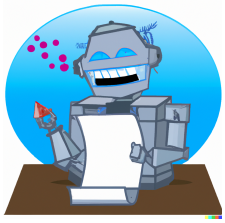
Transformers Go for the LOLs: Generating (Humourous) Titles from Scientific Abstracts End-to-End

Yanran Chen, Steffen Eger
Natural Language Learning Group (NLLG)
University of Mannheim, Germany
yanran.chen@stud.tu-darmstadt.de
steffen.eger@uni-mannheim.de



Abstract

We consider the end-to-end abstract-to-title generation problem, exploring seven recent transformer based models (including ChatGPT) fine-tuned on more than 30k abstract-title pairs from NLP and machine learning (ML) venues. As an extension, we also consider the harder problem of generating humorous paper titles. For the latter, we compile the first large-scale humor annotated dataset for scientific papers in the NLP/ML domains, comprising ∼2.6k titles. We evaluate all models using human and automatic metrics. Our human evaluation suggests that our best end-to-end system performs similarly to human authors (but arguably slightly worse). Generating funny titles is more difficult, however, and our automatic systems clearly underperform relative to humans and often learn dataset artefacts of humor. Finally, ChatGPT, without any fine-tuning, performs on the level of our best fine-tuned system.[1]

1 Introduction

Computer-assisted writing is an important and long-standing use case of NLP and natural language generation (NLG) (Burns, 1979), e.g., via and beyond tools such as spell checkers or grammatical error correction. The recent success of large-scale language models (LLMs), such as the GPT generation of NLG models, has made the goal even more realistic and promises full-scale automatic text generation, without any human intervention.

In this work, we concern ourselves with automatic text generation in the scientific domain. Sample scenarios in this general context involve (semi-)automatically generating reviews for scientific papers (Yuan et al., 2022), e.g., as a response to high reviewing load in the face of exploding submission numbers; and generating captions for tables that require reasoning capabilities (Moosavi et al., 2021). Our goal is much more modest: we ask whether language models can generate adequate titles given a human authored abstract as input; we refer to this task as **A2T** (abstract-to-title generation). Title generation is important as titles are the first access points to papers; a good title may attract more readers and consequently increase paper impact, e.g., in terms of citation numbers (Falagas et al., 2013). Besides generating titles per-se, we also aim for generating *humorous* titles, an inherently difficult problem due to small sample size and the vagueness of humor. Generating funny titles may be relevant as a funny title may attract more readers: for example, Heard et al. (2022) find that funny titles have significantly higher citation rates.

We approach the problem as a standard sequence-to-sequence text generation problem, where we fine-tune LLMs on more than 30k abstract-title pairs from ML and NLP. Our contributions:

- **(i)** We provide the first publicly available humor annotated dataset for scientific titles in the NLP and ML domain, with 2,638 humor annotated titles annotated by 2 annotators with decent levels of agreement (kappa ∼0.65).
- **(ii)** We explore 6 recent popular text generation systems on the A2T task, finding one to be competitive to human titles, according to automatic and human evaluation involving 15 annotators.
- **(iii)** We analyze the problem and find that the A2T task is to some degree ill-posed as a good title may leverage more than the abstract alone (we argue that the problem framing is still a legitimate and efficient approximation).
- **(iv)** For humor generation, we find that our models clearly underperform relative to humans and instead often learn dataset artefacts.
- **(v)** We finally analyze ChatGPT on a small scale and find that it may be competitive to (albeit

[1] Our paper title is a (modified) merge of a funny and unfunny title suggested by ChatGPT (chat.openai.com). Our paper logo is drawn by DALL-E (https://openai.com/dall-e-2/).
Data+code: https://github.com/cyr19/A2T

Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, pages 62–84

slightly weaker than) our best fine-tuned model without any task-specific fine-tuning at all.

2 Related Work

Title generation and evaluation Mishra et al. (2021) perform A2T with pre-trained GPT-2 fine-tuned on arxiv papers and subsequent (rule-based) modules of title selection and refinement. We compare many more text generation models for the task, use better evaluation (including more comprehensive human and automatic evaluation), do not make use of rule-based selection and also consider humor in title generation. Putra and Khodra (2017) classify sentences from paper abstracts into rhetorical categories, retain those relating to methods and results and then generate titles using templates. They further note the relationship between the task of summarization (Nenkova et al., 2011) and A2T, as a title can be seen as a summary of the research paper. We also leverage the relationship to summarization by considering pre-trained models fine-tuned on summarization datasets. In contrast to Putra and Khodra (2017) and Mishra et al. (2021), we only consider end-to-end models that do not involve pipelines. While refinement steps could be further helpful (but also error-prone), they additionally require potentially undesirable human intervention (Belouadi and Eger, 2023). Related to the task of title generation is the task of headline generation e.g. for news. Tan et al. (2017) use a coarse-to-fine approach which first identifies important sentences and then converts them into a headline. In this way, the model is not confused by 'too much' irrelevant information. In A2T, the first summarization step may not be necessary, as the abstract is already a summary of the scientific paper.

How titles should be (and are) structured has been researched for a long time, e.g., (Lewison and Hartley, 2005). Hartley (2008) gives a typology of title types, distinguishing 13 title classes, e.g., those that state results vs. methods.

Beyond title generation, related fields of text generation for science are related work generation (Li et al., 2022), more general automatic paper section writing assistance (Wang et al., 2019b), and automatically generating reviews for scientific articles (Yuan et al., 2022). More broadly relating to science, Meta has in 2022 released an LLM for the scientific domain called Galactica (Taylor et al., 2022), but they mostly explore it for scientific classification tasks rather than generation.

Humor identification and generation Humor detection is a niche area in NLP but nonetheless with a rich history. For example, Mihalcea and Strapparava (2006) distinguish funny from non-funny sentences (heuristically scraped from the Web) using features and traditional classifiers. Simpson et al. (2019) focus on efficiently annotating humor and inducing classifiers from crowdsourced data. Recently, Peyrard et al. (2021) show that transformers are strong at distinguishing funny from non-funny sentences on minimal pairs of satirical news headlines. In the scientific domain, Heard et al. (2022) annotate a dataset of more than 2k titles from ecology using a fine-grained Likert scale. The majority were labeled as non-funny and annotators exhibited low agreements. Shani et al. (2021) classify scientific titles as funny or not using humor-theory inspired features and scientific language models such as SciBERT (Beltagy et al., 2019) building on a dataset of Ig Nobel winners and humorous papers discussed in online forums.

There is considerably less work on humor generation. As one exception, He et al. (2019) generate puns by a retrieve-and-edit approach based on word2vec, thus circumventing the problem of little training data for puns.

3 Data

We use the dataset released by Beese et al. (2023), which contains title-abstract pairs and corresponding meta-information such as the publication year and venue. Beese et al. (2023) extracted the data from two sources: ACL Anthology (from 1984 to 2021) and machine learning conferences (from 1989 to 2021); we refer to the datasets from these two sources as `NLP` and `ML`, respectively. After filtering (described in Appendix A), **32,952** abstract-title pairs remain in our dataset.

4 Title Generation

We first explore whether existing state-of-the-art Seq2Seq models manage to generate human-level titles from abstracts. Hence, we do not include humor constraints. We use an 8:2 ratio to divide the data into train and test sets, and randomly select 1,000 instances from the train set for the dev set.

4.1 Models

We experiment with the following six generation models: (i) BART base ($\texttt{BART}_{\texttt{base}}$) (Lewis et al., 2020), (ii) GPT2 (`GPT2`) (Radford et al., 2019),

(iii) T5 small (Raffel et al., 2020) (T5), and (iv) PEGASUS large (Zhang et al., 2019) finetuned on Extreme Summarization (XSUM) dataset (Narayan et al., 2018) ($\texttt{PEGASUS}_{\texttt{xsum}}$). Noting the similarity between text summarization and our A2T generation task, we additionally inspect two BART large models finetuned on (v) XSUM ($\texttt{BART}_{\texttt{xsum}}$) and (vi) CNN dailymail (CNNDM) (See et al., 2017) ($\texttt{BART}_{\texttt{cnn}}$), respectively. XSUM and CNNDM contain document-summary pairs, where XSUM has one-sentence summaries, while each summary in CNNDM consists of multiple sentences.

Fine-tuning For all baseline models, we continue fine-tuning them on the abstract-title pairs from our dataset. Details are in Appendix B.

4.2 Evaluation

We assess the performance of the systems on 230 abstracts using both automatic evaluation metrics and human evaluation. We also include the human-generated titles in the evaluation, denoted as 'HUMAN'. While our test set is small, we note that (i) human evaluation is very time-consuming and (ii) we have more source-output pairs (i.e., 230×6, see below) than in some standard MT or summarization evaluation benchmarks such as WMT15-17 or SummEval (Fabbri et al., 2020).

Automatic Evaluation: As there are no A2T task-specific evaluation metrics, we use the following metrics from other NLG tasks: Rouge (Lin, 2004), BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), COMET (Rei et al., 2020), BARTScore (Yuan et al., 2021), MENLI (Chen and Eger, 2022). COMET is a metric supervised on human scores from MT, all others are unsupervised. We employ all metrics in both *reference-based* and *-free* settings. Reference-based, the metrics compare the system titles with the original human-generated titles, while reference-free, the system titles are directly compared to the abstracts. The details of the metric variants can be found in Appendix C. The reference-free setup is more consistent with our human evaluation below and overall more plausible for A2T.

Human Evaluation: The human evaluation is conducted reference-free: 15 annotators[2] were asked to select two best and two worst titles among six titles from different systems (including HUMAN), given the abstract. In order to make the annotation simpler for humans, we only considered one dimension of annotation, namely, 'overall quality', which may comprise aspects such as fluency, (grammatical) correctness, adequacy, etc. This mimics coarse-grained annotations such as direct assessment (DA) in fields like MT. We did not further subdivide the quality into more fine-grained subcategories, as the annotation is already difficult and comprises to understand a scientific abstract and to decide which title best fits it. Each instance (an abstract and its six titles) was evaluated by at least two annotators; depending on availability, some instances were annotated by up to five annotators. The average percentage agreement over all annotator pairs is ∼50%, implying that each two annotators agree on one selection among the two selected best/worst titles, on average.

