The Summary Evaluation Task in the MultiLing - RANLP 2019 Workshop

George Giannakopoulos
NCSR Demokritos, Greece
SciFY NPC, Greece
ggianna@iit.demokritos.gr

Nikiforos Pittaras
NCSR Demokritos, Greece
DIT, NKUA
pittarasnikif@iit.demokritos.gr

Abstract

This report covers the summarization evaluation task, proposed to the summarization community via the MultiLing 2019 Workshop of the RANLP 2019 conference. The task aims to encourage the development of automatic summarization evaluation methods closely aligned with manual, human-authored summary grades and judgements. A multilingual setting is adopted, building upon a corpus of Wikinews articles across 6 languages (English, Arabic, Romanian, Greek, Spanish and Czech). The evaluation utilizes human (golden) and machine-generated (peer) summaries, which have been assigned human evaluation scores from previous MultiLing tasks. Using these resources, the original corpus is augmented with synthetic data, combining summary texts under three different strategies (`reorder`, `merge` and `replace`), each engineered to introduce noise in the summary in a controlled and quantifiable way. We estimate that the utilization of such data can extract and highlight useful attributes of summary quality estimation, aiding the creation of data-driven automatic methods with an increased correlation to human summary evaluations across domains and languages. This paper provides a brief description of the summary evaluation task, the data generation protocol and the resources made available by the MultiLing community, towards improving automatic summarization evaluation.

1 Introduction and motivation

Automatic summary evaluation is related to the problem of how to automatically evaluate a summary of a larger source text. A body of work has produced popular methods, which build upon and rely on a small set human-authored summaries (often dubbed "golden" or "model" summaries) to be able to judge machine-generated summaries in an automated manner (e.g., (Lin, 2004; Hovy et al., 2005)). Additionally, there exists related work on fully automatic evaluation of summaries, without the need of model summaries (Louis and Nenkova, 2012; Saggion et al., 2010).

However, summary evaluation has remained an open problem in the summarization community for several years. Despite some progress in the engineered evaluation measures in producing results with an acceptable correlation with human judgements (Lin, 2004; Giannakopoulos et al., 2017; Giannakopoulos, 2009), application of these approaches in (a) multiple languages, and (b) multiple domains, illustrates that they may exhibit low robustness and consistency across these variable settings (Giannakopoulos et al., 2011). These pitfalls come to complement a set of other challenges that have been identified in the related literature, such as the usefulness in different variations of established methods (Rankel et al., 2013), the negligence over different components of human evaluation (Graham, 2015), the dangers of combining measures (Conroy and Dang, 2008), etc.

Given this set of issues, we extend previous work on summarization evaluation, including and focusing on the effect of sentence order on summary evaluation scores (Madnani et al., 2007). To this end, in this task we provide dataset resources rich with reordered sum-

Proceedings of the Multiling 2019 Workshop, co-located with the RANLP 2019 conference, pages 11–16
Varna, Bulgaria, September 6, 2019.
http://doi.org/10.26615/978-954-452-058-8_003

mary instances, ranging from single to multi-sentence shuffles and sentence swaps across summaries. Finally, our contribution adopts a multi-lingual setting, going beyond English summary data and including languages with far less resources in the NLP and summarization research community. We describe these contributions in detail, starting with the introduction of the summary evaluation task in Section 2, followed by a description of the data generation process in Section 3. We conclude with a discussion on the utility of the provided summary resources (Section 4) and conclude with an outline of this paper, along with future work and next steps of the MultiLing community.

2 The Summary Evaluation Task

Given the aforementioned issues and building on previous work of the MultiLing community as well as past efforts undertaken in previous MultiLing workshops, this year we relaunch the MultiLing Summary Evaluation task within and beyond the 2019 workshop. In the next paragraphs, we define the task, elaborate on the accompanying data and describe the evaluation methodology and utility of the provided resources.

2.1 Problem definition and scope

The summary evaluation tasks aims to incentivize the construction of automatic summary evaluation systems that produce judgements that correlate highly with corresponding feedback from human evaluators. As previously elaborated, such systems should wield desirable properties that go beyond existing work in summary evaluation methods, i.e.:

- Display a degree of robustness against multilingual application, being able to produce qualitative evaluations on a range of input languages.
- Be applicable in more than one domain. This trait could manifest itself as a language-agnostic pipeline, the application of transfer learning and domain adaptation, etc.

