Suitable Doesn't Mean Attractive. Human-Based Evaluation of Automatically Generated Headlines

Michele Cafagna[1,3], **Lorenzo De Mattei**[1,2,3], **Davide Bacciu**[1] **and Malvina Nissim**[3]

[1]Department of Computer Science, University of Pisa, Italy
[2]ItaliaNLP Lab, ILC-CNR, Pisa, Italy
[3]CLCG, University of Groningen, The Netherlands
{m.cafagna,m.nissim}@rug.nl, {lorenzo.demattei,bacciu}@di.unipi.it

Abstract

We train three different models to generate newspaper headlines from a portion of the corresponding article. The articles are obtained from two mainstream Italian newspapers. In order to assess the models' performance, we set up a human-based evaluation where 30 different native speakers expressed their judgment over a variety of aspects. The outcome shows that (i) pointer networks perform better than standard sequence to sequence models, creating mostly correct and appropriate titles; (ii) the suitability of a headline to its article for pointer networks is on par or better than the gold headline; (iii) gold headlines are still by far more inviting than generated headlines to read the whole article, highlighting the contrast between human creativity and content appropriateness.

1 Introduction and Background

Progress in language generation has made it really hard to tell if a text is written by a human or is machine-generated. The recently developed GPT-2 transformer-based language model (Radford et al., 2019), when prompted with an arbitrary input, is able to generate synthetic texts which are impressively human-like. But what makes generated text *good* text?

We investigate this question in the context of automatically generated news headlines.[1]

Headlines could be seen as very short summaries, so that one could use evaluation methods typical of summarisation (Gatt and Krahmer, 2018), but they are in fact a very special kind of summaries. In addition to being suitable in terms of content, newspaper titles must also be inviting towards reading the whole article. A model that, given an article, learns how to generate its title must then be able to cover both the summarisation as well as the luring aspect.

We collect articles from Italian newspapers online, and generate their headlines automatically. In contrast to the feature-rich approach of Colmenares et al. (2015), which requires substantial linguistic preprocessing for feature extraction, we rely on recent developments in language modelling, and train three different sequence-to-sequence models that learn to generate a headline given (a portion of) its article. We compare these generated headlines to one another and to the gold headline through a series of human-based evaluations which take several aspects into account, ranging from grammatical correctness to attractiveness towards reading the full article. The factors we measure are in line with the requirements for human-based evaluation mentioned by Gatt and Krahmer (2018), and are useful since it is known that standard metrics based on lexical overlap are not accurate indicators for the goodness of generated text (Liu et al., 2016).

Contributions We offer three main contributions: (i) a model which generates headlines from Italian news articles and which we make publicly available; (ii) a framework for human-based evaluation of generated headlines, which can serve as a blueprint for the evaluation of other types of generated texts; (iii) insights on the performance of different headline generators, and on the distinction between the concepts of suitable and attractive when evaluating headlines.

[1]A growing interest in headline generation is witnessed also in the organisation of a multilingual shared task at RANLP 2019, using Wikipedia data: http://multiling.iit.demokritos.gr/pages/view/1651/task-headline-generation

model	example generated headlines
s2s	Al Qaida : " L' Europa non è un pericolo per i nostri fratelli " la Samp batte la Sampdoria e la Samp non si ferma mai
pn	Teramo , bimbo di sei anni muore sotto gli occhi dei genitori mentre faceva il bagno Brescia , boa constrictor : sequestrati due metri e mezzo in un anno di animali
pnc	Argentina , Obama : " Paladino dei poveri e dei piu vulnerabili " . E il Papa si divide Cagliari , cane ha preferito rimandare il cane dal veterinario di Santa Margherita di famiglia

Table 1: Examples of headlines generated by the three models.

2 Task, Data, and Settings

The task is conceptually straightforward: given an article, generate its headline. Luckily, correspondingly straightforward is obtaining training and test data. We scraped the websites of two major Italian newspapers, namely *La Repubblica*[2] and *Il Giornale*[3], collecting a total of approximately 275,000 article-headline pairs. The two newspapers are not equally represented, with *Il Giornale* covering 70% of the data.

After removing some duplicates, and instances featuring headlines shorter than 20 characters (which are typically commercials), we were left with a total of 253,543 pairs, which we split into training (177,480), validation (50,709), and test (25,354) sets, preserving in each the proportion of the two newspapers.