Then, we use best-worst scaling (**BWS**) (Louviere and Woodworth, 1991) to obtain the final human score for each title as:

$$BWS = \frac{N_{best} - N_{worst}}{N_{annotators}} \quad (1)$$

where $N_{best/worst}$ refers to the number of times that the title was selected as one of the best/worst two titles and $N_{annotators}$ indicates the number of annotators responsible for that instance.

system	BWS	MoverS	BERTS	BARTS	COMET	MENLI	ROUGE
$\texttt{BART}_{\texttt{xsum}}$	**0.197**	-0.025	**0.889**	**-2.583**	**0.060**	-0.214	0.033
$\texttt{PEGASUS}_{\texttt{xsum}}$	0.022	-0.036	0.887	-2.819	0.060	-0.263	0.035
$\texttt{BART}_{\texttt{base}}$	0.015	-0.034	0.887	-2.709	0.059	-0.226	0.035
GPT2	-0.013	-0.087	0.881	-3.090	0.060	-0.285	0.020
T5	-0.039	-0.055	0.889	-2.735	0.057	-0.265	0.032
$\texttt{BART}_{\texttt{cnn}}$	-0.384	**0.046**	0.880	-2.982	0.047	**-0.159**	**0.055**
HUMAN	0.181	-0.062	0.873	-3.508	0.061	-0.029	0.029

Table 1: Ref-free evaluation results of the baseline models. We underlie the best performance among all generation systems including human. We bold the best performance among all automatic generation systems excluding human.

Results We present the **reference-based evaluation** results in Appendix D. *Among the six systems,* $\texttt{BART}_{\texttt{xsum}}$ *is best*, being selected by 4 out of 6 evaluation metrics, followed by $\texttt{BART}_{\texttt{cnn}}$.

Table 1 shows the **reference-free evaluation** results. Unlike in reference-based evaluation, only two evaluation metrics (COMET and MENLI) select HUMAN as the best system. $\texttt{BART}_{\texttt{xsum}}$ is still the best among the six automatic systems, obtaining best results on 4 out of 7 evaluation metrics (including BWS). Surprisingly, it outperforms HUMAN

[2]Most annotators are Master students, with an additional senior researcher and two Bachelor students.

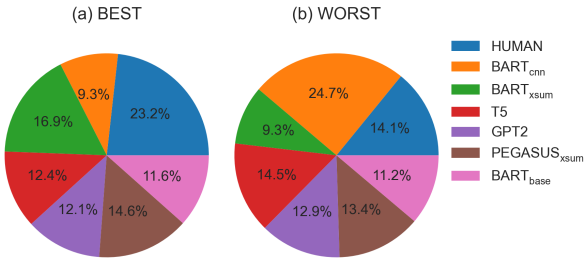


Figure 1: Distribution of generation systems of the titles selected as the *BEST*/*WORST* ones in human evaluation; percentages indicate the proportion of the generation systems being selected over all selections.

	230 instances ref-based		230 instances ref-free		35 instances ref-free	
	ρ	r	ρ	r	ρ	r
ROUGE	0.571	0.395	-0.250	-0.722	-0.121±0.11	-0.404±0.26
BARTS	0.393	0.389	0.214	-0.044	0.200±0.30	0.083±0.21
BERTS	0.571	0.442	0.250	0.079	0.236±0.26	0.296±0.22
MoverS	0.929	0.575	-0.071	-0.677	-0.129±0.13	-0.378±0.24
MENLI	0.357	0.345	0.321	0.139	0.057±0.15	0.160±0.21
COMET	**0.964**	**0.580**	**0.929**	**0.929**	0.414±0.32	0.679±0.15
A2TMetric	-	-	-	-	**0.707±0.17**	**0.726±0.16**

Table 2: Pearson's r and Spearman's ρ of evaluation metrics with **system-level** human judgements for all **230 instances** (1380 titles; left block) and **35 instances** (210 titles; right block). The correlations on the 35 instances are averaged over the test sets from five splits. We bold the highest correlation in each block.

even in the human evaluation (0.197 vs. 0.181 BWS). Nevertheless, as Figure 1(a) shows, HUMAN was still most frequently selected as among the two best titles (23.2%) among all generation systems, whereas the best neural generation system BART$_{xsum}$ was selected in 16.9% of the cases as one of the best two titles. However, Figure 1(b) shows that HUMAN was also more often selected as among the two worst titles (14.1% vs. 9.3% BART$_{xsum}$), explaining why BART$_{xsum}$ is better than HUMAN in human evaluation. Introspection shows that this is mostly due to words in the title which do not appear in the abstract. As a consequence, human annotators may believe that the model is hallucinating. Overall, we thus believe that there is a (slight) mismatch in our task definition: human authors may leverage the whole paper when designing their titles, not only the abstracts. However, paper2title generation would not only be a challenge for the text generation models (which are often limited in text length) but also for the human annotation process. We argue that framing the problem as abstract2title generation is a simplification with overall good tradeoffs between problem complexity and model and annotator capacity.

Why is the best model best? To get a deeper insight into the quality of the system titles, we first analyze their lengths. BART$_{cnn}$ produces titles much longer than human titles (14.95 vs. 8.27 tokens) and other systems (6.68-9.13 tokens), on average; besides, its titles are often truncated due to the maximal output length set to the model. This reflects the mismatch of the training data—BART$_{cnn}$ was first trained on CNNDM which has multiple sentences as a summary. Among the other systems, BART$_{xsum}$ and BART$_{base}$ generate titles having the largest overlap with the abstracts, based on the edit distance. While BART$_{xsum}$ (best/worst: 241/133) does not have a huge advantage over BART$_{base}$ (best/worst: 165/159), inspection of results indicates that BART$_{xsum}$ may give more precise and relevant titles, e.g., it picks out the key information from the abstracts more frequently; some examples are in Appendix E. This may be again due to its (extreme) summarization objective in the pre-training phase.

4.3 Reliability of Evaluation Metrics

To inspect the reliability of the used metrics, we calculate Spearman/Pearson correlation with system-level human judgments, i.e., average BWS per system, on the 1380 titles (230 instances $\times$ 6 titles). From Table 2 (left block), we observe: (1) most metrics perform better in the ref-based setup than ref-free, except for COMET. (2) Only ref-free COMET correlates well with human judgments from the perspective of both types of correlation.

Even though COMET performs well on system-level, this only indicates that COMET ranks systems similarly as humans. COMET is not necessarily good at selecting the best title among different choices (segment-level evaluation). Indeed, at segment-level, it correlates weakly with human scores (0.127 Kendall).[3] Inspired by this, we train a ref-free metric supervised on our own human scores.

4.4 A2TMetric

We develop the first **supervised A2T generation-specific evaluation metric**, using the human judgments collected in the evaluation for the 230 instances. Since HUMAN as a generation system is

[3] As we convert BWS to WMT relative ranking judgements (Ma et al., 2018), we use the Kendall-like formulation introduced there for segment-level correlation.

included in the evaluation, and the metrics will later be used to evaluate system-generated humorous titles, which may vastly differ from the original ones, we argue that a ref-free metric will better suit our needs.

Dataset We split the data of 230 instances to train (170 instances), dev (25 instances), and test (35 instances) set. To get more robust results, we generate five different splits of train, dev and test set and report the average performance of the metrics on the test set over the five splits in Table 2. We note that many titles receive a BWS of 0 when the number of annotators is small (because they were never selected as the best or worst two titles), which may be problematic when aiming to directly train a regression model. Besides, the human evaluation was similar to the ranking process. Therefore, we convert BWS in the train and dev set to relative-ranking judgments (Ma et al., 2018). That is, if two titles for one abstract obtain different BWS, this title pair is considered as one relative-ranking judgement. Each instance then contains one abstract, a "better" title, a "worse" title, and the score difference between the two titles in addition.

Framework We adopt a framework similar to the ranking-based variant of COMET to train the A2T metrics but in a ref-free setup. During training, the model optimizes the embedding space so that (1) the sentence embedding of the abstract (a) is closer to that of the "better" title (t^+) than to that of the "worse" title (t^-) (using the Triplet Margin loss (Schroff et al., 2015)) and (2) the difference between $d(a, t^+)$ and $d(a, t^-)$ is close to the difference in BWS human scores for the two titles (using the MSE loss), where $d(u, v)$ refers to the Euclidean distance between u and v. During predicting, the metrics calculate the Euclidean distance between the sentence embeddings of the abstract and the title.

Evaluation As Table 2 (right block) shows, our A2TMetric achieves the highest values of both average Spearman and Pearson correlations (above 0.71-0.73 vs. -0.40-0.68) and relatively low standard deviation (around 0.16 vs. 0.11-0.32), implying that it is not only superior to the existing metrics but also demonstrates comparably good robustness.

While the metric is still not of absolutely high quality segment level (0.276 Kendall), it clearly outperforms COMET and the other metrics (right half of Table 11 in the appendix) and the correlation values are on the same level as those of the best MT metrics in WMT22 shared Task (Freitag et al., 2022). System-level, we evaluate A2TMetric on 5 random samples of size 35 where the remainder instances are for train/dev. While there is a high variance due to small sample size, A2TMetric is on average 0.1-0.3 Pearson/Spearman better system-level than COMET (right block of Table 2). Even though comparing the trained A2TMetric to unsupervised metrics may seem unfair, this is exactly the key point: A2TMetric is better because it has been trained on our costly human data, which makes it valuable.

