Building a Text Summarization System in Multilingual Low Resource Settings

Petr Motlicek and Shantipriya Parida Idiap Research Institute Martigny, Switzerland





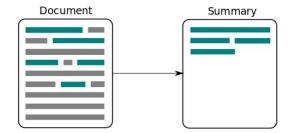
- Low Resource Challenges in Summarization
- Case Studies
 - Case 1 Summarization and Domain Adaptive Retrieval Across Languages (SARAL) - U.S. project
 - Case 2 ROXANNE EU project
 - Case 3 Usage of Synthetic Data for Text Summarization Swiss project
 - Case 4 Usage of OCR for Text Summarization
- Conclusion



Low Resource Challenges in Summarization

1. **Under-resourced languages:** owing to a shortage of quality linguistic data available for many NLP tasks including summarization.

2. Limited annotated summarization data: is difficult to use for abstract text summarization where size matters to train deep learning models.



Annotated summarization data include long text (document/paragraph) and its summary



- Low Resource Challenges in Summarization
- Case Studies
 - Case 1 Summarization and Domain Adaptive Retrieval Across Languages (SARAL) - U.S. project
 - Case 2 ROXANNE EU project
 - Case 3 Usage of Synthetic Data for Text Summarization Swiss project
 - Case 4 Usage of OCR for Text Summarization
- Conclusion



Case Studies

- The case studies include languages having limited or no summarization datasets available for building a summarization system.
- The case studies are based on the languages :
 - German (Indo-European language)
 - Odia (Indo-European language)
 - Tagalog (Austronesian language)
 - Swahili (African language)
 - Somali (Afroasiatic language)
 - Lithuanian (Eastern Baltic language)
 - Bulgarian (Slavic language)



- Low Resource Challenges in Summarization
- Case Studies
 - Case 1 Summarization and Domain Adaptive Retrieval Across Languages (SARAL) - U.S. project
 - Case 2 ROXANNE EU project
 - Case 3 Usage of Synthetic Data for Text Summarization Swiss project
 - Case 4 Usage of OCR for Text Summarization
- Conclusion



Case 1 - SARAL

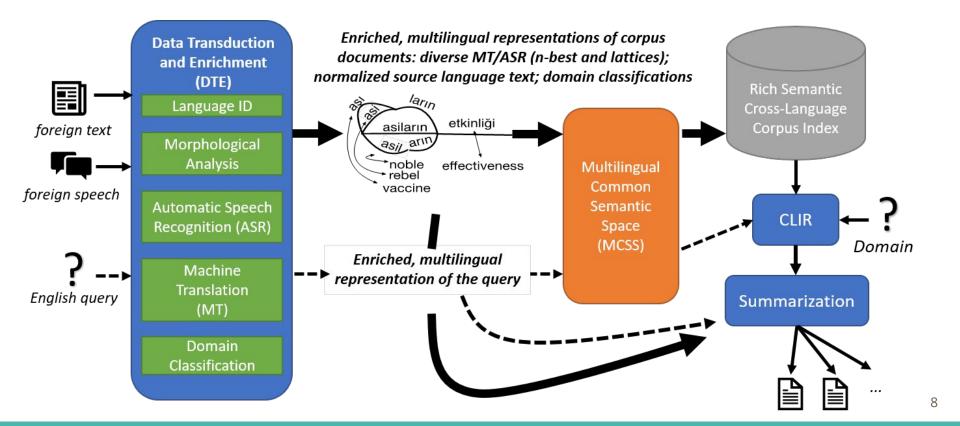
SARAL: Summarization and domain-Adaptive Retrieval Across Languages

सरल (saral): a Hindi word whose MATERIAL-relevant English translations include ingenious and simple



System Overview







Challenges in SARAL

- Limited training data: 30-100h of speech depending on the language
- Training data consists of conversational telephone speech only (8kHz)
- Evaluation data from 3 genres: conversational speech, news broadcast and topical broadcast
- WER affects machine translation and information retrieval
- New language every few months: Tagalog, Swahili, Somali, Lithuanian and Bulgarian
- Language-specific issues
 - Somali doesn't have a standardized spelling
 - Compound words
 - Choosing proper lexicon
- Two main approaches: multilingual training and semi-supervised training



