Building a Text Summarization System in Multilingual Low Resource Settings

Petr Motlicek and Shantipriya Parida
Idiap Research Institute
Martigny, Switzerland

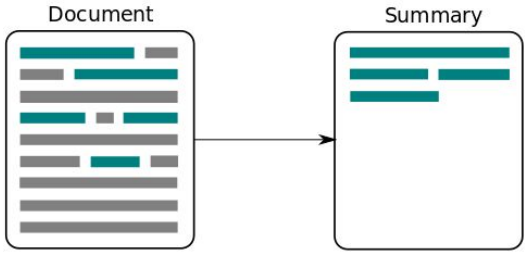
Agenda

- **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

Agenda

- 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)

Agenda

- 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

Information Sciences Institute
USC School of Engineering

idiap RESEARCH INSTITUTE

Massachusetts Institute of Technology

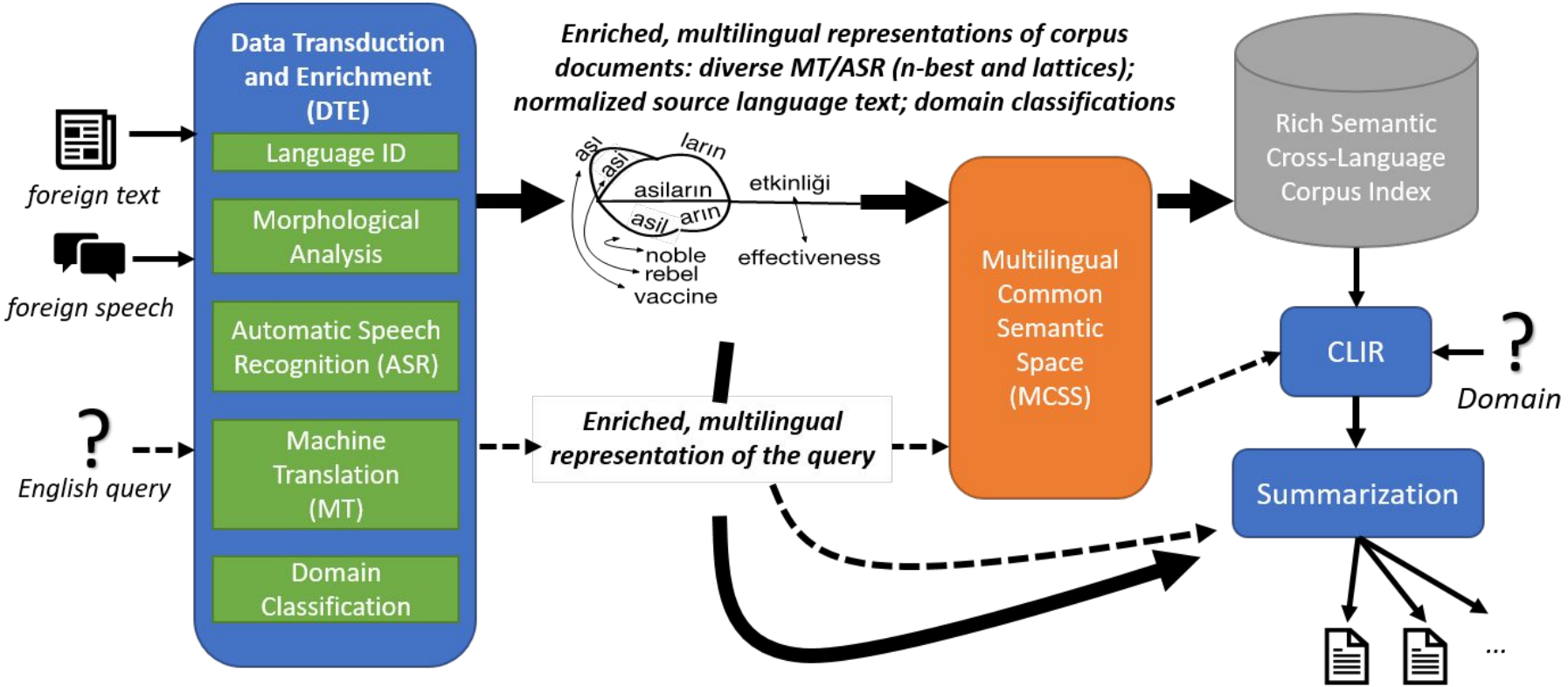
Northeastern

UNIVERSITY OF NOTRE DAME

Rensselaer

