A Master-Apprentice Approach to Automatic Creation of Culturally Satirical Movie Titles

Khalid Alnajjar
Department of Computer Science
and Helsinki Institute for IT
Faculty of Science
University of Helsinki, Finland
alnajjar@cs.helsinki.fi

Mika Hämäläinen
Department of Digital Humanities
Faculty of Arts
University of Helsinki, Finland
mika.hamalainen@helsinki.fi

Abstract

Satire has played a role in indirectly expressing critique towards an authority or a person from time immemorial. We present an autonomously creative master-apprentice approach consisting of a genetic algorithm and an NMT model to produce humorous and culturally apt satire out of movie titles automatically. Furthermore, we evaluate the approach in terms of its creativity and its output. We provide a solid definition for creativity to maximize the objectiveness of the evaluation.

1 Introduction

Movie theaters used to be prohibited in Saudi Arabia by law, however in mid-December 2017, Saudi Arabia officially declared they would abolish the law by 2018[1]. This gave birth to a movement on Twitter by the hashtag *#SaudiMovieTitles* where people would post alternative movie titles meant to be satirical towards the Saudi Arabian legislative system. An example of such a title is *I Know What You Ate Last Ramadan* for *I Know What You Did Last Summer*.

In this paper, we present a novel master-apprentice approach to computational creativity for the generation of such humoristic movie titles automatically. Our system consists of two creative agents: a computationally creative master that is implemented with a genetic algorithm approach and a neural network based apprentice that will learn from the master along with its peers, namely real people writing satirical movie titles on Twitter. We are introducing the apprentice to achieve creative autonomy, because of its nature of adjusting its standards by learning, a capability which is absent in the master.

Furthermore, we base our approach in terms of the design of its implementation and its evaluation on a solid definition of creativity that we build based on literature. This will make a reasoned evaluation of our system possible and set standards for future research in this topic.

For the purposes of our research, we understand cultural satire as a way of presenting critique towards a society in a humoristic fashion. The humoristic expression has to relate to Saudi Arabia in its expression - mostly by lexical choice. The relatedness has to be apparent when presented with the hashtag *#SaudiMovieTitles*.

Computational linguistic creativity of this nature poses a number of challenges, not only because it is a difficult NLG problem, but also because the generated output has to be humorous. Thus the goal of the system is very different from a more traditional NLG task where the system is to convey factual information such as timetables of a train or the current weather conditions in the form of natural language.

The message our system is to produce serves rather to convey emotion and provoke joy in the reader. Furthermore, the humor is to be delivered in a culture-specific way which combines cultural artifacts of the Western World, the Hollywood movie titles, with a Saudi Arabian cultural twist. The system will derive its satire from this juxtaposition of the two cultures following utterly conflicting norms.

2 Related Work

Automatic creation of humor has received some attention in the past. Valitutti et al. (2009) present a tool for interactive creation of puns. The sys-

[1]The press release of The Saudi Ministry of Culture and Information (MOCI) https://www.media.gov.sa/news/1101

Proceedings of The 11th International Natural Language Generation Conference, pages 274–283,
Tilburg, The Netherlands, November 5-8, 2018.

tem suggests funny replacement words for familiar expressions such as proverbs. The replacements are found by applying a phonetic similarity metric together with a Latent semantic analysis (LSA) based semantic similarity metric, which not only gives semantically related words but also semantically opposed ones.

A fully automated pun generator, presented in Valitutti et al. (2013), takes an English sentence as its input and changes one word in it based on three criteria: sound or spelling similarity, the replacement word has to be a taboo and that the word has to go well together with its immediate predecessor in the sentence. The system operates with a predefined set of taboo words and an n-gram model to assess how well the new words fit in the sentence.

A more recent take on the pun generation is that of Shah et al. (2016), which presents a template based approach on humor generation. The templates are filled by using WordNet relations for the input compound word. The system looks up a list of related words for the two words of the compound and forms a pun out of them.

