Deep Learning for Natural Language Understanding and Summarization

Candidate:

Moreno La Quatra

Supervisor:

Luca Cagliero

Examination Board:

Campos Ricardo, Instituto Politécnico De Tomar

Di Caro Luigi, Università di Torino

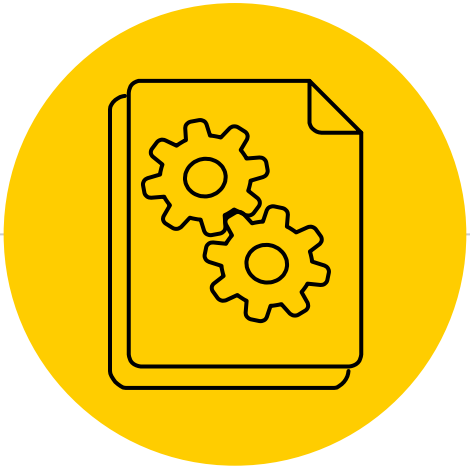
Mellia Marco, Politecnico di Torino

Papotti Paolo, École Nationale Supérieure des Télécommunications

Quintarelli Elisa, Università di Verona

Presentation outline

- Introduction to NLP and Text Summarization
- Analysis of scientific documents
- Temporal information in summarization
- Spoken content analysis
- Conclusions and future research directions



NLP and text summarization

Natural Language Processing

Natural language processing (NLP) is a field of **computer science**, **artificial intelligence**, and **linguistics** concerned with the interactions between computers and human (natural) languages.

In the last few years, there has been a shift towards **deep learning methods** in NLP. They are effective at many NLP tasks and are gradually becoming the standard approach.

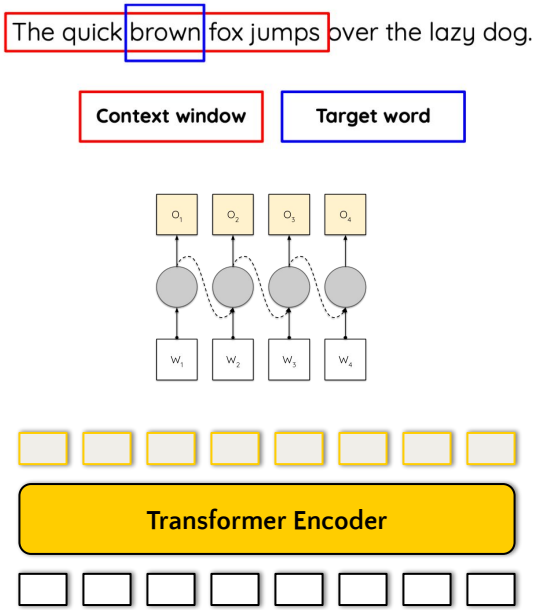
Self-supervised learning can be used to learn representations from data that are could be transferred to many different tasks.

Modern architectures for NLP

Word embedding models: words appearing in similar contexts have similar meanings.

Recurrent neural networks: specifically designed to handle sequential data.

Transformers: neural architecture that can effectively handle long-range dependencies.



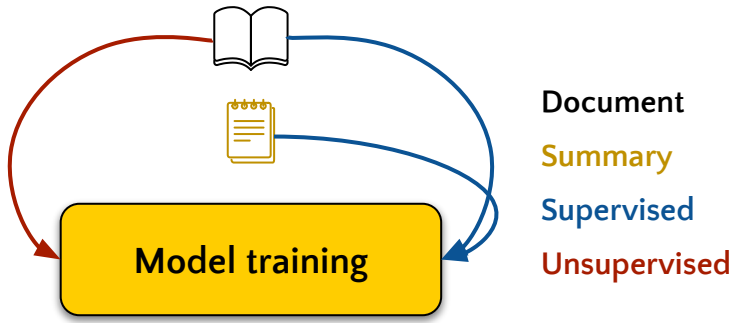
Automatic Summarization

It is the process of generating a **concise and fluent summary** from a document.

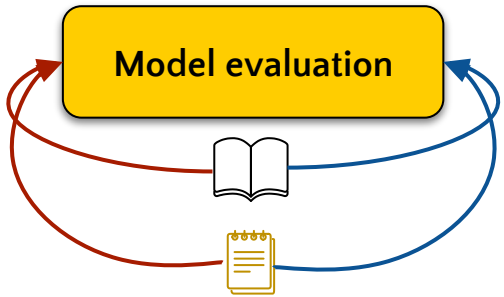
The pipeline can be customized according to the input data and the desired output. It usually involves some combination of the steps above.

Supervised vs Unsupervised

Supervised summarization models **require a training dataset** of input documents and their corresponding human-generated summaries.



Unsupervised summarization models do not require a training dataset. They usually **rely on heuristics** to identify the most important parts of the document.



Extractive summarization

Extractive summarization is a type of summarization where the summary is generated by **selecting the most important sentences** from the original text.

It requires document understanding to identify the key sentences that best represent the document but it does not require language generation.

Highlights are short sentences used to annotate scientific papers. They complement the abstract content by conveying the main result findings. To automate the process of paper annotation, highlights extraction aims at extracting from 3 to 5 paper sentences via supervised learning. Existing approaches rely on ad hoc linguistic features, which depend on the analyzed context, and apply recurrent neural networks, which are not effective in learning long-range text dependencies. This paper leverages the attention mechanism adopted in transformer models to improve the accuracy of sentence relevance estimation. Unlike existing approaches, it relies on the end-to-end training of a deep regression model. To attend patterns relevant to highlights content it also enriches sentence encodings with a section-level contextualization. The experimental results, achieved on three different benchmark datasets, show that the designed architecture is able to achieve significant performance improvements compared to the state-of-the-art.

Abstractive Summarization

Abstractive summarization involves both document understanding and language generation.

The model needs to be able to identify the key points in the document as well as **generate a summary** that is fluent and coherent.

