TECo: Exploring Word Embeddings for Text Adaptation to a given Context

Rui Mendes & Hugo Gonçalo Oliveira
CISUC, Department of Informatics Engineering
University of Coimbra, Portugal
rppm@student.dei.uc.pt, hroliv@dei.uc.pt

Abstract

TECo adapts well-known sayings (e.g., proverbs, movie titles) for a context, given by a textual input (e.g. news headline). For this, it may use one of three methods – word substitution, analogy and vector difference – which are different ways of exploiting word embeddings for word replacement towards a new text that should be semantically related to the input, in the sense that it could be used as a more creative sub-title or comment. These were further combined with two selection methods, based on word overlap and on sentence embeddings, and used in the production of text in context with a small set of Portuguese headlines. To better understand how well our purpose was suit, results were manually assessed. All methods produced text with both syntax and relatedness to the input above average, contrasting the underachieving funniness scores.

Introduction

To amplify the range of a given story, either real or made-up, authors commonly reuse expressions or sayings known by a general audience as a title or subtitle, sometimes also achieving a humorous effect. If the saying 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. Working on the automation of this process is thus natural. In fact, in scope of linguistic computational creativity, related systems have been developed for generating new creative headlines by resorting to figurative language (Alnajjar, Leppänen, and Toivonen 2019), or blending them with well-known expressions (Gatti et al. 2015); poetry inspired by news stories (Colton, Goodwin, and Veale 2012; Chrismartin and Manurung 2015) or Twitter trends (Gonçalo Oliveira 2017); or applying metaphors to the current news (Veale, Chen, and Li 2017). Having in mind the transformation of text with replacements constrained by the given intentions, operators on word embeddings were even formalised (Bay, Bodily, and Ventura 2017). Other systems simply recommend quotes to be used in dialogues (Ahn et al. 2016), or assign proverbs to news headlines (Mendes and Gonçalo Oliveira 2020).

As a complement to both previous works, and following the idea of using word embeddings, we propose three different methods that exploit this kind of word representation for adapting selected text, so that relatedness to the context increases: substitution of a word by another from or related to the context; substitution of two words by two analogously-related, one of which from the context; and substitution of two words related to the context, in such a way that their relation is preserved.

Although the proposed methods are language-dependent, this study is focused on Portuguese. We rely on Portuguese sayings, namely proverbs and movie titles, and integrate everything in a system dubbed TECo (standing for *Texto Em Contexto*, in English, Text in Context). Methods based on Term Frequency (TF-IDF) and on sentence embeddings – BERT encodings (Devlin et al. 2019) – were used for both selecting an initial set of sayings to be adapted, and selecting the final result to exhibit, out of all produced.

To better understand the potential of TECo, we ran different combinations of selection and adaptation methods and assessed their results for 30 headlines, manually. For all methods, syntax was generally good, relatedness was above average and funniness below. The most similar sayings in the original list, according to the selection methods, were assessed with the same criteria, with TF-IDF having comparable scores but BERT clearly lower. This suggests that the proposed adaptation methods are capable of creating new text, from a lower amount of original examples, and of comparable quality.

The paper is organised as follows: after this introduction, the proposed methods are described; results of the manual assessment are then presented and discussed; finally, we conclude with a brief discussion.

Methodology

The goal of TECo is to produce new text that resembles a known saying but is related, as much as possible, to an input short text, such that it can be used as a more creative way of transmitting the same idea, complementing it or just commenting on it. We propose three automatic methods for this adaptation: Substitution, Analogy and Vector Difference (hereafter, VecDiff). Besides a set of well-known sayings (TECO's knowledge base, hereafter KB), to be modified according to the input text (in this case, news headlines), all methods: (i) exploit a pre-trained model of static word embeddings, where words are represented by dense numeric vectors; (ii) assume that the most relevant words in a text are

Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)
ISBN: 978-989-54160-2-8

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 1: Running examples of the application of each adaptation method.

the open-class words (nouns, verbs, adjectives) that are used in a large corpus but have the lowest frequency (roughly, a high Inverse Document Frequency); (iii) go through all the sayings in a set and try to make adaptations focused on the most relevant words of both the sayings and the input texts. Methods only differ on the adopted strategies for selecting the word(s) to replace.