We used the training and validation sets to develop three different models that learn to generate a headline given an article. To keep training computationally manageable, each article was truncated after the first 500 tokens.[4] As an alternative to keep the text short but maximally informative, we also experimented with selecting relevant portions of the articles using the TextRank algorithm, a graph-model that ranks sentences in a text according to their importance (Mihalcea and Tarau, 2004). However, preliminary experiments on our validation set did not seem to yield better results than just selecting the first N-tokens of an article. Also, using TextRank would make a less natural comparison to the settings used for the human evaluation (see Section 4), so we did not pursue this option further.[5]

[2] https://www.repubblica.it

[3] http://www.ilgiornale.it

[4] We do not control for sentence endings, so the last sentence of each truncated article might get truncated.

[5] Each article is also equipped with a short summary, often complementary to the title in content. We do not use this text in the current experiments, but plan to exploit it in future work.

3 Models

The models that we trained and evaluated are described below. In Table 1 we show two generated examples for each of the three models to give an idea of their output.

Sequence-to-Sequence with Attention (S2S) We used a sequence-to-sequence model (Sutskever et al., 2014) with attention (Bahdanau et al., 2014) with the configuration used by See et al. (2017) but we used a bidirectional instead of a unidirectional layer. This choice applies to all the models we used. The final configuration is 1 bidirectional encoder-decoder layer with 256 LSTM cells each, no dropout and shared embeddings with size 128; the model is optimised with Adagrad with learning rate 0.15 and gradient clipped (Mikolov, 2012) to a maximum magnitude of 2. We experimented also with a version using pretrained Italian embeddings, but since some preliminary evaluation didn't show better results, we eventually decided not to use this other model.

Pointer Generator Network (PN) The hybrid pointer-generator network architecture See et al. (2017) can copy words from the source text via a *pointing mechanism*, and generate words from a fixed vocabulary. This allows for a better handling of out-of-vocabulary words, providing accurate reproduction information, while retaining the ability to reproduce novel words. The base architecture is a sequence-to-sequence model, except for the pointing mechanism and for the fact that the copy attention parameters are shared with the regular attention. An additional layer (so called *bridge* (Klein et al., 2017)) is trained between the encoder and the decoder and is fed with the latest encoder states. Its purpose is to learn to generate

initial states for the decoder instead of initialising them directly with the latest encoder states.

Pointer Generator Network with Coverage (PNC) This model is basically a Pointer Generator Network with an additional coverage attention mechanism that is intended to overcome the copying problem typical of sequence-to-sequence models (See et al., 2017). This is basically a vector, computed by summing up all the attention distributions over all previous decoder timesteps. This unnormalised distribution over the document words is expected to represent the degree of coverage that the words have received from the attention mechanism until then. This vector, called *coverage vector*, is used to penalise the attention over already generated words, to minimise the risk of generating repetitive text.

4 Evaluation

Evaluating automatically generated text is non-trivial. Given that many different generated texts can be correct, existing measures are usually deemed insufficient (Liu et al., 2016). The problem is even more acute for headline generation, since due to their nature and function, simple content evaluation based on word overlap is most likely not exhaustive. Human-based evaluation could provide a richer picture.

When discussing human-based (intrinsic) evaluation of summarisation models, Gatt & Krahmer (2018) mention two core aspects: *linguistic fluency or correctness*, and *adequacy or correctness relative to the input*, in terms of the system's rendition of the content. These also relate to the aspects examined in the context of evaluating the generation of the final sentence of a story, such as *grammaticality*, *(logical) consistency*, and *context relevance* (Li et al., 2018).

We took these factors into consideration when designing our evaluation settings. Since headlines must also carry some "attraction" factor to read the whole article, we included this aspect as well.

4.1 Settings

We call a case each set of an article and its four corresponding headlines to be evaluated, namely the three automatically generated ones, and the original (gold) title.

We prepared an evaluation form[6], which included five different questions for each case (see Figure 1). Each subject could see the four headlines and answer questions Q1–Q3. The corresponding article, in the truncated form that was also seen in training by the models, was only shown to the subjects after Q3, and they would then answer Q4–Q5. This choice was made in order to ensure that first questions were answered on the basis of the headlines only, especially for the validity of Q3. The order in which gold and generated titles were shown was randomised, though it was the same for each case for all participants.

