NUANS Homework 2: Extractive Summarization

Leonardo Lavalle
1838492
Sapienza, University of Rome
lavalle.1838492@studenti.uniroma1.it

1 Introduction

Automatic Text Summarization is an NLP process that focuses on reducing the amount of text from a given input while at the same time preserving key information and contextual meaning. There are two types of Automatic Text Summarization: Abstractive Summarization and Extractive Summarization. We'll deal with the Extractive one, which consists in selecting a subset of sentences from the text to form the summary. It is suited to cases where key sentences hold most of the text's information, such as news articles and not very appropriate to cases where information is thinly spread over the text (that can be the case of narrative stories/books). In fact, it'll be later discussed how it's not trivial at all to perform an extractive summary on our FairySum dataset. The stories assigned to me and on which

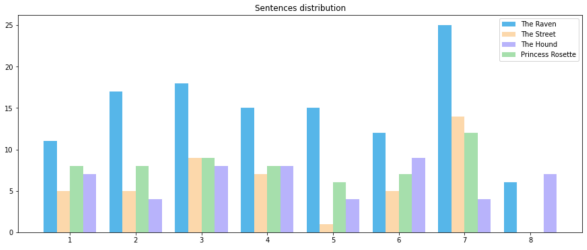

Figure 1: Sentences Distribution

I had to manually perform the extractive summarization were: *"The Street"*, *"The Raven"*, *"The Hound"* and *"Princesse Rosette"*. It's interesting to see for each story, how many sentences were selected and how much it was possible to shrink the summaries. The percentage reported in Table 1 indicates the portion of the original text kept.

As we can see in **Table 1**, the first three stories were reduced (more or less) by the same amount of percentage reduction, in contrast to the last one in which there were eliminated a larger portion of text. *But why?* This because the last story has an higher original text length, that means is a longer story than the others. This is rather intuitive: shorter stories, where the meaning is more condensate in fewer lines, are more difficult to skim. Most of their sentences are important for letting the reader understand the semantics of the plot and as a result they cannot be discarded.

Story	Text	Summ	%
The Raven	80	58	72.5%
The Hound	104	46	44.2%
The Street	89	51	57.3%
Princess Rosette	321	119	37%

Table 1: Percentage of sentences kept

Another interesting thing that can be analyzed is the distribution of the selected sentences, i.e., in which part of the texts the "most salient" sentences (the ones that we keep to produce the extractive summary) are present. This aspect is very important for a Deep Learning Model point of view, in particular for a Transformer one. In fact, if we consider the news domain, in which the salient information can be found in the first few sentences, summarization models will learn to pay attention only to the news introduction while ignoring the rest. In **Figure 1** we can see the results of this analysis on my assigned texts. The distributions are mostly constant, we don't have particular peacks, meaning that in the narrative domains the models don't need to focus on certain sections of the text. Perhaps it can be said that in subportion number 7 there is a tendency for salient sentences to be grouped together, denotating that towards the end of the story important events and plot twists occur. It should be also noted that in the very last section (number 8) of all texts very few sentences are picked. Another aspect emerges from the inspection of the distribution: *The Raven* story is the one

in which most of the original text were kept. That's because it is very dense with salient events and during the manually annotation it was very difficult to dispose of most sentences. As we'll discuss later, the output length of the summary is an open problem and find a way to compute the amount of salient sentences in a document is something very important to calculate it.

2 Issues and challenges during manual annotation

First of all I wanted to express my surprise in acknowledge the fact that performing an extractive summary by hand is not easy at all as I expected. Making a fluent summary of this type and at the same time shorter as possible without excluding the main events of a story is very difficult, also for humans and this was at the beginning quite daunting but at the same time exciting and inspiring. Stimulating because is still an unsolved and an on-going research problem, and the challenges to address are many. Since abstractive summaries require natural language generation and semantic representation are in general easier tasks for a human and in my opinion are more suitable for the narrative stories domain instead of the extractive ones. The latter instead fit better other type of texts such as news articles, scientific papers etc.

The main adopted technique was based on being generous in the sentences selection at first and then slowly skim them. More or less I did four selection phases of this kind for each story. The thing that helped me the most, without any doubts, was the possibility of having a brief **abstractive summary** for each story at hand to help with the selection of the most important sentences which denoted the main happenings within the plot. Indeed, one way that helped me a lot do summaries was to divide the abstractive summary into semantic blocks or blocks that encapsulate a macro event and select sentences in the original text that represent that event. And during the selection, to understand if a sentence was worth to mantain it, I had to keep attention if selected sentences were representing or not the same concept (semantically) of already selected ones. But let's now dive into the issues and challenges I had to face during the manual summarization:

- most of the stories weren't written in "common" English, or at any rate the English that we are used to reading at university, but it was literary, a sort of an archaic one.