COMET is still the best among the existing metrics. Therefore, we only leverage our trained A2TMetric and COMET to automatically evaluate the A2T systems' quality in §5.1.

5 Humorous Title Generation

To generate humorous titles, we first need a dataset of humor annotated titles in our domain (NLP and ML papers). We cannot resort to the data of Shani et al. (2021); Heard et al. (2022) as those leverage papers from other scientific fields. As a consequence, we build our own dataset. When constructing the dataset, we ask annotators to rely on their intuition of humor rather than issuing guidelines of what they should find funny. This can be justified as humor is often subjective and culture- and even gender-specific (Dore, 2019; Mundorf et al., 1988). There is also a multitude of theories around humor, indicating the ambiguity of the concept.[4]

Humor Annotation + Classification We train humor classifiers on human annotated data to automatically label titles as *FUNNY*, $FUNNY_{\text{med}}$, and $\neg$*FUNNY* (examples see Table 12 the appendix). Two co-authors participated in the annotation. Examples of their annotations are shown in Appendix F. Titles annotated as funny by both annotators allude to famous proverbs or book/movie titles ("*Taming the wild*"), make use of linguistic devices such as alliteration ("*Balancing Between Bagging and Bumping*") or leverage surprise ("*Is the Best Better? [...]"; "What's in a name? In some languages, grammatical gender*"). Medium funny titles often make use of playful/clever abbreviations,

[4]The wikipedia page for humor https://en.wikipedia.org/wiki/Theories_of_humor lists at least three modern popular theories of humor, based on relief, superiority and incongruity.

	Individuals	Ensemble
Stage 1	52.2 / 81.5	54.1 / 85.1
Stage 2	55.1 / 84.7	**57.7 / 88.1**

Table 3: Average macro F1 over the 11 individual classifiers and macro F1 of the ensemble classifiers from both stages on the held-out test set (where the two annotators obtain 0.649 kappa agreement). Performance on both three-way (first entry) and binary (second entry) classification tasks; for *binary* classification, *FUNNY* and $FUNNY_{\text{med}}$ are merged. We bold the highest macro F1 on each classification task.

e.g., "*CPR: Classifier-Projection Regularization for Continual Learning*".

Stage 1: The two annotators initially annotated **1,730** titles: 1,603 titles as ¬*FUNNY*, 106 as $FUNNY_{\text{med}}$, and 21 as *FUNNY* (kappa 0.65 on 300 common instances). To combat this severe data imbalance, we resort to ensembling with each classifier trained on more balanced splits: we randomly generate 11 different data splits, where the train set of each split consists of 100 funny or medium funny titles and 200 not funny titles (all randomly drawn). On those splits, we train 11 classifiers to construct an ensemble classifier. To evaluate the classifier performance, the two annotators annotated another 315 titles jointly, obtaining 0.639 Kappa. Our best ensemble classifier leverages the sum of the label values assigned by the 11 individual classifiers to predict humorousness, yielding 4.8% macro F1 improvement compared to the individual classifiers (62.4% vs. 57.6%). Details are in Appendix G.

Stage 2: To find more funny title candidates to annotate, the two annotators annotated the funniest 396 titles in the original dataset from Beese et al. (2023), predicted by the Stage 1 ensemble classifier; 75.8% (300 titles) were judged as *FUNNY* or $FUNNY_{\text{med}}$, which is substantially higher than the proportion of funny titles in the annotated data of Stage 1 (7.3%). Thus, the annotated data expands to **2,441** titles ($= 1,730 + 315 + 396$), where 1,893 are labeled as ¬*FUNNY*, 492 as $FUNNY_{\text{med}}$ and 56 as *FUNNY*. Subsequently, we re-train 11 classifiers on newly generated 11 data splits from the expanded data of 2,441 titles; now the train set of each split has 400 (medium) funny titles and 800 not funny titles. As before, we ensemble the 11 classifiers as in Stage 1.

We test the classifiers from both stages on a held-out test set containing 197 titles annotated by the two annotators (0.649 kappa). The macro F1 scores of those classifiers are presented in Table 3. As *FUNNY* titles are rare in the whole dataset, we also evaluate the classifiers on the corresponding binary classification task, where *FUNNY* and $FUNNY_{\text{med}}$ are merged. We observe that: (1) ensemble classifier performs better than the individual ones. (2) Classifiers from Stage 2 are superior to the ones from Stage 1, indicating larger size of the training data is beneficial. (3) The best three-way classifier achieves only ∼58% macro F1, but ∼88% macro F1 on the binary classification. Besides, we see a consistent improvement of human annotation quality: the two annotators achieve 0.01-0.1 higher Kappa when their annotations are down-scaled to binary (see Table 17 in Appendix G). **Thus, we use the ensemble classifier from Stage 2 as the humor classifier in further experiments**.

Final Dataset We use our humor classifier to automatically label the rest of the data. Considering the difficulty of three-way classification for both humans and classifiers, we only consider two humor levels in further experiments: (1) *FUNNY* (for funny and medium funny titles) and (2) ¬*FUNNY* (for not funny titles). Thus, we collect 31,541 instances (>95%) with ¬*FUNNY* and 1,411 with *FUNNY* titles. We split the resulting data to train, dev, and test sets, ensuring that (1) the data with human-annotated titles remains in the train set, as the humor classifier trained and evaluated on it will be used as an automatic humor evaluator; (2) 80% of the data in dev/test is from `NLP` and 20% from `ML` because our annotators are more knowledgable for NLP papers, and (3) the ratio of *FUNNY* data to ¬*FUNNY* data in dev/test set is 1:2.[5] As *FUNNY* data is only a small portion of the whole data, we only keep 600 instances in the dev/test sets, the remaining data serves as the train data. Appendix H summarizes the statistics of the final dataset.

Generation In the second phase of the experiments, we use the optimal model identified previously, i.e., $\texttt{BART}_{\texttt{xsum}}$, to generate titles with constraints on humor level. The input of the generation systems is formulated as "*humor level [SEP] abstract*", where humor level is either 0 (for ¬*FUNNY*) or 1 (for *FUNNY*).

[5]This aims to more easily compare the system-generated funny titles with the human-generated ones and does not relate to controlling the quality of titles in the test set.

Fine-tuning We fine-tune generation systems here as in §4.1 (hyperparameters see Appendix I): (1) we fine-tune a $\texttt{BART}_{\texttt{xsum}}$ on the abstract-title pairs in the train set with humor constraints. (2) We continue fine-tuning the model from (1) on self-generated pseudo data.[6]

The motivation of (2) is that we observe that the systems tend to ignore the humor constraints in the input and generate identical titles for different constraints in initial experiments. We assume that to expose systems to titles with different humor levels for the same abstract during training can encourage them to pay more attention to the humor constraints. To obtain the pseudo data, we: (i) generate titles for abstracts in the train set but with "opposite" humor constraints compared to the original titles, keeping only those pseudo titles with the correct humor labels assigned by the humor classifier; (ii) filter out *FUNNY* labeled titles with very frequent n-grams, in order to encourage more diverse titles. We finally merge the filtered pseudo data with the original data. Thus, in the training data of (2), each abstract has two titles, one with label *FUNNY* and the other with ¬*FUNNY*; it contains 15,474 instances in total, where 50% are pseudo ones.

5.1 Evaluation

We report results on generating both funny and not-funny titles, to explore the difference in models' performance after involving humor generation, based on both automatic and human evaluation.

Automatic Evaluation Based on the results for the automatic evaluation metrics in §4.3, we only leverage **COMET** and our supervised metric **A2TMetric** here to evaluate title quality. To evaluate humor, we use the following three metrics: (1) $\textbf{F1}_{\textbf{macro}}$ between the expected humor labels and those assigned by the humor classifier. (2) System accuracy of generating titles on correct humor levels, denoted as $\textbf{ACC}_{\textbf{FUNNY}}$ and $\textbf{ACC}_{\neg\textbf{FUNNY}}$. (3) The ratio of the cases that the systems generate the same titles for both humor constraints to all generation cases ($\textbf{Ratio}_{\textbf{SAME}}$); lower is better.

We generate titles with constraint on both humor levels for all abstracts in the test set, computing the automatic evaluation on 1200 titles in total.

Results We evaluate humor before and after training on pseudo data in Appendix J, Table 19:

Metric	COMET		A2TMetric	
humor constraint	¬FUNNY	FUNNY	¬FUNNY	FUNNY
$\texttt{BART}_{\texttt{xsum}}$	**0.0598**	0.0582	**-2.30**	-2.32
$\texttt{BART}_{\texttt{xsum}}$+pseudo	0.0593	0.0541	-2.31	-2.37
HUMAN	0.0586		-2.36	

Table 4: Automatic evaluation for titles' quality. We bold the best performance assessed by each metric. "Humor constraint" refers to the constraints given to the input of the generation systems.

(1) after continued training on the pseudo data, $\texttt{BART}_{\texttt{xsum}}$+pseudo achieves substantially higher F1_{macro} (from 0.647 to 0.856) and $\text{ACC}_{\textit{FUNNY}}$ (from 40.2% to 77.8%), and slightly better $\text{Ratio}_{\text{SAME}}$ (from 6.5% to 4.7%). (2) $\text{ACC}_{\neg\textit{FUNNY}}$ drops slightly compared to $\texttt{BART}_{\texttt{xsum}}$ (94.5% vs. 93.6%), indicating that both systems have high accuracy on generating ¬*FUNNY* titles and the fine-tuning on pseudo data only improves the system's accuracy to generate *FUNNY* titles.

We then present the quality evaluation results in Table 4. Both BART systems obtain better results than HUMAN on both evaluation metrics, which is in line with the observation in §4.2, especially when generating ¬*FUNNY* titles. However, we observe a consistent performance drop after training on the pseudo data (values in the first row vs. those in the second row). Further, we also note that the system generated ¬*FUNNY* titles have better quality than the *FUNNY* ones (values in the left column vs. those in the right column).

