...	migratory bird	**goose**, duck, heron...	as migratory as goose?
diamond	stone, material, gem...	precious, valuable, **hard**, crystalline...	material, **surface**...	hard surface	wood, wall, **concrete**...	as hard as concrete?

Table 2: Examples of comparisons obtained to be part of the riddles. Words in bold represent the choices made for each example.

5. **New category selection.** One of the new categories obtained in the previous step is randomly selected. For the current example, *food* is supposed to be the category selected.

6. **New query composition.** A new query for Thesaurus Rex is then composed by using the new category obtained in the previous step and the modifier selected in step 3. In the current example, we will assume this new query is *hot food*.

7. **Final concept selection.** With the query composed in the previous step, we obtain a list of concepts that belong to the category selected in step 5 (*food*) and at the same time present the property selected in step 3 (*hot*). This list is usually quite extensive, so the system randomly chooses among the results that have an associated weight among the $N\%$ of concepts with the highest weights. In our example, a possible final concept for *hot food* is *soup*.

8. **Comparison template.** With the final data obtained during this process, attribute and new concept, the template "as *attribute* as *concept*?" is filled. The result of this round is "as *hot* as *soup*?".

9. **Riddle composition.** Steps 3-8 are repeated as many times as desired, determined in the configuration of the system. In each round, a new comparison is generated and added to the final riddle. In our example, a possible riddle with three comparisons is the following:

   What is ...
   ... as hot as soup?
   ... as stellar as a galactic_nucleus?
   ... as yellow as a mango?

4. Evaluation

We have carried out an evaluation to test whether word associations obtained by our system provided useful information for riddle generation, and to assess the quality of the resulting riddles. In order to do that, human evaluators were asked to guess the initial concepts which were used to create the riddles. Then, we studied the rate of success obtained by the evaluators, while at the same time analyzing how many comparisons were required to obtain the correct answers in different riddles. Some issues related to ambiguity and contradiction appeared when creating the riddles, so we decided to create two different sets of riddles to perform the evaluation.

4.1. Design

Ten riddles were presented to human evaluators to see if they were able to find the initial target concepts. Riddles were presented in four phases, in order to know how many comparisons were needed to solve the riddle. In the first phase a single comparison was presented, in the second phase two comparisons were presented, three comparisons in the third phase and, finally, four comparisons in the fourth phase. The evaluation was carried out using Google Forms and some personal information was collected for statistical purposes (age, gender and riddle ability).
As explained in the previous section, the comparisons used in riddles are randomly chosen. When generating riddles for the evaluation, we realized that some of the comparisons do not add new information to previous ones, or the information added was contradictory or not valid due to the polysemy of some concepts. This is the case of *coke*, which is a flavoured carbonated drink and the street name for cocaine. Our system had generated "*is as carbonated as ...*" and "*is as hard as ...*". The second comparison is obtained due to the word association "*coke - hard drug*". Examples of contradictory comparisons are mostly related to attributes with imprecise values, like size or age. For example, Thesaurus Rex categorizes the concept *dog* as both small and large depending on the context. However, if our system chooses both attributes, we would have contradictory comparisons in the riddle.
In order to carry out a more detailed evaluation, we decided to create two different riddle sets. Using the same ten concepts, but with some differences in the provided comparisons, we created an original and a curated version of the riddles. For the first set, the resulting comparisons were randomly selected. For the other set, the four most significant comparisons were manually selected among seven generated using the described process in order to avoid not valid comparisons due to polysemy or semantic contradictions. The riddles used in the evaluation can be seen in an Appendix at the end of the paper.

4.2. Results

Both the evaluation with the random riddles and the one with the curated versions were performed in parallel by 12 different evaluators each, making a total of 24 participants in the experiment. The order of appearance for each riddle

Phases	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Phase 1	17%	0%	17%	8%	0%	0%	0%	8%	0%	0%
Phase 2	8%	0%	67%	0%	33%	25%	8%	17%	0%	0%
Phase 3	8%	0%	58%	33%	50%	92%	8%	67%	0%	0%
Phase 4	8%	0%	42%	17%	50%	67%	0%	75%	17%	0%

Table 3: Percentage of success for each riddle in random set

Phases	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Phase 1	0%	25%	0%	17%	0%	8%	17%	0%	33%	0%
Phase 2	33%	25%	75%	33%	25%	58%	67%	75%	33%	0%
Phase 3	42%	25%	75%	42%	58%	83%	75%	75%	67%	8%
Phase 4	42%	33%	92%	50%	58%	75%	75%	75%	67%	8%

Table 4: Percentage of success for each riddle in curated set

Phases	Random set	Curated set
Phase 1	5%	10%
Phase 2	16%	43%
Phase 3	32%	55%
Phase 4	28%	58%

Table 5: Percentage of guessed riddles

was fixed, so all the evaluators participating in each part of the experiment were presented exactly the same riddles. To view the results for each riddle, Tables 3 and 4 show disaggregated percentages of success for each riddle in each phase. The percentages of correctly guessed riddles in the evaluation for each phase are presented in Table 5.