- things got complicated when a salient concept/event were expressed in the story with direct speeches. Most of the time I was obligated to keep many lines because they were part of a big talk between characters. Moreover, many times these speeches were also not well tokenized.

- another problem was when an event is expressed by a sequence (also repeated) of minors happenings. The latter have to be included otherwise we lose the sense of the narrative and so was not very easy and clear how to make the right decisions. Think about *The Raven* when the young has to wait outside the house of the old lady three times or in *Princess Rosette* when the dog goes multiple times to steal from the king's cuisine. Instead the shorter and the standalone sentences, since they have no connections with the others are easier to be discarded.

- another difficulty depended on the type of texts. The narrative and more dynamic ones (*The Raven* and *Princess Rosette* for example) are easier while the descriptive and more discursive (*The Street*) are more difficult to summarize. If the salient events are not so many, what is the measure by which I adjust to make a descriptive text shorter? This is a direction that can be followed for deepen the issue.

3 Model

The choosen model for the task is Sentence-BERT (**SBERT**) [3], a modification of the BERT network using siamese and triplet networks that is able to derive semantically meaningful sentence embeddings. Extractive summary relies on the concept of sentence salience to identify the most important sentences in a document. *How do we select the most important sentences?* We use LexRank [1], a stochastic graph-based method for computing relative importance of sentences among a text.

3.1 Length of the Summary

The length of the summary we're going to produce remains an open question. It's an hyperparameter that has to be thoughtfully choosen and there is

no right answer to that. The ideal length depends on the context in which it'll be used, as well as the preferences of the people that will be reading the summary. In general, an extractive summary should be concise and to the point, but it should also include enough information to convey the main ideas of the original text. In our case, since we're dealing with short stories and fairy tales, it seemed right to decide the length of each summaries depending on the original text's length. As we've already discussed, by analyzing the summaries it's very clear that the longest the story the greater the decrease.

For this reason we won't select a *fixed* percentage of sentences, but a *variable* one which depends on each story's length. There can be many ways of doing that, but as a baseline we'll use the most simple (and probably less accurate) approach it can be thinked: (i) we set a *threshold*. (ii) if the length of a story is above it, we select the x percentage of its sentences. (iii) else, we select the y percentage. After few (and fast) trials it looked reasonable to keep the *30%* of sentences from stories which exceed the given threshold and the *60%* from the others in order to produce the extractive summaries. The choosen threshold was of 150 sentences.

3.2 Limitations and future improvements

Main approaches and SOTA models performed extractive summary by encode (usually with BERT-like architectures) and score sentences one by one from the original text, and then select the ones with an higher score to form a summary. This is exactly what SBERT does and is rather intuitive to understand that this is not an ideal way of proceeding if we want to obtain good summaries. Most of the time they don't make any sense if read consecutively. They lack a mechanism to ensure overall *coherence*. It's not necessarily true that the "best" sentences (where best are the ones with the highest score and the score is computed without seeing the context and the other sentences) will form the "best" summary. In fact previous works have pointed out the weakness of sentence-level extractors and in [2] was made a systematic analysis in this direction.

We can say that the actual extractive methods have a lot of *limits*. For example they don't take into account the relationships between adiacent sentences and it's difficult to understand which is the degree of summarization that we want to achieve. That's not easy even for humans: if you have a text you cannot easily answer which it'll be the length of its summary and how much it'll be decreased. As we talk through earlier there's no right answer to that, but for sure the answer is not having a fixed output length (as *SBERT* does).

I think the best approach for extrapolate summary length would be to consider and quantify three quantities: *original text length*, *typology* of the text (descriptive, discursive, narrative etc.) and the *concentration of important events* in the story (like the case of *The Raven* that is highly event condensed). One can learn the relation between these quantities and the ground-truth summaries (created by hand by the students of this course) length of the FairySum dataset by using advanced machine learning and regression techniques.