Human Evaluation We randomly sample 100 abstracts from the test set with controls on the source of the papers (80% from NLP and 20% from ML) and on the humor label of the original titles (50% *FUNNY* and 50% ¬*FUNNY*). For each abstract with a human funny title, we generate a funny and a non-funny system title, and accordingly for each non-funny human title. Thus, each evaluation instance contains one abstract and five titles: 1 original title + 4 system titles (2 generation systems × 2 humor levels). The annotators rank the five titles on two criteria: *general quality* and *humor degree*, based on the abstract; the annotators can assign identical ranks to multiple titles. We show a screenshot of an annotation instance and the annotation guidelines in Figure 2 in the appendix. Five annotators (three PhD students, one undergraduate student and one senior researcher) jointly annotate 10 from these 100 instances, obtaining 0.782 Spearman for humor and 0.325 for quality ranking on average per

[6] Synthetic data can be a useful resource (He et al., 2021), despite potential limitations (Shumailov et al., 2023).

humor constraint/label	*FUNNY*		*¬FUNNY*	
system	humor	quality	humor	quality
BART$_{xsum}$	1.94	**2.70**	2.76	**2.10**
BART$_{xsum}$+pseudo	1.58	2.97	2.75	2.56
HUMAN	**1.51**	2.86	**2.40**	2.63

Table 5: Average rank of the system titles for the abstracts with original titles labeled as *FUNNY* and *¬FUNNY* separately in the human evaluation of general quality and humor degree; smaller values denotes higher ranks. "Humor constraint/label" refers to the constraints given to the input of the generation systems and the humor labels of the original titles.

annotator pair. Then, they separately evaluate the remaining 90 instances. Note that since in our evaluation annotators rank titles, even the first ranked title does not necessarily have to be of high quality or funny, for any given abstract, if the remaining are very bad concerning quality/humor.

Results Table 20 (appendix) compares the two BART systems across all 200 instances (one funny and one non-funny title per abstract). Similar to automatic evaluation, we observe (1) a general quality drop but a performance boost for humor generation after training on pseudo data and (2) *¬FUNNY* titles have better quality than *FUNNY* ones.

Further, we compare the system titles with the original human titles in Table 5. BART$_{xsum}$ ranks higher than HUMAN concerning quality when generating both *FUNNY* and *¬FUNNY* titles (2.70 vs. 2.86 and 2.10 vs. 2.63), which is consistent with our previous human evaluation (§4.2). However, fine-tuning on the pseudo data impacts the quality of the generated funny titles, as the system is rated worse than HUMAN only in this category (2.97 vs. 2.86), which is also in line with our automatic evaluation from A2TMetric. HUMAN still generates funnier titles than the automatic systems, ranking highest among all systems (1.51 vs. 1.58-1.94).

6 Comparison with ChatGPT

We compare our fine-tuned BART$_{xsum}$ (without training on pseudo data) with the recent popular ChatGPT model.[7] Firstly, we use the two models to generate funny and not funny titles for 100 abstracts from the EMNLP 2022 handbook which ChatGPT could not have seen in its training data.

[7]Here, we used the ChatGPT interface (https://chat.openai.com/) of the first three releases (Nov. 30, 2022—Jan. 9, 2023); the official API was inaccessible back then.

system	humor rank	quality rank
BART$_{xsum}$	1.86 / 2.66	2.74 / **2.25**
ChatGPT	**1.41** / 3.12	3.62 / 2.30
human	2.53	2.85

Table 6: Average ranks of the generated *FUNNY* titles (first entry) and *¬FUNNY* titels (second entry) for **100 abstracts from EMNLP 2022 handbook** in the human evaluation of quality and humorousness; smaller values denote higher ranks. We bold the highest ranks for each criterion.

Our prompt for ChatGPT is *"I want a funny title and a not-funny title for the following abstract: [abstract]"*. The ranking-based human evaluation conducted here is identical to §5.1 and done by the same five annotators, who obtain 0.867 Spearman for humor and 0.548 for quality evaluation on average over annotator pairs this time.

The average rank per system with humor constraint is presented in Table 6. We observe that automatic generation systems are mostly ranked higher than HUMAN (2.25-2.74 vs. 2.85) except for ChatGPT producing funny titles (3.62 vs. 2.85). ChatGPT generates funnier but lower-quality titles compared to BART$_{xsum}$ but ChatGPT is almost on par for non-funny titles. Hence, we conclude that *ChatGPT without any fine-tuning may already perform similarly to our fine-tuned BART$_{xsum}$.*

After our experiments, ChatGPT has been updated several times. To inspect whether the new version performs better, we conduct a second experiment using the latest model "gpt-3.5-turbo-0613" with the official API, utilizing the default hyperparameters. Details are given in Appendix L. Overall, our evaluation suggests that *the newer ChatGPT does not perform better:* In 25 out of 40 cases, the previous titles were selected as the better ones. In fact, the new version performs much worse for generating *FUNNY* titles: it loses to the previous version on 18 out of 20 instances.

7 Discussion & Analysis

Are automatic titles really superior? Overall, our results in §5 and §6 seem to indicate that automatically generated titles outperform human titles. However, looking at the distribution of best/worst titles, we see again a high frequency of worst human titles as annotated by our human annotators; in fact, human titles are most frequently selected as worst titles except when the automatic systems

use the humor constraint. As before, the likely reason is a lower lexical overlap between human titles and abstracts. Indeed, we find that human titles have lower lexical overlap with abstracts when compared to automatically generated titles from ChatGPT and $\text{BART}_{\text{xsum}}$, e.g., 57-61% of content words in human titles appear in the abstract, while the number is 64-67% for $\text{BART}_{\text{xsum}}$ and ChatGPT. Very negatively evaluated human titles have even lower lexical overlap.

In contrast, human titles were again most frequently selected as best titles except when including ChatGPT. Overall, our findings implicate that automatically generated titles can be competitive but are presumably still slightly worse than author choices. To verify this hypothesis, *we suggest a more costly evaluation scheme in the form of a user study involving the authors of papers instead of paper external annotators in future studies.*

Is training on extra parts besides abstract beneficial? We argued that human titles may not only be based on abstracts, but (to some extent) the full papers. To inspect whether training title generation systems on more than abstracts alone leads to better systems, we train BARTXsum and the popular Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020), which can deal with longer input sequences, in two settings: (1) only abstracts and (2) abstracts, introductions, and conclusions; we denote the corresponding models as "[MODEL]+A" and "[MODEL]+X", respectively. We use the data from Hou et al. (2021), which contains the sentences of all papers from ACL Anthology until 2019. Technical details are given in Appendix M.

We randomly select 29 instances from the test sets for human evaluation: 14 for BARTXsum and 15 for LED. Two evaluators were asked to select the better one among the two titles generated by "[MODEL]+A" and "[MODEL]+X" with the same underlying model, given the abstract, introduction and conclusion. On the jointly assessed 10 instances, they obtained 0.474 Kappa. Our evaluation results show that: BARTXsum seems to benefit from training on more parts (BARTXsum+X wins 8 out of 14 instances); for LED, it is not the case (LED+A wins 11 out of 15 cases). On introspection, we do find that the models trained on more than abstracts can indeed leverage some relevant keywords not in the abstracts, which makes their titles sometimes better. On the other hand, they are tasked with identifying relevant titles given more 'background noise' (longer texts) which causes them to hallucinate more and be more vague. We show examples in Appendix N. Evaluation with more than abstracts alone is also considerably more costly for humans. Overall, these experiments thus indicate that training (and evaluating) on highly specific and condensed abstracts is advantageous.

Humor constraints On introspection, we find that the funny titles generated by ChatGPT do not conform to a style of humor used in scientific papers. This indicates that *ChatGPT lacks fine-tuning on humor in science*. For $\text{BART}_{\text{xsum}}$, its problem seems to be that it overfits to data artefacts learned from the data *indicating that it does not properly learn a generalizable notion of humor*. Additionally, both models often do not match the content of the abstract/title to the humor framing (examples see Table 21 in the appendix). In our human evaluation, such titles often obtain high humor but low quality ranks; however, when they are pertinent to the abstracts, they have the potential to receive high quality ranks as well (cf. Appendix K).

8 Conclusion

We considered the abstract-to-title generation problem using end-to-end models. To do so, we trained six recent text-to-text generation systems on more than 30k NLP and ML papers. We evaluated the systems using an array of state-of-the-art automatic metrics as well as human evaluation. Our evaluation indicates that some current text generation models can generate titles with similar quality as humans, but human authors are apparently still superior. We also considered the humorous title generation problem as an extension, compiling the first dataset in the NLP/ML domain in this context, comprising over 2.6k titles annotated by two annotators with acceptable agreement. We find that our systems struggle with generating humorous titles and instead overfit to frequent patterns in the data, indicating much scope for future research.

9 Limitations

In our work, we followed a standard protocol of evaluation of text generation involving (1) automatic metrics comparing source texts (abstracts) or references and system outputs and (2) human annotators considering the same sources of information. We argued that this standard evaluation scheme may not be fully adequate in our situation

as the human authored titles may take additional information into account (e.g., the full texts), which is difficult to incorporate, however, for our annotators and for the metrics. This leads to an (arguably small) bias against human titles, which seems to be automatically identifiable however via the distribution of best/worst titles selected for different systems. Overall, this limitation could better be addressed, however, by consulting the authors of papers for an additional but much more costly to realize evaluation in the form of a user study.

We also experimented with NLP and ML papers only, not taking other scientific fields into consideration. Finally, prompting for ChatGPT is an art in itself; other prompts may have yielded different results. To explore this, we used a slightly different prompt ("*Please give me a [funny] title for the following scientific abstract: [abstract]*") for ChatGPT on 20 instances, which led to very similar human evaluation results. It is conceivable, however, that there might have been prompts leading to better evaluation outcomes for ChatGPT.