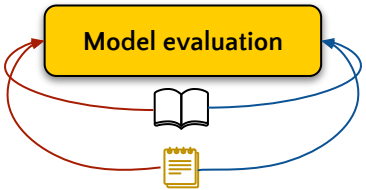
Highlights are short sentences used to annotate scientific papers. They complement the abstract content by conveying the main result findings. To automate the process of paper annotation, highlights extraction aims at extracting from 3 to 5 paper sentences via supervised learning. Existing approaches rely on ad hoc linguistic features, which depend on the analyzed context, and apply recurrent neural networks, which are not effective in learning long-range text dependencies. This paper leverages the attention mechanism adopted in transformer models to improve the accuracy of sentence relevance estimation. Unlike existing approaches, it relies on the end-to-end training of a deep regression model. To attend patterns relevant to highlights content it also enriches sentence encodings with a section-level contextualization. The experimental results, achieved on three different benchmark datasets, show that the designed architecture is able to achieve significant performance improvements compared to the state-of-the-art.

This paper proposes a novel methodology for selecting highlights from scientific papers. They are 3 to 5 sentences that convey the main result findings.The proposed model leverages the attention mechanism of transformers to improve sentence relevance estimation.

Evaluation Metrics

There exists a variety of evaluation metrics for automatic summarization:

- **ROUGE**: It leverages the **syntactic** overlap between candidate and reference summaries.
- **BERT-Score**: It takes into account the **semantic** similarity between the system and ground-truth summaries.
- **Human evaluation**: humans are asked to evaluate the system summaries. It is the **most expensive in terms of human efforts.**



Application domains - data types

Scientific documents

- Structured
- Explicit references
- Domain-specific

News

- Unstructured
- Implicit references
- Timestamped
- Multilingual

Spoken content

- Unstructured
- Multimodal
- Transcripted content
- Multilingual

Scientific Document Analysis

NLP for scientific summarization

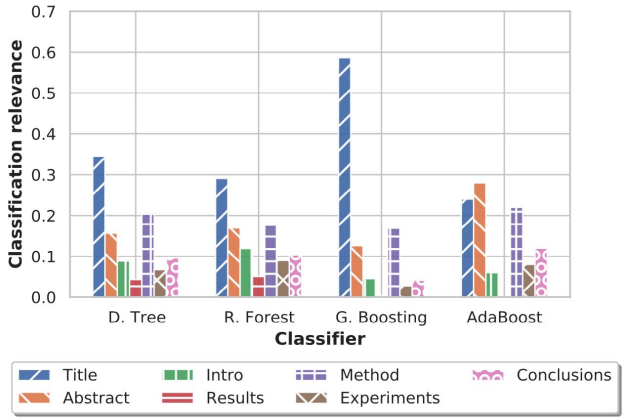
The diffusion of digital libraries has enabled the development of new methods for automatically analyzing large volumes of scientific literature.

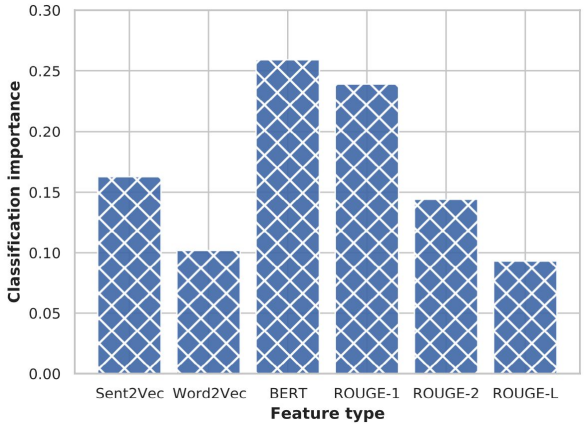
Scientific literature is growing at an exponential rate, making it increasingly difficult for researchers to keep up to date.

NLP can be used to **identify and extract essential information** from scientific texts leveraging both full-text and citation information.

Full-text analysis for citation context understanding

RQ1. How can we analyze the semantic correlation between citations and section in the reference paper?

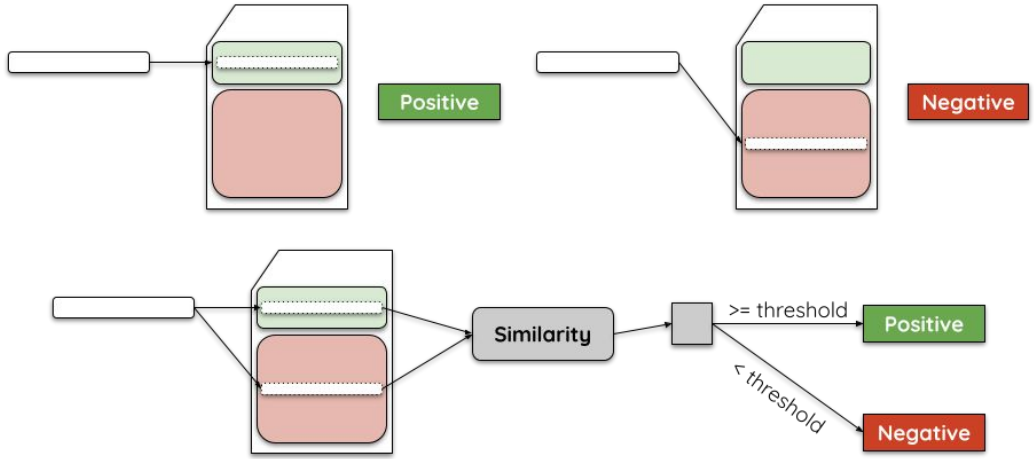




Semantic information extracted using deep learning models are relevant for the prediction task.

Citation context understanding - full-text analysis

RQ2: Can we use machine learning to identify citations that need full-text exploration, beyond just reading the abstract and title of cited articles?



Full-text analysis - results

Given a citation context in a reference paper, detect cases when reading open access sections of the target paper is not enough.