4.3. Discussion

The approach we have used to generate the riddles assures that all of them have a solution. In the random generation mode, all the participants of the evaluation guessed at least two riddles (5%) and thirteen at most (32.5%), with an average of eight riddles guessed (20%) per person. In the curated generation mode, the minimum amount of guessed riddles was 6 (15%) and the maximum was 24 (60%), with an average of 16.5 riddles guessed (41.3%) per person. The aim is not to get all the riddles solved, which would indicate that they are too easy.

As shown in Table 5 the results of the curated version of the riddles are significantly better than the ones of the random version. So, it is evident that a special selection of comparisons is needed in some cases.

Regarding the number of comparisons needed to guess the correct answer, a curious fact can be seen in Table 5. With just a single comparison, there is almost no chance of guessing the target concept. In most cases, people answer at random because there are lots of concepts that share the presented attribute. When providing two comparisons, users are able to multiply by four the number of correct guesses. When they are provided three comparisons, in the case of randomly chosen comparisons, they reach their maximum rate of success. In the case of manually selected comparisons, they guess 55% of riddles, which is almost the maximum success, because the difference with the last phase, where four comparisons are provided, is almost negligible.

At a more detailed level, Tables 3 and 4 shows that R10 has, in the best case scenario, a success rate of 8%. This is due to the fact that the attributes selected are not specific enough and there is a large amount of common properties with other concepts. In this example, the concept was *aircraft* and the attributes selected were: *mechanical*, *fast*, *mobile* and *complicated*. However, the third concept, *sun*, in the fourth phase of the curated set, had a success rate of 92%, as shown in Table 4. The reason for this is that the attributes selected for this concept were much more specific. For instance, an attribute that not many concepts share is *stellar*, which combined with *yellow*, *hot* and *central* limits the possible answers for this riddle.

From the point of view of the success rate in each set, in the last phase of the random set (Table 3), sometimes the percentage of correct answers decreases slightly. The reason for this, as explained by the participants in the evaluation, is that sometimes the last hints were contradictory, and users were confused and ended up changing their answer in the last attempt. However, in the curated set (Table 4), only R6 presents a decrease of the success rate in the last phase. This means that the last comparison in this case was confusing for some of the evaluators.

5. Conclusions and Future Work

An automated mechanism for riddle generation using Thesaurus Rex, a resource based on word associations, has been presented. Following the described process to generate the riddles, the subsequent evaluation points out that the word associations obtained by our system are useful for generating these riddles. However, the evaluation also shows that a manual selection of comparisons is useful because confusing comparisons may be generated when the target of the riddles is a polysemic concept or presents some contradictory attributes. Therefore, it is necessary to develop some mechanisms to select only the modifiers related to the sought meaning of the target concept, and consider in a special way attributes with imprecise values.

The results of the evaluation also suggest that the order in which the comparisons are provided is relevant in order to solve the riddle using less comparisons, so it may be useful to analyze the discriminating power of each attribute, so that the complexity of the riddles can be controlled. If this information is available, the system could select first (or last) the most discriminating attributes of the concept automatically. The underlying idea is that the higher the discriminating power of the attribute, the easier the riddle, as more concepts are excluded from the possible answers. Depending on the desired difficulty of the riddle, we can play with the order of the attributes according to their discriminating power.

As seen during evaluation, to make better guesses three or more comparisons are generally needed. In the future, we will evaluate with five or more comparisons to determine an optimal number in order to have riddles that are not impossible to guess, while at the same time are not too obvious. In addition, we would like to include riddles created by humans in future evaluations, so we can assess whether our riddles are easier or more difficult in comparison, and if they are considered natural in comparison with human-made ones.

In the future, we will explore the possibility of developing more creative riddles, for example with rhymes and a more elaborated selection of attributes and concepts.

One of our main concerns when developing the described riddle generator is the practical application for the resulting system. Currently, we are using a similar approach to generate rhetorical figures, such as analogies, similes and metaphors. In this case, instead of hiding the target concept and making the user guess it, the similarities between the two concepts are explored in order to create tropes that are as evocative and meaningful as possible. Hence, one of our current goals with riddle generation is to study the relationship between concepts through shared properties to gain a deeper insight that helps us generate better linguistic resources.

Another research line where we are starting to explore the applicability of the described techniques is accessibility, and more specifically, text simplification and text generation for users with cognitive disabilities. Other authors have already reported on the use of riddles to allow children with communication difficulties to develop their linguistic skills (Manurung et al., 2008). Following this idea, we aim at exploring the way in which riddles can be incorporated in the life of people with communication disabilities, supported, in addition, with the use of pictographs.

6. Acknowledgements

This work is funded by ConCreTe. The project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

7. References

Binsted, K. and Ritchie, G. (1997). Computational rules for generating punning riddles. *Humor - International Journal of Humor Research*, 10(1):25–76.

