Amplifying the Range of News Stories with Creativity: Methods and their Evaluation, in Portuguese

Rui Mendes
CISUC, Dept. Informatics Engineering
University of Coimbra, Portugal
rppm@student.dei.uc.pt

Hugo Gonçalo Oliveira
CISUC, Dept. Informatics Engineering
University of Coimbra, Portugal
hroliv@dei.uc.pt

Abstract

Headlines are key for attracting people to a story, but writing appealing headlines requires time and talent. This work aims to automate the production of creative short texts (e.g., news headlines) for an input context (e.g., existing headlines), thus amplifying its range. Well-known expressions (e.g., proverbs, movie titles), which typically include word-play and resort to figurative language, are used as a starting point. Given an input text, they can be recommended by exploiting Semantic Textual Similarity (STS) techniques, or adapted towards higher relatedness. For the latter, three methods that exploit static word embeddings are proposed. Experimentation in Portuguese lead to some conclusions, based on human opinions: STS methods that look exclusively at the surface text, recommend more related expressions; resulting expressions are somewhat related to the input, but adaptation leads to higher relatedness and novelty; humour can be an indirect consequence, but most outputs are not funny.

1 Introduction

Each minute, new stories are published online, listed in news aggregators, and spread through social media. With the amount of information each of us is constantly bombed with, most people end up looking at the headlines and, only sporadically, opening and reading the full text. We may thus say that headlines play a key role in this process: the more appealing they are, the higher the probability of someone actually reading their story. And while personal preferences are relevant, creativity and familiarity (e.g., text resembles a funny saying or situation) certainly contribute to higher appeal.

A common strategy for making the headline more catchy is to reuse expressions known by a general audience, sometimes also attempting at a humorous effect. If the expression is related enough, it can be used directly, but it may also suffer minor adaptations, to become more related to the context and still resemble the original saying. This is also common in news satire tv shows, like *The Daily Show*. While the host is telling a story, in one of the top corners of the screen, a short text (e.g., a book, movie title, proverb, idiomatic expression, or their adaptation), often accompanied by an image, complements the scene.

However, writing catchy headlines requires time and talent. Even if resorting to well-known expressions is a possible shortcut, it is still a knowledge intensive task (i.e., on folk or pop culture) and requires creativity skills. Therefore, in this paper, we propose to automate this process and assess different unsupervised methods for, given a short text (e.g., a news headline): (i) recommending a related known expression (e.g., a proverb) from a set; (ii) adapting a known expression so that it becomes related to the input. In any case, it should be possible to use the resulting expression as an alternative, though more creative, headline, sub-title, or, at least, a comment in social media. Methods tested for recommendation are mostly baselines for Semantic Textual Similarity (STS) (Agirre et al., 2012), including some based on the surface text, others based on static word embeddings, and on BERT (Devlin et al., 2019) contextual embeddings. As for the adaptation, static word embeddings are exploited for replacing some words in the expressions, taken directly from the headline, or based on similarity or analogy.

Proposed methods were tested with a set of news headlines, as input text, and a set of proverbs and movie titles. We have worked with Portuguese, and even though the resources and tools used are for the target language, theoretically, the methods are language-independent. The development of automatic approaches for producing creative artefacts, sometimes based on a given context or cur-

Proceedings of The 13th International Conference on Natural Language Generation, pages 252–262,
Dublin, Ireland, 15-18 December, 2020. ©2020 Association for Computational Linguistics

rent events, is not new. The proposed adaptation methods have some novelty, when compared to the well-established methods for recommendation. However, here they are used with proverbs, which poses additional challenges, such as the frequent utilisation of figurative language. Finally, we target Portuguese, a language for which work of this kind is still scarce.

For better understanding how successful this approach was, results of this Portuguese instantiation were assessed. Due to the underlying subjectivity in the appreciation of the results, evaluations (one for recommendation, another for adaptation) are based on the opinion of human judges, who scored pairs of headlines and expressions by different methods. This further enabled to compare methods and draw some conclusions, such as: on average, results of recommendation methods are not significantly different from each other, but using expressions that share words with the headline seems to increase the perceived relatedness; resulting expressions are somewhat related to the input, and adaptation leads to higher relatedness and novelty; humour can be an indirect consequence, but most outputs are not funny.

In the remainder of this paper, we overview work on the generation of creative text, focusing on those inspired by news or current events. We then describe our approaches for amplifying news stories with creative text and their evaluation: first, for the recommendation of known expressions, including the comparison of different methods; second, we propose three methods for adapting such expressions. We conclude with a brief discussion and some cues for future work.

2 Related Work

Computational Creativity (CC) is at the intersection of Artificial Intelligence and scientific areas like cognitive psychology, philosophy, and the arts. One of its ends is to develop creative systems, i.e., computational systems that exhibit behaviours deemed as creative by unbiased observers (Colton and Wiggins, 2012), e.g., they can produce new music or visual art, among others, including text.

