5th INTERNATIONAL CONFERENCE ON ELECTRICAL AND ELECTRONICS ENGINEERING | September 11–12, 2024

Organized by: Swinburne University of Technology, Melbourne, Australia
Approved Publication Partner: Springer Nature LNEE Series [Scopus/WoS]
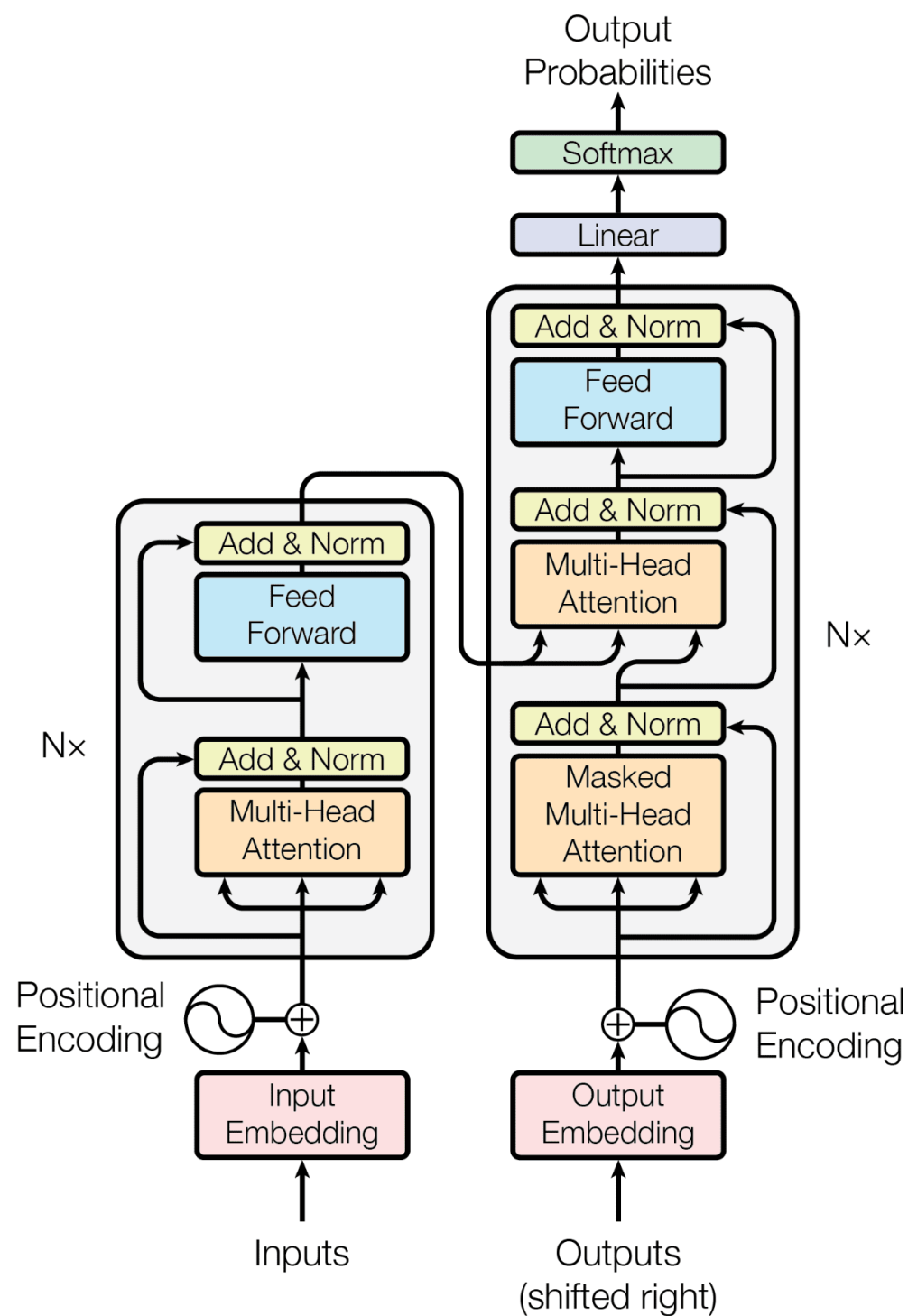
Paper Title:Text Summarization of Indo-Aryan Languages using Self-Attention Mechanism

Paper ID:192

Authors Name:Mr. Shubham Hadawle, Mr. Pranav Kotkar, Mr. Om Bhatia, Mr. Siddhant Dongre, Mr. Amit R. Singh

Presenter's Name: **Shubham Hadawle**
Presenter's Affiliation: **VESIT**

Using Transformers for Indic-Language Text Summarization

Exploring the latest advancements in using transformer models for improving accuracy and performance in Indic-Language text summarization.

What are LLMs like ChatGPT and Google's Bard based on?

Large Language Models (LLMs) like **ChatGPT** predominantly utilize the **transformer architecture**, which employs *self-attention mechanisms* to process natural language effectively. This architecture has revolutionized tasks like text generation and understanding in the field of artificial intelligence.

Why **Transformers**...?

- Transformers have revolutionized natural language processing (NLP) tasks due to their parallelization capabilities and efficient handling of long-range dependencies.

- LLMs like **ChatGPT** leverage the transformer's self-attention mechanism to generate accurate and contextually relevant responses.

- Without transformers, the current generative AI revolution wouldn't be possible [1].

