RANLP 2019 Multilingual Headline Generation Task Overview

Marina Litvak[1], John M. Conroy[2], and Peter A.Rankel[3]
[1]Shamoon College of Engineering, Beer Sheva, Israel
marinal@ac.sce.ac.il
[2]IDA/Center for Computing Sciences, 17100 Science Dr., Bowie, MD, USA
conroy@super.org
[3]Stratus Solutions Inc.
rankel@math.umd.edu

Abstract

The objective of the 2019 RANLP Multilingual Headline Generation (HG) Task is to explore some of the challenges highlighted by current state of the art approaches on creating informative headlines to news articles: non-descriptive headlines, out-of-domain training data, generating headlines from long documents which are not well represented by the head heuristic, and dealing with multilingual domain. This tasks makes available a large set of training data for headline generation and provides an evaluation methods for the task. Our data sets are drawn from Wikinews as well as Wikipedia. Participants were required to generate headlines for at least 3 languages, which were evaluated via automatic methods. A key aspect of the task is multilinguality. The task measures the performance of multilingual headline generation systems using the Wikipedia and Wikinews articles in multiple languages. The objective is to assess the performance of automatic headline generation techniques on text documents covering a diverse range of languages and topics outside the news domain.

1 Introduction

Headline Generation (HG) is an active area of research. A headline of a document can be defined as a short sentence that gives a reader a general idea about the main contents of the story it entitles. There have been many reported practical applications for headline generation (Colmenares et al., 2015; Buyukkokten et al., 2001; Linke-Ellis, 1999; De Kok, 2008; Gatti et al., 2016) or related tasks.

Automatic evaluation of automatically generated headlines is a highly important task, in its own right, where a candidate headline is assessed with respect to (1) readability (i.e. whether the headline is easy to understand), and (2) relevance (i.e. whether the headline reflects the main topic of an article).

The objective of the HG task is to stimulate research and assess the performance of automatic headline generation systems on documents covering a large range of sizes, languages, and topics. This report describes the task, how the datasets were created, the methods used to evaluate the submitted headlines, and the overall performance of each system.

2 Task and Datasets Description

The specific objective of each participant system of the task was to generate a headline/title for each document in one of two provided datasets, in at least three languages. No restrictions were placed on the languages that could be chosen. To remove any potential bias in the evaluation of generated headlines that are too small, the gold standard headline length in characters was provided for each test document and generated headlines were expected to be close to it. Two datasets were provided. Both are publicly available and can be downloaded from the MultiLing site.[1]

Wikipedia dataset

The dataset was created from the featured articles of Wikipedia, which are consists of over 13000 articles in over 40 languages. These articles are reviewed and voted upon by the community of Wikipedia editors who concur that they are the

[1]http://multiling.iit.demokritos.gr/pages/view/1651/task-headline-generation

Proceedings of the Multiling 2019 Workshop, co-located with the RANLP 2019 conference, pages 1–5
Varna, Bulgaria, September 6, 2019.
http://doi.org/10.26615/978-954-452-058-8_001

best and that the articles fulfill the Wikipedia's requirements in accuracy, neutrality, completeness, and style. As all featured article must have a summary, a subsets of these data were used at MulitLing 2013, 2015, and 2017 (Conroy et al., 2019). All the featured articles have titles for entire article and per section (sub-headings), thus, they also make an excellent corpus for research in headline generation. The Perl module Text::Corpus::Summaries::Wikipedia[2] is available and can be used to create an updated corpus. The testing dataset for this task was created from a subset of this corpus by requiring that each language has 30 articles and that the size of each article's body text be sufficiently large. A language was not select if the total number of remaining articles was less than 30.

Wikinews dataset

This dataset was created from the Wikinews articles. Since all featured articles have human-generated headlines, they make an excellent corpus for research in headline generation. The articles in this dataset do not have sub-headings, and only the main headline per article needed to be generated by participants in the provided test data. We manually assessed the collected data and filtered out files with small body or short and non informative headlines. The script for data collection is publicly available upon request. Table 1 shows the statistics about both datasets, including total number of documents, number of training and test documents per language, average document and headline length in characters (denoted by ADL and AHL, respectively).

3 Evaluation

3.1 Metrics

Both submissions were evaluated automatically, with help of the HEvAS system (Litvak et al., 2019). All headlines were evaluated in terms of multiple metrics, both from informativeness and readability perspectives. The informativeness metrics estimated the headlines quality at the lexical and semantic levels, by comparison to the content of gold standard headlines and the documents themselves.

