Probabilistic Context Extraction For Extractive Highlights Extraction

Pietro Cagnasso
Politecnico di Torino
Student ID: s300801
s300801@studenti.polito.it

Nicolò Vergaro
Politecnico di Torino
Student ID: s295633
s295633@studenti.polito.it

Luca Marcellino
Politecnico di Torino
Student ID: s292950
s292950@studenti.polito.it

Luca Villani
Politecnico di Torino
Student ID: s304992
s304992@studenti.polito.it

Abstract—**In this paper we present two works based on a prior work about extractive highlights extraction from scientific papers, namely THExt. As first extension, we perform abstractive title generation using BART. The model is fed with a concatenation of the paper's highlights and abstract. The highlights were used to express the main achievements, while the abstract was used to give a comprehensive view on the main concepts. This method proved to be effective, reaching remarkable results both in syntactic and semantics terms.**

THExt uses the abstract as a context to enrich the embeddings of each sentence during the regression step. In the second extension we focused our attention on the context. In order to do so, we had the intuition that there could exist high importance areas, in which it's likely that we will find sentences that are close to the highlights, ROUGE-2 wise. We computed these distributions and created a new context that is made of a concatenation of sentences picked according to the aforementioned distributions. This allows us to compose a context that is very likely to convey useful information. We called this procedure probabilistic context extraction (PCE). An important observation is that our proposed context is not static, unlike the abstract, thus we could perform hyper-parameter tuning to possibly enhance its performance. We obtained results comparable to the state of the art.

The code is available in this GitHub repository.

I. Introduction

This project focuses on text summarization, since we were asked to propose two extensions of La Quatra, Cagliero [6], the state of the art of highlights extraction by extractive summarization. The extensions we propose are applications of text summarization: abstractive headline generation and extractive highlight extraction, exploring a new context formulation.

Text summarization is a Natural Language Processing task that aims at condensing a text into a more concise and informative text. This helps reduce redundancy, maximize relevance and accessibility. We can identify two main approaches to address this task: abstractive summarization and extractive summarization. The former produces a new text that conveys the same meaning using different words from the source document to resemble a human generated summary. This involves understanding the context and contents of the source text. The latter produces a summary by selecting the most important sentences from the source document.

A. Headline generation

In the first extension we use abstractive summarization because the title is generated in natural language given the abstract and the highlights. The abstract was used to give a general context, while highlights were used to made the model aware of the main achievements. The model is then trained to generate a short sentence, which is the title. The model we used to carry out this task is BART (Bidirectional AutoRegressive Transformers) [7]. It uses a transformer-based encoder-decoder structure consisting of stacked layers of self-attention and a deep generative network. The encoder is a denoising bidirectional encoder and the decoder is an autoregressive decoder. BART is able to generate high-quality text by learning from large amounts of text data and has been found to outperform traditional abstractive summarization models.

B. Highlights extraction

In the second extension we use extractive summarization because we want to extract 3 to 5 sentence from a scientific paper and use them as highlights. The state of the art of extractive highlights extraction is THExt [6]. This model is based on BERT (Bidirectional Encoder Representations from Transformers) [5], specifically SciBERT [1], which is the version pre-trained on scientific texts. BERT is a multi-layer bidirectional Transformer encoder based on the original implementation described in Vaswani et al. [12]. This model is exploited by THExt to obtain embeddings of each sentence and its context (i.e., the abstract) to perform a regression task having as target the ROUGE-2 F1 measure of each sentence with respect to the true highlights. Sentences are then ranked and the top 3 are taken as highlights of the paper. For what concerns the second extension, we focused on the input part of THExt by changing the context, which in the original implementation is the abstract of the paper. As the authors found out, the abstract is in fact the best context that the model can use. Our intuition is based on the fact that, being the highlights oriented to results, we could improve the context (and thus performances) by combining information taken across different sections.

II. Methods and Experiments

Datasets: In both the tasks we performed, the datasets we used are CSPubSumm [3], BIOPubSumm, and AIPubSumm [2] as previously done by La Quatra, Cagliero. These are collections of scientific papers on computer science, bio-medical field, and artifical intelligence. During our tests we combined all these datasets, which we will call MiscPubSumm for brevity. This was done to make results comparable.

Evaluation metrics: To evaluate results we used syntactic and semantic measures. To evaluate *syntactically* we used the ROUGE metric (Recall-Oriented Understudy for Gisting Evaluation) [8]. It measures the similarity between a reference summary and a candidate summary based on the overlap of N-grams. We used ROUGE-1, ROUGE-2, and ROUGE-L to measure the overlap of uni-grams, bi-grams and the longest common sub-sequence, respectively. To evaluate *semantically* we exploited BERTScore [13]. It computes a similarity score for each token in the candidate sentence with each token in the reference sentence exploiting contextual embedding instead of exact matches. We used both metrics in the evaluation of the abstractive headline generation, while we used only ROUGE to evaluate the highlights extraction process, in order to have comparable results with the original paper.