Research has also been conducted in the vein of humoristic sarcasm generation, which is then served to the user as satire (Veale, 2018). The system generates sarcastic tweets that satirize a person with a knowledge-based approach and regenerative approach which regenerates sarcastic tweets giving the sarcasm a new contextual meaning. The system is online on Twitter and it actively engages into replying Tweets it has been mentioned in.

The related work, not unlike a great many publications in the field of computational creativity, overlooks an important aspect, which is a clear and motivated definition for creativity. This issue has been brought up in the past research as well (Jordanous, 2012), but yet a great many publications devote little to no comment on what creativity, the very thing that is computationally modeled, really means in the context of a creative system. This makes comparison to and building from previous research a difficult task. To provide an alleviation to this problem, we will provide a definition for creativity based on existing literature that we will follow in this paper.

3 Definition of Creativity

In order to say anything about the creativity of a computational system, it is important to define what creativity means in the context of the system. We do this by following the SPECS (Standardized Procedure for Evaluating Creative Systems) approach (Jordanous, 2012). In short, the approach requires us first to define what is needed for creativity in general and then what is needed for creativity in our case of producing cultural humorous movie titles in particular.

For the general definition, instead of undertaking defining something as abstract as creativity by ourselves, we opt for a well established definition in the field, namely the creative tripod (Colton, 2008). According to that definition, there are three indispensable components to creativity: *skill*, *imagination* and *appreciation*. All of them can come from the three different parties of the creative experience, the system, the programmer or the perceiver of the creative artifact. However, the system has to be perceived to exhibit all of them on its own to be considered creative.

To put the creative tripod in the context of our highly narrowed down task, we start by defining skill as an ability to produce a new title out of an existing one. To master this skill it is imperative that the original movie title still remain recognizable and that the new title deliver a humoristic joke that satirizes Saudi Arabia.

Imaginativeness requires that our system should at least achieve P-creativity. That is, according to Boden (2004), coming up with a surprising and valuable idea that is novel to the one who came up with it, as opposed to H-creativity which is novel in a more global context, i.e. no one else had thought of it before. What this means is that our system should come up with similar humorous titles as real people have in our dataset to achieve P-creativity, and preferably come up with humoristic titles nobody has ever tweeted about before to be H-creative.

Lastly, we define appreciation in this context as a computational capability of assessing various factors that affect on the quality of the created title. One of them is the ability of identifying puns by recognizing sound similarities on a phonetic level such as in *Sheikhs and the City* for *Sex and the City*. But a far more important requirement for appreciation is an automated assessment of the humorousness of the output. Two key components have been identified for humorousness in jokes: *surprise* and *coherence* (Brownell et al., 1983).

When the brain makes sense of the stimuli it

receives, it relies on prior experiences to predict what it is perceiving. When these predictions turn out to be wrong, surprise arises. This is even more clearly crystallized in the case of movie titles. When a title everyone has heard a plethora of times gets changed, it clearly breaks learned expectations of the brain of how the title should be, as it is the case for example with *How to Train Your Imam* from *How to Train Your Dragon*.

In the previous example, surprise arises from the fact that the brain was expecting to hear *dragon*, but got a wildly different word *imam*. This surprise, however, does not yet lead to humor if the surprising word is not interpreted in its context. This requires coherence; the word has to make sense in the context of the sentence, and also in the wider context of the movie the original title refers to. Part of the humoristic value comes from the thought of young Hiccup undertaking an adventurous task on training his imam, as he did in the original movie to train his dragon.

In addition to the requirements identified with the creative tripod framework, we also introduce the requirement of creative autonomy for our system. Creative autonomy (Jennings, 2010) requires *autonomous evaluation*, *autonomous change* and *non-randomness*. In other words, the system should be able to evaluate its output independently (this relates to the appreciation defined earlier) and change its standards on its own without instructions on how to do so. Neither of these two requirements can be based solely on a random choice.

We have defined the core requirements for our system to be considered to exhibit creative behavior. These requirements will be revisited in the Evaluation section of this paper to assess the extent to which they are met by the implementation. This is also required by the SPECS approach we have decided to follow in this paper.

4 Datasets

This section is devoted to the description of the datasets we have in use for the implementation of the system.