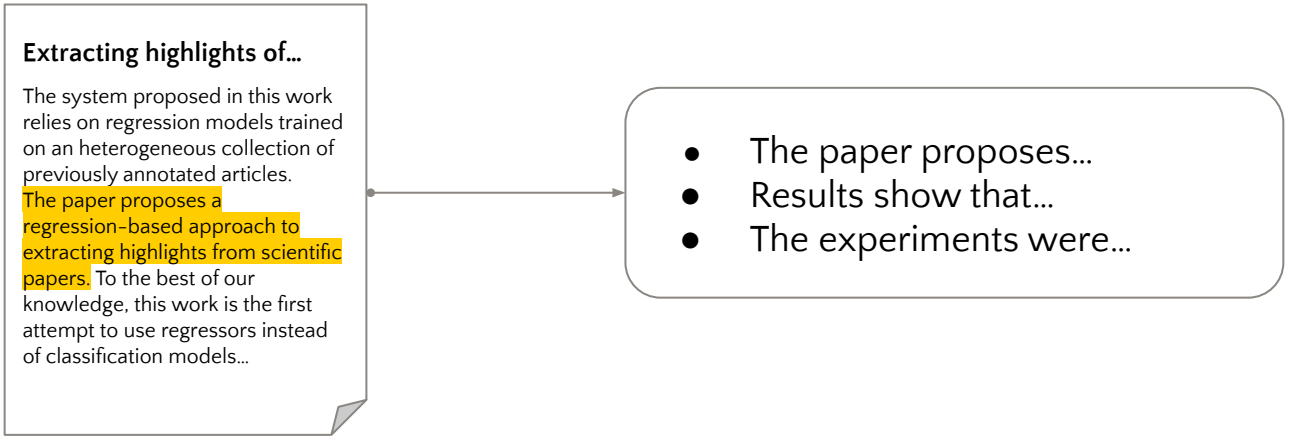
Lexical and semantic features are used to train a classification model to differentiate between cases that show clear benefits and those that do not.

Classifier	AUC	Accuracy	Negative class		
			Precision	Recall	F1-Score
AdaBoost	0.90	0.90	0.92	0.97	0.94
Gradient Boosting	0.91	0.91	0.94	0.96	0.95
Decision Tree	0.71	0.87	0.93	0.93	0.93
Random Forest	0.90	0.91	0.93	0.97	0.95

Highlights extraction from scientific publications

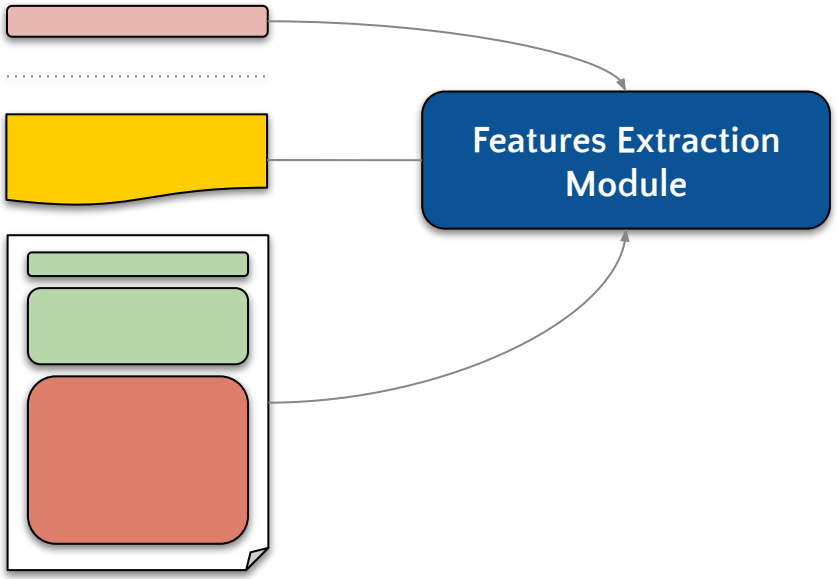
Researchers read a lot of papers to stay up to date in their field.

Scientific highlights extraction aims at automatically selecting a small set of sentences from a scientific publication that captures its main findings.



Highlights extraction - features extraction

We train a supervised model to rank sentences in a scientific publication according to their importance.

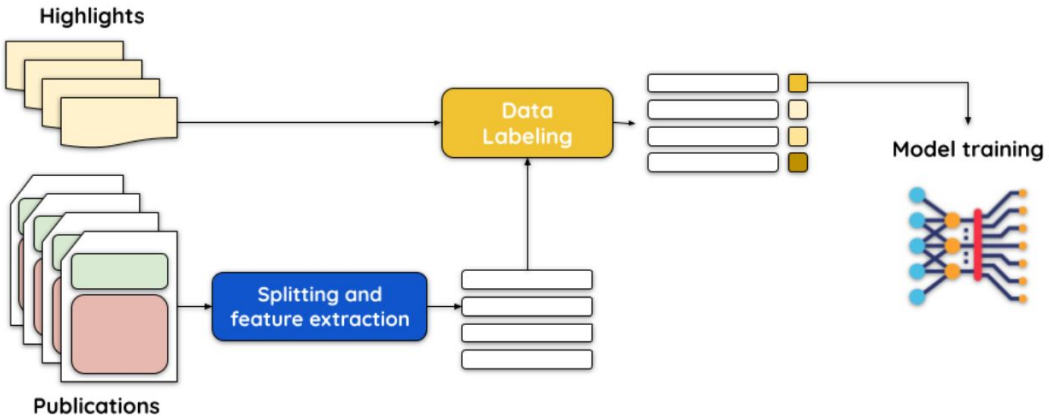


Sentence's features
Abstract Similarity
Title Similarity
Semantic Relevance
Sentence Position
Part-of-Speech count
Symbols count

Sentence label
Overlap with ground-truth highlights

Highlights extraction - regression model

We train a supervised model to rank sentences in a scientific publication according to their importance.