Experimental setup: To run the experiments we used a machine[1] equipped with an Intel Xeon, 13 GB of RAM, two Nvidia T4s, and running Ubuntu 20.04 LTS. The availability of the machine was subject to the following constraints:

- The number of CPU cores is limited to 2, down from 4, whenever running a machine with a GPU enabled
- The GPU is available for 30 hours a week
- The consecutive execution time is limited to 12 hours.

A. Headline generation

The intuition behind this extension is that the title can be seen as a highlight that summarizes the key point of the paper. These information can be especially found in the abstract and the highlights.

To choose the input to be fed into BART we performed an *ablation study* on a tenth of MiscPubSumm: we tested the highlights only, the abstract only and the concatenation of the two. To concatenate the three/five highlights or the highlights with the abstract in the first and latter case we used a separation token, which helped the model understand that those were not a continuous set of sentences. Based on the results of this study (Table I), the best input option is the concatenation of highlights and abstract. We somehow expected this result because the highlights can bring useful insights on the main results, while the abstract gives a context to those sentences.

After acknowledging the best input for the model to generate headlines, we performed 2 fine-tuning steps:

1) A first epoch on MiscPubSumm to let the model learn how headlines are generated with a global view on the landscape of scientific papers (although limited to computer science, biology, medicine and artificial intelligence)
2) A second epoch on each dataset separately to specialize the model on the specific topic.

We performed this pipeline starting from a version of BART pre-trained on the well-known CNN and Daily Mail dataset [11] and XSUM [10]. It can be found on HuggingFace under the name of `distilbart-cnn-12-6`. Both the epochs were carried out with 10^{-5} as learning rate, 10^{-2} as weight decay, and 4 as *per device* batch size (8 in total). The average required time was around 2 hours per topic.

TABLE I
F1 RESULTS OF THE ABLATION STUDY ON THE INPUT FED TO BART TO GENERATE HEADLINES.

Input	ROUGE-1	ROUGE-2	BERTScore
HL only	0.3528	0.1617	0.8791
Abs only	0.4031	0.2141	0.8865
HL + Abs	**0.4287**	**0.2287**	**0.8948**

B. Probabilistic Context Extraction

We based this extension on the assumption that there exists an underlying distribution, $\theta_{t,s}$, of important sentences over each section and topic. In order to change the context by creating a composition of sentences taken across different sections, we first had to understand how to pick sentences. We couldn't do it randomly, as not all sentences convey important information. We thought of estimating the probability distribution over sentences for each section and pick sentences according to it. To compute this empirical probability distribution, $\hat{\theta}_{t,s}$, we evaluated each sentence with the paper's highlights by looking at its ROUGE-2 F1 score, then we take the top-n, with $n = 20$ sentences and check to which section they belong. What we obtain are ”high importance areas” for each section, which were discretized in 20 bins, as we can see for example in Fig. 1. The number of bins equal to 20 was chosen after an empirical analysis.

This is not enough, in fact we need to understand how many sentences we need to pick from each section, for example we can think that the abstract has many important sentences, unlike the methods section. For this reason, we computed a table which tells us the *contribution* that each section gives to highlights. This was computed using the scores that we obtained in the previous step, but now we divide the number of contributions of each section by the total number of contributions of all sections. We can see all contributions for all categories in Table II. For example in CS almost 15% of the top-20 sentences closer to the highlights come from the discussion. These two steps were computed separately for CS, AI and BIO, obtaining in total three probability distributions and three contribution tables. We can observe the three probability distributions in Fig. 3.

TABLE II
CONTRIBUTION TABLE OF CS

Section	AI	BIO	CS
Abstract	0.1429	0.1906	0.1438
Introduction	0.5323	0.2627	0.5461
Methods	0.0679	0.0524	0.0730
Results	0.0929	0.1856	0.0898
Discussion	0.1639	0.3088	0.1473

After computing the probability distribution and the contribution table for each topic, we can now start to compose the

[1] Virtual Machine offered by Kaggle.

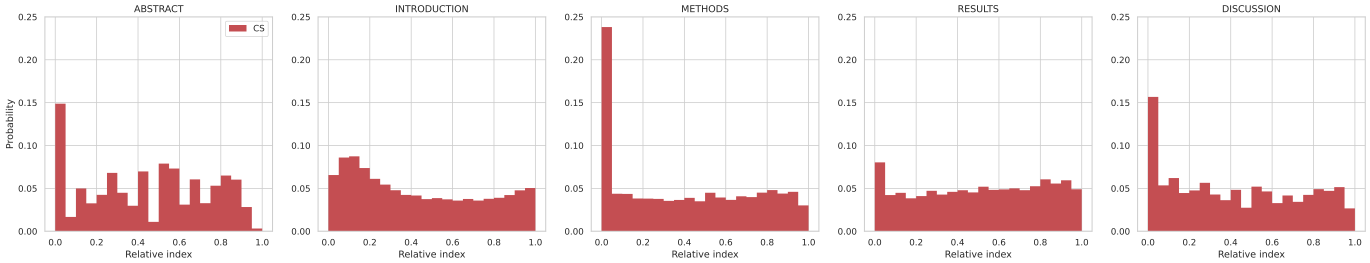

Fig. 1. Probability distribution of "high importance areas" against relative position inside the sections of CS

context by taking N sentences across sections. The parameter N can be fine-tuned to understand the optimal context length, we observed that the best results were obtained with $N = 15$. To obtain this parameter, we also tested the effect of varying the number K of selected highlights on the extraction performance (F1 score) and whether to use a separation token between sentences or just between sections. We obtained the same results of the authors which is $K = 3$. As we can see in Fig. 2, with $K = 3$ we have $N = 15$ and the best location to put a separation token results to be between each sentence. Further results of these experiments can be observed in Fig. 4.