4.1 Movie Data

For our case, we are only interested in the popular movies people might have heard of, as one of our requirements is the original title to be recognizable after it has been turned into a humoristic form. No matter how good the satire, the joke will not land, if the audience is unfamiliar with the original title.

To get indispensable information of the movies, we used the data dumps provided by the IMDB[2] to extract movie titles and their metadata. As the IMDB is a bit too extensive, sporting over a 5 million movies and TV shows in its database, we had to narrow it down to those that had got more than 100,000 votes from the IMDB. The number of votes does not tell anything about how well the movie was received by the audience, but it gives us a clue on its popularity. The more popular the movie, the more votes it will get. In total, we had a total of 1,661 movies.

As movies can be referred to without their episode name, we expand the list of movies by considering their titles without episode or sequel name. This is performed by stripping out words after the first dash or colon. For instance, the movie title *The Lord of the Rings: The Fellowship of the Ring* is converted into *The Lord of the Rings*.

4.2 Tweets

Another peculiar source of information are the tweets with the the hashtag *#SaudiMovieTitles*. We retrieved them by using the Twitter API, which resulted in a list of 2,445 tweets. However, the tweets are noisy because, more often than not, they contain other unrelated content (e.g. URLs, mentions and so on) in addition to the humoristic movie title. Moreover, the titles in the tweets might not be in the same format as the original movie title.

In order to serve us any use, the tweets needed to be processed so that what was left was just the humoristic title. This is done by converting the tweets into lowercase. Thereafter, the tokenization is applied, and any word starting with # or is filtered out to clean any hashtags and mentions, respectively. Any special characters and URLs are also deleted from the string. Finally, we remove all instances of the token "rt" from the string as it indicates a re-tweet.

Furthermore, we map the processed tweets to their original movie titles. The mapping phase is employed to backtrack any modifications performed on the original title for a humoristic effect. We achieve this by iterating over all tweets and calculating the edit distance on character and word level against all movie titles in the dataset. Movie

[2]https://datasets.imdbws.com/

titles with the least word edit distance, followed by character edit distance, are considered as the original title to tweets.

To reduce noise in the data (e.g. tweeted titles where all words have been changed), we keep the humoristic titles with at most 3 changed words, making still sure that not all the words have been changed in the tweeted title. Examples of mapped titles are *The Lord of the Thowbs* to *The Lord of the Rings* and *Gulf Fiction* to *Pulp Fiction*.

4.3 Saudi Arabia Related Vocabulary

We extract words, along with their part-of-speech tags that are related to Saudi Arabia from the tweets dataset. This is accomplished by parsing the tweeted titles and their mapped original counterparts with Spacy (Honnibal and Montani, 2017), followed by an analysis of their differences. We build the vocabulary by adding words that exist exclusively in the tweeted titles and not in the original ones. We also save the part-of-speech tags of the words.

Analyzing the vocabulary, it appears that all the added words, 1053 in total, are either nouns or adjectives. Examples of nouns are *thawab, imam* and *stone*, and adjectives are *saudi, shaytan* and *sunni*.

4.4 English Vocabulary of Arabic Origin

In order to produce titles relevant to the Saudi Arabian context, the system needs to have a list of English words related to the Arabian culture. The primary source of such words are the ones registered in the Oxford English Dictionary (OED, n.d.) as having Arabic as one of the languages in their etymology (514 in words in total). We have an access to the JSON files[3] of the OED, which made this task easier. The vocabulary also included the lexical categories of words.

5 Creating Movie Titles

In this section, we explain the method for creating movie titles. The method is divided into two submethods, (1) the master, which generates movie titles using genetic algorithms, and (2) the apprentice, which learns from the master and develops its own appreciation for generating movie titles.

5.1 The Master

The implementation of the master follows the one presented by Alnajjar et al. (2018) for slogan generation. The generator is a genetic algorithm that accepts an original movie title as an input and produces an entire population of satire movie titles based on the input.