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

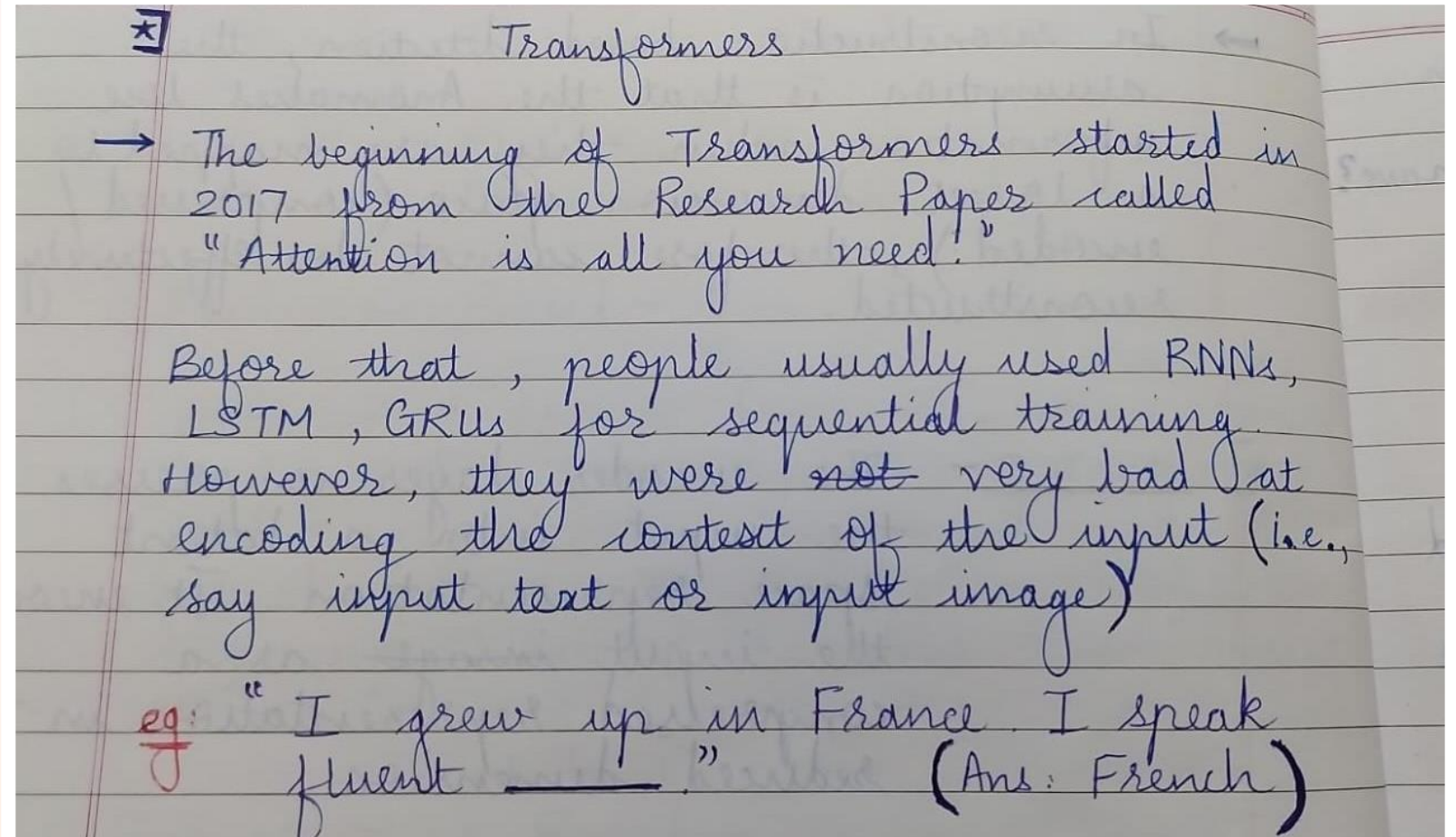**Illia Polosukhin*** [‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.
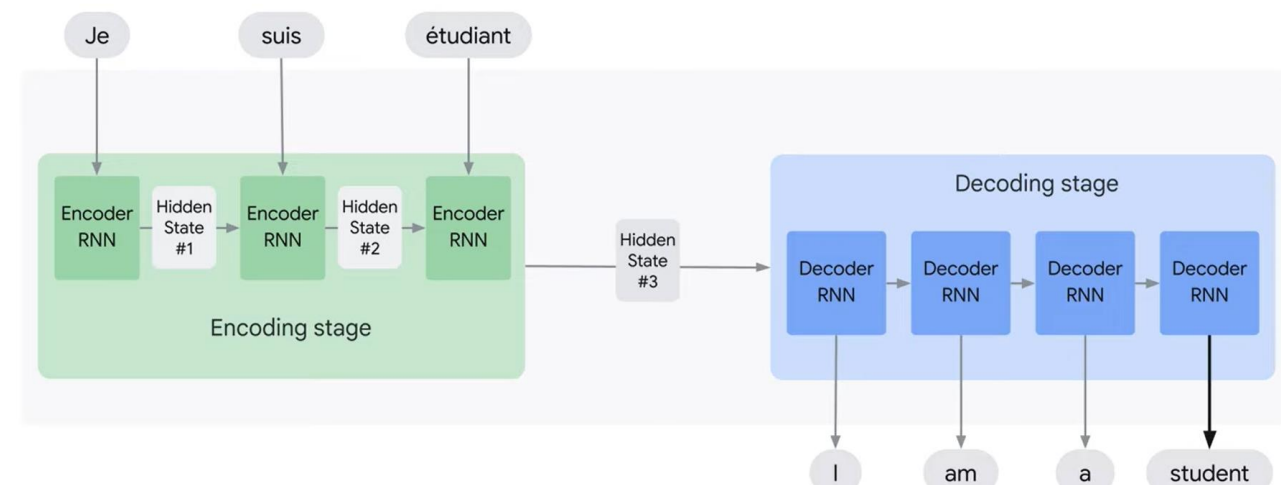
*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.
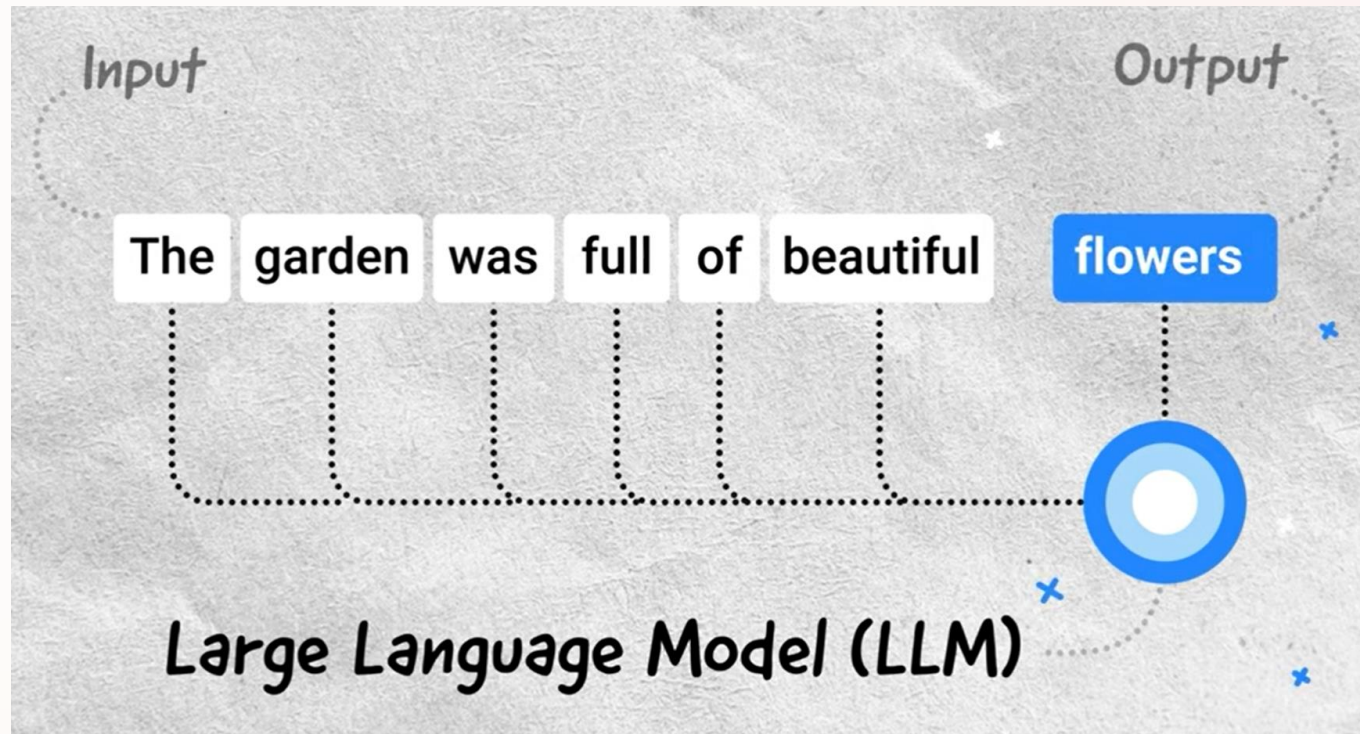
†Work performed while at Google Brain.
‡Work performed while at Google Research.