Work on linguistically-creative systems has tackled the generation of textual artefacts like narrative (Gervás et al., 2019), poetry (Colton et al., 2012; Chrismartin and Manurung, 2015; Gonçalo Oliveira, 2017), memorable news headlines (Gatti et al., 2015; Veale et al., 2017; Alnajjar et al., 2019) and slogans (Alnajjar and Toivonen, 2020), or humour (Binsted and Ritchie, 1997; Valitutti et al., 2016). Among those, some were adapted for producing new textual content based on current events. Poetry generation has been: guided by dependency relations in a single news story (Chrismartin and Manurung, 2015); inspired by the mood and related similes in a set of news stories, possibly including sentences from one of them (Colton et al., 2012); or inspired by Twitter trends, associated words and semantic relations (Gonçalo Oliveira, 2017).

Work on increasing the creativity, and thus memorability, of news headlines often resorts to well-known expressions, with which readers are familiar, including slogans, movie or song titles. On this context, the creativity of an automated journalism system was increased with the injection of known expressions and figurative language (Alnajjar et al., 2019), e.g., the headline "*Biggest gains for The Christian Democrats across Lapin vaalipiiri*" may become "*Biggest gains for The Christian Democrats, as powerful as a soldier, across Lapin vaalipiiri*'. Human-written headlines were also blended with well-known expressions, through the substitution of words in the expression with keywords from the headline (Gatti et al., 2015), e.g., "*What the Euro is coming to*" for the original headline "*UK anger at 1.7bn EU cash demand*". To reduce the risk of producing outputs with a detached meaning, a threshold was set on the cosine similarity between the original and the produced headline. Also relying on vector semantics and on the cosine, previously generated metaphors have been paired with the current news (Veale et al., 2017), e.g., the news tweet "*@newtgingrich says 'the country will become enraged' if the violent protests at @realDonaldTrump rallies continue*" was paired with metaphors like *What is a radical but a crusading demagogue?*.

When it comes to the generation of creative text, word embeddings are useful resources. In fact, the operations of similarity, neighborhood, theme and analogy have been formalised for adapting text by lexical replacement, with increased creativity, given a set of intentions (Bay et al., 2017). In our work, we also propose methods for adapting well-known expressions, based on a textual input (context), more like Gatti et al. (2015), with word embeddings exploited in the selection of suitable replacements. Although any text should work, news headlines were also used.

Humour can be a consequence of the former approaches, but it is rarely tackled specifically. There is, however, work on humour generation, also relying on lexical substitution in short texts (Valitutti et al., 2016). In this case, messages, e.g., "*Later we go where to eat?*" may become "*Later we go where to shit?*". Replacement words should have the same part-of-speech of the original, a similar sound or writing and, the more efficient constraint, is that they are taboos. Hossain et al. (2019) present a corpus of original news headlines and their manually-edited funny versions. While such a corpus may be useful for many tasks, including humor generation, funny headlines are often out-of-context (e.g., "*EU says summit with Turkey provides no answers to concerns*" becomes "*EU says gravy with Turkey provides no answers to concerns*"), which is different from our goal.

Another important difference of our work is that we experiment and present results for the Portuguese language, while the majority of the aforementioned systems processes and produces English text. For Portuguese, related work is less, but still covers: the automatic generation of poetry (Gonçalo Oliveira, 2017); riddles (Gonçalo Oliveira and Rodrigues, 2018), with the help of lexical-semantic knowledge; and memes (Gonçalo Oliveira et al., 2016), based on current news and rules for selecting an image macro and adapting the text.

A related task is that of headline generation for a given document (see e.g., (Takase et al., 2016)), which has some similarities with automatic summarization (Banko et al., 2000) and may involve a deeper understanding of the news story. This is, however, different from our goal, as our starting point are existing headlines, for which new creative alternatives are recommended or produced.

Alternatively to text generation, some systems simply recommend (famous) quotes (Ahn et al., 2016), or retrieve interactions in movie subtitles (Ameixa et al., 2014), to be used in dialogues. This may involve training an encoder-decoder network (Shang et al., 2015) in dialogues where target quotes are used. Though, an unsupervised retrieval-based approach is also possible, e.g., relying on vector semantics and pretrained language models for computing Semantic Textual Similarity (Agirre et al., 2012; Cer et al., 2017), and retrieving the most similar quote. In this work, we also test different unsupervised methods for the direct recommendation of well-known expressions.

3 Recommendation of Creative Text in Context

The first part of this work tackled the recommendation of suitable expressions for a given short text. A good example of our goal is taken from the 17th May 2020 edition of the Portuguese satire news tv show *Isto é Gozar com Quem Trabalha*: when presenting a story about the plan of the Portuguese President to take a swim in the sea after the Covid-19 lockdown, they used the text "*O Velho e o Mar*" (The Old Man and the Sea, a book by Ernest Hemingway).