The first method, Substitution, is the simplest. It replaces the most relevant word in the saying, a, by a word from the input text, b. Our intuition is that, by using a relevant word of the input text, the meaning of the saying 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, Yih, and Zweig 2013), phrased as b^* *is to* b *as* a^* *is to* a. The strategy is to use the two most relevant words in the saying as a and a^*, and the most relevant word in the input as b. Then: (i) from the previous three, compute a new word b^*; (ii) in the original saying, replace a and a^*, respectively by b and b^*. 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, 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 saying, a and a^*, such that $a - a^*$ maximises the (cosine) similarity with $b - b^*$; (iii) replace a and a^* respectively by b and b^*. 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.

Although the proposed methods are language-dependent, TECo is focused on Portuguese, our mother tongue. Its KB includes 1,600 Portuguese proverbs from project Natura[1] and over 3,000 movie titles in Portuguese, from IMDB[2]. Most should be well-known, as proverbs are part of the quotidian of most Portuguese people, being used to emphasize certain situations, usually implying some kind of humour. Moreover, these sayings are not usually to be taken literally, as they use several stylistic variations and their underlying meaning may not be clearly understood by a computer.

Table 1 illustrates some results of each method in this context, including an original headline, a proverb and the resulting output. Replaced words and their replacements are underlined. Results were produced with a pre-trained GloVe model of word embeddings, with 300-sized vectors (Hartmann et al. 2017), and relevant words were computed with the help of the newspaper corpus CETEMPúblico (Rocha and Santos 2000). In the first example, $b = bancos$ replaces $a = amigos$. In the second, $a = deixes$, $a^* = fazer$ and $b = apontar$, with $b* = b - a + a^* = comeces$. In the final example, $a = fere$, $a^* = ferido$, to which, out of the words in the headline, $b = finge$ and $b^* = detido$ is the pair with the most similar difference.

To avoid syntactic inconsistencies, for any method, replacement candidates must match the morphology of the replaced word, including part-of-speech (PoS), gender and number, obtained from the morphology lexicon LABEL-Lex (Ranchhod, Mota, and Baptista 1999). If necessary, it may suffer a disambiguation process with a PoS tagger, for which we used the one in NLPyPort (Ferreira, Gonçalo Oliveira, and Rodrigues 2019). The latter is also used for lemmatization enabling that, if morphology does not match, the lemma of the candidate can be inflected to the target form, with the help of the lexicon. If it is still not possible, the saying is just not considered. In any case, the set of possible replacements can be augmented by considering not only the relevant words in the input text, but also the most semantically-similar words, computed in the embeddings. For example, in the Substitution method, a can be replaced by a word different but semantically similar to b.

Finally, running through all sayings in the KB should result in several new texts. Even if, due to the morphology constraints, some sayings end up not being used, if similar words are considered for the same input, the same method may produce several variations of the same text. Therefore, a final step has to select the most similar text with the input, according to a sentence similarity method, such as those previously tested in a similar scenario (Mendes and Gonçalo Oliveira 2020). Also, to avoid that, for each input, all sayings in the KB are tested, an initial selection may also rely on such similarity methods or their combination.

Evaluation

To take initial conclusions, we ran all the methods in a set of 30 news headlines, with results of each method then manually assessed by two human judges. Besides some insights on the suitability of each method for our purpose, we tested

[1] https://natura.di.uminho.pt

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

Method	Syntax				Relatedness				Funniness			
	1(%)	2(%)	3(%)	Md	1(%)	2(%)	3(%)	Md	1(%)	2(%)	3(%)	Md
Substitution	0.0	6.2	93.8	3	35.8	21.7	42.5	2	41.7	45.8	12.5	2
Analogy	0.0	8.7	91.3	3	18.7	31.3	50.0	2.5	46.3	43.3	10.4	2
Vector Diff	0.0	12.9	87.1	3	13.3	22.5	64.2	3	31.7	55.4	12.9	2
TF-IDF	0.0	8.6	91.4	3	14.3	24.0	61.7	3	36.9	50.5	12.6	2
BERT	0.0	7.4	92.6	3	36.4	23.1	40.5	1	47.1	42.4	10.5	1

Table 2: Scores distribution.

combinations of two selection methods – TF-IDF and BERT – and included original sayings, directly selected by each of them, in the evaluation. Both get the most similar sayings to the input context. The difference is that TF-IDF represents each text as a weighted vector, based on all the sayings, while BERT encodes each text as a 768-sized vector according to pre-trained model covering 104 languages[3]. Those methods were also used before and after adaptation, for making the initial selection of sayings to use, and for selecting the final output, out of all adapted sayings. Therefore, a total of 14 texts were obtained for each headline: twelve ($3\times(2+2)$) produced by each method with a different combination of selection methods, plus two by each selection method alone. The initial selection contained 100 sayings, including the top-50 related according to the selection method and a random selection of other 50 sayings, for higher diversity.

