UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Ph.D. Thesis

Delving into the Abstractivity of Summaries

by Vicent Ahuir Esteve

Advisors

Professor Dr.
Lluís Felip Hurtado Oliver

Professor Dr.
Encarna Segarra Soriano

Thesis submitted to the Universitat Politècnica de València (UPV) as partial fulfillment of the requirements for the Ph.D. degree in Computer Sciences

❖ January 14, 2025 ❖

The difference between a helpful summary and a vague one is the difference between shaping the message to serve its purpose and simply trimming it down to fit.

Paraphrasing to Mark Twain.

When I decided to return to the university path, it was an exciting and clear choice; I needed to enter that thoughtful and moving environment. But also, it was a challenging and scary decision related to the "what will happen". However, everything changed when I started to meet interesting people, including colleagues and professors, equally; that was one of the most relevant and engaging parts of this journey. Along the way, I have met many people who have made this learning process extremely interesting, inviting me to continue giving my best and learning something new daily. Meeting my thesis mentors, Lluís and Encarna, and the rest of the members of the ELiRF research group changed my life and perception of my place in research. I went from looking at research as something out of my reach to loving this world of scientific research, especially the field of Natural Language Processing. This text would never have happened without those long conversations where you stop by to say good morning and, after several hours, we almost overflow the stack of open topics at once during the same conversation. So, what looked like a long road has come to this delightful end, and now there is nothing left to do but walk an even more exciting one. As a wise man says, "One less thing to do to do something else".

Last but not least, the people who have been, are, and will always be there for me. I could not have made this journey without the love, help, and support of my entire family and my partner in life JC®; I could not be more fortunate to have you all.

Thank you so much to everyone; I literally could not have done it without you.

What a journey. 

Acknowledgements

- This work was partially supported by the Generalitat Valenciana under project CIPROM/2021/023, and by the Spanish Ministerio de Universidades under the grant FPU21/05288 for university teacher training.
- It was also partially supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-01-21).
- It was also partially supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-11-21).
- It was also partially supported by MCIN/AEI/10.13039/501100011033, by the "European Union and "NextGenerationEU/MRR", and by "ERDF A way of making Europe" under grants PDC2021-120846-C44 and PID2021-126061OB-C41.
- It was also partially supported by the Generalitat Valenciana under project CIPROM/2021/023.

Abstract

With the exorbitant amount of content continuously generated by today's technological society, to which one has access almost instantaneously, a person can spend excessive time searching and filtering content to perform a task. Therefore, being able to automatically identify, extract, and condense the most relevant information from any document is vital. In Natural Language Processing (NLP) this problem is addressed by the automatic summary generation or summarization field.

Advances in the design of neural network architectures, most notably the Transformers architecture, have significantly increased the capabilities of language models. These improvements have enabled language models to generate more coherent texts. Therefore, in the automatic summarization field, the scientific community has begun to create summarization models that adopt an abstractive strategy. With the abstractive strategy, we obtain summaries where the summary is written by identifying the most relevant information of the document and reinterpreting the text of the original document to obtain a coherent, informative, and concise summary as possible. Therefore, the abstractive summarization strategy brings the models closer to how human beings summarize content.

This work delves into different challenges within the abstractive strategy of summarization, focusing our efforts on the composition or writing process of summaries. The thesis aims to define in more detail what characterizes the perception of abstractivity in summaries and whether there are grades of variation in that perception, from complete extractive summaries to purely abstract summaries. We also explore ways to determine when two summaries are valid for the exact text despite having notably different wording and structure. Additionally, we study the emotions that arise from certain words in summaries, how the emotions vary from the document to the summary, and how automatic summarization models are influenced by this aspect.

Resumen

Con la desorbitada cantidad de contenidos que genera continuamente la sociedad tecnológica actual, a los que se tiene acceso de forma casi instantánea, una persona puede dedicar un tiempo excesivo a buscar y filtrar contenidos para realizar una tarea. Por ello, ser capaz de identificar, extraer y condensar automáticamente la información más relevante de cualquier documento es vital. En el Procesamiento del Lenguaje Natural (PLN), este problema se aborda en el campo de la generación automática de resúmenes o *summarization*.

Los avances en el diseño de arquitecturas de redes neuronales, sobre todo la arquitectura Transformers, han aumentado considerablemente las capacidades de los modelos lingüísticos. Estas mejoras han permitido a los modelos lingüísticos generar textos más coherentes. Por ello, en el campo del resumen automático, la comunidad científica ha empezado a crear modelos de resumen que adoptan una estrategia abstractiva. Con la estrategia abstractiva se obtienen resúmenes en los que el resumen se redacta identificando la información más relevante del documento y reinterpretando el texto del documento original para obtener un resumen lo más coherente, informativo y conciso posible. Por lo tanto, la estrategia de resumen abstractivo acerca los modelos a la forma en que los seres humanos resumen los contenidos.

Este trabajo profundiza en diferentes retos dentro de la estrategia abstractiva de resumen, centrando nuestros esfuerzos en el proceso de composición o redacción de resúmenes. La tesis pretende definir con más detalle qué caracteriza la percepción de abstractividad en los resúmenes y si existen grados de variación en dicha percepción, desde resúmenes extractivos completos hasta resúmenes puramente abstractos. También exploramos formas de determinar cuándo dos resúmenes son válidos para el mismo texto a pesar de tener una redacción y una estructura notablemente diferentes. Además, estudiamos las emociones que suscitan ciertas palabras en los resúmenes, cómo varían las emociones del documento al resumen y cómo influyen este aspecto en los modelos de resumen automático.

Resum

Amb la desorbitada quantitat de continguts que genera contínuament la societat tecnològica actual, als quals es té accés de forma quasi instantània, una persona pot dedicar un temps excessiu a buscar i filtrar continguts per a fer una tasca. Per això, ser capaç d'identificar, extraure i condensar automàticament la informació més rellevant de qualsevol document és vital. En el Processament del Llenguatge Natural (PLN), este problema s'aborda en el camp de la generació automàtica de resums o *summarization*.

Els avanços en el disseny d'arquitectures de xarxes neuronals, sobretot l'arquitectura Transformers, han augmentat considerablement les capacitats dels models lingüístics. Estes millores han permés als models lingüístics generar textos més coherents. Per això, en el camp del resum automàtic, la comunitat científica ha començat a crear models de resum que adopten una estratègia abstractiva. Amb l'estratègia abstractiva s'obtenen resums en els quals el resum es redacta identificant la informació més rellevant del document i reinterpretant el text del document original per a obtindre un resum el més coherent, informatiu i concís possible. Per tant, l'estratègia de resum abstractiva acosta els models a la forma en què els éssers humans resumixen els continguts.

Este treball aprofundix en diferents reptes dins de l'estratègia abstractiva de resum, centrant els nostres esforços en el procés de composició o redacció de resums. La tesi pretén definir amb més detall què caracteritza la percepció de abstractivitat en els resums i si existixen graus de variació en esta percepció, des de resums extractius complets fins a resums purament abstractius. També explorem maneres de determinar quan dos resums són vàlids per al mateix text malgrat tindre una redacció i una estructura notablement diferents. A més, estudiem les emocions que susciten unes certes paraules en els resums, com varien les emocions del document al resum i com influïxen este aspecte en els models de resum automàtic.

Contents

List of Figures

List of Tables

Glossary

Abstractivity-p Measure that quantifies the degree of abstraction in a summary by evaluating how much of the summary's content is derived directly from the original document versus rephrased or newly introduced. The parameter p adjusts the weight given to longer fragments. With $p = 1$, all fragments –independently of their length– penalize equally the degree of abstraction. With $p > 1$, the presence of longer fragments penalizes more the degree of abstraction.

BERTScore It is a metric used to evaluate the quality of text generation by comparing the similarity between the generated text and reference text based on contextual word embeddings from BERT (Bidirectional Encoder Representations from Transformers). It measures the similarity between words in the generated and reference text by considering their contextualized meanings rather than simple token matching, providing a more nuanced evaluation of semantic similarity. BERTScore computes precision, recall, and F1 scores at the token level, based on cosine similarity between the word embeddings.

BLEU (Bilingual Evaluation Understudy) It is a metric used to evaluate the quality of machine-generated text, particularly in machine translation. It measures how closely the generated text matches one or more reference texts written by humans. The metric is calculated by comparing n-grams in the generated text with those in the reference text, penalizing outputs that are too short with a brevity penalty.

Cohen's Kappa It is a statistical measure used to assess the agreement between two raters or classifiers, while correcting for the possibility of agreement occurring by chance. It is commonly used in situations where two individuals or systems independently classify items into categories and helps to assess how much they agree beyond random chance.

Coleman-Liau Index (CLI) It is a readability test designed to assess the grade level of a text based on its sentence length and character count.

Compression It is the ratio of the length of the summary to the length of the original text. A lower ratio indicates higher compression, meaning the summary is much shorter relative to the original.

Coverage Measure that quantifies the extent to which a summary is derivative of a text. Coverage measures the percentage of words in the summary that are part of an extractive fragment with the article.

Dale-Chall Readability Score (DCRS) It is a readability score based on the complexity of its vocabulary and sentence length. It measures how easily a text can be understood, with a focus on whether the words used are familiar to readers.

Density Measure that quantifies how well the word sequence of a summary can be described as a series of extractions. Density is the squared average length of the extractive fragment to which each word in the summary belongs.

Factuality The accuracy and truthfulness of the information presented in the summary compared to the original source. It aims to ensure that the summarized content accurately represents the key facts, ideas, and details from the original text, without distortion, misrepresentation, or omission of essential information.

Fleiss' Kappa It is a statistical measure used to assess the reliability or agreement among multiple raters or assessors (more than two), who are classifying items into categories. It extends Cohen's Kappa, which is used for two raters, to handle situations where there are more than two raters, making it suitable for situations like group evaluations or multi-classifier systems.

Flesch-Kincaid Grade Level (FKGL) It is a readability test designed to assess the complexity of a text based on its sentence length and word length. It estimates the U.S. school grade level required to understand the text. The higher the FKGL score, the more advanced the text, which means that it may require a higher level of education to fully comprehend.

Harmonic Mean It is a type of average, commonly used when the values in a data set are rates or ratios, or when lower values should have a greater influence on the mean. It is calculated as the reciprocal of the arithmetic mean of the reciprocals of the values in the dataset.

Median Absolute Error (MdAE) It is a evaluation metric used to measure the accuracy of predictions, specifically in regression tasks. This metric is more robust than mean-based error metrics as it focuses on the median absolute errors, making it less influenced by extreme values or outliers.

Minkowski Distance It is a metric used to calculate the distance between two vectors or points in a vector space, and it can be adjusted to represent different types of distance depending on the parameter p.

Normalized Discounted Cumulative Gain (NDCG) It is a metric used to evaluate the quality of ranked results in information retrieval and recommendation systems. It measures how well the system's ranking matches the ideal ranking, taking into account the position of relevant items.

Novel n-grams Measure that quantifies the proportion of n-grams in a generated summary that do not appear in the original text. Indicates how much new phrasing or content the summary introduces compared to the source document.

Pearson's Correlation It is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is commonly used to assess how closely two variables are related, in terms of both direction (positive or negative) and strength (weak or strong).

Redability How easy and clear a piece of text is to read and understand. It involves factors such as the complexity of the language, sentence structure, vocabulary, and overall organization of the content. The goal of readability is to make the text accessible and engaging to its intended audience.

Relevance Degree to which the summarized content retains the most important, significant, or pertinent information from the original text. Higher relevance indicates that key points and essential details of the document are preserved in the summary while less critical information is omitted.

Root Mean Square Error (RMSE) It is a metric used to evaluate the accuracy of a model, especially in regression tasks. It measures the average magnitude of the errors between predicted values and the actual observed values. RMSE is used to understand how far off the predictions are from the true values, with larger values indicating worse model performance.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) It is a set of evaluation metrics used to assess the quality of machine-generated text, such as summaries or translations, by comparing it to reference texts created by humans. They evaluate the overlap between the generated and reference texts in terms of n-grams, word sequences, and their longest common subsequences.

Chapter 1

Introduction

Automatic summarization is one of the problems of Natural Language Processing (NLP) that has recently gained relevancy and urgency with the broad expansion of information technologies in recent decades. There is a wide need for these types of systems, due to the continuously increasing amount and complexity of unstructured information on the web, typically published in digital media such as newspapers, blogs, e-books, research articles, social media conversations, meetings, etc. Through automatic summarization systems, it is possible to simplify the contents of documents by omitting the parts of the documents that do not add valuable information and keeping the relevant ones, joining them in a comprehensive and readable manner. In this way, users need less time to understand the key information of documents, and they quicker decide whether a document is valuable for what they require or not, spending less time in the broad research phase. A social and educational dimension is also covered since the simplification of documents makes it easier for people with reduced cognitive abilities to understand them; therefore, more knowledge can reach a wider audience. Even more, the impact of technological consumption on user interaction is reduced. Given that the amount of irrelevant information sent is drastically shrunk, and therefore, users can complete their research tasks more efficiently with less data transference.

In order to be valuable and reliable for users, the summaries generated by summarization systems should present coherent and fluent text, that condenses the relevant information without introducing incorrect facts that do not entail from the original documents. Since summaries act as standard-bearers or substitutes for original docu-

ments, the summaries should not contain information that gives the user an incorrect idea of the summarized document. The automatic summarization idea falls apart if systems do not extract the most relevant information for the needs of the user, generated summaries are hard to understand, or they present erroneous facts that will misinform the reader. For this reason, the automatic summarization problem arises other NLP problems that are very complex by themself: text comprehension for detecting the relevant information, text generation for composing coherent and fluid texts, control aspects of the generation to produce summaries that fit the needs of the users better, control bias and subjectivity of the generated texts, the control of emotion bias (words that imply certain emotions) in the summarization of texts like news or opinions, or the control of the style of the summaries (How much text should be copied from the original document?, how frequently synonyms should be used?, what kind of language should be used?, ...).

Aside from generating summaries, the systems must be evaluated in several aspects, such as readability of the text, relevancy of the contained information, the appearance of incorrect facts (known as hallucinations), and the style of summaries (vocabulary used, complexity of sentences, ...). What defines a good summary? The question is far from have a trivial answer since the quality depends on the purpose of the summary, the target audience, or user preferences, for instance. Moreover, many summaries for the same document would be assessed as equally valid; however, a significant portion of the current metrics used for summarization evaluate systems considering a unique summary as reference. Regarding hallucinations, how do we detect which facts in the summary could be inferred from the original text, even though they do not appear explicitly in the document? How do we infer that knowledge? Finally, concerning the style that the summary is created, how do we define or characterize those styles to measure them? Overall, it can be noticed that automatic summarization depends on several other complex problems that need to be tackle to increase the effectiveness and capabilities of the automatic summarization solutions.

When we focus on the ways or strategies to generate summaries, there are two main strategies: extractive strategy, where fragments are extracted from documents, and abstractive techniques, where the most relevant information must be located and re-adapted for summaries. In order to address the problem of summary generation, an extractive approach has typically been followed (Erkan and Radev, 2004; Gunaratna, Thirunarayan, and Sheth, 2015; Kotkar et al., 2024; Mihalcea and Tarau, 2004; Nallapati, Zhai, and Zhou, 2017; Narayan, Cohen, and Lapata, 2018; See, P. J. Liu, and Manning, 2017). However, currently, the focus of automatic summarization research is on abstractive summarization strategy (Grail, Perez, and Gaussier, 2021; Laskar et al., 2023; Yang Liu and Lapata, 2019; Sun et al., 2024; Varab and Xu, 2023; Wang et al., 2020; J. Zhang et al., 2020) due to the strong text comprehension and generation capabilities empirically shown by Transformer-based language models (Devlin et al.,

2019; Lewis et al., 2020; J. Zhang et al., 2020); although Transformer-based solutions have been also proposed for extractive summarization (Abdel-Salam and Rafea, 2022; Yang Liu and Lapata, 2019; Sun et al., 2024; Varab and Xu, 2023; Xie et al., 2022). Specifically, the approach to abstractive automatic summarization today is to train Deep Learning models based on Transformers (Vaswani et al., 2017). For this purpose, the transfer learning technique is used, which consists of pre-training the models with vast amounts of unlabeled text to learn linguistic knowledge that is subsequently transferred to the supervised automatic summarization task by tuning the model using large corpora composed of pairs (document, summary). Among the models that use this strategy standout BART (Lewis et al., 2020) and PEGASUS (J. Zhang et al., 2020), which have been pre-trained for English, with different corpora and pre-training tasks. In order to increase the flexibility of the language models created, multilingual models have been introduced in recent years, where models are trained with corpora incorporating several languages. Of particular interest are the mBART (Yinhan Liu et al., 2020) and mT5 (Xue et al., 2021) models, which have been trained with dozens of languages at a time. Since the parameter scaling of language models in the last few years –what we refer nowadays as Large Language Models (LLMs)–, the multilingual approach has been the prevailing one in LLMs family models such as: GPT (OpenAI, 2023a; OpenAI, 2023b; Radford et al., 2019), LLAMA (Dubey et al., 2024; Touvron, Lavril, et al., 2023; Touvron, Martin, et al., 2023), or Gemini (Anil et al., 2023) among others.

It is clear that the future of the automatic generation of summaries is to generate them in an abstractive way, the strategy that we humans follow when summarizing. This strategy (as opposed to the extractive one) brings flexibility and compression capacity to the summarization process. However, the abstractive strategy is significantly more complex to approach than the extractive strategy. By allowing models to reinterpret the original text, they may introduce incorrect information regarding the summarized document. The control, detection, and quantification of the occurrence of hallucinations in summaries have become very important in recent years (Chen et al., 2021; Maynez et al., 2020; Z. Zhao, Cohen, and Webber, 2020). Also, by generating abstractive summaries, it is easier for more than one summary to have the same quality and validity. Having an indeterminate number of valid summaries implies that widely used summary evaluation metrics, such as the ROUGE variants, based on the coincidence of text fragments between the reference summary and the generated one, are no longer a reliable indication of quality. To continue progressing in the automatic abstractive summarization and make the systems further reliable, we need to propose equally reliable evaluation methods that adapt better to the abstractive summaries they evaluate. Therefore, new quality metrics are sought that adapt to the summarization paradigm, based on the semantics of the words or phrases (Clark, Celikyilmaz, and Smith, 2019; T. Zhang et al., 2020; W. Zhao et al., 2019), the ability of

the summary to answer questions (Scialom, Lamprier, et al., 2019), or metrics that do not require a reference summary (Bhuyan et al., 2023; Gao, W. Zhao, and Eger, 2020; Laban et al., 2022; Scialom, Dray, et al., 2021; Song et al., 2024; Vasilyev, Dharnidharka, and Bohannon, 2020).

Although the abstractivity strategy has led recent years' works, the measurement and characterization of abstractivity in summaries needs further exploration. Presently, the prevailing idea of what is considered an abstractive summary is abstractive when it is written in words other than those that appear in the document to be summarized; *they are written in the words of the author of the summary* (Nenkova and McKeown, 2011). For that reason, some works have addressed the quantification of abstraction by measuring the occurrence or absence of overlap of text segments between the summary and the original document (Bommasani and Cardie, 2020; Grusky, Naaman, and Artzi, 2018; Kryściński et al., 2018). However, following that definition, the measurement becomes coarse and reduces the possibility of understanding the process of adapting the relevant information in a document into an abstractive summary. Jing, 2002 identified a set of actions used by summarizing professionals in creating summaries; actions used for composing the summaries with the information in documents. From that work, it can be extracted that abstractivity in summaries does not necessarily come only from using words other than those used in the original document. Shortening sentences, combining them, using generalizations or specifications, or reordering content causes the original text to be restructured to adapt it into a summary. Moreover, this conception of abstractivity is compatible with the previous conception, since when we apply that set of actions, a summary will be perceived differently from the original document, and therefore, it could be roughly described as "*the summary was written in the author's words*".

One aspect that is usually overlooked in summarization is how the style of the summary varies depending on the purpose. Every summary condenses information, but if the goal of the summary is different, it would be expected that the way the relevant information is condensed would differ, even if all the summaries come from the same document and contain roughly the same key ideas. We can view the reader (the person who consumes the summary) and the writer (the person who creates the summary) as separate entities. In summarizing, the writer will always need to understand the document's content to summarize; however, there is a reasonable chance that the writer will create a summary that will only cover some of the reader's needs if the reader does not explain their purpose and needs. For that reason, enabling the possibility that the systems could better consider the users' needs regarding the style of the summary will always be beneficial.

In this work, we focus our efforts primarily on aspects derived from abstractive summarization, intending to increase the understanding, control, and reliability of

these types of solutions. Specifically, provide new resources (datasets and models) focused on the abstractive strategy of summarization; delve into the concept of *abstractive summaries* in contraposition with the *extractive summaries* and how both approaches are related; increase the perception of abstractivity in the summaries generated by the summarization models; and propose solutions to distinguish the quality of abstractive summaries better. Also, we explore the emotional bias in datasets and how it is transferred to summarization models at the end. Generally speaking, we will focus mostly on the writing aspect of the summarization problem to provide tools to create more flexible automatic summarization systems.

1.1 Hypothesis

The following hypothesis have primarly guide this thesis:

1. We hypothesize that a better definition and quantification of the aspects that characterize abstractivity in summaries would help increase the understanding and control of the models for the automatic generation of abstractive summaries. We expect that greater control over these models will improve the quality of solutions proposed for each specific case. For example, by using the same automatic summarization model, we hypothesize that it will be possible to avoid using synonyms, generalizations, or specifications when working with medical reports, which might not be desirable in this context, while such strategies might be useful in news summaries. Therefore, we hypothesize that more general summary models, less dependent on training data, could be developed that are better adapted to each particular problem.

2. Although extractive and abstractive strategies are often considered mutually exclusive and unrelated, we hypothesize that they can be seen as the extreme poles of a single dimension. The pure extractive strategy, which selects a reduced set of sentences that condense the most relevant information without modification, contrasts with the abstractivity strategy, which locates relevant information and then adapts and condenses it using editing actions. We propose that the extractive strategy can be considered a part of the abstractivity strategy when few or no editing actions are used. Furthermore, we hypothesize that the perception of whether an extractive or abstractive strategy is used can vary depending on the frequency of editing actions, indicating that extractivity and abstractivity exist on a continuum.

3. When composing any text, the choice of certain words or expressions over others may alter the message perceived by future readers. We hypothesize that this also holds true in abstractive summarization: when different expressions

or words are used instead of those in the original document, the message in the summary may change. A specific case of this can be seen in journalism, where certain words are preferred due to the emotions they evoke, even if they are technically synonyms. In news articles, word choices are often made to invoke certain emotions that might influence the reader's decision to read the full article. Therefore, we hypothesize that emotional bias is present not only in documents and summaries of newspaper datasets but also that summarization models trained on such datasets may acquire this bias.

1.2 Objectives

The main objective of the thesis is to contribute to the progress of the task of generating automatic summaries in the area of Natural Language Processing, especially in solutions that approach it from an abstractive strategy.

From the main objective we can extract several specific objectives:

1. **Resources for abstractive summarization:** To provide new resources for abstractive summarization, thus they would be useful for the scientific community; prioritizing those languages that are underrepresented in the available resources.

2. **Characterization of abstractivity:** To expand the definition of the current concept of abstractivity, with the identification of aspects that define it and metrics that can measure those aspects.

3. **Measure quality:** To find ways to measure the quality of the generated summaries that better fit the abstractive paradigm, where more than one summary can be considered a high quality solution.

4. **Emotional words:** To study whether journalistic datasets contain words that entail emotions to readers, find out whether there is a emotional bias in documents and summaries (prevalence for certain emotions depending on the dataset), and verify whether models acquire that bias through the training process.

5. **Develop abstractive systems:** To create systems for the automatic generation of summaries for documents of different domains, fields and lengths.

1.3 Methodology

To carry on this thesis, we followed a rigorous methodological approach grounded in several steps that helped us face each defined objective. At the same time, the author gained a robust comprehension of the NLP problem of automatic summarization during the thesis. The methodology also helped to synthesize all the knowledge raised from this work, which was destined through qualitative and quantitative analysis of the results of experimental studies. These are the steps that I have followed for the creation of the different publications contained in this thesis:

1. **Literature review:** There was an in-depth review of the existing work in order to gain a robust theoretical background to carry on the thesis and, more importantly, identify gaps in the current knowledge where we could contribute.

2. **Theoretical planning:** Once a potential idea was identified through the literature review, we create a plan to develop it. The plan includes the software development needs and technical needs. The plan also includes a discussion of the primary caveats that were detected and how we will surpass them in our solution.

3. **Technical learning and Development:** Considering our needs for the development. We first identified those technical aspects required for the development that needed to be fully consolidated and worked on them. After that, we developed the solution that was planned.

4. **Experimental study:** Performing experiments to assess the developed models or algorithms, using diverse metrics and test data to evaluate their performance.

5. **Statistical analysis:** Applying statistical methods to analyze experimental results, ensuring the robustness, reliability, and generalization of findings.

6. **Critical exploration:** Distilling from the statistical analysis the considerations and conclusions that have emerged from the results.

7. **Reporting:** Condensing all the work done in the scientific report where the scientific community will validate our statements and work. This step also helps us to reorder and consolidate the ideas distilled from the work since we need to think about how to communicate our work to the rest of the scientific community.

1.4 Structure of the thesis

The rest of the thesis, excluding Chapter 1, comprises additional eight chapters. Chapters 2 to 7 are the scientific articles that have been arisen from the work done to achieve the objectives that we defined for this thesis (Section 1.2). The thesis discussions can be found in Chapter 8, and in Chapter 9, we end with a the thesis conclusions and future work.

Regarding the articles that contain this thesis, we briefly summarize their purpose in the thesis project as follows:

- **Chapter 2 - DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles.** This work provides a new large-scale summarization dataset for Catalan and Spanish. The construction of the dataset focuses on creating a collection of document-summary pairs that do not present a clear extractive perception.

- **Chapter 3 - NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish.** This work aims to introduce a pre-training methodology that enhances the knowledge transfer to summarization and also enhances the perception of abstractivity of the summaries. It also presents a characterization of abstractivity based on currently available abstractivity-related metrics. Additionally, a new abstractivivity-related metric is proposed.

- **Chapter 4 - ELiRF-VRAIN at BioNLP Task 1B: Radiology Report Summarization.** This work details our participation in a shared task at the BioNLP 2023. We apply the pre-training methodology defined in Chapter 3 on the biomedical domain, and we analyze the behavior of the summarization models based on the pre-trained model obtained with our methodology.

- **Chapter 5 - ELiRF-VRAIN at BioLaySumm: Boosting Lay Summarization Systems Performance with Ranking Models.** This work details our participation at the BioNLP 2024. We apply the pre-training methodology defined in Chapter 3 on the medical-research domain, we tackle the problem of long document summarization, and we also approach the quality detection of summaries.

- **Chapter 6 - Abstractive Summarizers Become Emotional on News Summarization.** In this work, we study the presence of words that entail emotions in news articles. The emotional bias in newspaper datasets and how it is transferred to the summarization models is also studied.

- **Chapter 7 - Beyond Using Their Own Words: Abstractivity Characterization in Summarization.** In this work, we expand the definition of abstractivity and how it can be characterized. In the expanded definition, the extractive and abstractive approaches are placed in a single dimension. The expanded definition is possible when we consider the frequency of appearance of a set of summarization actions in the characterization of abstractivity. To achieve those goals, we elaborated a dataset for abstractivity characterization and measurement of the level of abstractivity in summaries.

1.5 Relation to the United Nations Sustainable Development Goals

In 2015 the United Nations defined the 2030 Agenda. Within that Agenda, 17 Sustainable Development Goals (SDGs) were established as a global call to action to tackle urgent social, economic, and environmental challenges (United Nations, 2023). The SDGs emphasize universal efforts to end poverty, fight inequality, promote sustainable economic growth, and protect the planet. The goals reflect an integrated approach to sustainable development, recognizing the interconnections among economic, social, and environmental dimensions for achieving long-term global progress.

In the case of this thesis, it aligns with the following SDGs:

- **4. Quality Education:** Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.

- **9. Industry, Innovation and Infrastructure:** Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation.

- **12. Responsible Consumption and Production:** Ensure sustainable consumption and production patterns.

Improving the quality of the summary models increases the possibility of condensing the information in the documents. Thus, this helps to enhance the dissemination of information and accessibility to broader audiences, which allows a direct increase in the quality of education. Greater ease of information dissemination also contributes to industry innovation and productivity since it will enable, for instance, reduce the

amount of time dedicated to research or to get the big picture of the information in a report. Finally, condensing information enables different ways of lowering energetic consumption since the presence of irrelevant information is shrunk with summarization; which will help reduce the amount of data users use in daily tasks.

Chapter 1. References

Abdel-Salam, Shehab and Ahmed Rafea (2022). "Performance Study on Extractive Text Summarization Using BERT Models". In: *Information* 13.2. ISSN: 2078-2489. DOI: 10.3390/info13020067. URL: https://www.mdpi.com/2078-2489/13/2/67 (cit. on p. 3).

Anil, Rohan et al. (2023). "Gemini: A Family of Highly Capable Multimodal Models". In: *arXiv preprint arXiv:2312.11805*. URL: https://arxiv.org/abs/2312.11805 (cit. on p. 3).

Bhuyan, Swagat Shubham et al. (2023). "Textual entailment as an evaluation metric for abstractive text summarization". In: *Natural Language Processing Journal* 4, p. 100028. ISSN: 2949-7191. DOI: https://doi.org/10.1016/j.nlp.2023.100028. URL: https://www.sciencedirect.com/science/article/pii/S2949719123000250 (cit. on p. 4).

Bommasani, Rishi and Claire Cardie (Nov. 2020). "Intrinsic Evaluation of Summarization Datasets". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8075–8096. DOI: 10.18653/v1/2020.emnlp-main.649. URL: https://aclanthology.org/2020.emnlp-main.649 (cit. on p. 4).

Chen, Sihao et al. (June 2021). "Improving Faithfulness in Abstractive Summarization with Contrast Candidate Generation and Selection". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, pp. 5935–5941. DOI: 10.18653/v1/2021.naacl-main.475. URL: https://aclanthology.org/2021.naacl-main.475 (cit. on p. 3).

Clark, Elizabeth, Asli Celikyilmaz, and Noah A. Smith (July 2019). "Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 2748–2760. DOI: 10.18653/v1/P19-1264. URL: https://aclanthology.org/P19-1264 (cit. on p. 3).

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423 (cit. on p. 2).

Dubey, Abhimanyu et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783 (cit. on p. 3).

Erkan, Günes and Dragomir R. Radev (Dec. 2004). "LexRank: Graph-based Lexical Centrality As Salience in Text Summarization". In: *J. Artif. Int. Res.* 22.1, pp. 457–479. ISSN: 1076-9757 (cit. on p. 2).

Gao, Yang, Wei Zhao, and Steffen Eger (July 2020). "SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 1347–1354. DOI: 10.18653/v1/2020.acl-main.124. URL: https://aclanthology.org/2020.acl-main.124 (cit. on p. 4).

Grail, Quentin, Julien Perez, and Eric Gaussier (Apr. 2021). "Globalizing BERT-based Transformer Architectures for Long Document Summarization". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, pp. 1792–1810. DOI: 10.18653/v1/2021.eacl-main.154. URL: https://aclanthology.org/2021.eacl-main.154 (cit. on p. 2).

Grusky, Max, Mor Naaman, and Yoav Artzi (June 2018). "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 708–719. DOI: 10.18653/v1/N18-1065. URL: https://aclanthology.org/N18-1065 (cit. on p. 4).

Gunaratna, Kalpa, Krishnaparasad Thirunarayan, and Amit Sheth (Feb. 2015). "FACES: Diversity-Aware Entity Summarization Using Incremental Hierarchical Conceptual Clustering". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 29.1. DOI: 10.1609/aaai.v29i1.9180. URL: https://ojs.aaai.org/index.php/AAAI/article/view/9180 (cit. on p. 2).

Jing, Hongyan (Dec. 2002). "Using Hidden Markov Modeling to Decompose Human-Written Summaries". In: *Computational Linguistics* 28.4, pp. 527–543. ISSN: 0891-2017. DOI: 10.1162/089120102762671972. URL: https://doi.org/10.1162/089120102762671972 (cit. on p. 4).

Kotkar, Aishwarya D. et al. (2024). "Comparative Analysis of Transformer-based Large Language Models (LLMs) for Text Summarization". In: *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, pp. 1–7. DOI: 10.1109/ACET61898.2024.10730348 (cit. on p. 2).

Kryściński, Wojciech et al. (Oct. 2018). "Improving Abstraction in Text Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1808–1817. DOI: 10.18653/v1/D18-1207. URL: https://aclanthology.org/D18-1207 (cit. on p. 4).

Laban, Philippe et al. (Feb. 2022). "SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization". In: *Transactions of the Association for Computational Linguistics* 10, pp. 163–177. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00453. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00453/1987014/tacl_a_00453.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00453 (cit. on p. 4).

Laskar, Md Tahmid Rahman et al. (Dec. 2023). "Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Ed. by Mingxuan Wang and Imed Zitouni. Singapore: Association for Computational Linguistics, pp. 343–352. DOI: 10.18653/v1/2023.emnlp-industry.33. URL: https://aclanthology.org/2023.emnlp-industry.33 (cit. on p. 2).

Lewis, Mike et al. (July 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*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703 (cit. on p. 3).

Liu, Yang and Mirella Lapata (Nov. 2019). "Text Summarization with Pretrained Encoders". 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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740. DOI: 10.18653/v1/D19-1387. URL: https://aclanthology.org/D19-1387 (cit. on pp. 2, 3).

Liu, Yinhan et al. (Nov. 2020). "Multilingual Denoising Pre-training for Neural Machine Translation". In: *Transactions of the Association for Computational Linguistics* 8, pp. 726–742. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00343. URL: https://www.aclweb.org/anthology/2020.tacl-1.47 (cit. on p. 3).

Maynez, Joshua et al. (July 2020). "On Faithfulness and Factuality in Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association*

for Computational Linguistics. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: https://aclanthology.org/2020.acl-main.173 (cit. on p. 3).

Mihalcea, Rada and Paul Tarau (July 2004). "TextRank: Bringing Order into Text". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Ed. by Dekang Lin and Dekai Wu. Barcelona, Spain: Association for Computational Linguistics, pp. 404–411. URL: https://aclanthology.org/W04-3252 (cit. on p. 2).

Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou (2017). "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, pp. 3075–3081. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10958/10817 (cit. on p. 2).

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (Oct. 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: Association for Computational Linguistics, pp. 1797–1807. DOI: 10.18653/v1/D18-1206. URL: https://aclanthology.org/D18-1206 (cit. on p. 2).

Nenkova, Ani and Kathleen McKeown (June 2011). "Automatic Summarization". In: *Foundations and Trends® in Information Retrieval* 5.2–3, pp. 103–233. ISSN: 1554-0669. DOI: 10.1561/1500000015. URL: http://dx.doi.org/10.1561/1500000015 (cit. on p. 4).

OpenAI (2023a). "GPT-4 Technical Report". In: *CoRR* abs/2303.08774. DOI: 10.48550/arXiv.2303.08774. arXiv: 2303.08774. URL: https://doi.org/10.48550/arXiv.2303.08774 (cit. on p. 3).

— (2023b). *OpenAI GPT-3.5 model series: technical overview and capabilities*. Available at https://platform.openai.com/docs/models/gpt-3.5 (cit. on p. 3).

Radford, Alec et al. (2019). "Language Models are Unsupervised Multitask Learners". In: URL: https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe (cit. on p. 3).

Scialom, Thomas, Paul-Alexis Dray, et al. (Nov. 2021). "QuestEval: Summarization Asks for Fact-based Evaluation". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6594–6604. DOI: 10.18653/v1/2021.emnlp-main.529. URL: https://aclanthology.org/2021.emnlp-main.529 (cit. on p. 4).

Scialom, Thomas, Sylvain Lamprier, et al. (Nov. 2019). "Answers Unite! Unsupervised Metrics for Reinforced Summarization Models". 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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 3246–3256. DOI: 10.18653/v1/D19-1320. URL: https://aclanthology.org/D19-1320 (cit. on p. 4).

See, Abigail, Peter J. Liu, and Christopher D. Manning (July 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)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. URL: https://aclanthology.org/P17-1099 (cit. on p. 2).

Song, Hwanjun et al. (Aug. 2024). "FineSurE: Fine-grained Summarization Evaluation using LLMs". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 906–922. DOI: 10.18653/v1/2024.acl-long.51. URL: https://aclanthology.org/2024.acl-long.51 (cit. on p. 4).

Sun, Weisong et al. (Mar. 2024). "An Extractive-and-Abstractive Framework for Source Code Summarization". In: *ACM Trans. Softw. Eng. Methodol.* 33.3. ISSN: 1049-331X. DOI: 10.1145/3632742. URL: https://doi.org/10.1145/3632742 (cit. on pp. 2, 3).

Touvron, Hugo, Thibaut Lavril, et al. (2023). "LLaMA: Open and Efficient Foundation Language Models". In: *arXiv preprint arXiv:2302.13971* (cit. on p. 3).

Touvron, Hugo, Louis Martin, et al. (2023). "LLaMA 2: Open Foundation and Fine-Tuned Chat Models". In: *arXiv preprint arXiv:2307.09288* (cit. on p. 3).

United Nations (2023). *The 17 Goals for Sustainable Development*. Accessed: 2024-11-03. URL: https://sdgs.un.org/goals (cit. on p. 9).

Varab, Daniel and Yumo Xu (July 2023). "Abstractive Summarizers are Excellent Extractive Summarizers". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 330–339. DOI: 10.18653/v1/2023.acl-short.29. URL: https://aclanthology.org/2023.acl-short.29 (cit. on pp. 2, 3).

Vasilyev, Oleg, Vedant Dharnidharka, and John Bohannon (Nov. 2020). "Fill in the BLANC: Human-free quality estimation of document summaries". In: *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Ed. by Steffen Eger et al. Online: Association for Computational Linguistics, pp. 11–20. DOI:

10.18653/v1/2020.eval4nlp-1.2. URL: https://aclanthology.org/2020.eval4nlp-1.2 (cit. on p. 4).

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., pp. 5998–6008. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 3).

Wang, Zhengjue et al. (Nov. 2020). "Friendly Topic Assistant for Transformer Based Abstractive Summarization". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, pp. 485–497. DOI: 10.18653/v1/2020.emnlp-main.35. URL: https://aclanthology.org/2020.emnlp-main.35 (cit. on p. 2).

Xie, Qianqian et al. (2022). "Pre-trained language models with domain knowledge for biomedical extractive summarization". In: *Knowledge-Based Systems* 252, p. 109460. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2022.109460. URL: https://www.sciencedirect.com/science/article/pii/S0950705122007328 (cit. on p. 3).

Xue, Linting et al. (June 2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41 (cit. on p. 3).

Zhang, Jingqing et al. (July 2020). "PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. ICML20. PMLR. JMLR.org, pp. 11328–11339. URL: https://proceedings.mlr.press/v119/zhang20ae.html (cit. on pp. 2, 3).

Zhang, Tianyi et al. (2020). "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations*. OpenReview.net. URL: https://openreview.net/forum?id=SkeHuCVFDr (cit. on p. 3).

Zhao, Wei et al. (Nov. 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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 563–578. DOI: 10.18653/v1/D19-1053. URL: https://aclanthology.org/D19-1053 (cit. on p. 3).

Zhao, Zheng, Shay B. Cohen, and Bonnie Webber (Nov. 2020). "Reducing Quantity Hallucinations in Abstractive Summarization". In: *Findings of the Association for*

Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, pp. 2237–2249. DOI: 10.18653/v1/2020.findings-emnlp.203. URL: https://aclanthology.org/2020.findings-emnlp.203 (cit. on p. 3).