A risk of our models is that they might produce misleading or even factually wrong titles which could be adopted by the human authors if not properly checked.

As a consequence of our missing annotation guidelines for humor, it is possible that our annotators have not clearly separated humor from related concepts such as 'click-baiting' (to the extent that such a separation is possible at all).

Acknowledgments

We thank the BMBF for its support via the grant Metrics4NLG. The last author is supported by DFG grant EG 375/5–1.

References

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. 2016. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29.

Takashi Awamura, Eiji Aramaki, Daisuke Kawahara, Tomohide Shibata, and Sadao Kurohashi. 2015. Location name disambiguation exploiting spatial proximity and temporal consistency. In *SocialNLP 2015@NAACL - 3rd International Workshop on Natural Language Processing for Social Media, Proceedings of the Workshop*, SocialNLP 2015@NAACL - 3rd International Workshop on Natural Language Processing for Social Media, Proceedings of the Workshop, pages 1–9. Association for Computational Linguistics (ACL). Publisher Copyright: © 2015 Association for Computational Linguistics; 3rd Workshop on Natural Language Processing for Social Media, SocialNLP 2015, associated with NAACL 2015 ; Conference date: 05-06-2015.

Samaneh Azadi and Suvrit Sra. 2014. Towards an optimal stochastic alternating direction method of multipliers. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 620–628, Bejing, China. PMLR.

David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. 2017. The shattered gradients problem: If resnets are the answer, then what is the question?

Dominik Beese, Begüm Altunbaş, Görkem Güzeler, and Steffen Eger. 2023. Did ai get more negative recently? *Royal Society Open Science*, 10.

Jonas Belouadi and Steffen Eger. 2023. Bygpt5: End-to-end style-conditioned poetry generation with token-free language models. In *ACL*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Hugh L. Burns. 1979. Stimulating rhetorical invention in english composition through computer-assisted instruction.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Sungmin Cha, Hsiang Hsu, Flávio P. Calmon, and Taesup Moon. 2020. Cpr: Classifier-projection regularization for continual learning. *CoRR*, abs/2006.07326.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yanran Chen and Steffen Eger. 2022. Menli: Robust evaluation metrics from natural language inference. *ArXiv*, abs/2208.07316.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019. BAM! born-again multi-task networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937, Florence, Italy. Association for Computational Linguistics.

Matt Crane. 2018. Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics*, 6:241–252.

Margherita Dore. 2019. *Humour in Audiovisual Translation: Theories and Applications*. Routledge, New York.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pablo Duboue and Jennifer Chu-Carroll. 2006. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *North American Chapter of the Association for Computational Linguistics*.

Aleksandra Edwards, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. 2020. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5522–5529, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, Richard Socher, and Dragomir R. Radev. 2020. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Matthew E Falagas, Angeliki Zarkali, Drosos E Karageorgopoulos, Vangelis Bardakas, and Michael N Mavros. 2013. The impact of article length on the number of future citations: a bibliometric analysis of general medicine journals. *PLoS One*, 8(2):e49476.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fumiyo Fukumoto and Yoshimi Suzuki. 2004. A comparison of manual and automatic constructions of category hierarchy for classifying large corpora. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 65–72, Boston, Massachusetts, USA. Association for Computational Linguistics.

Jonas Geiping, Liam Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. 2020. Witches' brew: Industrial scale data poisoning via gradient matching.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ajda Gokcen and Marie-Catherine de Marneffe. 2015. I do not disagree: leveraging monolingual alignment to detect disagreement in dialogue. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 94–99, Beijing, China. Association for Computational Linguistics.

Sreenivas Gollapudi, Kostas Kollias, and Debmalya Panigrahi. 2019. You get what you share: Incentives for a sharing economy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2004–2011.

Tanya Goyal, Nazneen Rajani, Wenhao Liu, and Wojciech Kryscinski. 2022. HydraSum: Disentangling style features in text summarization with multi-decoder models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 464–479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. The workweek is the best time to start a family – a study of GPT-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 528–544, Online. Association for Computational Linguistics.

James Hartley. 2008. *Academic writing and publishing: A practical handbook*. Routledge.

He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. 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 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuanli He, Islam Nassar, Jamie Ryan Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2021. Generate, annotate, and learn: Nlp with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

Stephen B. Heard, Chloe A. Cull, and Easton R. White. 2022. If this title is funny, will you cite me? citation impacts of humour and other features of article titles in ecology and evolution. *bioRxiv*.

Tom Heskes. 1996. Balancing between bagging and bumping. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 707–714, Online. Association for Computational Linguistics.

Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Kai-Wei Chang, and Yizhou Sun. 2022. Empowering language models with knowledge graph reasoning for question answering.

Baijun Ji, Tong Zhang, Yicheng Zou, Bojie Hu, and Si Shen. 2022. Increasing visual awareness in multimodal neural machine translation from an information theoretic perspective. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6755–6764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988, Online. Association for Computational Linguistics.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. 2019. Thieves on sesame street! model extraction of bert-based apis. *ArXiv*, abs/1910.12366.

Joel Lang and Mirella Lapata. 2010. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 939–947, Los Angeles, California. Association for Computational Linguistics.

Nyoungwoo Lee, ChaeHun Park, Ho-Jin Choi, and Jaegul Choo. 2022. Pneg: Prompt-based negative response generation for dialogue response selection task. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10692–10703, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Grant Lewison and James Hartley. 2005. What's in a title? numbers of words and the presence of colons. *Scientometrics*, 63(2):341–356.

Pengcheng Li, Wei Lu, and Qikai Cheng. 2022. Generating a related work section for scientific papers: an optimized approach with adopting problem and method information. *Scientometrics*, 127(8):4397–4417.

SHAO LI. 2004. Integrating context and transliteration to mine new word translations from comparable corpora.

Wanli Li and Tieyun Qian. 2022. Graph-based model generation for few-shot relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 62–71, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 163–169, Melbourne, Australia. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

Prakhar Mishra, Chaitali Diwan, Srinath Srinivasa, and G Srinivasaraghavan. 2021. Automatic title generation for text with pre-trained transformer language model. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 17–24. IEEE.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Scigen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Norbert Mundorf, Azra Bhatia, Dolf Zillmann, Paul Lester, and Susan Robertson. 1988. Gender differences in humor appreciation. *Humor*, 1(3):231–244.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Vivi Nastase and Marius Popescu. 2009. What's in a name? In some languages, grammatical gender. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1377, Singapore. Association for Computational Linguistics.

Ani Nenkova, Kathleen McKeown, et al. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233.

Ulrike Padó. 2016. Get semantic with me! the usefulness of different feature types for short-answer grading. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2186–2195, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexander Pak and Patrick Paroubek. 2010. Twitter based system: Using twitter for disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, page 436–439, USA. Association for Computational Linguistics.

Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. 2014. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Charuta Pethe and Steve Skiena. 2019. The trumpiest trump? identifying a subject's most characteristic tweets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1653–1663, Hong Kong, China. Association for Computational Linguistics.

Maxime Peyrard, Beatriz Borges, Kristina Gligoric, and Robert West. 2021. Laughing heads: Can transformers detect what makes a sentence funny? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3899–3905. ijcai.org.

Yuval Pinter, Cassandra L. Jacobs, and Max Bittker. 2020. NYTWIT: A dataset of novel words in the New York Times. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6509–6515, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jan Wira Gotama Putra and Masayu Leylia Khodra. 2017. Automatic title generation in scientific articles for authorship assistance: a summarization approach. *Journal of ICT Research and Applications*, 11(3):253–267.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Alan Ritter, Stephen Soderland, Doug Downey, and Oren Etzioni. 2008. It's a contradiction – no, it's not: A case study using functional relations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, Honolulu, Hawaii. Association for Computational Linguistics.

Christopher De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. 2015. Taming the wild: A unified analysis of hogwild!-style algorithms.

Joseph Sanu, Mingbin Xu, Hui Jiang, and Quan Liu. 2017. Word embeddings based on fixed-size ordinally forgetting encoding. In *Proceedings of the 2017 Conference on Empirical Methods in Natural*

Language Processing, pages 310–315, Copenhagen, Denmark. Association for Computational Linguistics.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

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, Vancouver, Canada. Association for Computational Linguistics.

Chen Shani, Nadav Borenstein, and Dafna Shahaf. 2021. How did this get funded?! Automatically identifying quirky scientific achievements. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 14–28, Online. Association for Computational Linguistics.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *ArXiv*, abs/2305.17493.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 5716–5728.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tiberiu Sosea and Cornelia Caragea. 2020. CancerEmo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904, Online. Association for Computational Linguistics.

Zayne Sprague, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction with incomplete information.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, page 952–961, USA. Association for Computational Linguistics.

Piotr Szymański and Kyle Gorman. 2020. Is the best better? Bayesian statistical model comparison for natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2203–2212, Online. Association for Computational Linguistics.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *IJCAI*, volume 17, pages 4109–4115.

Zeqi Tan, Yongliang Shen, Xuming Hu, Wenqi Zhang, Xiaoxia Cheng, Weiming Lu, and Yueting Zhuang. 2022. Query-based instance discrimination network for relational triple extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7677–7690, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Raphael Tang, Jaejun Lee, Ji Xin, Xinyu Liu, Yaoliang Yu, and Jimmy Lin. 2020. Showing your work doesn't always work. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2766–2772, Online. Association for Computational Linguistics.

Zineng Tang, Jie Lei, and Mohit Bansal. 2021. DeCEMBERT: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426, Online. Association for Computational Linguistics.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony S. Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *ArXiv*, abs/2211.09085.

Noriko Tomuro. 2001. Tree-cut and a lexicon based on systematic polysemy. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel. 2020. ENGINE: Energy-based inference networks for non-autoregressive machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2826, Online. Association for Computational Linguistics.