The lexical-level informativeness metrics employed are ROUGE (Lin, 2004; Colmenares et al., 2015) (ROUGE-1,2,SU,WSU) and averaged KL-Divergence (Huang, 2008). At the semantic level, we measured content overlap above abstract "topics" discovered by Latent Semantic Indexing (LSI) (Colmenares et al., 2015), Topic Modeling (TM) (Blei et al., 2003; Blei, 2012), and Word Embedding (WE) (Mikolov et al., 2013). The content overlap is calculated via comparison to the gold standard headlines (denoted by "similarity") and the document itself (denoted by "coverage").

The following readability metrics were computed: proper noun ratio (PNR) (Smith et al., 2012), noun ratio (NR) (Hancke et al., 2012), pronoun ratio (PR)(Štajner et al., 2012), Gunning fog index (Gunning, 1952), and average word length (AWL) (Rello et al., 2013).

The details about implementation of all these metrics can be found in (Litvak et al., 2019).

3.2 Baselines

For comparative evaluations and a possibility to get impression about relative performance of the evaluated systems, their scores were compared to five baselines that are implemented in HEvAS:
(1) ***First*** compiles a headline from nine first words; (2) ***Random*** extracts nine first words from a random sentence; (3) ***TF-IDF*** selects nine top-rated words ranked by their $tf - idf$ scores; (4) ***WTextRank*** generates a headline from nine words extracted by the TextRank algorithm (Mihalcea and Tarau, 2004) for the keyword extraction; and (5) ***STextRank*** extracts nine first words from the top-ranked sentence by the TextRank approach for extractive summarization.

3.3 Participants

Two teams submitted the results for the HG task. The teams are denoted by BUPT (Beijing University of Posts and Telecommunications) and NCSR (National Centre for Scientific Research "Democritos"). Table 2 contains the details about each team.

3.4 Results

Figure 1 and Figure 2 show the evaluation results of informativeness for the generated headlines by BUPT and NCSR, respectively. Figure 3 and Figure 4 show the evaluation results of readability for the generated headlines by BUPT and NCSR, respectively. Based on the results, we can see that neither of submissions outperformed all baselines

[2] https://goo.gl/ySgOS

Dataset	# documents	# languages	# training docs	# test docs	ADL	AHL	sub-titles
Wikipedia	9293	42	30–3793	30	32187.6	16.8	yes
Wikinews	3948	27	75–140	30	1450.8	40.7	no

Table 1: Dataset statistics.

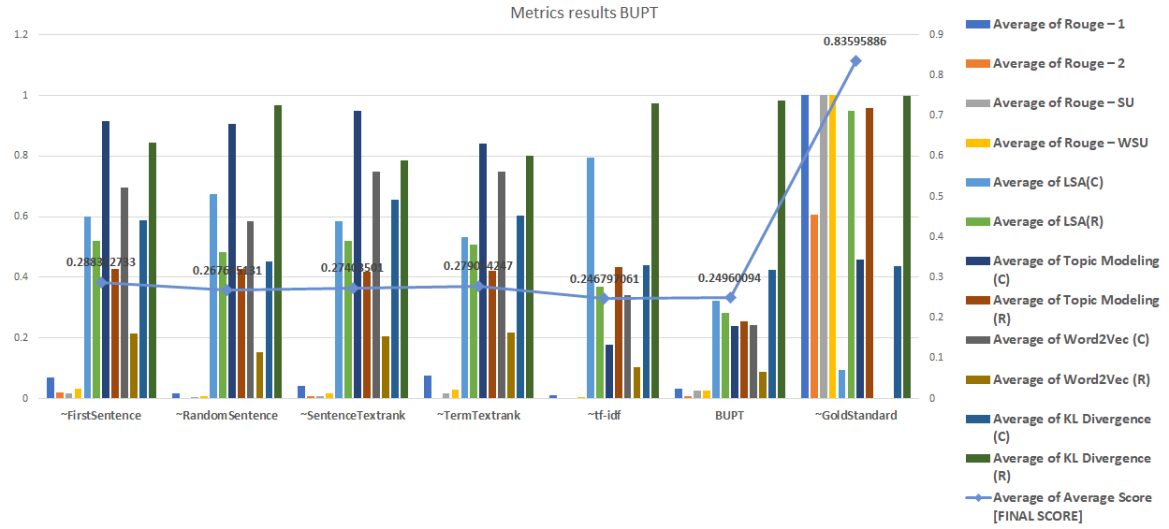

Figure 1: BUPT comparative results. Informativeness metrics.