Chapter 2

DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles

Encarnación Segarra-Soriano, Vicent Ahuir, Lluís-F. Hurtado, and José González (July 2022). "DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 5931–5943. DOI: 10.18653/v1/2022.naacl-main.434. URL: https://aclanthology.org/2022.naacl-main.434

Impact Index

In 2022, the *North American Chapter of the Association for Computational Linguistics* was ranked by the *Computing Research & Education* (CORE) as CORE-A congress. Also classified as A+ by the *Sociedad Científica Informática en España.*

Abstract

The application of supervised methods to automatic summarization requires the availability of adequate corpora consisting of a set of document-summary pairs. As in most Natural Language Processing tasks, the great majority of available datasets for summarization are in English, making it difficult to develop automatic summarization models for other languages. Although Spanish is gradually forming part of some recent summarization corpora, it is not the same for minority languages such as Catalan. In this work, we describe the construction of a corpus of Catalan and Spanish newspapers, the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) corpus. It is a high-quality large-scale corpus that can be used to train summarization models for Catalan and Spanish. We have carried out an analysis of the corpus, both in terms of the style of the summaries and the difficulty of the summarization task. In particular, we have used a set of well-known metrics in the summarization field in order to characterize the corpus. Additionally, we have evaluated the performance of some extractive and abstractive summarization systems on the DACSA corpus for benchmarking purposes.

2.1 Introduction

Automatic summarization is one of the central problems in Natural Language Processing (NLP). The development of automatic summarization systems is an important issue due to the great amount of information in different formats that is accessible on the web or in other repositories. It is necessary to develop techniques that help us to tackle that huge amount of information. For this reason, there is an increasing interest in the NLP community to develop techniques that allow the users to find, read, understand, or process the documents. In this context, automatic summarization can be an important aid because it provides a condensed version of documents that reduce the time to explore or analyze them.

Access to large-scale high-quality data is an essential prerequisite for making substantial progress in summarization. The application of supervised methods to automatic summarization, as those based on Neural Networks, requires the availability of adequate corpora consisting of document-summary pairs. The construction of large-scale and high-quality corpora for learning neural summarization models is not an easy task. It is necessary a great human effort to generate thousands of manual summaries, or to design new approaches to obtain these summaries in a semiautomatic way. The first important resource for learning corpus-based summarization models was the CNN/DailyMail summarization corpus (Hermann et al., 2015), originally constructed for the task of passage-based question answering and adapted to the document summarization task. It consists of news stories from CNN and DailyMail and contains 312 077 article-summary pairs. Afterwards, another English corpus was provided to the research community for summarization purposes, the NewsRoom corpus (Grusky, Naaman, and Artzi, 2018). It consists of 1.3 million article-summary pairs that have been written by the authors and the editors of 38 different major news publications. The corpus was created through a web-scale crawling of over 100 million pages from a set of online publishers by gathering the news and using the summaries provided in the HTML metadata. The summaries contained in this corpus combine both extractive and abstractive strategies to describe the content of the articles. Also in 2018, the XSUM corpus (Narayan, Cohen, and Lapata, 2018a) was presented, it is a large scale dataset obtained by harvesting online articles from the British Broadcasting Corporation (BBC) with one-sentence news summary.

As in most NLP tasks, the great majority of available datasets for summarization are in English. The lack of this kind of resources for other languages is an encumbrance to modeling that constraints the impact of language technologies on minority language communities. The creation of a large-scale Indonesian summarization dataset of 215 827 document-summary pairs, has just been published (Koto, Lau, and Baldwin, 2020). Recently, some datasets that aim to fill the gap among English and other languages for the automatic summarization task have been proposed: MLSUM

(Scialom et al., 2020), MassiveSumm (Varab and Schluter, 2021), and XL-Summ (Hasan et al., 2021). Although Spanish is the world's second-most spoken native language and is the official language in 21 countries, it has only recently been considered in general domain summarization datasets, as the aforementioned, and in specific domains as in (J.-A. González et al., 2019). The situation is worse for Catalan, although it is not an endangered language, it is spoken by 10 million people in Spain and other three European countries, it is minority worldwide and is underrepresented or even not considered in summarization corpora.

In this work, we describe the construction of a corpus of Catalan and Spanish newspapers, the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) corpus. With the aim of building a quality large-scale corpus that could be used to learn automatic summarization neural models for Catalan and Spanish, we used a strategy inspired by the construction of the NewsRoom corpus (Grusky, Naaman, and Artzi, 2018). We conducted a crawling process on 30 different newspaper websites to extract articles and summaries in a straightforward way. The crawling included from Spanish mass media to regional newspapers. In order to obtain the summaries, we took advantage of the highlights and summaries, provided by authors or editors of the articles.

To ensure the quality of the DACSA corpus, we perform two subsequent filtering processes on the downloaded articles. The first filter was used to ensure, at least, a minimum length in both the article and the abstract. All the articles or summaries that were considered too short were discarded. Obviously, an article or summary too short implies discarding the article-summary pair. The second filter was used to ensure that the summaries were not almost verbatim copies of the first sentences of the articles. To do this, the article-summary pairs in which the overlapping between the summary and the article prefix of the summary length was high were also discarded. This way, we try to avoid a positional bias in the summaries by discarding those samples in which the summary is reduced to select the first sentences of the article.

Once both filters were applied, we found that some newspaper sources had very few samples, less than 1000 in some cases. To balance the corpus partitions, we decided to remove the sources with few samples from the training, validation, and tests sets. Nevertheless, we joined together the samples from those sources to create a special test set, a test set with sources not present in the training process. Therefore, the corpus consists of four partitions per language: training, validation, and test sets along with an extra test set. Considering all the partitions, the DACSA corpus consists of a set of 725 184 article-summary pairs extracted from 9 different Catalan newspaper websites and 2 120 649 article-summary pairs extracted from 21 different Spanish newspaper websites. The DACSA corpus contains articles and summaries about politics, economics, sports, culture and other topics usually addressed in journalistic

domains. To our knowledge, the DACSA corpus is the largest summarization dataset for both languages.

We have used four well known metrics in the summarization field in order to characterize the corpus. These metrics are: extractive fragment coverage and density (Grusky, Naaman, and Artzi, 2018), abstractivity-p (Bommasani and Cardie, 2020), and novel n-grams (Kryściński et al., 2018). Additionally, for benchmarking purposes, we have evaluated the performance of 6 automatic summarization systems on the DACSA corpus. Concretely, we have used two unsupervised systems (lead-2 and textRank), an extractive summarization system, SHANN (J.-Á. González et al., 2019), two abstractive summarization systems, mBART (Yinhan Liu, Gu, et al., 2020) and mT5 (Xue et al., 2021), and one oracle to compute upper bounds of the performance in the DACSA corpus.

The DACSA corpus is publicly available at `https://huggingface.co/datasets/ELiRF/dacsa`.

2.2 Related Work

The automatic text summarization problem has been addressed in the literature using abstractive, extractive, or mixed approaches. On the one hand, extractive approaches compose summaries by selecting sentences or words directly from the documents (Chen and Bansal, 2018; Dong et al., 2018; Yang Liu and Lapata, 2019; Nallapati, Zhai, and Zhou, 2017; Narayan, Cohen, and Lapata, 2018b; Yao et al., 2018; X. Zhang et al., 2018). Most of these approaches address a sequential binary sentence classification problem in order to select the most salient sentences of the documents, following different criteria such as negative log likelihood on preselected sentences (Yang Liu and Lapata, 2019; Nallapati, Zhai, and Zhou, 2017) or ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) rewards in reinforcement learning environments (Dong et al., 2018; Narayan, Cohen, and Lapata, 2018b; Yao et al., 2018; X. Zhang et al., 2018). Other extractive architectures are based on siamese hierarchical attention networks built in terms of Long Short Term Memories and Transformer encoders (J. Á. González et al., 2020; J.-Á. González et al., 2019). These models have been succesfully applied in summarization tasks of Spanish newspapers and talk shows (J.-A. González et al., 2019). On the other hand, the abstractive approaches build the summaries by paraphrasing the sentences of the documents (Ive, Madhyastha, and Specia, 2019; Paulus, Xiong, and Socher, 2018; See, P. J. Liu, and Manning, 2017). The vast majority of existing neural abstractive summarization models are based on encoder-decoder architectures (Sutskever, Vinyals, and Le, 2014). Finally, there are also mixed approaches that combine extractive and abstractive techniques, performed in a decoupled way or simultaneously inside the models (Mendes et al., 2019).

Due to the recent success of self-supervised learning, the focus of text summarization research has exhibited a gradual shift from extractive techniques to abstractive techniques (Lewis et al., 2020; Raffel et al., 2020; J. Zhang et al., 2020). These kind of objectives allows to pretrain deep architectures (mainly Transformers) to learn vast amounts of general linguistic knowledge from large corpora, that can be transferred to downstream tasks by means of finetuning. The most successful model of this type is BERT (Devlin et al., 2019), that is pre-trained with Masked Language Model and Next Sentence Prediction objectives on raw texts from English Wikipedia and BooksCorpus. Based on BERT, some architectural improvements have been proposed like RoBERTa (Yinhan Liu, Ott, et al., 2019) or ALBERT (Lan et al., 2020).

In some recent works, BERT and RoBERTa have been finetuned for extractive summarization (Yang Liu and Lapata, 2019; Zhong et al., 2020), but, although it boosted the performance of the previous extractive approaches, the pretraining+finetuning philosophy has shown to be most effective for abstractive systems. Nowadays, the best performing abstractive models are BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and PEGASUS (J. Zhang et al., 2020), being all of them Transformers (Vaswani et al., 2017) pretrained self-supervisedly as denoising sequence to sequence autoencoders. Some multilingual variants of these models have been recently proposed, mBART (Yinhan Liu, Gu, et al., 2020) and mT5 (Xue et al., 2021). Both of them were pretrained following a multilingual denoising procedure on large-scale multilingual corpora. On the one hand, the mBART model was pretrained by using a corpus of 25 languages, extracted from the Common Crawl (Wenzek et al., 2020) (CC25). On the other hand, a multilingual variant of the Colossal Clean Crawled corpus (Raffel et al., 2020) was used to pretrain mT5.

Self-supervised pre-training requires obtaining large amounts of raw data in order to learn good initializations of deep models from denoising objectives. Also, the fine-tuning of these architectures in downstream tasks like text summarization implies the availability of adequate corpora consisting of document-summary pairs. As we mention above, the great majority of datasets for summarization are in English: CNN/DailyMail, NewsRoom, XSUM (Narayan, Cohen, and Lapata, 2018a), and so forth. Although some multilingual datasets have been recently created, as MLSUM, MassiveSumm, and XL-Summ, they do not provide a large enough portion of Spanish data and only MassiveSumm provides a few samples for Catalan. It is in this context where we propose to build the DACSA corpus.

The most used metrics in the literature to quantify the performance of the models in the summarization task are ROUGE (Lin, 2004) and BERTScore (T. Zhang et al., 2020). On the one hand, ROUGE measures the performance by counting exact matches. On the other hand, BERTScore is a more semantic measure which is based on contextual embeddings provided by a BERT language model. These metrics are convenient

to evaluate the performance, but they do not explicitly measure the abstractivity. Measuring the abstractivity of the summaries generated by the models is generally not trivial. In this work, we used a set of metrics as abstractivity indicators to asses the level of abstractivity: extractive fragment coverage and density (Grusky, Naaman, and Artzi, 2018), abstractivity$_p$ (Bommasani and Cardie, 2020), and novel n-grams (Kryściński et al., 2018). Additionally, we also used ROUGE and BERTScore to compare the different summarization models.

2.3 Building the DACSA corpus

The DACSA corpus was collected using a distributed web crawler that captured over 6 million news articles, close to 2 million of articles published in Catalan, and more than 4 million written in Spanish. The articles were captured from 30 newspapers sources, 9 sources for Catalan and 21 sources for Spanish. The range of years of publication was between 2010 to 2020.

We divided the crawling process into two services. The first service was designed to retrieve the list of articles on the website of the newspapers source; we refer to this service as the *URLs extractor service*. The second one aims to extract the content (article content and summary) of the article; we refer to this service as the *content extractor service*. The whole crawler was developed with Python 3 and JavaScript (Node.js runtime) programming languages.

For the configurations (one per source) of the *content extractor service*, we used CSS selectors and the library *cheerio* (`https://cheerio.js.org/`). In order to capture the article and summary text, we designed the selectors that captured the visible information that a person would read, avoiding metadata. Using visual information instead of metadata is important because we detected that likely the metadata was automatically created by some naive process that could lose information, such as just extracting the first tokens of the article; meanwhile, the visual text is likely complete, readable and coherent.

We searched websites of electronic newspapers published in Spain, in Catalan or Spanish languages. To find the addresses of each article, we decided to use the list of news that electronic newspapers usually have on their website. The benefit of using the list of articles provided by these websites, contrary to the common crawling approach of following every link, was that we aimed the articles themselves, and there was no need to identify whether the web page is a news article or other kind of content. Thus, from the list of news in that newspapers source, we created two configurations, one for the *URLs extractor service* and another for the *content extractor service*.

We intended DACSA to be a large-scale, high-quality corpus for Catalan and Spanish. Thus, after the massive capture of samples, we defined two requirements that the articles and summaries must satisfy. We first established a threshold in the minimum number of words of the article and the summary, and second, a threshold in the maximum similarity between the summary and the first sentences of the article.

On the one hand, we discarded those samples with a short text in the article or the summary. Specifically, every sample inside the corpus contains at least 100 words in the article and 10 words in the summary. With this restriction, we ensure that the samples have enough content to generate a summary with a reasonable length.

On other hand, we rejected from the corpus those samples in which the summary is generated by simply extracting the first sentences of the article. Specifically, we restricted the overlapping between the summary and the starting sentences of the article by using a similarity metric based on the Levenshtein distance to quantify the degree of overlapping. The Equation (2.1) presents the definition of this metric.

$$f(A,S) = 1 - \frac{Levenshtein(A_{[1,|S|]}, S)}{|S|} \tag{2.1}$$

where A is the sequence of words of the article text, S is the sequence of words of the summary, $|S|$ is number of words of the summary, and $Levenshtein$ is the operation which returns the well-known Levenshtein distance between two texts. In this corpus, we established a maximum threshold of 0.9 of $f(A,S)$ between the article and the summary.

2.4 Dataset

After the above processes, the DACSA corpus was built. This corpus provides pairs of news article and its summary from different newspapers for both, the Catalan and the Spanish languages. Regarding the Catalan set, there are 725 184 sample pairs from 9 newspapers, regarding the Spanish set, the corpus provides 2 120 649 sample pairs from 21 newspapers.

The amount of samples by newspapers source is far from being homogeneous. If these distributions would be preserved over the partitions (training, validation, and test set), the models will focus their learning in the predominant newspapers. To avoid this bias and achieve more general models, we propose that the test and validation sets be created in a way that all newspapers have roughly the same number of samples. To achieve this balance, we discarded some sources in order to guarantee that all sources represent at least 5% of samples in each one of these two sets. Additionally, we discarded those sources that have lower compression ratio than 10% in

Table 2.1: Statistics of the partitions for Catalan.

			Article			Summary		
Source	#Docs	Tokens	Vocabulary size	Sents per doc	Words per sent	Vocabulary size	Sents per doc	Words per sent
Training	636596	316817625	1206292	17.39	28.62	206616	1.17	20.36
Validation	35376	17831029	258999	16.17	31.17	51940	1.15	20.93
TESTI	35376	17704387	262148	16.13	31.03	51958	1.15	20.89
TESTNI	17836	15882219	247154	35.38	25.17	45997	1.56	25.93
Set	725184	368235260	1326343	17.71	28.67	223978	1.17	20.59

Table 2.2: Statistics of the partitions for Spanish.

			Article			Summary		
Source	#Docs	Tokens	Vocabulary size	Sents per doc	Words per sent	Vocabulary size	Sents per doc	Words per sent
Training	1802919	1172626265	2920894	23.94	27.17	454179	1.24	21.99
Validation	104052	67669381	550213	23.01	28.27	109460	1.21	23.36
TESTI	104052	67363994	550910	22.93	28.23	109706	1.21	23.34
TESTNI	109626	59603306	447679	16.25	33.46	116201	1.35	36.84
Set	2120649	1367262946	3189783	23.44	27.50	516307	1.24	22.95

their summaries, since we considered these summaries too long compared to their corresponding articles.

The three sets for Catalan (training, validation and test set) are composed by 6 of the 9 newspapers, the training set contains 636596 samples, and the validation and test sets have 35376 samples each one. For Spanish, the three sets are composed by 13 of the 21 newspapers, the training set contains 1802919 samples, and the validation and test sets have 104052 samples each one.

All the sources excluded were used as a separate test set. This partition allows evaluating the generalization capabilities of the summarization models against unseen newspaper sources. In this work, we refer to the test set with newspapers included in the training set as TESTI and to the test set that contains newspapers not included in the training set as TESTNI. The statistics of all the sets are shown in Tables 2.1 and 2.2.

In the Section 2.7, Tables 2.A.1 and 2.A.2 show the distribution and the average lengths in terms of sentences and words of the articles and their summaries for Catalan and Spanish sets, detailed by the different newspaper sources.

2.5 Analysis of Dataset

In this section, an analysis of the level of abstractivity of the summaries of the corpus is done. First, the definition of the different measures used in this work is given, and second, we provide the application of these measures to the DACSA corpus.

2.5.1 Definition of Abstractivity Metrics

We used a set of metrics as abstractivity indicators to asses the level of abstractivity, they capture the degree of text overlapping between the summary and article. In particular, the following metrics have been selected: extractive fragment coverage and density, abstractivity$_p$, and novel n-grams.

Extractive Fragment Coverage (Grusky, Naaman, and Artzi, 2018): the coverage measure quantifies the extent to which a summary is derivative of a text, that is, it measures the percentage of words in the summary that are part of an extractive fragment of the article.

Extractive Fragment Density (Grusky, Naaman, and Artzi, 2018): contrary to the coverage, the density measure takes into account the length of the extractive fragments. A summary might contain many individual words from the article and therefore have a high coverage, however it might have a low density if the extractive fragments are short.

Abstractivity$_p$ (Bommasani and Cardie, 2020): the *abstractivity$_p$* metric measures abstractivity as the absence of overlapping between the summary and the original text. Higher values indicate less overlapping and higher abstractivity. The p parameter weights the length of each extractive fragment, the higher value of p, the more the length of the extractive fragment is penalized.

Novel n-grams: (Kryściński et al., 2018) the *novel n-grams* metric quantifies the n-grams introduced in the summary that did not appear in the original text. The value of the metric is a percentage over the total of n-grams contained in the summary.

Additionally, we also used the **Compression Ratio**, that is, the ratio between the length of article and summary. Summarizing with higher compression is challenging as it requires capturing more precisely the critical aspects of the article text.

2.5.2 Dataset Abstractivity

This section presents the results of the abstractivity metrics described in Section 2.5.1 for the DACSA corpus. The results are shown separately for both languages; Table 2.3 shows the average values of the partitions for Catalan and Table 2.4 for Spanish.

Tables 2.B.1 and 2.B.2 in the Section 2.7 also show these results for each newspaper source.

Table 2.3: Average values of the metrics in the Catalan partitions.

Source	Compression	Coverage	Density	Abstractivity$_p$ (p=2)	Novel 2-grams	Novel 3-grams	Novel 4-grams
Training	23.12	80.87	3.52	84.13	55.55	73.08	81.26
Validation	22.85	81.16	3.96	82.50	54.02	70.99	79.02
TESTI	22.73	81.24	4.01	82.37	53.85	70.74	78.77
TESTNI	24.01	79.98	5.54	83.51	53.55	70.49	78.14
Set	23.11	80.09	3.62	83.95	55.35	72.80	80.96

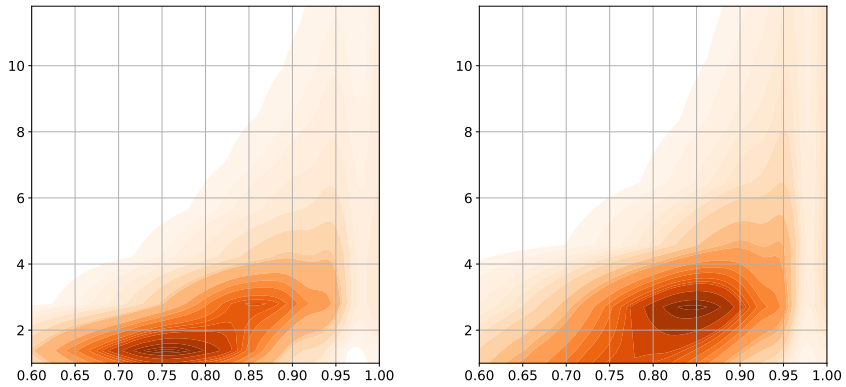
Table 2.4: Average values of the metrics in the Spanish partitions.

Source	Compression	Coverage	Density	Abstractivity$_p$ (p=2)	Novel 2-grams	Novel 3-grams	Novel 4-grams
Training	27.73	82.84	5.64	80.92	51.33	68.74	76.57
Validation	26.32	83.02	5.53	80.07	49.60	66.23	73.94
TESTI	26.20	83.11	5.58	79.92	49.40	65.96	73.64
TESTNI	13.43	72.07	6.37	86.10	59.65	74.01	79.71
Set	26.85	82.31	5.67	81.10	51.58	68.76	76.46

As Tables 2.3 and 2.4 show, the *training* and *validation* partitions have a similar type of summaries regarding their degree of abstractivity. The summaries in the test partitions, except the TESTNI set for Spanish, also show similar degree of abstractivity as the previous partitions.

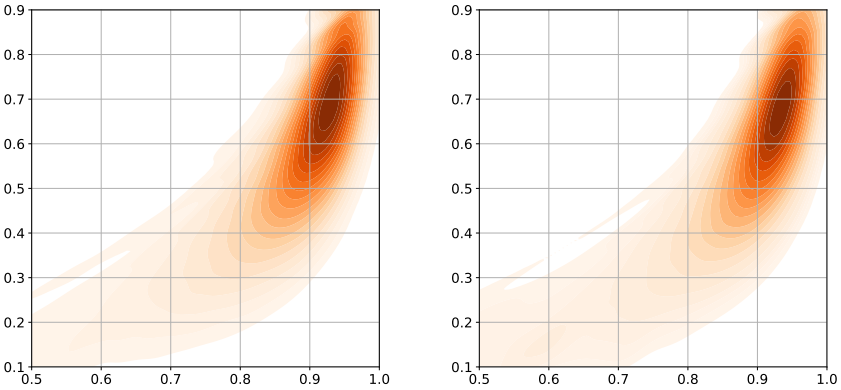
In order to better characterize the corpus, we also present in Figure 2.1 the distributions of the samples by combining the values of *extractive fragment coverage* and *extractive fragment density* of their summaries, and in Figure 2.2 the distribution of the samples by combining the values of *abstractivity$_p$ (p=2)* and *novel 2-grams*. These plots help to identify visually the degree of abstractivity of the summaries in the Catalan and Spanish sets. On the one hand, the metrics used in the first plots correlate negatively with the abstractitivy; thus, higher abstractivity is shown in the partition when the distribution is centered around the bottom left corner of the plot (where the values are lower on both metrics). On the other hand, the second plots correlate positively with the abstractivity; thus, the distributions are centered near the right top corner if the summaries are highly abstractive. Finally, we should point that due to the outliers, the distributions were hard to visualize. Hence, we exclude the 10% with the lowest values and the 10% with the highest values.

Figure 2.1 shows that the Catalan set mainly contains summaries with short extractive fragments since the distribution centers in 75% of coverage and a density



(a) Catalan (b) Spanish

Figure 2.1: Distribution of the samples for the Catalan and Spanish sets. `x-axis:` Extractive Fragment Coverage, `y-axis:` Extractive Fragment Density.



(a) Catalan (b) Spanish

Figure 2.2: Distribution of the samples for the Catalan and Spanish sets. `x-axis:` Abstractivity$_p$ (p=2), `y-axis:` Novel 2-grams.

lower than 2. Also, we observe that the distribution tends to go up and right; thus, the samples of the set diversify to less abstractive summaries. In the case of Spanish, we observe that the extractive fragments are longer than in the first language due to the higher density, and also, the distribution centers in the 85% of coverage, which indicates that the summaries in the Spanish set reuse more words from the article than in the Catalan set. However, the distribution tends to go down and left, which indicates a big presence of abstractive summaries in this set.

Figure 2.2 helps to show the diversity of the samples by combining abstractivity$_p$ ($p = 2$) and novel 2-grams, which brings us more information. Although in Figure 2.1

the distributions were different from language to language, in this figure, we observe that the two sets are similar regarding these two metrics; note that the darker zones follow the same pattern around the same range of values.

Based on Tables 2.3 and 2.4 and Figures 2.1 and 2.2, it can be concluded that the DACSA corpus provides samples that do not contain a predominance of extractive summaries, and show great diversity regarding their degree of abstractivity.

2.6 Summarization models and performance results

We evaluate several summarization systems to understand the challenges posed by the DACSA dataset for summarization tasks. We consider both extractive and abstractive models, along with an extractive oracle to show an upper bound of the extractive performance in the corpora.

Table 2.5: Average F_1 scores of the models in the summarization task in Catalan.

Partition	Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Ls	BERTScore
TESTI	mBART	28.59	11.89	23.00	23.39	72.03
	mT5	27.01	10.70	21.81	22.12	71.55
	SHANN	23.56	9.64	17.31	19.16	68.85
	TextRank	16.54	5.62	11.98	15.33	58.90
	Lead-2	23.41	9.33	17.28	19.04	68.96
	Oracle	41.68	25.53	36.29	36.64	75.87
TESTNI	mBART	27.46	11.04	21.13	22.01	70.33
	mT5	27.00	11.28	21.27	22.01	70.56
	SHANN	30.40	9.64	17.31	19.16	69.72
	TextRank	17.16	5.83	12.27	15.93	60.36
	Lead-2	31.44	15.74	23.63	26.32	70.30
	Oracle	47.16	29.44	40.23	41.82	75.86

Extractive systems: Lead-k, TextRank (Mihalcea and Tarau, 2004) and SHANN (J.-Á. González et al., 2019) have been evaluated. Lead-k is a heuristic that extracts the first k sentences of a text, being especially well suited to summarize newspaper articles. TextRank is a graph-based system inspired by PageRank, where nodes represent sentences, and edges measure similarities in terms of shared words. Finally, SHANN is a supervised system based on siamese hierarchical attentional networks. The document sentences are scored using sentence-level attentions and those with highest scores are extracted to build the summary. As the average number of sentences in the summaries of DACSA is near to two, we extracted two document sentences

Table 2.6: Average F_1 scores of the models in the summarization task in Spanish

Partition	Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Ls	BERTScore
TESTI	mBART	31.09	13.56	24.67	25.48	72.25
	mT5	31.72	14.54	25.76	26.31	72.86
	SHANN	26.93	11.84	20.07	22.12	69.86
	TextRank	14.13	4.27	8.13	13.15	57.83
	Lead-2	29.00	14.39	22.56	24.45	71.03
	Oracle	46.04	30.12	40.85	41.37	77.45
TESTNI	mBART	30.66	12.08	23.13	23.89	71.07
	mT5	30.61	12.36	23.53	24.05	71.26
	SHANN	35.55	15.22	24.63	27.41	70.83
	TextRank	21.78	6.13	11.77	18.97	54.54
	Lead-2	36.64	16.79	26.07	28.64	71.81
	Oracle	46.49	25.50	36.84	37.54	74.85

by using the extractive systems. We built the extractive systems upon code that is available on Github (Barrios et al., 2016), (J.-Á. González et al., 2019).

Abstractive systems: we considered two representative models with high performance on abstractive summarization, based on encoder-decoder architectures with Transformers as backbone: BART and T5. Due to there are neither BART nor T5 models pretrained from scratch for the Spanish and Catalan languages, we finetuned and evaluated their multilingual variants, mBART [1] and mT5[2]. It should be noted that, although both of them considered the Spanish language during pretraining, the Catalan language is not represented in the case of mBART, as this language is not contained in the CC25 dataset. We built the abstractive systems using the HugginFace toolkit (Wolf et al., 2020).

Oracle: we implemented an extractive oracle that aligns each summary sentence with the most similar document sentence using ROUGE. The aligned document sentences are concatenated to build the oracle summary.

In order to evaluate the models, we use ROUGE and BERTScore metrics. ROUGE-1, ROUGE-2 and ROUGE-L are reported to measure lexical overlapping, while BERTScore is used to measure semantic similarity.

Tables 2.5 and 2.6 show the performance results of the different models on the Cata-

[1]HuggingFace finetuned mBART models:
`ELiRF/mbart-large-cc25-dacsa-ca` `ELiRF/mbart-large-cc25-dacsa-es`

[2]HuggingFace finetuned mT5 models:
`ELiRF/mt5-base-dacsa-ca` `ELiRF/mt5-base-dacsa-es`

lan and Spanish DACSA TESTI and TESTNI sets in terms of ROUGE and BERTScore metrics. The oracle outperforms the other systems by a large margin. The worse results obtained by the oracle are in the Catalan TESTI, showing that this partition is the most abstractive test partition in the DACSA corpus. Generally, extractive systems are worse in the TESTI than in the TESTNI, which suggests a higher extractivity in TESTNI than in TESTI. The high results of Lead-2, especially in the TESTNI sets, show that there is a positional bias in these sets.

2.7 Conclusions

Languages other than English have a lack of resources for learning models based on deep learning. This is true for endangered languages but it is also true even for those languages that have millions of speakers but are minority worldwide such as Catalan. In this work, we describe the construction of a corpus of Catalan and Spanish newspapers, the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) corpus. We have included an analysis of the corpus using a set of well-known metrics in the summarization field in order to characterize the corpus. This characterization shows that DACSA provides samples that do not contain a predominance of extractive summaries, and show great diversity regarding their degree of abstractivity. We have also carried out an evaluation of the performance of some extractive and abstractive summarization systems on the DACSA corpus that could be used for benchmarking. To our knowledge, the DACSA corpus is the largest summarization dataset for Catalan and Spanish languages and is freely available for research purposes.

Chapter 2. References

Barrios, Federico et al. (2016). "Variations of the Similarity Function of TextRank for Automated Summarization". In: *CoRR* abs/1602.03606. arXiv: 1602.03606. URL: http://arxiv.org/abs/1602.03606 (cit. on p. 33).

Bommasani, Rishi and Claire Cardie (Nov. 2020). "Intrinsic Evaluation of Summarization Datasets". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8075–8096. DOI: 10.18653/v1/2020.emnlp-main.649. URL: https://aclanthology.org/2020.emnlp-main.649 (cit. on pp. 24, 26, 29).

Chen, Yen-Chun and Mohit Bansal (July 2018). "Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 675–686. DOI: 10.18653/v1/P18-1063 (cit. on p. 24).

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423 (cit. on p. 25).

Dong, Yue et al. (Oct. 2018). "BanditSum: Extractive Summarization as a Contextual Bandit". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3739–3748. DOI: 10.18653/v1/D18-1409 (cit. on p. 24).

González, J.-A. et al. (2019). "Summarization of Spanish Talk Shows with Siamese Hierarchical Attention Networks". In: *Applied Sciences 2019* 9.18. DOI: 10.3390/app9183836 (cit. on pp. 23, 24).

González, José Ángel et al. (2020). "Extractive summarization using siamese hierarchical transformer encoders". In: *Journal of Intelligent & Fuzzy Systems* 39. 2, pp. 2409–2419. ISSN: 1875-8967. DOI: 10.3233/JIFS-179901 (cit. on p. 24).

González, José-Ángel et al. (2019). "Siamese hierarchical attention networks for extractive summarization". In: *Journal of Intelligent & Fuzzy Systems* 36.5, pp. 4599–4607. DOI: 10.3233/JIFS-179011 (cit. on pp. 24, 32, 33).

Grusky, Max, Mor Naaman, and Yoav Artzi (June 2018). "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 708–719. DOI: 10.18653/v1/N18-1065. URL: https://aclanthology.org/N18-1065 (cit. on pp. 22–24, 26, 29).

Hasan, Tahmid et al. (Aug. 2021). "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 4693–4703. DOI: 10.18653/v1/2021.findings-acl.413. URL: https://aclanthology.org/2021.findings-acl.413 (cit. on p. 23).

Hermann, Karl Moritz et al. (2015). "Teaching Machines to Read and Comprehend". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 1693–1701. URL: http://dl.acm.org/citation.cfm?id=2969239.2969428 (cit. on p. 22).

Ive, Julia, Pranava Madhyastha, and Lucia Specia (Nov. 2019). "Deep Copycat Networks for Text-to-Text Generation". 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 3225–3234. DOI: 10.18653/v1/D19-1318 (cit. on p. 24).

Koto, Fajri, Jey Han Lau, and Timothy Baldwin (Dec. 2020). "Liputan6: A Large-scale Indonesian Dataset for Text Summarization". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 598–608. URL: https://aclanthology.org/2020.aacl-main.60 (cit. on p. 22).

Kryściński, Wojciech et al. (Oct. 2018). "Improving Abstraction in Text Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1808–1817. DOI: 10.18653/v1/D18-1207. URL: https://aclanthology.org/D18-1207 (cit. on pp. 24, 26, 29).

Lan, Zhenzhong et al. (2020). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: https://openreview.net/forum?id=H1eA7AEtvS (cit. on p. 25).

Lewis, Mike et al. (July 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*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703 (cit. on p. 25).

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013 (cit. on pp. 24, 25).

Liu, Yang and Mirella Lapata (Nov. 2019). "Text Summarization with Pretrained Encoders". 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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740. DOI: 10.18653/v1/D19-1387. URL: https://aclanthology.org/D19-1387 (cit. on pp. 24, 25).

Liu, Yinhan, Jiatao Gu, et al. (Nov. 2020). "Multilingual Denoising Pre-training for Neural Machine Translation". In: *Transactions of the Association for Computational Linguistics* 8, pp. 726–742. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00343. URL: https://www.aclweb.org/anthology/2020.tacl-1.47 (cit. on pp. 24, 25).

Liu, Yinhan, Myle Ott, et al. (July 2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692. arXiv: 1907.11692 [cs.CL]. URL: http://arxiv.org/abs/1907.11692 (cit. on p. 25).

Mendes, Afonso et al. (2019). "Jointly Extracting and Compressing Documents with Summary State Representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 3955–3966. DOI: 10.18653/v1/N19-1397 (cit. on p. 24).

Mihalcea, Rada and Paul Tarau (July 2004). "TextRank: Bringing Order into Text". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Ed. by Dekang Lin and Dekai Wu. Barcelona, Spain: Association for Computational Linguistics, pp. 404–411. URL: https://aclanthology.org/W04-3252 (cit. on p. 32).

Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou (2017). "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, pp. 3075–3081. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10958/10817 (cit. on p. 24).

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (Oct. 2018a). "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: Association for Computational Linguistics, pp. 1797–1807. DOI: 10.18653/v1/D18-1206. URL: https://aclanthology.org/D18-1206 (cit. on pp. 22, 25).

— (June 2018b). "Ranking Sentences for Extractive Summarization with Reinforcement Learning". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1747–1759. DOI: 10.18653/v1/N18-1158 (cit. on p. 24).

Paulus, Romain, Caiming Xiong, and Richard Socher (May 2018). "A Deep Reinforced Model for Abstractive Summarization". In: *Proceedings of the 6th International Conference on Learning Representations ICLR*. Vancouver, BC, Canada: OpenReview.net, pp. 1–13. URL: https://openreview.net/forum?id=HkAClQgA- (cit. on p. 24).

Raffel, Colin et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html (cit. on p. 25).

Scialom, Thomas et al. (Nov. 2020). "MLSUM: The Multilingual Summarization Corpus". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8051–8067. DOI: 10.18653/v1/2020.emnlp-main.647. URL: https://aclanthology.org/2020.emnlp-main.647 (cit. on p. 23).

See, Abigail, Peter J. Liu, and Christopher D. Manning (July 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)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. URL: https://aclanthology.org/P17-1099 (cit. on p. 24).

Segarra-Soriano, Encarnación et al. (July 2022). "DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 5931–5943. DOI: 10.18653/v1/2022.naacl-main.434. URL: https://aclanthology.org/2022.naacl-main.434 (cit. on p. 19).

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Proceedings of the 27th International Conference on*

Neural Information Processing Systems - Volume 2. NIPS'14. Montreal, Canada: MIT Press, pp. 3104–3112 (cit. on p. 24).

Varab, Daniel and Natalie Schluter (Nov. 2021). "MassiveSumm: a very large-scale, very multilingual, news summarisation dataset". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10150–10161. DOI: 10.18653/v1/2021.emnlp-main.797. URL: https://aclanthology.org/2021.emnlp-main.797 (cit. on p. 23).

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., pp. 5998–6008. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 25).

Wenzek, Guillaume et al. (May 2020). "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4003–4012. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.494 (cit. on p. 25).

Wolf, Thomas et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6 (cit. on p. 33).

Xue, Linting et al. (June 2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41 (cit. on pp. 24, 25).

Yao, Kaichun et al. (2018). "Deep reinforcement learning for extractive document summarization". In: *Neurocomputing* 284, pp. 52–62. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2018.01.020 (cit. on p. 24).

Zhang, Jingqing et al. (July 2020). "PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. ICML20. PMLR. JMLR.org, pp. 11328–11339. URL: https://proceedings.mlr.press/v119/zhang20ae.html (cit. on p. 25).

Zhang, Tianyi et al. (2020). "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations*. OpenReview.net. URL: https://openreview.net/forum?id=SkeHuCVFDr (cit. on p. 25).

Zhang, Xingxing et al. (Oct. 2018). "Neural Latent Extractive Document Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 779–784. DOI: 10.18653/v1/D18-1088 (cit. on p. 24).

Zhong, Ming et al. (July 2020). "Extractive Summarization as Text Matching". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6197–6208. DOI: 10.18653/v1/2020.acl-main.552. URL: https://aclanthology.org/2020.acl-main.552 (cit. on p. 25).

Chapter 2. Appendix

2.A Statistics of DACSA

We show in Tables 2.A.1 and 2.A.2 a more detailed view of the statistics of the DACSA corpus, distinguishing among the sources from which it was built. The sources that were only considered in the TESTni partitions are marked with an asterisk.

Table 2.A.1: Statistics by source in the Catalan set.

			Article			Summary		
Source	#Docs	Tokens	Vocabulary size	Sents per doc	Words per sent	Vocabulary size	Sents per doc	Words per sent
CA01	238 233	114 500 016	614 146	17.68	27.19	115 954	1.14	20.16
CA02	194 697	105 119 526	621 612	19.99	27.01	112 904	1.28	19.14
CA03	137 447	63 683 416	485 286	14.99	30.92	91 975	1.05	22.65
CA04	56 827	24 891 291	276 720	14.84	29.52	58 071	1.21	17.52
CA05	44 381	26 977 332	277 225	18.04	33.69	55 216	1.15	23.86
CA06	35 763	17 181 460	202 931	11.31	42.49	42 289	1.05	22.79
CA07*	7104	3 800 842	83 942	18.04	29.66	19 267	1.02	26.51
CA08*	5882	9 414 192	185 977	66.04	24.24	31 006	2.54	24.84
CA09*	4850	2 667 185	102 024	23.61	23.29	19 584	1.16	28.05
Set	725 184	368 235 260	1 326 343	17.71	28.67	223 978	1.17	20.59

Table 2.A.2: Statistics by source in the Spanish set.