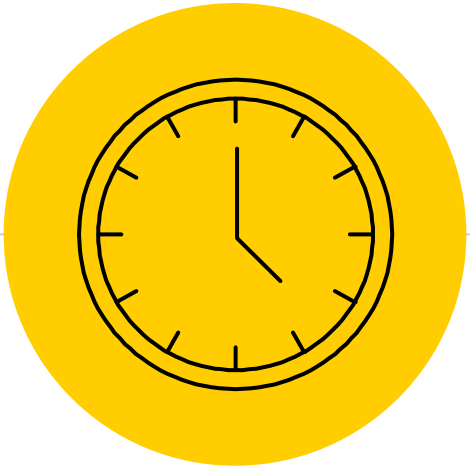


Highlights extraction - results

We compare the results of our model with state-of-the-art classification models and unsupervised baselines.

K	Sub-Modular	Text-Rank	Lex-Rank	DT CLF	RF CLF	MLP CLF	GB CLF	LSTM CLF	DT REG	RF REG	MLP REG	GB REG
						CSPubSumm						
3	0.235*	0.209*	0.257*	0.276*	0.298	0.272*	0.273*	0.295	0.303*	0.313	0.309*	**0.316**
4	0.228*	0.205*	0.237*	0.258*	0.284*	0.254*	0.254*	0.278*	0.291*	0.297*	0.297*	**0.303**
5	0.213*	0.193*	0.217*	0.239*	0.265*	0.239*	0.240*	0.256*	0.270*	0.278*	0.278*	**0.284**
						BioPubSumm						
3	0.221*	0.208*	0.227*	0.248*	0.253*	0.250*	0.250*	0.243*	0.259*	0.275*	0.278	**0.28**
4	0.215*	0.197*	0.223*	0.236*	0.241*	0.239*	0.238*	0.231*	0.250*	0.265*	0.27	**0.271**
5	0.204*	0.185*	0.199*	0.222*	0.227*	0.225*	0.224*	0.219*	0.237*	0.249*	**0.258**	0.257
						AIPubSumm						
3	0.180*	0.195*	0.225*	0.256	0.277	0.27	0.268	0.235*	0.256	0.283	0.28	**0.289**
4	0.175*	0.187*	0.212*	0.247	0.271	0.252	0.253	0.226*	0.256	0.274	0.267	**0.281**
5	0.166*	0.177	0.201*	0.227*	0.263	0.235	0.236	0.215*	0.244	0.263	0.259	**0.266**

Regression models learn to model the importance of publication sentences and can **rank them effectively** for highlights extraction.

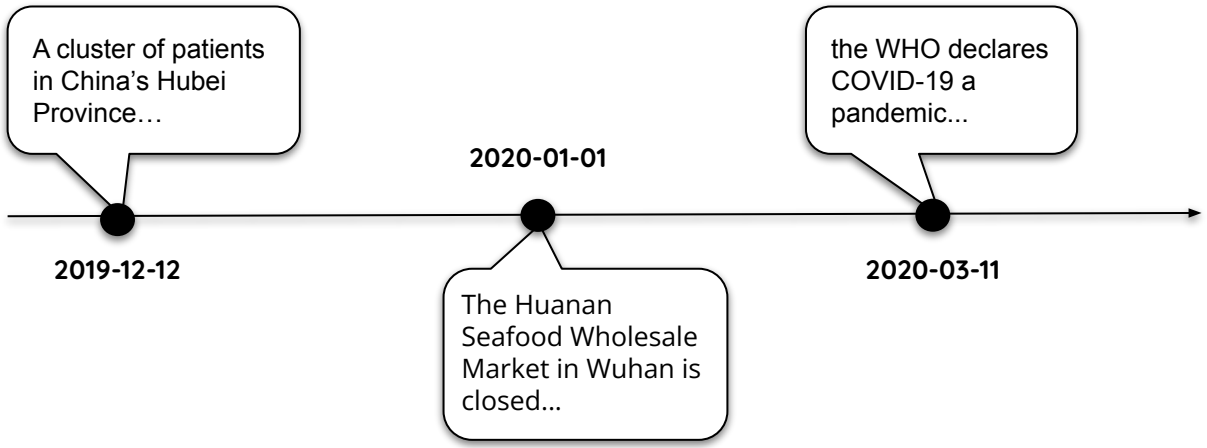


Temporal information in summarization

Timeline summarization

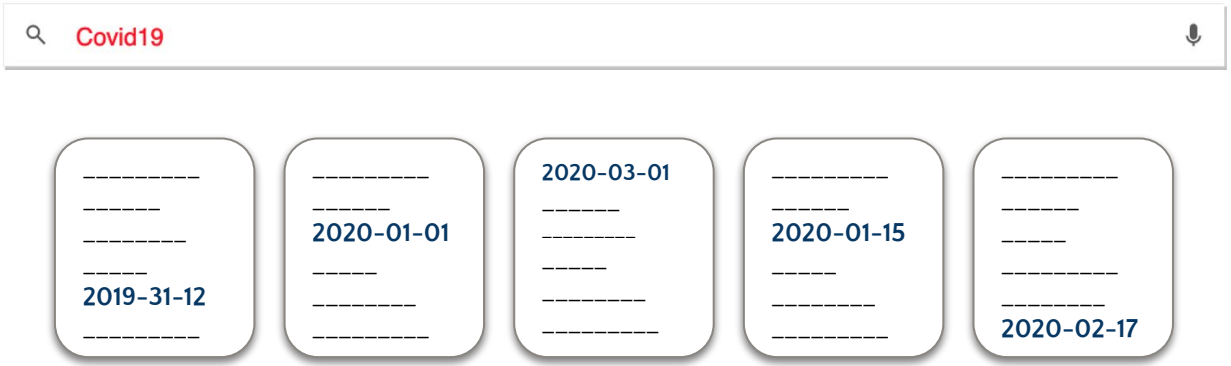
Events happening in the world are often complex and hard to follow.

A timeline of events can provide a concise and easy-to-understand overview of a sequence of events.



Timeline Summarization

The task usually takes a set of news as input and output a summary consisting of a set of dates with a short description of the event for each date.