The master operates on a semantic space of words related to Saudi Arabia to make the substitution of content words in the input with cultural words about Saudi Arabia possible. The semantic space is a combination of the vocabularies described in Section 4.3 and 4.4.

5.1.1 Evolutionary algorithm

Our algorithm starts by producing an initial population which undergoes an evolutionary process throughout a certain number of generations. The evolutionary algorithm in place is a standard (μ + λ)[4] which applies mutation and crossover on the current population to generate λ number of offspring. Subsequently, the algorithm evaluates the fitness of the individuals in the current population and their offspring, and selects μ number of the fittest individuals to survive to the next generation. Once the specified number of generations is reached, the evolutionary process terminates and returns the final population.

5.1.2 Initial Population

The method produces μ copies of the input movie title. For each copy, the method substitutes a randomly selected content word, i.e. not a stop word, with a random word from the semantic space. The substitution is done in such a fashion that the parts-of-speech of the original word and its substitute have to match. Furthermore, the word is inflected if necessary. The resulting titles form the initial population.

5.1.3 Mutation and Crossover

We define one type of mutation and crossover. The mutation procedure follows the same substitution approach performed during the construction of the initial population. The crossover employed in our function is a single-point crossover where words before and after a randomly selected point in two individuals are swapped.

5.1.4 Evaluation as Appreciation

The appreciation is implemented in the master by four internal evaluation dimensions, which are (1) prosody, (2) semantic relatedness to Saudi Arabia, (3) semantic similarity, and (4) number of al-

[3] We used the JSON files updated on the 14 of February in 2018

[4] We use the value 100 for both μ and λ

tered words. The first two dimensions are maximized, whereas the last two are minimized. Additionally, a dimension can be represented by the weighted sum of multiple sub-functions.

The prosody dimension assesses the sound similarity of the original word and its replacement. The dimension is composed of four prosody features, namely consonance, assonance, rhyme and alliteration. We utilize *espeak-ng tool*[5] to produce IPA transcriptions for words to better evaluate how they sound when pronounced. The tool is capable of producing IPA even for non-English words, such as *jahannam*.

To measure the semantic relatedness of words to Saudi Arabia, we build a semantic relatedness model following the model described in Xiao et al. (2016). Using the model, we measure the dimension as the maximum semantic relatedness of any content words in the generated title to the words "Saudi" and "saudi".

We employ a word2vec model trained on News dataset by Google (Mikolov et al., 2013) to measure the semantic similarity between two words. The dimension is represented as the mean of the semantic similarity of each introduced word to its original word. This dimension is minimized to increase surprise, with the idea that a lower semantic relatedness between the original word and its substitute would result in a bigger surprise.

The last dimension monitors the number of words altered in the input title. Minimizing this dimension motivates that less substitutions are made to the title, which makes it more recognizable.

These are the criteria based on which the fitness of individuals is evaluated at the end of each generation to let only the best ones survive to the next generation. Also, the master uses this exact same functionality when it is to show appreciation to titles outside of its own creations such as those created by the apprentice.

5.1.5 Selection and Filtering

To reduce having a dominating dimension and motivate generating titles with diverse and balanced scores on all four dimensions, we opt for a non-dominant sorting algorithm –*NSGA-II*– (Deb et al., 2002) as the selection algorithm.

During each iteration of the evolution, the current population and its offspring go through a filtering phase which filters out any duplicate titles.

[5] https://github.com/espeak-ng/espeak-ng

5.2 The Apprentice

The apprentice is a sequence-to-sequence neural network model, which will be trained by the parallel data of the original titles and their humoristic counter-parts produced by the master. Furthermore, the apprentice is trained with the parallel titles extracted from the tweets. We use a general purpose NMT library called OpenNMT (Klein et al., 2017) for this task.

Similar sequence-to-sequence based approaches have been used in the past for text paraphrasing task (Brad and Rebedea, 2017; Sleimi and Gardent, 2016), which shares its similarities with the task we are set to solve. This gives us a reason to believe that sequence-to-sequence approach is a viable way of implementing the apprentice.