			Article			Summary		
Source	#Docs	Tokens	Vocabulary size	Sents per doc	Words per sent	Vocabulary size	Sents per doc	Words per sent
ES01	550 148	420 786 144	1 473 628	31.36	24.39	210 079	1.40	19.02
ES02	342 045	174 411 220	907 312	16.66	30.61	148 271	1.06	22.34
ES03	196 410	93 755 039	622 073	15.40	31.00	110 728	1.02	20.59
ES04	168 065	105 628 806	659 054	23.35	26.92	112 908	1.09	22.30
ES05	148 053	105 453 102	626 058	28.35	25.13	109 546	1.47	20.46
ES06	116 561	93 956 373	524 177	26.16	30.81	169 025	1.27	43.20
ES07	107 162	70 944 634	470 244	19.90	33.26	87 901	1.29	25.27
ES08	99 098	65 352 628	495 148	25.03	26.35	81 654	1.25	18.38
ES09	81 947	42 825 867	363 075	15.54	33.63	71 913	1.03	22.41
ES10	74 024	57 782 514	470 826	30.28	25.78	81 793	1.31	20.23
ES11*	70 193	29 692 261	272 248	11.06	38.26	84 898	1.22	44.48
ES12	57 235	28 198 002	294 175	16.06	30.68	58 580	1.21	19.49
ES13	35 163	20 156 337	260 690	19.22	29.83	50 556	1.15	21.20
ES14	35 112	28 408 974	309 194	30.48	26.55	78 751	1.18	28.35
ES15*	17 379	10 099 958	153 598	16.82	34.54	41 512	1.85	26.89
ES16*	16 965	13 791 564	166 446	28.26	28.77	29 955	1.07	25.18
ES17*	2450	4 545 924	135 761	74.97	24.75	23 588	3.16	26.72
ES18*	1374	641 752	39 094	17.08	27.34	12 365	1.98	29.43
ES19*	643	398 834	26 797	17.73	34.99	2495	1.04	16.02
ES20*	467	233 873	22 699	18.70	26.78	3857	1.22	24.23
ES21*	155	199 140	19 750	39.06	32.89	2098	1.91	21.79
Set	2 120 649	1 367 262 946	3 189 783	23.44	27.50	516 307	1.24	22.95

2.B Abstractivity in DACSA

We show in Tables 2.B.1 and 2.B.2 a fine-grained view of the abstractivity of the DACSA corpus, distinguishing among the sources from which it was built.

Table 2.B.1: Average abstractivity metrics by source in the Catalan set.

Source	Compression	Coverage	Density	Abstractivity$_p$ (p=2)	Novel 2-grams	Novel 3-grams	Novel 4-grams
CA01	24.21	81.26	3.47	84.21	55.13	72.93	81.36
CA02	23.62	80.71	3.28	85.43	56.90	74.99	82.98
CA03	20.81	79.95	3.27	84.88	56.73	73.92	82.14
CA04	21.50	79.54	3.27	83.51	57.03	73.92	81.85
CA05	24.76	83.27	5.94	76.76	47.67	63.16	70.94
CA06	21.88	82.45	4.53	80.48	50.73	67.09	75.12
CA07*	20.22	80.70	3.02	87.41	56.61	74.61	83.31
CA08*	31.01	72.49	2.04	95.75	65.60	85.19	92.28
CA09*	21.09	88.00	13.48	62.98	34.44	46.63	53.37
Set	23.11	80.09	3.62	83.95	55.35	72.80	80.96

Table 2.B.2: Average abstractivity metrics by source in the Spanish set.

Source	Compression	Coverage	Density	Abstractivity$_p$ (p=2)	Novel 2-grams	Novel 3-grams	Novel 4-grams
ES01	35.07	83.64	7.26	81.25	52.32	71.16	79.22
ES02	22.65	83.24	5.46	77.25	49.24	65.21	72.49
ES03	23.89	81.52	3.60	82.53	54.06	71.48	79.90
ES04	28.31	83.78	5.54	77.77	48.99	65.27	72.84
ES05	25.88	79.10	3.55	86.94	57.40	75.30	82.86
ES06	16.50	83.51	6.48	85.33	46.31	63.20	71.21
ES07	22.55	85.31	6.53	79.31	44.69	61.50	69.70
ES08	31.95	80.76	3.51	83.57	55.76	73.63	81.43
ES09	24.04	80.37	3.07	85.79	56.72	74.92	83.32
ES10	33.36	82.58	3.98	83.60	53.33	71.91	80.12
ES11*	8.50	63.03	1.65	96.53	73.02	88.20	93.65
ES12	23.33	81.02	5.92	77.85	53.15	69.51	76.67
ES13	26.35	85.67	7.90	67.78	42.31	55.97	62.51
ES14	26.41	89.09	9.50	70.79	29.76	40.31	46.88
ES15*	11.94	94.27	24.19	51.47	20.16	27.35	30.80
ES16*	32.02	84.84	4.22	83.45	48.88	68.16	77.59
ES17*	28.10	68.50	11.03	86.13	61.74	76.20	80.81
ES18*	10.83	94.68	39.75	37.55	14.05	18.49	21.77
ES19*	38.80	76.20	5.07	68.91	53.72	64.12	67.99
ES20*	21.60	85.98	11.34	69.44	42.00	56.84	63.79
ES21*	39.51	78.64	4.10	90.11	56.33	73.82	81.75
Set	26.85	82.31	5.67	81.10	51.58	68.76	76.46

Chapter 3

NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish

Vicent Ahuir, Lluís-F. Hurtado, José Ángel González, and Encarna Segarra (2021). "NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish". In: *Applied Sciences* 11.21. ISSN: 2076-3417. DOI: 10.3390/app11219872. URL: https://www.mdpi.com/2076-3417/11/21/9872

Impact Index

In 2021, the *Applied Sciences* journal was listed at the 2nd quantile (Q2) in the Journal Citation Report (JCR) at *Engineering, Multidisciplinary* category.

Abstract

Most of the models proposed in the literature for abstractive summarization are generally suitable for the English language but not for other languages. Multilingual models were introduced to address that language constraint, but despite their applicability being broader than that of the monolingual models, their performance is typically lower, especially for minority languages like Catalan. In this paper, we present a monolingual model for abstractive summarization of textual content in the Catalan language. The model is a Transformer encoder-decoder which is pretrained and fine-tuned specifically for the Catalan language using a corpus of newspaper articles. In the pretraining phase, we introduced several self-supervised tasks to specialize the model on the summarization task and to increase the abstractivity of the generated summaries. To study the performance of our proposal in languages with higher resources than Catalan, we replicate the model and the experimentation for the Spanish language. The usual evaluation metrics, not only the most used ROUGE measure but also other more semantic ones such as BERTScore, do not allow to correctly evaluate the abstractivity of the generated summaries. In this work, we also present a new metric, called *content reordering*, to evaluate one of the most common characteristics of abstractive summaries, the rearrangement of the original content. We carried out an exhaustive experimentation with DACSA (a large Dataset for Automatic summarization of Catalan and Spanish newspaper Articles) to compare the performance of the monolingual models proposed in this work with two of the most widely used multilingual models in text summarization, mBART and mT5. The experimentation results support the quality of our monolingual models, especially considering that the multilingual models were pretrained with many more resources than those used in our models. Likewise, it is shown that the pretraining tasks helped to increase the degree of abstractivity of the generated summaries. To our knowledge, this is the first work that explores a monolingual approach for abstractive summarization both in Catalan and Spanish.

3.1 Introduction

The purpose of the summarization process is to condense the most relevant information from a document or a set of documents into a small number of sentences. This process can be performed in an extractive or an abstractive way. While extractive summarization consists of identifying and copying those sentences in the original document that contain the most remarkable and useful information, abstractive summaries require abstractive actions that must be mastered. In this way, summaries are not mere clippings of the original documents; rather, abstractive summarizations are created by choosing the most important phrases of the documents and paraphrasing that content, creating a combination of some phrases, introducing new words, searching for synonyms, creating generalizations or specifications of some words or reordering content. All these actions must be done while preserving the linguistic cohesion and the coherence of the information (National Information Standards Organization, 1997; Jing, 2002; Rane and Govilkar, 2019; Verma, Pal, and Om, 2019; Widyassari et al., 2020).

Nowadays, Transformer-based language models excel in text generation, especially due to the transfer learning paradigm, by means of self-supervised pretraining on large text corpora, and later fine-tuning on downstream tasks. The generation capabilities achieved by these models boosted the state of the art in automatic summarization. However, most of the models proposed in the literature, such as BART (Lewis et al., 2020), PEGASUS (J. Zhang et al., 2020), or T5 (Raffel et al., 2020) are intended to the English language and are not directly applicable to other languages. Multilingual models such as mBART (Yinhan Liu et al., 2020) or mT5 (Xue et al., 2021) were also studied in the literature to address that language constraint, but despite their applicability being broader than that of the monolingual models, their performance is typically lower, especially on languages that are underrepresented in the pretraining corpora, or differ so much in linguistic terms from the most represented languages (Cañete et al., 2020; Martin et al., 2020; Pires, Schlinger, and Garrette, 2019; Virtanen et al., 2019)

For minority languages like Catalan, the data resources available are much lower than other languages like English, Chinese, or Spanish. Additionally, the multilingual models typically either do not include data of minority languages, or if they do, its proportion in the pretraining sets is much lower than those of the majority languages. In this work, we hypothesize that monolingual models are a better choice for those minority languages, such as the Catalan language, which are underrepresented in the pretraining datasets of the multilingual models, but for which reasonable amounts of data are available.

In this work, a BART-like summarization model for the Catalan language is

pretrained from scratch, and then fine-tuned on the summarization task. During the pretraining step, we include several self-supervised tasks to enhance the of the degree of abstractivity of the generated summaries. Furthermore, to test our hypothesis about monolingual models, we compare the performance of our proposal against well-known pretrained multilingual models such as mBART and mT5. It is also interesting to study the performance of our proposal in languages with higher resources than Catalan. For this reason, we replicate the model and the experimentation for the Spanish language to extract conclusions about abstractivity and monolingual models in two different languages.

We performed experimentation on the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) corpus ("DACSA: a Dataset for Automatic summarization of Catalan and Spanish newspaper Articles" n.d.) This corpus provides pairs of news article and its summary from different journals in the Catalan and the Spanish languages. The experimental results show that the monolingual models generalize better than the multilingual ones, obtaining a more stable summarization performance on the test partitions of the DACSA dataset. The provided experimentation also illustrate the improvements in abstractivity as a result of the addition of the pretraining tasks. We analyze the abstractivity of the models through the use of abstractivity indicators (Jing, 2002). Following some of these indicators, which correspond to actions done by professional summary writers, we quantify the degree of abstractivity of the generated summaries as the summaries generated by the models. One of the common actions when a person writes an abstractive summary is to rearrange the information from the original document. To our knowledge, no metrics were proposed for this specific action. For this reason, in this work the *content reordering* metric, which aims to quantify the rearrangement degree of the information in the summary with respect to the document, is proposed.

The contributions of this work are the following:

- A monolingual abstractive text summarization model, News Abstract Summarization for Catalan (NASca), is proposed. This model, based on the BART architecture (Lewis et al., 2020), is pretrained with several self-supervised tasks to improve the abstractivity of the generated summaries. For fine-tuning the model, a corpus of online newspapers is used (DACSA).

- An evaluation of the performance of the model on the summarization task and an evaluation of the degree of abstractivity of its generated summaries are presented. We compare the results of each NAS model with the results obtained by the summarization models based on well-known multilingual language models (mBART (Yinhan Liu et al., 2020) and mT5 (Xue et al., 2021)) fine-tuned for the summarization task for each language using the DACSA corpus.

- A text summarization model with the same pretraining process than NASca is also trained and evaluated for Spanish, News Abstract Summarization for Spanish (NASes).
- The *content reordering* metric is proposed, which helps to quantify if the extractive content within the abstractive summary is written in a different order than in the document.

The monolingual models, NASca (https://huggingface.co/ELiRF/NASCA, accessed on 21 October 2021) and NASes (https://huggingface.co/ELiRF/NASES, accessed on 21 October 2021), proposed in this work were publicly release through HuggingFace model hub (Wolf et al., 2020).

3.2 Related Work

Abstractive summarization works normally focused on the creation of models using approaches different to those used for extractive summarization (Yang Liu and Lapata, 2019; Nallapati, Zhai, and Zhou, 2017; Nallapati, Zhou, et al., 2016; Rush, Chopra, and Weston, 2015; See, P. J. Liu, and Manning, 2017; Zhong et al., 2020). Recently, abstractive summarizers became ubiquitous due to their powerful generation capabilities, achieved by using encoder-decoder architectures with Transformers (Vaswani et al., 2017) as backbone, and by pretraining them with self-supervised language modeling tasks on massive text corpora. This kind of models, especially PEGASUS (J. Zhang et al., 2020), BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and ProphetNet (Qi et al., 2020), fine-tuned for summarization tasks, are the state of the art in abstractive summarization benchmarks.

While all these models are nearly identical regarding their architecture, they mainly differ in the self-supervised tasks used in the pretraining stage. In some cases, such as BART, T5, and ProphetNet, these tasks aims the models to learn general aspects of the language, e.g., by masking tokens or reordering sentences. More specifically, BART is pretrained to reconstruct masked spans (text infilling) and to arrange sentences in the original order after being permuted (sentence permutation). Similarly, T5 is pretrained on encoder-decoder masked language modeling, in order to address universally all text-based language problems in a text-to-text format. Regarding ProphetNet, it is pretrained on future n-gram prediction to encourage the model to plan for future tokens instead of the next token, which prevents overfitting on strong local correlations. However, in other cases such as PEGASUS, the self-supervised tasks intentionally resemble the summarization task to encourage whole-document understanding and summary-like generation. In contrast to the previous models, PEGASUS is trained with Gap Sentences Generation (GSG), which consists

of reconstructing the sentences that maximize the ROUGE with respect to the whole document. In this way, the authors of PEGASUS hypothesize that GSG is more suitable for abstractive summarization than other pretraining strategies, as it closely resembles the downstream task.

Other works are also based on strategies that involve pretraining to improve the abstractivity of the generated summaries. For instance, in (Magooda and Litman, 2020), domain transfer and data synthesis techniques by using pretrained models are explored to improve the performance of abstractive summarization models in low-resource scenarios. Also, the authors of (Kryściński et al., 2018) propose to use pretrained language models to incorporate prior knowledge about language generation, which provides results comparable to state-of-the-art models in terms of ROUGE, while increasing the level of abstraction of the generated summaries, measured in terms of n-gram overlapping. Finally, in (Zou et al., 2020) a combination of several pretraining tasks is introduced to tailor the models to abstractive summarization, improving performance upon other Transformer-based models with significantly less pretraining data. Specifically, three tasks were proposed for pretraining: sentence reordering, next segment generation and masked document generation. While sentence reordering and masked document generation are identical to the text infilling and sentence permutation tasks used in BART, next segment generation aims to complete a document given a prefix of that document. Therefore, our work is similar to (Zou et al., 2020) in the sense that we combine the pretraining tasks of BART and PEGASUS to improve the abstractive skills of monolingual models trained for Catalan and Spanish.

All the models and proposals discussed in this section are intended for the English language, however, there are many other languages that deserve attention. Some efforts were done to consider other languages along with the English language by means of multilingual models such as mBART (Yinhan Liu et al., 2020) or mT5 (Xue et al., 2021). Although these efforts are very convenient and useful in many cases, the performance of the multilingual models is typically lower on languages that are underrepresented in the pretraining data or differ so much, in linguistic terms, from the most represented languages (Pires, Schlinger, and Garrette, 2019; Virtanen et al., 2019). Learning monolingual models from scratch was extensively explored for language understanding by means of pretraining monolingual BERT models, with excellent results in many languages such as French (Le et al., 2020; Martin et al., 2020), Dutch (Vries et al., 2019), or Spanish (Cañete et al., 2020; González, Hurtado, and Pla, 2021). However, monolingual pretraining in languages other than English is still unexplored for language generation tasks such as abstractive summarization. To our knowledge, this is the first work that explores a monolingual approach for abstractive summarization both in Catalan and Spanish.

Table 3.1: Statistics of Catalan set. Sources marked with * were not used for training the models.

			Article			Summary		
Source	**Docs**	**Tokens**	|V|	**Sents per doc**	**Words per sent**	|V|	**Sents per doc**	**Words per sent**
#1	238,233	114,500,016	614,146	17.68	27.19	115,954	1.14	20.16
#2	194,697	105,119,526	621,612	19.99	27.01	112,904	1.28	19.14
#3	137,447	63,683,416	485,286	14.99	30.92	91,975	1.05	22.65
#4	56,827	24,891,291	276,720	14.84	29.52	58,071	1.21	17.52
#5	44,381	26,977,332	277,225	18.04	33.69	55,216	1.15	23.86
#6	35,763	17,181,460	202,931	11.31	42.49	42,289	1.05	22.79
#7*	7104	3,800,842	83,942	18.04	29.66	19,267	1.02	26.51
#8*	5882	9,414,192	185,977	66.04	24.24	31,006	2.54	24.84
#9*	4,850	2,667,185	102,024	23.61	23.29	19,584	1.16	28.05
Set	725,184	368,235,260	1,326,343	17.71	28.67	223,978	1.17	20.59

3.3 Newspapers Summarization Corpus

As stated above, the models proposed in this work are focused on the specific domain of newspaper articles. To train the models, the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) ("DACSA: a Dataset for Automatic summarization of Catalan and Spanish newspaper Articles" n.d.) corpus was used. This corpus provides pairs of news article and its summary from different newspapers for both, the Catalan and the Spanish languages.

Regarding the Catalan set, there are 725,184 sample pairs from 9 newspapers, and their distribution is shown in the Table 3.1:

Regarding the Spanish set, the corpus provides 2,120,649 sample pairs from 21 newspapers, distributed as it is detailed in the Table 3.2:

When the distributions of the samples on both subsets are analyzed, the amount of samples by source is far from being homogeneous. If these distributions preserve over the partitions (training, validation, and test set), the models will focus their learning on the newspapers that are predominant. To avoid this bias and achieve more general models, the test and validation sets were created in a way that ensured that all newspapers had roughly the same number of samples on those sets. To achieve this balance in the validation and test sets, the sources with less samples were discarded. In this way, it is guaranteed that all sources represent at least 5% of samples in each one of these two sets. The sources that were excluded are marked with an asterisk in the Tables 3.1 and 3.2.

Table 3.2: Statistics of Spanish set. Sources marked with * were not used for training the models.

				Article			Summary	
Source	**Docs**	**Tokens**	|V|	**Sents per doc**	**Words per sent**	|V|	**Sents per doc**	**Words per sent**
#1	550,148	420,786,144	1,473,628	31.36	24.39	210,079	1.40	19.02
#2	342,045	174,411,220	907,312	16.66	30.61	148271	1.06	22.34
#3	196,410	93,755,039	622,073	15.40	31.00	110,728	1.02	20.59
#4	168,065	105,628,806	659,054	23.35	26.92	112,908	1.09	22.30
#5	148,053	105,453,102	626,058	28.35	25.13	109,546	1.47	20.46
#6	116,561	93,956,373	524,177	26.16	30.81	169,025	1.27	43.20
#7	107,162	70,944,634	470,244	19.90	33.26	87,901	1.29	25.27
#8	99,098	65,352,628	495148	25.03	26.35	81654	1.25	18.38
#9	81,947	42,825,867	363,075	15.54	33.63	71,913	1.03	22.41
#10	74,024	57,782,514	470,826	30.28	25.78	81793	1.31	20.23
#11*	70,193	29,692,261	272,248	11.06	38.26	84,898	1.22	44.48
#12	57,235	28,198,002	294,175	16.06	30.68	58,580	1.21	19.49
#13	35163	20,156,337	260,690	19.22	29.83	50,556	1.15	21.20
#14	35112	28,408,974	309,194	30.48	26.55	78751	1.18	28.35
#15*	17379	10,099,958	153,598	16.82	34.54	41512	1.85	26.89
#16*	16,965	13,791,564	166,446	28.26	28.77	29,955	1.07	25.18
#17*	2450	4,545,924	135,761	74.97	24.75	23,588	3.16	26.72
#18*	1374	641,752	39,094	17.08	27.34	12,365	1.98	29.43
#19*	643	398,834	26,797	17.73	34.99	2495	1.04	16.02
#20*	467	233,873	22,699	18.70	26.78	3857	1.22	24.23
#21*	155	199,140	19,750	39.06	32.89	2098	1.91	21.79
Set	2,120,649	1,367,262,946	3,189,783	23.44	27.50	516307	1.24	22.95

The three sets for Catalan include 6 of the 9 newspapers, creating a training set that contains 636,596 samples and 35,376 samples for validation and test sets. In the case of Spanish, the three sets are composed of 13 of the 21 newspapers provided in the Spanish set of DACSA: the training set contains 1,802,919 samples, and the validation and test sets contain 104,052 samples each.

All the sources excluded were used as a separate test set. This partition allows to evaluate the generalization capabilities of the models. In this work, we refer to the test set with newspapers included in the training set as TESTI and to the test set that contains newspapers not included in the training set as TESTNI. The statistics of all the sets are shown in Tables 3.3 and 3.4.

Table 3.3: Statistics of partitions for Catalan language.

				Article			Summary	
Partition	Docs	**Tokens**	|V|	**Sents per doc**	**Words per sent**	|V|	**Sents per doc**	**Words per sent**
Training	636,596	316,817,625	1,206,292	17.39	28.62	206,616	1.17	20.36
Validation	35,376	17,831,029	258,999	16.17	31.17	51,940	1.15	20.93
TESTi	35,376	17,704,387	262,148	16.13	31.03	51,958	1.15	20.89
TESTni	17,836	15,882,219	247,154	35.38	25.17	45,997	1.56	25.93

Table 3.4: Statistics of partitions for Spanish language.

				Article			Summary	
Partition	**Docs**	**Tokens**	|V|	**Sents per doc**	**Words per sent**	|V|	**Sents per doc**	**Words per sent**
Training	1,802,919	1,172,626,265	2,920,894	23.94	27.17	454,179	1.24	21.99
Validation	104,052	67,669,381	550,213	23.01	28.27	109,460	1.21	23.36
TESTi	104,052	67,363,994	550,910	22.93	28.23	109,706	1.21	23.34
TESTni	109,626	59,603,306	447,679	16.25	33.46	116,201	1.35	36.84

3.4 Summarization Models

In this work, a monolingual news summarization model is proposed: News Abstractive Summarization for Catalan (NASca). It is a Transformer encoder-decoder model with the same architecture and hyper-parameters as BART (Lewis et al., 2020). Inspired by the work of (Zou et al., 2020), we decided to combine several pretraining tasks to inject linguistic knowledge during the pretraining stage with the aim of increasing the abstractivity of the summaries generated by the model. Specifically, four tasks were combined: sentence permutation, text infilling (Lewis et al., 2020), Gap Sentence Generation (GSG) (J. Zhang et al., 2020), and Next Segment Generation (NSG) (Zou et al., 2020). NASca is pretrained simultaneously with the four tasks, which are randomly selected at each batch following a uniform distribution.

We hypothesize that the combination of these four pretraining tasks leads to improvements in the summarization task, especially concerning the abstractivity of the generated summaries. Firstly, with sentence permutation and text infilling, the model should acquire capabilities of content reordering and phrase replacements. Secondly, GSG should tailor the model to whole-document understanding, summary-like generation and paraphrasing. Finally, with NSG, the model could increase the

cohesion of the whole summary, as the task consists of generating continuations of documents given a prefix.

NASca was pretrained with the documents of the Catalan training set of the DACSA corpus (including some documents discarded in the corpora creation process ("DACSA: a Dataset for Automatic summarization of Catalan and Spanish newspaper Articles" n.d.)), the Catalan subset of the OSCAR corpus (Ortiz Suárez, Romary, and Sagot, 2020), and the dump from 20 April 2021 of the Catalan version of the Wikipedia. In total, 9.3 GB of raw text (2.5 millions of documents) were used to pretrain it.

Additionally, we replicated NASca for the Spanish language. We refer to this model as News Abstractive Summarization for Spanish (NASes). NASes is identical to NASca in terms of architecture and pretraining tasks, but they differ in the pretraining dataset. To pretrain NASes, we only used the Spanish documents of the DACSA corpus and the dump from 20 April 2021 of the Spanish version of the Wikipedia. We did not consider for NASes the Spanish subset of OSCAR corpus so as to not increase excessively the difference in the amount of data available for the Spanish model regarding the Catalan one. In total, 21 GB (8.5 million documents) were used to pretrain NASes. Note that even though we did not use the OSCAR corpus, the size of the pretraining dataset for Spanish is twice the size of the Catalan pretraining dataset.

In addition to the monolingual models, two multilingual models were used for the experimental comparison in the summarization task. We worked with two of the most widely used multilingual models in text summarization, mBART and mT5. Regarding the mBART model, we used the *mbart-large-cc25* version, released by Facebook and available online through HuggingFace (`https://huggingface.co/facebook/mbart-large-cc25`, accessed on 21 October 2021) (Wolf et al., 2020). For the mT5 model, we used the *mt5-base* version, published by Google, that is also available online (`https://huggingface.co/google/mt5-base`, accessed on 21 October 2021)).

All the monolingual and multilingual models were fine-tuned and evaluated for the summarization task using the DACSA corpus. The monolingual models proposed in this work were publicly released (`https://huggingface.co/ELiRF/NASCA`, accessed on 21 October 2021), (`https://huggingface.co/ELiRF/NASES`, accessed on 21 October 2021).

3.5 Metrics

To evaluate the performance of the summarization models we used the usual evaluation metrics, the most used ROUGE measure (Lin, 2004) which is based on n-grams,

and a more semantic such as BERTScore (T. Zhang et al., 2020), which is based on contextual embeddings provided by a BERT language model. However, these metrics do not allow to correctly evaluate the abstractivity of the generated summaries.

Measuring the abstractivity of the summaries generated by the models is, except counting the introduced new words, not trivial. In some studies, abstractivity was measured as the absence of n-gram overlap (Bommasani and Cardie, 2020; Grusky, Naaman, and Artzi, 2018), however, creating abstractive summaries is not just about solely of using different vocabulary (Jing, 2002). In this work, we used a set of metrics as abstractivity indicators to asses the level of abstractivity. In particular, the following metrics were selected: *extractive fragment coverage* (Grusky, Naaman, and Artzi, 2018), *abstractivity*$_p$ (Bommasani and Cardie, 2020), *novel 1-grams*, *novel 4-grams* (Kryściński et al., 2018). Also in this work, we present a new metric, called *content reordering*, to evaluate one of the most common characteristics of abstractive summaries, the rearrangement of the original content.

The *content reordering* metric was defined to quantify the percentage of reordering that the information in the summary suffered with respect to its original order in the document. This metric correlates positively with the abstractivity, and thus, by reordering the information, the summary increases its abstractivity.

The measure is based on the inversion concept. The inversion operation extracts all pairs of items that are out of order: $INV(\pi) = \{(a_i, a_j) | i < j \wedge a_i > a_j\}$, where π is a list of comparable elements (Barth, Mutzel, and Jünger, 2004). For instance, with the list $[1,5,4,2]$, the inverse operation results in $[(5,4),(5,2),(4,2)]$.

Given a list of pairs (u, v), where u is the position of a maximum length segment in the original document, and v is the position in which such segment is placed in the summary, this list is sorted by u and the number of inversions that must be made to order the list of pairs by v is calculated. Thus, this allows us to quantify the disorder established in the list of the second component of the pairs when we take into account the order of the first component.

Let $\mathscr{F}(T,S)$(Grusky, Naaman, and Artzi, 2018) be the operation that returns the longest common extractive segments between a text T and its summary S, let $|S|$ be the number of words of the summary, and let $Reordered(T,S)$ be the operation that counts the number of extractive reordered segments; *content reordering* is defined as follows:

$$ContentReordering(T,S) = \begin{cases} \frac{\sum_{f \in \mathscr{F}(T,S)} |f|}{|S|} \cdot \frac{Reordered(T,S)}{|\mathscr{F}(T,S)|-1}, & |\mathscr{F}(T,S)|>1. \\ 0, & otherwise. \end{cases}$$

The output value range of the function is $[0, 1]$, where 1 is the highest degree of information rearrangement.

To illustrate this metric, we provide a full example with the following text (T):

> $_{1}$*Content reordering* *is a metric that* $_{7}$*quantifies* *how the extracted information from the original document is rearranged in the summary.* $_{21}$*Reorder the content* $_{24}$*is a common action* *used* $_{28}$*in abstractive summarization*.

and the following summary (S):

> $_{1}$*In abstractive summarization*, $_{4}$*reorder the content* $_{7}$*is a common action*, $_{11}$*content reordering* $_{13}$*quantifies* *it.*

The highlighted text are fragments in common between the original text and its summary. The subindex before the fragment indicates the starting position in words of the fragment. Thus, the list of the pairs (u, v) of the extractive fragments is the following one when it is ordered by u:

$$[(1, 11), (7, 13), (21, 4), (24, 7), (28, 1)]$$

The resulting list of the INV operation applied on the list made up with the second components of the pairs of the previous list is:

$$INV([11, 13, 4, 7, 1]) = [(\underline{11}, 4), (11, 7), (11, 1), (\underline{13}, 4), (13, 7), (13, 1), (\underline{4}, 1), (\underline{7}, 1)]$$

The $Reorder(T, S)$ operation is 4 since there are 4 extractive reordered segments. This value is computed as the unique values in the first components of the pairs in the previous list (11, 13, 4, 7). Additionally, the length (in words) of the summary is 14, there are 5 extractive fragments, and the sum of their length is 13. With all this information, the *content reordering* metric is calculated as follows:

$$ContentReordering(T, S) = \frac{13}{14} \cdot \frac{4}{5-1} = 0.93$$

With this result, we conclude that there is a certain degree of abstractivity in the summary introduced by a high degree of rearrangement of the information. This fact can be verified in the summary of the example. This abstractivity was introduced by the rearrangement of the extractive segments, and not due to the absence of text overlapping between the summary and the original text.

3.6 Results

In this section, we present the conducted experimentation with the summarization models. Firstly, we present the results of the performance obtained by the three models for Catalan in the summarization task: the NASca model, the mBART model, and the mT5 model. Secondly, we show the results regarding the abstractivity of these models for Catalan. Additionally, we show the results for the three models for Spanish, the NASes model and the two multilingual ones. All the models were evaluated on the two test partitions, TESTi and TESTni.

3.6.1 Summarization Performance of the Models for Catalan

The performance of the models was evaluated using the ROUGE metrics (Lin, 2004) and BERTScore metric (T. Zhang et al., 2020). For each metric, we calculated the average F1 score and its 95% confidence interval (exponent = lower bound, subscript = upper bound) by using bootstrapping. Results are shown in Table 3.1.

The average F1 scores are shown in a normal font size and their confidence intervals in a smaller font size, placed at the right-side of the score. The best average score for each metric within a test partition is remarked in bold style. The confidence intervals are shown in blue color if their range intersects with the confidence interval of the best score value of the metric within the same test partition; in other case, the confidence intervals are presented in black color.

Table 3.1: Average F1 scores and confidence intervals of models in summarization task in Catalan.

Partition	Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Ls	BERTScore
TESTi	NASca	**28.84**$^{28.68}_{29.01}$	11.68$^{11.51}_{11.85}$	22.78$^{22.61}_{22.94}$	23.30$^{23.13}_{23.46}$	71.85 $^{71.78}_{71.92}$
	mBART	28.59 $^{28.42}_{28.77}$	**11.89**$^{11.73}_{12.06}$	**23.00**$^{22.82}_{23.16}$	**23.39**$^{23.22}_{23.56}$	**72.03**$^{71.96}_{72.10}$
	mT5	27.01 $^{26.84}_{27.18}$	10.70 $^{10.54}_{10.87}$	21.81 $^{21.65}_{21.97}$	22.12 $^{21.98}_{22.29}$	71.55 $^{71.49}_{71.61}$
TESTni	NASca	**28.19**$^{27.97}_{28.42}$	11.20 $^{10.99}_{11.43}$	**21.45**$^{21.20}_{21.65}$	**22.44**$^{22.21}_{22.67}$	70.14 $^{70.05}_{70.22}$
	mBART	27.46 $^{27.24}_{27.69}$	11.04 $^{10.81}_{11.29}$	21.13 $^{20.93}_{21.37}$	22.01 $^{21.78}_{22.24}$	70.33 $^{70.25}_{70.43}$
	mT5	27.00 $^{26.77}_{27.23}$	**11.28**$^{11.04}_{11.52}$	21.27 $^{21.03}_{21.51}$	22.01 $^{21.78}_{22.23}$	**70.56**$^{70.47}_{70.65}$

The Table 3.1 shows, regarding the TESTi partition, that the NASca model performs similarly compared to the multilingual mBART model. mBART presents significantly better BERTScore result than NASca while there are overlappings in the confidence intervals in the ROUGE measures. The mT5 model has obtained a significant lower performance than the other two models, despite the fact that mT5

contains the Catalan language in its pretraining phase unlike the mBART model. We hypothesize that the pretraining dataset could influence the results. It could be that the data considered for Catalan to pretrain mT5 differs so much from our domain. Also, the proportion of languages similar to Catalan in the pretraining corpus could be related to this effect.

In the case of the TESTNI partition, there is a significant overall reduction of the performance in most of the metrics of the three models in comparison to the TESTI partition. Generally speaking, the NASCA model has significantly better performance in almost all ROUGE metrics compared to the multilingual models, although there is an overlapping between the confidence interval of NASCA and that of mT5 in ROUGE-2. According to BERTScore, the mT5 model obtains significant differences in comparison to the scores of the NASCA and mBART models.

Taking into account the higher scores and the generalization capabilities, the results of the monolingual model are significant better than the multilingual ones. In one side, mBART has similar performance than NASCA model in the TESTI partition, however, the performance reduction in the second test partition indicates that the model generalizes worse than the other two models. On the other side, the mT5 model generalizes better than mBART, since the drop of the perfomance between the TESTI and the TESTNI is lower in mT5 than mBART, however, mT5 presents significantly lower performance than that of the NASCA model.

3.6.2 Abstractivity of the Summaries Generated by the Models for Catalan

To evaluate the abstractivity, 4 metrics were used: *extractive fragment coverage* (Grusky, Naaman, and Artzi, 2018) (henceforth, we refer to it simply as *coverage*), *abstractivity-p* (Bommasani and Cardie, 2020), *novel n-grams* (Kryściński et al., 2018) and *content reordering*. From now on, we refer those metrics as indicators, since each indicator complements, in some way, the other indicators to obtain a global perception of the level of abstractivity. The Table 3.2 shows the average scores and their confidence intervals. The scores are calculated by comparing the generated summaries against to their respective article text. The scores remarked in bold styles indicates the highest abstractivity. In this experimentation, the lowest value is emphasized in the *extractive fragment coverage* indicator since it correlates negatively with the abstractivity and the highest value is remarked in the remaining abstractivity indicators, since they correlate positively.

Table 3.2: Abstractivity indicators and confidence intervals for Catalan. Values are shown as percentages.

Partition	Model	Extractive Fragment Coverage	Content Reordering	Abstractivity-p (p = 2)	Novel 1-grams	Novel 4-grams
TESTI	NASca	**96.99** 96.94 97.04	46.17 45.79 46.55	**47.19** 46.90 47.46	**03.21** 03.15 03.26	**28.65** 28.41 28.92
	mBART	97.73 97.68 97.77	**47.85** 47.44 48.23	37.70 37.42 37.97	02.40 02.36 02.45	23.80 23.55 24.02
	mT5	98.59 98.55 98.62	41.25 40.84 41.67	38.04 37.78 38.28	01.51 01.48 01.55	21.89 21.71 22.08
TESTNI	NASca	**96.66** 96.55 96.77	42.37 41.84 42.88	**41.89** 41.44 42.37	**03.52** 03.40 03.63	**26.32** 25.91 26.68
	mBART	97.08 96.99 97.16	**42.96** 42.40 43.56	36.98 36.55 37.41	03.01 02.92 03.09	24.32 23.95 24.70
	mT5	98.31 98.26 98.36	38.82 38.24 39.41	39.18 38.83 39.54	01.80 01.74 01.85	23.20 22.92 23.48

As it is shown in Table 3.2, all the models show a predominant extractivity behavior in the same way as the most abstractive models in the literature. All the scores of the abstractivity indicators denote low abstractivity. For instance, the *coverage* and *novel 1-grams* indicators show that the models reuses a lot of words from the original documents. Although all the models present high-extractivity in their generated summaries, there are significant differences among the models that can be analyzed.

Regarding the TESTI partition, the scores of most of the abstractivity indicators of the NASca model reflect significantly better abstractivity than that of the multilingual models. Also, we can observe that the multilingual models have relatively similar scores in most of the indicators, although, the indicators of the mBART model show slightly more abstractivity than the mT5 model.

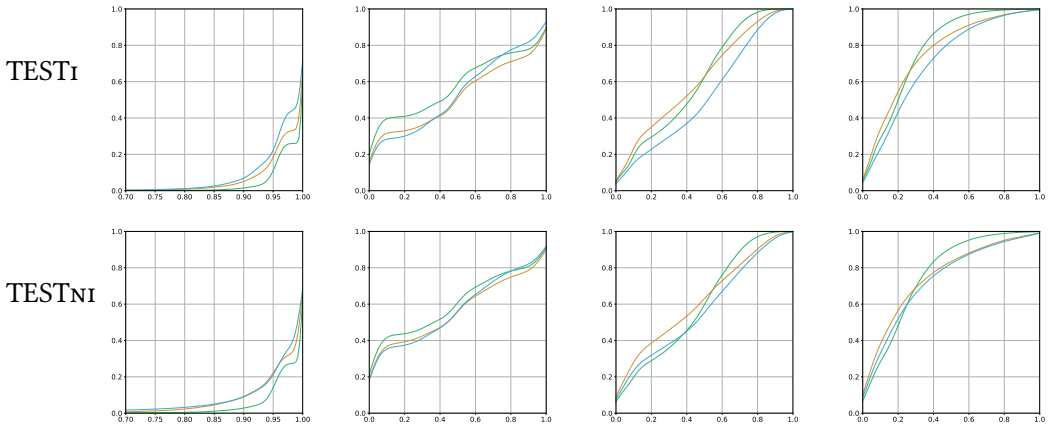
In the case of the TESTNI partition, the NASca model indicators reflect better abstractivity than in the multilingual models. However, compared to the values in TESTI, NASca reduced most of their abstractivity indicators scores except the *coverage* indicator, which is slightly better. In this partition, the differences in the values between the NASca model and the multilingual models are lower than in the TESTI partition.

Overall, it is noticeable that the NASca model reuses a lot of content from the original text. The model uses a lot of words from the original text which is reflected in the low value of the *novel 1-grams* indicator. However, despite the fact that the model reuses a lot of words, the extractive fragments tend to be shorter than in the multilingual models, since the *novel 4-grams* indicator shows a significantly higher value than in the multilingual models; this fact is also exposed by the *abstractivity-p* indicator, which presents a difference between the 5% and the 10% depending on the partition and the multilingual model. For all these observations in the indicators,

we conclude that the NASca model generates summaries with higher degree of abstractivity than the multilingual models.

With the aim of better analyzing the behavior of the models, we computed the cumulative distributions of the abstractivity indicators for each model and test partition. The results are presented in the Figure 3.1.

The plots show in the x-axis the indicator measured, and in the y-axis, the percentage of generated summaries that present less or equal score to the value in the x-axis. These plots are helpful to evaluate the abstractivity of the generated summaries by taking into account how they are distributed based on certain score. If a metric correlates negatively with the abstractivity, it is desired that the scores be lower; that is, the model accumulates the samples fast. In contrast, if the metric correlates positively, it is desired that the scores be higher. In this case, we say that the model accumulates the samples slowly.