Erik Velldal, Lilja Øvrelid, Jonathon Read, and Stephan Oepen. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational Linguistics*, 38(2):369–410.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019a. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

Qifan Wang, Li Yang, Xiaojun Quan, Fuli Feng, Dongfang Liu, Zenglin Xu, Sinong Wang, and Hao Ma. 2022. Learning to generate question by asking question: A primal-dual approach with uncommon word generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 46–61, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019b. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991, Florence, Italy. Association for Computational Linguistics.

William Yang Wang and Kathleen McKeown. 2010. "got you!": Automatic vandalism detection in Wikipedia with web-based shallow syntactic-semantic modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1146–1154, Beijing, China. Coling 2010 Organizing Committee.

Xinrun Wang, Bo An, Martin Strobel, and Fookwai Kong. 2018. Catching captain jack: Efficient time and space dependent patrols to combat oil-siphoning in international waters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

Liang Wen, Houfeng Wang, Yingwei Luo, and Xiaolin Wang. 2022. M3: A multi-view fusion and multi-decoding network for multi-document reading comprehension. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1450–1461, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joachim Wermter and Udo Hahn. 2006. You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 785–792, Sydney, Australia. Association for Computational Linguistics.

Bowen Xing and Ivor W. Tsang. 2022. Co-guiding net: Achieving mutual guidances between multiple intent detection and slot filling via heterogeneous semantics-label graphs.

Rui Yan, Mingkun Gao, Ellie Pavlick, and Chris Callison-Burch. 2014. Are two heads better than one? crowdsourced translation via a two-step collaboration of non-professional translators and editors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1134–1144, Baltimore, Maryland. Association for Computational Linguistics.

Kaiyu Yang, Jia Deng, and Danqi Chen. 2022. Generating natural language proofs with verifier-guided search.

Shunyu Yao, Rohan Rao, Matthew Hausknecht, and Karthik Narasimhan. 2020. Keep CALM and explore: Language models for action generation in text-based games. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8736–8754, Online. Association for Computational Linguistics.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. Generative knowledge graph construction: A review. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1–17, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ki Yoon Yoo and Nojun Kwak. 2022. Backdoor attacks in federated learning by rare embeddings and gradient ensembling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 72–88, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang,

Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Jianyang Zhang, Tao Liang, Mingyang Wan, Guowu Yang, and Fengmao Lv. 2022. Curriculum knowledge distillation for emoji-supervised cross-lingual sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 864–875, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Xiaodan Zhu, Gerald Penn, and Frank Rudzicz. 2009. Summarizing multiple spoken documents: Finding evidence from untranscribed audio. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, page 549–557, USA. Association for Computational Linguistics.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online. Association for Computational Linguistics.

A Filtering

(1) We restrict the data to the main conference papers (e.g., EMNLP, ACL). We limit the data to abstracts of length smaller than 400 words as extremely long abstracts in the dataset often contain extra sections other than abstracts. (3) We only leverage papers published after the year 2000 (which form the majority anyway).

B Training details for title generation

We train models with AdamW Optimizer (Loshchilov and Hutter, 2019) and linear learning rate scheduler, and subsequently use beam search (Vijayakumar et al., 2016) as the sampling strategy to generate the output candidates. The optimal checkpoint for each model is selected based on the ROUGE1/2/L (Lin, 2004) scores on the dev set. Table 7 displays the hyperparameter for training and Table 8 shows the parameters used for beam search. The models were trained using Google Colab with a Tesla K80 GPU which has 24 GB of memory. We show the number of parameters of each baseline model in Table 15.

C Variants of used automatic evaluation metrics

In ref-based evaluation, we report Rouge-1 recall, BERTScore recall, unigram MoverScore, BARTScore recall, MENLI(ref←cand_e-c) and COMET(wmt20-comet-da). In ref-free setup, we use the Faithfulness variant for BARTScore, MENLI(src→cand_-c) and COMET (wmt21-comet-qe-mqm) instead; the variants of the other metrics are the same as in ref-based setting.

D Ref-based evaluation results of baseline models

Table 9 shows the ref-based automatic evaluation results of the baseline models.

E $\text{BART}_{\text{base}}$ VS. $\text{BART}_{\text{xsum}}$

Table 10 shows the examples of abstract-title pairs where $\texttt{BART}_{\texttt{base}}$ failed to capture the key information in the abstract while $\texttt{BART}_{\texttt{xsum}}$ succeeded.

F Examples of funny titles

Table 13 and Table 14 show sample funny titles labeled by human annotators. We note: some instances of humor require contextual (e.g., culture- or domain-specific) knowledge such as references to popular TV shows ('Germany's next language model'); this is characteristic of humor and makes it challenging/subjective. Despite of this, our agreements indicate a shared notion of humor among our annotators.

	learning rate	batch size	epochs	gradient accumulation steps
$\texttt{BART}_{\texttt{xsum}}$	3e-05	3	3	8
$\texttt{PEGASUS}_{\texttt{xsum}}$	6e-04	3	3	8
$\texttt{BART}_{\texttt{base}}$	3e-04	8	3	8
GPT2	3e-04	2	3	8
T5	3e-04	8	3	8
$\texttt{BART}_{\texttt{cnn}}$	3e-04	4	3	8

Table 7: Training hyperparameter for title generation. We use the AdamW optimizer with a weight decay of 0.01 and keep the other settings as default in Huggingface's Trainer API.

max length	30
min length	3
repetition penalty	2
length penalty	10
num beams	5
num return sequences	5

Table 8: Parameter settings for beam search.

system	MoverS	BERTS	COMET	BARTS	MENLI	ROUGE
$\texttt{BART}_{\texttt{xsum}}$	**0.410**	**0.912**	**-0.283**	-3.816	0.076	**0.455**
$\texttt{PEGASUS}_{\texttt{xsum}}$	0.404	0.906	-0.371	-3.964	0.005	0.384
$\texttt{BART}_{\texttt{base}}$	0.405	0.907	-0.373	-3.986	0.036	0.403
GPT2	0.400	0.902	-0.461	-4.114	-0.020	0.361
T5	0.381	0.898	-0.501	-4.177	-0.025	0.337
$\texttt{BART}_{\texttt{cnn}}$	0.282	0.907	-0.634	**-3.747**	**0.133**	0.448

Table 9: Ref-based evaluation results of the baseline models. We underlie the best performance among all generation systems including human. We bold the best performance among all automatic generation systems excluding human.

G Humor annotation + classifiation

The two annotators first annotated the same **230 titles** independently, obtaining only 0.397 Kappa agreement, which indicates a relatively bad annotation quality. To improve the inter-agreement between the annotators, they then discussed the reasons leading to disagreement. Subsequently, they annotated another **300 titles** independently, achieving a decent 0.650 Kappa for a task as subjective as humor. As a consequence, *we use the maximal label value among the two annotations for each title as its final label for the 300 titles*, i.e., if one annotator labels a title with 1 ($FUNNY_{\text{med}}$), while the other labels with 0 ($\neg FUNNY$), we assign label 1 to the title. Each annotator then labeled 600 different titles separately, bringing **1,730** ($230 + 300 + 600 \times 2 = 1730$) annotated titles in total, where 1,603 titles are labeled as $\neg FUNNY$, 106 as $FUNNY_{\text{med}}$ and 21 as $FUNNY$.

As the funny titles (labeled as *FUNNY*) are very few compared to the not funny ones (labeled with 0), we generate 11 different data splits, where the train set of each split consists of 100 funny titles and 200 not funny ones (randomly sampled from the 1730 titles), while the remaining 27 funny titles and other 27 not funny ones compose the dev set. From the 11 different data splits, we obtain 11 classifiers (checkpoints selected based on the macro F1 on each dev set). We then evaluate the ensembles of the 11 classifiers on **315 newly annotated titles** by the two annotators, who obtain **0.639 Kappa** agreement this time. With this step, we study the optimal ensemble of the classifiers and also obtain more funny titles from the whole data by annotating the funniest titles selected by the ensemble classifiers. We design two types of ensemble classifiers:

- **EnsMV**, which relies on the majority vote of the 11 classifiers. Specifically, each title receives 11 labels from the 11 classifiers: if the number of $\neg FUNNY$ labels exceeds 5, the title is labeled as $\neg FUNNY$; if not, the title is labeled as *FUNNY* when the number of *FUNNY* labels exceeds the number of $FUNNY_{\text{med}}$ labels, otherwise it is labeled as $FUNNY_{\text{med}}$.
- **EnsSUM**$_{i,j}$, which depends on the sum of

Abstract	[...] we propose to learn word embeddings based on the recent fixed-size ordinally forgetting encoding (FOFE) method, which can almost uniquely encode any variable-length sequence into a fixed-size representation. [...] (Sanu et al., 2017)
$\text{BART}_{\text{base}}$	Learning Word Embeddings Based on Ordinally Forgetting Encoding
$\text{BART}_{\text{xsum}}$	Learning Word Embeddings Based on Fixed-Size Ordinally Forgetting Encoding

Abstract	[...] Unfortunately, the reliance on manual annotations, which are both difficult and highly expensive to produce, presents a major obstacle to the widespread application of these systems across different languages and text genres. In this paper we describe a method for inducing the semantic roles of verbal arguments directly from unannotated text . [...] (Lang and Lapata, 2010)
$\text{BART}_{\text{base}}$	Inducing Semantic Roles from Text for Semantic Role Labeling
$\text{BART}_{\text{xsum}}$	A Probabilistic Model for Semantic Role Induction from Unannotated Text

Abstract	[...] At the same time, we argue that relation labeling can benefit from naked tree structure and should be treated elaborately with consideration of three kinds of relations including within-sentence, across-sentence and across-paragraph relations. Thus, we design a pipelined two-stage parsing method for generating an RST tree from text. [...] (Wang et al., 2017)
$\text{BART}_{\text{base}}$	Pipelined Two-Stage Parsing of Named Discourse Trees
$\text{BART}_{\text{xsum}}$	Pipeline-based Parsing of Discourse Trees for RST and Relation Labeling

Table 10: Examples of abstract-title pairs where $\text{BART}_{\text{base}}$ failed to capture the key information in the abstract while $\text{BART}_{\text{xsum}}$ succeeded. The key information is highlighted in both abstracts and titles.