- Low Resource Challenges in Summarization
- Case Studies
 - Case 1 Summarization and Domain Adaptive Retrieval Across Languages (SARAL) - U.S. project
 - Case 2 ROXANNE EU project
 - Case 3 Usage of Synthetic Data for Text Summarization Swiss project
 - Case 4 Usage of OCR for Text Summarization
- Conclusion

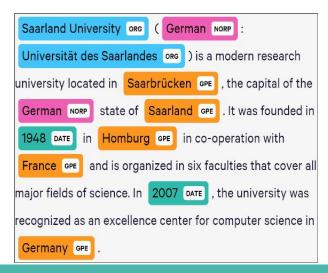
Case 2 - ROXANNE



(Real time netwOrk, teXt, and speaker ANalytics for combating orgaNized crimE)

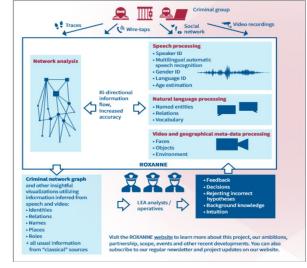
- FCT H2020 project (<u>http://roxanne-euproject.org</u>).
- Various NLP technologies applied (potentially including summarization, entity detection, and topic detection).

Entity Detection



Relations • Names Places . Roles **Topic Detection** + all usual information from "classical" sour manson Currently we are using BERT encodinas. bicycle chotograp pictures 2 vehicle car butter motorcyc' Traditional BoW Bag-of-Concepts







- Low Resource Challenges in Summarization
- Case Studies
 - Case 1 Summarization and Domain Adaptive Retrieval Across
 Languages (SARAL) U.S. project
 - Case 2 ROXANNE EU project
 - Case 3 Usage of Synthetic Data for Text Summarization Swiss project
 - Case 4 Usage of OCR for Text Summarization
- Conclusion



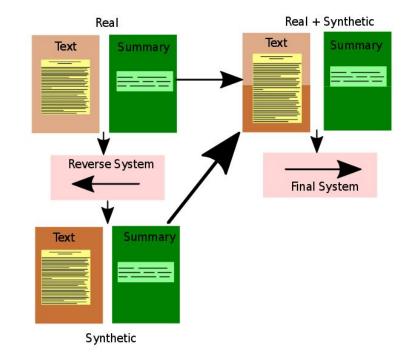
Case 3 - Usage of Synthetic Data for Text Summarization

- Based on Idiap participation in the SwissText 2019 challenge (100'000/2'000) paragraphs and summaries for training/evaluation.
- Use of synthetic data: a popular approach in machine translation for the low resource conditions to improve the quality.
- Can such approaches work for the text summarization task?.



Method

- Use a state-of-the-art "Transformer Model" as implemented in OpenNMT-py.
- Different experiments performed based on real and synthetic data.
- Synthetic data used to increase the size of the training data.
- To generate synthetic data :
 - 1. A system is trained in reverse direction i.e. source as summary and target as text.
 - 2. The reverse system is used to generate text for the given summary. Now, synthetic data is ready.
 - 3. Mix the real and synthetic data and train the final system.



Generation of synthetic data using reverse system.



Dataset

Real data (SwissText dataset)

• Synthetic data (Common Crawl)

- 1. Build Vocabulary (using SwissText dataset, most frequent German words).
- 2. Select sentences based on the prepared Vocabulary. From the selected sentences, randomly choose 100K.
- 3. Generate synthetic data by using 100K sentences to input to the reverse trained model.

| Dataset | #Text | #Summaries |
|-----------------|-------|------------|
| Train | 90K | 90K |
| Dev | 5К | 5К |
| Test | 5K | 5К |
| Test Evaluation | 2К | - |

Statistics of experimental data (real) including the number of text and summaries.

| Dataset | #Text | #Summaries |
|---------|-------|------------|
| Train | 190K | 190K |

Statistics of experimental data (real + synthetic) including the number of text and summaries.