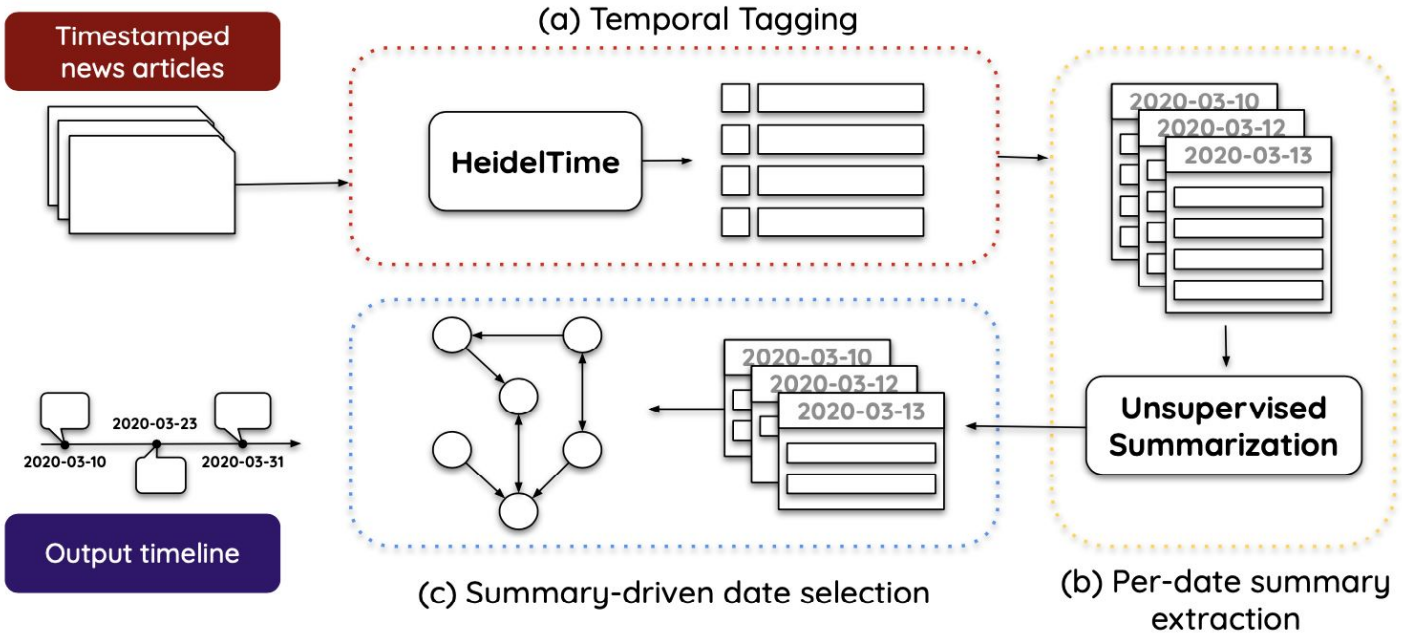


Summarize dates first

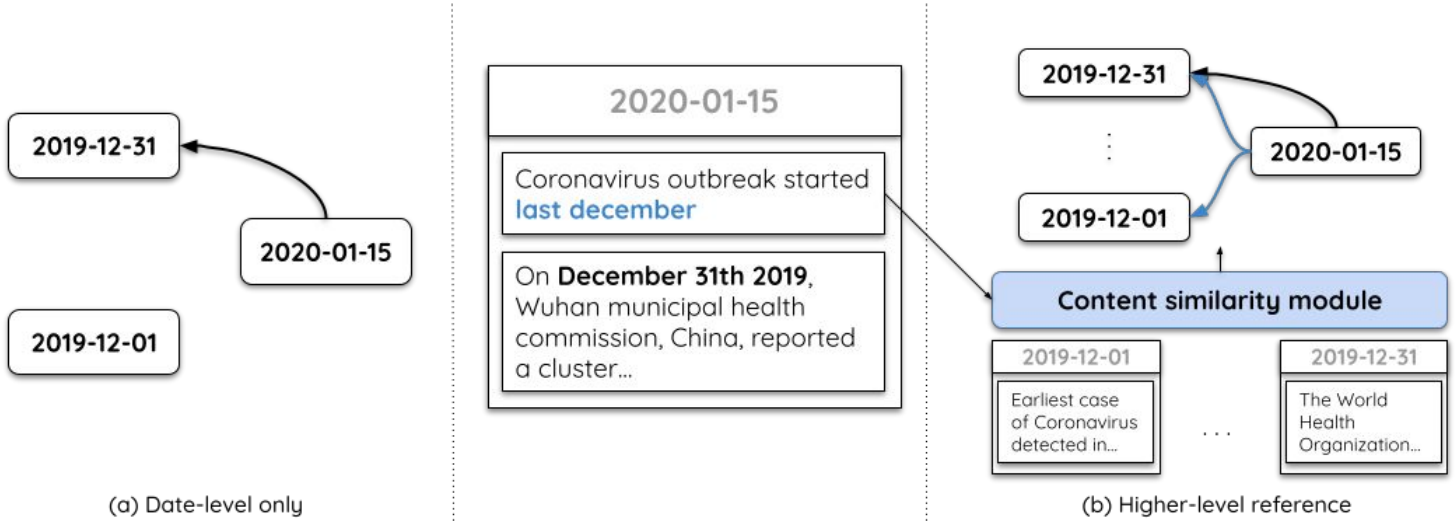
We propose to first summarize the events for each date, and then select a subset of dates that are considered to be important.

- Date selection could leverage the summarized event descriptions.
- It can exploit *high-level* temporal references.
- Disentangling summarization and selection allow the exploration of different methods for each sub-task.

Summarize dates first - pipeline overview



Summarize dates first - graph enrichment



Summarize dates first - results

- The in-degree of a node is the most effective relevance metric.
- Leveraging high-level temporal references can improve the results on benchmarks covering *long* events (e.g., more than 250 days)
- Unsupervised summarization models are able to effectively summarize dates.