UMASS AMHERST

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

Agenda

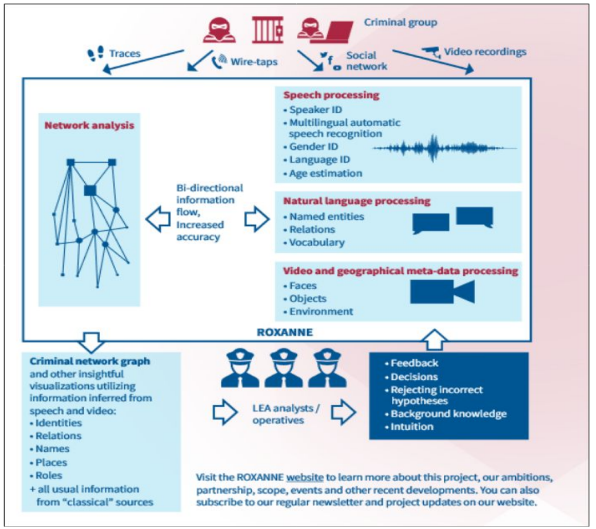
- Low Resource Challenges in Summarization
- Case Studies
  - Case 1 - Summarization and Domain Adaptive Retrieval Across Languages (SARAL) - U.S. project
  - **Case 2 - ROXANNE - EU project**
  - Case 3 - Usage of Synthetic Data for Text Summarization - Swiss project
  - Case 4 - Usage of OCR for Text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

Case 2 - ROXANNE

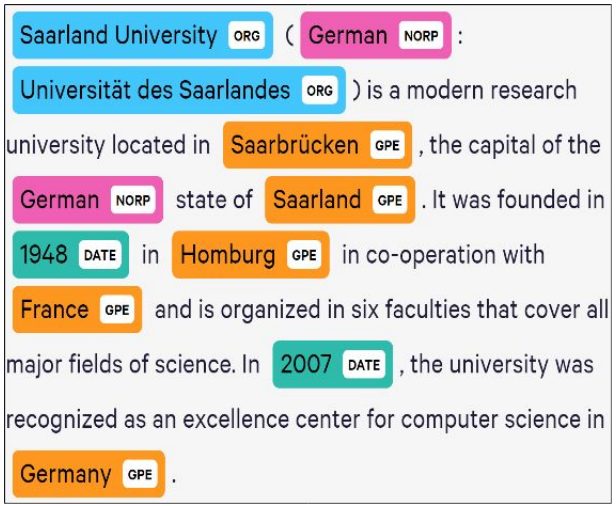
(Real time netwOrk, teXt, and speaker ANalytics for combating orgaNized crimE)