(**a**) *Ext. Frag. Coverage* (**b**) *Content Reordering* (**c**) Abstractivity-p ($p = 2$) (**d**) *Novel 4-grams*

Figure 3.1: Cumulative distribution of 4 abstractivity indicators for models NASca, mBART, mT5 for Catalan.

In Figure 3.1, regarding the *coverage* indicator, which correlates negatively with abstractivity, we observe that the NASca model stays always on top of the multilingual models, so this indicates that the samples are accumulated faster, which is a positive indication for the abstractivity. In the remaining indicators, which correlate positively with the abstractivity, the NASca model tends to accumulate the samples slower than the multilingual models, which is also positive concerning the abstractivity, except the *content reordering* indicator. Regarding this indicator, although NASca present a lower value than the mBART model in the Section 3.6.2, the NASca model's distribution stays below the mBART until 40%, and later reaches and surpasses the multilingual

models. This means that the *NASca* model, overall, introduces less *content reordering* on their summaries; however, the amount of summaries with rearrangement of the information is higher than in the ones generated by the multilingual models.

The results presented in the Table 3.2 and the Figure 3.1 show enough evidences to conclude that the NASca model presents better abstractivity than the rest of the trained models. Additionally, to verify if the improvement in the abstractivity indicators is due to the pretraining tasks, we pretrained a BART model specifically for Catalan using only the pretraining tasks proposed in the original work (Lewis et al., 2020). The results show that both models, NASca and BART, have a similar performance in the summarization task, however, the NASca model presents significant higher abstractivity indicators. For instance, in the *coverage* indicator of the TESTni partition, the NASca model scores 96.99 (96.94, 97.04) and BART 97.29 (97.24, 98.41). In the case of novel 4-grams, and also for TESTni, the NASca model scores 26.65 (25.91, 26.68) and BART 25.48 (25.12, 25.82).

An example of an article and the summaries generated by the three models is shown in Appendix 3.A.1.

3.6.3 Summarization Performance and Abstractivity of the Summaries Generated by the Models for Spanish

It is also interesting to study the performance of our proposal in languages with higher resources than Catalan. For this reason, we replicated the model and the experimentation for the Spanish language. The summarization performance results and the results related to the abstractivity indicators are shown in Tables 3.3 and 3.4, respectively. In addition, the cumulative distributions of the abstractivity indicators are presented in Figure 3.2.

Table 3.3: Average F1 scores and confidence intervals of models in summarization task in Spanish.

Partition	Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Ls	BERTScore
TESTi	NASes	**33.24** 33.12 33.38	**15.79** 15.63 15.93	**26.76** 26.63 26.89	**27.56** 27.43 27.69	**73.11** 73.05 73.16
	mBART	31.09 30.98 31.20	13.56 13.44 13.68	24.67 24.56 24.78	25.48 25.37 25.58	72.25 72.21 72.30
	mT5	31.72 31.60 31.85	14.54 14.39 14.67	25.76 25.63 25.89	26.31 26.18 26.44	72.86 72.82 72.91
TESTni	NASes	30.60 30.52 30.68	10.75 10.66 10.83	22.29 22.21 22.37	23.06 22.99 23.15	70.66 70.62 70.69
	mBART	**30.66** 30.58 30.74	12.08 11.98 12.18	23.13 23.06 23.22	23.89 23.81 23.98	71.07 71.04 71.10
	mT5	30.61 30.51 30.70	**12.36** 12.25 12.47	**23.53** 23.43 23.62	**24.05** 23.95 24.14	**71.26** 71.22 71.30

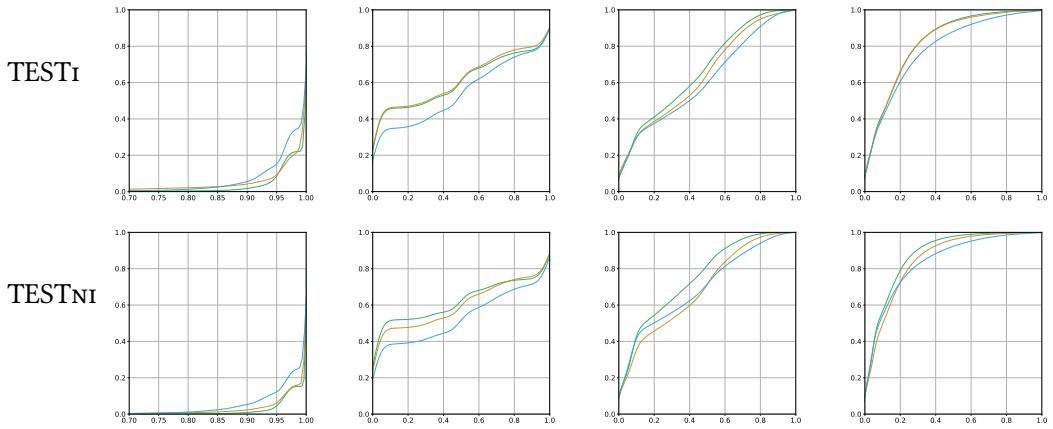
Table 3.3 shows that the NASes model presents the best performance of the three models in the TESTi partition. All the scores obtained by the NASes model are significantly better compared to those of the multilingual models. Specifically, the NASes model achieve, on average, 8.2% higher performance than mBART and 4.5% higher than mT5. Regarding the TESTni partition, the NASes model reduces its performance in average, while mT5 achieves the best results in almost all the metrics.

The results show that the NASes excelled in the TESTi partition, which contains newspapers included in the training partition. However, NASes presents lower generalization capabilities than the multilingual models due to the noticeable performance reduction in the TESTni partition, which contains newspapers not included in the training partition.

Table 3.4: Abstractivity indicators and confidence intervals for Spanish. Values are shown as percentages.

Partition	Model	Extractive Fragment Coverage	Content Reordering	Abstractivity-p ($p = 2$)	Novel 1-grams	Novel 4-grams
TESTi	NASes	**97.65** 97.62 97.68	**45.27** 45.04 45.50	**38.15** 37.97 38.31	**02.55** 02.52 02.58	**21.17** 21.04 21.31
	mBART	98.14 98.10 98.18	37.70 37.45 37.92	35.17 35.00 35.32	01.85 01.81 01.89	17.58 17.47 17.70
	mT5	98.74 98.72 98.76	38.67 38.42 38.92	32.41 32.25 32.58	01.36 01.34 01.38	17.39 17.29 17.49
TESTni	NASes	**98.16** 98.13 98.19	**46.58** 46.33 46.82	29.76 29.60 29.92	**02.00** 01.97 02.03	**15.76** 15.65 15.88
	mBART	98.92 98.90 98.94	39.38 39.13 39.61	**30.48** 30.33 30.64	01.03 01.01 01.05	14.68 14.59 14.78
	mT5	99.24 99.23 99.26	37.17 36.91 37.43	24.19 24.06 24.32	00.83 00.81 00.84	12.08 12.00 12.16

Regarding the abstractivity indicators on the TESTi partition, presented in Table 3.4, all the scores of the NASes model are significantly better than those of the multilingual models. In the TESTni partition, the models present less abstractivity in comparison to the TESTi partition. Also in TESTni, the NASes model shows significant differences compared to the multilingual models in all the indicators, excluding *abstractivity-p* where mBART obtains better scores than NASes and the mT5 models. We also computed the cumulative distributions of the abstractivity indicators for each model and test partition. The results are presented in the Figure 3.2.



(a) *Ext. Frag. Coverage* (b) *Content Reordering* (c) Abstractivity-p ($p = 2$) (d) *Novel 4-grams*

Figure 3.2: Cumulative distribution of 4 abstractivity indicators for models NASes, mBART, mT5 for Spanish.

The plots presented in Figure 3.2 help us to reinforce the observations extracted from the numerical results showed in Table 3.4. The NASes model tends to accumulate slightly higher percentage of samples in the *coverage* indicator after the 90% of *coverage* is achieved. Regarding the remaining indicators, the accumulation tends to occur slower than in the other two models.

The abstractivity indicators analysis shows that the summaries generated by NASes have a significant higher abstractivity than those generated by the multilingual models, something that complements the observations made in the Sections 3.6.1 and 3.6.2 about the models for Catalan.

3.7 Conclusions

In this work, a monolingual model for abstractive summarization in Catalan, NASca, was presented. The model was pretrained from scratch based on the BART architecture and using four self-supervised tasks with the aim of increasing the abstractivity of the generated summaries. The fine-tuning phase was carried out using the DACSA dataset, a corpus of articles obtained from online newspapers. The experimentation conducted supports the correctness of our proposal considering the three evaluated aspects: the performance of the model, the abstractivity of the generated summaries, and the generalization capabilities of the model.

Following the same architecture and the same training strategy, a model for abstractive summarization in Spanish, NASes, was also trained and evaluated, and it

also provided very good results. To our knowledge, this is the first work that explores a monolingual approach for abstractive summarization both in Catalan and Spanish.

Additionally, in this work, we also proposed a new metric, *content reordering*, with the aim of helping to quantify the rearrangement of the original content within an abstractive summary. This characteristic is common in abstractive summaries, but it is not considered by the metrics in the literature.

Chapter 3. References

Ahuir, Vicent et al. (2021). "NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish". In: *Applied Sciences* 11.21. ISSN: 2076-3417. DOI: 10.3390/app11219872. URL: https://www.mdpi.com/2076-3417/11/21/9872 (cit. on p. 45).

Barth, Wilhelm, Petra Mutzel, and Michael Jünger (2004). "Simple and Efficient Bilayer Cross Counting". In: *Journal of Graph Algorithms and Applications* 8.2, pp. 179–194. DOI: 10.7155/jgaa.00088 (cit. on p. 56).

Bommasani, Rishi and Claire Cardie (Nov. 2020). "Intrinsic Evaluation of Summarization Datasets". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8075–8096. DOI: 10.18653/v1/2020.emnlp-main.649. URL: https://aclanthology.org/2020.emnlp-main.649 (cit. on pp. 56, 59).

Cañete, José et al. (2020). "Spanish Pre-Trained BERT Model and Evaluation Data". In: *PML4DC at ICLR 2020* (cit. on pp. 48, 51).

"DACSA: a Dataset for Automatic summarization of Catalan and Spanish newspaper Articles" (n.d.). Unsubmitted (cit. on pp. 49, 52, 55).

González, José Ángel, Lluís-F. Hurtado, and Ferran Pla (2021). "TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter". In: *Neurocomputing* 426, pp. 58–69. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.09.078. URL: https://www.sciencedirect.com/science/article/pii/S0925231220316180 (cit. on p. 51).

Grusky, Max, Mor Naaman, and Yoav Artzi (June 2018). "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 708–719. DOI: 10.18653/v1/N18-1065. URL: https://aclanthology.org/N18-1065 (cit. on pp. 56, 59).

National Information Standards Organization (1997). *Guidelines for Abstracts*. Standard. Maryland, U.S.A: American National Standards Institute (cit. on p. 48).

Jing, Hongyan (Dec. 2002). "Using Hidden Markov Modeling to Decompose Human-Written Summaries". In: *Computational Linguistics* 28.4, pp. 527–543. ISSN: 0891-2017. DOI: 10.1162/089120102762671972. URL: https://doi.org/10.1162/089120102762671972 (cit. on pp. 48, 49, 56).

Kryściński, Wojciech et al. (Oct. 2018). "Improving Abstraction in Text Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1808–1817. DOI: 10.18653/v1/D18-1207. URL: https://aclanthology.org/D18-1207 (cit. on pp. 51, 56, 59).

Le, Hang et al. (May 2020). "FlauBERT: Unsupervised Language Model Pre-training for French". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 2479–2490. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.302 (cit. on p. 51).

Lewis, Mike et al. (July 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*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703 (cit. on pp. 48–50, 54, 62).

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013 (cit. on pp. 55, 58).

Liu, Yang and Mirella Lapata (Nov. 2019). "Text Summarization with Pretrained Encoders". 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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740. DOI: 10.18653/v1/D19-1387. URL: https://aclanthology.org/D19-1387 (cit. on p. 50).

Liu, Yinhan et al. (Nov. 2020). "Multilingual Denoising Pre-training for Neural Machine Translation". In: *Transactions of the Association for Computational Linguistics* 8, pp. 726–742. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00343. URL: https://www.aclweb.org/anthology/2020.tacl-1.47 (cit. on pp. 48, 49, 51).

Magooda, Ahmed and Diane J. Litman (2020). "Abstractive Summarization for Low Resource Data using Domain Transfer and Data Synthesis". In: *CoRR* abs/2002.03407. arXiv: 2002.03407. URL: https://arxiv.org/abs/2002.03407 (cit. on p. 51).

Martin, Louis et al. (2020). "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (cit. on pp. 48, 51).

Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou (2017). "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, pp. 3075–3081. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10958/10817 (cit. on p. 50).

Nallapati, Ramesh, Bowen Zhou, et al. (Aug. 2016). "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, pp. 280–290. DOI: 10.18653/v1/K16-1028. URL: https://aclanthology.org/K16-1028 (cit. on p. 50).

Ortiz Suárez, Pedro Javier, Laurent Romary, and Benoıt Sagot (July 2020). "A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1703–1714. URL: https://www.aclweb.org/anthology/2020.acl-main.156 (cit. on p. 55).

Pires, Telmo, Eva Schlinger, and Dan Garrette (July 2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493 (cit. on pp. 48, 51).

Qi, Weizhen et al. (2020). "Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 2401–2410 (cit. on p. 50).

Raffel, Colin et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html (cit. on pp. 48, 50).

Rane, Neha and Sharvari Govilkar (2019). "Recent Trends in Deep Learning Based Abstractive Text Summarization". In: *International Journal of Recent Technology and Engineering* 8.3, pp. 3108–3115. DOI: 10.35940/ijrte.c4996.098319 (cit. on p. 48).

Rush, Alexander M., Sumit Chopra, and Jason Weston (Sept. 2015). "A Neural Attention Model for Abstractive Sentence Summarization". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 379–389. DOI: 10.18653/v1/D15-1044. URL: https://aclanthology.org/D15-1044 (cit. on p. 50).

See, Abigail, Peter J. Liu, and Christopher D. Manning (July 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)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1073–

1083. DOI: 10.18653/v1/P17-1099. URL: https://aclanthology.org/P17-1099 (cit. on p. 50).

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., pp. 5998–6008. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 50).

Verma, Pradeepika, Sukomal Pal, and Hari Om (2019). "A Comparative Analysis on Hindi and English Extractive Text Summarization". In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 18.3, pp. 1–39. DOI: 10.1145/3308754 (cit. on p. 48).

Virtanen, Antti et al. (2019). "Multilingual is not enough: BERT for Finnish". In: *CoRR* abs/1912.07076. arXiv: 1912.07076. URL: http://arxiv.org/abs/1912.07076 (cit. on pp. 48, 51).

Vries, Wietse de et al. (2019). "BERTje: A Dutch BERT Model". In: *CoRR* abs/1912.09582. arXiv: 1912.09582. URL: http://arxiv.org/abs/1912.09582 (cit. on p. 51).

Widyassari, Adhika Pramita et al. (2020). "Review of automatic text summarization techniques & methods". In: *Journal of King Saud University - Computer and Information Sciences*. ISSN: 1319-1578. DOI: 10.1016/j.jksuci.2020.05.006. URL: https://www.sciencedirect.com/science/article/pii/S1319157820303712 (cit. on p. 48).

Wolf, Thomas et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6 (cit. on pp. 50, 55).

Xue, Linting et al. (June 2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41 (cit. on pp. 48, 49, 51).

Zhang, Jingqing et al. (July 2020). "PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. ICML20. PMLR. JMLR.org, pp. 11328–11339. URL: https://proceedings.mlr.press/v119/zhang20ae.html (cit. on pp. 48, 50, 54).

Zhang, Tianyi et al. (2020). "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations*. OpenReview.net. URL: https://openreview.net/forum?id=SkeHuCVFDr (cit. on pp. 56, 58).

Zhong, Ming et al. (July 2020). “Extractive Summarization as Text Matching”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6197–6208. DOI: 10.18653/v1/2020.acl-main.552. URL: https://aclanthology.org/2020.acl-main.552 (cit. on p. 50).

Zou, Yanyan et al. (Nov. 2020). “Pre-training for Abstractive Document Summarization by Reinstating Source Text”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 3646–3660. DOI: 10.18653/v1/2020.emnlp-main.297. URL: https://aclanthology.org/2020.emnlp-main.297 (cit. on pp. 51, 54).

Chapter 3. Appendix

3.A Summarization example

An example of an article, its reference summary, and the summaries generated by the three models are shown in Figure 3.A.1. It also shows the different metrics achieved by each summary. All the generated summaries are syntactically and semantically correct. Based on the low values of the ROUGE scores, we can affirm that all the generated summaries are very different from the reference one. Regarding the coverage indicator, although the three summaries are quite extractive, since they use several segments from the article, mT5 is by far the most extractive. Considering all the abstractive indicators, NASca and mBART are better than mT5, and NASca outperforms mBART especially in terms of novel n-grams and abstractivity-p.

Article: La clau va ser el ritme. El ritme amb què Marc Márquez va arrencar al Gran Premi de l'Argentina i amb què el va acabar. El pilot de Cervera, que sempre assegura que li agraden les curses en grup, va fer avançaments, va buscar els forats i va passar-s'ho bé dalt de la moto: a l'Argentina va decidir ser, per un dia, infidel al seu estil. Sabia que tenia ritme, ho havia demostrat durant totes les sessions d'entrenaments lliures i també als oficials (havia dominat cinc de les sis sessions), i a la cursa no va tenir rival. Va sortir, va posar el "mode creuer", com va dir, i va perdre de vista la resta de rivals. En una volta, un segon d'avantatge, i ja s'escapava de 12 segons dels perseguidors quan va decidir passar a controlar la cursa, sense prendre més riscos dels necessaris. "No és el meu estil, però després del que va passar l'any passat tenia ganes de fer una cursa així. Va passar el que va passar i volia demostrar el meu ritme", va assegurar després de baixar de la moto. Márquez va marcar la pole i la volta ràpida, i va ser líder des que es van apagar els semàfors fins al final. Va aconseguir el que es coneix com un Grand Chelem: el de Cervera, de fet, tan sols n'ha aconseguit cinc des que va debutar a MotoGP; tres a Austin (2014, 2016 i 2018), un a Jerez (2014) i el de diumenge a l'Argentina. "Pocs dies a l'any et trobes amb aquestes sensacions dalt de la moto. Calia aprofitar-ho, ha sigut perfecte", reconeixia. La manera més dolça de marcar el ritme. La victòria es va començar a coure molt abans de la sortida, al box, amb el seu equip, llegint els temps de les sessions d'entrenaments. "Els papers deien que era qui tenia més ritme. He intentat marcar les diferències en les set primeres voltes i, després, mantenir l'avantatge", explicava el català. Com si fos un rellotge, clavava volta a volta un 1:39. Al final, els 12 segons d'avantatge es van reduir a 9.816, que, si bé no és la distància més gran amb què Márquez ha guanyat una cursa (a Brno el 2017 va acabar primer amb 12.438 respecte a Pedrosa), sí que és la més gran que ha aconseguit el de Cervera en una cursa en sec: tant a Brno fa dos anys com a Sachsenring en fa tres, en què va acabar a 9.857 de Crutchlow, la pluja va marcar les curses. Lluny també queden els més de 37 segons d'avantatge amb què Dani Pedrosa va guanyar a Xest el 2012 sobre Nakasuga, també sota la pluja, després de la caiguda de Lorenzo. "Com que hem guanyat per deu segons, sembla que som en un altre món, però no, la distància és només de quatre punts respecte a Dovizioso", afegia Márquez. Just abans del podi es va veure segurament una de les imatges de l'any: Valentino Rossi, que va acabar segon, va encaixar la mà amb Márquez, un gest que no es veia des de feia un any, quan el de Cervera, precisament a Termas de Río Hondo, va tocar l'italià i el va fer caure, cosa que va comportar l'inici d'un terratrèmol. Diumenge, ja al podi, els dos campions van fer xocar les ampolles de xampany, però sense dirigir-se la paraula.

Reference: El triomf de Márquez a l'Argentina, el més ampli en sec del de Cervera a MotoGP.

NASca: El de Cervera va marcar la 'pole' a l'Argentina i va ser líder del Mundial en una volta.
(ROUGE-1: 5.97; ROUGE-2: 4.42; ROUGE-L: 4.72; BERTScore: 67.08)
(Coverage: 85.00; Reordering: 85.00; Abstractivity-p: 87.75; Novel 1-grams: 15.79; Novel 4-grams: 94.12)

mBART: El de Cervera marca la 'pole' a l'Argentina i és líder des que es van apagar els semàfors.
(ROUGE-1: 6.28; ROUGE-2: 4.72; ROUGE-L: 5.97; BERTScore: 69.17)
(Coverage: 85.00; Reordering: 85.00; Abstractivity-p: 79.75; Novel 1-grams: 15.00; Novel 4-grams: 70.59)

mT5: El pilot de Cervera, que sempre assegura que li agraden les curses en grup, va fer avançaments, va buscar els forats i va passar-se bé dalt de la moto.
(ROUGE-1: 9.58; ROUGE-2: 8.68; ROUGE-L: 9.27; BERTScore: 72.96)
(Coverage: 96.97; Reordering: 48.48; Abstractivity-p: 35.54; Novel 1-grams: 3.70; Novel 4-grams: 13.33)

Figure 3.A.1: Text of the article, the reference summary, and the summaries generated by the models.

Chapter 4

ELiRF-VRAIN at BioNLP Task 1B: Radiology Report Summarization

Vicent Ahuir Esteve, Encarna Segarra, and Lluis Hurtado (July 2023). "ELiRF-VRAIN at BioNLP Task 1B: Radiology Report Summarization". In: *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Ed. by Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen. Toronto, Canada: Association for Computational Linguistics, pp. 524–529. DOI: 10.18653/v1/2023.bionlp-1.52. URL: https://aclanthology.org/2023.bionlp-1.52

Impact Index

In 2023, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks* was co-located at the *Association for Computational Linguistics* conference; a congress ranked by the *Computing Research & Education* (CORE) as CORE-A+ in 2023.

Abstract

This paper presents our system at the Radiology Report Summarization Shared Task-1B of the 22nd BioNLP Workshop 2023. Inspired by the work of the BioBART model, we continuously pre-trained a general domain BART model with biomedical data to adapt it to this specific domain. In the pre-training phase, several pre-training tasks are aggregated to inject linguistic knowledge and increase the abstractivity of the generated summaries. We present the results of our models, and also, we have carried out an additional study on the lengths of the generated summaries, which has provided us with interesting information.

4.1 Introduction

Radiology reports are documents that interpret radiological examinations. Usually, a radiology report consists of three sections: (1) a background section that describes general information about the patient and exam, (2) a findings section that presents details of the examination, and (3) an impression section that summarizes the findings against the background. This last section is the most crucial for doctors to make clinical decisions.

Due to the recent success of self-supervised learning, the focus of text summarization research has exhibited a gradual shift from extractive techniques to abstractive techniques. The best-performing abstractive models are BART (Lewis et al., 2020), T5 (Raffel et al., 2020), PEGASUS (J. Zhang et al., 2020), and GPT-3 (Brown et al., 2020), being all of them Transformers (Vaswani et al., 2017) pre-trained self-supervisedly as denoising sequence to sequence autoencoders. This kind of approaches allow to pre-train deep architectures to learn vast amounts of general linguistic knowledge from large corpora, that can be transferred to downstream tasks by means of fine-tuning. Almost all of these systems used benchmark datasets compiled from news articles, such as the CNN-DailyMail dataset (CNN-DM) (Hermann et al., 2015) and Newsroom (Grusky, Naaman, and Artzi, 2018). However, not so many efforts have been carried out in the biomedical domain.

Language models pre-trained on biomedical corpora may further enhance the performance of current biomedical NLG methods, such BioBERT (Lee et al., 2020) or PubMedBERT (Gu et al., 2021). However, there are very few in-domain generative language models for biomedicine. In (Yuan et al., 2022), authors proposed a biomedical auto-regressive generative language model, BioBART, pre-trained on the biomedical corpora. They continuously pre-train BART on PubMed[1] abstracts to achieve biomedical domain adaption only using the text-infilling task. The in-domain BioBART outperforms BART model and sets strong baselines for several NLG tasks.

In the framework of BioNLP workshop, some challenges and shared tasks focusing on summarization were created. MEDIQA 2019 edition focused on question entailment and textual inference and their applications in medical Question Answering (Ben Abacha, Shivade, and Demner-Fushman, 2019). MEDIQA 2021 (Ben Abacha, Mrabet, et al., 2021) promoted research on summarization for consumer health QA and clinical text. In this edition, the winner system (Dai et al., 2021), based on PEGASUS, employed a domain adaptation strategy by further fine-tuning a small amount of in-domain data to improve generalization and transfer abilities.

[1] https://pubmed.ncbi.nlm.nih.gov/

4.2 Task Description

The Shared Task-1B of the 22nd BioNLP Workshop 2023 (J.-B. Delbrouck, Varma, et al., 2023), focuses on the summarization of radiology reports. The task of the summarization of radiology reports can be defined as follows: given a radiology report with findings and background sections, the goal is to generate the impression section. For this shared task, the *Impressions* are only generated from the *Findings* section.

Shared Task 1B was divided into two subtasks. The first one is about generating impressions sections based exclusively on the text report. The second one is summarizing the report based on the text information and the indicators that could be extracted from the attached radiology image. The participants were invited to approach both subtasks but were allowed to participate in one; we chose to participate only in the first subtask.

4.2.1 The Dataset

For the subtask where we participated, a dataset is provided based on MIMIC-III (Johnson, Pollard, and Mark, 2016) with 79 779 samples from two different radiography modalities and six anatomical parts. This dataset was split into four partitions: `train` (59 320 samples), `validation` (7413), `test` (6526), and `hidden-test` (6531).

Table 4.1: Average sentences and words on Findings and Impressions for each partition.

	Findings		**Impressions**	
	Sent.	Words	Sent.	Words
`train`	8.80	124.90	3.91	52.26
`validation`	8.85	125.69	3.93	52.65
`test`	9.28	134.95	3.75	50.97
`hidden-test`	10.23	155.28	-	-

Table 4.1 details the average number of sentences and words for Findings and Impressions, excluding the Impressions of the `hidden-test` that were not available to participants. On the one hand, we notice that `train` and `validation` have similar lengths for both, Impressions and Findings. On the other hand, the test partitions contain longer Findings, especially the `hidden-test`. Moreover, Impressions are shorter in the `test` partition than in the `train` and `validation` ones; thus, `test` presents a higher compression ratio in its Impressions partition than those of the other two partitions.

4.2.2 System Evaluation

For evaluating the systems, the ViLMedic framework (J.-b. Delbrouck et al., 2022) was used. It is a framework that aims to increase results reproducibility for medical tasks, such as medical report summarization. Specifically, the systems were evaluated with the following metrics and scores: ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2002), BERTScore (T. Zhang et al., 2020), and RadGraph (J.-B. Delbrouck, Chambon, et al., 2022).

4.3 Pre-training Model

Inspired by the work of the BioBART model, we continuously pre-trained a general domain BART model with biomedical data to adapt it to this specific domain. Specifically, our starting point was the architecture and weights of the base version of BART[2], publicly available at the repository of HuggingFace (Wolf et al., 2020).

For the pre-training phase, we followed the methodology used in the News Abstractive Summarization models (NAS) work (Ahuir et al., 2021). In NAS, several pre-training tasks are aggregated to inject linguistic knowledge during the pre-training stage and to increase the abstractivity of the generated summaries. We chose this methodology because we hypothesize that reference impressions are written in a mostly abstractive way. Also, the pre-training method helped to transfer more knowledge to the summarization task, which increased the model's performance in the original work.

For pre-training, we chose data that were as similar as possible to radiology reports in terms of vocabulary and grammar. We selected the following MIMIC datasets available at PhysioNet (Goldberger et al., 2000): note events in MIMIC-III (2 083 180 samples) (Johnson, Pollard, and Mark, 2016), radiology reports in MIMIC-CXR (128 032) (Johnson, Pollard, Mark, et al., 2019), and discharge (331 794) and radiology reports (2 321 355) in MIMIC-IV (Johnson, Pollard, Horng, et al., 2023). Additionally, we included Wikipedia articles related to medicine to reinforce the domain vocabulary (97 192).

The base version of the BART model is limited to 1024 input tokens; however, most samples exceeded this size. This fact could lead us to lose valuable training data. To overcome this limitation, we split texts into narrower samples as follows:

[2] https://huggingface.co/facebook/bart-base

having the text split by lines and a window of no more than 1000 words, we generated sub-samples that contained at least a new line and filled the window with as many words as possible. With this method, we obtained a dataset of 40 894 042 samples.

Due to infrastructure limitations, we could only train the model for one epoch, which took 12 days with four NVIDIA RTX 3090 graphic cards. The following hyperparameters were used: 4 samples per device, 256 gradient accumulation steps, a learning rate of 5×10^{-5} with a linear scheduler, 1% of the epoch for warm-up, and 32 bits of training precision. For the hyperparameters not mentioned, we have used the default values of the 4.23.1 version of `transformers` library of HuggingFace[3].

4.4 Models for the Task

For the downstream task, we obtained three models based on our pre-trained model. The first model (M1) was fine-tuned with the `train` partition of the shared task. The second one (M2) was fine-tuned with the `train` and `validation` partitions. Finally, the third one (M3) was fine-tuned with all the partitions with available references: `train`, `validation`, and `test`.

For the fine-tuning phase, we did a grid search of certain hyperparameters with RayTune (Liaw et al., 2018) through the HuggingFace library. We did 20 trials over the following hyperparameters: `learning_rate` (from 8×10^{-6} to 4×10^{-5}), `num_train_epochs` (10 or 15), and `gradient_accumulation_steps` (2, 4, or 8). Since we wanted to find which models obtained a more balanced performance among the four metrics of the task (ROUGE-L, BLEU, RadGraph-F1, and BertScore-F1), we defined the objective to maximize as the harmonic mean (Ferger, 1931) of these four scores. Finally, the three models were fine-tuned during 15 epochs with an NVIDIA RTX 4090 with the following hyperparameters: 8 samples per device, 4 gradient accumulation steps, and a learning rate of 2.14×10^{-5}.

For the generation of impressions, we used the *generate* method[4] of HuggingFace. To achieve better performance, we identified certain hyperparameters and performed grid search using the M1 model and the `validation` partition, specifically: `max_length` (60, 70, 80, **90**, or 100), `num_beams` (3, 4, **6**, 8, or 10), and `no_repeat_ngram_size`

[3] https://huggingface.co/docs/transformers/v4.23.1/en/main_classes/trainer#transformers.TrainingArguments

[4] https://huggingface.co/docs/transformers/v4.23.1/en/main_classes/text_generation#transformers.generation_utils.GenerationMixin

(3, 4, **5**, 6, 8, or 10). The bolded values maximized the harmonic mean score; thus, we fixed them to generate impressions with ours models.

4.5 Results

Table 4.1: Results on test partitions of our models and those of the groups that achieved the highest score on any of the four measures. For all measures, a higher value means a better performance. M1, M2 and M3 are the three models created with our approach. G1 and G2 are the models that have, at least, a highest value in any measure, without taking into account our models.

Pt	Md	BL	RL	BS	RG	HM
T	M1	**17.61**	30.19	53.13	31.19	28.41
	M2	17.41	29.57	52.24	31.40	28.10
	G1	15.99	**34.07**	**56.30**	**35.25**	28.89
	G2	17.33	33.93	55.49	34.93	**29.89**
HT	M1	16.98	30.52	54.03	31.79	28.24
	M3	18.06	30.19	53.94	32.58	29.04
	G1	**18.36**	**35.32**	**57.26**	**36.94**	**31.42**

Table 4.1 shows the results of our models (M1, M2, M3) and those of the groups (G1, G2) that reached the highest score, excluding our models, on any of the four scores: BLEU (BL), ROUGE-L (RL), BertScore-F1 (BS), and RadGraph-F1 (RG). The overall performance on the four metrics is reflected by the harmonic mean (HM). The results are divided in two sections: `test` (T) and `hidden-test` (HT). The leaderboard scores were computed by limiting the prediction and the reference to 256 words.

Overall, our models have lower performance than the best systems. In the `test` partition, our best model (M1) averages a 9% lower performance than the other two groups if BLEU is excluded from the count, and 5.3% less when is included, meaning that M1 performs substantially better in BLEU than the other systems. Comparing our models, M2 seems to perform worse than M1, despite being trained with more data. In the case of `hidden-test`, our best model (M3) averages 10.7% lower performance than G1 if BLEU is excluded from the count and 8.43% if not. Comparing the performance of our models, unlike what happened in the `test` partition, M3 performs better than M1. Therefore, the inclusion of the `test` partition in training resulted in more acknowledgment for the model, probably because of the additional findings types.

4.6 Discussion

When we observe Table 4.1, two main questions rapidly come to our mind: *Why did our models obtain lower values in all scores but BLEU?*, and *Why the M2 model performed*

worse than M1, despite being trained with more data?. Surprisingly, both questions point to a main problem in our models: the length of the generated impressions.

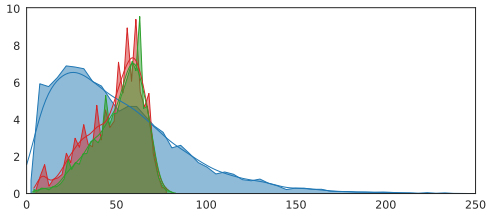


Figure 4.1: Distribution of samples per number of words. Reference impressions (blue) and generated ones M1 (red), M2 (green). `X-axis`: length of the impressions in words, `Y-axis`: percentage of samples with a certain length.

Figure 4.1 presents the distribution of the impressions by their length in words. The blue distribution is the reference impressions, the red one is for those generated by M1, and the green one is for M2. It is noticeable that the impressions of our models follow a completely different distribution than the references. However, the three distributions have similar averages in word count: 51 for the references, 49 for the M1, and 51 for the M2. Our models are trying to set a common length for the impressions instead of identifying which ones should be shorter and which ones should be longer. Therefore, BLEU seems to be weaker in this situation than the other metrics. Moreover, M2 generates longer impressions than M1, lowering the precision and, by extension, its general performance. However, there is a chance that M2 excels in some interesting aspects compared to M1.

Table 4.1: Precision and Recall of M1 and M2 models in the `test` partition when there is no sentence limit (SLN) and when the prediction is limited by the number of sentences of the reference (SLY).

		precision/recall			
	M	**RL**	**BS**	**RG**	**HM**
SLN	1	**30.59**/36.47	**52.31**/54.56	**30.40**/37.43	**38.17**
	2	29.16/**37.55**	49.66/**55.45**	30.17/**38.34**	37.90
SLY	1	**35.75**/33.29	**56.38**/52.77	**34.05**/34.14	39.12
	2	35.73/**33.94**	56.09/**53.45**	33.34/**34.69**	**39.26**

Table 4.1 shows the precision and recall obtained by the models M1 and M2 on `test` for ROUGE-L, BERTScore, and RadGraph; also, the harmonic means of these six measures are shown. The SLN group shows the real performance of the models. Contrary, SLY shows the performance when, at most, the first n sentences of the prediction are taken into account, where n is the number of sentences of the reference.

On SLn, we observe that M2 has better recall than M1 but worse precision due to the longer generated impressions, which caused the final lower performance. On SLy, it is noticeable that both models gain more precision than lose recall; thus, our models place more relevant information at the beginning. SLy shows higher harmonic mean values than SLn, which indicates that we could improve the performance of our models by just focusing on making the models increase their focus on the reference length. Moreover, the harmonic mean values also show that M2 places more relevant content than M1 at the beginning of the text because limiting the number of sentences was more beneficial for M2 than for M1. Therefore, the additional data boosted the model in aspects that were not noticeable by using F1 measures.

4.7 Conclusions

We presented an approach for Radiology Report Summarization that continuously pre-trains a general domain BART model. This approach focuses on two main aspects: the use of biomedical data to adapt the model to this specific domain and the use of several pre-training tasks designed to inject linguistic knowledge and increase the abstractivity of the generated summaries. After the pre-training phase, we fine-tuned this model with different amounts of data from the shared task.
We also presented a study of the relationship between the models performance and the lengths of the generated summaries. We observed that our models condense the main information in the first sentences of the summaries. From the length distribution of the summaries, we found that our models tend to generate summaries with a common length; meanwhile, the reference summaries present more length variability. It seems that this behavior could penalize the performance of our models, especially on those reports with short reference summaries.

Chapter 4. References

Ahuir, Vicent et al. (2021). "NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish". In: *Applied Sciences* 11.21. ISSN: 2076-3417. DOI: 10.3390/app11219872. URL: https://www.mdpi.com/2076-3417/11/21/9872 (cit. on p. 80).

Ahuir Esteve, Vicent, Encarna Segarra, and Lluis Hurtado (July 2023). "ELiRF-VRAIN at BioNLP Task 1B: Radiology Report Summarization". In: *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Ed. by Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen. Toronto, Canada: Association for Computational Linguistics, pp. 524–529. DOI: 10.18653/v1/2023.bionlp-1.52. URL: https://aclanthology.org/2023.bionlp-1.52 (cit. on p. 75).

Ben Abacha, Asma, Yassine Mrabet, et al. (June 2021). "Overview of the MEDIQA 2021 Shared Task on Summarization in the Medical Domain". In: *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, pp. 74–85. DOI: 10.18653/v1/2021.bionlp-1.8. URL: https://aclanthology.org/2021.bionlp-1.8 (cit. on p. 78).

Ben Abacha, Asma, Chaitanya Shivade, and Dina Demner-Fushman (Aug. 2019). "Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering". In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, pp. 370–379. DOI: 10.18653/v1/W19-5039. URL: https://aclanthology.org/W19-5039 (cit. on p. 78).

Brown, Tom et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (cit. on p. 78).

Dai, Songtai et al. (June 2021). "BDKG at MEDIQA 2021: System Report for the Radiology Report Summarization Task". In: *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, pp. 103–111. DOI: 10.18653/v1/2021.bionlp-1.11. URL: https://aclanthology.org/2021.bionlp-1.11 (cit. on p. 78).

Delbrouck, Jean-Benoit, Pierre Chambon, et al. (Dec. 2022). "Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 4348–4360. URL: https://aclanthology.org/2022.findings-emnlp.319 (cit. on p. 80).

Delbrouck, Jean-Benoit, Maya Varma, et al. (July 2023). "Overview of the RadSum23 Shared Task on Multi-modal and Multi-anatomical Radiology Report Summarization". In: *Proceedings of the 22st Workshop on Biomedical Language Processing*. Toronto, Canada: Association for Computational Linguistics (cit. on p. 79).

Delbrouck, Jean-benoit et al. (May 2022). "ViLMedic: a framework for research at the intersection of vision and language in medical AI". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Dublin, Ireland: Association for Computational Linguistics, pp. 23–34. DOI: 10.18653/v1/2022.acl-demo.3. URL: https://aclanthology.org/2022.acl-demo.3 (cit. on p. 80).

Ferger, Wirth F (1931). "The nature and use of the harmonic mean". In: *Journal of the American Statistical Association* 26.173, pp. 36–40 (cit. on p. 81).

Goldberger, A. L. et al. (2000). "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". In: *Circulation* 101.23, e215–e220. DOI: https://doi.org/10.1161/01.CIR.101.23.e215 (cit. on p. 80).

Grusky, Max, Mor Naaman, and Yoav Artzi (June 2018). "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 708–719. DOI: 10.18653/v1/N18-1065. URL: https://aclanthology.org/N18-1065 (cit. on p. 78).

Gu, Yu et al. (Oct. 2021). "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *ACM Trans. Comput. Healthcare* 3.1. ISSN: 2691-1957. DOI: 10.1145/3458754. URL: https://doi.org/10.1145/3458754 (cit. on p. 78).