Model		Type	Timeline 17	Crisis	Entities	CovidTLS
Chieu & Lee [26]		U	0.230*	0.166*	0.09*	0.176
Martschat & Markert [112]	ASMDS	U	0.531	0.278	0.163	0.685
	TLS-constraint	U	0.527	0.266	0.180	0.679
Proposed method	In-degree	U	0.549	**0.302**	**0.197**	**0.689**
	HITS	U	**0.553**	0.206	0.095	0.679
	Pagerank	U	0.537	0.175	0.161	0.623
	Degree	U	0.532	0.275	0.117	0.679
DateWise [51]		S	0.544	0.295	0.205	0.679

Spoken content summarization

Spoken content analysis

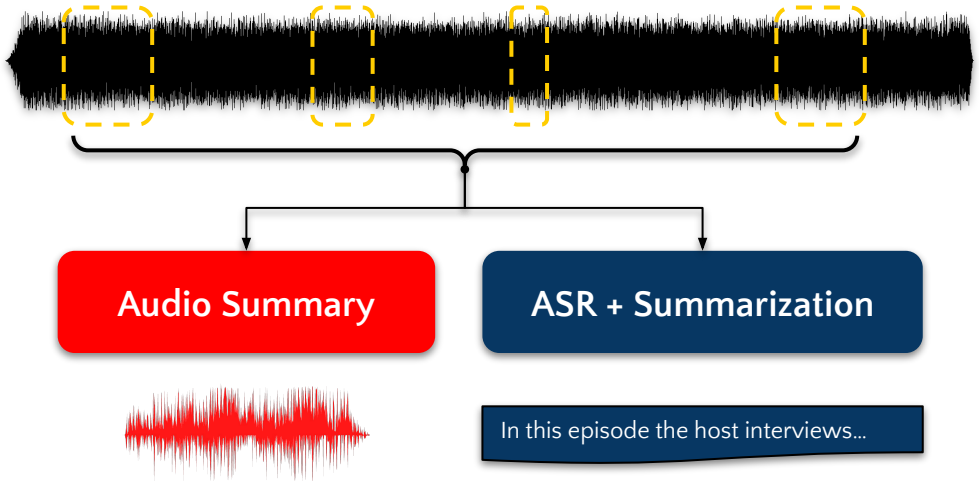
Natural language is not limited to the written form but is also expressed through other modalities such as recorded speeches.

Audio modality provides complementary information to the text and can be used to improve the understanding of the content.

Multimodal approaches combining both text and audio information have been shown to be effective in several tasks (e.g., emotion recognition)

Spoken content summarization

Spoken content summarization consists in generating a summary of an audio recording. It can be both a text transcript or an audio summary.

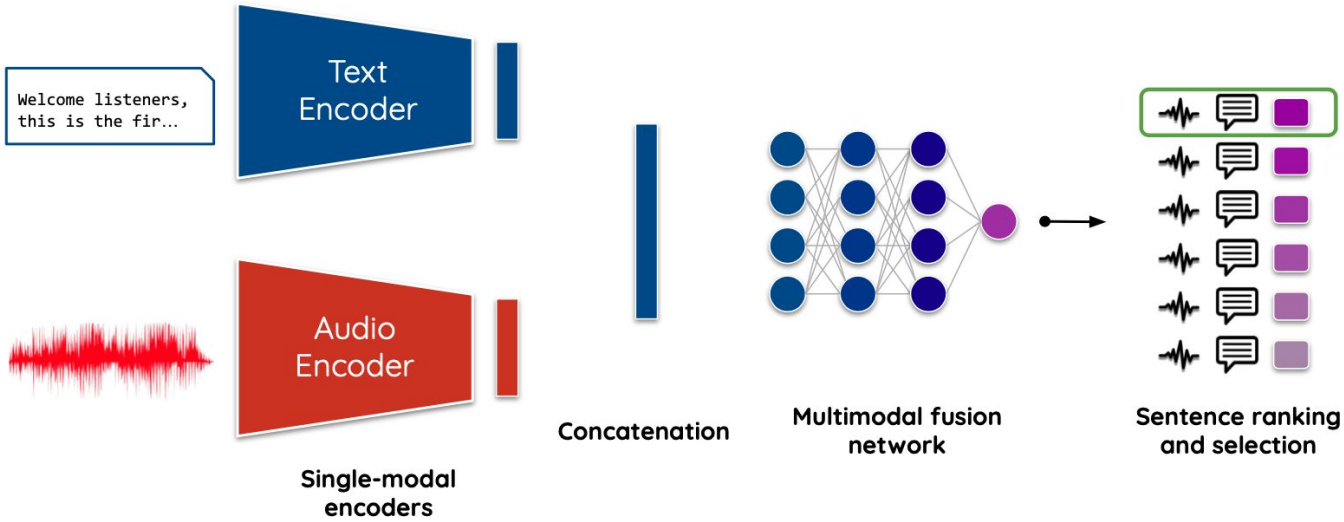


MATeR: Multimodal Audio-Text Regressor

MATeR is multimodal architecture for the extractive summarization of podcasts. It aims at providing a quick overview of the podcast content.

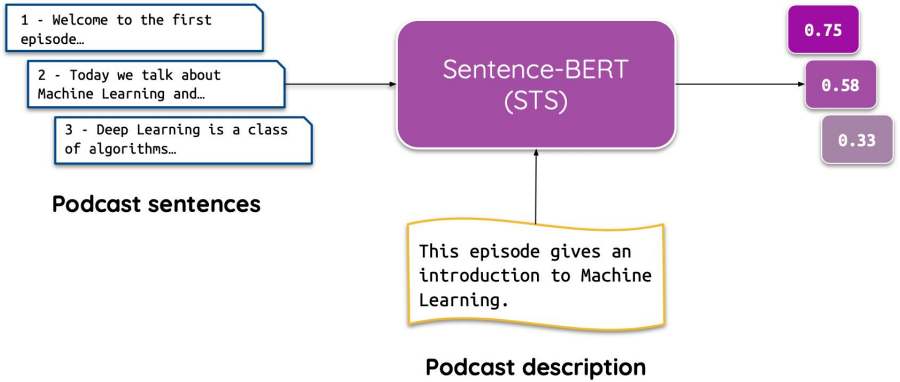
- **Text:** convey the main ideas of the podcast. It provides information on the topic and the key points made by the speaker.
- **Audio:** provide an overview of the soundscape of the podcast. It can convey the emotions and the overall tone of the talk.
- **Information fusion:** provide an integrated view of the podcast. It can highlight the most important aspects of the talk both in terms of content and delivery.