For a possible method of extractive summarization one idea may be to link phrases that refer to the same entities in such a way as to prevent one phrase from referring to another that has been deleted. As another option it would be interesting to find a mechanism for *coreference* so that sentences with pronouns referring to deleted sentences are not left separated (this happened frequently). This to avoid one of the many issues I experienced during the manual annotations: when I was selecting a salient sentence, very often it had a pronoun "hanging" on it, and so I was forced to also take the preceding (not salient) sentences that made it clear what the pronoun referred to.

Being the aspect that helped me the most, I'm sure that the employment of the *abstractive summary* (computed with a SOTA model) in someway (also as an additional input) would be beneficial to the performance of overall system. It could be leveraged by highlighting parts of the original text that are also expressed in it in order to engage a *inter-attention* mechanism between the abstractive summary and the text. I'll explore this improving performance direction in the mini project because I'm very trustful that somehow it'll help in extracting better sentences.

4 HUMAN-HUMAN comparison

We received the other summaries executed from the other students. Specifically I got two more for *The Hound*, two more for *The Street*, one more for *Princess Rosette*, and no one for *The Raven* (I won't make obviously any comparison of it). The difference that immediately stands out is their length. As we discussed earlier it is a variable that

can be influenced by many factors like the type of the text or simply the different perception that a human has about how much a summary has to be shrinked to be considered as such. In other words: it depends on the individual and dissimilar degree of summarization that each person wants to achieve.

Let's talk first about *Princess Rosette*: being a long story and full of meaningful events, the extraction was not the hardest one because the main happenings were easy to detect. In fact the other summary is not very different form mine and also the lengths are similar: 119 sentences mine and 91 the other.

The scenario completely changed with *The Street*: a very descriptive story, full of metaphors and historical references most of the times unclear. My summary is of 51 lines and the other two respectively of 22 and 21. I thought that maybe I did something wrong, that I didn't performed the task properly. But that is precisely the task: understand how humans make extractive summaries and why the keep more sentences or less. In general, descriptive texts leave ample room for choice and this happened here. I read the other two summaries and are not even that bad as I thought, but I think they overlooked one aspect of the story that I considered central for the short story: the fact that the street has a soul and that after all the buildings around the street collapsed, roses and trees blossomed. So I wanted to the keep track of all their references along the story and maybe for this reason I produced a longer summary.

On the other hand, for *The Hound* we have one summary that is very similar to mine in terms of selected sentences and number of them: mine is 46 and that one is 52 lines long. Instead the other provided summary has only 21 selected phrases. It's objectively a bad summary because it's not fluent at all, sentences are badly connected and its difficult to understand the plot of the story. The original text has some descriptive parts and of course different people can extract different salient sentences depending on their feelings, but in my opinion in this case it wasn't done a great job.

5 HUMAN-SYSTEM comparison

For evaluating quantitatively the summary extraction system we'll use the **ROUGE** metric. In particular the *F1 score* for ROUGE-1, ROUGE-2 and ROUGE-L. The metrics compare an automatically produced summary against a reference (human-produced) summary. In this case the reference summaries are composed by mine and by the ones produced by other students. In the cases in which the references summaries to a certain text are more than one, a simple average of the output measure will be computed. We are de facto considering all the students extractive summaries equally.

	ROUGE		
Story	**R-1**	**R-2**	**R-L**
The Raven	0.88	0.83	0.83
The Hound	0.67	0.60	0.62
The Street	0.65	0.56	0.57
Princess Rosette	0.78	0.55	0.53

Table 2: ROUGE results

In **Table 2** we can see the results. The performance of the system, depending on the ROUGE metric, it seems to work quite well. The ROUGE-L, which tends to capture sentence structure more accurately than the other two because it considers the longest common subsequence, shows good results, this is clear. But it must be said that ROUGE is an intrinsic evaluation measure, so it doesn't consider the *semantics* of summaries. This lack can be seen in the fact that *The Raven* has the highest scores in all the three metrics; precisely the story that has the lowest sentences reduction ratio. Having a reference summary with many sentences of the original text, it'll inevitably lead to an increasement of ROUGE (just look at its formula). So in the future it'll be fundamental, for a more reasonable evaluation of this task, to find new evaluation metrics which take into account the semantics of the texts.

References

[1] Dragomir R. Radev Gunes Erkan. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization.

[2] Yiran Chen Danqing Wang Xipeng Qiu Xuanjing Huang Ming Zhong, Pengfei Liu. 2020. Extractive summarization as text matching.

[3] Iryna Gurevych Nils Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.