Exploration of Cross-lingual Summarization for Kannada-English Language Pair

Vinayaka R Kamath **Rachana Aithal K R** **K Vennela** **Mamatha HR**
Department of Computer Science & Engineering
PES University, India
vinayakarkamath@pesu.pes.edu
{rachanaaithal88, kvennela1998}@gmail.com
mamathahr@pes.edu

Abstract

Cross-lingual summarization(CLS) is a process in which given a document in source language aims at generating summary in a different, destination language. Low resource languages like Kannada greatly benefit from such systems because they help in delivering a concise representation of the same information in a different popular language. We propose a novel dataset generation pipeline and a first of its kind dataset that will aid in CLS for both English-Kannada and Kannada-English pair. This work is also an attempt to inspect the existing systems and extend them to the Kannada-English language pair using our dataset.

1 Introduction

With the advancement in technology, language should not be a barrier to gain knowledge when everyone has access to the same information. The need for an intelligent system that understands and analyzes text in low resource languages while delivering concise representation of the text in a well known language without losing out on any information is paramount. Cross-lingual summarization systems fit these requirements perfectly. For a given document in the source language, the primary objective of the system is to generate meaningful summary in the target language (different from source language) without discarding any crucial information. These systems extend resources available in low resources languages like Kannada to everyone who can understand a well known language such as English.

Monolingual summarization is extensively studied due to the availability of resources while cross-lingual summarization systems are relatively unpopular due to the lack of training corpus. To tackle this, we present a one of it's kind dataset for training a CLS system for Kannada-English language pair along with our experimentation. A robust pipeline for the generation of dataset is designed using round trip translation strategy from (Zhu et al., 2019) and a back translation strategy proposed by (Duan et al., 2019) using the Newsroom dataset from (Grusky et al., 2018) as our primary backbone. We have successfully extended several methods from (Wan, 2011), (Jhaveri et al., 2019) and some baselines from (Shen et al., 2018) to Kannada-English language pair using our dataset. Section 2 describes the dataset generation pipeline and the experimentation carried out. The results and inferences are discussed in section 3 while the conclusions are briefed in section 4.

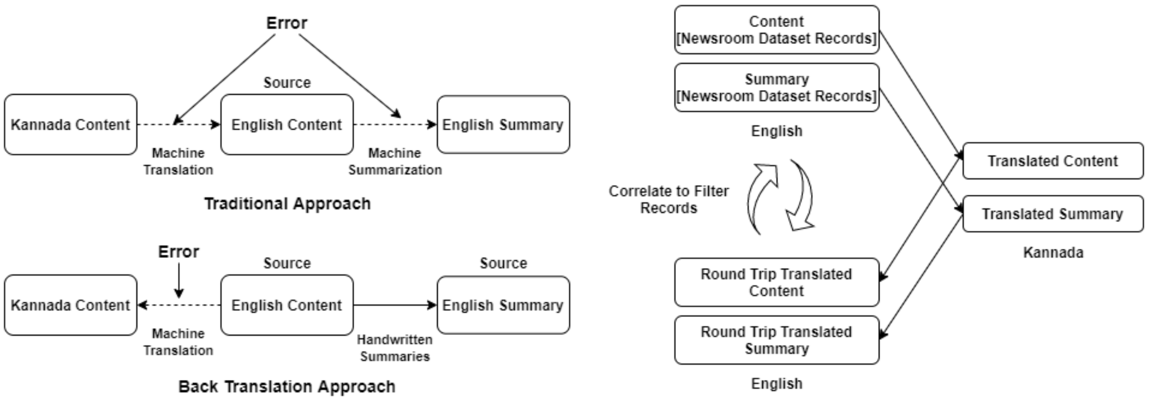
2 Experimentation

2.1 Dataset Construction

This section describes the methods used to construct the very first Kannada-English summarization dataset. The absence of a CLS corpus for Kannada-English language pair is a significant hindrance. To overcome this, we propose a novel pipeline and a newly constructed high quality dataset that will aid in CLS for Kannada-English language pair.