To aid the construction of such systems, we provide a collection of resources along with the support and expertise of the MultiLing community. Specifically, as part of the MultiLing2019 effort, we have generated and made publicly available a diverse multilingual dataset (as well as a collection of tools, services and web infrastructure, expected to be finalized within the year) described in the following sections.

3 A Synthetic Summary Evaluation Dataset

Continuing from the 2017 workshop, we have renewed the data generation architecture and methodology, paired with an updated infrastructure support roadmap for the task.

3.1 Source data

We utilize compiled datasets from previous MultiLing tasks (Giannakopoulos et al., 2011; Kubina et al., 2013; Giannakopoulos et al., 2015; Giannakopoulos et al., 2017), composed of multilingual news articles from Wikinews[1]. Each article is paired with model ("golden") summaries, as well as graded, machine-generated summaries from past MultiLing participants. Specifically, we use the source documents and golden summaries of the MultiLing 2013 multilingual and multi-document summarization task. The data consists of a collection of 15 topics with source articles for a number languages. We select languages with coverage over the entirety of the topics, arriving at a total of 6 languages with approximately 5 source articles each, i.e. Arabic, Czech, English, Greek, Romanian, Spanish.

Additionally, we utilize the automatic summaries generated by participant systems in the workshop of that year (Kubina et al., 2013; Giannakopoulos, 2013) along with human-annotated grades. The total number of files (summaries and full source texts) per language are listed in Table 1.

3.2 Synthetic Data Generation

Using the dataset described above, we apply data augmentation methods to produce additional summaries, via an application of an array of summary transformation or "scrambling" mechanisms. The purpose of these op-

[1] `https://en.wikinews.org/wiki/Main_Page`

	Original input dataset					
split	train			test		
language	sources	models	peers - scores	sources	models	peers - scores
Arabic	75	60	150	75	30	75
Czech	75	60	90	75	30	45
English	75	60	148	75	30	74
Greek	75	60	90	75	30	45
Romanian	75	60	90	75	30	45
Spanish	75	60	90	75	30	45

Table 1: Total number of train / test set source documents, model summaries and evaluated peer summaries, per language in the summary evaluation task input dataset, across all 15 document topics.

	Original input dataset					
split	train			test		
type	sources	models	peers - scores	sources	models	peers - scores
count	450	180	329	450	90	390
	Synthetic dataset					
split	train			test		
type	sources	models	peers - scores	sources	models	peers - scores
count	N/A	6300	11515	N/A	3150	13650

Table 2: Total source documents, model and evaluated peer summaries, for all languages and topics. We provide the counts for (a) the original MultiLing summary and source data in the input dataset (top), (b) the total data produced by processing the input via the synthetic generation process (bottom), for each input summary type.

erators are to introduce noise in a systematic manner, with the amount of such disturbances affecting the original summary quality in a predictable way. Each such process utilizes input summary data to produce a new synthetic summary, by introducing randomness at the sentence level. The input to this process is either a single summary or a combination of multiple summaries, as outlined in the method descriptions below:

1. **Sentence reordering (`SO`)**: this method operates at the level of a single summary. Given an input summary S in the form of a collection of sentences s_i, $S = \{s_1, s_2, \ldots, s_N\}$, `SO` scrambling produces an output summary $F_{SO}(S)$, where $F()$ is a random shuffle operation assuming the form of a derangement (de Montmort, 1713) – i.e. identity mappings of the source elements are avoided. Evaluation on output data from this strategy should capture the impact of sentence order in summary evaluation methods.

2. **Sentence replacement (`SR`)**: here, the output summary is produced by two steps of random selection. First, a number of sentences $s_i \in S$ are randomly chosen from the input summary to be replaced. Subsequently, replacement sentences are randomly picked from other summary files, which is implemented as follows. First, all available tuples (S_r, s_j) are generated, with S_r denoting other summaries (different than S) in the available pool for the same topic and language as S, and s_j a sentence in S_r. We then randomly select one replacement tuple for each input sentence marked for replacement, swapping the latter with the corresponding summary / sentence source contained in the tuple. This strategy extends upon `SO` by also considering content scrambling across different summaries, along sentence order within the input summary; this is meant to identify how overall quality of the constituent