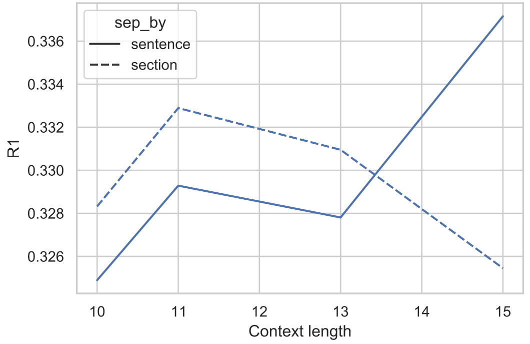

Fig. 2. Tuning of N and separation token location (F1 score)

These tests were conducted on AI for computing power and time reasons and were used also on CS and BIO. After fixing N, we use the contributions table c to understand the number of sentences that we will pick from each section and topic $n_{t,s} = [N \cdot c_{t,s}]$. Next we choose with replacement, according to the probability distribution $\hat{\theta}_{t,s}$, the same number of bins as the required number of sentences from each section. The final stage is to pick one sentence from each extracted bin to obtain the desired number of sentences to compose the context. If there is exactly one sentence in each bin and we require more sentences than the number of bins available, we take all sentences; on the other hand, if there are multiple sentences in each bin, we perform without replacement the *picking strategy*:

- Random: randomly pick a sentence inside the bin
- ROUGE2: rank each sentence inside the bin using ROUGE-2 F1 score with the abstract and pick the first
- Best: don't select the bin first, but directly select the $n_{t,s}$ best sentences from each section using ROUGE-2 F1 score against the abstract.

The second and third strategies aim at maximizing the informative content of the new context exploiting the section La Quatra and Cagliero found out to be the best in this task.

After all these steps, we finally have the context of the desired length. We then feed the new context into a pre-trained version of THExt available on HuggingFace for each topic: AI, BIO and CS. As we can observe, this extension brings us more flexibility in creating the context, as we have two parameters that can be fine-tuned, which are the *context size* and *picking strategy*.

To carry out these experiments we performed one epoch on each topic's dataset. We used 10^{-5} as learning rate, 10^{-2} as weight decay and 16 as *per device* batch size (32 in total). It took us around 8 hours of training for each topic (3) and each picking strategy (3). Another time consuming step is the creation of the dataset for the "best" and "ROUGE-2" strategies; we created a dataset for each topic with the contexts for each strategy for each paper. This was computed beforehand, so we could remain in the time constraints of Kaggle during training (12 hours). The duration of the datasets creation varies based on the topic but it took on average around 50 minutes per dataset per topic.

As final note, we can incorporate some prior knowledge or other techniques to manually change the distributions or the contributions (e.g., set abstract to 1 in c to reproduce THExt).

C. Direct extraction

To further prove our assumption of existence of high importance areas, we tried to extract highlights directly using the previously computed distribution and contribution table. We wanted to see if it can compete with more complex and time consuming methods. In order to do so, we extracted 15 bins as if we were creating a context with random picking strategy. Then we computed the ROUGE-2 F1 score of each selected sentence with the abstract to rank them. We tried to select $K = 1, .., 5$ sentences from the 15 to choose the best K and found that $K = 2$ is the best for CS, while $K = 3$ is the best for BIO and AI.

III. Results

A. Headline generation

We tested our trained models to assess their capabilities on all the datasets. Given the high quality of the extractive

TABLE III
F1 RESULTS IN HEADLINE GENERATION FOR ALL THE FINE-TUNED MODELS ON THE RESPECTIVE DATASET.

Model	HL	R-1	R-2	R-L	BERTScore
AI	Real	0.4332	0.2240	0.3607	0.9064
	THExt	0.4408	0.2378	0.3671	0.9075
BIO	Real	0.4580	0.2541	0.3961	0.9027
	THExt	0.4344	0.2455	0.3869	0.8992
CS	Real	0.5584	0.3818	0.5012	0.9233
	Thext	0.5178	0.3301	0.4545	0.9175

highlights produced by THExt we have also tested our models giving in input those highlights instead of the original ones. In Table III we can observe the results in terms of syntactic and semantic similarity.

The fine-tuned models are available on HuggingFace: Misc, AI, BIO, CS. To give the opportunity to test those models in an easy and effective way we set up an HuggingFace space. The model fine-tuned on AIPubSumm generated the title of this paper. In Table VII we provide a few examples of generated titles compared to the original ones.