The apprentice was trained for 50 epochs with the titles the master had produced for a random set of the most popular IMDB movie titles. This set consisted of 6568 humoristic titles. After this, the model was trained for 50 additional epochs with the data from the 1,483 tweeted titles. This made it possible for the apprentice to learn to a set of standards for humoristic titles from its master and adjust those standards with the peer data. In both cases, we use a 25% of the data in validation. The high number of epochs together with a larger validation set seemed[6] to make the model learn more given the limitation imposed by the scarce training data.

5.3 The Symbiotic Nature

Currently, we cannot argument for the apprentice developing appreciation that matches the requirements we have set for it in the definition of creativity, albeit it will learn some kind of an appreciation because it is capable of giving a confidence score to its creations. However, as the nature of such appreciation is not assessed in this paper, we opt for a symbiotic approach in the appreciation of the full system.

The apprentice will only be allowed to learn from its peers if the master shows high enough appreciation towards the creations of its peers. Furthermore, when producing its creative output, the apprentice consults the master for its opinion on which output should be picked and presented to

[6] We tried training with fewer epochs and fewer titles in the validation, but the model failed to learn anything meaningful.

the audience.

The apprentice might be dependent on the master in terms of appreciation, but just as much the master is dependent on its apprentice. The master possesses no capability of adjusting its standards, because it exhibits no learning from others. This is where the apprentice can learn its own standards from its peers, and is thus the only one responsible of the creative autonomy of the entire system.

6 Evaluation

In this part, we are presenting an evaluation of the system from two points of view. Firstly, we will assess critically how the definition of creativity with its requirements is met in the implementation of the system. Secondly, we validate this assessment by having ordinary people evaluate the output of the system by answering to questions on a 5-point Likert scale. These questions are derived from the definition of creativity in our context by following the SPECS approach.

6.1 Evaluation of the Creative Process

The skill of converting a movie title into a humorous one can be shown on the implementation level of the master. It takes an existing title, uses it as a skeleton and outputs a new title. The apprentice model also clearly demonstrates this lowest level requirement for skill. By nature, the NMT model produces an output based on its input. However, we can say little about the fulfillment of the further criteria for skill just by looking at the creative process. Recognizability of the original title, humorousness and satire towards Saudi Arabia are highly subjective notions and thus they will be assessed by evaluators in the next section.

The system can be shown to achieve P-creativity with a rather easy test by looking at the output in relation to the data the systems were given initially. For example the master produced a title *The Hobbit: An Unexpected Desert* from *The Hobbit: An Unexpected Journey* even though the master was not given knowledge of such a possibility for a humoristic title. The apprentice was able to produce the title *The Amazing Spider Mosque* for *The Amazing Spiderman* even thought its training data did not provide it with this mapping.

In order to analyze the imagination any further, we have to assess also the H-creativity of the produced titles. We could perform a Google search with some of the generated titles and claim H-creativity if it did not return any hits. However, we feel that this is not quite enough as the same kind of a joke might have come up elsewhere with a slightly different context. This is the reason why we have to verify the H-creativity with evaluators.

Appreciation can be more easily assessed from the point of view of the master. The master has been programmed to look at sound similarity which covers the optional requirement for a pun. Surprise is modeled in the master by it minimizing the semantic similarity of the original word and the new replacement. If we define surprise as a failed prediction done by the brain, we can back the master's way of producing it in neuroscience. Research (Lau et al., 2008) has shown that words the brain expects to hear in a sentence cause a lower N400 response than unexpected words. This is because when the brain fails at its expectations, it has to activate the new surprising concepts together with those close to it semantically. This might not, however, be the only explanation for surprise nor a sufficient one. Therefore, the requirement for surprise, though met, has room for improvement.

The last requirement for appreciation was coherence. This has been implemented on the level of semantic coherence to Saudi Arabia. However, as some of the titles written by people, such as *Sheikhs and the City*, show, the contextual coherence is next to impossible to assess without the wider context of the movie itself. We feel that this kind of pragmatic coherence is such a wide task to tackle that it is deserving of a dedicated paper on its own right and thus is beyond the scope of this research. However, it is an important question for the future as it has been shown that humor of the kind we are focusing on in this paper derives its meaning greatly from its pragmatic context (Hämäläinen, 2016).