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

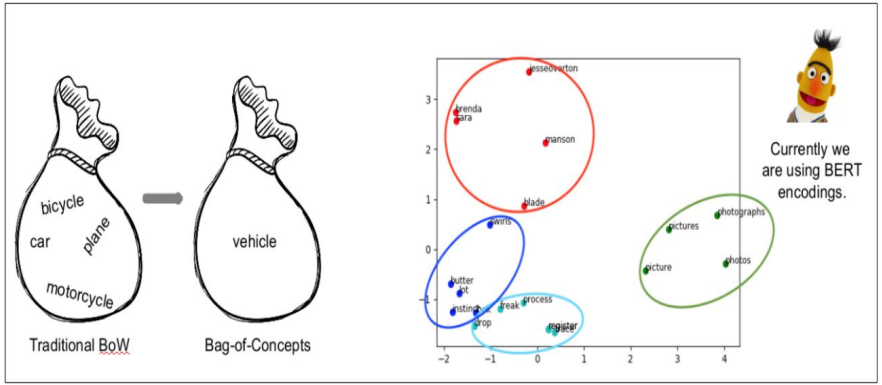
Project Overview



Entity Detection



Topic Detection



Agenda

- 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

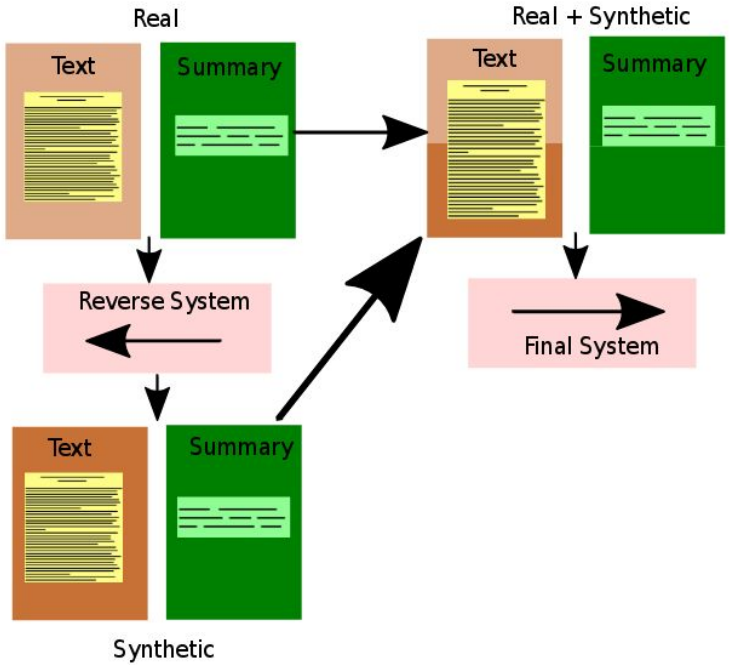
- Based on Idiap participation in the SwissText 2019 challenge (100’000/2’000) paragraphs and summaries for training/evaluation.

- *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)

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

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

- 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.

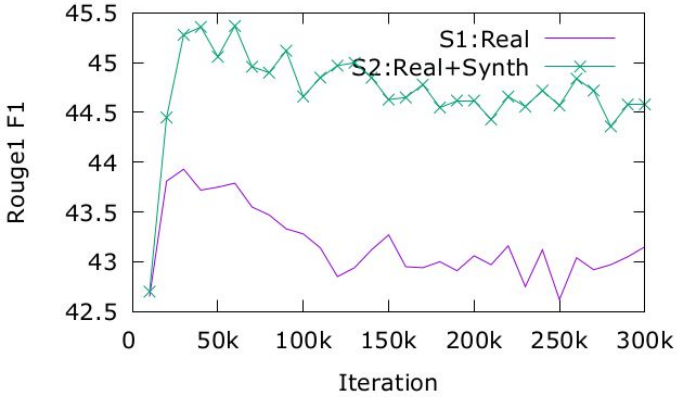
Dataset	#Text	#Summaries
Train	190K	190K