B. Probabilistic Context Extraction

We tested the new context on all datasets to compare the results with THExt's. We must first notice that, due to updates to some core libraries, the results of THExt described in the paper can't be reproduced; the new results can be observed in Table IV. As we can see in Table IV, we outperformed THExt-Abstract in each topic and measure (except for AI in ROUGE2, where the difference is anyways little, -0.2%). Results on the three datasets are in the Appendix: Figures 5, 6 and 7. The fine-tuned models are available on HuggingFace: AI, BIO, CS. To give the opportunity to test those models in an easy and effective way we set up an HuggingFace space. In Tables VIII, IX and X we provide a few examples of highlights extracted by each model, in Table VIII we report also the highlights generated for this paper.

TABLE IV
THEXT PERFORMANCE COMPARISON BETWEEN THE ORIGINAL CONTEXT AND THE ONE OBTAINED USING PCE WITH K=3.

Dataset	Method	R-1	R-2	R-L
AI	THExt-Abstract	0.3350	**0.1253**	0.2998
	THExt + PCE-random	0.3353	0.1213	0.3056
	THExt + PCE-ROUGE2	0.3372	0.1178	0.3038
	THExt + PCE-best	**0.3415**	0.1250	**0.3111**
BIO	THExt-Abstract	0.3002	0.1017	0.2716
	THExt + PCE-random	0.3308	0.1195	0.3000
	THExt + PCE-ROUGE2	0.3326	0.1196	0.3025
	THExt + PCE-best	**0.3335**	**0.1222**	**0.3038**
CS	THExt-Abstract	0.3138	0.1204	0.2899
	THExt + PCE-random	0.3647	0.1485	0.3361
	THExt + PCE-ROUGE2	0.3676	0.1510	0.3390
	THExt + PCE-best	**0.3738**	**0.1613**	**0.3443**

C. Direct extraction

Results of direct extraction are impressive (especially R1 and RL ones), considering that it takes no training or inference time, but is just an extraction following a probability distribution with random picking strategy inside the bin. The results can be observed in Fig. 5, 6 and 7 compared to THExt-Abstract and THExt-PCE; while we report just the best ones in Table V. Further comparisons with other methods can be found in Table VI.

TABLE V
BEST RESULTS OF DIRECT EXTRACTION

Dataset	K	R-1	R-2	R-L
AI	3	0.2850	0.0863	0.2633
BIO	3	0.2738	0.0805	0.2565
CS	2	0.2940	0.0875	0.2653

IV. DISCUSSION

A. Headline generation

The first thing that comes to mind after observing the examples provided in the Results section, is that AI, CS and BIO results are different, in fact the model learnt that in the different categories titles are given in different ways (e.g., in AI the name of the method is often included in the title).

The final observation considers the outcomes obtained using the highlights extracted by THExt. These results are impressive given that the highlights used are extractive, which could potentially be considered a limitation.

We noticed a clear limitation of the model dealing with our paper, completely ignoring one of the two task we carried out. We imagine this can be due to the fact that the presence of multiple tasks is not a typical scenario, at least in the available training sets.

B. Probabilistic context extraction

As we were able to observe in the results, our method, especially the ”best” picking strategy, outperforms the reference paper, which is currently the state of the art. We can assume that at least a part of the improvements are due to performing an additional epoch on each dataset with respect to the pre-trained models available on HuggingFace. We think that an even better context than the abstract is feasible, so more information about the paper can be incorporated. We also must notice that with a deeper hyper-parameter tuning, which we didn't perform for time reasons, we could possibly obtain even better results.

Another possible improvement could be achieved by performing a deeper cleaning on text, for example by removing or standardizing references to figures, tables or citations (e.g., [2] becomes just ”CITATION” or Fig./Figure all become ”FIGURE”). We can also decode Unicode characters that are left in the sentences (e.g., ”\u00a0”).

C. Direct extraction

We saw that direct extraction performs surprisingly well on the task and we can think that results can be further improved by using the ”best” picking strategy instead of random and by tuning the number of sentences considered. This could take a little more time but results may be better.

REFERENCES

[1] Iz Beltagy, Arman Cohan, and Kyle Lo. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676, 2019.

[2] Luca Cagliero and Moreno La Quatra. Extracting highlights of scientific articles: A supervised summarization approach. *Expert Systems with Applications*, 160:113659, 2020.

[3] Ed Collins, Isabelle Augenstein, and Sebastian Riedel. A supervised approach to extractive summarisation of scientific papers. *CoRR*, abs/1706.03946, 2017.

[4] Ed Collins, Isabelle Augenstein, and Sebastian Riedel. A supervised approach to extractive summarisation of scientific papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 195–205, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[6] Moreno La Quatra and Luca Cagliero. Transformer-based highlights extraction from scientific papers. *Knowledge-Based Systems*, 252:109382, 2022.

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019.

[8] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[9] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders, 2019.

[10] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018.

[11] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368, 2017.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[13] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019.