For this, we tested different methods for computing the semantic similarity between news headlines, the input text, and Portuguese proverbs, the expressions to output. All methods are unsupervised and can be seen as baselines for STS. Once all available proverbs are ranked according to the input, the first, which maximises similarity, is recommended. This section describes the tested methods, how they were used in this experimentation, their results and evaluation by human judges.

3.1 Methods

The following STS methods were covered: shallow methods that consider only the surface text; methods based on sparse vector representations of the text, with word counts and TF-IDF weighting; methods based on static word embeddings, with sentences represented by the average embedding, weighted with TF-IDF or not; and methods based on contextual embeddings that encode each sentence in a single vector. Overall, eight methods were tested: Jaccard Similarity; CountVectorizer; TfIdfVectorizer; GloVe (Pennington et al., 2014) embeddings, with and without TF-IDF; FastText-CBOW (Bojanowski et al., 2017) embeddings with and without TF-IDF; and BERT (Devlin et al., 2019) embeddings. All methods return a value between 0 and 1, proportional to the similarity between the two sentences. By definition, Jaccard Similarity already does this, while, for the other methods, this value is given by the cosine similarity between the vector representations of each text.

3.2 Experimentation Setup

With the News API[1], a set of 60 news headlines was gathered from online editions of Portuguese news-

[1] https://newsapi.org/

papers. The source of expressions was a corpus of 1,617 Portuguese proverbs, obtained from project Natura[2]. For all methods but BERT, headlines and proverbs were tokenized with the NLPyPort package (Ferreira et al., 2019), a layer on top of NLTK (Loper and Bird, 2002) for better handling Portuguese.

Sparse-vector representations of sentences and the computation of TF-IDF were based on the corpus of proverbs. For GloVe and FastText, we used pretrained models for Portuguese, both with 300-sized vectors, respectively from NILC (Hartmann et al., 2017) and fastText.cc[3] repositories. Finally, for BERT, we used a pretrained multilingual model that covers 104 languages, including Portuguese, BERT-Base, Multilingual Cased[4].

All methods were implemented in Python, with the help of the following packages: scikit-learn (Pedregosa et al., 2011), for the CountVectorizer (Count) and TfIdfVectorizer (TFIDF); gensim (Řehůřek and Sojka, 2010), for handling the static word embeddings; and bert-as-a-service[5], for loading and using BERT for encoding sentences.

3.3 Evaluation

As it happens with many other creative outputs, assessing the suitability of a proverb to a headline is a subjective task, which cannot be automatised. Therefore, to compare the performance of each method in the proposed scenario, we relied on human opinions. For this, proverbs were recommended by each method for each of the 60 headlines in the gathered set. Several surveys were then created with Google Forms, each having ten of those headlines followed by the headline's recommended proverbs, at most eight, randomly distributed, without repetitions or any identification with regard to the method.

We then asked 24 Portuguese-speaking volunteers to answer the surveys. Each survey was answered by four different judges, who scored the proverbs recommended for ten headlines, according to their relatedness and funniness, with the following questions: (a) "*How would you rate the relation between the proverbs and the news title?*" [Not related (1); Remotely related (2); Considerably related (3); Extremely related (4)]; (b) "*In relation to the headline, how funny is each proverb?*" [Not funny (1); Remotely funny (2); Considerably funny (3); Extremely funny (4)]. Volunteers were not informed that the expressions were proverbs nor that they had been recommended automatically.

Table 1 shows the distribution of scores for recommendations by each method, according to human opinions. It also includes the median ($\tilde{x}$), which is, nevertheless, not very discriminant. We omit both GloVe and FastText with TF-IDF because their scores were not much different.

Out of the possible scores for relatedness, not related (1) was always the most common. A curious outcome is that pretrained embeddings, like GloVe, FastText and, especially, BERT, do not make much qualitative difference on the results. In fact, most judges gave higher scores to proverbs that share one or more words with the headline, which does not always happen when word or sentence embeddings are involved (e.g., first two examples in table 2). We recall that most proverbs are highly figurative, meaning that models trained in general language may struggle to interpret them.

A deeper look shows that only two methods had at least 10% recommendations with average relatedness scores higher than 3.5, namely the Jaccard Similarity (12%) and TFIDF (10%); and only three recommendations had the highest average relatedness, namely the first three examples in table 2.

Funniness scores are not much different, and only two recommendations got the highest average score from all judges, including the last two examples in table 2. Though not intended, both use taboo words, which should have contributed to their high score. This also suggests that we can increase funniness by forcing the presence of such words (Valitutti et al., 2016). On the other hand, it comes at the cost of lower relatedness.

4 Adaptation of Expressions to a Context

According the previous evaluation, it is not easy to find suitable proverbs for a given headline. This could possibly be improved if the corpus of proverbs were increased. However, headlines can be significantly different and, virtually, on any topic, so another option is to start from any known expression and adapt it, so that it becomes more related to the headline, while still resembling the original expression.