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



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

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

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

Agenda

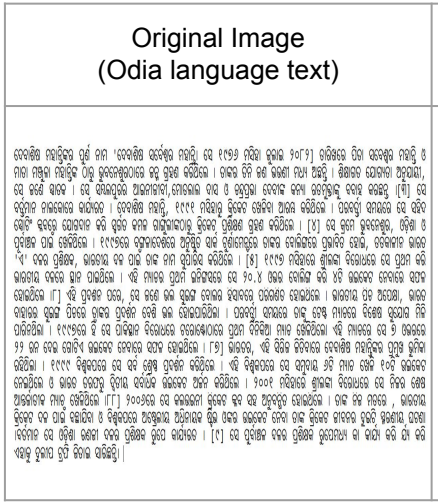
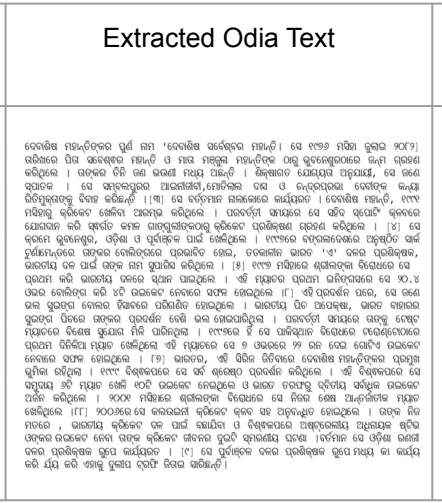
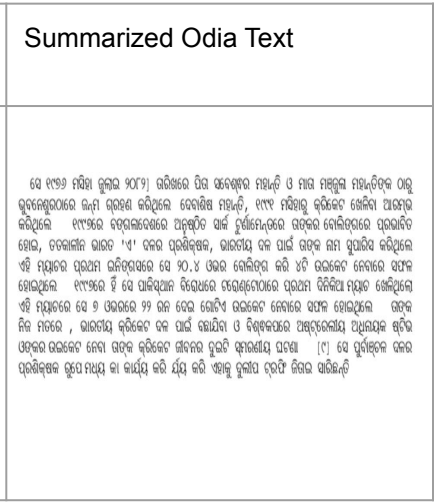
- Low Resource Challenges in Summarization
- Case Studies
  - Case 1 - Summarization and Domain Adaptive Retrieval Across Languages (SARAL) - U.S. project
  - Case 2 - ROXANNE - EU project
  - Case 3 - Usage of Synthetic Data for Text Summarization - Swiss project
  - **Case 4 - Usage of OCR for Text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

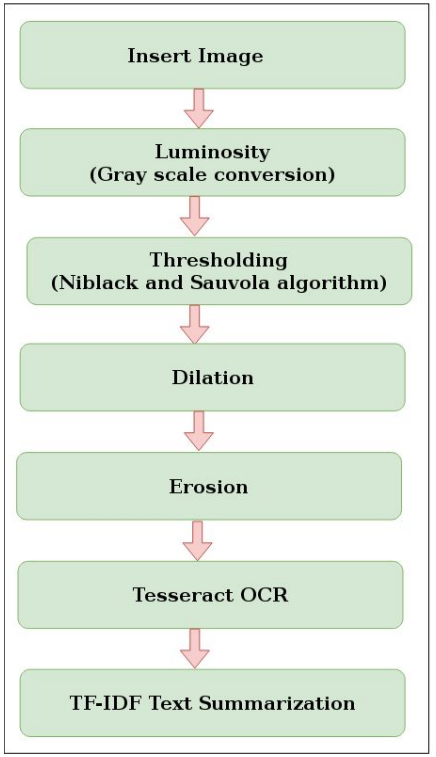
Case 4 - Usage of OCR for Text Summarization

- 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

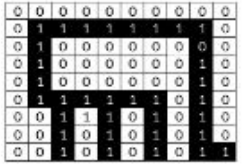
- Odia language text extracted from the image files using Tesseract optical character recognition (OCR).
- Summarize the obtained text using extractive summarization techniques.

Original Image (Odia language text)	Extracted Odia Text	Summarized Odia Text
		



Block diagram of the text summarization process

Example

Step 1 : Suppose the letter in Odia (ଈ)	
Step 2: Missing pixels during capturing or thresholding. This time if Tesseract operates then it may detect (ଈ.)	
Step 3: Dilation operation performed for pixel expansion.	
Step 4: To obtain the accurate shape, it needs to sink pixels so, Erosion performed.	

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%)

Human Evaluation Rating Table

Agenda

- 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.


idiap
Thank You