APPENDIX

TABLE VI
DIRECT EXTRACTION WITH RANDOM STRATEGY PERFORMANCE COMPARISON WITH OTHER METHODS

Dataset	Method	R-1	R-2	R-L
AI	Liu and Lapata [9]	0.266	0.059	0.238
	Collins et al. [4]	0.279	0.069	0.235
	THExt + Abstract	0.3350	0.1253	0.2998
	Direct extraction (K=3)	0.2850	0.0863	0.2633
BIO	Liu and Lapata [9]	0.249	0.059	0.224
	Collins et al. [4]	0.287	0.087	0.243
	THExt + Abstract	0.3002	0.1017	0.2716
	Direct extraction (K=3)	0.2738	0.0805	0.2565
CS	Liu and Lapata [9]	0.252	0.058	0.228
	Collins et al. [4]	0.339	0.127	0.295
	THExt + Abstract	0.3138	0.1204	0.2899
	Direct extraction (K=2)	0.2940	0.0875	0.2653

TABLE VII
EXAMPLES OF PAPER ORIGINAL TITLES AND GENERATED TITLES FOR EACH TOPIC

	Model trained on MiscPubSumm
Original	Transformer-based highlights extraction from scientific papers
Generated	A Transformer-based Highlights Extractor (THExt)
	Model trained on AIPubSumm
Original	Formal verification of ethical plan selection in autonomous systems
Generated	Formal verification of ethical choices in autonomous systems
Original	A bilevel approach to enhance prefixed traffic signal optimization
Generated	A bilevel approach to determine optimal number of traffic signal optimization
	Model trained on BIOPubSumm
Original	Calcifying fibrous tumor and inflammatory myofibroblastic tumor are epigenetically related: A comparative genome-wide methylation study
Generated	Molecular characterization of calcifying fibrous tumor and inflammatory myofibroblastic tumor
Original	Reparative effects of interleukin-1 receptor antagonist in young and aged/co-morbid rodents after cerebral ischemia
Generated	Interleukin-1 receptor antagonist (IL-1Ra) promotes neurogenesis after experimental stroke in young and aged rats
	Model trained on CSPubSumm
Original	Muscle fatigue based evaluation of bicycle design
Generated	Evaluation of on-road bicycle design using surface electromyography
Original	Optimal protruding node length of bicycle seats determined using cycling postures and subjective ratings
Generated	Effects of protruding node length on body posture, subjective discomfort, and stability during cycling

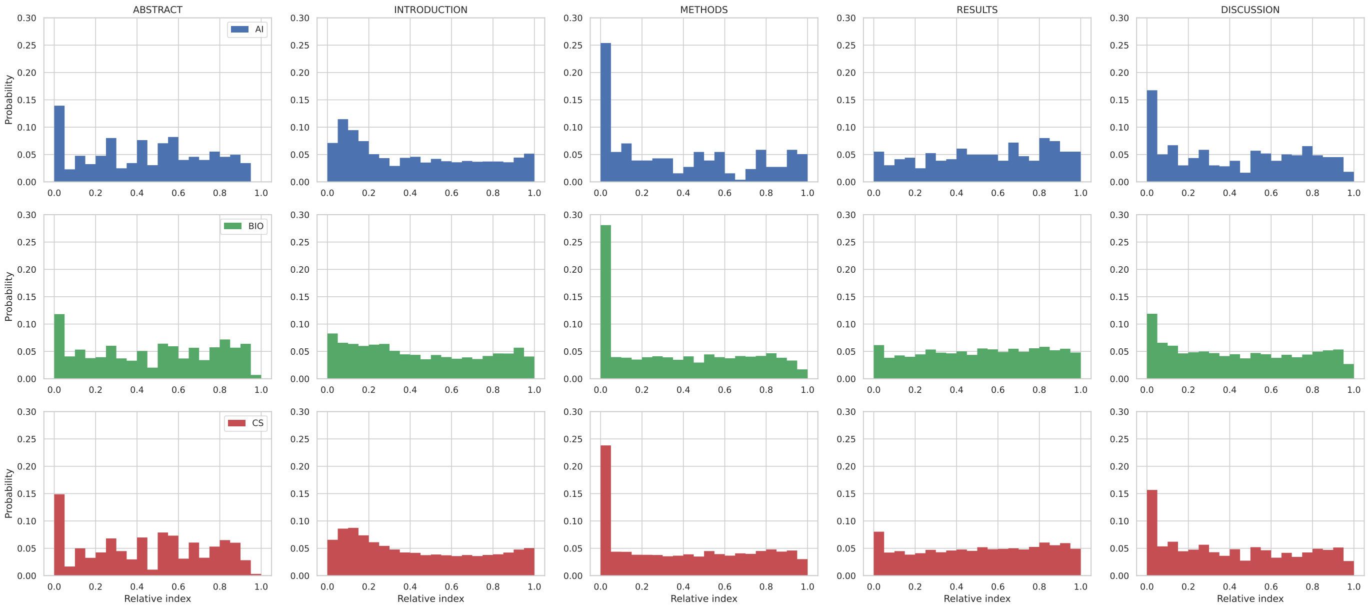

Fig. 3. Probability distribution of ”high importance areas” against relative position inside the sections for each topic

