MultiLing 2017 Overview

George Giannakopoulos
NCSR "Demokritos", Greece

John M. Conroy
IDA / Center for Comp. Sciences, U.S.A

Jeff Kubina
U.S. Dep. of Defense, U.S.A

Peter A. Rankel
Elder Research, U.S.A

Elena Lloret
Univ. of Alicante, Spain

Josef Steinberger
Univ. of West Bohemia, Czech Republic

Marina Litvak
Shamoon College of Engineering, Israel

Benoit Favre
LIF, France

Abstract

In this brief report we present an overview of the MultiLing 2017 effort and workshop, as implemented within EACL 2017. MultiLing is a community-driven initiative that pushes the state-of-the-art in Automatic Summarization by providing data sets and fostering further research and development of summarization systems. This year the scope of the workshop was widened, bringing together researchers that work on summarization across sources, languages and genres. We summarize the main tasks planned and implemented this year, also providing insights on next steps.

1 Overview

MultiLing covers a variety of topics on Natural Language Processing, focused on the multilingual aspect of summarization:

- **Multilingual summarization across genres and sources:** Summarization has been receiving increasing attention during the last years. This is mostly due to the increasing volume and redundancy of available online information but also due to the user created content. Recently, more and more interest arises for methods that will be able to function on a variety of languages and across different types of content and genres (news, social media, transcripts).

  This topic of research is mapped to different community tasks, covering different genres and source types:

  - Multilingual single-document summarization (Giannakopoulos et al., 2015);
  - user-supplied comments summarization (OnForumS task (Kabadjov et al., 2015));
  - conversation transcripts summarization (see also (Favre et al., 2015)).

  The spectrum of the tasks covers a variety of real settings, identifying individual requirements and intricacies, similarly to previous MultiLing endeavours (Giannakopoulos et al., 2011a; Giannakopoulos, 2013; Elhadad et al., 2013; Giannakopoulos et al., 2015).

- **Multilingual summary evaluation:** Summary evaluation has been an open question for several years, even though there exist methods that correlate well to human judgement, when called upon to compare systems. In the multilingual setting, it is not obvious that these methods will perform equally well to the English language setting. In fact, some preliminary results have shown that several problems may arise in the multilingual setting (Giannakopoulos et al., 2011a). The same challenges arise across different source types and genres. This aspect of the workshop aims to cover and discuss these research problems and corresponding solutions.

The workshop builds upon the results of a set of research **community tasks**, which are elaborated on in the following paragraphs. However, this year MultiLing also hosts works beyond the tasks themselves, but still within the scope of automatic summarization and evaluation in different genres and settings.

2 Community Tasks

In this year's MultiLing community effort we are implementing the following tasks:

Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres, pages 1–6,
Valencia, Spain, April 3, 2017.

- Multilingual Single-Document Summarization (MSS)
- Multilingual Summary Evaluation (MSE)
- Online Forum Summarization (OnForumS)
- Call Centre Conversation Summarization (CCCS)
- Headline Generation Task (HG)

Due to time limitations, all but the MSS and OnForumS tasks will run beyond the workshop timespan, thus the proceedings will be complemented by the proceedings addendum [1], containing system reports and evaluation results.

3 Multilingual Single-Document Summarization Task Overview

The Multilingual Single-document Summarization 2017 posed a task to measure the performance of multilingual, single-document, summarization systems using a dataset derived from the featured articles of 41 Wikipedias. The objective was to assess the performance of automatic summarization techniques on text documents covering a diverse range of languages and topics outside the news domain. This section describes the task, the dataset and the methods to be used to evaluate the submitted summaries. To give amble time for evaluation the results and analysis will be presented at the workshop and published later. The objective of this task, like the 2015 Multilingual Single-document Summarization Task, was to stimulate research and assess the performance of automatic single-document summarization systems on documents covering a large range of sizes, languages, and topics.

3.1 Task and Dataset Description

Each participating system of the task was to compute a summary for each document in at least one of the datasets 41 languages. To remove any potential bias in the evaluation of generated summaries that are too small, the human summary length in characters was provided for each test document and generated summaries were expected to be close to it.

The testing dataset was created using the same steps as reported in Section 2 of (Giannakopoulos et al., 2015) and excluded the articles in the training dataset (which was the testing dataset for the task in 2015). For each language Table 1 contains the mean character size of the summary and body of the articles selected for the test dataset. Within the dataset there is no correlation between the summary and body size of the articles, in fact, the variance in the summary size is small. This is likely because Wikipedia style requirements dictate that a summary be at most four paragraphs,[2] regardless of article size, and paragraphs be reasonably sized.[3]

3.2 Preprocessing and Evaluation

For the evaluation the baseline summary for each article in the dataset was the prefix substring of the article's body text having the same length as the human summary of the article. An oracle summary was also computed for each article using the combinatorial covering algorithm in (Davis et al., 2012) by selecting sentences from its body text to cover the tokens in the human summary using as few sentences as possible until its size exceeded the human summary, upon which it was truncated.