Transformers

→ The beginning of Transformers started in 2017 from the Research Paper called "Attention is all you need."

Before that, people usually used RNNs, LSTM, GRUs for sequential training. However, they were not very bad at encoding the context of the input (i.e., say input text or input image)

eg: "I grew up in France. I speak fluent _____." (Ans: French)

Traditional RNN encoder-decoder

Je   suis   étudiant

Encoder RNN — Hidden State #1 — Encoder RNN — Hidden State #2 — Encoder RNN

Encoding stage

Hidden State #3

Decoding stage

Decoder RNN → Decoder RNN → Decoder RNN → Decoder RNN

I   am   a   student

Input / Output

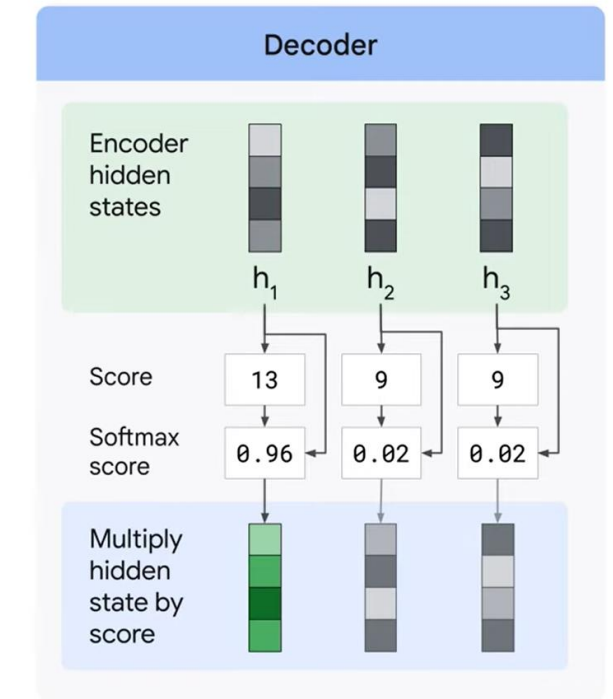The garden was full of beautiful **flowers**

Large Language Model (LLM)

To focus on the most relevant parts of the input:

1. Look at the set of encoder hidden states that it received.

2. Give each hidden state a score.

3. Multiply each hidden state by its soft-maxed score.

Decoder

	h_1	h_2	h_3
Encoder hidden states			
Score	13	9	9
Softmax score	0.96	0.02	0.02
Multiply hidden state by score			

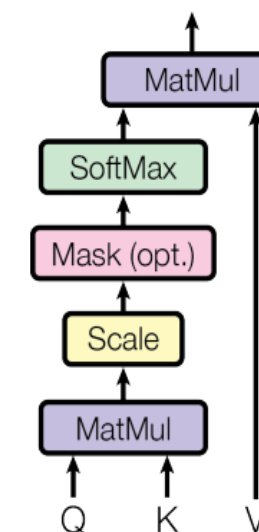$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
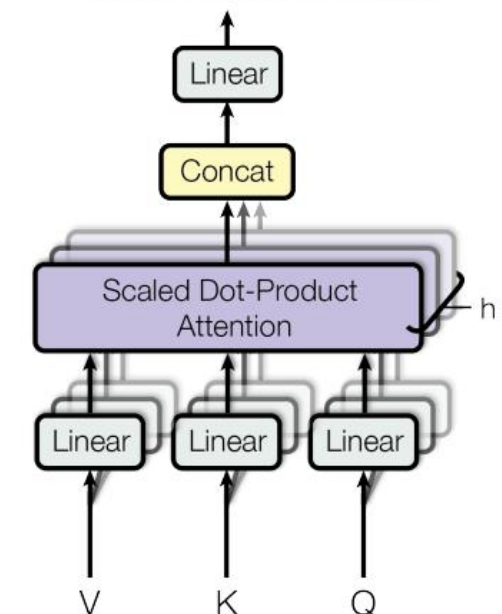
Q
K
V