	Composite Dataset, v1							
split	train				test			
type	sources	models	peers - scores	synth	sources	models	peers - scores	synth
count	450	180	329	0	450	90	390	1890
	Composite Dataset, v2							
split	train				test			
type	sources	models	peers - scores	synth	sources	models	peers - scores	synth
count	450	180	329	17815	450	90	390	16800

Table 3: Total source documents, model summaries, evaluated peer summaries and synthetic summaries, for all languages and topics in the provided dataset versions. The current version of the composite dataset (v1, top) includes a subset of the synthetic data in the test portion of the dataset. Version v2 (bottom) contains the entirety of the generated synthetic data.

summaries tends to influence the resulting mixed summary.

3. **Summary merging (`ME`)**: The merging scrambling method is the final operator examined in our approach, and is a coarse-grained version of `SR`. Here, the scrambling does not operate on the sentence level, but splits the entire summary into two halves. The split is computed with respect to number sentences, not characters, i.e. $S_{first} = \{s_i \in S : i \leq |S|/2\}$ and $S_{second} = \{s_i \in S : i > |S|/2\}$ where $|S|$ denotes the sentence set cardinality for the summary. One of the two halves is subsequently randomly selected to be replaced with a corresponding half (e.g. a first (second) half is only replaced with another first (second) half) from another randomly selected summary for the same topic and language. This approach extends on `SO` and `SR` by introducing a potential change in the overall length of the summary, along with random replacement of summary content.

Having these scrambling options, we generate 5 randomized samples per strategy and summary file of the compiled input dataset described above. For each of the 5 samples, how each scrambling strategy is applied (e.g. which sentences are reordered in `SO` and how, which replacement summaries are selected in `SR` and `ME`, and so on) is randomly decided, leading to variations between them. Additionally, for the modification strategies that operate on the sentence level (i.e. `SO` and `SR`), we vary the percentage p of the sentences affected, $p \in \{20, 40, 60\}$. For example, for $p = 20$, approximately 20% of the summary sentences are randomly reordered in `SO` scrambling, while 20% of source sentences are replaced when `SR` scrambling is applied. The percentage determines the amount of scrambling noise the post-processing step introduces, which is expected to be associated with a corresponding change in quality in the synthetic output summary.

3.3 Available datasets

The two configuration modifiers (i.e. the amount of noise and number of repetitions) combined with the three strategies described above, result in the generation of 35 synthetic samples, for each summary in the original input dataset. The total number of synthetic data generated is detailed in table 2 and compared with the counts of the original assembled source data described in Section 3.1. The augmentation process results in a well-populated collection of summaries; we estimate that this volume of data will be able to leverage and support a productive and fruitful summary evaluation task.

In the following weeks, the MultiLing community will launch a large-scale human evaluation effort in order to annotate the synthetic summaries with manual evaluation scores. Until the completion of this task, we provide two dataset versions to the summarization community. These datasets are illustrated in Table 3. The compacted version (v1) consists of the original source data, with the test set extended with a small, representative sample of the synthetic data. This sample is

extracted by including 3 random representatives for each scrambling strategy and noise strength for each topic / language pair, resulting in a total of $6 \times 15 \times (2 \times 3 + 1) \times 3 = 1890$ synthetic summaries for the test set. We do not extend the training portion, given the lack of human evaluation scores for the synthetic data. However, we provide the full composite dataset to interested parties, amounting to a total of 17815 and 16800 synthetic summaries, for the train and test portion of the dataset, respectively. Both dataset versions are publicly available in the MultiLing community website [2].

3.4 Implementation

We used Python v3.7 to generate the synthetic summaries. Language-aware sentence splitting was performed using the Stanford CoreNLP library[3](Manning et al., 2014), along with the pycountry[4] library for locale processing. The NLTK [5] (Loper and Bird, 2002) package was used for generic text processing and manipulation tasks.

3.5 Evaluation plan

As mentioned previously, the manual evaluation of the synthetic data is currently in progress, utilizing resources and expertise within the MultiLing community. The available datasets will be incrementally updated with evaluation scores, as the latter are being aggregated and incorporated. Additionally, in the immediate future, MultiLing will further support the summary evaluation task by introducing an automatic evaluation platform on the MultiLing website [6], along with an array of usability, user experience and interface improvements to the community webpage. Further, we will examine providing means and support for crowd-sourcing (Pittaras et al., 2019), to aid and reduce the cost of human evaluation in summarization tasks.