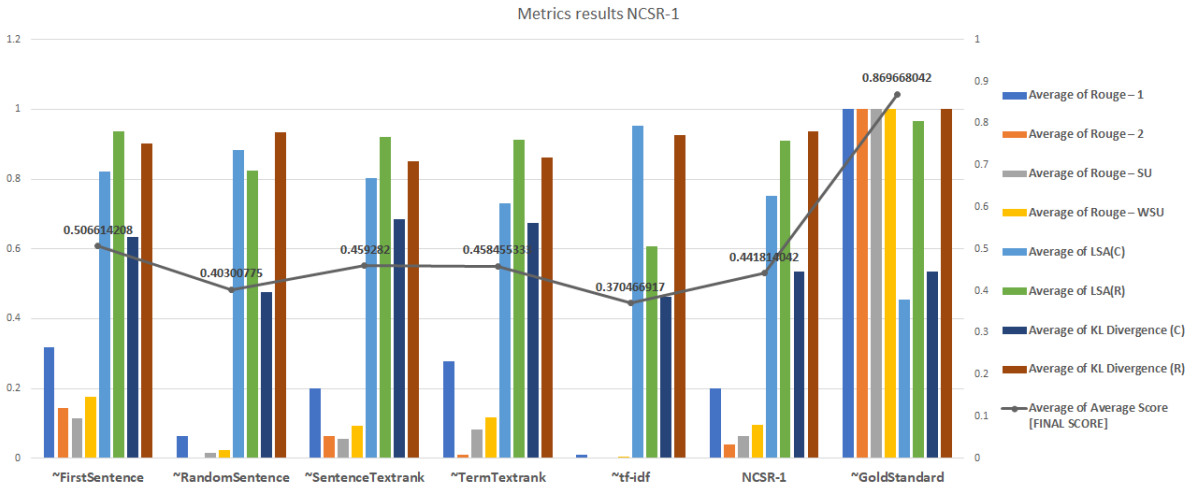

Figure 2: NCSR comparative results. Informativeness metrics.

Team	dataset	# languages	method
BUPT	Wikipedia	41	extractive
NCSR	Wikinews	3	abstractive

Table 2: Teams statistics.

in informativeness metrics. Because BUPT extracted entire sentences, their headlines are less informative but most readable. The NCSR headlines, conversely, are more informative than headlines produced by some baselines but not readable.

4 Conclusions

The Multilingual Headline Generation task presented the first open evaluation of multilingual headlines. Wikinews and the Wikipedia feature articles, both which have been used in previous multilingual summarization tasks proved again to be a great source of pre-marked data. In this first evaluation two teams submitted systems, one for each

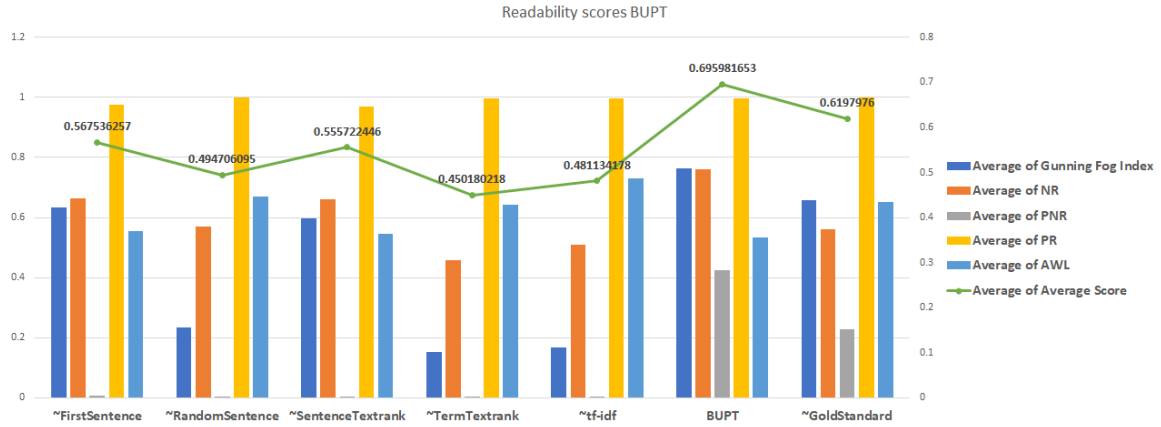

Figure 3: BUPT comparative results. Readability metrics.