Query : what are the things that I'm looking for?
Key : what are the things that I have?
Value : what are the things that I will communicate?
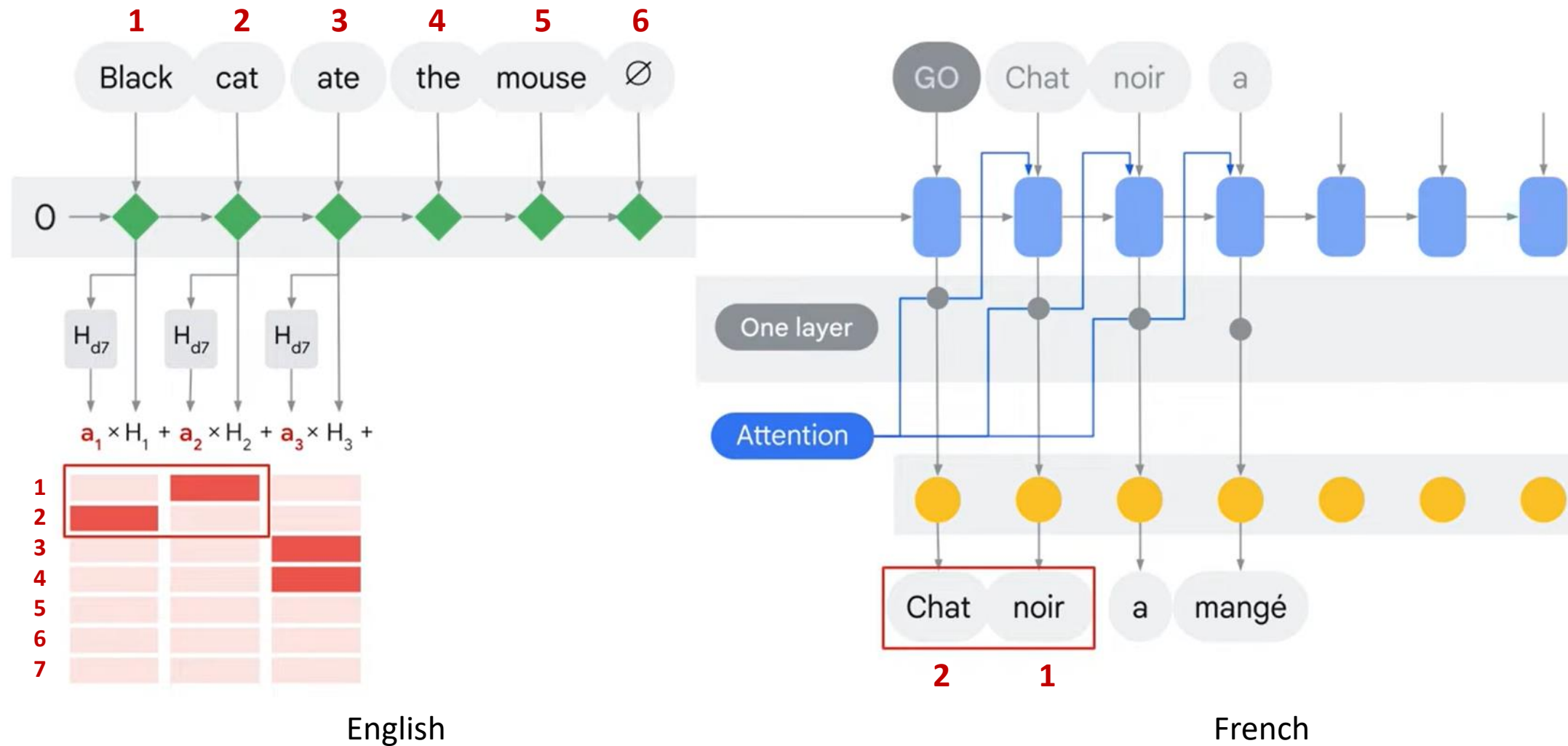
Scaled Dot-Product Attention

MatMul
SoftMax
Mask (opt.)
Scale
MatMul

Q K V

Multi-Head Attention

Linear
Concat
Scaled Dot-Product Attention
Linear Linear Linear

V K Q

Improve translation with attention network

Current Challenges in Indic-Language Summarization

Variations in Language Language Structures Structures

There are challenges posed by the unique linguistic structures and characteristics of Indic languages, such as Hindi, making summarization a complex task.

Inflectional Nature of Indic Languages

The inflectional nature of Indic languages is rich, which introduces additional complexity in generating accurate summaries.

eg: **Cat** in English, in Hindi बिल्ली (Billi) & बिल्ला (Billa)

Resource Limitations

Limited availability of high-quality training data for Indic languages, hindering the development and performance of text summarization models.

We have done **Text Summarization** of <u>two</u> *Indo-Aryan* Languages:
1. Hindi
2. Gujarati

Link to Dataset

Indian Language News Text Summarization Dataset

▲ 1   New Notebook   ⬇ Download (42 MB) ▾ ⋮

Data Card   Code (3)   Discussion (0)

Hin_train.csv (50.48 MB)   ⬇ ⛶ ❯

Detail   Compact   Column   5 of 5 columns ⌄

🔗 id ☰	🔗 Link ☰	A Heading ☰	A Summary ☰	A Article ☰
7957 unique values	**7957** unique values	**7957** unique values	**7922** unique values	**7936** unique values
a3d72cd62e7e1b5529dd 254a85fa2b5a	https://www.indiatv. in/india/national/in dia-successfully- test-fires-the-new- generation-agni- prime-mi...	भारत की बढ़ी ताकत, 'अग्नि प्राइम' मिसाइल का हुआ सफल परीक्षण, मारक क्षमता 1000 से 2000 किलोमीटर	भारत ने शनिवार को ओडिशा के बालासोर तट पर 'अग्नि प्राइम' मिसाइल का सफल परीक्षण किया। सरकारी अधिकारियो...	भुवनेश्वर : भारत ने शनिवार को ओडिशा के बालासोर तट पर 'अग्नि प्राइम' मिसाइल का सफल परीक्षण किया। सरका...
f8904fd63c1d257dc5ee 9b529df4a097	https://www.indiatv. in/india/national- covid-19- vaccination-india- has-administered- above-40-crore- dos...	Covid-19 Vaccination: देश में कोविड टीके की 40 करोड़ से अधिक खुराक दी गई, स्वास्थ्य मंत्रालय ने दी ज...	केंद्रीय स्वास्थ्य मंत्रालय के मुताबिक, आज शाम 7 बजे तक भारत में COVID वैक्सीनेशन के अंतर्गत कुल 40,...	नई दिल्ली। देश में कोरोना की संभावित तीसरी लहर की संभावनाओं के बीच कोरोना वैक्सीनेशन पर बहुत अधिक जो...
4e89ee2300613ae0c0e8 db4ed0e2c7aa	https://www.indiatv. in/india/uttar- pradesh/bjp-is- preparing-to-cater- to-brahmin-voters-	यूपी में ब्राह्मण वोटरों को साधने की तैयारी कर रही है बीजेपी, लोगों से घर-घर जाकर मुलाकात करेंगे कें...	बीजेपी के केंद्र सरकार में ब्राह्मण मंत्रियों, सांसद और विधायक प्रबुद्ध ब्राह्मण समाज के लोगों से घर...	बीजेपी यूपी में जनवरी के दूसरे हफ्ते से ब्राह्मण वोटरों को साधने और उनकी नाराज़गी को दूर करने के लिए...