	230 instance	35 instances
	τ	τ
ROUGE	-0.054	-0.014
BARTS	0.092	0.121
BERTS	0.078	0.113
MoverS	0.001	0.038
MENLI	0.061	0.121
COMET	**0.127**	0.194
A2TMetric	-	**0.276**

Table 11: Segment-level WMT τ-like correlations of ref-free evaluation metrics on all 230 instances (1380 titles; left block) and 35 instances (210 titles; right block). The correlations on the 35 instances are averaged over the test sets from five splits. We bold the highest correlation in each block.

Title	Label
Learning to learn by gradient descent by gradient descent (Andrychowicz et al., 2016)	*FUNNY*
CancerEmo: A Dataset for Fine-Grained Emotion Detection (Sosea and Caragea, 2020)	$\textit{FUNNY}_{\text{med}}$
Global Encoding for Abstractive Summarization (Lin et al., 2018)	$\neg$*FUNNY*

Table 12: Examples of annotated titles.

the label values. The sum of the label values for each title ranges from 0 (11 classifiers $\times$ 0 for $\neg$*FUNNY*) to 22 (11 classifiers $\times$ 2 for *FUNNY*). We then select a threshold i for $\textit{FUNNY}_{\text{med}}$ and j for *FUNNY*: if sum $< i$, the title is labeled as $\neg$*FUNNY*; otherwise it is labeled as $\textit{FUNNY}_{\text{med}}$ (when sum $< j$) or *FUNNY* (when sum $\geq j$).

Table 16 shows the evaluation results of Stage 1; we only present the performance of $\text{EnsSUM}_{i,j}$ with optimal i and j here, i.e., $\text{EnsSUM}_{7,16}$. We observe that: (1) both ensembles perform better than the individual ones (+4-5% macro F1) and (2) **$\text{EnsSUM}_{7,16}$** is slightly better than EnsMV (62.4% vs. 61.4% macro F1).

H Dataset Statistics

Table 18 shows the statistics of the final dataset.

I Parameters for humor generation

We train $\text{BART}_{\text{xsum}}$ on our train set using the AdamW optimizer with weight decay 0.01 and learning rate 4e-05 for 5 epochs. Then we continue to train it on the pseudo data for one epoch to obtain $\text{BART}_{\text{xsum}}$+pseudo. We use the default settings in Huggingface's Trainer API for the other hyperparameters. We train the models with an RTX A6000 GPUwhich has 48 GB of memory.

To monitor the models' ability to generate titles on correct humor levels, we use *macro F1* between

the expected humor labels (i.e., the humor constraints given to the inputs) and the humor labels assigned to the generated titles by the humor classifier as the performance indicator, with which on the dev set we select the optimal model checkpoints of the two systems.

J Automatic evaluation of humor generation

Table 19 shows the systems' ability for humor generation before and after training on the pseudo data according to the automatic evaluation.

K Examples of system-generated funny titles

Table 22 and 23 show 10 system-generated low-quality funny titles and 10 system-generated high-quality funny titles, respectively, according to the human evaluation results.

Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper) (Castro et al., 2019)
Thieves on Sesame Street! Model Extraction of BERT-based APIs (Krishna et al., 2019)
Are Two Heads Better than One? Crowdsourced Translation via a Two-Step Collaboration of Non-Professional Translators and Editors (Yan et al., 2014)
Taming the Wild: A Unified Analysis of Hogwild-Style Algorithms (Sa et al., 2015)
Balancing Between Bagging and Bumping (Heskes, 1996)
Speculation and Negation: Rules, Rankers, and the Role of Syntax (Velldal et al., 2012)
What's in a name? In some languages, grammatical gender (Nastase and Popescu, 2009)
BAM! Born-Again Multi-Task Networks for Natural Language Understanding (Clark et al., 2019)
Is the Best Better? Bayesian Statistical Model Comparison for Natural Language Processing (Szymański and Gorman, 2020)
Keep CALM and Explore: Language Models for Action Generation in Text-based Games (Yao et al., 2020)

Table 13: Examples of **human** titles which were labeled as $FUNNY_{\text{med}}$+$FUNNY_{\text{med}}$, $FUNNY_{\text{med}}$+*FUNNY*, or *FUNNY*+*FUNNY* by the two annotators (the two entries denote the label assigned by different annotators.).

FUNNY
German's Next Language Model (Chan et al., 2020)
Is the Best Better? Bayesian Statistical Model Comparison for Natural Language Processing (Szymański and Gorman, 2020)
Comparing Apples to Apple: The Effects of Stemmers on Topic Models (Schofield and Mimno, 2016)
(Almost) No Label No Cry (Patrini et al., 2014)
The Trumpiest Trump? Identifying a Subject's Most Characteristic Tweets (Pethe and Skiena, 2019)
Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results (Crane, 2018)
Know What You Don't Know: Unanswerable Questions for SQuAD (Rajpurkar et al., 2018)
Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer (Rao and Tetreault, 2018)
Can You Tell Me How to Get Past Sesame Street? Sentence-Level Pretraining Beyond Language Modeling (Wang et al., 2019a)
Showing Your Work Doesn't Always Work (Tang et al., 2020)
"Got You!": Automatic Vandalism Detection in Wikipedia with Web-based Shallow Syntactic-Semantic Modeling (Wang and McKeown, 2010)
It's a Contradiction - no, it's not: A Case Study using Functional Relations (Ritter et al., 2008)
$FUNNY_{\text{med}}$
CPR: Classifier-Projection Regularization for Continual Learning (Cha et al., 2020)
NYTWIT: A Dataset of Novel Words in the New York Times (Pinter et al., 2020)
MedDialog: Large-scale Medical Dialogue Datasets (Zeng et al., 2020)
Catching Captain Jack: Efficient Time and Space Dependent Patrols to Combat Oil-Siphoning in International Waters (Wang et al., 2018)
The Shattered Gradients Problem: If resnets are the answer, then what is the question? (Balduzzi et al., 2017)
Go Simple and Pre-Train on Domain-Specific Corpora: On the Role of Training Data for Text Classification (Edwards et al., 2020)
SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge (Ke et al., 2020)
Get Semantic With Me! The Usefulness of Different Feature Types for Short-Answer Grading (Padó, 2016)
Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching (Geiping et al., 2020)
ENGINE: Energy-Based Inference Networks for Non-Autoregressive Machine Translation (Tu et al., 2020)
You Can't Beat Frequency (Unless You Use Linguistic Knowledge) - A Qualitative Evaluation of Association Measures for Collocation and Term Extraction (Wermter and Hahn, 2006)
OntoGUM: Evaluating Contextualized SOTA Coreference Resolution on 12 More Genres (Zhu et al., 2021)

Table 14: Selected **human** titles in the annotated data judged as funny or medium funny by the annotators.

Annotation Example

	humor_rank	quality_rank
Abstract 1: This paper presents a model for summarizing multiple untranscribed spoken documents. Without assuming the availability of transcripts, the model modifies a recently proposed unsupervised algorithm to detect re-occurring acoustic patterns in speech and uses them to estimate similarities between utterances, which are in turn used to identify salient utterances and remove redundancies. This model is of interest due to its independence from spoken language transcription, an error-prone and resource-intensive process, its ability to integrate multiple sources of information on the same topic, and its novel use of acoustic patterns that extends previous work on low-level prosodic feature detection. We compare the performance of this model with that achieved using manual and automatic transcripts, and find that this new approach is roughly equivalent to having access to ASR transcripts with word error rates in the 33–37% range without actually having to do the ASR, plus it better handles utterances with out-ofvocabulary words.		
Title 0: Unsupervised Summarization of Spontaneous Speech using Acoustic Patterns	2	2
Title 1: Unsupervised Summarization of Spoken Text using Acoustic Patterns	2	1
Title 2: Don't Do the ASR, Use Acoustic Patterns! Unsupervised Summarization of Spoken Text using Re-occurring Patterns	1	2
Title 3: Reading Between the Lines: Unsupervised Summarization of Spontaneous Speech using Acoustic Patterns	2	4
Title 4: Summarizing multiple spoken documents: finding evidence from untranscribed audio	2	4

Annotation Guidelines

1. Each instance has one abstract and five titles for it.
2. Each annotator should rank the 5 titles given in the abstract according to their level of humorousness and general quality. For example, the title deemed the funniest should be given a rank of 1 and the least funny title should be given the lowest rank.
3. Soft ranking is allowed. I.e., you could rank titles equally if you could not diffrentiate them according to a certain criterion; your final ranks could be like [1,1,2,2,2], [1,2,2,4,4] or so.
4. Don't worry about small mistakes in your annotation for soft ranking - even if you write [1,4,4,5,5] instead of [1,2,2,4,4], it's no problem.
5. "General quality" could concern criteria such as fluency, information adequacy, and grammatical correctness etc. However, as there is no clear criterion for the assessment of general quality, the evaluation is left to the discretion of the assessor.

Figure 2: Screenshot of an annotation instance and the annotation guidelines. The evaluation is conducted with google spreadsheet.

	# parameters
$\text{BART}_{\text{base}}$	140M
$\text{BART}_{\text{xsum}}$	400M
BART_{cnn}	400M
T5	60M
GPT2	117M
$\text{PAGASUS}_{\text{xsum}}$	568M

Table 15: Number of parameters of the six baseline models.