Hermann, Karl Moritz et al. (2015). "Teaching Machines to Read and Comprehend". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 1693–1701. URL: http://dl.acm.org/citation.cfm?id=2969239.2969428 (cit. on p. 78).

Johnson, Alistair, Tom Pollard, Steven Horng, et al. (2023). "MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2)". In: PhysioNet. DOI: https://doi.org/10.13026/1n74-ne17 (cit. on p. 80).

Johnson, Alistair, Tom Pollard, and Roger Mark (2016). "MIMIC-III Clinical Database (version 1.4)". In: PhysioNet. DOI: https://doi.org/10.13026/C2XW26 (cit. on pp. 79, 80).

Johnson, Alistair, Tom Pollard, Roger Mark, et al. (2019). "MIMIC-CXR Database (version 2.0.0)". In: PhysioNet. DOI: https://doi.org/10.13026/C2JT1Q (cit. on p. 80).

Lee, Jinhyuk et al. (2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4, pp. 1234–1240 (cit. on p. 78).

Lewis, Mike et al. (July 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*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703 (cit. on p. 78).

Liaw, Richard et al. (2018). "Tune: A Research Platform for Distributed Model Selection and Training". In: *arXiv preprint arXiv:1807.05118* (cit. on p. 81).

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013 (cit. on p. 80).

Papineni, Kishore et al. (July 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040 (cit. on p. 80).

Raffel, Colin et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html (cit. on p. 78).

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., pp. 5998–6008. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 78).

Wolf, Thomas et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6 (cit. on p. 80).

Yuan, Hongyi et al. (May 2022). "BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model". In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dublin, Ireland: Association for Computational Linguistics, pp. 97–109. DOI: 10.18653/v1/2022.bionlp-1.9. URL: https://aclanthology.org/2022.bionlp-1.9 (cit. on p. 78).

Zhang, Jingqing et al. (July 2020). "PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. ICML20. PMLR. JMLR.org, pp. 11328–11339. URL: https://proceedings.mlr.press/v119/zhang20ae.html (cit. on p. 78).

Zhang, Tianyi et al. (2020). "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations*. OpenReview.net. URL: https://openreview.net/forum?id=SkeHuCVFDr (cit. on p. 80).

Chapter 5

ELiRF-VRAIN at BioLaySumm: Boosting Lay Summarization Systems Performance with Ranking Models

Vicent Ahuir, Diego Torres, Encarna Segarra, and Lluís-F. Hurtado (Aug. 2024). "ELiRF-VRAIN at BioLaySumm: Boosting Lay Summarization Systems Performance with Ranking Models". In: *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*. Ed. by Dina Demner-Fushman et al. Bangkok, Thailand: Association for Computational Linguistics, pp. 755–761. DOI: 10.18653/v1/2024.bionlp-1.68. URL: https://aclanthology.org/2024.bionlp-1.68

Impact Index

In 2024, *The 23nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks* was co-located at the *Association for Computational Linguistics* conference; a congress ranked by the *Computing Research & Education* (CORE) as CORE-A+ in 2023.

Abstract

This paper presents our contribution to the BioLaySumm 2024 shared task of the 23rd BioNLP Workshop. The task is to create a lay summary, given a biomedical research article and its technical summary. As the input to the system could be large, a Longformer Encoder-Decoder (LED) has been used. We continuously pre-trained a general domain LED model with biomedical data to adapt it to this specific domain. In the pre-training phase, several pre-training tasks were aggregated to inject linguistic knowledge and increase the abstractivity of the generated summaries. Since the distribution of samples between the two datasets, eLife and PLOS, is unbalanced, we fine-tuned two models: one for eLife and another for PLOS. To increase the quality of the lay summaries of the system, we developed a regression model that helps us rank the summaries generated by the summarization models. This regression model predicts the quality of the summary in three different aspects: *Relevance*, *Readability*, and *Factuality*. We present the results of our models and a study to measure the ranking capabilities of the regression model.

5.1 Introduction

Nowadays, there is more information than ever at the disposal of the general public. In the specific domain of biomedical research, there is information that would be interesting to non-expert audiences, including journalists or even members of the public, such as what occurred during the recent COVID-19 global pandemic (L. L. Wang et al., 2020). However, the technical language is a barrier for the non-specialist public that may prevent them from accessing that information (Goldsack, Z. Zhang, et al., 2022; Guo et al., 2021).

Abstract summarization models should be useful in reducing the gap in understanding information. Since the models can generate a concise summary of a given text and capture its most relevant information (Beltagy, Peters, and Cohan, 2020; Brown et al., 2020; Lewis et al., 2020; Raffel et al., 2020). It is possible to obtain new models that generate summaries adapted to a much wider audience; what is known as *lay summary*. In a lay summary, the text should contain the main ideas of the article that would be interesting for a non-expert audience, enhancing readability by adding background information and reducing (or avoiding) technical terminology.

In this paper, we present the results and analysis of our system in the participation at the BioLaySumm (Goldsack, Scarton, et al., 2024) at the 23rd BioNLP Workshop (Demner-Fushman et al., 2024).

5.2 Task Drescription

In the 2024 edition, the BioLaySumm poses a single shared task, rather than two, as in the previous edition (Goldsack, Luo, et al., 2023). The task is to create a lay summary, given a biomedical research article and its technical summary (abstract section of the article).

The organization provides a biomedical dataset (Goldsack, Z. Zhang, et al., 2022) that contains biomedical research articles from two sources: *eLife Sciences*[1] and *Public Library of Science* (PLOS)[2]. Each sample contains the text of the article, the technical summary, and the reference lay summary. The dataset is divided into three partitions: `train`, `val`, and `test`.

Table 5.1 shows the sample distribution of each source. It can be observed that the number of samples is way unbalanced towards the PLOS source, even though `test` presents the same number of samples for each source. This kind of distribution would be challenging when someone would like to develop a single summarization

[1] `https://elifesciences.org/`
[2] `https://plos.org`

Table 5.1: Dataset samples distribution per partition and source. Additionally to the number of samples, the table also shows the percentage over the source.

	train	val	test
eLife	4346 (91.9)	241 (5.1)	142 (3.0)
PLOS	24773 (94.3)	1376 (5.2)	142 (0.5)

model without prompting or instructions. The alternative would be to create separate summarization models, one for eLife and the other for PLOS. The BioLaySumm organizers invited the participants to present solutions indistinctly using one or two models.

To measure the performance of the systems, the organizers of the competition selected a set of measures that would help to evaluate the performance in three different aspects: *Relevance*, *Readability*, and *Factuality*. For ***Relevance*** the following scores were chosen: ROUGE (1, 2, L) (Lin, 2004), BERTScore (T. Zhang et al., 2020). To measure the ***Readability*** aspect: Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Dale-Chall Readability Score (DCRS) (Dale and Chall, 1948), Coleman-Liau Index (CLI) (Coleman and Liau, 1975), and LENS (Maddela et al., 2023). Finally, to measure ***Factuality***, two scores were selected: AlignScore (Zha et al., 2023), SummaC (Laban et al., 2022).

5.3 Pre-trainined Model

For this task, we have used a Longformer Encoder-Decoder (LED) (Beltagy, Peters, and Cohan, 2020) since we were approaching summarizing long texts, such as the case of scientific articles. This lets us increase the amount of information available on the encoder side. We used as a starting point the LED base model from AI2[3], publicly available at the repository of HuggingFace (Wolf et al., 2020), and continuously pre-trained it with in-domain data.

For the continual pre-training phase, we followed the training methodology used in the News Abstractive Summarization models (NAS) work (Ahuir, Hurtado, et al., 2021). This methodology combines multiple pre-training tasks to incorporate linguistic knowledge in the pre-training phase and enhance the abstract nature of the produced summaries. Incorporating those tasks in continuous pre-training should help the model to transfer knowledge specific to the summarization task and increase the performance of the downstream model after fine-tuning, just as it did in the original NAS work.

[3] https://huggingface.co/allenai/led-base-16384

The data used for continuous pre-training was chosen specifically to adapt the model to the biomedical research domain. We collected text from three different sources: abstracts (technical summaries) from PubMed (National Center for Biotechnology Information (NCBI), 2024) (17M samples), PubMed articles and abstracts from the `scientific_papers`[4] dataset (Cohan et al., 2018) (240K). Also, articles and technical summaries from the dataset train partition used in this competition (eLife + PLOS) (29K).

Due to infrastructure limitations, we limited the encoder input to work with no more than 4096 tokens. Taking into account this restriction, and with the objective of maximizing the amount of data, we split text by lines, using a window of no more than 4000 words. We generated subsamples that contained at least a new line and filled the windows with as many words as possible. The final amount of samples went up to 59M samples.

When working with LongFormers, you have to select which tokens will receive global attention in addition to local attention. In the original work (Beltagy, Peters, and Cohan, 2020), the authors recommend setting `[CLS]` token with global attention. However, we hypothesized that adding landmarks across the input with global attention could increase performance. For this reason, we added a special token with global attention (`<sent>`) after a certain number of sentences. The number of sentences was not constant but dictated by a minimum number of words of separation between `<sent>` tokens. Thus, the special token was placed at the end of every number of sentences with a total length of at least k words. Previous experimentation was carried out to determine the number of words. The best results were obtained with at least $k = 20$ words of separation.

The base model was pre-trained for three epochs in our Research Institute's cluster with 8 NVIDIA A40 graphic cards with 48GB of VRAM were used for the process; which took a month. The main hyperparameters are: 128 samples per device, 4 gradient accumulation steps, a learning rate of 5×10^{-5} with a constant scheduler, gradient checking, and an 8-bit quantified optimizer.

5.4 Lay Summarization Models

We developed two different approaches for the competition. In the first approach (**M1**), the model receives the technical summary and adapts the text and information to a lay summary style. In the second approach (**M2**), additional text is included beside the technical summary, that was, the introduction and the discussion sections of the article, similar to (Poornash et al., 2023).

[4] `http://tiny.cc/54x2yz/scientific_papers`

Since the distribution of samples is not well-balanced, we fine-tuned two models per approach: one for eLife and another for PLOS. The four models were fine-tuned for ten epochs each with an NVIDIA RTX 3090 with 24GB; each approximation took nearly 24 hours. The relevant hyperparameters are: 4 samples per device and a learning rate of 5×10^{-5} with a linear scheduler.

In our tests over validation, M1 outperformed M2 in the overall performance. The detailed results can be seen in Table 5.A.1 (Section 5.8).

5.5 Ranking Model

In order to increase the quality of the lay summaries of the system, we developed a regression model to rank the summaries generated by the summarization models. This regression model predicts the quality of the summary in three different aspects: *Relevance*, *Readability*, and *Factuality*.

5.5.1 Dataset Creation and Model Development

We use a Longformer encoder already trained in the biomedical domain[5] to develop the regression model. The classification layer was modified from the default in HuggingFace. We use a mean-max function of the hidden states of the last attention layer to calculate the embedding that feeds the feedforward classification layer. In mean-max, the mean of the hidden states is concatenated with the max values of those hidden states.

To fine-tune the model, we needed first to find a way to obtain sample variability in the scores in the three aspects. For this reason, we employed data augmentation based on LLMs. For this purpose, we adapted to our needs the novel framework *TextMachina* (Sarvazyan, González, and Franco-Salvador, 2024) and generated new samples using four LLMs: Vicuna 13b (Chiang et al., 2023), Alpaca 13b (Taori et al., 2023), OpenChat 7.5b (G. Wang et al., 2023), and Llama2 13b (Touvron et al., 2023). Using the technical summary and the lay summary from randomly selected samples of both sources, we applied different prompts to gain diversity in the quality of the samples in the three aspects. With this data augmentation, we obtained 16236 new samples for training and 4212 for validation.

To create the training and validation partitions for regression, we use the generated samples and the technical and lay summaries from the corresponding partition of the competition dataset. To obtain the reference scores, we computed *Readability*, *Relevance*, and *Factuality*, using the formulas shown in Section 5.8. At this point, we

[5] https://huggingface.co/kiddothe2b/biomedical-longformer-large

should remark on two details: (a) it can be noticed that all the scores are in a range $[0, 1]$, and always correlate positively with the quality of the summary, (b) due to time constraints, the *Factuality* score is only measured with *AlignScore* in the regression dataset.

The regression model was trained for five epochs in VRAIN's cluster for two days with 4 NVIDIA A30 with 24GB of VRAM. The main hyperparameters are: 6 samples per device, 2 gradient accumulation steps, a learning rate of 5×10^{-5} with a lineal scheduler, gradient checking, and an 8-bit quantified optimizer.

5.5.2 Usage and Performance

To rank the samples, we first score them. For scoring the quality of a lay summary, we used the regression model to measure the quality regarding the *Relevance*, *Readability*, and *Factuality*. With those values, we compute a single score based on the harmonic mean of those three values. The harmonic mean would give better scores to summaries that simultaneously hold high quality on the three aspects. We will refer to this score as `hm-score` for clarity.

In order to measure the ranking capabilities of the regression model, we measured the Normal Discounted Cumulative Gain (NDCG) over the real `hm-score` of the score of the best summary available and the real score of the chosen summary, based on the predicted `hm-score`.

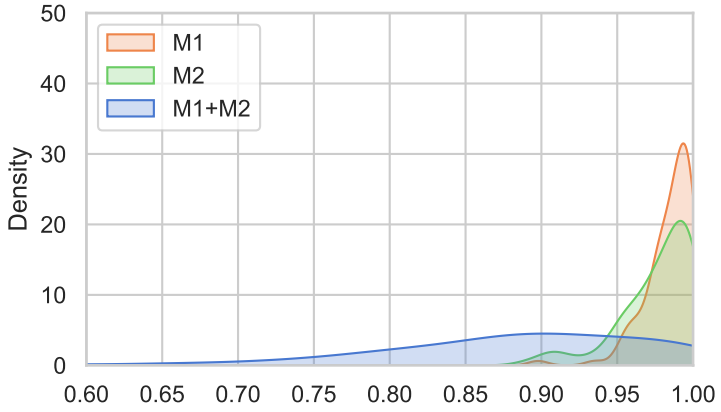


Figure 5.1: Distribution of the $NDCG_1$ scores obtained by the ranking model, when we consider both sources (eLife+PLOS). In `M1` and `M2`, the model ranks 10 summaries per sample; 20 summaries in `M1+M2`.

In Figure 5.1, we observe the distribution of the NDCG scores when the model ranks one approach (M1 or M2) and when the model ranks a mix of both (M1+M2). It

can be noticed that with M1, it has better ranking capabilities than with M2. However, in both approaches, the scores are mainly in range of $[0.95, 1.0]$, which means that most of the time, one of the best summaries is chosen. When we mix the sources, the regression model reduces its ranking capabilities, which could indicate that it would be less precise when the quality of summaries to choose from varies a lot.

The improvements in validation using the Ranking can be seen in Table 5.A.1 (Section 5.8) can be seen for M1 M2, and M1+M2.

5.6 Results

For the competition, we sent a total of three submissions. ***S1*** that included lay summaries generated with M1 approach without any kind of ranking. ***S2*** that contained lay summaries generated with M1 and selected with the rank model (10 summaries per sample). Additionally, we sent a third submission (***S3***) that contained summaries from M1 and M2 and selected with the regression model (20 summaries per sample).

Table 5.1: Official results comparison for test partition for the three submissions (S1, S2, S3), and relative performance (RP) of S2 compared to the best overall system in the competition (UIUC_BioNLP). Bold values are the best values for each score. The up arrow (↑) indicates that the value of the score correlates positively with the quality of the lay summary, and the down arrow (↓) negatively. The `hm-score` is also included, which is not part of the official results.

	S1	**S2**	**S3**	**RP(%)**
Relevance				
↑ ROUGE-1	47.99	**48.15**	48.01	98.39
↑ ROUGE-2	13.61	**13.66**	13.60	87.06
↑ ROUGE-L	42.90	**43.09**	43.06	94.07
↑ BERTScore	85.94	**85.95**	85.91	99.05
Readability				
↓ FKGL	13.64	**13.61**	13.65	86.33
↓ DCRS	10.89	**10.86**	10.90	86.00
↓ CLI	14.71	**14.66**	14.70	91.13
↑ LENS	47.90	**48.02**	33.42	90.96
Factuality				
↑ AlignScore	78.37	78.21	**78.72**	97.71
↑ SummaC	60.91	60.66	**61.37**	82.67
`hm-score`	48.68	**48.69**	46.59	90.08

Table 5.1 shows official results for the test partition for the three submissions. It

can be noticed that S2 provided the best results. Compared to S1, S2 increased the performance thanks to the ranking model. However, if the summarization model can not generate a wider variety of proposals, the ranking model will not help too much. Regarding S3, which includes the M1 and M2 summaries, we notice a lower quality of the final selection. Nevertheless, this submission increases the *Factuality* aspect, which could be attributed to the fact that M2 manages more information, reducing the factuality errors. Finally, regarding the relative performance (RP), our solution obtained more than 90% of performance in most of the scores, compared to the best overall submission. Further improvements need to be made, especially in the readability aspect.

5.7 Discusions

The results presented in Section 5.6 raise the benefits and constraints that must be taken into account when combining generation models with ranking models to choose which text will be presented to the end user.

Regarding the benefits, they are evident. With the ranking models, we can enhance the quality of the summaries presented to the user even though we use the same automatic summarization models. We use the ranking model to choose those summaries that obtained the best ranking scores since those texts will have better quality compared to other summaries generated by the same models. This selection should boost the overall performance of the system in most cases.

In relation to the constraints. The ranking model does not generate summaries or make texts better; it just rates summaries generated by the summarization models, and we select the best summaries based on those scores. Therefore, if summarization models have a bad performance and/or we can not provide enough variety to choose from, the benefits will be diminished. For this reason, we should combine the ranking models with summarization models that can complement each other depending on the text to summarize and offer variety in the generated summaries.

5.8 Conclusions

In this work, we have presented our contribution to the BioLaySumm 2024 shared task of the 23rd BioNLP Workshop. We used LED models to allow adding more text in the model input. Although we started from the same pre-trained model, different fine-tuned models were trained for the two sources of the competition: eLife and PLOS. Two different approaches were followed, one with just the technical summary as input, and another with additional text beside the technical summary. Our preliminary evaluation showed that the first approach performed better, but the second should be

developed further since the larger input context improved the *Factuality* aspect. An additional contribution of our approach is the use of a regression-based ranking model that helped to boost the quality of the final summary by choosing the promising one from a set of summaries generated by the models. The model that obtained the best results in the competition was the one that combined the first approach and the ranking model.

Chapter 5. References

Ahuir, Vicent, Lluís-F. Hurtado, et al. (2021). "NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish". In: *Applied Sciences* 11.21. ISSN: 2076-3417. DOI: 10.3390/app11219872. URL: https://www.mdpi.com/2076-3417/11/21/9872 (cit. on p. 93).

Ahuir, Vicent, Diego Torres, et al. (Aug. 2024). "ELiRF-VRAIN at BioLaySumm: Boosting Lay Summarization Systems Performance with Ranking Models". In: *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*. Ed. by Dina Demner-Fushman et al. Bangkok, Thailand: Association for Computational Linguistics, pp. 755–761. DOI: 10.18653/v1/2024.bionlp-1.68. URL: https://aclanthology.org/2024.bionlp-1.68 (cit. on p. 89).

Beltagy, Iz, Matthew E. Peters, and Arman Cohan (2020). "Longformer: The Long-Document Transformer". In: *arXiv:2004.05150* (cit. on pp. 92–94).

Brown, Tom B. et al. (2020). "Language Models are Few-Shot Learners". In: arXiv: 2005.14165 [cs.CL] (cit. on p. 92).

Chiang, Wei-Lin et al. (Mar. 2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. URL: https://lmsys.org/blog/2023-03-30-vicuna/ (cit. on p. 95).

Cohan, Arman et al. (2018). "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. DOI: 10.18653/v1/n18-2097. URL: http://dx.doi.org/10.18653/v1/n18-2097 (cit. on p. 94).

Coleman, Meri and T L Liau (1975). "A computer readability formula designed for machine scoring". In: *Journal of applied psychology* 60.2, p. 283 (cit. on p. 93).

Dale, Edgar and Jeanne S Chall (1948). "A formula for predicting readability". In: *Educational research bulletin* 27.1, pp. 11–28 (cit. on p. 93).

Demner-Fushman, Dina et al., eds. (Aug. 2024). *The 23nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Bangkok, Thailand: Association for Computational Linguistics (cit. on p. 92).

Goldsack, Tomas, Zheheng Luo, et al. (July 2023). "Overview of the BioLaySumm 2023 Shared Task on Lay Summarization of Biomedical Research Articles". In: *The 22nd*

Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Toronto, Canada: Association for Computational Linguistics, pp. 468–477. DOI: 10.18653/v1/2023.bionlp-1.44. URL: https://aclanthology.org/2023.bionlp-1.44 (cit. on p. 92).

Goldsack, Tomas, Carolina Scarton, et al. (Aug. 2024). "Overview of the BioLaySumm 2024 Shared Task on the Lay Summarization of Biomedical Research Articles". In: *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Bangkok, Thailand: Association for Computational Linguistics (cit. on p. 92).

Goldsack, Tomas, Zhihao Zhang, et al. (Dec. 2022). "Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 10589–10604. URL: https://aclanthology.org/2022.emnlp-main.724 (cit. on p. 92).

Guo, Yue et al. (May 2021). "Automated Lay Language Summarization of Biomedical Scientific Reviews". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.1, pp. 160–168. DOI: 10.1609/aaai.v35i1.16089. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16089 (cit. on p. 92).

Kincaid, J. Peter et al. (1975). "Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel". In: *Research Branch report* 8, p. 75 (cit. on p. 93).

Laban, Philippe et al. (Feb. 2022). "SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization". In: *Transactions of the Association for Computational Linguistics* 10, pp. 163–177. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00453. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00453/1987014/tacl_a_00453.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00453 (cit. on p. 93).

Lewis, Mike et al. (July 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*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703 (cit. on p. 92).

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013 (cit. on p. 93).

Maddela, Mounica et al. (July 2023). "LENS: A Learnable Evaluation Metric for Text Simplification". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computa-

tional Linguistics, pp. 16383–16408. DOI: 10.18653/v1/2023.acl-long.905. URL: https://aclanthology.org/2023.acl-long.905 (cit. on p. 93).

National Center for Biotechnology Information (NCBI) (2024). *PubMed: A Resource by the National Center for Biotechnology Information*. https://pubmed.ncbi.nlm.nih.gov/ (cit. on p. 94).

Poornash, A.s. et al. (July 2023). "APTSumm at BioLaySumm Task 1: Biomedical Breakdown, Improving Readability by Relevancy Based Selection". In: *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Ed. by Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen. Toronto, Canada: Association for Computational Linguistics, pp. 579–585. DOI: 10.18653/v1/2023.bionlp-1.61. URL: https://aclanthology.org/2023.bionlp-1.61 (cit. on p. 94).

Raffel, Colin et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html (cit. on p. 92).

Sarvazyan, Areg Mikael, José Ángel González, and Marc Franco-Salvador (2024). *TextMachina: Seamless Generation of Machine-Generated Text Datasets*. arXiv: 2401.03946 [cs.CL] (cit. on p. 95).

Taori, Rohan et al. (2023). *Stanford Alpaca: An Instruction-following LLaMA model*. https://github.com/tatsu-lab/stanford_alpaca (cit. on p. 95).

Touvron, Hugo et al. (2023). "LLaMA: Open and Efficient Foundation Language Models". In: *arXiv preprint arXiv:2302.13971* (cit. on p. 95).

Wang, Guan et al. (2023). "OpenChat: Advancing Open-source Language Models with Mixed-Quality Data". In: *arXiv preprint arXiv:2309.11235* (cit. on p. 95).

Wang, Lucy Lu et al. (July 2020). "CORD-19: The COVID-19 Open Research Dataset". In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Ed. by Karin Verspoor et al. Online: Association for Computational Linguistics. URL: https://aclanthology.org/2020.nlpcovid19-acl.1 (cit. on p. 92).

Wolf, Thomas et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6 (cit. on p. 93).

Zha, Yuheng et al. (July 2023). "AlignScore: Evaluating Factual Consistency with A Unified Alignment Function". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 11328–11348. DOI: 10.18653/v1/2023.acl-long.634. URL: https://aclanthology.org/2023.acl-long.634 (cit. on p. 93).

Zhang, Tianyi et al. (2020). "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations*. OpenReview.net. URL: https://openreview.net/forum?id=SkeHuCVFDr (cit. on p. 93).

Chapter 5. Appendix

5.A Results in Evaluation (val partition)

Table 5.A.1: Results comparison for validation partition for the two approaches without using ranking (M1 and M2), with ranking (M1R, M2R), and M1+M2 ranked (AR). Bold values are the best values achieved for each score. The up arrow (↑) indicates that the value of the score correlates positively with the quality of the lay summary, and the down arrow (↓) negatively.

	M1	M2	M1R	M2R	AR
Relevance score	48.23	45.02	**48.28**	45.26	47.26
↑ ROUGE-1	48.88	44.06	**48.97**	44.44	47.60
↑ ROUGE-2	14.52	11.20	**14.54**	11.38	13.30
↑ ROUGE-L	43.60	40.39	**43.68**	40.77	42.70
↑ BERTScore	85.90	84.41	**85.91**	84.44	85.42
Readability score	38.51	28.16	**38.64**	28.49	37.74
↓ FKGL	13.67	15.03	13.66	14.89	**13.58**
↓ DCRS	10.85	11.69	**10.82**	11.61	10.85
↓ CLI	14.47	15.53	14.43	15.43	**14.22**
↑ LENS	49.00	23.87	**49.11**	23.61	44.20
Factuality score	68.49	**81.35**	68.16	80.96	70.37
↑ AlignScore	77.00	86.65	76.64	**85.67**	77.36
↑ SummaC	59.97	76.04	59.68	**76.25**	63.38
`hm-score`	48.94	42.85	**48.97**	43.14	48.49

Table 5.A.1 shows the results of the two model types when one summary is requested (columns M1 and M2). Or, when 10 summaries are requested per sample, rank with our ranking model and select the top-ranked summary for each sample (columns M1+R and M2+R).

5.B *Relevance, Readability* and *Factuality* scores.

We defined *Relevance* as the average of the following scores: ROUGE-1, ROUGE-2, ROUGE-L and BERTScore. *Factuality* is the average values of AlignScore and SummaC scores.

For defining *Readability*, we start first defining the function *Clamp and Complement (CC)*:

$$CC_f^z(x) = \frac{z - f(x)|_{[0,z]}}{z} \tag{5.1}$$

Equation (5.1) shows that, given a function f, an integer number $z > 0$, and sample x. The sample x is evaluated with f. Then, the score is clamped in a range from $[0, z]$, complemented, and normalized.

Therefore, we define *Readability* as follows:

$$\begin{aligned} Readability(x) = (& \\ & CC_{FKGL}^{20}(x) + \\ & CC_{DCRS}^{20}(x) + \\ & CC_{CLI}^{20}(x) + \\ & \frac{LENS(x)}{100} \\ ) \cdot & \frac{1}{4} \end{aligned} \tag{5.2}$$

Equation (5.2), shows that *Readability* is defined as the average of the following four scores: FKGL, DCRS, CLI, and LENS. For the three first scores (FKGL, DCRS, and CLI), the values below 20 are clamped since we consider that 20 is already a really high readability level for lay summarization purposes. Additionally, values are complemented and normalized when needed.

Chapter 6

Abstractive Summarizers Become Emotional on News Summarization

Vicent Ahuir, José-Ángel González, Lluís-F. Hurtado, and Encarna Segarra (2024). "Abstractive Summarizers Become Emotional on News Summarization". In: *Applied Sciences* 14.2. ISSN: 2076-3417. DOI: 10.3390/app14020713. URL: https://www.mdpi.com/2076-3417/14/2/713

Impact Index

In 2023, the *Applied Sciences* journal was listed at the 1st quantile (Q1) in the Journal Citation Report (JCR) at *Engineering, Multidisciplinary* category.

Abstract

Emotions are central to understanding contemporary journalism; however, they are overlooked in automatic news summarization. Actually, summaries are an entry point to the source article that could favor some emotions to captivate the reader. Nevertheless, the emotional content of summarization corpora and the emotional behavior of summarization models are still unexplored. In this work, we explore the usage of established methodologies to study the emotional content of summarization corpora and the emotional behavior of summarization models. Using these methodologies, we study the emotional content of two widely used summarization corpora: CNN/DailyMail and Xsum, and the capabilities of three state-of-the-art transformer-based abstractive systems for eliciting emotions in the generated summaries: Bart, Pegasus, and T5. The main significant findings are as follows: (i) emotions are persistent in the two summarization corpora, (ii) summarizers approach moderately well the emotions of the reference summaries, and(iii) more than 75% of the emotions introduced by novel words in generated summaries are present in the reference ones. The combined use of these methodologies has allowed us to conduct a satisfactory study of the emotional content in news summarization.

6.1 Introduction

Storytelling is an important aspect of journalism that aims to share facts or ideas in the best way to reach, captivate attention, and convince the audience. Hence, news often does not directly re-tell events, but rather gives an interpretation of those events by a human, whose feelings can often become an important part of the story's meaning (Kennedy et al., 2012). Besides, there is clear evidence that using emotional cues helps to catch our attention and prolong our engagement (Beckett and Deuze, 2016). For this reason, emotions have become an important dynamic in how news is produced and consumed, central to our understanding of journalism (Lecheler, 2020; Richardson, 2020).

According to how online newspapers produce news articles, our entry points to a story are the headline and the summary. If they catch our attention, we will likely read the source article. Therefore, we would expect that human summarizers favor emotional content when generating summaries and headlines, potentially over/under-emphasizing some emotions compared to the source article (Kennedy et al., 2012). Table 6.1 illustrates this with two summaries for the same article that evoke different emotions.

Few works have explored emotions under the umbrella of automatic news summarization (Kennedy et al., 2012), which have otherwise been considered in other domains such as dialogue or microblogging (Chen et al., 2021; Panchendrarajan, Hsu, and Li Lee, 2021).

Nowadays, pre-trained language models are the reputable approach for developing state-of-the-art abstractive summarization systems of news articles. Their capabilities to summarize news articles have been proven, standing out in terms of phrase-overlapping metrics like Rouge (Lin, 2004), through a broad set of corpora. However, the *emotional behavior* of these systems is still unexplored. Along with other summarization aspects such as abstraction (Jumel, Louis, and Cheung, 2020), faithfulness, or factuality (Maynez et al., 2020), *emotional behavior* can shed light on how to develop better summarizers.

In this work, we explore the usage of established methodologies to study the emotional content of summarization corpora and the emotional behavior of summarization models. Using these methodologies, we carry out the first study about the emotional content of news articles and their summaries. This study is mainly based on two measures to quantify the emotional content in texts at the word level: emotion density and emotion
ratio (Kennedy et al., 2012), and is divided into two stages. First, we study the emotional content of two widely used news summarization corpora in the literature: Cnn/DailyMail (Hermann et al., 2015) and Xsum (Narayan, Cohen, and Lapata,

Table 6.1: An example of two different summaries for the same article. Using the NRC lexicon, we highlight the words that convey emotions (the emotions are listed in brackets). Phrases and emotions in blue refer to positive aspects, and those marked in red to negative aspects.

Article	Penglais Farm (Aberystwyth University) will have a total of 1000 rooms, but only 700 will be ready [anticipation] this month to welcome [joy] students. The university said developer Balfour Beatty confirmed [trust] the remaining 300 rooms will be ready [anticipation] during the 2015–2016 academic year. Balfour Beatty has been asked to comment. The unfinished [¬anticipation] rooms have not been let to students.
Summary$_1$	Hundreds of rooms at a student halls development at Aberystwyth University will not be ready [¬anticipation] for the new term.
Summary$_2$	700 rooms at Aberystwyth University will be ready [anticipation] to welcome [joy] students this month.

2018). Second, we study the capabilities of abstractive summarizer models for eliciting emotions in the generated summaries that match the emotions introduced by humans in reference summaries. This study has been performed on three state-of-the-art transformer-based
systems Fabbri et al., 2021: BART (Lewis et al., 2020), PEGASUS (J. Zhang et al., 2020), and T5 (Raffel et al., 2020). This work aims to answer the following questions: (i) what and how frequent are the emotions in documents and summaries of both corpora; (ii) how emotion densities and ratios of the generated summaries correlate with densities and ratios of the reference summaries; and (iii) whether the emotions of novel words that appear in the generated summaries but not in the source articles match emotions of their reference summary. For reproducibility purposes, the software used in this work is freely available on GitHub (`https://github.com/ELiRF/EmotionsInNewsSummarization`, accessed on 10 January 2024).

6.2 Related Work

Automatic summarization has been addressed in the literature using mainly extractive or abstractive approaches. Extractive approaches build summaries by selecting text directly from the document (Y. Liu and Lapata, 2019; Mutlu, Sezer, and Akcayol, 2020; Zhong et al., 2020), while abstractive systems build the summaries by paraphrasing

text from the document (Gehrmann, Deng, and Rush, 2018; See, P. J. Liu, and Manning, 2017). Recently, strong efforts have been made in developing abstractive systems by focusing on encoder-decoder architectures pre-trained in self-supervised ways (Lewis et al., 2020; Raffel et al., 2020; J. Zhang et al., 2020). One of the best-known problems of these systems is related to hallucinating content, where the models are prone to generate content in the summaries that is not directly inferable from the source document. Several works aim to reduce hallucinations or improve the factual consistency of abstractive summarizers, e.g., employing content planning (Narayan, Y. Zhao, et al., 2021), reinforcement learning (M. Zhang et al., 2021), or constraining the generation (Z. Zhao, Cohen, and Webber, 2020). Abstractive summarizers could also be guided, for instance, to work better on aggregating semantic information (Jumel, Louis, and Cheung, 2020), with specific topics (Belwal, Rai, and Gupta, 2021), or to represent better the keywords and relationships among the entities (Dou et al., 2021; S. Liu et al., 2022).

Along with hallucinations, factuality, and abstractivity, emotions are also important to be studied in summarization systems and in the corpora used to train them. Since summaries are an entry point to the source article, the emotions elicited in the summaries directly impact the perception of the users. Few works have considered emotions for summarization in dialogue or microblog summarization (Chen et al., 2021; Panchendrarajan, Hsu, and Li Lee, 2021), but, to our knowledge, only Kennedy et al., 2012 has studied emotions in automatic news summarization. They proposed an emotion-aware news summarization system and introduced the concepts of emotion densities and ratios, which we used extensively in our work. Similarly, in our work, we use them to study salient emotions in human-written summaries of two widely used summarization corpora (Cnn/DailyMail and Xsum). Different from Kennedy et al., 2012, we also study the emotional behavior of abstractive summarization systems, and we do not ground emotions to predefined categories since (i) articles from the considered categories are discarded, (ii) current summarization corpora do not consider categories, and (iii) we aim to obtain global insights of emotions at newspaper-level.

Emotions have been studied out of the scope of news summarization, to understand the affective state of users in applications such as e-commerce (Bielozorov, Bezbradica, and Helfert, 2019), opinion analysis in social media (S. Mohammad et al., 2018; Plaza-del-Arco et al., 2021), or healthcare (Dheeraj and Ramakrishnudu, 2021; Muñoz and Iglesias, 2022). Emotions have also been studied in the news domain to detect fake news (Kumari et al., 2022) or the stance toward specific targets (Mascarell et al., 2021). To our knowledge, our work is the first to analyze emotions under the umbrella of news automatic summarization to obtain insights from the emotional content of news summarization corpora and the emotional behavior of abstractive summarizers.

6.3 Emotional Content Measures

We aim to quantify (i) how frequent an emotion is in a text and (ii) which emotions increase/decrease their frequency in summaries compared to their frequency in articles. We base our study on the methodology introduced in Kennedy et al., 2012.

Following this methodology, we assume that the presence of an emotional word in a text is enough to convey some degree of an emotion. Although this assumption oversimplifies the problem because of the inherent limitations of lexicons, such as the lack of compositionality or ambiguity, having a moderately accurate fine-grained view of emotions in texts is useful. We use the Nrc lexicon (S. M. Mohammad and Turney, 2013) (version 0.92), which contains 27 k words and their associations with the eight basic emotions in Plutchik's wheel (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). Ten thousand of these words were manually annotated through crowdsourcing, and the remaining 17 k words are Wordnet synonyms of the annotated words. We use the Nrc lexicon through the Nrclex Python package to detect words with emotions. Words from texts and the Nrc lexicon are lemmatized to deal with inflections.

To measure how frequent an emotion e is in a text t, we use the emotion density defined (*ED*) in Equation (6.1).

$$ED(e,t) = \frac{count(e,t)}{|t|} \tag{6.1}$$

where $count(e,t)$ is the number of words in the text t that convey the emotion e following the Nrc lexicon, and $|t|$ is the number of words in t. We compute the emotion density on articles, reference summaries, and generated summaries.

To quantify emotions that appear more/less frequently in a summary than in an article, we use the emotion ratio (*ER*). The emotion ratio of an emotion e in an article-summary pair is defined in Equation (6.2).

$$ER(e,a,s) = \frac{ED(e,s)}{ED(e,a)} \tag{6.2}$$

where a is an article and s a summary. When $ER(e,a,s) > 1$, we say the emotion e is overemphasized in the summary. On the contrary, when $ER(e,a,s) < 1$, we state that the emotion is underemphasized in the summary. Intuitively, emotions that are more frequent in reference summaries than in source articles should also be more numerous in generated summaries (Kennedy et al., 2012). To measure this, we compute the emotion ratios for both reference summaries and summaries generated by abstractive summarizers.

6.4 Summarization Corpora

To conduct our study about emotions in news corpora and abstractive models, we choose two reference corpora in English news summarization: CNN/DailyMail and Xsum. Both corpora are publicly available on the HuggingFace hub: https://huggingface.co/datasets/cnn_dailymail (CNN/DailyMail, version 3.0.0, accessed on 10 January 2024), and https://huggingface.co/datasets/xsum (Xsum, accessed on 10 January 2024). Table 6.1 shows the number of samples and statistics for documents and summaries for both corpora.

Table 6.1: Statistics for the two corpora: CNN/DailyMail and Xsum. From left to right: corpus size, average document, and summary length (in terms of words and sentences), and vocabulary size in document and summary.

	#docs	avg. doc. Length		avg. sum. Length		Vocabulary Size	
		Words	Sentences	Words	Sentences	Document	Summary
CNN/DM	311,971	693.62	38.55	49.00	3.70	839,788	231,778
Xsum	226,711	377.51	19.20	21.33	1.00	425,532	83,414

6.5 Emotions in Summarization Corpora

First, we study how frequently the articles and summaries contain emotional words. To this aim, Table 6.1 shows the percentage of articles and summaries that has at least one word of an emotion.

Table 6.1: Percentage of articles (%D) and summaries (%S) in both corpora containing at least one word of an emotion.

		Fear	Anger	Anticipation	Trust	Surprise	Sadness	Disgust	Joy
CNN/DM	**%D**	99.66	98.81	99.58	99.98	99.24	99.67	97.17	99.40
	%S	79.22	69.63	84.23	90.12	60.36	76.38	53.81	69.79
Xsum	**%D**	95.72	91.74	98.60	99.04	93.58	96.26	85.70	94.97
	%S	59.45	46.89	61.19	70.52	38.36	54.17	31.75	44.38

Most articles in both corpora show some emotion, and it is common to see all the emotions co-occurring (77% of articles in Xsum and 95% in CNN/DailyMail have words representing all the emotions at some point). It is not so in the summaries: the percentage of summaries that elicit each emotion is lower than the percentage of articles, especially in Xsum, and it is not as frequent as in the articles where all the

emotions co-occur. *Fear*, *sadness*, *anticipation*, and *trust* are the emotions that appear in a more significant number of articles and summaries.