Colton, S. (2002a). Automated puzzle generation. In *Proceedings of the AISB'02 (Symposium on AI and Creativity in the Arts and Science)*.

Colton, S. (2002b). *Automated Theory Formation in Pure Mathematics*. Springer series on distinguished dissertations.

Cunningham, H., Maynard, D., and Tablan, V. (2000). JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, November.

De Palma, P. and Weiner, E. (1992). Riddles: accessibility and knowledge representation. In *Proceedings of Coling'92*, pages 1121–1125.

Guerrero, I., Verhoeven, B., Barbieri, F., Martins, P., and Pérez y Pérez, R. (2015). Theriddlerbot: A next step on the ladder towards creative twitter bots. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 315–322.

Manurung, R., Ritchie, G., Pain, H., Waller, A., O'Mara, D., and Black, R. (2008). The construction of a pun generator for language skills development. *Applied Artificial Intelligence*, 22(9):841–869.

Pintér, B., Voros, G., Szabo, Z., and Lorincz, A. (2012). Automated word puzzle generation using topic models and semantic relatedness measures. In *Joint Conference on Mathematics and Computer Science (MACS), Siófok, Hungary*.

Ritchie, G. (2003). The jape riddle generator: technical specification. EDI-INF-RR 0158, School of Informatics, University of Edinburgh.

Veale, T. and Li, G. (2013). Creating similarity: Lateral thinking for vertical similarity judgments. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 660–670.

Waller, A., Black, R., O'Mara, D. A., Pain, H., Ritchie, G., and Manurung, R. (2009). Evaluating the standup pun generating software with children with cerebral palsy. *ACM Transactions on Accessible Computing*, 1(3):16:1–16:27.

Appendix: Evaluation Riddles

The complete list of riddles used in the evaluation is presented in Table 6. The first column, *Riddle*, shows the position of the riddle in the evaluation, and refers to the riddles shown in Tables 3 and 4. The second column, *Word*, shows the target concept that had to be guessed by the users. In the third column, *Random riddle*, we show the sequence of clues, in the form of riddles, that were given to the users to guess the target word in the random version of the evaluation. Finally, in the last column, *Curated riddle*, we show the sequence of clues that were provided to the users in the curated version of the evaluation.

Riddle	*Word*	*Random riddle*	*Curated riddle*
R1	*Ant*	*... as tiny as an isopod?* *... as terrestrial as a bird?* *... as small as a rabbit?* *... as social as a wedding?*	*... as invertebrate as a crab?* *... as social as a wedding?* *... as tiny as an isopod?* *... as annoying as a slug?*
R2	*Devil*	*... as useless as sand?* *... as common as chromium?* *... as evil as envy?* *... as invisible as joy?*	*... as evil as envy?* *... as supernatural as a deity?* *... as powerful as extreme anger?* *... as bad as a war?*
R3	*Sun*	*... as stellar as a galactic nucleus?* *... as hot as a soup?* *... as natural as wood?* *... as gravitational as a planet?*	*... as hot as a soup?* *... as stellar as a galactic nucleus?* *... as yellow as a mango?* *... as central as a living-room?*
R4	*Car*	*... as large as a horse?* *... as physical as a hardness?* *... as private as a hotel?* *... as technical as a medicine?*	*... as mechanical as a gear?* *... as everyday as clothing?* *... as heavy as lead?* *... as large as a horse?*
R5	*Whale*	*... as marine as a barnacle?* *... as large as a furniture?* *... as migratory as a goose?* *... as aquatic as a fish?*	*... as large as a furniture?* *... as migratory as a goose?* *... as marine as a barnacle?* *... as aquatic as a fish?*
R6	*Diamond*	*... as transparent as a hair?* *... as pure as gold?* *... as costly as a car?* *... as simple as a screwdriver?*	*... as hard as concrete?* *... as transparent as a hair?* *... as precious as silver?* *... as geometric as a circle?*
R7	*Milk*	*... as liquid as methanol?* *... as raw as cotton?* *... as natural as wood?* *... as everyday as clothing?*	*... as white as a pollock?* *... as liquid as methanol?* *... as natural as wood?* *... as raw as cotton?*
R8	*Shark*	*... as large as a horse?* *... as dangerous as scissors?* *... as marine as a barnacle?* *... as aquatic as a fish?*	*... as dangerous as scissors?* *... as marine as a barnacle?* *... as predatory as a cheetah?* *... as big as a tiger?*
R9	*Coke*	*... as commercial as a supermarket?* *... as hard as concrete?* *... as carbonated as a cooler?* *... as cool as damp soil?*	*... as carbonated as a cooler?* *... as commercial as a supermarket?* *... as dark as a fig?* *... as cool as damp soil?*
R10	*Aircraft*	*... as physical as swimming?* *... as mobile as a truck?* *... as modern as a restaurant?* *... as fixed as a tree?*	*... as mechanical as a pump?* *... as fast as a hamburger?* *... as mobile as a truck?* *... as complicated as a ship?*

Table 6: Riddles used in the evaluation