The pipelined process of translation followed by summarization of the content in source language introduces a lot of noise in the generated data. Moreover, the process thoroughly utilises a third party ”automated machine translation system”. To elevate the quality, back translation strategy was implemented using the English content as pivot as specified in Fig.1a. This ensured that there was little dependency on third party systems and minimal noise in the generated data. However, this method can only be successful if a good quality source dataset is used. Cornell newsroom is one such backbone that we used while building our dataset. The Cornell Newsroom dataset(Grusky et al., 2018) is a large monolingual dataset for training and evaluating summarization systems. It con-

Proceedings of the 17th International Conference on Natural Language Processing, pages 144–148
Patna, India, December 18 - 21, 2020. ©2020 NLP Association of India (NLPAI)



(a) Translation Approach Comparison

(b) Round Trip Translation Procedure

Figure 1: Illustrations of methods used to create the dataset

tains 1.3 million articles and summaries written by authors and editors in the newsrooms of 38 major publications.

Filtering noisy outliers is crucial to ensure highest quality possible. The Round Trip Translation(RTT) mechanism described in (Zhu et al., 2019) is used to achieve this. The RTT strategy is used to acquire high-quality large-scale cross lingual summarization dataset from existing large-scale monolingual dataset(Fig.1b). It can be observed that the current monolingual translators are not very proficient. This may result in addition of considerable amount of noise in the dataset if it is constructed by direct translation. Therefore, to improve the quality of parallel corpus the RTT mechanism is employed. This involves calculation rouge score between the reference content and hypothesis content which we obtained after round trip translation. This is proceeded by the filtering of the dirty samples by choosing samples above a threshold value. This ensures that the dataset is reliable and efficient. The threshold is decided by manually sampling the records of the dataset at different values of the rouge scores and visually inspecting the quality of the record. The records are assessed for sentence completion, preservation of semantic meaning and external noise. A suitable threshold is which acts as a cutoff value for noisy records. A heuristic threshold is chosen by trading off the number of records in the dataset to the peak quality of the records. The same procedure is followed to generate summaries(including RTT summaries) as well, this is to make sure that the dataset maps both English content to it's corresponding Kannada summary as well as the Kannada content to it's English summary. This ensures that the dataset is capable of aiding in training both Kannada to English as well as English to Kannada CLS systems.

As a result of our proposed system, we constructed a high quality dataset of 23,113 records that supports interchangeability of source-target languages. The whole dataset is made publicly available along with the rouge scores to filter the records as per the requirements of the application.

2.2 Cross Lingual Summarization Systems

Extending the current state-of-the-art CLS methodologies to the regional language of Kannada can accelerate the process of designing a robust system that can be used to generate good quality summaries for the content in Kannada. This section briefly describes the methods extended to Kannada as illustrated in Figure 2.

2.2.1 Baselines

Early translation systems are used as baselines. During early translation, the process of translation and summarization are stacked in that order to form a simple cross-lingual summarizer. This system relies on the good quality translators and summarizers available for a high resource language like English. In our work, four mono-lingual summarization algorithms namely LSA, LexRank, Luhn and TextRank are used to extract summaries from the translated documents.

Latent Semantic Analysis (LSA) (Steinberger and Jezek, 2004) is an algebraic-statistical method that extracts hidden semantic structure of words and sentences. It relies on Term Frequency-Inverse Document Frequency (TF-IDF) and Singular Value Decomposition (SVD) to achieve summarization