The appreciation will only be discussed here from the point of view of the master, as the design choice of the system was to give the master the responsibility of appreciation. However, an interesting question for the future is the appreciation learned by the apprentice. Since an NMT model can score its predictions, it has to have learned a kind of an appreciation, but the nature of it is not discussed here. Although this gives an interesting direction for the future research on the topic.

We can demonstrate that the system is capable

of achieving creative autonomy because it can appreciate autonomously its own creations and those of its peers. Furthermore, the changes it makes to its standards are guided by observations made on the artifacts of its peers that have received enough appreciation from the master. Even though the master will follow a limited way of generating humorous titles, its apprentice is liberated from such limits thanks to its neural network architecture.

6.2 Evaluation of the Output

The creative tripod, through which we have defined creativity, requires there to be perceivably skill, appreciation and imagination in the system. The existence of these in the output of the master and the apprentice is evaluated by formulating questions based on the requirements we set for this particular creative task.

6.2.1 Skill

Skill is probably the leg of the tripod that most vocally calls for evaluation from people. In the previous evaluation section, we could only clearly demonstrate that the system fills the most basic requirement, that is to produce a new title out of the existing one.

The further requirements, recognizability of the original title, whether the new title is humoristic and whether it satirizes Saudi Arabia, are beyond any justified assessment without resorting to people's opinion. Thus we need to asses them with the following questions.

1. The title is humorous
2. The original title is recognizable
3. The humor in the title relates to Saudi Arabia
4. The title is critical towards Saudi Arabia

The first question can be asked directly, there is no need to find a better way to ask whether the original title is still recognizable or not. It is important to note that when we are evaluating the skill, we only want to know whether the skill of producing humor exists. The quality of the humor is left for the evaluation of the appreciation.

As for the other questions we want to know whether humor is perceived and whether a relation to Saudi Arabia is perceived. Furthermore, we are curious to see whether criticism is perceived. We will not ask directly whether the title is satirical for two reasons, firstly in order to understand anything as satire, further context is needed and secondly terms such as satire, sarcasm and irony are difficult for an ordinary person to grasp and get often mixed up with one another in people's minds.

6.2.2 Imagination

We took Boden's P- and H-creativity as the basis of the imagination quality of the system. We defined the system to be imaginative enough if it can come up with something new to itself and to exceed the expectations of its imagination if it can come up with something novel in a greater context.

The previous evaluation of the creativity in the system, we have showed that the system is capable of P-creativity and we have shown examples of its H-creativity. While P-creativity, as it is limited to the system itself, has been inarguably proven, the H-creativity calls for further validation. Therefore, we formulate the following evaluation questions to assess the imaginativeness of the system.

5. The joke in the title sounds familiar
6. The joke in the title is obvious

These questions get to the core of what is required from H-creativity, it has to be novel in a global context, i.e. it cannot sound familiar to another slightly different joke the evaluator might have heard before. Also, the joke cannot be too obvious, because if its perceived as an obvious one, people would likely come up with it easily and thus it can hardly be seen H-creative.

6.2.3 Appreciation

Appreciation is something we have discussed to a great extent in the previous evaluation section. As for appreciation, we are not interested to see whether the system gets a high grade from the people for each evaluation question. Instead, we are more interested in seeing the extent to which they are in line. Does the appreciation of the system predict human appreciation in the same variables?

The constituents of the appreciation we identified earlier were pun detection, and humorousness. The latter was then further divided into two appreciable features: surprise and coherence. To put these into the form of evaluation questions, we resulted in the following ones.

7. The title delivers a pun

8. The joke in the title is surprising

9. The joke of the title makes sense in the context of the original movie

The three evaluation questions are meant to evaluate the three requirements respectively. For coherence, we are interested in how coherent the humor is in the context of the movie. A title might be perfectly humoristic if it was in the context of one movie, where as it might make little to no sense in the case of another movie.