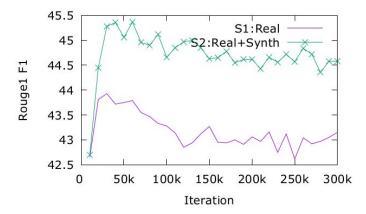
Evaluation

| Setting | Dataset | Rouge_1_F1 | Rouge_2_F1 |
|---------|---------|------------|------------|
| S1 | Dev | 43.9 | 28.5 |
| | Test | 39.7 | 22.9 |
| S2 | Dev | 45.4 | 29.8 |
| | Test | 55.7 | 41.8 |

Evaluation results of our models

| Team | Rouge_1 | Rouge_2 |
|---|---------|---------|
| Shantipriya Parida, and Petr Motlicek (s2) | 40.2 | 22.2 |
| Dmitrii Aksenov, Georg Rehm, Julian Moreno Schneider | 40.4 | 21.9 |
| Nikola Nikolov | 34.7 | 19.3 |
| Valentin Venzin, Jan Deriu, Didier Orel, Mark Cieliebak | 39.8 | 23.4 |
| Pascal Fecht | 40.9 | 23.5 |

SwissText 2019 Text Summarization Challenge Result Source: http://ceur-ws.org/Vol-2458/summarization_challenge.pdf



Learning curves in terms of Rouge 1 F1 Score on dev set

• Evaluations made using Rouge (Recall-Oriented Understudy for Gisting Evaluation) score, a popular metric for text summarization.



- Low Resource Challenges in Summarization
- Case Studies
 - Case 1 Summarization and Domain Adaptive Retrieval Across Languages (SARAL) - U.S. project
 - Case 2 ROXANNE EU project
 - Case 3 Usage of Synthetic Data for Text Summarization Swiss project
 - Case 4 Usage of OCR for Text Summarization
- Conclusion



Case 4 - Usage of OCR for Text Summarization

• Odia is categorized as a classical Indian language.

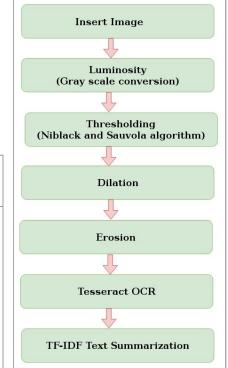
• Although Odia language has a rich cultural heritage, this is not digitized or accessible, resulting in a lack of web resources.

• How to build a summarization system for such low-resource language ?.

Model

- Odia language text extracted from the image files • using Tesseract optical character recognition (OCR).
- Summarize the obtained text using extractive summarization techniques.