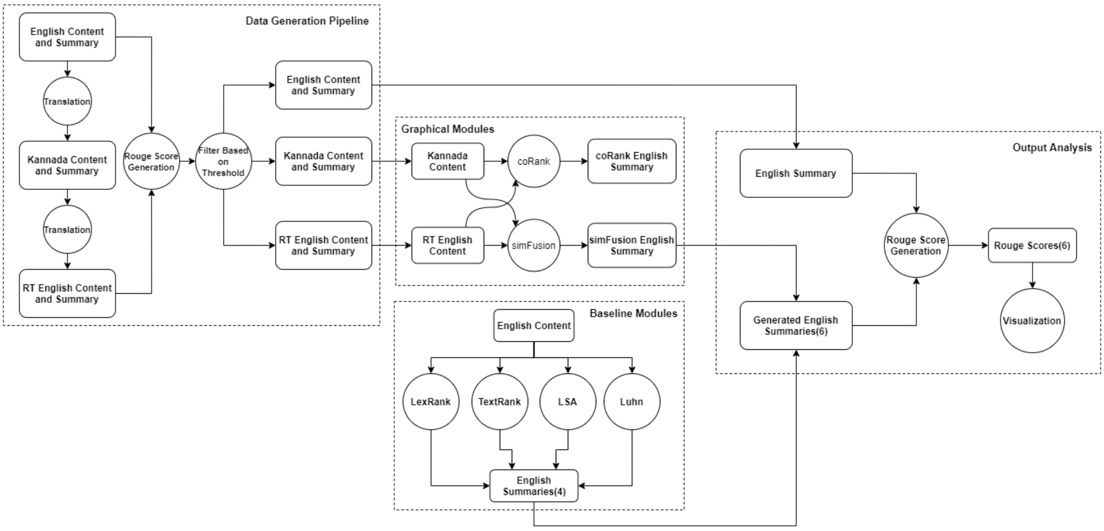


Figure 2: Schematic of the complete experimentation stack

of the text. The matrix constructed using TF-IDF is subjected to decomposition using SVD, there after the topic method is used to extract concepts and sub-concepts to select important sentences. These selected sentences are ranked and presented as a concise summary. LexRank (Erkan and Radev, 2004) is an unsupervised approach to text summarization based on graph-based centrality scoring of sentences. The algorithm recommends sentences that are very similar to the others in the document, thus curbing redundant information in the output. Luhn algorithm (Torres-Moreno, 2014) takes a naive approach based on TF-IDF that concentrates on the window size of non-important words between words of high importance. The summary is generated by assigning weights to the words and recommending sentences with maximum weight values. TextRank (Mihalcea and Tarau, 2004) is a graph based sentence ranking algorithm that uses PageRank algorithm to attain convergence. The algorithm is very similar to LexRank but uses simpler methods to accelerate the computation.

2.2.2 Extended Models

This section describes the sub-modular function maximization based summarization algorithms that were adopted and used for cross-lingual settings. coRank(Jhaveri et al., 2019) and simFusion(Xi et al., 2005) were extended to Kannada-English language pair and a thorough analysis of the results was performed to understand the semantic suitability of the techniques.

coRank method leverages both the Kannada lexicon information and the English-side information in a co-ranking way. The source Kannada sentences and the translated English sentences are simultaneously ranked in a unified graph-based algorithm. The saliency of each Kannada sentence relies not only on the Kannada sentences linked with it, but also the corresponding English sentences associated with it and the same holds true for English sentences. simFusion algorithm uses the Kannada side information for English sentence ranking in the graph-based framework. The similarity value between two English sentences is computed by linearly fusing the similarity value between the corresponding two Kannada sentences with its very own. The graph is constructed using the similarity in both the source and the target languages. In both the methods, the sentences with the highest saliency scores are compiled together to give the summary in the target language.

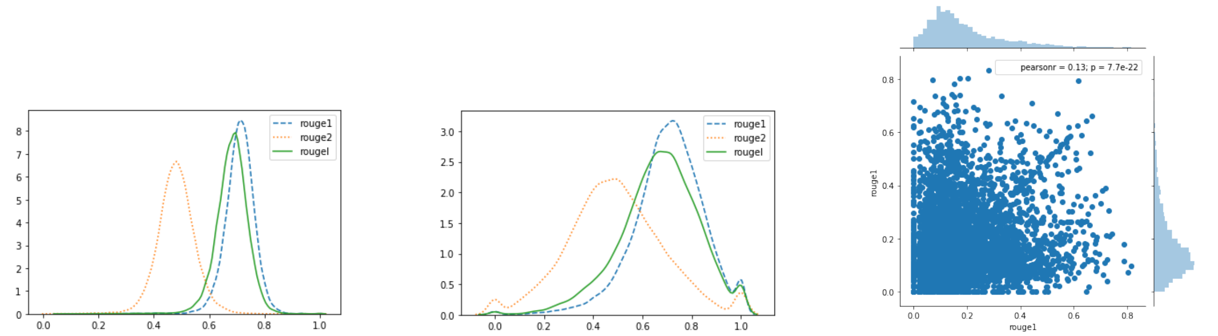
3 Results

The rouge scores of the round trip translated corpus with the records from the newsroom dataset were recorded. This was done in order to inspect the records and set an appropriate threshold to filter out the noisy records. Removing these unwanted records helped in increasing the quality of the dataset. The same observation was recorded for the summaries as well. This enabled the dataset to be more flexible with the interchangeability of the source and destination languages.