Version 1 (163.93 MB)

▥ Eng_train.csv
▥ Eng_val_article.csv
▥ Guj_train.csv
▥ Guj_val_article.csv
▥ Hin_train.csv
▥ Hin_val_article.csv
▥ sample_submission.csv

Home > International Symposium on Intelligent Informatics > Conference paper

Abstractive Text Summarization of Hindi Corpus Using Transformer Encoder-Decoder Model

Rashi Bhansali ✉, Anushka Bhave, Gauri Bharat, Vedant Mahajan & Manikrao Laxmanrao Dhore

Conference paper | First Online: 05 April 2023

141 Accesses

Part of the Smart Innovation, Systems and Technologies book series (SIST,volume 333)

Abstract

Text Summarization based on Abstraction is the task of generating a concise summary that captures the principal ideas of the source text. It potentially contains new phrases that do not appear in the original text. Although it is widely studied for languages like English and French, owing to the scarcity of data on regional vernacular languages like Hindi, the research in this area is still in the primitive stages. We propose a novel approach for building an Abstractive Text Summarizer for Hindi corpus using the Transformer encoder-decoder architecture. Firstly, efficient pre-trained word representations are generated using Facebook's fastText model. Next, the Transformer model is employed to extract contextual dependencies and yield better

IndicBART is a multilingual, sequence-to-sequence pre-trained model focusing on Indic languages and English. It currently supports 11 Indian languages and is based on the mBART architecture. You can use IndicBART model to build natural language generation applications for Indian languages by finetuning the model with supervised training data for tasks like machine translation, summarization, question generation, etc. Some salient features of the IndicBART are:

- Supported languages: Assamese, Bengali, Gujarati, Hindi, Marathi, Odiya, Punjabi, Kannada, Malayalam, Tamil, Telugu and English. Not all of these languages are supported by mBART50 and mT5.

- The model is much smaller than the mBART and mT5(-base) models, so less computationally expensive for finetuning and decoding.

An abstractive text summarization technique using transformer model with self-attention mechanism

Original Article | Published: 01 June 2023

Volume 35, pages 18603–18622, (2023) | Cite this article

Neural Computing and Applications

Aims and scope →

Submit manuscript →

Access this article

Sandeep Kumar ✉ & Arun Solanki

726 Accesses | Explore all metrics →

Abstract

Creating a summarized version of a text document that still conve incredibly complex endeavor in natural language processing (NLP summarization (ATS) is the process of using facts from source sen them into concise representations while maintaining the content Manually summarizing large amounts of text are challenging and humans. Therefore, text summarization has become an exciting r This research paper proposed an ATS model using a *Transformer T Attention Mechanism (T2SAM)*. The self-attention mechanism is a

---

Abstractive Text Summarization for Hindi Language using IndicBART

Arjit Agarwal[1,†], Soham Naik[1,†] and Sheetal Sonawane[1]

[1]SCTR's Pune Institute of Computer Technology, Pune, India

Abstract

Text summarization is an important application of natural language processing (NLP) especially in this era where there is an abundance of information on the internet. In such a scenario, it will be easier to navigate useful information quickly if a clear and concise summary of articles (or other text sources) can be generated. It is time consuming to give this task to humans as it involves scanning thousands of words and documents. But by using the advancements in natural language processing, models can be constructed for text summarization, that generate summaries in an adept and concise manner. There is a big scope for implementing these advanced natural language processing techniques for a low-resource language like Hindi because of its popularity and the fact that relatively less research work is done in this field. This paper is a part of the ILSUM shared task whose main focus is to generate abstractive text summaries using textual data in Hindi language. The accuracy of the generated summaries are checked using the ROUGE evaluation metric. We have achieved a ROUGE-1 Fscore of 0.544 on the testing dataset by using the IndicBART model for training.
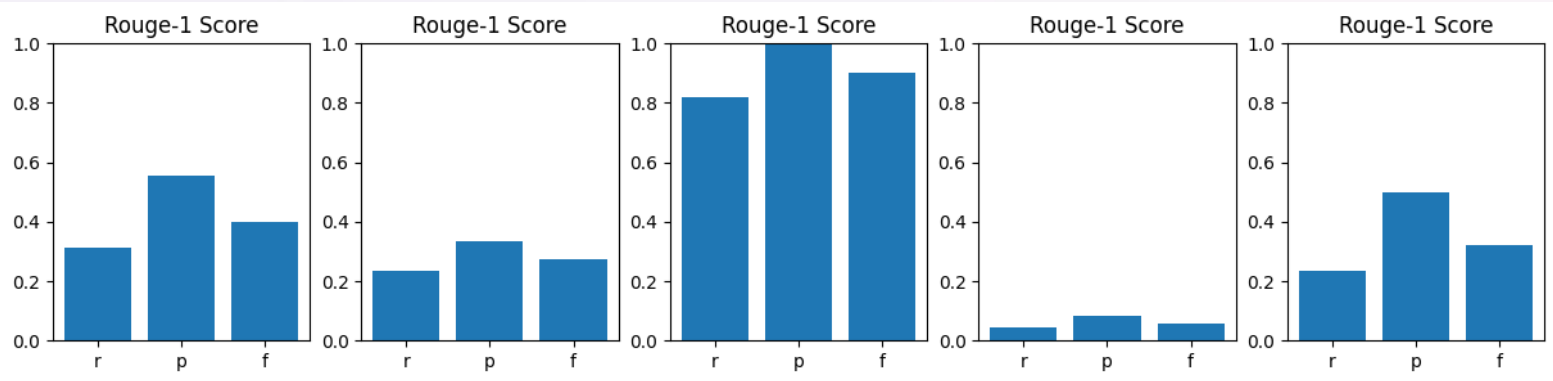
Keywords

Text summarization, Abstractive, Natural language processing, Bart, Rouge

Table 1 (Gujarati)