| Original Image (Odia language text) | Extracted Odia Text | Summarized Odia Text | Dilation |
|---|---|---|---------------------------|
| | | | |
| ଦେଇଞିକ ମହାନ୍ତିକର ପୂର୍ବ ନମା 'ଦେଇଞିକ ରହର୍ବିକ ମହାନ୍ତି। ସେ ୧୯୭୬ ମରିସା କୁରାଜ ୬୦୮୨) ଗରିଶରେ ତିନା ଉତନ୍ତ୍ୱର ମହାନ୍ତି ଓ ମାନ ମହାକା ମହାହିବ ପାନୁ ଭବନେକ୍ରପାରେ ବହୁ ହୁଧକ ବସିହିରେ । ତାବର ତିନି ଜଣ ଭେଣା ମଧ୍ୟ ଅନ୍ତି । ଶିଗରତ ଯୋକନା ମୁହରଣା, ସେ ଜଣି କାତନ । ସେ ସାରସରୁର ଅନ୍ତମାନାଙ୍କ, ମେହାନ ମହିନ୍ଦୁ ବିଜେଗ ଜେବିମ ଅପର କରିଥିରେ । ପରେବି (କରର ସେ ଭେତ ବନ୍ତମାକ ମଇଜନରେ ଗର୍ଯାରଣ । ବେମ୍ବିଷ ନହାନ୍ତି, ୧୯୧୧ ମହିନ୍ଦୁ ବିଜେଗ ଜେବିମ ଅପର କରିଥିରେ । ପରେବି (କରର ସେ ଭେଁ ସେ ଜଣି କାତନ କରି ସ୍ୱାର୍ଚ୍ଚ ଦେନାର ସନୁସାସାଦରୁ କେତନ ପିନ୍ତାର ହୁଏକ ସେଥିରେ । ପରେବି (କରରରେ ସେ ଭିନିତ ସେ ଜଣି ବିଜେଗ, ଭାରତସା ବଳ ସାହୁଣ କରି ବିଜେଗ ଜଣିକା ସାହର କରିଥିରେ । ପରେବି (କରରରେ ସେ ଭିନିତ ସାଜିଙ୍କର ପାର୍ଚ୍ଚ ଭୋରତମ କରି ସ୍ୱାର୍ଚ୍ଚ ଦେନାର ସନୁସାସାରେ (କେତନ ପିନ୍ତର ଜଣିକା ସାରେକରେ ପ୍ରକାର କରେ । "ଏ କେତନ ପିକ୍ରର ଭାରତ ସେ ସେ ସାହର ଜଣେ କରି ପୁରସ କରିର ସେ ସେମିର ସେ ମିହି କେସେରେ ସେ ସେ ସେ ପାରେ କରିଥିରେ । ସାହରେ ପାରସାରେ ସେ ସେ ସେ ସେ ସୋଇଟିଲେ ମା ଏହି ପ୍ରକତ ପରେ, ସେ କରେ ଜଣ ପୁରସ କରିସ ହେର ପିଦରରେ ପରସନ ସେଯିରେ । ଭାରତୀ ସେ ଅପରେ ସେ ସେ ପାରେ ସେଇଟିଲେ । ମା ଏହି ସହରେ ପରେ କରେ କୋଳ ସୁରସିନ କରି ସେରେମେ ସିହାନ ପରସନ କୋଇନିକ କିନ୍ତ ମହିକ କେକରେ ସେହାର ସେଇଟିଲେ । କରି ପରସନ ଭାନର ସହାନ ସହାନ କରି କରେ ସେରେମିରା । ସେ ମହାର ସେନିକରେ ସେ କାନାନେରେ ସହ ସେଇଟିଲି । ୧୯୧୭ରେ ସେ ସେ ପରେ କେତ କେହାର ସହାର ପରସନେ ଅପନେ ପରସନେ ସେ ସେନେ ପରସନେ କେଇ କେକରେ । କନେରେ ହେ ସେ ସେ କେରେ ସେହିନିଲା । ୧୯୧୭ରେ ସେ ସେ ପରେ କେତ କେହାର ସେନ ସେ ସେହାର ପରସନ ସେ ସେନ ପରସନେ କେକ ସହାନ ସେ ସେ କେନେରେ ସେହିନିଲା । ୧୯୧୭ରେ ସେ ସେ ସହାନ ସହାର ସେହାର ପରସନ ସେ ସେ ସେହାର ସେ ସେନେ ସେ ସେନ କେଇ କେନିନାରେ କେ । କେନେ କେରେ କେ କେରେ ସେହିନିଲା । ୧୯୧୭ରେ ସେ ସେ ସହାର ସେହାର ସେନ ସେ ସେ ସେହାରେ ସେନେ କେ ସେ ସେ ସେ ସେ ସେହାରେ ସେ । କେନେ କେ ସେ ସେହା ସେ ସେନେ ସେକ କେ କେ ସେ ସେନେ ସେକ କେ ସେହିରେ କେ ସାର କେ ସେହାର ସେ । କିତନ କେରେ ସେ କେ କେ କେ ସେକରେ ସେହା ସେକ ରେ ସେହାର ସେହାର ସେ ସହର କେ ସେହାର କେ ସେ କେ କେ କରେ କେ କେ କେ କେ କେ ସହର ସେ ସହ ସେରି ସେ ସହ ଅନେ କ କରେ ପରେ କରେ ସେହାର ସେହାର ସେ ସେକ ସେକ ସେ ସେହର ସେ ସେହାର ସେହାର ସେହାର ସେ ସେ ସେ ସେ ସେହାର ସେହା ସେହାର ସେହାର ସେହାର ସେହାର ସେହାର ସେହାର ସେ ସେହାର ସେହାର ସେହାର ସେହାର ସେହାର ସେହାର ସେହା ସେହାର ସେହା ସେହାର ସହା । । କରି ସେହାର ସେ | ් අධ්යක් 1 රට ශාල්පත්වන් ධර්ෂාක්ෂණ වන්නා සංක්ෂ යට ය. හැදගවරයා වන්නේදී දේශය ප්රිස්තුවෙන්දී දේශය පරියියාදී (11) යෙ ප්රරාහන කියන්නෙ කරේවුන් වෙන්න පතියුද්ද අධ්යක් දේශය පරියියාදී දේශය ප්රිස්තාන් දේශය ප්රිස්තාන් දේශය ප්රිස්තාන් දේශය කොහොස කර කත්වේ කොහොඩ කරේදීමක් පාර්තාන් ප්රිස්තාන් දේශය