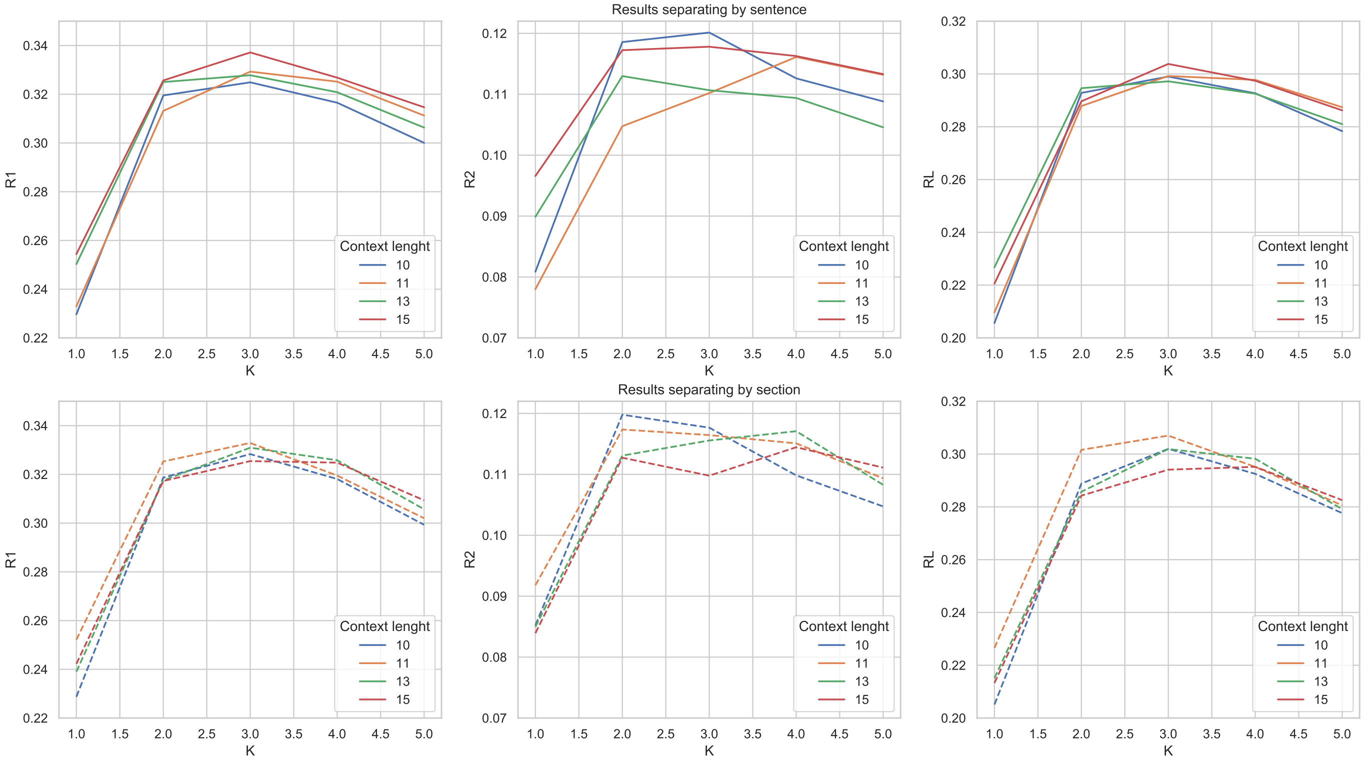

Fig. 4. Tuning K, length of context N and the separation method (F1 score)

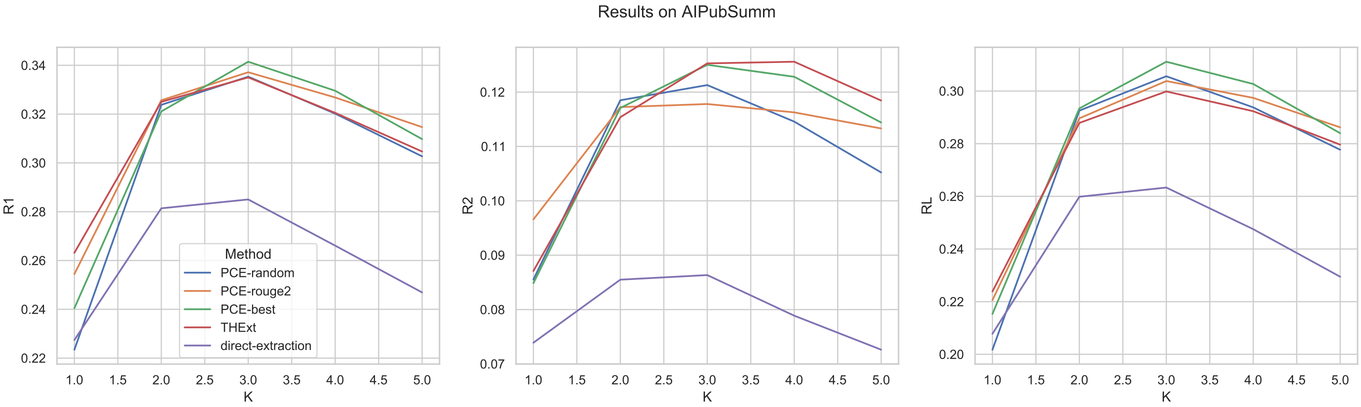

Fig. 5. Results of THExt+PCE and THExt+Abstract on AI (F1 score)