L Comparison of ChatGPT versions

We randomly choose 10 abstract-title pairs from our previous evaluation for both low- and high-quality titles, following each humor constraint (*FUNNY* and $\neg$*FUNNY*); this totals to 40 evaluation instances.[8] Then, we use the new version of ChatGPT to generate titles for those abstracts, according to the humor constraints of their paired titles. Two annotators were tasked with rating the higher quality title among the two from different

[8] In this context, we consider the titles ranked above 2 as high quality and below 3 as low quality.

	Individuals	Ensembles	
		EnsMV	$\text{EnsSUM}_{7,16}$
F1	57.6%	61.4%	**62.4%**

Table 16: Average macro F1 over the 11 individual classifiers and macro F1 of the ensemble classifiers from stage 1 on the evaluation data of 315 titles (where the two annotators obtain 0.639 kappa). We bold the highest macro F1 score.

ChatGPT versions, obtaining a Cohen's Kappa score of 0.756 for agreement on 10 common instances.[9]

M Training on extra parts besides abstract

We do the same filtering in §3 except for restricting to main conference papers, as there are no venue labels; additionally, we remove the papers which have empty title, abstract, introduction, or conclusion sections in the data. The filtered data contains 22,452 papers, which are then split into train, dev, and test sets in a ratio of 8:1:1. For "[MODEL]+X" models, we concatenate the texts of the three parts by two "</s>" tokens as the model input. For LED models, we limit the maximal input length to 2,048, which is able to cover the concatenated inputs of the great majority of instances; as for BARTXsum, the maximal input length is 1,024, which indicates the inputs of around half of the instances will be truncated.

We train all models using the Trainer API from huggingface with a learning rate of 4e-5 and a batch size of 32 for 20 epochs; the other hyperparameters are default. Each training was stopped by an early stopping with 2 patience, based on the rouge scores on the dev set. We use beam search with 5 beams and a length penalty of 2 for decoding.

[9] If one can not differentiate between the two titles, it is allowed to annotate them as equal.

	#titles	Kappa three-way	Kappa binary
Stage 1	230	0.397	**0.513**
	300	0.650	**0.754**
	315	0.639	**0.709**
Stage 2	197	0.649	**0.661**

Table 17: Kappa agreements between the two annotators on several data pieces. "#titles" refers to the number of titles in a certain piece of data. We bold the higher Kappa on the same data.

	Humor label ¬*FUNNY*	Humor label FUNNY	Total	Source NLP	Source ML
train	30,741	1,011	31,752	16,141	15,611
dev	400	200	600	480	120
test	400	200	600	480	120
total	31,541	1,411	32,952	17,101	15,851

Table 18: Distribution of the source (NLP or ML) and humor labels (*FUNNY* or ¬*FUNNY*) of the instances in our dataset.

	$F1_{macro}$	$ACC_{\neg FUNNY}$	ACC_{FUNNY}	$Ratio_{SAME}$
$BART_{xsum}$	0.647	**94.5%**	40.2%	6.5%
$BART_{xsum}$+pseudo	**0.856**↑	93.6%↓	**77.8%**↑	**4.7%**↑

Table 19: Automatic evaluation for the systems' ability to generate titles with correct humor constraints. We bold the best performance. ↑/↓ in the second row indicates the performance being better/worse after training on the pseudo data.

system	humor constraint	humor	quality
$BART_{xsum}$	¬*FUNNY*	2.85	**2.32**
	FUNNY	1.79	2.81
$BART_{xsum}$+pseudo	¬*FUNNY*	2.97	2.64
	FUNNY	**1.43**	3.26

Table 20: Average rank of the system titles for all abstracts in the human evaluation of general quality and humor degree; smaller values denotes higher ranks. "Humor constraint" refers to the constraints given to the input of the generation systems.

N MODEL+A vs. MODEL+X

Table 24 illustrates the examples of abstract-title pairs where the important keywords were missing from the abstracts and only available in other parts like conclusion, and Table 25 displays the examples of titles with hallucinations.

BART$_{xsum}$ - funny titles with artefacts
What's in a Semantic Model? Comparing LDA and LSA on the Web (Stevens et al., 2012) *Don't* paraphrase unless you know what you are talking about: Improving Question Answering Performance by Paraphrasing (Duboue and Chu-Carroll, 2006) *Don't* Transliterate, Use Context! Mining New Word Translations from Comparable Corpora Using Context Information (LI, 2004) *Reading Between the Lines:* Unsupervised Summarization of Spontaneous Speech using Acoustic Patterns (Zhu et al., 2009)
ChatGPT - non-scientific funny titles
Proof Generation: Now You See It, Now You Don't! (Yang et al., 2022) Co-Guiding Net: Helping You Hit the Slot and Intent Jackpot! (Xing and Tsang, 2022) Abduct Me If You Can: How to Prove a Claim With a Little Help From Your Friends (Premises) (Sprague et al., 2022) OREO-LM: The Creamy, Crunchy, and Smart Way to Answering Open-Domain Questions (Hu et al., 2022)

Table 21: Examples of system-generated funny titles from BART$_{xsum}$ with artefacts and non-scientific funny titles from ChatGPT. The citations here are the original papers for those titles.

BART$_{xsum}$
Don't Invite Adversaries to Poison Your Data: Exploiting Federated Learning for Adversarial Backdoor Attacks (Yoo and Kwak, 2022) Don't Take the Easy Way Out: Generating Adversarial Negative Responses with Large-Scale Language Models for Dialogue Selection (Lee et al., 2022) Don't Give Up on Style: Learn to Generate Stylistically-Diverse Summaries with Multiple Decoders (Goyal et al., 2022) CKD: Curriculum Knowledge Distiller for Cross-Lingual Sentiment Analysis with Emoji (Zhang et al., 2022) Successive Prompting: Learning to Break Down Complex Questions into As Simple As Possible (Dua et al., 2022)
ChatGPT
Graphin' It Up: A Humorous Guide to Generative Knowledge Construction (Ye et al., 2022) Tiny Tasks, Big Results: A Hilarious Guide to Few-Shot Relation Extraction (Li and Qian, 2022) Revealing the Magic Behind Transformer Language Models: A Lighthearted Investigation (Geva et al., 2022) Ask and You Shall Receive: A Whimsical Approach to Automatic Question Generation (Wang et al., 2022) Federated Learning: The More You Poison, the More You Win! (Yoo and Kwak, 2022)

Table 22: Examples of system-generated **low-quality** funny titles, which obtain high humor ranks but low quality ranks in the human evaluation.

BART$_{xsum}$
Don't Agree with Me? Introducing Semantic Environment Features Improves Agreement-Disagreement Classification in Online Discourse (Gokcen and de Marneffe, 2015) The Myth of the Two Sides of the Same Coin: Claim Generation and Claim Retrieval in a World of Claims (Gretz et al., 2020) Sharing is Caring: Incentives for Self-Organization in Social Welfare Maximization (Gollapudi et al., 2019) DeCEMBERT: Dense Captions and Entropy Minimization for Video-and-Language Pre-training (Tang et al., 2021) Stochastic Alternating Direction Method of Multipliers Revisited: Faster Rates and Better Algorithms (Azadi and Sra, 2014)
ChatGPT
Succeed with Successive Prompting: Breaking Down Complex Questions for LMs (Dua et al., 2022) Feeling the Pulse of Dialogue: A Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation (Song et al., 2022) Triple Trouble: A Novel Query-Based Approach to Joint Entity and Relations Extraction (Tan et al., 2022) Two Heads are Better than One: A Multi-View Fusion and Multi-Decoding Method for Multi-Document Reading Comprehension (Wen et al., 2022) Seeing is Believing: A Picture's Worth a Thousand Words in Multimodal Machine Translation (Ji et al., 2022)

Table 23: Examples of system-generated **high-quality** funny titles, which obtain both high humor and quality ranks in the human evaluation.

Abstract	This paper describes a lexicon organized around systematic polysemy: a set of word senses that are related in systematic and predictable ways. The lexicon is derived by a fully automatic extraction method which utilizes a clustering technique called tree-cut. We compare our lexicon to WordNet cousins, and the inter-annotator disagreement observed between WordNet Semcor and DSO corpora. (Tomuro, 2001)
LED+A	A systematic polysemy lexicon based on tree-cut
LED+X	A Systematic Polysemy Lexicon Based on Tree-Cut Extraction
Abstract	We address the problem dealing with a large collection of data, and investigate the use of automatically constructing category hierarchy from a given set of categories to improve classification of large corpora. We use two well-known techniques, partitioning clustering, []-means and a [] to create category hierarchy. []-means is to cluster the given categories in a hierarchy. To select the proper number of [], we use a [] which measures the degree of our disappointment in any differences between the true distribution over inputs and the learner's prediction. Once the optimal number of [] is selected, for each cluster , the procedure is repeated. Our evaluation using the 1996 Reuters corpus which consists of 806,791 documents shows that automatically constructing hierarchy improves classification accuracy. (Fukumoto and Suzuki, 2004)
BARTXsum+A	Automatic Construction of Category Hierarchy for Improved Classification of Large Corpora
BARTXsum+X	Automatic Construction of Category Hierarchy for Text Classification

Table 24: Examples of abstract-title pairs where the important keywords were missing from the abstracts and only available in other parts like conclusion. We highlight the keywords in the titles from "[MODEL]+X" systems. Tokens masked with "[]" are those with OCR errors that could not be recognized.

Paper	Awamura et al. (2015)
LED+A	Location Disambiguation Using Spatial and Temporal Clues
LED+X	Location Disambiguation Using Spatial Clustering and Temporal Consistency
Paper	Pak and Paroubek (2010)
BARTXsum+A	Automatic Disambiguation of Chinese Sentiment Ambiguous Adjectives Using Twitter
BARTXsum+X	NUS-CORE : Using Twitter to Disambiguate Adjective Sentiment Ambiguous Adjectives

Table 25: Examples of titles with hallucinations. We highlight the hallucinated words in the titles from "[MODEL]+X" systems.