Judges were presented with headlines, followed by the list of texts by each combination, and were asked to use a 3-point Likert scale for ranking: syntax (1, text has several grammatical and/or structural issues, and may be difficult to interpret; 2, text has minor issues regarding grammar and structure, but is still understandable; 3, text does not have any grammatical or structural issues); relatedness (1, minimal or no relation at all between the text and the headline; 2, text somewhat related to the input; 3, relation between the text and the headline is clear / could be used as a substitute or a comment); and funniness (1, not funny and will not make anyone laugh; 2, somewhat funny and could be potentially be funny, depending on the reader's subjective view; 3, very funny, with a great potential to make people laugh). Table 2 shows the distribution of scores and their median (Md), for texts produced by each adaptation method, regardless of the judge and selection methods, plus the output of the two selection methods, when applied directly to the full KB.

Judge agreement, measured with Cohen's Kappa, was 0.57 (moderate) for syntax, 0.35 (fair) for relatedness and 0.17 (fair) for funniness. Syntax is more objective, and thus agreement was higher. On the other hand, the other two aspects, especially funniness, are highly subjective, also due to the structure and figurative language of the Portuguese proverbs and vagueness of some movie titles.

According to the scores, syntax is not severely affected by the adaptations, meaning that the produced text is generally grammatical. Few exceptions occur in the adaptation of verbs. Specifically, in Portuguese, the same verb has often different forms for different tenses, genders and numbers, but the same form may also work for different tenses. Thus, incorrectly identifying the tense in the original saying may result in using an incorrect form in the adapted text.

Regarding relatedness, scores are above average. We highlight VecDiff with 64% texts clearly related with the headline and only 13% with no relation. Analogy got 50% clearly related and Substitution 42%, with 35% not related. This makes sense because, while Substitution makes a single replacement, the other two replace two words that, nevertheless, try to keep the original relation, with VecDiff using two words from the original context. On the selection methods, TF-IDF surprisingly got 61% clearly related selections, but BERT selected the lowest proportion of related sayings.

On funniness, results were not as good, as very funny texts were not much more than 10%. This may be due to the aforementioned subjectivity of humour, which may hamper the judge's decision to give the maximum funniness to a text, because actually making other people laugh depends on many variables. Moreover, the capability of producing content with humoristic value is highly dependent on the context of the input, e.g. it is harder to generate content regarding sad news headlines.

Table 3 illustrates some of the results produced. The first three got the maximum score in all aspects by all judges, and the final two got the lowest scores in relatedness and funniness. Furthermore, it is important to state that most of the resulting texts with minimum scores in both relatedness and funniness used BERT for their final selection, as it seems to suffer from the figurative language used in the sayings, and often selects one that is too distant from the headline, with less focus on shared words. On the other hand, TF-IDF tends to select expressions that share words with the input, thus increasing their relatedness and achieving scores similar to the adaptation methods. This should, however, be analysed more deeply in the future.

Conclusion

Briefly, this study proposed three text adaptation methods to bring a well-known saying closer to a given context, with positive results on syntax and relatedness to the context, but not so much on funniness. When compared to the usage of existing sayings, selected with TF-IDF, with no adaptation, scores are very similar. This shows that the adaptation methods are indeed capable of creating new syntactically-correct and related text, and thus a good option when the number of sayings is limited. We recall that selection methods were applied to the full KB, with 4,600 sayings, while adaptation used only a subset of 100, which they were able to adapt for increased relatedness.

For future endeavours, it would be prolific to test and as-

[3] https://github.com/google-research/bert

Method	Headline	Proverb	Output
Substit + TF-IDF	*Rooney sobre cortes de salários: 'Porque é que são os futebolistas os bodes expiatórios?'* (Rooney about salary cuts: 'Why are footballers the scapegoats?)	*Os amigos são para as ocasiões* (Friends are for the ocasions)	*Os expiatórios são para as ocasiões.* (Scapegoats are for the ocasions)
Analogy + 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)	*paga o justo pelo pecador* (The fair pays for the sinner)	*paga o definido pelo pecador* (The defined pays for the sinner)
VecDiff + BERT	*Ronaldo juntou a família na quarentena para cantar os parabéns à sobrinha* (Ronaldo brings family together during quarentine to sing happy birthday to his niece)	*papagaio come o milho, periquito leva a fama.* (Parrot eats the corn, but the parakeet gets the fame)	*papagaio come o milho, sobrinhinho leva a fama.* (Parrot eats the corn, but the little nephew gets the fame)
Substit + BERT	*Finge ter Covid-19 no Facebook e acaba detido* (Pretends to have Covid-19 on Facebook and ends up arrested)	*Nem por ser Natal* (Not even for being Christmas)	*nem por ter natal* (Not even for having Christmas)
Substit + BERT	*Trabalhadores da hotelaria e turismo há quase dois meses sem salários* (Workers from hotels and tourism have been without salary for two months)	*Mãe só há uma* (There is only one mother)	*Semana só há uma* (There is only one week)