MATeR - architecture



MATeR - labeling process

The relevance score of a sentence in a podcast episode is computed using the semantic similarity of each sentence with the podcast description.



Training data consists of text, audio pairs with estimated relevance score.

MATeR - results

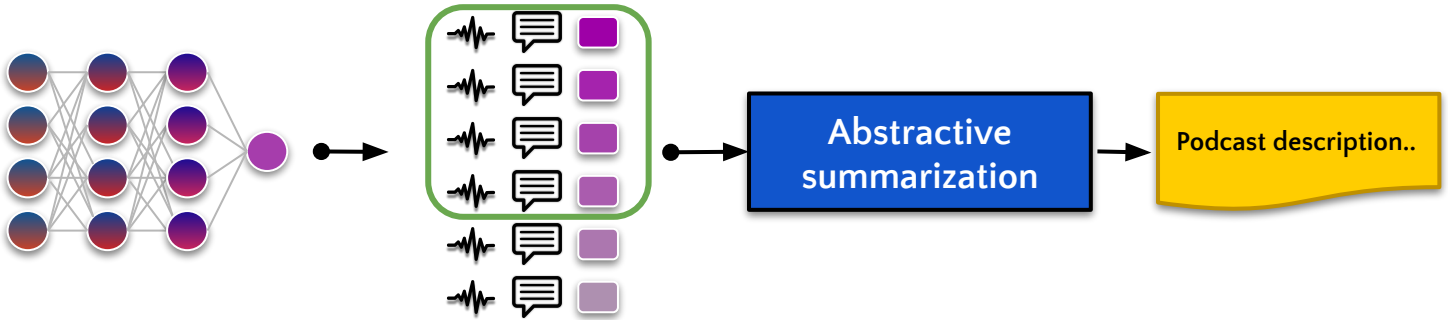
Method	R1-P	R1-R	R1-F1	R2-P	R2-R	R2-F1	RL-P	RL-R	RL-F1	SBERT
LEAD-1	0.150*	0.170*	0.142*	0.014*	0.013*	0.011*	0.129*	0.147*	0.122*	0.350*
TextRank	0.154*	0.177*	0.147*	0.015*	0.016*	0.013*	0.133*	0.154*	0.127*	0.363*
CoreRank	0.176*	0.179*	0.157*	0.030*	0.024*	0.023*	0.152*	0.154*	0.135*	0.418*
TextRank-BM25	0.156*	0.203*	0.159*	0.017*	0.020*	0.015*	0.132*	0.174*	0.135*	0.414*
HiBERT	0.186*	0.219*	0.184	0.036*	0.033*	0.031*	0.162*	0.191*	0.160	0.482
MATeR-text	0.162*	0.168*	0.143*	0.016*	0.016*	0.013*	0.140*	0.146*	0.123*	0.348*
MATeR	**0.193**	**0.225**	**0.188**	**0.042**	**0.041**	**0.036**	**0.168**	**0.197**	**0.164**	**0.490**

- MATeR outperforms its text-only counterpart as well as strong baselines designed for extractive summarization.
- Both lexical and semantic evaluation show that MATeR generates summaries that are more informative and better reflect the content of the podcast.

Select-and-rewrite for podcast summarization

We extend the framework using a two-step pipeline approach:

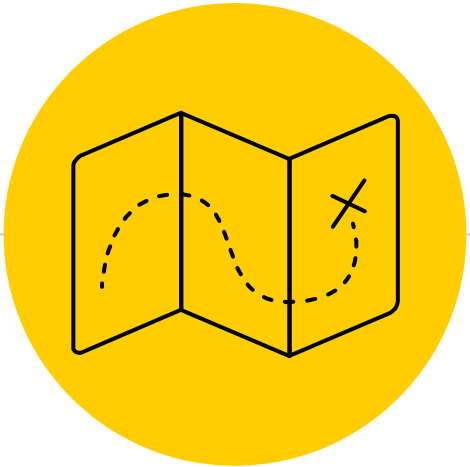
1. The most relevant sentences are selected from podcast transcript.
2. The selected sentences are fed into an abstractive summarization model to generate the final summary.



Select-and-rewrite - evaluation

The model has been evaluated at the TREC 2021 podcast track:

- The system ranked first according to the quality of the audio summaries.
- It is able to generate fluent summaries including the main topics of the podcast, thus providing relevant context for the listeners.
- The abstractive approach, however, is not always able to provide summaries that are **factually correct**. This aspect require further investigation in future works.