This is not new and is also a commonly adopted strategy in news satires. For instance, in the 5th

[2] natura.di.uminho.pt/wiki/doku.php

[3] https://fasttext.cc/

[4] github.com/google-research/bert

[5] github.com/hanxiao/bert-as-service

Method	Relatedness (%)					Funniness (%)				
	1	2	3	4	$\tilde{x}$	1	2	3	4	$\tilde{x}$
Jaccard	29	24	22	25	2	22	29	28	22	2
Count	35	26	25	14	2	25	34	24	17	2
TFIDF	33	25	26	17	2	28	31	29	13	2
GloVe	34	29	25	13	2	33	24	33	11	2
FastText	41	27	22	11	2	39	28	25	7	2
BERT	41	24	25	11	1	37	27	19	17	2

Table 1: Human evaluation of recommendation approaches.

Headline	Proverb	Method	Rel	Fun
Malásia devolve 150 contentores ilegais de lixo a países subdesenvolvidos (Malaysia returns 150 illegal trash containers to underdeveloped countries)	*Quem faz de si lixo, pisam-no as galinhas* (Whoever makes themselves trash, will be trampled by the chicken)	TFIDF	4	3.5
Tempestade 'Glória' fez 12 mortos em Espanha. Governo culpa alterações climáticas ('Gloria' storm made 12 casualities in Spain. Government blames climate change)	*A culpa morre solteira* (Guilt dies single)	TFIDF	4	3.25
Ainda não é demasiado tarde para salvarmos os oceanos (It is not too late to save the oceans)	*Não deixe para amanhã o que você pode fazer hoje* (Do not leave for tomorrow what you can do today)	BERT	4	2.5
Veredicto abre a porta a protecção para 'refugiados climáticos' (Veredict opens door for protection to 'climate refugees')	*Para trás mija a burra.* (The female donkey pisses backwards)	Jaccard	2.5	4
Judoca Jorge Fonseca galardoado com o prémio Ética no Desporto de 2019 (Judoka Jorge Fonseca is awarded with the prize 'Sport Ethics 2019')	*Não contes com o ovo no cu da galinha.* (Do not count with the egg in the chicken's butt.)	Jaccard	1.5	4

Table 2: Highest-scored recommendations.

April 2020 edition of *Isto é Gozar com Quem Trabalha*, the expression "*Droga de Elite*" (Elite drug), an adaptation of the Brazilian movie title "*Tropa de Elite*" (Elite Squad), illustrated a story on Covid-19 in the Brazilian favelas, where drug dealers were ensuring that residents followed the sanitary rules.

This section describes three methods that explore static word embeddings in the automatic adaptation of a given expression, to suit, as much as possible, an input short text, in such a way that it can be used for transmitting or complementing the same idea, though more creatively. After presenting the methods, we describe an experiment where Portuguese proverbs and movie titles were adapted for news headlines, followed by their results and evaluation.

4.1 Methods

We propose three automatic methods for adapting a known expression for a given short text. Besides a list of well-known expressions, to be modified according to the input text (e.g., news headline), all methods: (i) exploit a pretrained model of static word embeddings; (ii) assume that the most relevant words in a text are previously computed; (iii) go through all the expressions in a list and try to make adaptations guided by the most relevant words of both the expressions and the input texts. Methods only differ on the adopted strategies for selecting the word(s) to replace.

The first method, Substitution, replaces the most relevant word in the expression, a, by a word from the input text, b, or by a word similar to b. Our intuition is that, by using a relevant word of the input text, the meaning of the expression becomes more semantically-related to the given context.

The second method, Analogy, relies on a common operation for computing analogies in word embeddings, i.e., $b - a + a^* = b^*$ (Mikolov et al., 2013), phrased as b^* *is to* b *as* a^* *is to* a. The strategy is to use the two most relevant words in the expression as a and a^*, the most relevant word in the input as b, and then: (i) from the previous three, compute a new word b^*; (ii) in the original expression, replace a and a^* by b and b^*, respectively. Given that both pairs of words are analogously-related, our intuition is that the result will still make sense and be more related to the input text.

The third method, Vector Difference (VecDiff), also selects the two most relevant words in the input text, b and b^*, and then: (i) computes the vector between the previous $b - b^*$; (ii) identifies the pair of open-class words in the expression, a and a^*, such that $a - a^*$ maximises the (cosine) similarity with $b - b^*$; (iii) replaces a and a^* by b and b^*, respectively. Our intuition is that the new text will not only use two words of the input, and thus be more related, but also that they will be included in such a way that their relation is roughly preserved.