Preprocessing of all the submitted and human summaries was performed using the Natural Language Toolkit (Bird et al., 2009). Sentence splitting was done using *punkt()*. Models based on the Wikipedia data were built for each language. For each summary the pre-processing steps were:

1. all multiple white-spaces and control characters are convert to a single space
2. any leading space is removed
3. the resulting text string is truncated to the human summary length
4. the text is tokenized and, if possible, lemmatized
5. all tokens without a letter or number are discarded
6. all remaining tokens are lowercased.

[1] Cf. http://multiling.iit.demokritos.gr/pages/view/1638/multiling-2017-proceedings-addendum.

[2] https://en.wikipedia.org/wiki/Wikipedia:LEAD

[3] https://en.wikipedia.org/wiki/Wikipedia:WBA

Table 1: Dataset Languages and Sizes

ISO	LANGUAGE	SUMMARY	BODY	ISO	LANGUAGE	SUMMARY	BODY
af	Afrikaans	1743 (784)	32407 (20378)	ka	Georgian	1114 (682)	23626 (23018)
ar	Arabic	2129 (1045)	38682 (16354)	ko	Korean	905 (491)	15723 (7098)
az	Azerbaijani	1375 (937)	48687 (45855)	li	Limburgish	569 (237)	14177 (16326)
bg	Bulgarian	1451 (782)	29421 (10774)	lv	Latvian	1334 (514)	25292 (13464)
bs	Bosnian	1275 (801)	26497 (15319)	mr	Marathi	970 (653)	14727 (8438)
ca	Catalan	1733 (906)	28536 (14460)	ms	Malay	1420 (952)	22820 (16851)
cs	Czech	1947 (745)	33751 (24010)	nl	Dutch	1316 (562)	36638 (18062)
de	German	1122 (470)	42838 (30382)	nn	Norwegian	965 (493)	17772 (9073)
el	Greek	1582 (905)	36081 (16652)	no	Nor.-Bok.	1808 (913)	37128 (22024)
en	English	1878 (735)	20683 (9644)	pl	Polish	1470 (687)	31460 (16319)
eo	Esperanto	1286 (875)	22905 (10279)	pt	Portuguese	2247 (759)	37189 (16777)
es	Spanish	2083 (892)	47670 (39981)	ro	Romanian	2204 (710)	38973 (20349)
eu	Basque	1105 (742)	23558 (16672)	ru	Russian	1855 (915)	59337 (27360)
fa	Persian	1850 (581)	29525 (13172)	simple	Simp. Eng.	973 (351)	9793 (7027)
fi	Finnish	1135 (406)	23971 (10538)	sk	Slovak	1104 (631)	26102 (11024)
fr	French	1924 (884)	65960 (41289)	th	Thai	1851 (951)	30549 (15203)
hr	Croatian	1398 (1119)	22430 (13583)	tr	Turkish	2059 (807)	32240 (23667)
id	Indonesian	1813 (964)	26634 (18564)	tt	Tagalog	1149 (779)	23648 (14139)
it	Italian	1743 (701)	51461 (20832)	uk	Ukrainian	1023 (758)	35552 (32014)
ja	Japanese	383 (275)	21349 (14694)	zh	Chinese	662 (245)	10614 (6338)
jv	Javanese	1118 (855)	14033 (10810)				

Table 1: The table lists the languages in the dataset with the first column containing the ISO code for each the language, the second column the name of the language, and the remaining columns containing the mean size, in characters, and standard deviation, in parentheses, of the summary and body of the article. For example, for English the mean size of the human summaries is 1,857 characters.

As of the time of publication of the proceedings, three teams have participated and automatic methods of scoring the subumissions, using ROUGE (Lin, 2004) and MeMoG (Gia,), are underway and will be presented at the EACL 2017 workshop. A human evaluation will proceed afterwards.

4 OnForumS Task

Further to the pilot of OnForumS in 2015, we organized the task again in 2017 with a brand new dataset. The OnForumS task investigates how the mass of comments found on news providers web sites can be summarized. We posit that a crucial initial step towards that goal is to determine what comments link to, be that either specific news snippets or comments by other users. Furthermore, a set of labels for a given link may be articulated to capture phenomena such as agreement and sentiment with respect to the comment target. Solving this labelled linking problem can enable recognition of salience (e.g., snippets/comments with most links) and relations between comments (e.g., agreement).

The OnForumS task is a particular specification of the linking task, in which systems take as input a news article with comments and were asked to link and label (sentiment, argument) each comment to sentences in the article, to the article topic as a whole or to other comments. The set of possible labels is for sentiment is [POS, NEUT, NEG] and the set of possible argument labels is [IN FAVOR, AGAINST, IMPARTIAL].

This year we focus on English (The Guardian) and Italian (La Repubblica) as in the previous edition and we released the 2015 test data as training data.

The 2017 text collection contains 19 English and 19 Italian articles. This year we had 4 participanting teams and together with two baselines we received 9 runs. The evaluation focuses on how many of the links and labels were correctly identified, as in the previous OnForumS run. The next step is to manually validate the links and labels using CrowdFlower.