6.2.4 Results

The master was made to produce populations[7] for the most popular movie titles. Out of these, we picked at random 50 titles that exhibited appreciation based on the master's own standards. Furthermore, we picked another 50 titles at random that were produced by the apprentice and appreciated by the master.

The questions defined earlier were presented for each title produced by the master and the apprentice to evaluators on a platform called Figure-Eight. Due to the way the platform operates, every title was not presented to every evaluator, but they appeared at random in such a way that each individual title was evaluated by 20 different evaluators. All in all 48 evaluators participated. As knowledge of English is entailed by the system, we defined a further requirement for the evaluators' language to be Arabic. This way they should be familiar with the specialized cultural vocabulary exhibited by the humoristic titles and know enough English to understand the title.

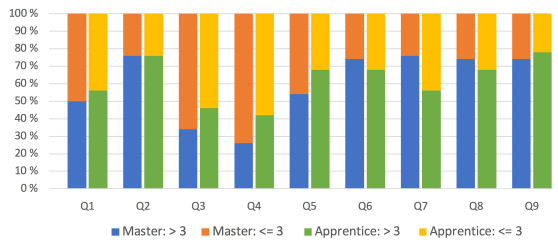

Figure 1: Results of the evaluation for the questions for both the master and the apprentice

For each evaluated title, we calculated the average of the judgments received on each question. Furthermore, judgments are considered agreeing if their average evaluation score is greater than 3, i.e. above neutral. Figure 1 shows the percentage of agreements to neutral or disagreement judgments. Examples of the generated titles can be seen in Table 1,

From the results, it is evident that both methods, the master and the apprentice, were capable of producing humorous titles. Furthermore, it appears that the original movie titles were usually recognized, with 76%, form generated titles. Nevertheless, a low percentage, $\leq 45\%$, of the generated titles were perceived as related to or critical towards Saudi Arabia. In terms of the imagination questions, Q5 and Q6, statistics illustrate that the methods generated familiar and obvious jokes, most of the time; which suggests low H-creativity. Despite that, the methods have received high agreement on the appreciation questions. These results demonstrate that both methods have produced creative alternatives of original movies; however, further development and evaluation is needed to explore the methods and enhance their current state of creativity.

Comparing the results of the master and the apprentice, we notice that generated titles by the apprentice have received different judgments than the master, on most of the questions. Such distinction shows that the apprentice has developed its own internal appreciation. As a result, the apprentice exhibits characteristics of creative autonomy, and, with continuous data from the master and peers, the apprentice will adaptively adjust its appreciation.

7 Conclusions and Final Remarks

We have presented a novel approach for generating satirical humor. Evaluating it critically against the definition of the creativity we established, we identified two clear shortcomings of the approach: firstly a better definition for surprise is needed to better model it in the future, and secondly coherence calls for a better contextual model relating the coherence in the context of the original movie where the title has occurred. Both of these shortcomings require additional research and are worthy of own publications dedicated to the topics.

Our definition of creativity that has been built on top of existing theories gives us a good starting point to conduct any future research on the topic. Also the evaluation questions deriving from the definition provides us with a way of comparing the results of any future improvements to the ones

[7] The evaluated titles are the ones the apprentice was trained with. See Section 5.2

Question	High Scoring Ones	Low Scoring Ones
Q1	*The Lord of the Lambs: The Return of the King* *My Big Saudi Wedding*	*Iraqi of Arabia* *Empire of the State*
Q2	*Muslim League* *Captain Tanker*	*The Sunni: Part III* *Iraqi-man: Houthi*
Q3	*My Big Saudi Wedding* *The Good the Haram and the Saudi*	*Iraqi Buyers Club* *The Wedding Attack*
Q4	*Serabs of Saudi York* *Transformers Saudi of Extinction*	*La La Kill* *The Wedding Attack*
Q5	*Night at The Arabian* *My Big Saudi Wedding*	*La La Invasion* *Attack Powers: The Oil dependence Who Shagged Me*
Q6	*Night at the Arabian* *The Saudi Runner*	*The Sunni Life of Izars* *La La Invasion*
Q7	*Sunnah Squad* *The Twilight Bomb Eclipse*	*Muslim Story 3* *Harry Potter and the Revenge's Sulham*
Q8	*The Twilight Bomb Eclipse* *The Saudi Runner*	*The Imam Ultimatum* *Muslim Fu Panda*
Q9	*The Amazing Surah-man* *The Sound of Jihad*	*The Amazing Spider Mosque* *Harry Potter and the Revenge's Sulham*

Table 1: Examples of high and low scoring titles based on the evaluators' judgment

presented in this paper.