To avoid syntactic inconsistencies, in any method, replacement candidates must match the morphology of the replaced word, including part-of-speech (PoS), gender and number, according to a morphology lexicon or a PoS tagger. If morphology does not match, the lexicon can be further used to inflect the candidate to the target form. Moreover, the set of possible replacements can be augmented by considering not only the relevant words in the input, but also their most semantically-similar, computed in the embeddings, e.g., in the Substitution method, a can be replaced by a word different but semantically similar to b.

Table 3 shows an example of the application of each method, including the original headline, a proverb and the resulting output. Replaced words and their replacements are underlined. In the first example, $b = bancos$ replaces $a = amigos$. In the second, $comeces = apontar - deixes + fazer$. In the third, $fere - ferido \approx finge - detido$.

4.2 Experimentation Setup

Although the proposed methods are language dependent, we focused again on Portuguese. This time, we used the same ≈1,600 proverbs as in the previous section, but added about 2,500 movie titles in Portuguese, obtained from IMDB[6]. For better identification of the titles, we discarded the 25% oldest and all others with less than four words. Moreover, to avoid the inclusion of English names, all words in the title had to be in a Portuguese morphology lexicon.

Regarding pre-processing, NLPyPort (Ferreira et al., 2019) was used for tokenization, PoS tagging and lemmatization, and the morphology lexicon LABEL-Lex (Ranchhod et al., 1999) for information on word inflections. Relevant words were considered to be the least frequent, but appearing, in the newspaper corpus CETEMPúblico (Rocha and Santos, 2000). For word embeddings, we used the same GloVe model as in the previous section.

For experimentation, we used a set of 100 news headlines on different topics, posted between April and May 2020, in the Twitter accounts of Portuguese newspapers. An initial set was randomly selected, but darker headlines (e.g., about death) were manually excluded later, to make 100. For each headline, new texts were produced with the previous three methods. However, application to each headline could result in several new texts. Even if, due to the morphology constraints, some expressions end up not being used, others will. Plus, for each relevant word, we considered the top-5 similar. Thus, the same method may produce several variations of the same text for the same input, one for each.

For selecting a single expression, the recommendation methods described in section **??** are used in the beginning and in the end of the process. First, given a headline, they select the subset of the full corpus of expressions, in which each adaptation method will be applied to. Since the previous evaluation (section 3.3) did not help in the choice of a single recommendation method, we used two significantly different methods for this selection: TF-IDF and BERT. More precisely, given the headline, this subset will contain the top-30 expressions recommended by TF-IDF, the top-30 recommended by BERT, plus a random selection of 30, for higher diversity. This should still result in several (adapted) expressions, where a single one has to be selected from. This final selection is again made with the help of TF-IDF or BERT, which rank the results, allowing us to use only the top-ranked, i.e., the most similar to the headline, as long as it is not equals to an existing expression. Thus, depending on the final recommendation method, the resulting expression may be different.

4.3 Evaluation

To get insights on the suitability of the proposed methods, we relied again on human opinions. Since text was being changed automatically, it was now important to assess syntax, i.e., whether the resulting expression had grammatical or structural issues, that could make it difficult to interpret. However, syntax is less subjective than the other criteria, so it was assessed by the authors in a set of expressions produced for another subset of 30 headlines. This was enough to conclude that there were only syntactic issues in a minority of the produced expressions – 3% for Substitution, 8% for Analogy, 13% in Vector Difference – and the majority of these issues did not have much impact on interpretability.

Furthermore, we decided to assess an important aspect of creativity, which is novelty. In addition to being related to the input text, results should also be novel, in the sense that they have not been produced before or associated with the input text. Ideally, in addition to those, it would be nice again if the new expression had the potential of making

[6] https://www.imdb.com/interfaces/

Method	Headline	Proverb	Output
Substit	*Bancos preparam-se para dar menos crédito às famílias* (Banks preparing to give less credit to families)	*amigos, amigos, negócios à parte* (Friends, friends, business apart)	*bancos, bancos, negócios à parte* (Banks, banks, business apart)
Analogy	*EUA estão a apontar para o pior número de desemprego da sua história* (USA are pointing out to the worst unemployment numbers in their history)	*não deixes para amanhã o que podes fazer hoje* (Do not leave for tomorrow what can be done today)	*não comeces para amanhã o que podes apontar hoje* (Do not start tomorrow what you can point out today)
VecDiff	*Finge ter Covid-19 no Facebook e acaba detido* (Pretends to have Covid-19 on Facebook and ends up arrested)	*quem com ferro fere, com ferro será ferido* (Those who hurt with iron, with iron shall be hurt)	*quem com ferro finge, com ferro será detido* (Those that pretend with iron, with iron shall be arrested)

Table 3: Running examples of the application of each adaptation method.

the reader laugh (funniness).

To have resulting expressions scored according to the evaluation criteria, a survey was again created, with a headline in each page, followed by expressions from eight different processes, namely the result of adapting the TF-IDF or BERT recommendations to the output of the three adaptation methods (6), plus expressions recommended directly by TF-IDF and BERT, with no adaptation (2). The latter were included to enable a comparison between reusing well-known expressions directly or with an adaptation, also having in mind that, in section 3, the headlines were different and so was the evaluation scale.