The threshold was decided after inspecting the distribution of the rouge scores between round trip translated data and the original summary. Our ex-

	Algorithm	Rouge 1	Rouge 2	Rouge l
CLS Baselines	LSA	17.911	5.611	12.4041
	LexRank	19.6429	6.2439	14.2806
	Luhn	19.1805	6.2301	13.8486
	TextRank	19.3113	6.2806	13.5528
CLS Graphical	coRank	18.7916	6.3136	13.1178
	simFusion	18.3779	6.0298	12.8286
Popular Summarization Systems on Newsroom	Seq2Seq + Attention (Rush et al., 2015)	5.99	0.37	5.41
	Fast-RL (Keneshloo et al., 2019)	21.93	9.37	19.61
	ExtConSumm (Mendes et al., 2019)	39.40	27.80	36.20
	Modified P-G (Shi et al., 2019)	39.91	28.38	36.87

Table 1: Results from Experimentation.



(a) Distribution of Content Rouge Scores

(b) Distribution of Summary Rouge Scores

(c) Correlation between coRank and simFusion Results

Figure 3: Analysis of the results

perimentation of manual inspection by evenly sampling 30% of the records by 3 volunteers yielded 0.24 with variance of 0.03, the next stages of the experiments were carried forward with this threshold in mind. Figures 3a and 3b helped in filtering the records by providing the overview of the distribution. The attempt to check for redundancy in the outputs between the results exhibited by the simFusion and coRank algorithms were done. Although both of them follow a graphical approach, Figure 3c proved that the information captured by these algorithms were quite different. There was very little correlation between the scores of two algorithms for the same set of records, this implies that an ensemble of both these models can provide maximum outreach.

The rouge scores achieved from the experimentation with the different CLS methods are depicted in Table 1. The set of CLS experiments were conducted with Kannada as the source language and English as the destination language for the task at hand. These results are compared among themselves as well as other popular summarization systems on Newsroom dataset, since a cross-lingual summarization system for Kannada-English pair does not exist yet.

4 Conclusion

In this paper, we proposed a first of its kind dataset that consists of content-summary mappings for the Kannada-English language pair. Since Kannada is a low resource language, the dataset can aid for further experimentation on cross lingual applications. The newly designed dataset generation pipeline has also been proven to generate high quality records considering the CLS methods that has been successfully extended to Kannada-English language pair using the dataset. Table 1 shows that the results from our experiments are comparable to that of those that have used the same corpus for designing systems. These results can act as a solid foundation for further exploration. The results from the Table 1 also act as a proof of correctness for the experimental setup. We believe that the first set of CLS experiments for Kannada presented in this paper has set reasonable benchmarks with adequate resources to carry forward computational linguistic experiments for a low resource language. We intend to design/implement systems that use the translated content along with the source content to perform better at CLS tasks centered around Kannada as a part of our future work.

References

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Max Grusky, Mor Naaman, and Yoav Artzi. 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*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Nisarg Jhaveri, Manish Gupta, and Vasudeva Varma. 2019. clstk: The cross-lingual summarization toolkit. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 766–769.

Yaser Keneshloo, Naren Ramakrishnan, and Chandan K Reddy. 2019. Deep transfer reinforcement learning for text summarization. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 675–683. SIAM.

Afonso Mendes, Shashi Narayan, Sebastião Miranda, Zita Marinho, André FT Martins, and Shay B Cohen. 2019. Jointly extracting and compressing documents with summary state representations. *arXiv preprint arXiv:1904.02020*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, Mao-song Sun, et al. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327.

Tian Shi, Ping Wang, and Chandan K. Reddy. 2019. LeafNATS: An open-source toolkit and live demo system for neural abstractive text summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 66–71, Minneapolis, Minnesota. Association for Computational Linguistics.

Josef Steinberger and Karel Jezek. 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4:93–100.

Juan-Manuel Torres-Moreno. 2014. *Automatic text summarization*. John Wiley & Sons.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1546–1555. Association for Computational Linguistics.

Wensi Xi, Edward A. Fox, Weiguo Fan, Benyu Zhang, Zheng Chen, Jun Yan, and Dong Zhuang. 2005. Simfusion: Measuring similarity using unified relationship matrix. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 130–137, New York, NY, USA. Association for Computing Machinery.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. *arXiv preprint arXiv:1909.00156*.