Each form comprised 20 cases to evaluate, and was sent to 3 participants. We created 10 different forms, thus obtaining judgements for 200 total cases with 30 different participants (600 separate judgements). The participants are all native speakers of Italian, and balanced for gender (15F/15M). We also aimed at a wide range of ages (17–77) and education levels (middle school diploma to PhD). This variety was sought in order to prevent as much as possible judgements that are based too strongly on personal biases, taste, and familiarity with specific topics over others.

The headlines used for this evaluation exercise were randomly selected from the test set. When extracting them though, we excluded all cases where at least one model produced a headline containing at least an unknown word (represented with the special token $< UNK >$), since this would make the headline look too weird and not much comprehensible. This led to excluding approximately 50% of the samples. The model with the highest proportion of headlines with at least one UNK was the S2S (37%), followed by the PNC (31%), and the PN (30.2%). In terms of topics, random picking ensured a variety of topics; manual inspection anyway showed that most news were mainly about chronicle facts, and international politics.

4.2 Analysis

We discuss the results in detail for questions Q1, Q3, Q4, Q5. For Q2, we simply note that the most similar in content are always the two pointer networks, and the most dissimilar are all three pairs that involve the gold headlines. This suggests that human titles focus on aspects of the article that are different from those picked by the generator, most likely as humans can abstract away from the actual text and use much more creativity.

[6] An example can be found here: `https://forms.gle/MB31uEGT856af2MP7`

The four titles are shown (repeated for each question below)

A. Usa , la fabbrica del vetro d' aria per il telefono d' aria in Usa
B. Se il lavoro va ai robot : un automa vale sei operai
C. Usa , Trump : " Trump si difende l' occupazione e l' economia nazionale "
D. Usa , la beffa del condizionatore d' aria " made in Usa " : " Ecco come si difende "

And the following questions are then asked:

[at this stage the subjects only see titles, without the article]	
Q1. Questi titoli sono scritti correttamente?	yes,no for each
Q2. Secondo te, questi titoli parlano dello stesso articolo?	yes,no for pairs of titles
Q3. Quale di questi titoli ti invoglia maggiormente a leggere l'intero articolo?	pick one
[now the subjects also see the (truncated) article]	
New York . Chiamiamola la beffa del condizionatore d' aria " made in Usa " . La marca è Carrier , filiale della multinazionale United Technologies . Un caso ormai celebre , che Donald Trump addita come un esempio della sua azione efficace a tutela della classe operaia . A novembre , appena eletto presidente (ma non ancora in carica) , Trump si occupa dello " scandalo Carrier " : vogliono chiudere una fabbrica di condizionatori a Indianapolis per trasferirla in Messico , delocalizzando a Sud del confine 800 posti di lavoro . Il presidente - eletto fa fuoco e fiamme , chiama il chief executive dell' azienda . Forse interviene la casa madre , United Technologies , che ha grosse commesse per l' esercito e non vuole inimicarsi il neo - presidente . Sta di fatto che Carrier cede alle pressioni , fa dietrofront : la fabbrica resta sul suolo Usa , nello Stato dell' Indiana . Tripudio di Trump che canta vittoria via Twitter : " Ecco come si difende l' occupazione e l' economia nazionale " . Passano i mesi e il caso viene dimenticato . Fino a quando il chief executive Greg Hayes rivela ai sindacati che i 16 milioni di investimento nella sede di Indianapolis vanno tutti in robotica , automazione : " Alla fine ci saranno meno posti di prima . Dobbiamo ridurre i costi , per essere competitivi " . La morale è crudele , la vittoria di Trump si [. . .]	
Q4. Ritieni che il titolo sia appropriato all'articolo?	yes,no for each
Q5. Quale ti sembra più adatto? Ordinali	rank 1–4

Figure 1: Sample evaluation case. Subjects are presented with the gold and generated headlines in random order, and must answer a progression of questions, without and with seeing the article. Q1 targets correctness, Q2 targets the similarity in topic focus, Q3 targets attractiveness, Q4 and Q5 target appropriateness (absolute, and relative to one another). In this example, A=s2s, B=gold, C=pnc, D=pn.

Grammatical Correctness (Q1) When asked to evaluate whether the headlines were written correctly, the participants assessed all headlines as correct more frequently than not correct, with Gold and PN having the best ratio of yes vs no (Figure 2). What is, however, interesting is that even Gold headlines are frequently judged as not correct, implying that either the participants were very strict, or correctness is not a necessary or particularly typical feature of newspaper headlines. While it is important for us to assess how well the generators perform also in terms of well-formed sequences, if (grammatical) correctness is not strictly a property of newspaper headlines, this evaluation question might have to be formulated differently. In any case, among the models, for the current question, the PN behaves almost on par with the gold headlines.