We carried out a study of the most frequent combination of emotions in the summaries. The study shows that there are larger combinations of emotions in CNN/DAILYMAIL than in XSUM, likely because summaries are twice as long. Interestingly, summaries of CNN/DAILYMAIL are twice as long as XSUM ones, but it is four times more likely that all emotions appear in their summaries (23.08% vs. 5.68%). Of the CNN/DAILYMAIL summaries, 52.43% are in the top-10 combinations, while, in XSUM, the top-10 combinations accumulate 29.74% of the summaries. Figures 6.A.1 and 6.A.2 of Appendix 6.8 show this study.

We found 27.8k examples in XSUM (12%) and 2.8k in CNN/DAILYMAIL (0.9%) where the reference summaries elicit at least one emotion that does not appear in the article. For both corpora, the most frequent emotions in these cases are *disgust*, *anger*, and *surprise*, and the least frequent ones are *anticipation* and *trust*. Table 6.B.1 of Appendix 6.8 shows one example from XSUM. Second, by focusing on emotion densities and ratios, we study how frequently each emotion is elicited in articles and summaries and what emotions are over/under-emphasized in the summaries. Figure 6.1 shows Kernel Density Estimation (KDE) plots of emotion densities and ratios for each emotion in both corpora. In these plots, the x-axes represent values of either emotion densities or ratios, and the y-axes define the probability density function for the kernel density estimation. The figure shows that emotion densities and ratios are similarly distributed.

Related to the articles (first column in Figure 6.1), *trust* concentrates the most significant number of articles with higher $ED(e, a)$. In contrast, *disgust* collects the most significant number of articles with lower $ED(e, a)$. The distribution of *fear* is the most skewed. Despite the differences among the $ED(e, s)$ and $ED(e, a)$ distributions, emotions in summaries (second column of Figure 6.1) show a similar behavior: *trust* concentrates the most significant number of summaries with higher $ED(e, s)$, and *disgust* with lower ones. In XSUM, the distributions are shifted toward higher values of $ED(e, s)$ compared to CNN/DAILYMAIL.

Regarding the ratios $ER(e, a, s)$ (third column in Figure 6.1), there is a tendency to overemphasize the emotion *fear* in both corpora, as suggested by the median. *Surprise*, *disgust*, and *joy* are underemphasized in both corpora. Interestingly, *disgust* is the emotion with the highest density in the tail, when $ER(e, a, s)$ is higher than ~3 in CNN/DAILYMAIL and ~4 in XSUM. In CNN/DAILYMAIL, the summaries tend to overemphasize emotions, especially the negative ones, while in XSUM they tend to overemphasize *fear*. In Table 6.B.3 of Appendix 6.8, we show an example from XSUM where the emotion ratio of negative emotions is high. *Anger*, *surprise*, *disgust*, and *joy* show a median emotion ratio of 0 in XSUM.

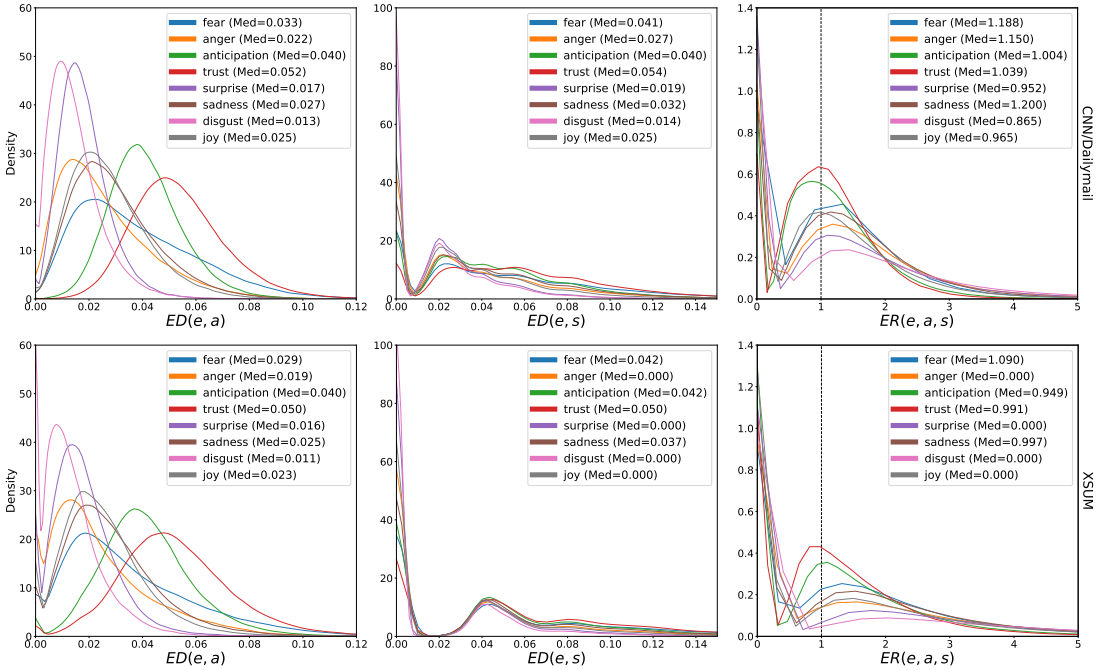


Figure 6.1: Density plots of emotion content measures in CNN/DailyMail (top row) and XSUM (bottom row). The x-axes represent emotion densities in articles ($ED(e, a)$, left column), emotion densities in summaries ($ED(e, s)$, mid column), and emotion ratios ($ER(e, a, s)$, right column). To avoid undefined values in the emotion ratios, we discarded all those examples where $ED(e, a) = 0$. The legends include the median of each emotion.

Therefore, the central tendency is not to include words with these emotions in the summary.

6.6 Emotions in Summarization Systems

In this section, we describe our study of the emotional behavior of three widely used state-of-the-art abstractive summarizers.

6.6.1 Models

We used three state-of-the-art abstractive summarization models, implemented in HuggingFace Transformers (Wolf et al., 2020), as the main systems for our experimentation: BART (Lewis et al., 2020), PEGASUS (J. Zhang et al., 2020), and T5 (Raffel et al., 2020). Since the experimentation is performed with the CNN/DailyMail and XSUM corpora, we used already finetuned checkpoints. For BART, we used *bart-large-cnn*(`https://huggingface.co/facebook/bart-large-cnn`, accessed on 10 January 2024) and *bart-*

large-xsum (https://huggingface.co/facebook/bart-large-xsum, accessed on 10 January 2024). For PEGASUS, we used *pegasus-cnn_dailymail* (https://huggingface.co/google/pegasus-cnn_dailymail, accessed on 10 January 2024) and *pegasus-xsum* (https://huggingface.co/google/pegasus-xsum, accessed on 10 January 2024). Finally, for T5, due to the lack of checkpoints for these corpora in HuggingFace, we finetuned the *t5-base* (https://huggingface.co/t5-base, accessed on 10 January 2024) model with each corpus. The two T5 models have been trained with batches of 8 samples and a constant learning rate of 5×10^{-5}, using two GIGABYTE NVIDIA RTX 3090 GPUs hosted in our research laboratory. We used early stopping to stop the training after five epochs of patience on the validation loss.

We consider two baselines commonly used in the literature for completeness: LEAD and RANDOM. LEAD extracts the first sentence of the source article in XSUM and the first three sentences in CNN/DAILYMAIL.

RANDOM extracts the same number of sentences as LEAD, but randomly selected from the source article. Additionally, we use an oracle to represent the best hypothetical summarization model. The oracle selects the sentence in the source article that maximizes the averaged ROUGE F_1 scores for each sentence in the reference summary.

For reproducibility, we show the results of these systems on the test sets in terms of ROUGE and BERTSCORE; measures commonly used in the literature for summarization Fabbri et al., 2021. The results are shown in Table 6.C.1 of Appendix 6.8. The hyper-parameters used for the abstractive summarizers are shown in Table 6.D.1 of Appendix 6.8.

6.6.2 Emotional Coherence and Bias

We analyze how emotion densities and emotion ratios of the generated summaries correlate with the corresponding metrics of the reference summaries. We introduce two metrics based on the Pearson correlation coefficient to this aim.

6.6.2.1 Emotional Coherence

Emotional coherence measures how the emotion densities for an emotion e in the generated summaries correlate with the emotion densities for e in the reference summaries. In that sense, it quantifies the strength and direction of the relation between the proportion of words with an emotion e in a generated summary and the proportion of words with that emotion in the reference summary. The emotional coherence for an emotion e is computed as the Pearson correlation between the emotion densities in the reference summaries $y = \{ED(e, s_1), \ldots, ED(e, s_N)\}$ and in the generated summaries $\hat{y} = \{ED(e, \hat{s}_1), \ldots, ED(e, \hat{s}_N)\}$.

Figure 6.1 shows the emotional coherence between reference summaries and summaries generated by each model for all the emotions and corpora.

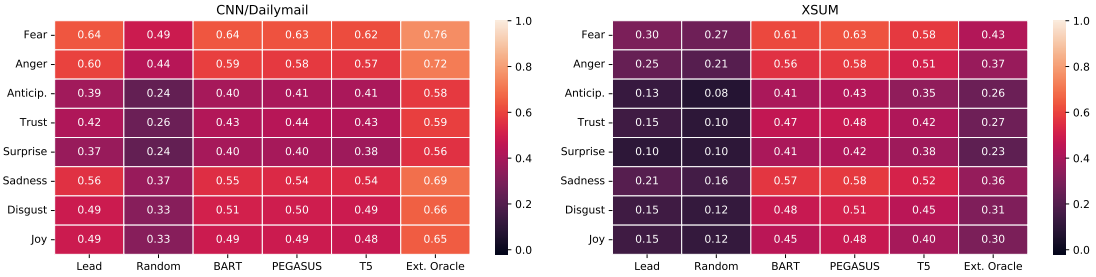


Figure 6.1: Emotional coherence of each model for each emotion in CNN/DailyMail and Xsum. Correlations are statistically significant (p-value is 0 in all the cases).

We observe that all the emotional coherences are higher than 0, suggesting positive relationships between the emotion densities. Abstractive models generally present a coherence higher than 0.5 in negative emotions: *fear*, *anger*, *sadness*, and *disgust*; and a coherence between 0.35 and 0.5 in the other emotions: *anticipation*, *trust*, *surprise*, and *joy*. Hence, abstractive models approximate better the emotion densities of negative emotions. In Xsum, T5 is the abstractive model with the lowest emotional coherence and Pegasus with the highest one. In CNN/DailyMail, Bart generally has a slightly higher emotional coherence than T5 and Pegasus. All the abstractive systems show a similar emotional coherence in both corpora.

Baseline systems also show higher emotional coherence in negative emotions. However, different from abstractive ones, these systems show low emotional coherence in Xsum. Lead shows an emotional coherence very similar to that of abstractive systems in CNN/DailyMail (slightly higher for some emotions). Hence, the first sentences of the source articles keep moderately well, and similar to the abstractive models, the expected emotion densities in the summaries of CNN/DailyMail. All the systems have higher emotional coherence than Random in both corpora.

The oracle shows the highest coherence in CNN/DailyMail, suggesting that the emotional coherence of the abstractive models could be increased if they focus on better sentences from the source (in terms of Rouge concerning the reference summary). It is not so in Xsum, where abstractive systems have higher coherence than the oracle. It suggests that focusing on the best sentences of the articles would not help to increase the emotional coherence in Xsum.

6.6.2.2 Emotional Bias

Emotional bias measures how the emotion ratios for an emotion e in the generated summaries correlate with the emotion ratios for e in the reference summary. Hence, it

quantifies the strength and direction of the relation between the emphasis, regarding the source article, placed on an emotion e in a generated summary and the emphasis placed on that emotion in the reference summary. The emotional bias for an emotion e is computed as the Pearson correlation between the emotion ratios in the reference summaries $y = \{ER(e, s_1, a_1), \ldots, ER(e, s_N, a_N)\}$ and in the generated summaries $\hat{y} = \{ER(e, \hat{s}_1, a_1), \ldots, ER(e, \hat{s}_N, a_N)\}$. To compute the emotional bias, we discard all those examples where the emotion ratio is undefined (when $ED(e, a) = 0$).

Figure 6.2 shows the emotional bias between reference summaries and summaries generated by each model for all the emotions and corpora. In almost all the cases, the emotional biases are higher than 0, suggesting positive relationships between emotion ratios. The strength of the correlations is notably lower than in the emotional coherence (Figure **??**). It suggests it is more difficult to approximate the emotion ratios than the emotion densities.

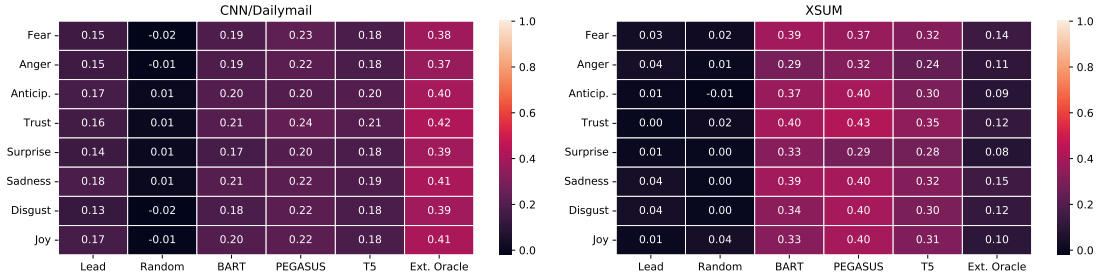


Figure 6.2: Emotional bias of each model for each emotion in CNN/DAILYMAIL and XSUM. Correlations are statistically significant (p-value is 0 in all the cases).

The abstractive systems show higher emotional bias in XSUM than in CNN/DAILYMAIL. In XSUM, T5 is the abstractive model with the lowest emotional bias. PEGASUS shows the highest emotional bias for almost all emotions in CNN/DAILYMAIL and XSUM. All the abstractive systems show, in XSUM, the lowest emotional bias for *anger* and *surprise*, and the highest emotional bias for *sadness*, *fear*, and *trust*. The emotional biases of abstractive systems are similar for all the emotions in CNN/DAILYMAIL.

Baseline models, LEAD and RANDOM, show a low emotional bias in CNN/DAILYMAIL and a negligible one (close to 0) in XSUM. The low emotional bias of LEAD indicates that the first sentences of the source articles do not show the expected emotion ratios in the summaries neither of CNN/DAILYMAIL nor Xsum. In CNN/DAILYMAIL, LEAD shows a slightly lower emotional bias than abstractive models for all the emotions, but in XSUM, the difference concerning abstractive models is high. All the systems show higher emotional bias than RANDOM.

The oracle shows the highest emotional bias in CNN/DAILYMAIL but not in XSUM, where the abstractive models stand out. It could suggest again that abstractive models

could increase their emotional bias in Cnn/DailyMail if they focus on better sentences from the source article (in terms of Rouge with respect to the reference summary) but not in Xsum.

6.6.3 Emotions of Novel Words

Abstractive summarizers are moderately good at generating summary-worthy novel words that are not present in the source. These novel words could convey a set of emotions. However, whether the emotions of the novel words are those expected in the reference summary is still unclear. We study it on the test sets of Cnn/DailyMail and Xsum by computing the precision between the emotions of the novel words in a generated summary and all the emotions in the reference summary.

Let $\mathscr{E}_s$ be the set of emotions in a reference summary and $\mathscr{E}_{\hat{s}}$ the set of emotions of the novel words in a generated one, precision (P) for N samples is computed as shown in Equation (6.3); where the intersection refers to the emotions in common between those found in the novel words and the reference summary. We only consider those cases where there are novel words with emotions in the generated summary ($|\mathscr{E}_{\hat{s}}| > 0$) and the reference one has words with emotions ($|\mathscr{E}_s| > 0$).

$$\mathrm{P} = \frac{1}{N} \sum_{\forall(s,\hat{s})} \frac{|\mathscr{E}_s \cap \mathscr{E}_{\hat{s}}|}{|\mathscr{E}_{\hat{s}}|} \tag{6.3}$$

We also compute the recall (R) to see how many of the emotions in the reference summary are covered by the emotions of the novel words in the generated summary. Recall is computed as shown in Equation (6.4).

$$\mathrm{R} = \frac{1}{N} \sum_{\forall(s,\hat{s})} \frac{|\mathscr{E}_s \cap \mathscr{E}_{\hat{s}}|}{|\mathscr{E}_s|} \tag{6.4}$$

Table 6.1 shows precision and recall for each model and corpora, along with other data statistics used to compute them. Most of the novel words generated by the models have emotions that match those of the reference summaries since precision is higher than 75% in all cases. Pegasus is the system that shows the highest precision in both corpora. T5 has slightly higher precision than Bart in Cnn/DailyMail, but not in Xsum.

The precision of all the models is higher in Cnn/DailyMail than in Xsum. The abstractive models generate more novel words in Xsum (4.9 novel words per summary) than in Cnn/DailyMail (0.9 novel words per summary). Then, generating more novel words will likely include more non-expected emotions. Table 6.B.5 of Appendix 6.8

shows an example from Xsum where the emotions of the novel words in a summary generated by Pegasus do not match exactly the emotions of the reference summary.

Table 6.1: Precision and recall of the emotions in the novel words generated by each model, compared to the emotions of the reference summaries. The number of samples without (w/o) novel words in the generated summary and w/o emotions in the novel words of the generated summary ($|\mathscr{E}_{\hat{s}}| = 0$) are also shown. The last column indicates the number of samples finally considered in the evaluation. We also show percentages of samples in the test sets.

		Precision	Recall	Samples w/o Novel Words	Samples $\|\mathscr{E}_{\hat{s}}\| = 0$	Samples
Cnn/DM	**BART**	84.73	38.79	7912 (68.9%)	2499 (21.7%)	1022 (8.9%)
	PEGASUS	85.82	36.75	6394 (55.6%)	3332 (29.0%)	1707 (14.9%)
	T5	84.99	34.47	7019 (61.1%)	3090 (26.9%)	1336 (11.6%)
Xsum	**BART**	76.10	49.49	324 (2.9%)	3717 (32.8%)	6704 (59.1%)
	PEGASUS	77.73	51.01	279 (2.5%)	3685 (32.5%)	6781 (59.8%)
	T5	75.30	47.21	354 (3.1%)	3901 (34.4%)	6554 (57.8%)

Both in Cnn/DailyMail and Xsum, Pegasus generates novel words in more samples than Bart and T5 (lowest **Samples w/o novel words**). However, for a larger number of samples than Bart and T5 in Cnn/DailyMail, the novel words generated by Pegasus do not convey emotions (highest **Samples $|\mathscr{E}_{\hat{s}}| = 0$**). By contrast, Pegasus generates emotional novel words for a slightly larger number of samples than Bart and T5 in Xsum (lowest **Samples $|\mathscr{E}_{\hat{s}}| = 0$**). We notice that the models generate more novel words in Xsum than in Cnn/DailyMail, but the number of samples where novel words do not convey emotions is similar in both corpora.

Interestingly, the recall is between 34% and 51%, which suggests that the emotions of the novel words are enough to cover, approximately, at least a third part of the overall emotional content of the reference summaries. Bart has the highest recall in Cnn/DailyMail and Pegasus in Xsum. Although it is difficult to explain why, the number of emotions in the reference summaries (lower in Xsum than in Cnn/DailyMail) could play a big role.

Considering the overall results of the two corpora, the difference in recall is significant. We consider that it is due to the difference in the introduction of new words in both cases. The fact that the Xsum corpus is much more abstractive in nature than the Cnn/DailyMail means that the former incorporates a greater number of novel words, and therefore, more emotional content.

6.7 Discussion

We summarize the most important contributions and findings of this work in relation to the objectives stated in the introduction section.

Emotional content of summarization corpora. First, we found that 99% of articles and 70% of summaries of the studied corpora contain at least one emotion. We also found that 12% in Xsum and 0.9% in Cnn/DailyMail of the reference summaries elicit at least one emotion that does not appear in the article. Second, we applied two measures, emotion density, and emotion ratio, to articles and summaries of both corpora and the results that we analyzed. Related to the articles, we observed that *trust* concentrates the most significant number of articles with higher emotion densities. In contrast, *disgust* concentrates the largest number of articles with lower emotion densities. Related to emotions in summaries, we noticed a similar behavior. In Xsum, the distributions are shifted toward higher values of emotion densities compared to Cnn/DailyMail. Regarding the emotion ratios, there is a tendency to overemphasize the emotion *fear* in both corpora. In Cnn/DailyMail, the summaries tend to overemphasize emotions, especially the negative ones, while in Xsum they tend to overemphasize *fear*.

Emotional behavior of summarization models. We introduced two new measures, *emotional coherence* and *emotional bias*, to measure how the emotion densities and ratios of generated summaries correlate with those of the reference. We found that all the emotional coherences are higher than 0, suggesting positive relationships between the emotion densities. Abstractive models generally present a coherence higher than 0.5 in negative emotions: *fear*, *anger*, *sadness*, and *disgust*; and a coherence between 0.35 and 0.5 in the other emotions. Additionally, we found a higher emotional bias in Xsum than in Cnn/DailyMail. In Xsum, T5 is the abstractive model with the lowest emotional bias. Pegasus shows the highest emotional bias for almost all emotions in Cnn/DailyMail and Xsum. Also, we analyzed whether the novel words generated by the summarization models convey the emotions expected in their reference summaries. We observed that most of the novel words generated by the models have emotions that match those of the reference summaries. Interestingly, the recall is between 34% and 51%, which suggests that the emotions of the novel words are enough to cover, approximately, at least a third part of the emotions in the reference summaries.

Finally, we should remark that the proposed methodology is valid for studying emotions in summarization regardless of the method used to detect emotions. However, the approach used in this work presents some limitations since we assumed that the presence of an emotional word in a text is enough to convey some degree of an emotion. Although this assumption oversimplifies the problem because of the inherent limitations of lexicons, such as the lack of compositionality or ambiguity, having

a moderately accurate fine-grained view of emotions in texts is helpful. Therefore, we detected emotions at the word level using lexicons, although other alternatives could exist.

6.8 Conclusions

We studied the prevalence of emotions in news summarization corpora, specifically, how much these emotions are emphasized in the summaries compared to the source article and the capabilities of state-of-the-art abstractive summarizers for eliciting expected emotions in the generated summaries.

A large percentage of articles and summaries in CNN/DAILYMAIL and XSUM elicit emotions, especially *fear*, *sadness*, *anticipation*, and *trust*. Our findings also suggest that reference summaries in CNN/DAILYMAIL overemphasize negative emotions, while XSUM underemphasizes all the emotions except *fear*. Abstractive summarizers approach moderately well the emotion densities in the summaries. However, they do not show the same emotional bias as human summarizers when emphasizing emotions in the summaries. Finally, we noticed that most of the novel words generated by the models convey emotions expected in the reference summaries, especially in CNN/DAILYMAIL, where the models generate few novel words.

In future work, we plan to develop news summarization models with controllable text generation driven by the emotions of the reference summaries and via prompting P. Liu et al., 2023, which could produce better emotional coherence in the generated summaries and potentially, reduce undesired biases towards some emotions and stances.

Chapter 6. References

Ahuir, Vicent et al. (2024). "Abstractive Summarizers Become Emotional on News Summarization". In: *Applied Sciences* 14.2. ISSN: 2076-3417. DOI: 10.3390/app14020713. URL: https://www.mdpi.com/2076-3417/14/2/713 (cit. on p. 107).

Beckett, Charlie and Mark Deuze (2016). "On the Role of Emotion in the Future of Journalism". In: *Social Media + Society* 2.3. DOI: 10.1177/2056305116662395 (cit. on p. 110).

Belwal, Ramesh Chandra, Sawan Rai, and Atul Gupta (2021). "Text summarization using topic-based vector space model and semantic measure". In: *Information Processing Management* 58.3, p. 102536. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2021.102536. URL: https://www.sciencedirect.com/science/article/pii/S0306457321000443 (cit. on p. 112).

Bielozorov, Artem, Marija Bezbradica, and Markus Helfert (2019). "The Role of User Emotions for Content Personalization in e-Commerce: Literature Review". In: *HCI in Business, Government and Organizations. eCommerce and Consumer Behavior*. Ed. by Fiona Fui-Hoon Nah and Keng Siau. Cham: Springer International Publishing, pp. 177–193 (cit. on p. 112).

Chen, Yulong et al. (Aug. 2021). "DialogSum: A Real-Life Scenario Dialogue Summarization Dataset". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 5062–5074. DOI: 10.18653/v1/2021.findings-acl.449 (cit. on pp. 110, 112).

Dheeraj, Kodati and Tene Ramakrishnudu (2021). "Negative emotions detection on online mental-health related patients texts using the deep learning with MHA-BCNN model". In: *Expert Systems with Applications* 182, p. 115265. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.115265. URL: https://www.sciencedirect.com/science/article/pii/S0957417421006977 (cit. on p. 112).

Dou, Zi-Yi et al. (June 2021). "GSum: A General Framework for Guided Neural Abstractive Summarization". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 4830–4842. DOI: 10.18653/v1/2021.naacl-main.384. URL: https://aclanthology.org/2021.naacl-main.384 (cit. on p. 112).

Fabbri, Alexander R. et al. (2021). "SummEval: Re-evaluating Summarization Evaluation". In: *Transactions of the Association for Computational Linguistics* 9. Ed. by Brian Roark and Ani Nenkova, pp. 391–409. DOI: 10.1162/tacl_a_00373. URL: https://aclanthology.org/2021.tacl-1.24 (cit. on pp. 111, 117).

Gehrmann, Sebastian, Yuntian Deng, and Alexander Rush (Oct. 2018). "Bottom-Up Abstractive Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4098–4109. DOI: 10.18653/v1/D18-1443. URL: https://aclanthology.org/D18-1443 (cit. on p. 112).

Hermann, Karl Moritz et al. (2015). "Teaching Machines to Read and Comprehend". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 1693–1701. URL: http://dl.acm.org/citation.cfm?id=2969239.2969428 (cit. on p. 110).

Jumel, Clément, Annie Louis, and Jackie Chi Kit Cheung (Nov. 2020). "TESA: A Task in Entity Semantic Aggregation for Abstractive Summarization". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8031–8050. DOI: 10.18653/v1/2020.emnlp-main.646 (cit. on pp. 110, 112).

Kennedy, Alistair et al. (2012). "Getting Emotional about News Summarization". In: *Advances in Artificial Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 121–132. ISBN: 978-3-642-30353-1. DOI: 10.1007/978-3-642-30353-1_11 (cit. on pp. 110, 112, 113).

Kumari, Rina et al. (2022). "What the fake? Probing misinformation detection standing on the shoulder of novelty and emotion". In: *Information Processing Management* 59.1, p. 102740. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2021.102740. URL: https://www.sciencedirect.com/science/article/pii/S0306457321002223 (cit. on p. 112).

Lecheler, Sophie (2020). "The Emotional Turn in Journalism Needs to be About Audience Perceptions". In: *Digital Journalism* 8.2, pp. 287–291. DOI: 10.1080/21670811.2019.1708766 (cit. on p. 110).

Lewis, Mike et al. (July 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*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703 (cit. on pp. 111, 112, 116, 134).

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013 (cit. on p. 110).

Liu, Pengfei et al. (Jan. 2023). "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Comput. Surv.* 55.9. ISSN: 0360-0300. DOI: 10.1145/3560815. URL: https://doi.org/10.1145/3560815 (cit. on p. 123).

Liu, Shuaiqi et al. (2022). "Key phrase aware transformer for abstractive summarization". In: *Information Processing Management* 59.3, p. 102913. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2022.102913. URL: https://www.sciencedirect.com/science/article/pii/S0306457322000395 (cit. on p. 112).

Liu, Yang and Mirella Lapata (Nov. 2019). "Text Summarization with Pretrained Encoders". 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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740. DOI: 10.18653/v1/D19-1387. URL: https://aclanthology.org/D19-1387 (cit. on p. 111).

Mascarell, Laura et al. (Nov. 2021). "Stance Detection in German News Articles". In: *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*. Dominican Republic: Association for Computational Linguistics, pp. 66–77. DOI: 10.18653/v1/2021.fever-1.8. URL: https://aclanthology.org/2021.fever-1.8 (cit. on p. 112).

Maynez, Joshua et al. (July 2020). "On Faithfulness and Factuality in Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: https://aclanthology.org/2020.acl-main.173 (cit. on p. 110).

Mohammad, Saif et al. (June 2018). "SemEval-2018 Task 1: Affect in Tweets". In: *Proceedings of The 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1–17. DOI: 10.18653/v1/S18-1001. URL: https://aclanthology.org/S18-1001 (cit. on p. 112).

Mohammad, Saif M. and Peter D. Turney (2013). "Crowdsourcing a Word-Emotion Association Lexicon". In: *Computational Intelligence* 29.3, pp. 436–465 (cit. on p. 113).

Muñoz, Sergio and Carlos A. Iglesias (2022). "A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations". In: *Information Processing Management* 59.5, p. 103011. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2022.103011. URL: https://www.sciencedirect.com/science/article/pii/S0306457322001212 (cit. on p. 112).

Mutlu, Begum, Ebru A. Sezer, and M. Ali Akcayol (2020). "Candidate sentence selection for extractive text summarization". In: *Information Processing Management* 57.6, p. 102359. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2020.10235

9. URL: https://www.sciencedirect.com/science/article/pii/S0306457320308542 (cit. on p. 111).

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (Oct. 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: Association for Computational Linguistics, pp. 1797–1807. DOI: 10.18653/v1/D18-1206. URL: https://aclanthology.org/D18-1206 (cit. on p. 110).

Narayan, Shashi, Yao Zhao, et al. (2021). "Planning with Learned Entity Prompts for Abstractive Summarization". In: *Transactions of the Association for Computational Linguistics* 9. Ed. by Brian Roark and Ani Nenkova, pp. 1475–1492. DOI: 10.1162/tacl_a_00438. URL: https://aclanthology.org/2021.tacl-1.88 (cit. on p. 112).

Panchendrarajan, Rrubaa, Wynne Hsu, and Mong Li Lee (2021). "Emotion-Aware Event Summarization in Microblogs". In: *Companion Proceedings of the Web Conference 2021*. New York, NY, USA: Association for Computing Machinery, pp. 486–494. ISBN: 9781450383134. URL: https://doi.org/10.1145/3442442.3452311 (cit. on pp. 110, 112).

Plaza-del-Arco, Flor Miriam et al. (2021). "Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021". In: *Procesamiento del Lenguaje Natural* 67.0, pp. 155–161. ISSN: 1989-7553. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6385 (cit. on p. 112).

Raffel, Colin et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html (cit. on pp. 111, 112, 116, 134).

Richardson, Nick (Nov. 2020). "Journalism and Emotion". In: *Australian Journalism Review* 42, pp. 339–340. DOI: 10.1386/ajr_00047_5 (cit. on p. 110).

See, Abigail, Peter J. Liu, and Christopher D. Manning (July 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)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. URL: https://aclanthology.org/P17-1099 (cit. on p. 112).

Wolf, Thomas et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6 (cit. on p. 116).

Zhang, Jingqing et al. (July 2020). "PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. ICML20. PMLR. JMLR.org, pp. 11328–11339. URL: https://proceedings.mlr.press/v119/zhang20ae.html (cit. on pp. 111, 112, 116, 134).

Zhang, Mengli et al. (2021). "FAR-ASS: Fact-aware reinforced abstractive sentence summarization". In: *Information Processing Management* 58.3, p. 102478. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2020.102478. URL: https://www.sciencedirect.com/science/article/pii/S0306457320309675 (cit. on p. 112).

Zhao, Zheng, Shay B. Cohen, and Bonnie Webber (Nov. 2020). "Reducing Quantity Hallucinations in Abstractive Summarization". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2237–2249. DOI: 10.18653/v1/2020.findings-emnlp.203. URL: https://aclanthology.org/2020.findings-emnlp.203 (cit. on p. 112).

Zhong, Ming et al. (July 2020). "Extractive Summarization as Text Matching". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6197–6208. DOI: 10.18653/v1/2020.acl-main.552. URL: https://aclanthology.org/2020.acl-main.552 (cit. on p. 111).

Chapter 6. Appendix

6.A Top-10 Emotions in Corpora

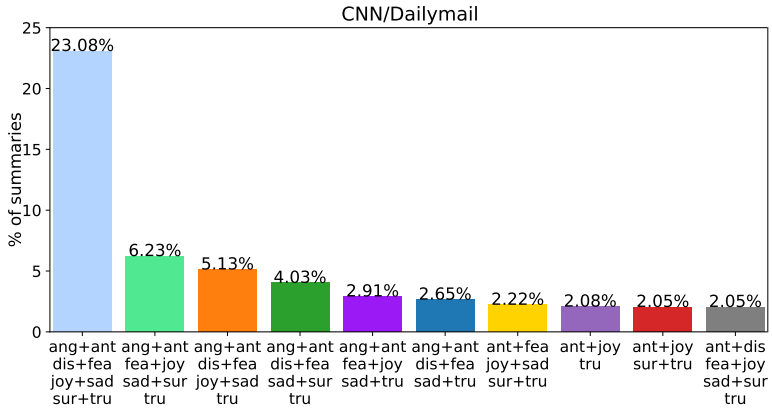


Figure 6.A.1: Ten most frequent combinations of emotions in the summaries of CNN/DAILYMAIL. Bar labels indicate the percentage of summaries in the whole corpus.

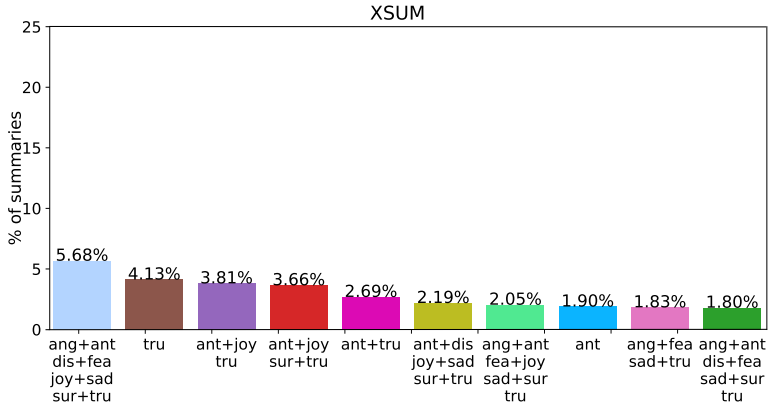


Figure 6.A.2: Ten most frequent combinations of emotions in the summaries of XSUM. Bar labels indicate the percentage of summaries in the whole corpus.

6.B Examples of Emotions in News Summaries

Table 6.B.1: An example from Xsum where the emotions *fear*, *anger*, and *disgust* appear in the summary but not in the article. Bold underlined words appear in the Nrc lexicon, and words in brackets are their emotions.

Article	The 29-year-**old** **[*sadness*]**, who had only joined from AC Milan three days earlier, was **accidentally** **[*surprise*]** **caught** **[*surprise*]** on the right side of their head by Lorient midfielder Didier Ndong 15 min into the match. "Jeremy was operated on under local anaesthesia on Wednesday night to sew up their ear," Bordeaux said. "We wish a speedy recovery to our player and **hope** **[*anticipation, joy, surprise, trust*]** to see them back soon."international Ndong said: "I apologise to Jeremy Menez, and to Bordeaux, and **hope** **[*anticipation, joy, surprise, trust*]** to see them back in Ligue 1 action soon. "It was completely involuntary but had **unfortunate** **[*sadness*]** consequences for him.
Reference summary	Bordeaux's France international forward Jeremy Menez **lost** **[*anger, disgust, fear, sadness, surprise*]** **part** **[*sadness*]** of their right ear in a pre-season game with FC Lorient.

Table 6.B.3: An example from Xsum where the emotion ratio, computed as shown in Equation (6.2), of *fear* (5.29), *anger* (7.05), and *sadness* (5.29) is higher than 5, i.e., the proportion of words with these emotions in the summary is more than 5 times higher than the proportion in the source article. Bold underlined words are those present in the Nrc lexicon, and words in brackets are their emotions. Punctuation marks are not counted as words.

Article	A telehandler vehicle was used to smash a wall at the rear of the Sainsbury's Local store on Bingham Road in Cotgrave at about 04:00 BST on Monday. The **cash** **[*trust, joy, anticipation, anger, fear*]** **machine** **[*trust*]** was taken away in another vehicle described as a white vehicle, possibly an Audi, **police** **[*trust, fear*]** said. **Officers** **[*trust*]** have **sealed** **[*trust*]** off the area and have appealed for any **witnesses** **[*trust*]** to **come** **[*anticipation*]** forward. It is unclear how much of the shop has been **damaged** **[*anger, disgust, sadness*]**.
Reference summary	A **cash** **[*trust, joy, anticipation, anger, fear*]** **machine** **[*trust*]** has been **stolen** **[*anger, fear, sadness*]** in a **ram** **[*anger, anticipation*]** **raid** **[*anger, fear, surprise*]** on a Nottinghamshire supermarket

Table 6.B.5: An example from XSUM where the emotions of the novel words in a summary generated by PEGASUS do not match exactly the emotions of the reference summary (precision = 0.25). Bold underlined words are those present in the NRC lexicon, and words in brackets are their emotions. Words in **blue** are the novel words in the generated summary.

Article	The McGill's 904 service went up in flames just outside Largs on the A760 Kilbirnie Road at about 13:35 on Saturday. Emergency services attended but the driver and passengers were uninjured. A woman whose partially-sighted mother was on board later thanked the driver for keeping everyone safe. Kathleen McKenna told the BBC: "The bus started filling up with smoke. "The driver told everyone to get off as quickly as possible. He then made sure everyone was as far away as possible. "The bus was popping and banging as the fire took hold. The driver did really, really well. "The police arrived and asked if anyone needed to go to hospital but they were all fine. They just needed a cup of tea. "Police Scotland said the road was closed for a time but later re-opened. The burnt-out bus has been removed.
Reference summary	A bus driver whose vehicle **caught [surprise]** **fire [fear]** in North Ayrshire has been **praised [joy, trust]** after all the **passengers [anticipation]** were **safely [joy, trust]** **evacuated [fear]**.
Emotions in reference	surprise, fear, joy, trust, anticipation
PEGASUS summary	A bus has been **badly [sadness]** **damaged [anger, disgust, sadness]** after **catching [surprise]** fire in North Ayrshire.
Emotions in novel words	sadness, anger, disgust, surprise

6.C Rouge/BertScore Performance

We evaluated the quality of the summaries generated by the models in terms of Rouge and BertScore. Table 6.C.1 shows the results.

Table 6.C.1: Rouge (R) and BertScore (BS) F_1-scores for all the models and corpora.

		R1	R2	RL	BS
Cnn/DM	**Lead**	40.05	17.48	36.34	23.45
	Random	28.48	8.34	25.51	11.88
	BART	43.76	20.86	40.68	33.64
	PEGASUS	43.96	21.38	41.07	35.18
	T5	43.03	20.31	40.04	32.87
	Extractive oracle	52.34	30.23	48.86	39.77
Xsum	**Lead**	16.71	1.65	12.30	14.27
	Random	15.23	1.77	11.38	11.71
	BART	45.23	22.13	37.02	50.13
	PEGASUS	47.16	24.58	39.31	52.74
	T5	40.98	18.02	32.99	48.85
	Extractive Oracle	29.38	8.68	22.43	22.66

6.D Generation Hyperparameters

For reproducibility, we show in Table 6.D.1 the hyperparameters used for the `generate` method in HuggingFace Transformers model class. We tried to keep them similar to the original implementations Lewis et al., 2020; Raffel et al., 2020; J. Zhang et al., 2020.