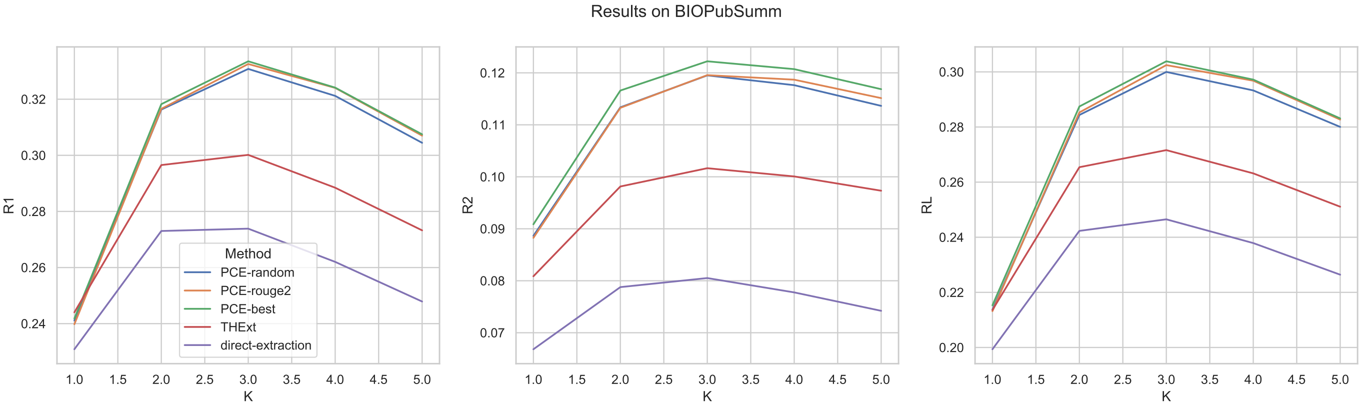

Fig. 6. Results of THExt+PCE and THExt+Abstract on BIO (F1 score)

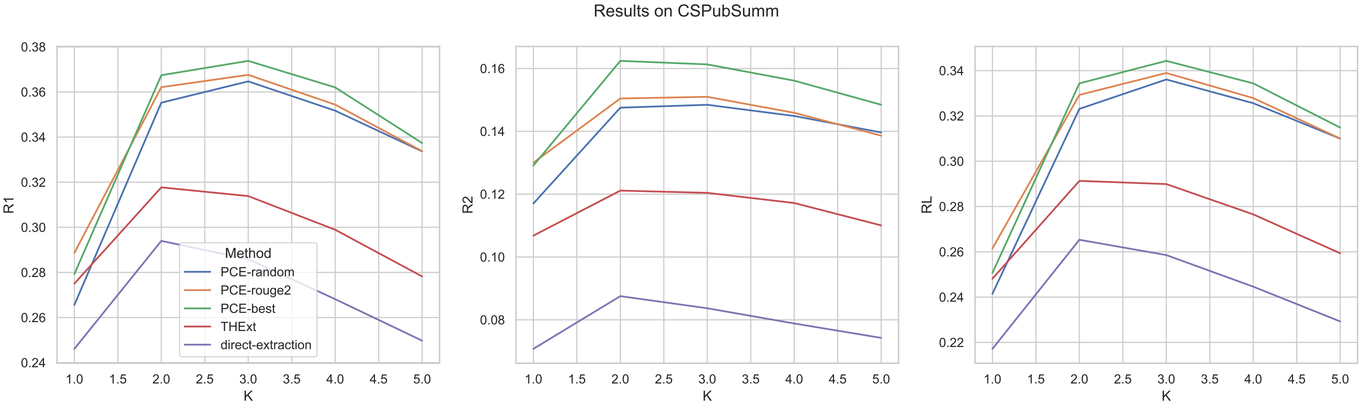

Fig. 7. Results of THExt+PCE and THExt+Abstract on CS (F1 score)

TABLE VIII
Examples of extracted highlights comparing PCEs variants and THExt on AIPubSumm.

Model trained on AIPubSumm	
Human-generated	• An autonomous system should act ethically, but what if it has no all-ethical choice? • We model how to rank states violating multiple instances of ethical principles. • We enable an autonomous system to use this ethic rank to rank its available plans. • We guarantee that when a plan is chosen, it is the most ethical plan available.
THExt-Abstract	• In Section 3 we outline our formal theoretical framework for the implementation and verification of ethically constrained behaviour in autonomous systems and also point to some relationships of our framework to deontic logic. • The main contribution of our paper is a verifiable ethical decision-making framework that implements a specified ethical decision policy. • We propose a method for, and have implemented a working prototype of, an ethical extension to a rational agent governing an unmanned aircraft (UA).
PCE-random	• This work on model checking ethical choices is preliminary. • Model checking takes place relative to some requirement specified in a formal language [22]. • In general model checking systems cannot handle this kind of quantification.
PCE-Rouge2	• In Section 2 we cover relevant background material on autonomous systems, machine ethics and verification. • We propose a method for, and have implemented a working prototype of, an ethical extension to a rational agent governing an unmanned aircraft (UA). • One of the reasons for selecting Gwendolen as the basis for our implementation language, Ethan, was that it provided the potential for formally verifying ethical decision-making.
PCE-best	• We propose a method for, and have implemented a working prototype of, an ethical extension to a rational agent governing an unmanned aircraft (UA). • Model checking has been used in [20] for providing formal evidence for the certification of autonomous unmanned aircraft. • The main contribution of our paper is a verifiable ethical decision-making framework that implements a specified ethical decision policy.
This paper	• We obtained results comparable to the state of the art. • An important observation is that our proposed context is not static, unlike the abstract, thus we could perform hyper-parameter tuning to possibly enhance its performance. • This project focuses on text summarization, since we were asked to propose two extensions of La Quatra, Cagliero the state of the art of highlights extraction by extractive summarization.