4 Discussion

The generated dataset provides summaries of variable quality, spread across multiple, identifiable noise categories (e.g. sentence order, sentence replacement and merging). We expect this engineered feature to aid the development of evaluation approaches and measures that attempt to capture and highlight such artifacts, as an additional stepping stone to arriving at high correlation to human judgements. Specifically, we emphasize the importance of detection and quantification of the degree of alignment of such automatic evaluations and human grades. This alignment should capture, encapsulate and be influenced by details of the synthetic generation process of a summary (i.e. which scrambling method is applied), the amount of noise introduced (e.g. number and distance of reordered sentences), the evaluated quality of the source summary / summaries (e.g. a combination of the grades of two merged summary parts), etc. Finally, additional avenues for alignment to human scores (e.g. degrees of qualitative deviation, corresponding to the aforementioned factors) could be discovered on top of the provided ones, via engineered or automatic methods.

5 Future work and conclusions

In this paper we have provided a brief description of the summary evaluation task, bootstrapped in the MultiLing 2019 workshop. We have described in detail a synthetic data generation process, making publicly available two versions of a composite dataset (containing synthetic and non-synthetic data) that is produced from it. We believe that these data can be utilized towards generating efficient and robust summary evaluation approaches.

Within the next months, we will work on the human evaluation task of the generated synthetic data. Additionally, we will implement the evaluation steps outlined in Section 3.5, in order to create an accessible benchmark towards incentivizing the improvement automatic summary evaluation methods. Furthermore, we will make available a corresponding augmented dataset using domains different from news articles, utilizing MultiLing corpora from other workshop tasks. Additionally, appropriate dissemination and outreach steps will be taken to further encourage participation in the summary evaluation task within and beyond the MultiLing community.

[2] http://multiling.iit.demokritos.gr/
[3] https://stanfordnlp.github.io/CoreNLP/
[4] https://pypi.org/project/pycountry/
[5] https://www.nltk.org/
[6] http://multiling.iit.demokritos.gr/

References

John M. Conroy and Hoa Trang Dang. 2008. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 145–152, Manchester, UK, August. Coling 2008 Organizing Committee.

Pierre Rémond de Montmort. 1713. *Essay d'analyse sur les jeux de hazard*. C. Jombert.

George Giannakopoulos, Mahmoud El-Haj, Benoit Favre, Marina Litvak, Josef Steinberger, and Vasudeva Varma. 2011. Tac 2011 multiling pilot overview.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. Multiling 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274.

George Giannakopoulos, John Conroy, Jeff Kubina, Peter A Rankel, Elena Lloret, Josef Steinberger, Marina Litvak, and Benoit Favre. 2017. Multiling 2017 overview. In *Proceedings of the MultiLing 2017 workshop on summarization and summary evaluation across source types and genres*, pages 1–6.

George Giannakopoulos. 2009. Automatic summarization from multiple documents. *Ph. D. dissertation*.

George Giannakopoulos. 2013. Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 20–28.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal, September. Association for Computational Linguistics.

E. Hovy, C. Y. Lin, L. Zhou, and J. Fukumoto. 2005. Basic elements.

Jeff Kubina, John Conroy, and Judith Schlesinger. 2013. Acl 2013 multiling pilot overview. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 29–38.

C. Y. Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Annie Louis and Ani Nenkova. 2012. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300, Aug.

Nitin Madnani, Rebecca Passonneau, Necip Fazil Ayan, John M Conroy, Bonnie J Dorr, Judith L Klavans, Dianne P O'Leary, and Judith D Schlesinger. 2007. Measuring variability in sentence ordering for news summarization. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 81–88. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Nikiforos Pittaras, Stefano Montanelli, George Giannakopoulos, Alfio Ferrara, and Vangelis Karkaletsis. 2019. Crowdsourcing in single-document summary evaluation: The argo way. *Multilingual Text Analysis: Challenges, Models, And Approaches*, page 245.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136, Sofia, Bulgaria, August. Association for Computational Linguistics.

H. Saggion, J. M. Torres-Moreno, I. Cunha, and E. SanJuan. 2010. Multilingual summarization evaluation without human models. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, page 1059–1067.