	Input	Reference Summary	Model Output Summary
0	પકડાયેલા આરોપીઓની તસવીર પરંતુ પોલીસને પાકી બાતમી મળી હોવાથી પારડી પોલીસની ટીમે ટેમ્પોને પોલીસ સ્ટેશન લાવી અને અંદરથી તપાસ કરતા તેમાં એક ચોરખાનું બનાવેલું મળ્યું હતું.. આ ચોર ખાના ને ખોલી ને જોતા પોલીસ પણ ચોંકી ગઈ હતી. કારણ કે ઉપરથી ખાલી દેખાતા ટેમ્પોમાં બનાવેલા ચોર ખાનામાં વિદેશી દારૂની 816 બોટલો મળી આવી હતી.. આથી પોલીસે વિદેશી દારૂ નો જથ્થો અને ટેમ્પો ચાલક અને તેના સાથીદારની ધરપકડ કરી હતી. પોલીસની આ કાર્યવાહીમાં વિદેશી દારૂ અને ટેમ્પો મળી અંદાજે 2.41 લાખ રૂપિયાનો મુદ્દામાલ કબજે કર્યો છે. અને ટેમ્પો ચાલક પ્રવીણ પાટીલ અને તેની સાથે રહેલા કોમલ	Valsad News: ખાલી ટેમ્પો લઇને પોલીસ પહોંચી સ્ટેશન, તપાસ કરતા નીકળ્યો 816 બોટલ દારૂ, પોલીસ ચોંકી ગઈ	ખાલી ટેમ્પો લઇને પોલીસ પહોંચી સ્ટેશન પહોંચી ગઈ
1	અમદાવાદ સહિત ગુજરાતમાં કેસમાં ઉત્તરોત્તર વધારો નોંધાઈ રહ્યો છે. ત્યારે 2020માં શરૂ થયેલી કોરોના મહામારી 2022ના આરંભે પણ કહેર વર્તાવી રહી છે. નવા વેરિયન્ટ ઓમિક્રોનના કેસની સ્થિતિ આવનારા સમયમાં ગંભીર બનશે એવી આશંકા સિવિલ હોસ્પિટલના સુપરિન્ટેન્ડેન્ટ ડો. રાકેશ જોશીએ દિવ્ય ભાસ્કર સાથેની વાતચીતમાં કરી છે. એવામાં કોરોનાની પહેલી લહેર 2020માં નવેમ્બરમાં હાઈ રહી હતી, તો 2021માં માર્ચ-એપ્રિલમાં પણ હાઈ રહી હતી. આ દરમિયાન 2021ના ડિસેમ્બરમાં અંતમાં અને 2022ની શરૂઆતમાં જાન્યુઆરીમાં ત્રીજી લહેરનું મંડાણ થયાં છે. કોરોનાના કેસમાં રોજેરોજ નવા રેકોર્ડ બની રહ્યા છે. તેવામાં કોરોનાના નવા કેસ 8 હજારના આંકડાને પણ અમદાવાદમાં વટાવી દે એવી શક્યતા છે.બીજી અને ત્રીજી લહેરની	ઓમિક્રોનનું સુપરફાસ્ટ સ્પ્રેડિંગ:ત્રીજી લહેરમાં પણ અમદાવાદ જ એપીસેન્ટર, આ ગતિએ તો 5 દિવસમાં જ 5800ના સિંગલ ડે ઓલટાઇમ હાઇનો આંક ક્રોસ થઇ જશે	ત્રીજી લહેરમાં પણ અમદાવાદ જ દિવસમાં જ કોરોનાના માત્ર સંક્રમણ થઇ જશે
2	અમદાવાદ : વિશ્વ માનવ અધિકાર દિવસ (World Human Rights Day-2021)અને પવિત્ર દલાઈ લામાને (Dalai Lama)એનાયત કરાયેલા 32માં નોબેલ શાંતિ પુરસ્કાર દિવસ પર રિવરફ્રન્ટ તિબેટિયન સ્વેટર માર્કેટ (Riverfront Tibetan Sweater Market)ખાતે કાર્યક્રમ યોજાયો હતો. ભારત તિબેટ મૈત્રી સંઘ અને તિબેટિયન સ્વેટર માર્કેટના સંયુક્ત પ્રયાસથી વિશ્વ માનવ અધિકાર દિવસ અને પવિત્ર દલાઈ લામાને એનાયત કરાયેલા 32માં નોબેલ શાંતિ પુરસ્કાર દિવસ પર યોજાયેલા કાર્યક્રમમાં મેયર કિરીટ પરમાર (Mayor Kirit Parmar), એમ જ પુસ્તકાલયના સભ્ય ડો હેમંત ભટ્ટ તેમજ ભારત તિબેટ મૈત્રી સંઘના રાષ્ટ્રીય ઉપાધ્યક્ષ ડો.અમિત જ્યોતિકર ઉપસ્થિત રહ્યા હતા. આ પ્રસંગે મેયર કિરીટ પરમારે તિબેટની સંસ્કૃતિ તેમજ ભારત સાથેના સંબંધો વિશે	Human Rights Day 2021: અમદાવાદ : તિબેટની આઝાદીમાં લેવો જોઇએ ભાગ, જાણો શા માટે	શા માટે રૂપાણીએ ચાર યુવકે સાયકલ

rouge_score

[{'rouge-1': {'r': 0.375, 'p': 0.8571428571428571, 'f': 0.5217391262003781},
'rouge-2': {'r': 0.25, 'p': 0.5714285714285714, 'f': 0.34782608272211724},
'rouge-1': {'r': 0.375, 'p': 0.8571428571428571, 'f': 0.5217391262003781}},
{'rouge-1': {'r': 0.3181818181818182,
 'p': 0.6363636363636364,
 'f': 0.4242424197979798},
'rouge-2': {'r': 0.22727272727272727,
 'p': 0.4545454545454545,
 'f': 0.3030302985858586},
'rouge-1': {'r': 0.3181818181818182,
 'p': 0.6363636363636364,
 'f': 0.4242424197979798}},
{'rouge-1': {'r': 0.14285714285714285,
 'p': 0.3333333333333333,
 'f': 0.19999999580000008},
'rouge-2': {'r': 0.07692307692307693, 'p': 0.2, 'f': 0.11111110709876558},
'rouge-1': {'r': 0.14285714285714285,
 'p': 0.3333333333333333,
 'f': 0.19999999580000008}},
{'rouge-1': {'r': 0.36363636363636365,
 'p': 0.5714285714285714,
 'f': 0.44444443969135805},
'rouge-2': {'r': 0.2, 'p': 0.3333333333333333, 'f': 0.24999999531250006},
'rouge-1': {'r': 0.36363636363636365,
 'p': 0.5714285714285714,
 'f': 0.44444443969135805}},
{'rouge-1': {'r': 0.75, 'p': 0.9230769230769231, 'f': 0.8275862019500595},
'rouge-2': {'r': 0.6, 'p': 0.6923076923076923, 'f': 0.6428571378826532},
'rouge-1': {'r': 0.75, 'p': 0.9230769230769231, 'f': 0.8275862019500595}}]