For each resulting expression, human judges were asked to score the relatedness, by selecting one of the following: (1) *There is no relation at all between the generated expression and the input*; (2) *The expression is somewhat related to the input, because it shares one or two words, or other contextual aspects*; (3) *The expression is clearly related to the input's context, could replace it or be used as a comment*. Novelty, could be scored as: (1) *I knew the expression well before reading it*; (2) *Reminds me of some expression, but has some changes*; (3) *The expression is completely new to me*. Finally, for funniness, the options were: (1) *The expression is not funny and should make no one laugh*; (2) *The expression is somewhat funny or could be, depending on the reader's subjective view*; (3) *The expression is very funny and has a great potential to make people laugh*.

This time, we had 100 Portuguese-speaking volunteers, each answering the previous questions for five headlines and their eight expressions, in such a way that each expression was scored by five different judges, which were not informed that the expressions were produced automatically. Table 4 has the distribution of scores given in the human evaluation, for text produced by each method.

We recall that one of the motivations for adapting text instead of simply reusing it, was to increase relatedness. Even though the impact was not as high as expected, this trend is confirmed by the higher proportion of the highest scores (2 and 3) for the adaptation methods on this criteria. This is more clear if we look at results using the same method for the final recommendation. Moreover, figures suggest again that TF-IDF leads to better relatedness, because texts recommended by this method have a higher proportion of results scored with 3 than those by BERT. In section 3.3, we pointed out that this can be due to BERT recommendations sharing less words with the context. Yet, in the future we should also explore different ways of using BERT for this purpose.

When using TF-IDF in the final recommendation, about a third of the results is highly related to the headline, which is ok, especially if, before utilisation, results can be manually selected out of the top-ranked. Curiously, the proportion of completely related results is higher for the simplest method, Substitution, and lower for the Vector Difference. The latter is also the only of the three methods that, in this scenario, has more than a third of results considered unrelated to their headline. Table 5 presents three high-scored examples, with the adaptation method, the input headline, the original expression and the output expression, with considered words underlined and average scores in the three assessed criteria.

On the remaining criteria, adapted expressions are more novel, a key aspect for a creative system that shows we can go further than simply reusing text. This was expected because we are comparing newly created expressions with proverbs and movie titles, most of which are part of our culture and known by many people. At the same time, even though they are new, resulting expressions should still resemble existing expressions, which is

Method	Relatedness (%)				Novelty (%)				Funniness (%)			
	1	2	3	x̃	1	2	3	x̃	1	2	3	x̃
Final recommendation by TF-IDF												
Substitution	25.8	38.0	36.2	2	17.6	40.8	41.6	2	45.2	29.4	25.4	2
Analogy	29.0	37.8	33.2	2	17.4	36.4	46.2	2	44.6	29.8	25.6	2
Vector Difference	34.6	35.6	29.8	2	17.0	35.4	47.6	2	53.4	27.0	19.6	1
Recommendation only	38.4	32.0	29.6	2	33.8	34.6	31.6	2	53.0	27.8	19,2	1
Final recommendation by BERT												
Substitution	44.8	34.0	21.2	2	22.0	36.4	41.6	2	49.2	30.6	20.2	1
Analogy	38.8	36.6	24.6	2	20.0	33.8	46.2	2	52.0	28.2	19,8	1
Vector Difference	38.4	35.6	26.0	2	16.4	32.2	51.4	3	51.8	24.2	24.0	1
Recommendation only	52.5	27.8	19.8	1	22.0	35.0	43.0	2	46.0	29.8	24.2	2

Table 4: Human evaluation of the adaptation and recommendation approaches.

why not all are scored with 3. For the same reason, novelty of the recommendation methods was lower, but not always 1, possibly because some judges did not know all the expressions. This, however, could have also influenced, positively, the scores of the adaptation methods, i.e., if the judge does not know the original expression, they will also not associate its adaptation with an expression they previously heard. On this criteria, BERT did not make much difference when applied to the adaptation methods, but novelty of its recommendations is significantly higher than for TF-IDF. We also note that the ranking of the adaptation methods according to novelty is the inverse of the relatedness, possibly because some judges rated the novelty when compared to the headline. The middle example in table 5 is one of the best scored results on novelty.

About a half of the results of each method is just not funny. On the other hand, the proportion of clearly funny results is between 19% and 26%, which, considering that this criteria is not explicitly tackled, is not bad, and suggests that, indeed, humour can be a consequence of this word-play. All the examples in table 5 have average funniness of 2 or higher. In the first, funniness is more subjective, but the result may suggest too much promiscuity between banks and the government. The last two make unexpected associations, like *people* with a *herd*, or *honey / love* with *despair*.