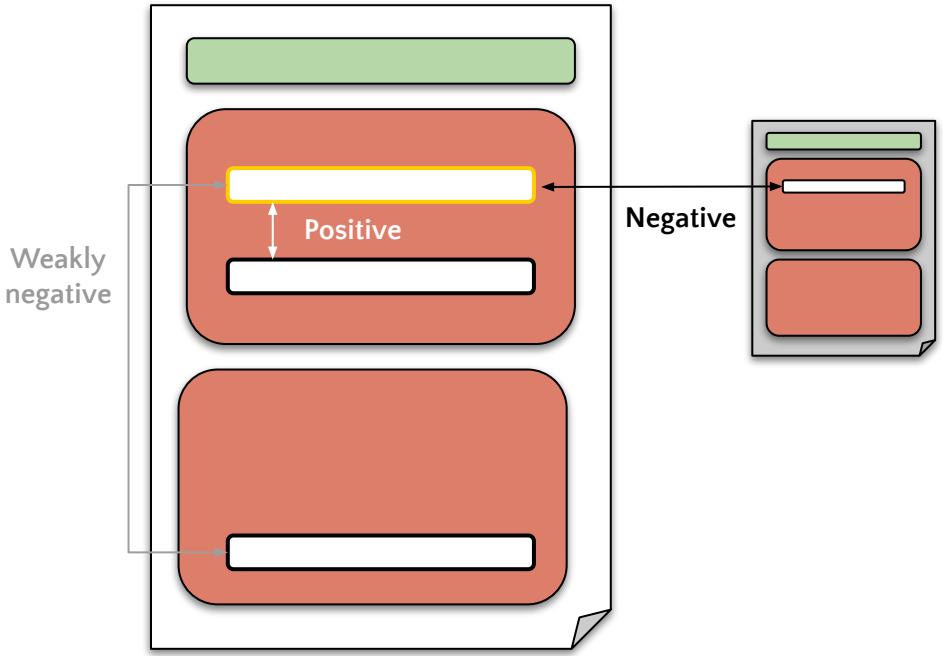


Future directions

Induction bias for contrastive learning

Contrastive learning allows the injection of inductive bias to generate training signal for self-supervised learning.

Scientific publications, do not only contain text, but also have a **structure that can be used to learn representations**.



Factual consistency in text generation

Factual consistency is a key desideratum for text generation. We may want to **constrain a text generation model** so that it does not output sentences that are factually incorrect with respect to a given context.

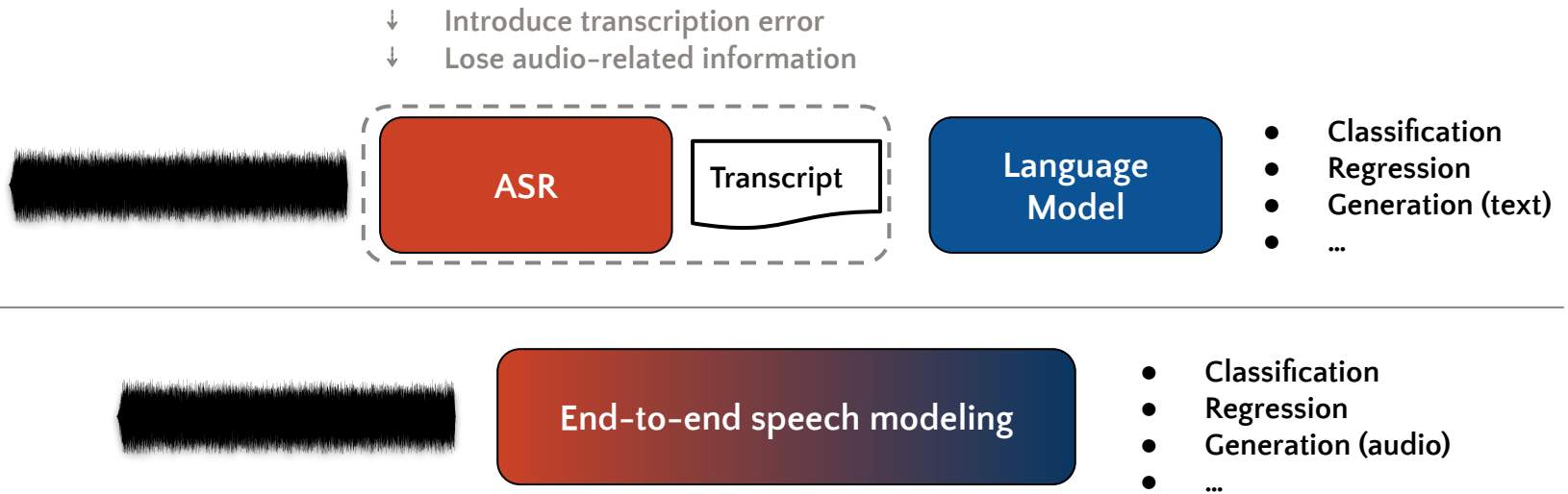
From **Microsoft 's Steve Ballmer** to Cisco Systems ' John Chambers to **IBM 's Samuel Palmisano** , chief executives of the nation's largest tech firms have written checks to the Bush campaign this election cycle .
While there are **exceptions -- such as Apple Computer's Steve Jobs** , a non - contributor recently named by John Kerry's presidential campaign as an economic adviser --

Steve Jobs has been a big donor to the Bush campaign this election cycle, but he's not the only one. **Apple's Steve Ballmer** and **Cisco's Samuel Palmisano** have also contributed to the campaign, reports the Wall Street Journal.

Real results from preliminary exploration of **abstractive timeline summarization.**

Textless NLP

Develop NLP models that can operate directly on the audio domain, without the need for transcribed text.



References

Scientific document analysis

La Quatra, M., Cagliero, L., and Baralis, E. (2021a). Leveraging full-text article exploration for citation analysis. Scientometrics, 126(10):8275–8293.

Cagliero, L. and La Quatra, M. (2020). Extracting highlights of scientific articles: A supervised summarization approach. Expert Systems with Applications, 160:113659.

Timeline summarization

La Quatra, M., Cagliero, L., Baralis, E., Messina, A., and Montagnuolo, M. (2021b). Summarize dates first: A paradigm shift in timeline summarization. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 418–427.

Spoken content analysis

Vaiani, L., La Quatra, M., Cagliero, L., and Garza, P. (2022). Leveraging multimodal content for podcast summarization. In Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, pages 863–870.

Vaiani, L., La Quatra, M., Cagliero, L., and Garza, P. (2021). Polito at trec 2021 podcast summarization track.

Thanks!

Candidate:

Moreno La Quatra

Supervisor:

Luca Cagliero

Examination Board:

Campos Ricardo, Instituto Politécnico De Tomar

Di Caro Luigi, Università di Torino

Mellia Marco, Politecnico di Torino

Papotti Paolo, École Nationale Supérieure des Télécommunications

Quintarelli Elisa, Università di Verona