ප්රිස්තාන් කරීම කර්දා දේශය කොහොස කර කත්වේ කෙන සාධාන්ත කරීමක් සංක්ෂ කර්දා කොහොස දේශය ක්රීමක් දේශය ප්රිස්තාන් ප්රිස්තාන් ප්රිස්තාන්ත් ප්රිස්තාන් දේශය ප්රිස්තාන්ත කරන්නේ කරන්නේ කරන්න කරීමක් ප්රිස්තාන්ත් ප්රිස්තාන්ත කරීමක් කරන්නේ කරන්නේ කරන්නේ ප්රිස්තාන්ත කරන්නේ කරන්න කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ ප්රිස්තාන්ත කරන්නේ කරන්නේ ප්රිස්තාන්ත කරන්නේ කරන්නේ කරන්නේ කරන්නේ ප්රිස්තාන්ත කරන්නේ කරන්නේ ප්රිස්තාන්ත කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ ප්රිස්තාන්ත කරන්නේ කරන්නේ ප්රිස්තාන්ත කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ ප්රිස්තාන්ත කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ ප්රිස්තාන්ත කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ ප්රිස්තාන්තාන්තාන්ත කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ ප්රිස්තාන්ත කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ ප්රිස්තාන්තාන්ත කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ ප්රිස්තාන්තාන්තාන්තාන්තය කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කත්තාන්ත් කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කත්තාන්ත් කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කත්තාන්ත් කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කත්තාන්ත් කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කත්තාන්ත් කරන්නේ කරන්නේ කත්තාන්ත් කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කත්තාන්ත් කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කත්තාන්ත් කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කත්තාන්ත් කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කරන්නේ කත්තාන්ත් කරන්නේ කරන්නේ කත්තාන්ත් ක | ହୋଇଥିଲେ ୧୯୯୬ରେ ହିଁ ସେ ସାକିସ୍ଥାନ ବିରେଧରେ ବେଭଣ୍ଟମେଠାରେ ସ୍ରରଥମ ସିନିକିସ ମୁହାତ ଖେଳିଥିଲେ ଏହି ମ୍ୟାରତେ ସେ ୬ ଓଉରତେ ୨୨ ରହି ସେଇ ଗୋଟିଏ ଭରକରେ କେମରେ ସଫଳ ହୋଇଥିଲେ ଗଙ୍କ ନିର ମରରେ , ଭାରସୀୟ କ୍ରିଭିକେଟ ଦଳ ପାଇଁ କୋଣ୍ଡିସା ଓ ବିଷ୍କଳପରେ ଅଙ୍କରେଲୀୟ ଅଧିଲୟକ କ୍ରିସିର ସଙ୍କର ଭଇକେଟ ନେଶ ସହାଜ କ୍ରିଭିକେଟ ସିନେର ଦୁଇଥି ବିନ୍ମାରଣୀୟ ସାରଖା (୮) ସେ ପୁର୍ସାଞ୍ଚଳ କଳର ସାଦଶିକଳ କେସ ମଧ୍ୟର କା କାର୍ଯ୍ୟ କରି ସିଯ କରି ଏହା କମସା ଦେସଟି କିଲା ଭାଜିଲ <i>ି</i> | Erosion |
| | | | |
| | | | Tesseract OCR |
| | | | |
| | | | TF-IDF Text Summarization |
| | ଓଡ଼େଇ ଦେଇବେ ନେବା ସହକୁ କରିକେଟ ଗନେର ସୁରନ୍ଧି ସ୍ୱମର୍ଭଣୟ ସମନ୍ଧା 'ଦେମନା ନେସ ପରିଭ ଭଣଣ ବଳର ପରିଗିବେଳ ଭୁରେ ସାଯିତକ । (୯) ସେ ପୁରିଞ୍ଚେଳ ବଳର ପରିଗିବ୍ଞକ ଜୁସେ ମଧ୍ୟ କା କାର୍ଯ୍ୟ କରି ଯିଣ୍ଟ କରି ଏହାକୁ ଦୁରୀସ ଦ୍ୱରଫି ଟିସର ଭରିଛନ୍ତି । | | Block diagram of the text |