The master-apprentice approach makes it possible for us to study the creativity and its development in the apprentice from a multi-agent perspective. This evokes interesting questions such as: What if there were multiple master-apprentice pairs and they would function as each other's peers? What if an apprentice took classes of multiple masters simultaneously? What if the masters were experts on different fields such as humor and poetry? Would the apprentice then learn to generate based on both fields? We are also interested in diving into the black box of the NMT architecture of the apprentice to see what kind of an appreciation it can develop.

8 Acknowledgements

This work has been supported by the Academy of Finland under grants 276897 (CLiC) and 293009 (STRATAS).

References

Khalid Alnajjar, Hadaytullah Hadaytullah, and Hannu Toivonen. 2018. "Talent, Skill and Support." A method for automatic creation of slogans. In *Proceedings of the Ninth International Conference on Computational Creativity (ICCC 2018)*, pages 88–95, Salamanca, Spain. Association for Computational Creativity (ACC).

Margaret A Boden. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Florin Brad and Traian Rebedea. 2017. Neural paraphrase generation using transfer learning. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 257–261, Santiago de Compostela, Spain. Association for Computational Linguistics.

Hiram H Brownell, Dee Michel, John Powelson, and Howard Gardner. 1983. Surprise but not coherence: Sensitivity to verbal humor in right-hemisphere patients. *Brain and Language*, 18(1):20 – 27.

Simon Colton. 2008. Creativity Versus the Perception of Creativity in Computational Systems. In *AAAI Spring Symposium: Creative Intelligent Systems*, Technical Report SS-08-03, pages 14–20, Stanford, California, USA.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.

Mika Hämäläinen. 2016. Reconocimiento automático del sarcasmo - ¡Esto va a funcionar bien! Master's thesis, University of Helsinki, Finland. URN:NBN:fi:hulib-201606011945.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *To appear*.

Kyle E. Jennings. 2010. Developing Creativity: Artificial Barriers in Artificial Intelligence. *Minds and Machines*, 20(4):489–501.

Anna Jordanous. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, 4(3):246–279.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Ellen F Lau, Colin Phillips, and David Poeppel. 2008. A Cortical Network for Semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, 9:920–933.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

OED. n.d. OED Online. Oxford University Press. Http://www.oed.com/.

Priyanshi R Shah, Chintan D Thakkar, and Swati Mali. 2016. Computational creativity: Automated pun generation. *International Journal of Computer Applications*, 140(10).

Amin Sleimi and Claire Gardent. 2016. Generating Paraphrases from DBPedia using Deep Learning. In *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, pages 54–57.

Alessandro Valitutti, Oliviero Stock, and Carlo Strapparava. 2009. Graphlaugh: A tool for the interactive generation of humorous puns. pages 1 – 2.

Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, and Jukka M Toivanen. 2013. "Let Everything Turn Well in Your Wife": Generation of Adult Humor Using Lexical Constraints. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 243–248, Sofia, Bulgaria.

Tony Veale. 2018. A Massive Sarcastic Robot: What a Great Idea! - Two Approaches to the Computational Generation of Irony. In *Proceedings of the Ninth International Conference on Computational Creativity (ICCC 2018)*, pages 120–127, Salamanca, Spain. Association for Computational Creativity (ACC).

Ping Xiao, Khalid Alnajjar, Mark Granroth-Wilding, Kathleen Agres, and Hannu Toivonen. 2016. Meta4meaning: Automatic metaphor interpretation using corpus-derived word associations. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC 2016)*, Paris, France. Sony CSL, Sony CSL.