Attractiveness (Q3) In the large majority of the cases, the gold headline was chosen as the most inspiring for reading the whole article (Figure 3). Among the models, the headlines generated by the PN is mostly chosen, followed by the PNC, and lastly by the S2S. Such results suggest that there is something in the way experts create headlines, most likely related to human creativity, rhetoric and communication strategies, which systems are

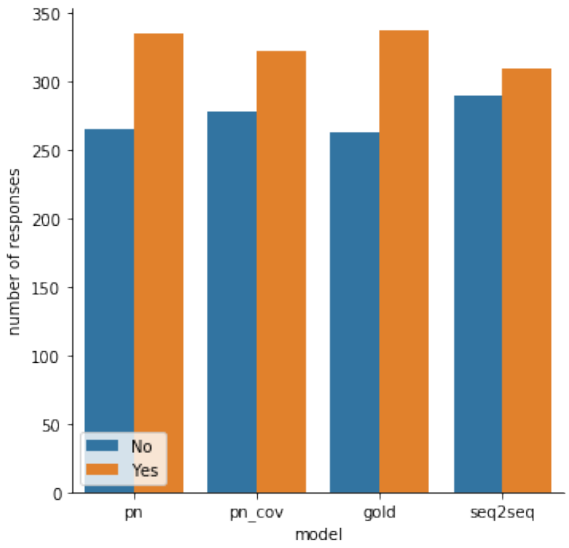

Figure 2: Correctness judgments (Q1)

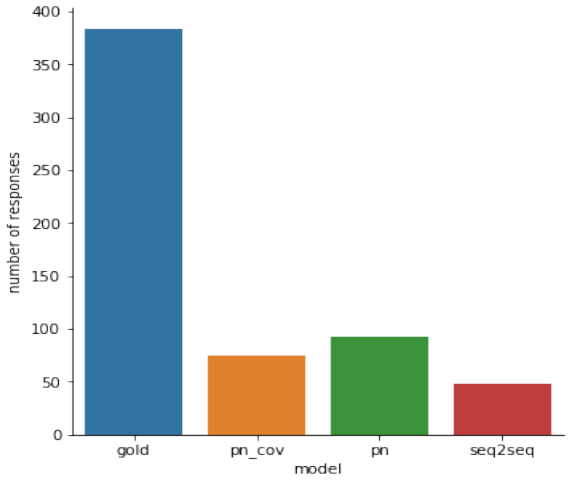

Figure 3: Attractiveness judgements (Q3)

not yet able to reproduce. Additionally, some online newspapers' business models can be heavily clickbait-based, causing headlines to be more sensational than faithful to the article's actual contents.

Suitability (Q4-Q5) There are two results to be analysed in the context of assessing how appropriate a headline is with respect to its article. In terms of a binary evaluation for each headline (Figure 4, left), in all cases, including gold, the headline is deemed not appropriate more than the times is deemed appropriate. In the case of gold, this could be due to the fact that excessive creativity to make the title attractive can make it less adherent to the actual content. In the case of the generated headlines, they might just not be good enough.

	G	S2S	PN	PNC	tot
correctness	0.439	0.427	0.345	0.337	0.387
attractiveness	–	–	–	–	0.120
suitability	0.349	0.354	0.374	0.313	0.348
suitability-rank	0.444	0.364	0.339	0.398	0.389

Table 2: Krippendorf's alpha scores for the human annotations. The rightmost column shows the agreement over all systems plus gold headlines.

The rank shows a possibly unexpected trend (Figure 4, right side). The headline chosen as most appropriate (ranked 1st) is most of the times the one produced by the PN model, even more so than the gold. Not only, the gold is also the headline that features last (ranked 4th, thus least suitable) more than any of the other titles. This is reflected in the average rank (see caption of Figure 4), as the gold headline comes in last, and the PN-generated title is comparatively the most preferred.