summarization process





| Step 1 : Suppose the letter in Odia (ଲ) | $ \begin{array}{c ccccccccccccccccccccccccccccccccccc$ |
|---|--|
| Step 2: Missing pixels during capturing or thresholding. | 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 0 1 0 1 0 1 0 |
| This time if Tesseract operates then it may detect (ה .) | |
| Step 3: Dilation operation performed for pixel expansion. | 0 1 |
| Step 4: To obtain the accurate shape, it needs to sink pixels so, Erosion performed. | $ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0$ |

Result (Human Evaluation)



- A manual evaluation performed to evaluate generated summaries.
- Odia extracted text about different eminent persons along with generated summaries are provided to four experts.
- Experts have gone through the prepared evaluation criteria and provided the results in the scale of **[1-100].**

•

| Parameter | Description | | | |
|-------------|---|--|--|--|
| Parameter 1 | Is the summary related to the given topic ?. | | | |
| Parameter 2 | Is the summary contain relevant information about the person mentioned in the text ?. | | | |
| Parameter 3 | Are the Bag of words, in summary, providing a relatable meaning ?. | | | |
| Parameter 4 | Is the summary length enough ?. | | | |
| Parameter 5 | The Overall quality of the output. | | | |

Human Evaluation Parameter

| Human Evaluator | Topic Name (in English) | Parameter 1 | Parameter 2 | Parameter 3 | Parameter 4 | Parameter 5 |
|--------------------|----------------------------|-------------|-------------|-------------|-------------|-------------|
| Evaluator1 | Navin Pattnaik | 100% | 80% | 55% | 55% | Good (75%) |



- Low Resource Challenges in Summarization
- Case Studies
 - Case 1 Summarization and Domain Adaptive Retrieval Across
 Languages (SARAL) U.S. project
 - Case 2 ROXANNE EU project
 - Case 3 Usage of Synthetic Data for Text Summarization Swiss project
 - Case 4 Usage of OCR for Text Summarization
- Conclusion



Conclusion

- Usage of synthetic data for the abstract text summarization under low resource condition found effective.
- In the case of low resource settings and lack/unavailability of online content, OCR based techniques for text summarization can be useful. And, it also useful to generate summarization data to train deep learning models.



Any Questions ?



References

[1] Parida, S., & Motlicek, P. (2019, November). **Abstract Text Summarization: A Low Resource Challenge**. 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**) (pp. 5996-6000).

[2] Parida, S., & Motlicek, P. (2019). Idiap Abstract Text Summarization System for German Text Summarization Task. In *Proceedings of the 4th edition of the Swiss Text Analytics Conference*.

[3] Pattnaik, P., Mallick, D. K., Parida, S., & Dash, S. R. (2019, December). **Extractive Odia Text Summarization System: An OCR Based Approach**. In *International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making* (pp. 136-143). **Springer**, Cham.