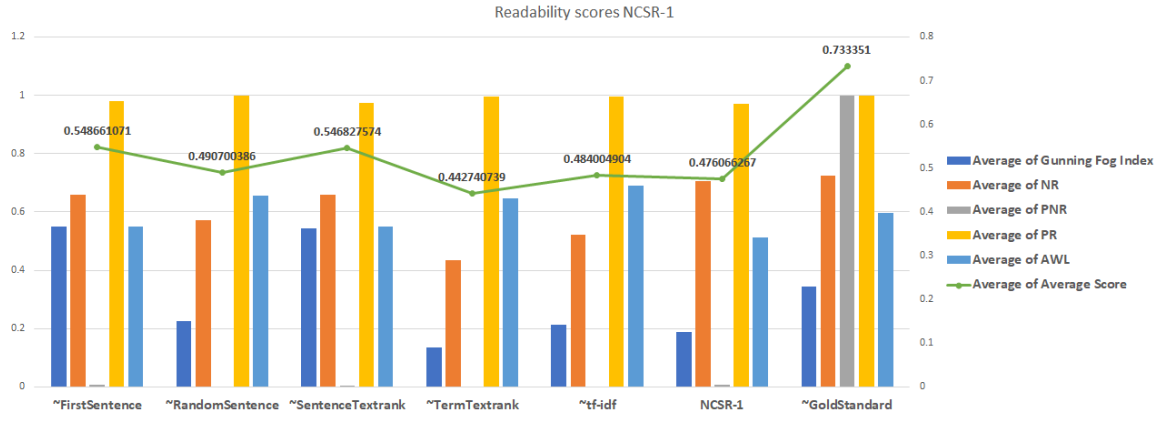

Figure 4: NCSR comparative results. Readability metrics.

task. Their systems were able to improve over some of the baselines. Further analysis of the submitted headlines, both system and baselines can be done to aid in development of stronger methods for automatic multilingual headline generation.

References

D M Blei. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Orkut Buyukkokten, Hector Garcia-Molina, and Andreas Paepcke. 2001. Seeing the whole in parts: text summarization for web browsing on handheld devices. In *Proceedings of the 10th international conference on World Wide Web. ACM*.

Carlos A Colmenares, Marina Litvak, Amin Mantrach, and Fabrizio Silvestri. 2015. Heads: Headline generation as sequence prediction using an abstract feature-rich space. In *Proceedings of the 2015 Conference of the NAACL: HLT*. pages 133–142.

John M. Conroy, Jeff Kubina, Peter A. Rankel, and Julia S. Yang. 2019. *Multilingual Summarization and Evaluation Using Wikipedia Featured Articles*, chapter Chapter 9, pages 281–336.

D De Kok. 2008. Headline generation for dutch newspaper articles through transformation-based learning. *Master's thesis* .

Lorenzo Gatti, Gozde Ozbal, Marco Guerini, Oliviero Stock, and Carlo Strapparava. 2016. Heady-lines: A creative generator of newspaper headlines. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces*. ACM, pages 79–83.

Robert Gunning. 1952. The technique of clear writing. .

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. *Proceedings of COLING 2012* pages 1063–1080.

Anna Huang. 2008. Similarity measures for text document clustering. In *sixth New Zealand computer science research student conference (NZCSRSC2008)*. pages 49–56.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out workshop*.

N. Linke-Ellis. 1999. Closed captioning in america: Looking beyond compliance. In *Proceedings of the TAO Workshop on TV Closed Captions for the Hearing Impaired People, Tokyo, Japan*. pages 43–59.

Marina Litvak, Natalia Vanetik, and Itzhak Eretz Kdosha. 2019. Hevas: Headline evaluation and analysis system. In *Recent Advances in Natural Language Processing (RANLP)*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the EMNLP*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*. Springer, pages 203–219.

Christian Smith, Henrik Danielsson, and Arne Jönsson. 2012. A good space: Lexical predictors in vector space evaluation. In *LREC 2012*. Citeseer, pages 2530–2535.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility*. Citeseer, pages 14–22.