TABLE IX
Examples of extracted highlights comparing PCEs variants and THExt on BIOPubSumm.

Model trained on BIOPubSumm	
Human-generated	• It has been hypothesized that calcifying fibrous tumoris the late regressive stage of inflammatory myofibroblastic tumor. • By genome-wide methylation assay we could provide evidence that these lesions are a spectrum of one entity. • The well-known fusion genes ALK, ROS1 and RET, a hallmark of IMT, were not find in our CFT.
THExt-Abstract	• The most convincing argument that CFT and IMT form a spectrum with CFT being the burned out end we found by methylation profiling showing overlapping methylation patterns. • The control group consisted of IMT (n=7), leiomyoma (n=7), angioleiomyoma (n=9), myopericytoma (n=7) and reactive soft tissue lesions (n=10). • Four of the five CFTs formed a homogeneous methylation group with the IMTs by clustering and t-SNE analysis (Fig. 2), which remained stable when varying the number of CpGs (Cytosine-phosphatidyl-Guanine) used.
PCE-random	• Calcifying fibrous tumor (CFT) is a rare benign mesenchymal lesion occurring in both children and (young) adults [1-5]. • Calcifying fibrous tumors (CFT) were first described in 1988 by Rosenthal and Abdul-Karim under the name ”childhood fibrous tumor with psammoma bodies”. • Furthermore, we performed FISH for ALK, ROS1 and RET on CFTs as rearrangements of these genes are the genetic hallmark of IMT.
PCE-Rouge2	• Calcifying fibrous tumor (CFT) is a rare benign mesenchymal lesion occurring in both children and (young) adults [1-5]. • Later on, these lesions were considered neoplastic and designated as calcifying fibrous tumor [6]. • Calcifying fibrous tumors (CFT) were first described in 1988 by Rosenthal and Abdul-Karim under the name ”childhood fibrous tumor with psammoma bodies”.
PCE-best	• The most convincing argument that CFT and IMT form a spectrum with CFT being the burned out end we found by methylation profiling showing overlapping methylation patterns. • This study investigated the relation between CFT and IMT by comparing the genome-wide methylation patterns of both CFT and IMT alongside myopericytoma, angioleiomyoma, leiomyoma and reactive soft tissue lesions. • Leiomyoma and reactive lesions clustered separately.

TABLE X

Examples of extracted highlights comparing PCEs variants and THExt on CSPubSumm.

	Model trained on CSPubSumm
Human-generated	• Evaluation of on-road bicycle design was performed using surface EMG on 12 male volunteers. • Three types of bicycle design, i.e., rigid frame, suspension and sports were studied. • Bicycles with suspension were found to have lesser rider muscle fatigue.
THExt-Abstract	• This study used percentage deviation in MPF and RMS to significantly differentiate the muscle activity before and after 30 min of cycling; slope of MPF and RMS to significantly evaluate the fatigue rate during on-road cycling on the three different bicycle designs. • These significant differences among the three bicycle designs have been demonstrated by means of MVC test before and after cycling and real-time monitoring of muscle activity during on-road cycling. • The results suggested that there is higher muscle fatigue in low back pain group when compared to their cohorts.
PCE-random	• The results suggested that there is higher muscle fatigue in low back pain group when compared to their cohorts. • All the participants in the study being right handed, the results eventually expressed a significantly higher fatigue in the right side when compared to the left. • The results also suggest that the SU bicycle shows evidence of how well the suspension prevents muscular fatigue and vibration-induced low-back pain.
PCE-Rouge2	• The results suggested that there is higher muscle fatigue in low back pain group when compared to their cohorts. • This study used percentage deviation in MPF and RMS to significantly differentiate the muscle activity before and after 30min of cycling; slope of MPF and RMS to significantly evaluate the fatigue rate during on-road cycling on the three different bicycle designs. • The percentage deviations extracted from MVC test and the slopes of RMS and MPF of sEMG signals extracted from real-time monitoring during on-road cycling were statistically analysed.
PCE-best	• Changes in muscle activity during on-road cycling were quantified each time a participant rode a bicycle. • This study used percentage deviation in MPF and RMS to significantly differentiate the muscle activity before and after 30 min of cycling; slope of MPF and RMS to significantly evaluate the fatigue rate during on-road cycling on the three different bicycle designs. • These significant differences among the three bicycle designs have been demonstrated by means of MVC test before and after cycling and real-time monitoring of muscle activity during on-road cycling.