5 MultiLingual Summary Evaluation

The summary evaluation task revisits the multilingually applicable evaluation challenge. The aim is to introduce novel, automatic evaluation methods of summary evaluation. Even though, currently, systems are evaluated using the ROUGE (Lin, 2004) and MeMoG (Gia,) metrics, there exists a big gap between automatic methods and manual annotations, especially in non-English settings (Giannakopoulos et al., 2011b).

This year's task reuses the MultiLing 2013 and 2015 single-document and multi-document summarization corpora and evalautions. Furthermore, we generate summary variations (often through inducing "noise"), which the evaluation systems will be asked to grade. These variations include:

- Sentence re-ordering;
- Random sentence replacement;
- Merging between different summaries.

All the above changes will be studied, to understand the strengths and weaknesses of different evaluation methods with respect to these synthetic deviations. Then, a human evaluation will be conducted to see whether humans respond similarly to the automatic methods with respect to the different noise types.

The aim of this task and study is to understand how variations of text change its perceived quality of a summary. It also aims to highlight the (in)sufficiency of existing methods in the multilingual setting and promote new, more robust approaches for summary evaluation.

6 Tasks in preparation

6.1 Headline Generation

The Headline Generation (HG) task aims 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, and generating headlines from long documents which are not well represented by the head heuristic. This task has been previously addressed in past summarization challenges, such as the well-known Document Understanding Conferences (DUC) for the 2002, 2003 or 2004 editions.

With the high-rate of information increase, novel summarization methods that could condense and extract relevant information in just one sentence (i.e., headlines) would perfectly fit in today's society for creating better information access and processing tools. We will rerun the headline generation task in DUC[4] 2002, 2003, 2004 conditions

[4]Cf. `http://duc.nist.gov//`

in order to create comparable results, and determine to what extent the techniques and methods have improved with respect to former participants.

Moreover, we will encourage multilingual or cross-lingual approaches able to generate headlines for at least two languages. We expect to make available a large set of training data for headline generation, and create evaluation conditions to objectively assess and compare different approaches.

6.2 Call Centre Conversation Summarization

The Call Centre Conversation Summarization (CCCS) task — run for the first time as a pilot task in 2015 — consists in automatically generating summaries of spoken conversations in the form of textual synopses that shall inform on the content of a conversation and might be used for browsing a large database of recordings. As in CCCS 2015, participants to the task shall generate abstractive summaries from conversation transcripts that inform a reader about the main events of the conversations, such as the objective of the participants and how they are met. Evaluation will be performed by ROUGE-like measures based on human-written summaries as in CCCS 2015, and — if possible — will be coupled by manual evaluation, depending on the funding we can secure for the task.

7 Conclusion

This year MultiLing covers a number of challenging problems related to summarization. In the proceedings (and the addendum) one can find various methods using deep learning and word embeddings, topic modeling, optimization and other approaches to achieve summarization and summary evaluation across settings.

The rest of the proceedings will allow the reader to examine interesting challenges related to abstractive summarization, argument labeling, multi-genre, multi-document and query-based summarization. They will also identify and attemp to tackle important challenges related to summary evaluation beyond English.

We hope that the conclusion of the tasks after the workshop will provide the grounds for further research and open systems development, revising and improving the way summarization is modeled, faced, evaluated and implemented in the years to come.

8 Acknowledgements

This work was supported by project MediaGist, EUs FP7 People Programme (Marie Curie Actions), no. 630786, MediaGist.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Sashka T. Davis, John M. Conroy, and Judith D. Schlesinger. 2012. OCCAMS - an optimal combinatorial covering algorithm for multi-document summarization. In *12th IEEE International Conference on Data Mining Workshops, ICDM Workshops, Brussels, Belgium, December 10, 2012*, pages 454–463.

Michael Elhadad, Sabino Miranda-Jiménez, Josef Steinberger, and George Giannakopoulos. 2013. Multi-document multilingual summarization corpus preparation, part 2: Czech, hebrew and spanish. *MultiLing 2013*, page 13.

Benoit Favre, Evgeny Stepanov, Jérémy Trione, Frédéric Béchet, and Giuseppe Riccardi. 2015. Call centre conversation summarization: A pilot task at multiling 2015. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 232.

Summarization system evaluation variations based on n-gram graphs.

George Giannakopoulos, Mahmoud El-Haj, Benoit Favre, Marina Litvak, Josef Steinberger, and Vasudeva Varma. 2011a. Tac2011 multiling pilot overview. In *TAC 2011 Workshop*.

George Giannakopoulos, Mahmoud El-Haj, Benoit Favre, Marina Litvak, Josef Steinberger, and Vasudeva Varma. 2011b. Tac2011 multiling pilot overview. In *TAC 2011 Workshop*.

George Giannakopoulos, Jeff Kubina, John M. 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 *SIGDIAL 2015*.

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.

Mijail Kabadjov, Josef Steinberger, Emma Barker, Udo Kruschwitz, and Massimo Poesio. 2015. Onforums: The shared task on online forum summarisation at multiling'15. In *Proceedings of the 7th*

Forum for Information Retrieval Evaluation, pages 21–26. ACM.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.