5 Conclusions

Aiming at the amplification of news stories, we explored a set of automatic methods for making headlines more creative and appealing, with the exploitation of well-known expressions and word embeddings. When tested with Portuguese news headlines, human opinions suggest that our goals were achieved with relative success. Results are somewhat related to input headlines, especially if known expressions are adapted, and not used directly, which also increases novelty, intimately related to creativity. For the best adaptation methods, about a third of the results was clearly related to the headline. Although, in a few cases, humour was an indirect consequence, most of the outputs were not so funny. In the future, funniness may benefit from ranking candidate results with a humour classifier, based on different humour-relevant features, such as the one recently proposed for Portuguese (Clemêncio et al., 2019).

All the methods are unsupervised and exploring supervision was never our goal, mainly because the lack of training data, e.g., different Portuguese headlines for the same news scored according to their creativity, possibly also including some created specifically for this purpose. While such a dataset is not available for Portuguese, for English, a crowdsourced corpus of 15k original news headlines and their manually-edited funny versions was recently presented (Hossain et al., 2019), making thus room for a data-driven approach for our task.

Another observation of our work was that, in this context, methods that consider exclusively the surface text retrieve proverbs that are perceived to be more related. This happens because recommended proverbs tend to use some words of the input, immediately suggesting some relation. Although state-of-the-art BERT embeddings would be better at representing sentence meanings, they are not apt to deal with the figurative language in proverbs, at least if they are not fine-tuned for this, or possibly trained only for the target language. In the future, it would be interesting to make a deeper study on STS and figurative language.

The proposed methods can be integrated in a tool for suggesting alternative headlines that are still related to a story, possibly useful for journalists. Even if not all results have the necessary

Method	Input	Original Expression	Output	Scores
Subs + **TF-IDF**	*Bancos dizem que as condições das linhas de crédito foram definidas pelo governo* (Banks claim that credit conditions were defined by the government)	*Dar a César o que é de César e a Deus o que é de Deus* (To Caesar what is Caesar's and to God what is God's)	*Dar a governo o que é de governo e a Deus o que é de Deus* (To the government what is the government's and to God what is God's)	**Rel**= 3 **Nov**= 2 **Fun**= 2.5
Analogy + **BERT**	*Uma simples conversa gera gotículas que podem ficar suspensas no ar até 14 minutos* (A simple conversation generates droplets that can be suspended in the air for up to 14 minutes)	*Uma ovelha má põe o rebanho a perder* (A bad sheep makes the herd lose)	*Uma gotícula suspensa põe o rebanho a perder* (A suspended droplet makes the herd lose)	**Rel**= 2.25 **Nov**= 2.75 **Fun**= 2.5
VecDiff + **BERT**	*Noivos desesperam por terminar lua de mel, mesmo estando "presos" nas Maldivas* (Newlyweds are desperate to end their honeymoon, even though they are "stuck" in the Maldives)	*Ouro é o que ouro vale* (Gold is what gold is worth)	*Mel desespera o que mel vale* (Honey despairs honey is worth)	**Rel**= 2.25 **Nov**= 2.5 **Fun**= 2

Table 5: Examples of high-scored expressions according to the human judges.

quality, different options may be produced (e.g., the top-n for each method), faster than if it were a human, who may still make the final decision or additional adaptations. To some extent, enabling human interaction would make such a tool co-creative, as other interactive systems developed for writing stories (Roemmele and Gordon, 2015), song lyrics (Watanabe et al., 2017) or poetry (Gonçalo Oliveira et al., 2017), among others.

Finally, this work should also contribute to the development of more creative chatbots, e.g., capable of providing creative responses to out-of-domain interactions. In the meantime, the TECo Twitterbot[7] is already following several Portuguese newspapers and retweeting some of their publications together with comments produced by the proposed methods.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of 1st Joint Conference on Lexical and Computational Semantics-Vol. 1: Proceedings of main conference and shared task, and Vol. 2: Proceedings of Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Yeonchan Ahn, Hanbit Lee, Heesik Jeon, Seungdo Ha, and Sang-goo Lee. 2016. Quote recommendation for dialogs and writings. In *Proceedings of 3rd Workshop on New Trends in Content-Based Recommender Systems, CBRecSys@ RecSys*, pages 39–42.

Khalid Alnajjar, Leo Leppänen, and Hannu Toivonen. 2019. No time like the present: Methods for generating colourful and factual multilingual news headlines. In *Proceedings of 10th International Conference on Computational Creativity*, ICCC 2019, pages 258–265. Association for Computational Creativity.

Khalid Alnajjar and Hannu Toivonen. 2020. Computational generation of slogans. *Natural Language Engineering*, pages 1–33.

David Ameixa, Luisa Coheur, Pedro Fialho, and Paulo Quaresma. 2014. Luke, i am your father: dealing with out-of-domain requests by using movies subtitles. In *International Conference on Intelligent Virtual Agents*, pages 13–21. Springer.