Table 3: Examples of produced texts, along with their adaptation and selection methods.

sess social reactions to the produced text. The proposed methods could be integrated in a chatbot, as a possible conversational aid, or as a creativity booster, in areas like journalism. In the meantime, TECo is working as a Twitter bot, *@TextoEmContexto*[4], that regularly reads the headlines of Portuguese newspapers and posts resulting text, as the example in Figure 1.

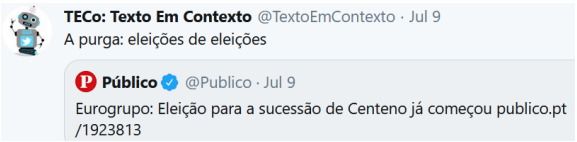

Figure 1: Example of a tweet posted by the Twitter bot.

Acknowledgements: This work was partially supported by FCT's INCoDe 2030 initiative, in the scope of the demonstration project AIA, "Apoio Inteligente a Empreendedores (*chatbots*)".

References

Ahn, Y.; Lee, H.; Jeon, H.; Ha, S.; and Lee, S.-g. 2016. Quote recommendation for dialogs and writings. In *CBRecSys@ RecSys*, 39–42.

Alnajjar, K.; Leppänen, L.; and Toivonen, H. 2019. No time like the present: Methods for generating colourful and factual multilingual news headlines. In *Procs of 10th ICCC*, 258–265. ACC.

Bay, B.; Bodily, P.; and Ventura, D. 2017. Text transformation via constraints and word embedding. In *Procs. of 8th ICCC*, ICCC 2017, 49–56.

Chrismartin, B., and Manurung, R. 2015. A chart generation system for topical meaningful metrical poetry. In *Procs. of 6th ICCC*, ICCC 2015, 308–314.

Colton, S.; Goodwin, J.; and Veale, T. 2012. Full FACE poetry generation. In *Procs. 3rd ICCC*, ICCC 2012, 95–102.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Procs. NAACL 2019*, NAACL-HLT 2019, 4171–4186. ACL.

Ferreira, J.; Gonçalo Oliveira, H.; and Rodrigues, R. 2019. Improving NLTK for processing Portuguese. In *Procs. SLATE 2019*, volume 74 of *OASIcs*, 18:1–18:9. Schloss Dagstuhl.

Gatti, L.; Özbal, G.; Guerini, M.; Stock, O.; and Strapparava, C. 2015. Slogans are not forever: Adapting linguistic expressions to the news. In *Procs 24th IJCAI*, IJCAI 2015, 2452–2458. AAAI Press.

Gonçalo Oliveira, H. 2017. O Poeta Artificial 2.0: Increasing meaningfulness in a poetry generation Twitter bot. In *Procs. of the INLG Workshop CC-NLG*, 11–20. Santiago de Compostela, Spain: ACL.

Hartmann, N. S.; Fonseca, E. R.; Shulby, C. D.; Treviso, M. V.; Rodrigues, J. S.; and Aluísio, S. M. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proc. 11th STIL*, STIL 2017.

Mendes, R., and Gonçalo Oliveira, H. 2020. Comparing different methods for assigning portuguese proverbs to news headlines. In *Procs. of 11th ICCC*, ICCC 2020, This volume.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Procs. of NAACL-HLT*, NAACL 2013, 746–751. ACL.

Ranchhod, E.; Mota, C.; and Baptista, J. 1999. A computational lexicon of Portuguese for automatic text parsing. In *Procs. SIGLEX99 Workshop: Standardizing Lexical Resources*. ACL.

Rocha, P. A., and Santos, D. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *V PROPOR 2000*, 131–140. São Paulo: ICMC/USP.

Veale, T.; Chen, H.; and Li, G. 2017. I read the news today, oh boy. In *Procs. of Intl. Conf. on DAPI*, 696–709. Springer.

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