Table 2 (Hindi)

	Input	Reference Summary	Model Output Summary
0	नई दिल्ली: आज देशभर में रामनवमी का पर्व मनाया जा रहा है। लेकिन इस पावन मौके पर देश में कई जगह जुलूस के दौरान हिंसा की खबरें सामने आई हैं। गुजरात, झारखंड से लेकर पश्चिम बंगाल और मध्य प्रदेश से हिंसा की अलग-अलग घटनाएं सामने आई हैं। जहां एक ओर गुजरात के हिम्मतनगर और आणंद में हिंसक झड़प देखने को मिली है, तो वहीं मध्य प्रदेश के बड़वानी और झारखंड के लोहरदगा और पश्चिम बंगाल के दक्षिण हावड़ा में भी जुलूस पर हमला हुआ।गुजरात में जुलूस पर हमला-साबरकांठा जिले के हिम्मतनगर के छपरिया क्षेत्र में रामनवमी के मौके पर जुलूस	Rama Navami: रामनवमी के मौके पर देश के कई इलाकों में शोभायात्रा पर हमले, जगह-जगह पथराव और आगजनी	देश के इलाकों पर हमले के बाद देखें और फायरिंग
1	नई दिल्ली: जम्मू-कश्मीर में टारगेटेड किलिंग के बीच आज से गृह मंत्री अमित शाह का जम्मू-कश्मीर दौरा शुरू हो रहा है। जम्मू-कश्मीर से आर्टिकल 370 और 35A हटाये जाने के बाद पहली बार अमित शाह श्रीनगर पहुंच रहे हैं। शाह जम्मू-कश्मीर में तीन दिन तक रहेंगे और वहां सुरक्षा इंतजामों के साथ साथ विकास की योजनाओं का भी जायजा लेंगे। गृह मंत्री का ये दौरा ऐसे समय में हो रहा है जब एक ओर पुंछ और राजौरी में आतंकियों के खिलाफ बड़े ऑपरेशन चलाए जा रहे हैं तो दूसरी ओर जम्मू-कश्मीर में आम नागरिकों के खिलाफ आतंकी वारदातें बढ़ गईं	आज कड़ी सुरक्षा के बीच श्रीनगर पहुंचेंगे गृहमंत्री अमित शाह, Article 370 हटने के बाद पहला कश्मीर दौरा	श्रीनगर में आतंकी हमले के बाद गृह मंत्री अमित शाह भी शामिल
2	चामराजनगर. कर्नाटक के चामराजनगर के जिला अस्पताल में ऑक्सीजन की कमी की वजह से बड़ा हादसा हुआ है। यहां पिछले 24 घंटों में ऑक्सीजन की किल्लत होने के कारण 24 मरीजों की मौत हो गई है, इन मरीजों में कोरोना के मरीज भी शामिल हैं। चामराजनगर के जिला अस्पताल में ऑक्सीजन की किल्लत की वजह से हुई मौतों पर जिला प्रभारी मंत्री सुरेश कुमार ने कहा कि हम डेथ ऑडिट रिपोर्ट का इंतजार कर रहे हैं। हादसे पर सीएम येदियुरप्पा ने चामराजनगर के जिलाधिकारी से बात की है और एक ईमरजेंसी कैबिनेट मीटिंग बुलाई है।कांग्रेस ने बोला हमलाचामराजनगर. कर्नाटक के	कर्नाटक के चामराजनगर में ऑक्सीजन की कमी से 24 मरीजों की मौत	कर्नाटक के में ऑक्सीजन की कमी से मरीजों की मौत
3	IndiGo bars specially-abled child: विमानन कंपनी इंडिगो द्वारा रांची एयरपोर्ट पर एक दिव्यांग बच्चे को विमान में सवार होने से रोके जाने के एक दिन बाद नागरिक उड्डयन मंत्री ज्योतिरादित्य सिंधिया ने सोमवार को कहा कि ऐसे बर्ताव को बिल्कुल भी बर्दाश्त नहीं किया जाएगा और किसी भी इंसान को ऐसी स्थिति से गुजरना नहीं पड़े। सिंधिया ने साथ ही कहा कि वह खुद घटना की जांच कर रहे हैं। इसके बाद एयरलाइन के CEO ने माफी मांगी है।बता दें कि इंडिगो ने दिव्यांग बच्चे को रांची एयरपोर्ट पर विमान में सवार होने से रोक दिया क्योंकि वह "घबराया" हुआ	IndiGo bars specially-abled child: इंडिगो ने दिव्यांग बच्चे को फ्लाइट में चढ़ने से रोका, तो भड़क गए सिंधिया, बोले- खुद करूंगा जांच	अब से नहीं किया जाएगा का जहाज अधिकारी से संपर्क करने वाले पर