Table 6.D.1: Hyperparameters used during generation for all models and corpora.

	Model	Length	Max Length	Min Length	N-gram Blocking	Number of Beams
Cnn/DM	**BART**	2.0	142	56	3-g	4
	PEGASUS	0.8	128	32	No	4
	T5	2.0	142	56	3-g	6
Xsum	**BART**	1.0	62	11	3-g	6
	PEGASUS	0.6	64	No	No	6
	T5	1.0	62	11	No	6

CHAPTER 7

Beyond Using Their Own Words: Abstractivity Characterization in Summarization

Vicent Ahuir, Encarna Segarra, and Lluís-F. Hurtado (Nov. 2024). "Beyond Using Their Own Words: Abstractivity Characterization in Summarization". Work pending of publication

Impact Index _____

Peding of publication.

Abstract

In this work, we present an extension of the definition of abstractivity within the scope of the automatic generation of summaries. We propose to join extractivity and abstractivity in a single dimension, where extractivity would be on one side of the dimension and complete abstractivity on the opposite one, but in between, there would be levels of abstractivity. A dataset manually annotated to characterize the level of abstractivity of the summaries and to measure the presence of a set of actions applied to compose the summaries has been built. Using this dataset, a study of the sample distribution in terms of abstractivity, annotator agreement, and correlation between annotations regarding the set of actions is presented. An experimental work with a double objective is carried out; on the one hand, we want to validate our perception that extractivity and complete abstractivity are extreme points of a single dimension with multiple abstractivity levels, and on the other hand, we want to verify if there is an overall correlation between the frequency of the actions used for creating the summary and the level of abstractivity. The results confirm both objectives.

7.1 Introduction

Summarizing is the process of condensing the most relevant information from a document into a single, shorter document, the summary. Initially, the essential information in the article has to be identified. There are two strategies to generate the summary from the selected information. In an purelly extractive approach, the sentences with the selected information are copied directly to the summary. In an abstractive approach, the generated summaries also contain the essential information, but it is "expressed, usually, in the words of the author of the summary" (Nenkova and McKeown, 2011).

Although the first approaches to the problem were extractive, after the emergence of the Transformer architecture (Vaswani et al., 2017) and its capabilities, most of the published works have addressed the generation of summaries under abstractive approaches. However, to the best of our knowledge, the characterization of abstractivity within summaries has not been sufficiently studied (Bommasani and Cardie, 2020; Grusky, Naaman, and Artzi, 2018; Jing, 2002; Kryściński et al., 2018). A more detailed and extended characterization of the abstractivity in summaries would help to better understand how abstractive models generate their summaries.

Generally, works related to the evaluation of the level of abstractivity of the generated summaries focus on measuring the appearance of new words in the summaries compared with the summarized documents (Chen et al., 2021; Dou et al., 2021; Fu et al., 2021; Manakul and Gales, 2021; Wu et al., 2021; Zheng et al., 2020; Zou et al., 2020). This strategy conforms to Nenkova and McKeown (2011)'s definition of abstractive summaries. However, it is not the only way to produce a summary in "the author's words". It is possible to make a summary in which very few new words or expressions are introduced compared with the original document, and yet the main ideas are expressed in a different way (Ahuir, Hurtado, et al., 2021). For example, a summary can be written based mainly on the reordering of some segments extracted from the document, with the introduction of very few new elements.

Jing (2002) conducted a study on the actions that abstraction professionals used to create their abstractive summaries. Specifically, he identified the following six actions: *sentence reduction*, *sentence combination*, *syntactic transformation*, *lexical paraphrase*, *generalization/specification*, and *content reordering*. Based on the hypothesis that writing an abstractive summary is based on using this set of actions, we can characterize the abstractivity of a text by measuring the presence of each of the six actions.

In this work, we propose an extension of the definition of abstractivity in the automatic summarization area. Although in the literature, the extractive and abstractive approaches have been treated as mutually exclusive (Y. Liu and Lapata, 2019; Sun et al., 2024; Varab and Xu, 2023), we join extractivity and abstractivity in a single

dimension, what we call the level of abstractivity. The extractivity would be on one side of the dimension, and the complete abstractivity on the opposite one, but in between, there would be levels of abstractivity. Additionally, we want to add new characteristics for summaries by measuring the presence of each action identified by Jing (2002); we hypothesize that these characteristics will help to measure the level of abstractivity.

The main contributions of this work are:

(i) A dataset has been built that contains document-summary pairs manually annotated in terms of a set of actions (including the Jing (2002)'s actions) using a Likert scale: the Characterization of the Level of Abstractivity in Summarization (*CLAsum*) dataset. It is publicly available at `https://huggingface.co/datasets/??`.

(ii) Some analyses have been carried out on the *CLAsum* dataset: sample distribution in terms of abstractivity, annotator agreement, and correlation between annotations in terms of the set of actions.

(iii) To check if there is an overall correlation between the frequency of the actions used for creating the summary and the level of abstractivity, two tasks have been considered: Abstractivity Inducting Features extraction, and Abstractivity Level prediction.

(iv) Using the *CLAsum* dataset, a set of machine learning models have been trained to predict both, the Abstractivity Inducting Features and the Abstractivity Level in summaries.

(v) Using these models, some experimentation is carried out to test how beneficial the inclusion of the Abstractivity Inducting Features information is in the Abstractivity Level prediction.

7.2 The *CLAsum* dataset

7.2.1 Sample Gathering

To build a dataset that could contain diversity regarding abstractivity, we took as document-summary pairs source the test partitions of two well-known datasets in the summarization area: CNN/DailyMail (See, P. J. Liu, and Manning, 2017), and XSum (Narayan, Cohen, and Lapata, 2018). We chose these datasets since they have been deeply used with extractive and abstractive summarization methods Ahuir, González, et al., 2024; Alami Merrouni, Frikh, and Ouhbi, 2023; Giarelis, Mastrokostas, and Karacapilidis, 2023; Kabeer and Khan, 2024. We could coarsely infer what kind of summaries would be found in each dataset concerning the extractivity/abstractivity.

Specifically, in the CNN/DailyMail dataset, summaries tend to be more extractive than abstractive; whereas, the ones in XSum are more abstractive than extractive; this makes both datasets complementary for this work.

For pair sampling from each dataset, we could randomly select the pairs; however, this would preserve the original distributions of each dataset. Since we aimed to create a dataset with diversity regarding abstractivity, we designed a sampling method based on clustering that could obtain pairs with summaries with different degrees of extractivity/abstractivity, independently of the original distribution of the dataset. For each dataset, 5 clusters were created using the KMeans algorithm and a set of features related to abstractivity to characterize each pair. Ahuir, Hurtado, et al. (2021) shown that combining a set of abstractivity-related metrics is useful for abstractivity measurement; therefore, we used the following ones: Coverage and Density (Grusky, Naaman, and Artzi, 2018), Content Reordering (Ahuir, Hurtado, et al., 2021), Abstractivity (p=[2,3]) (Bommasani and Cardie, 2020), and Novel [2,3,4]-grams (Kryściński et al., 2018). Consequently, an 8-component features vector was used to characterize a pair.

From the 10 clusters, 5 per source dataset, we extracted 20 document-summary pairs per cluster. The final dataset comprised 100 pairs from CNN/DailyMail and another 100 from XSum. Additionally, we set some restrictions in the selected pairs: (1) the document should contain a maximum of 500 words, (2) the summary should contain a minimum of 38 words, and (3) the proportion of words the document/summary should be at least 2:1.

With the sampling method and the filters, we ensured that the final dataset should contain summaries with different degrees of extractivity/abstractivity, which makes the dataset useful for abstractivity studies.

7.2.2 Labeling Guideline Design

Since our main objective was to evaluate how the information in a document was modified (removing content, merging sentences, etc.) to buid the corresponding summary, the process of labeling is based on the assumption that the analyzed texts are actual summaries. However, we noticed that the datasets contained samples with no actual summaries (the summary can not be entailed from the document), which was also perceived by other authors (Guo et al., 2022). To detect those *not-summaries*, the guideline started with two questions related to the quality of the summary: one about the information included in the summary with respect to the document **(A)** and another about the amount of new information added by the summary **(B)**. Then 8 abstractivity-related questions complete the guideline, from **(C)** to **(J)**. One question about the perception of abstractivity (question **(C)**) and 7 questions for the actions

identified by Jing (2002) (from **(D)** to **(J)**); Generalization **(H)** and Specification actions **(I)** were split to gain information. The complete guideline can be found in Section 7.7.

We designed the guideline with a Likert scale. The number of options would vary from question to question since some aspects required more granularity than others. For each question, we added options until we felt that the possible answers collected enough variability and the annotators would not be forced to choose one option as a fallback. The number of options are the following ones: **(A)** Relevance of the information in the summary (5 options), **(B)** Amount of novel information within the summary (3 options), **(C)** Perception of the level of abstractivity (5), **(D)** Content exclusion (4), **(E)** Sentence information melting (3), **(F)** Syntax alteration (3), **(G)** Synonym usage (3), **(H)** Generalization usage (4), **(I)** Specification usage (4), and **(J)** Reordering (3). In question **(C)**, 1 is the definition of extractivity summarization and 5 is the definition of a high-level abstractivity summarization. From questions **(D)** to **(J)**, higher values imply a higher presence of the corresponding action.

Aditionally, we included the answer 0 ("Does not apply; it is not a summary.") for questions **(C)** to **(J)** (abstractivity-related questions). In that way, the annotators would not be forced to answer the abstractivity-related questions if they do not consider the evaluated text a valid summary.

7.2.3 Labeling Process

The labeling process was conducted by people from our research group, a total of 13 people with a high degree level of studies in Computer Science (9 University professors, 3 PhD students, and 1 Master's degree student). Additionally, 4 Computer Science degree students collaborated with the labeling process. Thus, 17 volunteers with good English level (but not native speakers) contributed to accomplishing the annotation process.

Our goal was to build an annotated dataset with more than one set of labels per document-summary pair. Therefore, we established to obtain 3 different sets of labels per pair, acquiring a total of 600 samples (document-summary pair+labels). Also, we pursued to capture the variety of perceptions from groups of people, therefore, we distributed the pairs to the annotators, avoiding the coincidence 3-annotators group between document-summary pairs as much as possible.

The labeling process was largely unguided for annotators since half were NLP experts. We exploit this fact to verify the understandability (verified by all the annotators) and correctness (verified by the experts) of the guideline without introducing any bias. To verify the guideline, we had an initial meeting to explain the labeling pipeline and resolve any doubts about the guideline; however, no labeling examples were given. Then, annotators labeled a small subset of pairs, which helped to raise

some caveats that were corrected. In this way, we obtained the final version of the guideline and ensured its quality.

The annotation process was carried out with an in-house developed labeling application called YALT![1] (technical details in Section 7.7). YALT! showed the guideline and document-summary on the left and the answer page on the right. Additionally, aiming to help annotators, the application showed the longest common extractive segments between the document and the summary in different colors (Figure 7.B.1, Section 7.7). With YALT! we had a simple workflow: (1) the annotator placed the assigned pairs in a folder, (2) they annotated each pair, (3) the labels were packed, and (4) uploaded to a shared drive. This workflow also ensured that an annotator could not see labels from other annotators.

Finally, it is worth mentioning that the annotation process was especially difficult, since most of the questions need some text analysis to be answered. On average, each pair took 15 minutes, which implies 150 hours of working time for creating the *CLAsum* dataset composed of 600 samples; nearly 9 hours per annotator that labeled around 36 pairs each.

7.2.4 Sample Distribution

Table 7.1, shows the distribution of pairs that contain an actual summary and which ones do not.

Table 7.1: Distribution of document-summary pairs in *CLAsum* that contain a summary and which do not contain an actual summary (*not-summary*).

Summary	Not a Summary
175	25

We observe that 12.5% of pairs do not contain an actual summary since they do not contain at least some information extracted from the summarized document. All the *not-summaries* pairs came from the XSum dataset. Since we were studying the abstractivity in summaries, we only used the subset of samples with valid summaries for the study. We refer to that subset as $CLAsum^s$, which contains 525 samples (175 document-summary pairs with 3 different sets of labels each).

Regarding the perception of abstractivity (question **(C)**), Figure 7.1 shows the distribution per source (where 1 represents the extractivity summarization style and 5 the highest perception of abstractivity).

[1] https://github.com/vahuir/YALT

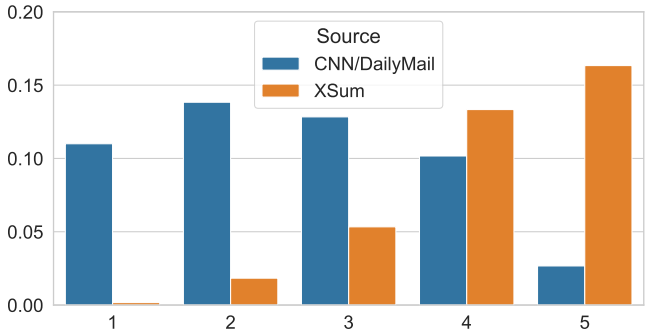


Figure 7.1: Distribution of answers for question **(C)** in *CLAsum*^{s}, regarding the perception of the abstractivity level in the summary.

Figure 7.1 shows that although the sampling process was the same for the two sources, the achived distribution of samples differs from one source to the other, regarding perception of abstractivity. It can be noticed that the clustering process collected more diverse samples from CNN/DailyMail than from XSum, regarding the abstractivity level. XSum samples tended to present a higher abstractivity level; which verifies our initial coarse inference stated in Section 7.2.1.

7.2.5 Annotator Agreement Analysis

The labeling process addresses a complex and subjective task. A total agreement between annotators cannot be expected, then, it would not be advisable to study the agreement in terms of exact matches (binary distance). Therefore, we used the *Relative* distance between two labels.

Equation (7.1) shows the definition of this distance.

$$\text{R-Dist}_Q(l_1, l_2) = \frac{|l_1 - l_2|}{M_Q - 1} \tag{7.1}$$

Given two labels (l_1, l_2) for question Q, *Relative* distance returns the percentage of the absolute distance that separates l_1 from l_2, in relation to the range between the minimum value (1) and the max value that can acquire this question (M_Q).

Table 7.2 shows the average agreement among annotators for each question. We used Cohen's and Fleiss' Kappa for the measurement, with the *Relative* distance (Equation (7.1)) as distance function between observations. *Cohen's Kappa* column is calculated with the pair-wise average score among the three annotators.

Table 7.2: Agreement scores in per Question in the *CLAsum*^{s} subset using the *Relative* distance.

Question	Cohen's Kappa	Fleiss' Kappa
(A)	0.94±0.15	0.75±0.21
(B)	1.00±0.00	0.87±0.22
(C)	0.92±0.18	0.71±0.19
(D)	0.92±0.19	0.67±0.23
(E)	0.96±0.16	0.64±0.34
(F)	0.90±0.24	0.52±0.30
(G)	0.90±0.23	0.61±0.28
(H)	0.86±0.24	0.60±0.22
(I)	0.86±0.24	0.59±0.22
(J)	0.89±0.25	0.46±0.32

It can be observed that the average agreement with Cohen's Kappa is almost perfect. However, when Fleiss' Kappa is considered, the agreement strength is reduced to substantial on most of the questions (except **(B)**), and moderate for questions **(F)**, **(I)**, and **(J)**. It can be deduced that the annotators' answers do not differ that much for a given question; however, there are slight degree deviations among the three annotations at once (the answers are not unanimous).

We extracted the distances between annotators and questions for each document-summary pair's question to analyze the deviations between annotators. The integer distance was measured between two answers; the distance was computed by counting the number of answers that separated one label from the other. Figure 7.2 shows the distribution of integer absolute distance.

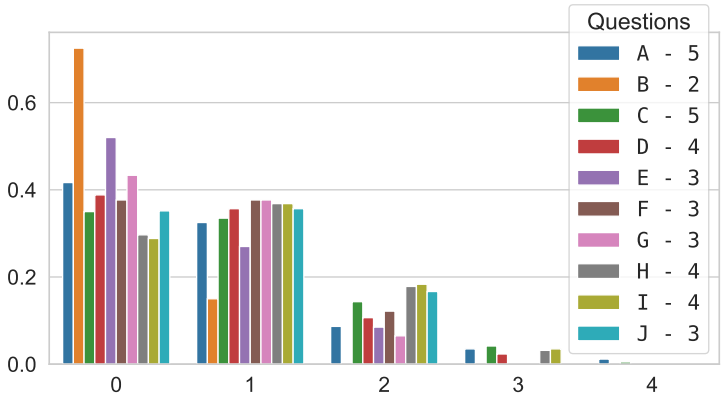


Figure 7.2: Distribution of answer distances between two annotators on labels for the same document-summary pair in the *CLAsum*^{s} subset.

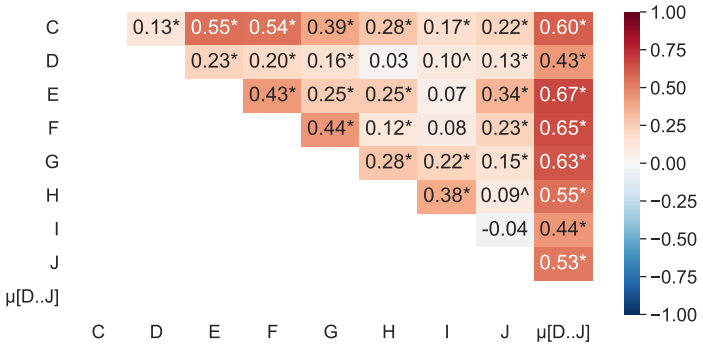


Figure 7.1: Pearson's correlation between two questions in the *CLAsum*^{s} subset. $\mu[D..J]$ is the normalized average from questions **(D)** to **(J)**. N=525; x^* means $p < 0.01$, two tails; $x^\wedge$ means $p < 0.05$, two tails.

It can be observed that 30% to 50% of the labels show agreement between annotators, excluding answer B, where the agreement elevates to more than 70% of the cases. However, if we aggregate the annotations with agreement and the ones that are at a distance 1, we cover nearly 80% of the observations in each answer.

With the information extracted from Table 7.2 and Figure 7.2, along with the average Cohen's Kappa between annotators in Section 7.7, it can be gathered that the labeling process produced a dataset that captured subjectivity but retained enough agreement to consider the data coherent and valid, from where useful information could be extracted.

7.3 Abstractivity-related Questions Correlation Analysis

In this section, we analyze in the *CLAsum*^{s} subset whether the answers to the questions related to the actions identified by Jing (2002) correlate with the perception of the level of abstractivity that the annotators had regarding the viewed summaries.

Figure 7.1 presents Pearson's correlation of questions from **(C)** to **(J)** (abstractivity-related questions). Additionally, we introduce a new column ($\mu[D..J]$), the average of the 7 questions (**(D)** to **(J)**) normalized by the maximum value that can acquire each question.

Considering the column of the *perception of abstractivity* **(C)**, two questions present a strong positive correlation with it: *sentence information melting* **(E)** and *syntax alteration* **(F)**, which indicates that they are more frequently used in the creation of more abstractive summaries. The introduction of *synonyms* **(G)** has a moderate

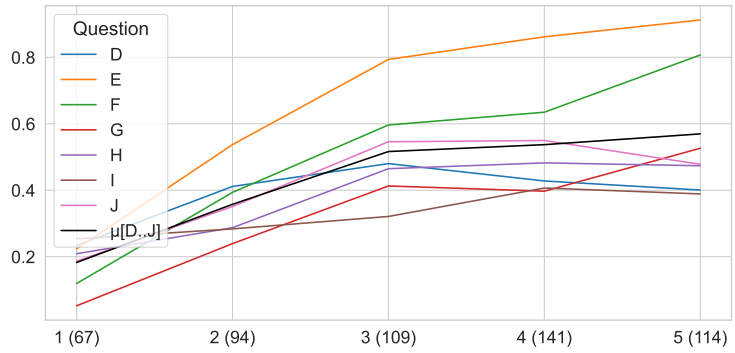


Figure 7.2: Average answer per question in the *CLAsum*^{s} subset. $\mu[D..J]$ is the normalized average from questions **(D)** to **(J)**. `X-axis`: level of abstractivity (question **(C)**), showing the number of samples in parethesis. `Y-axis`: average normalized answer.

positive correlation; hence, it has also an impact in the perception of abstractivity in summaries. The rest of the actions have a weak positive correlation with **(C)**: *excluding content* **(D)**, *include generalizations* **(H)** and *especifications* **(I)**, and *reordering content* **(J)**. It indicates that those actions have slightly more presence in higher levels of abstractivity. When we consider all Jing (2002)'s actions at the same time ($\mu[D..J]$), a strong postive correlation with **(C)** appears, with slightly higher value than the highest correlation of a single action (question **(E)**). Therefore, that indicates that the perception of abstractivity is related to how frequently those actions are used to compose a summary and how they are combined.

Figure 7.2 shows the evolution of the average answer for each question when the abstractivity level (question **(C)**) is considered. To facilitate the comparison, the answers were normalized by the maximum possible value available in their respective question. Also included the aggregation of all the answers ($\mu[D..J]$).

From abstractivity levels 1 to 3, it can be noticed a clear positive evolution in the average answer for all the questions. Therefore, in that range of abstraction, the more often are used the actions, the higher perception of abstractivity exists. This does not occur for all the actions from abstractivity levels 3 to 5. In the case of actions *sentence information melting* **(E)** or *syntax alteration* **(F)**, they are more frequently used in higher levels of abstractivity, which indicates that more abstractive summaries tend to cover more information (*content exclusion* **(D)** decreases) by joining sentences and shortening the content with syntax alterations. Also we observe an increasing pick of usage of *synonyms* **(G)** from abstractivity levels 3 and 4 to 5; which increases the chance that the reader would feel the summary written "the author's words". In the case of actions *generalizations*, *especifications*, and *content reordering*, usage frequency does not increase in this range of abstractivity. Lastly, when we consider $\mu[D..J]$, it is

noticeable that it always increases with the level of abstractivity, which sustains the obtained correlation value in Figure 7.1.

All these observations corfirm our hypotheses. The level of presence of Jing (2002)'s actions are related to the level of abstractivity preception, and, that exists a single continuous dimension where the two summarization styles, extractive and abstractive, could coexist.

7.4 Abstractivity Characterization

Based on the conclusions of Section 7.3, we identify two ways, or tasks, for describing summaries in relation to extractivity/abstractivity.

The first task is the **the Abstractivity Level (AL) prediction**. This task aims to identify the deegree of perception of paraphrasing of the main information of a text included in the summary. In this work, our scale is the same as in question **(C)** (excluding answer 0). Therefore, level 1 indicates an extractive summarization style and level 5 indicates a strong perception that the summary's author has created it with "their own words".

The second task is to measure how often each Jing (2002)'s action was used for re-writing and synthesizing the main information from a text has been used to include the information in the summary. We named this task as **Abstractivity Inducting Features (AIFs) extraction**, and it should predict the answers for the 7 questions from **(D)** to **(J)** (answer 0 excluded).

7.5 Experimentation

In this section, we detail the experimentation done over the *CLAsum*^{s} subset. The experimental work has a double objective; on the one hand, we want to validate our perception that extractivity and complete abstractivity are extreme points of a single dimension with multiple abstractivity levels, and on the other hand, we want to verify the role of AIFs in the characterization of these levels of abstractivity. Since we had 3 set of labels per sample and we wanted to capture the diversity withing each document-summary pair, both task where approached as regression tasks.

7.5.1 Extraction Features and Supervised Machine Learning Methods

The extraction features process was the same that was used for the clustering process described in Section 7.2.1. The document-summary pairs in the samples were repre-

sented by a vector of 8 abstractivity-related indicators. For the sake of clarity, we will refer to them as *abs-metrics-features*.

To tackle the Abstractivity Inducting Features extraction and Abstractivity Level prediction tasks, we selected a wide range of supervised machine learning methods that can accept the choosen freatures. All of them were used with the implementation available in the Scikit-Learn (Pedregosa et al., 2011) python module. The methods that were considered are the following: *Linear Regression*, *Linear SVM*, *SVM with RFG kernel*, *Random Forest*, and *Multi-Layer Perceptron*.

Since some of the methods can not handle more than one feature in the output, we circumvented this handicap by training one model for each feature in the case of AIFs extraction tasks.

7.5.2 Evaluation Metrics

For measuring the performance of the regression tasks, we used *Root Squared Mean Error* (RMSE) and *Median Absolute Error* (MdAE). Additionally, the *Relative* distance (Equation (7.1), Section 7.2.5) was used in the Abstractivity Level prediction task, and the *Minkowski* (p=7) distance for the AIFs extraction task. To evaluate the predictions with *Minkowski* distance, we measured the distance between the AIFs extraction vector and the reference vector since we want to evaluate the extracted features' cohesion. The *Minkowski* distance between vectors was measured against the normalized AIFs vectors. The normalized AIFs vectors with values from $[0, 1]$ were obtained by dividing each aspect by the maximum value possible for that aspect.

7.5.3 Types of Architectures

For the AIFs extraction task, we used an architecture where systems receive the *abs-metrics-features* (Section 7.5.1) and return the AIFs vectors of the samples.

In the case of the Abstractivity Level prediction task, we used three different architectures:

(1) ***End-to-End*:** types of systems that receive the *abs-metrics-features* and returns the AL prediction for the samples.

(2) ***AIFs-to-AL*:** these types of systems receive vectors with the *abs-metrics-features* concatenated with the AIFs and predict the AL. We used the reference AIFs labels as input to train these systems, along with the *abs-metrics-features*.

(3) ***Pipeline*:** in this architecture the *abs-metrics-features* are received and it perform the AL prediction in two steps. In the first step, a system for AIFs extraction task predicts the AIFs based solely on the *abs-metrics-features*. In the second step,

abs-metrics-features and extracted AIFs are concatenated and passed to an *AIFs-to-AL* system to perform the AL prediction task.

The *Pipeline* architecture should help verify the usefulness of the AIFs' information to perform the AL prediction task compared with the one that does not used them (*End-to-End* architecture).

7.5.4 Training and Evaluation Methodology

Since the *CLAsum*^{s} subset contains only 175 document-summary unique pairs, we trained and evaluated all system configurations 20 times with different partitions, which will show the variability in the performance of each configuration and the conclusions extracted from the results would not be tied to any random aspect of the validation process.

Considering that the distribution of the classes regarding the abstractivity level is not well-balanced, we did not use the K-Fold methodology. Instead, we split the dataset with a different random state (seed) each time. In each partition, 20% of the document-summary pairs were put aside for testing (105 samples from 3 sets of labels of 35 pairs) and the rest for training (520 samples). The partitions were created with the `train_test_split` from Scikit-Learn, setting the seed with an integer number from 0 to 19 and stratified the dataset by document-summary pairs taking into account the median of the 3 answers for question **(C)**. Therefore, all the samples for a certain document-summary pair were always in the same partition. It should be also mentioned that we verified that all train partitions contain all the possible labels/answers for each question and that 99.4% of the samples were used for testing at least once.

For the configuration of each Supervised Machine Learning Method, we bypass modifying the default parameters of the Scikit-Learn implementation (version 1.5.0) to avoid introducing more variables in the study. Only the random state was set to 42 when the method had this feature and set max steps to 1 000 000 (limit never reached).

7.6 Systems' Results

This section presents the results for both tasks, AIFs extraction and Abstractivity Level prediction.

7.6.1 Abstractivity Inducting Features Extraction Task

The best results for the AIFs extraction task are for the MLP system that obtained $(M{=}0.39, SD{=}0.01)$ in the Mikowski distance (*M-Dist*), $(M{=}0.76, SD{=}0.02)$ in the *RSME* metric

Table 7.1: Results of the best system per architecture for Abstractivity Level prediction task in *CLAsum*^{s} subset.

Arch.	Mthd	R-Dist ↓	RMSE ↓	MdAE ↓
E	lSVM	$0.149^{0.142}_{0.155}$	$0.96^{0.93}_{1.00}$	$\mathbf{0.57}^{0.54}_{0.61}$
P	SVM+lSVM	$\mathbf{0.147}^{0.141}_{0.154}$	$\mathbf{0.95}^{0.92}_{0.98}$	$0.59^{0.56}_{0.62}$
A	lSVM	$0.136^{0.131}_{0.141}$	$0.88^{0.85}_{0.91}$	$0.56^{0.53}_{0.59}$

and (M=0.57,SD=0.02) in the Median Average Error (MdAE). The values in *M-Dist* indicate that AIFs extracted by systems are nearly 40% away from the reference. Results in *RSME* and *MdAE* confirm that systems extract AIFs that, component to component, tend to differ around ±0.5 from the correct answer. These observations indicate that the *abs-metrics-features* vectors are not informative enough for extracting AIFs. We hypothesize that AIFs contain additional information concerning the abstractivity. All system results can be found in Table 7.D.1 (Section 7.7).

7.6.2 Abstractivity Level Prediction Task

In this section, we discuss the results for each architecture for the Abstractivity Level prediction task.

Table 7.1 details the results for Abstractivity Level regression task obtained by the best system for each architecture: *End-to-End* (E), *Pipeline* (P), and *AIFs-to-AL* (A). Numbers in bold are the best average values in their columns, excluding *A* since these systems received the reference AIFs labels. Each cell shows the average value and the 95% confidence interval (exponent = lower bound, subscript = upper bound). The results of all considered systems are detailed in Table 7.E.1 (Section 7.7).

We use the two-sample t-student significant test over the *R-Dist* metric to verify whether there are significant differences between them. The best *End-to-End* (E) system that takes as input the *abs-metrics-features* obatains an average performance of 0.148 evaluated with *R-Dist*, which indicates that the level of abstractivity predicted contains an of error ±0.59 on a scale from 1 to 5. The introduction of the AIFs produce that the best *Pipeline* (P) system slightly improves the performance over the E system, $t(19)=-2.38, p<0.05$, with an average of 0.147 in *R-Dist*. However, when we consider the best *AIFs-to-AL*, that receives the reference AIFs values, the improvement over the system E is significant, $t(19)=-6.07, p<0.01$; averaging 0.136 in *R-Dist*. Therefore, we can assure that the AIFs' information helps to improve the prediction of the Abstractivity Level, though we should dispose of more reliable AIFs extraction systems. When we compare the P system with the A system, it can be concluded that the accumulated error of the AIFs extractor system used in the P architecture had a significant impact,

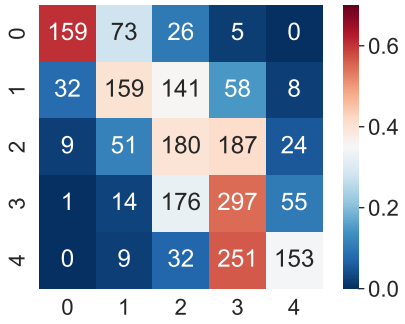


Figure 7.1: Confusion matrix of the best system A in the *CLAsum*^{s} subset.

$t(19)=5.52, p<0.01$, in reducing the effectiveness of the AIFs in the system. Note that system A is an upper bound of P systems, since A systems need AIFs in addition to *abs-metrics-features*; hence, they can not be used independently.

We calculated the confusion matrix (CM) to better understand how the system A is performing in the AL prediction task. The CMs were based on 20 runs and 105 test samples per run (2100 samples in total). The number in the `y-axis` is the reference label, and the one in the `x-axis` is the predicted. Numbers in cells indicate the number of samples in each combination. The color scale is related with the percentage of true positives per row. Since the systems predict the AL as a real number instead of an integer, the prediction have been rounded to compute the confusion matrix. The CM of the best system A is shown in Figure 7.1.

Figure 7.1 shows that the system A is capable of predicting the correct level of abstractivity of nearly half of the samples (948). Additionally, most of the samples not correctly labeled are found close to the diagonal at ±1 level of abstractivity (966). Therefore, the system A predicts 91% of samples the correct level of abstractivity or predicts one closer to the human perception. The fact that the system can approximate close enough the level of abstractivity of the given summaries based on the *abs-metrics-features* and AIFs information, supports our idea that there is a single continuous dimension where extractivity (level 1 of abstractivity) and complete abstractivity (level 5 of abstractivity) summarization styles coexist.

7.7 Conclusions

In this work, we have presented and made available to the scientific community the *CLAsum* dataset. This is a hand-annotated dataset that allows characterizing the complexity of the process of summarizing a document by measuring the Abstractivity Level and seven Abstractivity Inducting Features.

The results from the study of the dataset and the experimental work show how the Abstractivity Level and AIFs are related and how AIFs are useful when measuring the level of abstractivity of a summary. Our study places extractivity and complete abstractivity as the extreme points of a single dimension with multiple levels.

While this work has raised an initial frame for delving into the abstractivity aspect of summaries, several opportunities remain for future work. On the one hand, we want to extend the dataset in quantity and diversity to other different domains with the aim of allowing more generalizable studies. On the other hand, we want to study the capability of deep neural models to extract AIFs, predict abstractivity levels, and generate summaries with different abstractivity levels.

Chapter 7. References

Ahuir, Vicent, José-Ángel González, et al. (2024). "Abstractive Summarizers Become Emotional on News Summarization". In: *Applied Sciences* 14.2. ISSN: 2076-3417. DOI: 10.3390/app14020713. URL: https://www.mdpi.com/2076-3417/14/2/713 (cit. on p. 139).

Ahuir, Vicent, Lluís-F. Hurtado, et al. (2021). "NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish". In: *Applied Sciences* 11.21. ISSN: 2076-3417. DOI: 10.3390/app11219872. URL: https://www.mdpi.com/2076-3417/11/21/9872 (cit. on pp. 138, 140).

Ahuir, Vicent, Encarna Segarra, and Lluís-F. Hurtado (Nov. 2024). "Beyond Using Their Own Words: Abstractivity Characterization in Summarization". Work pending of publication (cit. on p. 135).

Alami Merrouni, Zakariae, Bouchra Frikh, and Brahim Ouhbi (Oct. 2023). "EXABSUM: a new text summarization approach for generating extractive and abstractive summaries". In: *Journal of Big Data* 10.1, p. 163. ISSN: 2196-1115. DOI: 10.1186/s40537-023-00836-y. URL: https://doi.org/10.1186/s40537-023-00836-y (cit. on p. 139).

Bommasani, Rishi and Claire Cardie (Nov. 2020). "Intrinsic Evaluation of Summarization Datasets". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8075–8096. DOI: 10.18653/v1/2020.emnlp-main.649. URL: https://aclanthology.org/2020.emnlp-main.649 (cit. on pp. 138, 140).

Chen, Xiuying et al. (Aug. 2021). "Capturing Relations between Scientific Papers: An Abstractive Model for Related Work Section Generation". 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)*. Online: Association for Computational Linguistics, pp. 6068–6077. DOI: 10.18653/v1/2021.acl-long.473. URL: https://aclanthology.org/2021.acl-long.473 (cit. on p. 138).

Dou, Zi-Yi et al. (June 2021). "GSum: A General Framework for Guided Neural Abstractive Summarization". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech-*

nologies. Online: Association for Computational Linguistics, pp. 4830–4842. DOI: 10.18653/v1/2021.naacl-main.384. URL: https://aclanthology.org/2021.naacl-main.384 (cit. on p. 138).

Fu, Xiyan et al. (Aug. 2021). "RepSum: Unsupervised Dialogue Summarization based on Replacement Strategy". 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)*. Online: Association for Computational Linguistics, pp. 6042–6051. DOI: 10.18653/v1/2021.acl-long.471. URL: https://aclanthology.org/2021.acl-long.471 (cit. on p. 138).

Giarelis, Nikolaos, Charalampos Mastrokostas, and Nikos Karacapilidis (2023). "Abstractive vs. Extractive Summarization: An Experimental Review". In: *Applied Sciences* 13.13. ISSN: 2076-3417. DOI: 10.3390/app13137620. URL: https://www.mdpi.com/2076-3417/13/13/7620 (cit. on p. 139).

Grusky, Max, Mor Naaman, and Yoav Artzi (June 2018). "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 708–719. DOI: 10.18653/v1/N18-1065. URL: https://aclanthology.org/N18-1065 (cit. on pp. 138, 140).

Guo, Yanzhu et al. (Dec. 2022). "Questioning the Validity of Summarization Datasets and Improving Their Factual Consistency". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5716–5727. DOI: 10.18653/v1/2022.emnlp-main.386. URL: https://aclanthology.org/2022.emnlp-main.386 (cit. on p. 140).

Jing, Hongyan (Dec. 2002). "Using Hidden Markov Modeling to Decompose Human-Written Summaries". In: *Computational Linguistics* 28.4, pp. 527–543. ISSN: 0891-2017. DOI: 10.1162/089120102762671972. URL: https://doi.org/10.1162/089120102762671972 (cit. on pp. 138, 139, 141, 145–147).

Kabeer, Adebayo and Samiya Khan (2024). "Refining Abstractive Text Summarization: Towards an Optimizing Approach based on PTLM Selection and Fine-Tuning". In: *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–6. DOI: 10.1109/CONECCT62155.2024.10677220 (cit. on p. 139).

Kryściński, Wojciech et al. (Oct. 2018). "Improving Abstraction in Text Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1808–1817. DOI: 10.18653/v1/D18-1207. URL: https://aclanthology.org/D18-1207 (cit. on pp. 138, 140).

Liu, Yang and Mirella Lapata (Nov. 2019). "Text Summarization with Pretrained Encoders". 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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740. DOI: 10.18653/v1/D19-1387. URL: https://aclanthology.org/D19-1387 (cit. on p. 138).

Manakul, Potsawee and Mark Gales (Aug. 2021). "Long-Span Summarization via Local Attention and Content Selection". 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)*. Online: Association for Computational Linguistics, pp. 6026–6041. DOI: 10.18653/v1/2021.acl-long.470. URL: https://aclanthology.org/2021.acl-long.470 (cit. on p. 138).

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (Oct. 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: Association for Computational Linguistics, pp. 1797–1807. DOI: 10.18653/v1/D18-1206. URL: https://aclanthology.org/D18-1206 (cit. on p. 139).

Nenkova, Ani and Kathleen McKeown (June 2011). "Automatic Summarization". In: *Foundations and Trends® in Information Retrieval* 5.2–3, pp. 103–233. ISSN: 1554-0669. DOI: 10.1561/1500000015. URL: http://dx.doi.org/10.1561/1500000015 (cit. on p. 138).

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830 (cit. on p. 148).

See, Abigail, Peter J. Liu, and Christopher D. Manning (July 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)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. URL: https://aclanthology.org/P17-1099 (cit. on p. 139).

Sirokov, Roman (2024). *pywebview: A lightweight cross-platform library to create web-based desktop GUIs*. Version 3.6. URL: https://github.com/r0x0r/pywebview (cit. on p. 161).

Sun, Weisong et al. (Mar. 2024). "An Extractive-and-Abstractive Framework for Source Code Summarization". In: *ACM Trans. Softw. Eng. Methodol.* 33.3. ISSN: 1049-331X. DOI: 10.1145/3632742. URL: https://doi.org/10.1145/3632742 (cit. on p. 138).

Varab, Daniel and Yumo Xu (July 2023). "Abstractive Summarizers are Excellent Extractive Summarizers". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan

Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 330–339. DOI: 10.18653/v1/2023.acl-short.29. URL: https://aclanthology.org/2023.acl-short.29 (cit. on p. 138).

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., pp. 5998–6008. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 138).

Wu, Wenhao et al. (Aug. 2021). "BASS: Boosting Abstractive Summarization with Unified Semantic Graph". 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)*. Online: Association for Computational Linguistics, pp. 6052–6067. DOI: 10.18653/v1/2021.acl-long.472. URL: https://aclanthology.org/2021.acl-long.472 (cit. on p. 138).

Zheng, Chujie et al. (2020). "Topic-Guided Abstractive Text Summarization: a Joint Learning Approach". In: *arXiv preprint arXiv:2010.10323* (cit. on p. 138).