Michele Banko, Vibhu O Mittal, and Michael J Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 318–325.

Benjamin Bay, Paul Bodily, and Dan Ventura. 2017. Text transformation via constraints and word embedding. In *Proceedings of 8th International Conference on Computational Creativity*, ICCC 2017, pages 49–56.

Kim Binsted and Graeme Ritchie. 1997. Computational rules for generating punning riddles. *Humor: International Journal of Humor Research.*

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. Association for Computational Linguistics.

[7] https://twitter.com/TextoEmContexto

Berty Chrismartin and Ruli Manurung. 2015. A chart generation system for topical meaningful metrical poetry. In *Proceedings of The 6th International Conference on Computational Creativity*, ICCC 2015, pages 308–314, Park City, UT, USA.

André Clemêncio, Ana Alves, and Hugo Gonçalo Oliveira. 2019. Recognizing humor in Portuguese: First steps. In *Proceedings of 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3-6, 2019, Part II*, volume 11805 of *LNCS/LNAI*, pages 744–756. Springer.

Simon Colton, Jacob Goodwin, and Tony Veale. 2012. Full FACE poetry generation. In *Proceedings of 3rd International Conference on Computational Creativity (ICCC)*, pages 95–102, Dublin, Ireland.

Simon Colton and Geraint A Wiggins. 2012. Computational creativity: the final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 21–26. IOS Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

João Ferreira, Hugo Gonçalo Oliveira, and Ricardo Rodrigues. 2019. Improving NLTK for processing Portuguese. In *Proceedings of 8th Symposium on Languages, Applications and Technologies (SLATE 2019)*, volume 74 of *OASIcs*, pages 18:1–18:9. Schloss Dagstuhl.

Lorenzo Gatti, Gözde Özbal, Marco Guerini, Oliviero Stock, and Carlo Strapparava. 2015. Slogans are not forever: Adapting linguistic expressions to the news. In *Proceedings 24th International Joint Conference on Artificial Intelligence*, IJCAI 2015, pages 2452–2458. AAAI Press.

Pablo Gervás, Eugenio Concepción, Carlos León, Gonzalo Méndez, and Pablo Delatorre. 2019. The long path to narrative generation. *IBM Journal of Research and Development*, 63(1):8–1.

Hugo Gonçalo Oliveira. 2017. O Poeta Artificial 2.0: Increasing meaningfulness in a poetry generation Twitter bot. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, pages 11–20, Santiago de Compostela, Spain. Association for Computational Linguistics.

Hugo Gonçalo Oliveira, Diogo Costa, and Alexandre Pinto. 2016. One does not simply produce funny memes! – explorations on the automatic generation of Internet humor. In *Proceedings of 7th International Conference on Computational Creativity*, ICCC 2016, pages 238–245, Paris, France.

Hugo Gonçalo Oliveira, Tiago Mendes, and Ana Boavida. 2017. Co-PoeTryMe: a co-creative interface for the composition of poetry. In *Proceedings of 10th International Conference on Natural Language Generation*, INLG 2017, pages 70–71, Santiago de Compostela, Spain. ACL Press.

Hugo Gonçalo Oliveira and Ricardo Rodrigues. 2018. Exploring lexical-semantic knowledge in the generation of novel riddles in Portuguese. In *Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation*, CC-NLG 2018, pages 17–25, Tilburg, The Netherlands. Association for Computational Linguistics.

Nathan S. Hartmann, Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues, and Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of 11th Brazilian Symposium in Information and Human Language Technology (STIL 2017)*.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "President Vows to Cut <Taxes> Hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of 2013 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 746–751. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Elisabete Ranchhod, Cristina Mota, and Jorge Baptista. 1999. A computational lexicon of Portuguese for automatic text parsing. In *Proceedings of SIGLEX99 Workshop: Standardizing Lexical Resources*. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Paulo Alexandre Rocha and Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pages 131–140, São Paulo. ICMC/USP.

Melissa Roemmele and Andrew S Gordon. 2015. Creative help: a story writing assistant. In *International Conference on Interactive Digital Storytelling*, pages 81–92. Springer.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1054–1059.

Alessandro Valitutti, Antoine Doucet, Jukka M. Toivanen, and Hannu Toivonen. 2016. Computational generation and dissection of lexical replacement humor. *Natural Language Engineering*, 22(5):727–749.

Tony Veale, Hanyang Chen, and Guofu Li. 2017. I read the news today, oh boy. In *International Conference on Distributed, Ambient, and Pervasive Interactions*, pages 696–709. Springer.

Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, Tomoyasu Nakano, Satoru Fukayama, and Masataka Goto. 2017. Lyrisys: An interactive support system for writing lyrics based on topic transition. In *Proceedings of the 22nd international conference on intelligent user interfaces*, pages 559–563.