[{'rouge-1': {'r': 0.3125, 'p': 0.5555555555555556, 'f': 0.399999995392},
'rouge-2': {'r': 0.1111111111111111,
 'p': 0.05882352941176470,
 'f': 0.07692307239644997},
'rouge-1': {'r': 0.3125, 'p': 0.5555555555555556, 'f': 0.399999995392}},
{'rouge-1': {'r': 0.23529411764705882,
 'p': 0.2758620641141499,
 'f': 0.2758620641141499},
'rouge-2': {'r': 0.05882352941176470,
 'p': 0.09090909090909091,
 'f': 0.07142856665816359},
'rouge-1': {'r': 0.17647058823529413, 'p': 0.25, 'f': 0.20689654687277062}},
{'rouge-1': {'r': 0.8181818181818182, 'p': 1.0, 'f': 0.8999999950500001},
'rouge-2': {'r': 0.6363636363636364,
 'p': 0.7777777777777778,
 'f': 0.6999999950500001},
'rouge-1': {'r': 0.8181818181818182, 'p': 1.0, 'f': 0.8999999950500001}},
{'rouge-1': {'r': 0.045454545454545456,
 'p': 0.08333333333333333,
 'f': 0.05882352484429102},
'rouge-2': {'r': 0.0, 'p': 0.0, 'f': 0.0},
'rouge-1': {'r': 0.045454545454545456,
 'p': 0.08333333333333333,
 'f': 0.05882352484429102}},
{'rouge-1': {'r': 0.23529411764705882, 'p': 0.5, 'f': 0.319999995648},
'rouge-2': {'r': 0.05882352941176470,
 'p': 0.14285714285714285,
 'f': 0.08333332920138908},
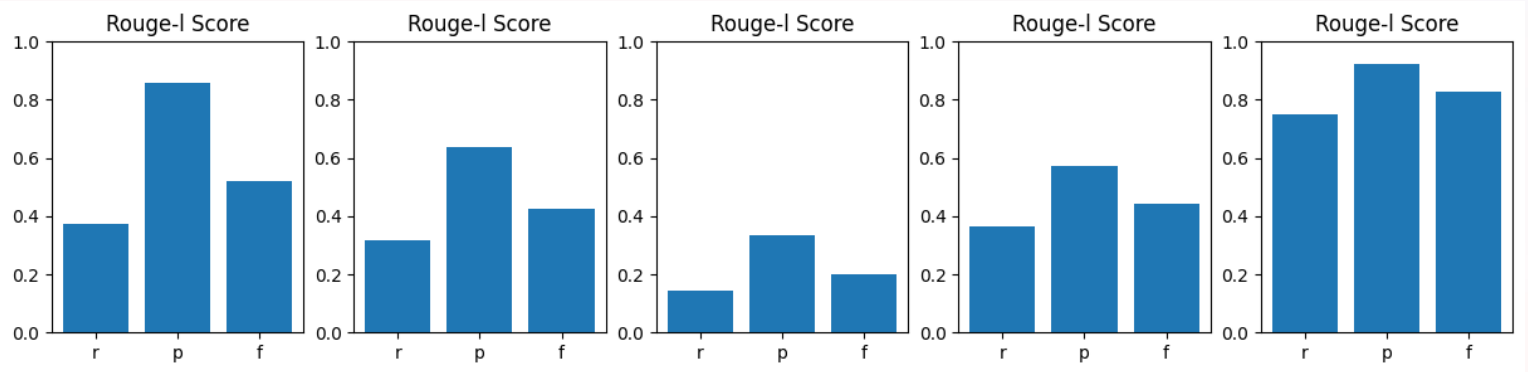'rouge-1': {'r': 0.11764705882352941, 'p': 0.25, 'f': 0.1599999956480001}}]

Hindi Text Summarization Rouge Scores

Gujarati Text Summarization Rouge Scores

ROUGE, or *Recall-Oriented Understudy for Gisting Evaluation*, is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing.

Research Paper 3 Results

ROUGE score is the quotient of the matching words under the total count of words in reference sentence (summurization).

	Metrics	Validation Dataset	Test Dataset
Rouge-1	**F1-Score**	**0.551466**	**0.544284**
	Precision	0.498257	0.489564
	Recall	0.657105	0.652449
Rouge-2	**F1-Score**	0.457683	0.443253
	Precision	0.413388	0.396962
	Recall	0.546676	0.534722
Rouge-3	**F1-Score**	0.431636	0.414829
	Precision	0.389511	0.370719
	Recall	0.517708	0.503161
Rouge-4	**F1-Score**	0.417699	0.399905
	Precision	0.376664	0.356794
	Recall	0.503248	0.487699

Table 1
ROUGE Score

ROUGE, or *Recall-Oriented Understudy for Gisting Evaluation*, is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing.

In its simplest form ROUGE score is the quotient of the matching words under the total count of words in **reference** sentence (summurization). Regarding the denominator ROUGE is a recall oriented metric.

$$ROUGE_1 = \frac{\sum_{unigram \in reference} Count_{match}(unigram)}{\sum_{unigram \in reference} Count(unigram)} = \frac{\sum_{unigram \in reference} Count_{match}(unigram)}{\ell_{ref}^{unigram}}$$

ROUGE-1 example

ROUGE-1 is the ROUGE-N metric applied with unigrams.

Type	Sentence	Length
Reference	The way to make people trustworthy is to trust them.	$\ell_{ref}^{unigram} = 10$
Hypothesis	To make people trustworthy, you need to trust them.	$\ell_{hyp}^{unigram} = 9$

The following illustrates the computation of ROUGE-1 on the summurization sentences

reference ⟶ The way to make people trustworthy is to trust them
$\ell_{ref}^{unigram} = 10$

hypothesis ⟶ To make people trustworthy, you need to trust them
$\ell_{hyp}^{unigram} = 9$
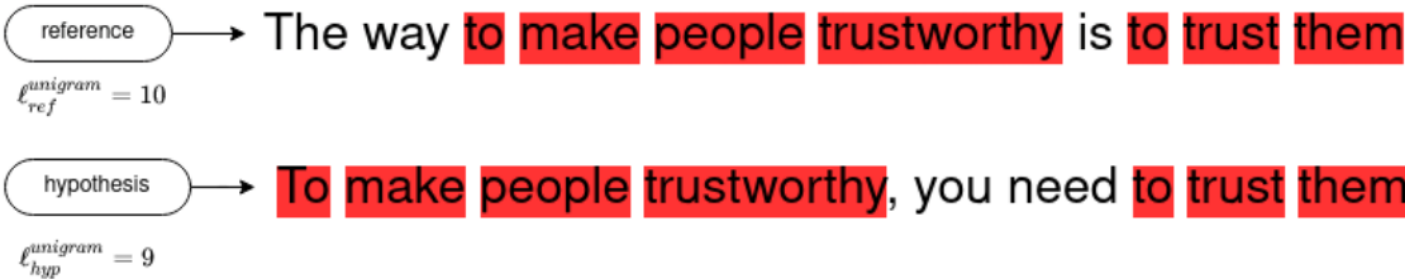
$$ROUGE_1 = \frac{7}{10} = 0.7$$

```
str = [
    "the quick brown animal jumped over the lazy dog"
    "the quick brown fox jumped over the lazy dog"];
references = tokenizedDocument(str)

references =

  2x1 tokenizedDocument:

    9 tokens: the quick brown animal jumped over the lazy dog
    9 tokens: the quick brown fox jumped over the lazy dog
```

Calculate the ROUGE score between the candidate document and the reference documents.

```
score = rougeEvaluationScore(candidate,references)

score = 0.8889
```

Thank You