4.3 Agreement

Given that we obtained three separate judgments per case, in addition to the separate evaluations, we can also assess how much the subjects agree with one another. Table 2 shows the values for Krippendorf's alpha over all of the annotated aspects. Low scores suggest that the task is highly subjective, and this is especially true for the evaluation of how attractive a headline is towards reading the whole article. Possibly surprising is the score regarding the evaluation of the headline's correctness, which could be viewed as a more objective feature to assess. Such relatively low score could be due to the vagueness of Q1, in combination with the nature of headlines, which even in their human version might be formulated in ways that do not necessarily abide to grammatical rules.

5 Conclusions

The quality of three different sequence-to-sequence models that generate headlines starting from an article was comparatively assessed through human judgement, which we contextually used to evaluate the original headlines as well. The best system is a pointer network model, with correctness judgements on par with the gold headlines. Evaluating the generated output on different levels, especially attractiveness, which typically characterises news headlines, uncovered an interesting aspect: gold headlines appear to be the most attractive to read the whole article, but are not con-

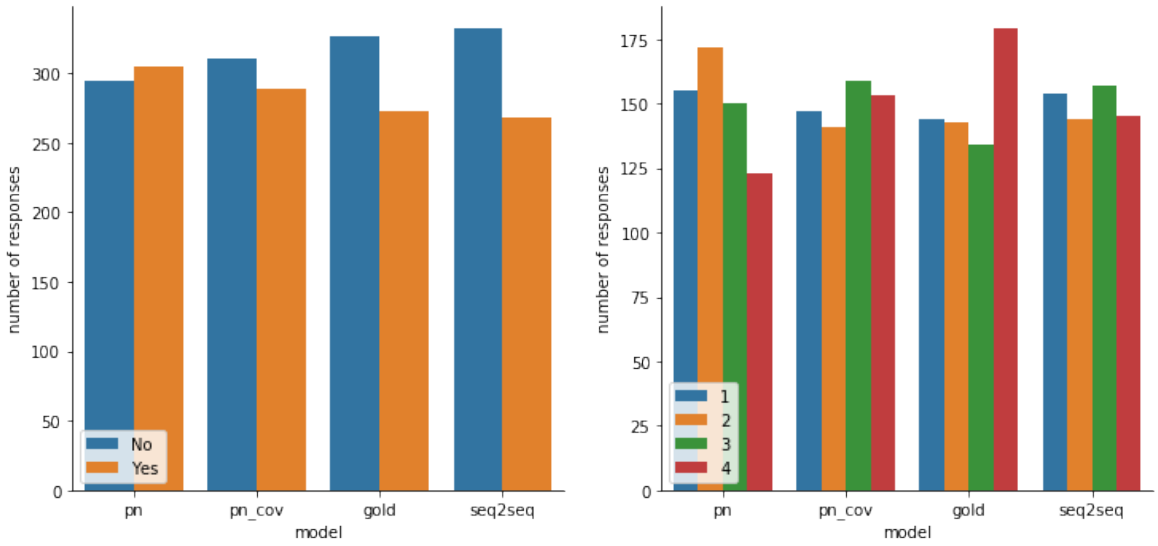

Figure 4: Suitability. Left: suitability judgment for each headline (yes/no). Right: headlines are ranked according to most (1) to least (4) appropriate for each corresponding article. Average ranking: PN=2.401; Seq2Seq=2.488; PN_C=2.530; GOLD=2.580

sidered the most suitable, on the contrary, they are judged as the most unsuitable of all. Therefore, when automatically generating headlines, just relying on content might never lead us to titles that are human-like and attractive enough for people to read the article. This should be considered in any future work on news headline generation. At the evaluation stage, it would also be beneficial to involve professional journalists. A first contact with one of the newspapers at the early stages of our evaluation experiments did not yet yield any concrete collaboration, but expert judgement on the quality of the generated headlines is something we would like to include in the future.

One aspect that we have not explicitly considered in our experiments is that the headlines come from different newspapers (positioned at opposite ends of the political spectrum), and can carry newspaper-specific characteristics. Robust headline generation should consider this, too.

Acknowledgments

We are deeply grateful to all of the participants to our evaluation. We also would like to thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine high performance computing cluster. A heartfelt thank you also to Angelo Basile, with whom we discussed both theoretical and implementation aspects of this work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Carlos A Colmenares, Marina Litvak, Amin Mantrach, and Fabrizio Silvestri. 2015. HEADS: Headline generation as sequence prediction using an abstract feature-rich space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 133–142.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Generating reasonable and diversified story ending using sequence to sequence model with adversarial training. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1033–1043.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Tomáš Mikolov. 2012. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.