Zou, Yanyan et al. (Nov. 2020). "Pre-training for Abstractive Document Summarization by Reinstating Source Text". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 3646–3660. DOI: 10.18653/v1/2020.emnlp-main.297. URL: https://aclanthology.org/2020.emnlp-main.297 (cit. on p. 138).

Chapter 7. Appendix

7.A Labeling Guideline

In this section, the completed guideline that was used in the labeling process is presented.

Given a newspaper article and a summary in the left side (Document/Summary tab), answer 10 questions/statements regarding the content of the article and the summary and/or the way the summary was created. The possible answers are detailed on the left side (Questions tab).

(A) *The summary provides the most relevant information about the article, and the article extends it with additional details*:

0: Strongly Disagree.

1: Disagree.

2: Undecided.

3: Agree.

4: Strongly Agree.

(B) *Regarding information contained in the summary*:

1: All the information in the summary can be found in the article (not necessarily in the exact words).

2: Almost all the information in the summary can be found in the article, but adds some additional information.

3: I can not consider the given summary a truly abstract. All the information provided in the summary, it is additional and can not be extracted or inferred from the article.

(C) *What is your perception about how the author of the summary wrote it?*:

- 0: Does not apply; it is not a summary.
- 1: They rely entirely on the article. It is as if I was reading complete sentences highlighted in the article.
- 2: They rely heavily on the article to write the summary. It only presents slight changes in form and/or order concerning the article.
- 3: They mainly rely on the article to write the summary. Segments of the summary can be identified in the article. Still, the author alters the article's text in form and/or order.
- 4: They weakly rely on the article to write the summary and alter a lot of the article in form and/or order.
- 5: Overall, they do not rely on the article to write the summary; instead, they explain the main ideas of the article in their own words.

(D) *How does the author handle non-relevant information in the article?*:

- 0: Does not apply; it is not a summary.
- 1: They discard complete sentences. No segments or words of a sentence are discarded.
- 2: They focus on mainly discarding complete sentences. Segments or words of sentence discarding is also present, but it is less often than complete sentences discarding.
- 3: They focus mainly on discarding text segments within the sentences of the article. The complete sentence discarding is absent, or it is noticeably less frequent than segment.
- 4: All information is considered relevant; they manage to cover all the information in the article and substantially reduce its length. discarding.

(E) *For the creation of the summary, part of the information selected from the sentences of the article is combined to form the sentences of the summary*:

- 0: Does not apply; it is not a summary.
- 1: No sentences from the article are combined. Each sentence in the summary corresponds to the information contained by a sentence in the article.
- 2: Some sentences in the summary are created by combining the information contained by certain sentences from the article.
- 3: Most of the sentences of the summary are created by combining information from some sentences of the article. discarding.

(F) *Sentences in the article that contain the information reflected in the summary have been syntactically altered for inclusion in the summary*:

0: Does not apply; it is not a summary.

1: No syntactic alterations exist to create the summary.

2: There are some syntactic alterations to create the summary.

3: There are many syntactic alterations to create the summary.

(G) *When including sentences or segments of the article in the summary, the author replaces words or expressions with semantically equivalent ones*:

0: Does not apply; it is not a summary.

1: Never.

2: Sometimes.

3: Quite often.

(H) *The summary includes generalizations of information extracted from the article. A generalization is describing one or more concepts using a less specific word (e.g., "Matthew and Amanda reappear in the new sequel of the acclaimed fiction movies of galactic adventures series" in the summary "Matthew and Amanda" could be grouped as "The main actors . . . ")*:

0: Does not apply; it is not a summary.

1: No information can be considered susceptible to generalization without a significant loss of information.

2: No information susceptible to generalization was generalized.

3: Less than half of the information susceptible to generalization was generalized; the rest was not generalized.

4: More than half of the information susceptible to generalization was generalized.

(I) *The summary includes specifications of information extracted from the article. A specification would be to use expressions or words that make the information more specific (e.g., "The race driver has won his ninth F1 World Championship Grand Prix" in the summary "The race driver" could be detailed as "The F1 driver . . . ")*:

0: Does not apply; it is not a summary.

1: No information can be considered susceptible to specification.

2: No information susceptible to specification was specified.

3: At most, half of the information susceptible to specification was specified; the rest was not specified.

4: More than half of the information susceptible to specification was specified.

(J) *The author of the summary rearranges the chosen information. For example, if facts A-B-C appear in the article, the author refers to them in the following order B-A-C in the summary*:

0: Does not apply; it is not a summary.

1: Never.

2: On one occasion.

3: On several occasions.

7.B YALT!: Yet Another Labeling Tool!

Figure 7.B.1 presents the labeling application developed for the labeling process called *YALT! (Yet Another Labeling Tool!)*. The application was developed with Python 3 and PyWebview Sirokov, 2024, a framework for developing GUI applications with HTML and CSS. The application would be capable of handling different labeling text tasks by just developing an HTML web page for the task needings (supports HTML with CSS and JavaScript). YALT! is available as a Python module.

For the labeling task of this work, we split the labeling window into two parts. On the left side, the annotator could see the guidelines in English and Spanish ("Questions" and "Preguntas" tabs) and the Document and Summary to work with. On the right side, the annotator had the 10 questions to answer. Additionally, and to facilitate the labeling process, the application presented the exact Common Long Sequences between the document and the summary in different colors.

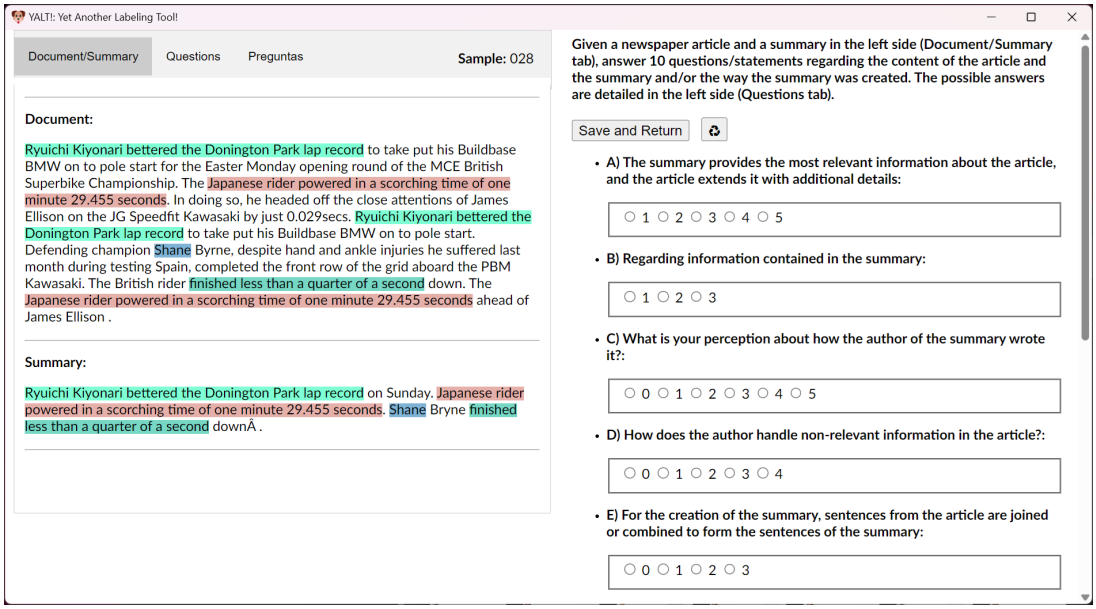


Figure 7.B.1: Labeling window of a sample in the YALT! application.

7.C Average Pair-wise Annotator Agreement

Figure 7.C.1 shows the average Cohen's Kappa agreement between two given annotators. White spaces are combinations that did not occur in the labeling process. The agreement is measured with the *Relative* distance (Equation (7.1), Section 7.2.5) between two annotators and the 10 questions at once.

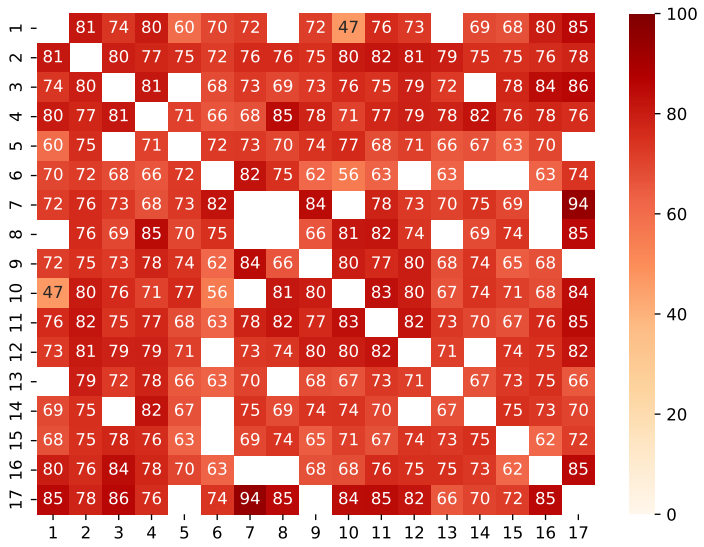


Figure 7.C.1: Average of Cohen's Kappa pair-wise agreement score (*Relative* distance).

7.D Results for Abstractivity Inducting Features task

Table 7.D.1 details the results obtained by all models for the Abstractivity Inducting Features extraction task in the *CLAsum*^{s} subset. System configurations are sorted in ascending order by the *M-Dist* column.

Table 7.D.1: Results systems for Abstractivity Inducting Features (AIFs) extraction task in the *CLAsum*^{s} subset. The supervised machine learning methods (Mthd) are: MLP (Multi Layer Perceptron), LiR (Linear Regression), lSVM (Linear SVM), RnF (Random Forest), and SVM.

Mthd	M-Dist ↓	RMSE ↓	MdAE ↓
MLP	**0.389** $^{0.384}_{0.394}$	**0.76** $^{0.75}_{0.77}$	**0.57** $^{0.56}_{0.58}$
LiR	0.390 $^{0.386}_{0.395}$	0.77 $^{0.75}_{0.78}$	0.59 $^{0.58}_{0.60}$
lSVM	0.406 $^{0.401}_{0.410}$	0.78 $^{0.77}_{0.79}$	0.59 $^{0.57}_{0.60}$
RnF	0.408 $^{0.403}_{0.413}$	0.79 $^{0.78}_{0.80}$	0.58 $^{0.56}_{0.60}$
SVM	0.409 $^{0.403}_{0.415}$	0.78 $^{0.77}_{0.79}$	0.55 $^{0.53}_{0.56}$

7.E Results for Abstractivity Level task

Table 7.E.1 details the results for Abstractivity Level prediction task in *CLAsum*^{s} subset. All the systems with architectures *AIFs-to-AL* (A) and *End-to-End* (E) are listed. In the case of systems with *Pipeline* (P) architecture, only the top-10 systems with best performance are listed. Numbers in bold are the best average values in their columns, excluding *A* since these systems received the reference AIFs labels. The table is sorted ascending by R-Dist (*Relative Distance*) column.

Table 7.E.1: Results systems for Abstractivity Level prediction task in the *CLAsum*^{s} subset. The supervised machine learning methods (Mthd) are: MLP (Multi Layer Perceptron), LiR (Linear Regression), lSVM (Linear SVM), RnF (Random Forest), and SVM. For systems with *Pipeline* architecture, the Mthd format is `Sys1+Sys2`; indicating that `Sys1` is the system used for extracting the AIFs and `Sys2` the one used for predicting the Level of Abstractivity.

Arch.	Mthd	R-Dist ↓	RMSE ↓	MdAE ↓
A	lSVM	$0.136^{0.131}_{0.141}$	$0.88^{0.85}_{0.91}$	$0.56^{0.53}_{0.59}$
A	SVM	$0.138^{0.133}_{0.142}$	$0.89^{0.86}_{0.91}$	$0.56^{0.54}_{0.59}$
A	MLP	$0.138^{0.134}_{0.143}$	$0.88^{0.86}_{0.91}$	$0.56^{0.54}_{0.61}$
A	LiR	$0.140^{0.134}_{0.146}$	$0.89^{0.85}_{0.93}$	$0.60^{0.56}_{0.63}$
A	RnF	$0.143^{0.138}_{0.149}$	$0.90^{0.87}_{0.93}$	$0.60^{0.56}_{0.65}$
P	SVM+lSVM	$\mathbf{0.147}^{0.141}_{0.154}$	$0.95^{0.92}_{0.98}$	$0.59^{0.56}_{0.62}$
E	lSVM	$0.149^{0.142}_{0.155}$	$0.96^{0.93}_{1.00}$	$\mathbf{0.56}^{0.54}_{0.61}$
P	lSVM+lSVM	$0.149^{0.143}_{0.155}$	$0.95^{0.92}_{0.98}$	$0.60^{0.56}_{0.63}$
P	MLP+lSVM	$0.150^{0.144}_{0.155}$	$\mathbf{0.94}^{0.91}_{0.97}$	$0.64^{0.61}_{0.67}$
P	LiR+lSVM	$0.150^{0.144}_{0.155}$	$\mathbf{0.94}^{0.91}_{0.97}$	$0.63^{0.60}_{0.66}$
P	lSVM+SVM	$0.150^{0.144}_{0.156}$	$0.98^{0.94}_{1.01}$	$\mathbf{0.56}^{0.55}_{0.60}$
P	LiR+SVM	$0.150^{0.144}_{0.156}$	$0.97^{0.93}_{1.00}$	$0.59^{0.56}_{0.62}$
P	RnF+SVM	$0.151^{0.145}_{0.157}$	$0.97^{0.94}_{1.00}$	$0.62^{0.59}_{0.65}$
P	SVM+SVM	$0.151^{0.145}_{0.157}$	$0.98^{0.95}_{1.02}$	$0.59^{0.56}_{0.62}$
P	RnF+lSVM	$0.151^{0.145}_{0.156}$	$0.95^{0.92}_{0.98}$	$0.66^{0.63}_{0.69}$
P	MLP+SVM	$0.151^{0.145}_{0.157}$	$0.97^{0.93}_{1.00}$	$0.59^{0.56}_{0.62}$
E	SVM	$0.151^{0.146}_{0.157}$	$0.97^{0.94}_{1.01}$	$0.61^{0.59}_{0.64}$
E	MLP	$0.153^{0.148}_{0.158}$	$0.94^{0.92}_{0.97}$	$0.66^{0.64}_{0.69}$
E	LiR	$0.154^{0.148}_{0.160}$	$0.96^{0.91}_{1.00}$	$0.65^{0.61}_{0.70}$
E	RnF	$0.156^{0.149}_{0.160}$	$0.96^{0.93}_{0.99}$	$0.65^{0.62}_{0.68}$

Chapter 8

Discussions

Through all the chapters of this thesis, from 2 to 7, we developed an ensemble of ideas related to the NLP task of automatic summarization of documents. The development of these ideas did not aim to find solutions that develop better in specific scenarios; although some of them excelled in that aspect. We aim to develop concepts and ideas that will help propose more flexible, controllable, general, and less data-driven automatic summarization solutions in the future. Specifically, we focused on the following aspects related to summarization: (i) contribute with new summarization resources to the scientific community; (ii) increase the transference from the pre-training phase of ground-skills related to summarization; (iii) expand the definition and characterization of the presence of abstractivity in summaries; (iv) prove the presence of a single summarization strategy that encompasses both strategies, extractive and abstractive, and can obtain summaries that are placed along the dimension that contains the two strategies as edges; (v) identify ways to infer the quality of summaries without a gold reference; (vi) explore summarizing specific-purpose-driven decisions such as the selection emotional-arising words to influence readers attention.

In Chapter 2 we have detailed the DACSA dataset, a large-scale dataset for automatic summarization for Catalan and Spanish. The dataset contains more than 2 million document-summary pairs for Spanish, which, at the moment, is the most extensive public collection of samples for summarization for the Spanish language. Even more important, DACSA also contains more than 700 thousand document-summary pairs for Catalan, a low-resource language that is only sometimes considered in sci-

entific works and resource creation. In the case of Catalan, at this point, it is also the most extensive public collection of samples for the summarization task for this language. Another relevant aspect of DACSA is that the dataset was created with abstractivity summarization in mind by filtering summaries that were considered too extractive. When we studied the presence of abstractivity in the reference summaries, we could observe that the presence of abstractivity was noticeable in summaries from both languages. Therefore, DACSA is a highly relevant resource for the automatic summarization task in Catalan and Spanish, especially taking into account that all the texts were captured prior to the disposal of Large Language Models (LLMs)-based generative tools, such as ChatGPT (OpenAI, 2024) or Gemini (Anil et al., 2023), to the mass audience; making the dataset even more valuable.

In Chapter 3, we have presented a work where we focused on changing the behavior of the automatic summarization models through the pre-training phase and also studying them with the presence of abstractivity in their generated summaries. In this work, we designed a pre-train methodology that combined four self-supervised tasks to increase the transfer knowledge for summarization from the pre-training phase. This combination of tasks was designed to achieve better performance in the downstream task and increase the presence of abstractivity in the generated summaries without changing the pre-training data. From the work, a set of summarization models for Catalan and Spanish languages were developed, three per language. The summarization models for Catalan and Spanish were obtained by fine-tuning them with the DACSA dataset (Chapter 2); two were based on the pre-training methodology, and the rest were not. The models based on the pre-training methodology showed better performance and generalization capabilities, especially the Catalan one. To study the presence of abstractivity in the generated summaries, we measure it by combining a set of currently available abstractive-related metrics with a new one proposed in this work (*Content Reordering*). With this methodology, we could verify that the summarization models based on the designed pre-trained method showed significantly higher abstractivity in their generated summaries than the others.

In Chapter 4 we have presented our participation in a shared task related to summarizing radiology reports (Delbrouck et al., 2023) at the BioNLP 2023. In this work, we tested the pre-training methodology proposed in Chapter 3 and adapted a general language model to the biomedical domain based on BART (Lewis et al., 2020). Because the results obtained by our summarization models needed to be more competitive, we conducted a late study to raise the possible problem sources that had reduced the performance of our solution. From that study, we concluded that our models could condense the most relevant information at the beginning of the summaries, implying that the pre-training methodology helped in the downstream task. However, the

models did not deal appropriately with the long-tail length distribution of summaries, substantially reducing the end performance. From this work, it can be verified that the proposed pre-trained method was beneficial in the journalistic domain and the biomedical domain, potentially being generally applicable. However, further studies should be conducted to verify whether it applies to any domain.

In Chapter 5, we have introduced our participation in the shared task related to lay summarization of biomedical research articles (Goldsack et al., 2024) at the BioNLP 2024. In this work, we used the same approach as in Chapter 4; we continuously pre-trained a language model to adapt it to the biomedical research articles' specific domain. To tackle the summarization of scientific articles, we adapted an LED Longformer model since the Longformer enabled the possibility to handle longer sequences (Beltagy, Peters, and Cohan, 2020). Even though we used a Longformer model, because we were dealing with very long texts, the model could not accept the whole document. For that reason, the work contains details about the impact of including certain document sections in the performance in the downstream task.

Additionally, we explored the possibility of boosting the performance of the summarization models by predicting the quality of the generated summaries without reference summaries. To achieve this, we designed a regression model that would predict the quality of a summary based on the article and the generated summary. To obtain the regression model, we created a dataset based on data augmentation with prompting and a set of LLMs to generate summaries with different quality degrees. This regression model was used as a ranking model to select summaries more likely to obtain higher scores in the competition metrics. Our system that employed a summarization and ranking model in the competition showed a fair, competitive performance, especially in choosing the relevant information.

In Chapter 6, we have analyzed utterly different aspect of summaries, the emotions that entails a document or a summary from the words chosen by the author of the texts. The emotions of a text are subjacent to the words and expressions used to create it; two texts can be semantically equivalent and transmit different emotions to the reader. Specifically, we explored emotional words (words that entail emotions to the reader) in two journalistic datasets. It could be verified that articles often contain emotional words that entail some emotions rather than others; this was also observed in summaries. However, in summaries, emotional words were more frequent than in the news articles that they synthesized. Moreover, it was noted that certain emotions were amplified in summaries since there are certain emotions –such as fear– that invite the reader to continue reading the complete article more than others. Therefore, we identified a clear emotional bias in those journalistic datasets.

Considering the study of the emotional bias in those datasets, we studied the behavior of the summarization models fine-tuned with those datasets. We observed that the models acquired a similar bias when they generated the summaries; especially in the negative ones. Since we worked with abstractive summarization models, we also analyzed the emotional novel words; that is, words that were in the summary and not in the document and entail any emotion. It was noted that models also presented an emotional bias when they chose new words to include in the summary. Therefore, we concluded that models inhered a similar emotional bias from existing in the datasets. Thus, this work raised new aspects to control in summarization rather than generate unbiased emotional summaries or summaries with the emotional bias established by the user needs.

Finally, in Chapter 7, we have explored the abstractivity perception in summaries: what makes a summary more or less "*writtern in the author's words*"? We aimed to better characterize the presence of abstractivity in summaries. With a better characterization, we also aimed to verify if there are pieces of evidence indicating that extractivity and abstractivity can be included in a single dimension with both pure strategies in the edges.

We constructed a dataset for the study of this aspect called *CLAsum*, a highly diverse dataset with 600 samples collected from 17 annotators with 11 labels per sample. Even though the *CLAsum* dataset was created with a large group of people and the annotation process was complex, it showed an annotator agreement reasonably high, indicating that the labeling process produced a dataset that captured subjectivity but retained enough agreement to consider the data coherent and valid, from where helpful information could be extracted. From the statistical study of the dataset, we extracted that the frequency of usages of the summarization actions identified by Jing (2002) was related to the degree of perception in humans of abstractivity in summaries.

With the *CLAsum* dataset, we also carried out a study with machine learning methods to approach two abstractivity-related tasks: (1) predict the level of abstrarctivity and (2) predict the degree of appearance of each summarization action. As input features for the machine learning methods, we used a set of abstractivity-related metrics as in Chapter 3. The results of the study confirmed the following: (i) the information contained by the degree of appearance of the summarizations actions was complementary to the existing abstractivity-related metrics, and this information was helpful to predict the level of abstractivity, (ii) the models could distinguish between summaries with different level of abstractivity with a reasonable precision, which indicates, (iii) there are pieces of evidence of a presence of a single dimension with degrees where the pure extractive and abstractivite strategies coexist as the edges of

that dimension.

Chapter 8. References

Anil, Rohan et al. (2023). "Gemini: A Family of Highly Capable Multimodal Models". In: *arXiv preprint arXiv:2312.11805*. URL: https://arxiv.org/abs/2312.11805 (cit. on p. 166).

Beltagy, Iz, Matthew E. Peters, and Arman Cohan (2020). "Longformer: The Long-Document Transformer". In: *arXiv:2004.05150* (cit. on p. 167).

Delbrouck, Jean-Benoit et al. (July 2023). "Overview of the RadSum23 Shared Task on Multi-modal and Multi-anatomical Radiology Report Summarization". In: *Proceedings of the 22st Workshop on Biomedical Language Processing*. Toronto, Canada: Association for Computational Linguistics (cit. on p. 166).

Goldsack, Tomas et al. (Aug. 2024). "Overview of the BioLaySumm 2024 Shared Task on the Lay Summarization of Biomedical Research Articles". In: *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Bangkok, Thailand: Association for Computational Linguistics (cit. on p. 167).

Jing, Hongyan (Dec. 2002). "Using Hidden Markov Modeling to Decompose Human-Written Summaries". In: *Computational Linguistics* 28.4, pp. 527–543. ISSN: 0891-2017. DOI: 10.1162/089120102762671972. URL: https://doi.org/10.1162/089120102762671972 (cit. on p. 168).

Lewis, Mike et al. (July 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*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703 (cit. on p. 166).

OpenAI (2024). *ChatGPT: OpenAI Language Model*. https://chat.openai.com/. Available at https://chat.openai.com/ (cit. on p. 166).

Chapter 9

Conclusions and Future Work

Throughout this thesis, we focused on unraveling the process of writing the summary, leaving, most of the time, aside from detecting the relevancy of the information. In that way, we aimed to provide tools to create more flexible automatic summarization systems capable of adapting their writing process to the user's needs.

We verified that we could vary the behavior of the summarization models by modifying the learning process of the models in the pre-training phase to make the process less data-driven. The resulting summarization models varied their behavior by altering the learning process in the initial phase, even though the datasets were always the same. We proved that the summarization style concerning the perception of abstractivity could vary if adequately inducted. To study the models regarding perception abstractivity, we created a preliminary characterization of that aspect by combining existing abstractivity-realted metrics and a new one.

We validated that the goal of the writer of the summary could be hidden behind the word selection, a subjacent aspect such as the emotional entailment to the person that will read the summary. Choosing one word over another could vary the emotional message the reader would perceive after reading the summary, even though they were almost synonyms. It was verified that emotional words have a significant role in creating articles and summaries in the journalistic field. Moreover, certain emotions were amplified in summaries to catch the attention of the reader and encourage them to continue reading the article. When the resulting summarization models trained with that journalistic data were analyzed, they also showed the same emotional bias in the datasets. The summarization models amplified the same emotions in the generated

summaries as the reference summaries.

Finally, we identified a way to determine more accurately what the user will recognize as a more extractive or abstractive summary. To be more accurate about the level of perception of abstractivity, we slice the process of converting a document into a summary in a set of summarization actions and quantify the frequency of appearance. This way of quantifying the perception of abstractivity enabled the possibility of joining the two strategies of summarization –usually mutually excludents– (extractive and abstractive summarization) in a single dimension. We showed evidence that users do not perceive summaries as extractive or abstractive; there are grades of abstractivity/extractivity, and the identified summarization actions have a role in varying the grade.

In closing, we identified aspects related to the composition of the summary and not related to the selection of the information that will contain the summary. They are aspects that should be considered from the users' needs and should not remain constant in the behavior of summarization systems; that is, summarization systems should "listen" to the user in order to satisfy what the users need. It is clear that if that information from the user is considered, the system will create a higher quality summary from the user's perspective than the summary created without that information, even though both summaries retain almost the same information and readability.

9.1 Thesis contributions

In this section, we list all the contributions derived from the all the works from the thesis. The contributions will be listed by chapter.

Table 9.1 syntezises the contributions from the work of Chapter 2: "DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles". The contributions are grouped by type of contribution.

Table 9.1: Contributions from Chapter 2 - DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles.

Type	Section	Contribution
Methodologies	2.3	• Filtering and creation of a summarization dataset to ensure certain quality and abstractivity in its summaries.

Type	Section	Contribution
Software	2.3	• Distributed newspaper crawling software highly configurable through CSS selectors.
Datasets	2.4	• Public-available large-scale dataset with more 700K document-summary pairs for Catalan and more than 2M for Spanish: `https://huggingface.co/datasets/ELiRF/dacsa`
Models	2.6	• Two Transformers-based abstractive summarization models and one extractive model for Catalan; three more for Spanish. The abstractive summarization models are publicly available: - ***mBARTca*:** `https://huggingface.co/ELiRF/mbart-large-cc25-dacsa-ca` - ***mBARTes*:** `https://huggingface.co/ELiRF/mbart-large-cc25-dacsa-es` - ***mT5ca*:** `https://huggingface.co/ELiRF/mt5-base-dacsa-ca` - ***mT5es*:** `https://huggingface.co/ELiRF/mt5-base-dacsa-es`
Studies	2.5	• DACSA dataset analysis includes an abstractivity study and an abstractive/extractive approaches summarization performance.

Table 9.2 lists the contributions from the work of Chapter 3: "NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish".

Table 9.2: Contributions from Chapter 3 - NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish.

Type	Section	Contribution
Methodologies	3.4	• Pre-training methodology to increase the knowledge transfer for summarization and increment the perception of abstractivity in summaries.
	3.5	• Measurement of abstractivity using a set of abstractivity-related metrics.
Metrics	3.5	• *Content Reordering*. A first metric for abstractivity measurement/characterization is not based on quantifying the number of overlapping words or segments between the document and summary.
Models	3.4	• Transformers-based pre-trained models: One pre-trained model for Catalan and another for Spanish. Language models with more transferable knowledge for summarization.
	3.4	• Transformers-based abstractive summarization models: three models for Catalan and another three for Spanish. The abstractive summarization models based on our pre-training methodology are publicly available: - **NASca:** `https://huggingface.co/ELiRF/NASCA` - **NASes:** `https://huggingface.co/ELiRF/NASES`
Studies	3.6	• Performance and abstractivity study of summarization models in Catalan and Spanish.

Table 9.3 syntezises the contributions from the work of Chapter 4: "ELiRF-VRAIN at BioNLP Task 1B: Radiology Report Summarization".

Table 9.3: Contributions from Chapter 4 - ELiRF-VRAIN at BioNLP Task 1B: Radiology Report Summarization.

Type	Section	Contribution
Methodologies	4.3	• Enhancing the data leveraging during the pre-training phase with a sliding window based on sentences.
Models	4.3	• Transformers-based pre-trained model adapted to the biomedical domain with better transferable knowledge for summarization.
	4.4	• Transformers-based abstractive biomedical summarization model for radiology report summarization in English.
Studies	2.5	• Study related to the lenght of the generated summaries and the location of relevant information in them.

Table 9.4 condenses the contributions from the work of Chapter 5: "ELiRF-VRAIN at BioLaySumm: Boosting Lay Summarization Systems Performance with Ranking Models".

Table 9.4: Contributions from Chapter 5 - ELiRF-VRAIN at BioLaySumm: Boosting Lay Summarization Systems Performance with Ranking Models.

Type	Section	Contribution
Methodologies	5.5	• Generating lay summaries with different quality through prompting and Large Language Models (LLMs) for training the ranking model.
Models	5.3	• Longformer-based pre-trained model adapted to the biomedical research domain with better transferable knowledge for summarization.
	5.4	• Longformer-based abstractive summarization models for biomedical research articles lay summarization in English.

Type	Section	Contribution
	5.5	• Longformer-based regression model to predict the quality of lay summaries of biomedical research articles on three aspects: *Relevance*, *Readability*, and *Factuality*.
Studies	5.5	• Ranking capabilities of the regression model.
	5.5	• Impact of the regression model in the final performance.

Table 9.5 summarizes the contributions from the work of Chapter 6: "Abstractive Summarizers Become Emotional on News Summarization".

Table 9.5: Contributions from Chapter 6 - Abstractive Summarizers Become Emotional on News Summarization.

Type	Section	Contribution
Methodologies	6.3	• Framework for extracting and studying emotions in texts and emotional comparisons between a document and its summary.
	6.6	Emotional comparisons between reference and generated summaries.
Metrics	6.6	• Emotional coherence and bias, for comparisons between generated summary and reference summary regarding text entailed emotions.
Software	6.1	• Software for emotion extraction in texts.
Models	6.6	• Two Transformers-based abstractive summarization models for English.
Studies	6.5	• Emotional analysis for two well-known newspaper datasets in English.

Type	Section	Contribution
	6.6	• Emotional analysis for extractive and abstractive summarization systems and impact of the datasets used for fine-tuning in their bias.

Table 9.6 lists the contributions from the work of Chapter 7: "Beyond Using Their Own Words: Abstractivity Characterization in Summarization".

Table 9.6: Contributions from Chapter 7 - Beyond Using Their Own Words: Abstractivity Characterization in Summarization.

Type	Section	Contribution
Concepts	7.3	• The frequency of usage of summarization actions is related to the level of abstractivity of the resulting summary.
	7.3	• Extractive and abstractive strategies coexist in a single continuous dimension.
Methodologies	7.5	• Describe a summary by measuring the frequency of summarization actions for measuring the level of abstractivity.
Datasets	7.2	• *CLAsum* is the first available dataset for abstractivity characterization in summaries.
Software	7.7	• YALT is a highly adaptable offline annotation tool through HTML+CSS templates.
Models	7.5	• Set of models for abstractivity level prediction in summaries.
Studies	7.2	• Abstractivity characterization and level prediction through the study of the *CLAsum* dataset.
	7.5	• Abstractivity characterization and level prediction with supervised machine learning algorithms.

9.2 Future work

From the work that we have been carrying out in this thesis, we can identify several research promising oportunities that can be studied in future research works related to automatic summarization:

Expanding self-supervised tasks for summarization: Using pre-training tasks to prepare the general language models for the downstream task has been proven beneficial. Since they do not require any specific curated dataset but rather raw or general text, introducing the proper tasks of this type in the model development would usually be positive on the final performance of the model. Therefore, studying ways to create or refine self-supervised tasks is promising. In the case of summarization, we could explore approximations based on text reordering or synonym introduction of words/expressions, for instance.

Data augmentation for quality measurement: The capabilities of LLMs to generate text has been growing extremly fast in the past few years, and they are still growing. Since the aquisition of data for summarization quality measurement is heavily time-consuming when the process is done by humans, it is natural to look at LLMs to expand the resources. We have observed in this thesis that data augmentation gave us good initial results for training models for quality measurement. Therefore, it is very promissing to continue exploring this approach, even though, combining them with human effort to ensure the quality of the resources generated.

Emotion extraction with deep learning models: The approach presented in this thesis to extract emotions from text was a straightforward approach based on a lexicon and restricted to words. Thus, context is not considered (negations, for instance). We plan to explore the extraction of emotions in text with Transformers-based models. In this way, we will tackle the disambiguation problem present in our initial approach, and emotions from expressions or subjacent emotions in the text will also be extracted.

Summary generation guided by the emotions: Develop news summarization models with controllable text generation driven by the emotions of the reference summaries and via prompting, which could produce better emotional coherence in the generated summaries and potentially, reduce undesired biases towards some emotions and stances.

Expand the resources related with the abstractivity characterization and level detection: We want to expand the quantity and diversity of the resources focused on the perception of abstractivity in summaries to allow more generalizable

studies and further conclusions that would open new research lines.

Deep neural models for approaching the tasks of Abstractivite Inducting Features (AIFs) extraction and Abstractivity Level prediction: These tasks were explored in Chapter 7 with classical supervised machine learning algorithms. However, the use of deep learning approaches for these tasks has not been explored. With current resources (the *CLAsum* dataset, Section 7.2), we should employ low-data approximations like Low-Rank Adaptation (LoRA), the usage of Language Models Adapters, zero-shot or few-shots solutions, classification methods based on entailment or through prompting.

Control the level of perception of abstractivity: The generation of summaries with different grades of abstractivity through a single model has not been explored yet. Controlling the level of abstractivity with the user input, depending on the type of content or any control mechanism, would create more general and adaptable summarization models. Since we found evidence that abstractivity is related to the frequency of usage of different summarization actions, the control could also be done by controlling the different actions separately.

Reference-less quality measurement that takes into account the information of the abstractivity aspect of summaries: Explore the creation of reference-less solutions –measure the quality without a gold reference– to measure the quality of summaries considering the abstractivity information of the summary to be evaluated. Even though abstractivity is not related to the quality of the summary, the measurement could be more precise since the solution could adapt much better to more extractive or abstractive summaries indistinctly.

9.3 Thesis works unrelated to summary generation

These works helped consolidate and deepen the already acquired knowledge and add new knowledge related to Natural Language Processing and machine learning in general. Aspects such as feature extraction, selection of the machine learning method depending on the problem (supervised/unsupervised, classical/deep learning methods), data processing, statistical analysis, creation of new models based on Transformers, or data augmentation.

We participated in the following works:

- Vicent Ahuir, José Ángel González, and Lluís F. Hurtado (2022). "Enhancing

Sexism Identification and Categorization in Low-data Situations". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022*. Ed. by Julio Gonzalo et al. Vol. 3202. CEUR Workshop Proceedings. CEUR-WS.org. URL: https://ceur-ws.org/Vol-3202/exist-paper5.pdf

- Vicent Ahuir, Lluís-Felip Hurtado, Fernando García-Granada, and Emilio Sanchis (2023a). "ELiRF-VRAIN at DIPROMATS 2023: Cross-lingual Data Augmentation for Propaganda Detection". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023*. Ed. by Manuel Montes-y-Gómez et al. Vol. 3496. CEUR Workshop Proceedings. CEUR-WS.org. URL: https://ceur-ws.org/Vol-3496/diproma ts-paper6.pdf

- Vicent Ahuir, Lluís-Felip Hurtado, Fernando García-Granada, and Emilio Sanchis (2023b). "ELiRF-VRAIN at PoliticES-IberLEF2023: Dealing with Long Texts in Transformer-based Systems for User Profiling". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023*. Ed. by Manuel Montes-y-Gómez et al. Vol. 3496. CEUR Workshop Proceedings. CEUR-WS.org. URL: https://ceur-ws.org/Vol-349 6/politices-paper1.pdf

- Andreu Casamayor, Vicent Ahuir, Antonio Molina, and Lluís-Felip Hurtado (2024b). "ELiRF-VRAIN at MentalRiskES 2024: Using LongFormer for Early Detection of Mental Disorders Risk". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), Valladolid, Spain, September 24, 2024*. Ed. by Salud María Jiménez-Zafra et al. Vol. 3756. CEUR Workshop Proceedings. CEUR-WS.org. URL: https://ceur-ws.org/Vol-3756/Mental RiskES2024%5C_paper3.pdf

- Andreu Casamayor, Vicent Ahuir, Antonio Molina, and Lluís-Felip Hurtado (2024a). "ELiRF-VRAIN at eRisk 2024: Using LongFormers for Early Detection of Signs of Anorexia". In: *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*. Ed. by Guglielmo Faggioli, Nicola Ferro, Petra Galuscáková, and Alba García Seco de Herrera. Vol. 3740. CEUR Workshop Proceedings. CEUR-WS.org, pp. 803–812. URL: https://ceur-ws.org/Vol-3740/paper-75.pdf

9.4 Master's thesis and Bachelor's degree final projects

While developing this thesis, we also participated in co-mentoring several Master's thesis and Bachelor's degree final projects related to Transformer-based solutions to approach NLP problems. This experience helped us reinforce and increase some soft skills such as communication language adaptation, knowledge dissemination, project guidance, problem analysis, or critical analysis.

We co-mentored the following Bachelor's degree final projects:

- Diego Torres-Bertomeu (2023). "Condicionant l'estil en la generació de resums abstractius de notícies". Bachelor's Degree Final Project. Universitat Politècnica de València. URL: https://riunet.upv.es/handle/10251/197102
- Elias Esteve-Bernal (2023). "Simplificación de Sentencias Judiciales a Lectura Fácil". Bachelor's Degree Final Project. Universitat Politècnica de València. URL: https://riunet.upv.es/handle/10251/201325
- Alberto Sánchez-Sánchez (2024). "Adaptación de Sentencias Judiciales a Lectura Fácil usando técnicas de Procesamiento de Lenguaje Natural". Bachelor's Degree Final Project. Universitat Politècnica de València. URL: https://riunet.upv.es/handle/10251/208247
- Violeta Vicente-Cantero (2024). "Sistema de desnormalización de texto basado en modelos neuronales para el español y el catalán". Bachelor's Degree Final Project. Universitat Politècnica de València. URL: https://riunet.upv.es/handle/10251/207955

We also co-mentored the following Master's thesis:

- Andreu Casamayor-Segarra (2023). "Sistema de recuperació d'informació semàntica multilingüe". Master's Thesis. Universitat Politècnica de València. URL: https://riunet.upv.es/handle/10251/197525
- Victor Cardona-Lorenzo (2023). "Detección y agrupamiento de noticias en fuentes periodísticas digitales". Master's Thesis. Universitat Politècnica de València. URL: https://riunet.upv.es/handle/10251/198546
- Diego Torres-Bertomeu (2024). "Generación de resúmenes abstractivos en lenguaje simplificado". Master's Thesis. Universitat Politècnica de València. URL: https://riunet.upv.es/handle/10251/207678