SIGDIAL 2015 Multilingual Single-Document Summarization Task Overview

Jeff Kubina
U.S. Department of Defense
9800 Savage Rd., Ft. Meade, MD
jmkubin@tycho.ncsc.mil

John M. Conroy
IDA/Center for Computing Sciences
17100 Science Dr., Bowie, MD
conroy@super.org

Abstract

The 2015 SIGDIAL Multilingual Single-document Summarization Task posed a task to measure the performance of multilingual, single-document, summarization systems using a dataset derived from the featured articles of 38 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 report describes the task, the dataset, the methods used to evaluate the submitted summaries, and the overall performance of each participant's system.

1 Introduction

Document summarization is an active area of research. The ACM Digital Library has over 800 reports on the subject published since 1993 with over half of them appearing in the last six years. While the initial impetus for much of this research was the annual Text Analysis Conference (TAC) Workshop on Document Summarization, many conferences now accept reports on document summarization techniques. The objective of this task, like the 2013 Multilingual Single-document Summarization Pilot 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. This report describes the task, how the dataset was created, the methods used to evaluate the submitted summaries, and the overall performance of each system.

2 Task and Dataset Description

Each participanting system of the task was to compute a summary for each document in at least one of the dataset's 38 languages. No restrictions were placed on the languages that could be chosen (though all participants chose English as one of their 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 (Kubina et al., 2013) but excluded the articles in the training dataset (which were the testing dataset for the pilot task in 2013). First, the body and summary of each article is compressed to approximate their information content size. Next, articles that have a compressed body size less than five times their compressed summary size are discarded. This is done to ensure there is sufficient information in the body to generate a summary. Finally, to select articles with reasonable summary and body sizes, within each language the median of the ratio of compressed body size to compressed summary size was computed and only the 30 articles closest to the median were included in the dataset. A language was not selected if the number of remaining articles after the selection process was less than 30. For each language Table 1 contains the mean character size of the summary and body of the articles selected for the test dataset.

3 Teams

Seven teams submitted the results for over 23 summarization systems. The teams are denoted by BGU-SCE-M, BGU-SCE-P, CCS, EXB, LCS-IESI, NTNU, and UA-DLSI; for brevity their associated systems are denoted by a number appended to the team name. Table 3 contains the total systems and languages submitted for each team.

Table 1: Dataset Languages and Sizes

ISO	LANGUAGE	SUMMARY	BODY	ISO	LANGUAGE	SUMMARY	BODY
af	Afrikaans	1199 (218)	26295 (14335)	ja	Japanese	378 (143)	18715 (7652)
ar	Arabic	1877 (141)	44144 (20993)	ka	Georgian	1003 (98)	18076 (10113)
bg	Bulgarian	1415 (169)	26582 (7984)	ko	Korean	796 (239)	16636 (9731)
ca	Catalan	1531 (86)	26992 (13635)	ms	Malay	1309 (644)	19233 (9047)
cs	Czech	2003 (160)	34268 (17078)	nl	Dutch	1147 (137)	32450 (15081)
de	German	1070 (80)	38200 (20293)	no	Nor.-Bok.	1581 (143)	35747 (13497)
el	Greek	1681 (284)	33400 (16174)	pl	Polish	1174 (84)	26407 (17249)
en	English	1857 (111)	25782 (13713)	pt	Portuguese	2000 (110)	30793 (11553)
eo	Esperanto	1172 (134)	24898 (11884)	ro	Romanian	1673 (126)	30540 (12815)
es	Spanish	2044 (129)	38368 (21978)	ru	Russian	1430 (100)	45118 (24491)
eu	Basque	1033 (155)	23893 (16282)	sh	Serbo-Croat.	1353 (704)	28302 (13304)
fa	Persian	1648 (262)	25781 (9292)	sk	Slovak	1475 (618)	32428 (15070)
fi	Finnish	1176 (95)	30116 (11169)	sl	Slovenian	1195 (113)	20756 (11459)
fr	French	1792 (95)	55805 (27157)	sr	Serbian	1677 (183)	37107 (12465)
he	Hebrew	908 (75)	21856 (12509)	sv	Swedish	1495 (87)	24509 (9114)
hr	Croatian	1093 (92)	22160 (8792)	th	Thai	1894 (426)	27409 (6688)
hu	Hungarian	1450 (81)	30170 (14321)	tr	Turkish	1889 (287)	30871 (14854)
id	Indonesian	1500 (159)	27260 (9245)	vi	Vietnamese	2094 (174)	36893 (13833)
it	Italian	1217 (77)	36173 (18601)	zh	Chinese	636 (55)	14050 (6269)

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.

TEAM	SYSTEMS	LANGUAGES
BGU-SCE-M	5	ar, en, he
BGU-SCE-P	3	ar, en, he
CCS	5	all
EXB	1	all
LCS-IESI	1	all
NTNU	1	all
UA-DLSI	6	de, en, es

Table 3: The table lists the team names, the total systems submitted, and the languages covered by the systems.

4 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. It is included in the evaluation to show the approximate maximum score achievable using extractive summarization methods.

Preprocessing of all the submitted and human summaries was performed, depending on the language, either by the Basis Technology's Rosette software (Basis Technology, 2015) or the Natural Language Toolkit (Bird et al., 2009). Table 2 lists the software package used for each language and if lemmatization was performed. For each summary the preprocessing 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.

5 Results

Summaries were automatically evaluated against the human summary of each article using ROUGE-1, 2, 3, 4, (Lin, 2004) and MeMoG (Giannakopoulos et al., 2008). For MeMoG the character n-gram size used for each language is the same as in the 2013 pilot task, which are listed in Table 3 of (Kubina et al., 2013).

For each language and each metric (ROUGE-1, 2, 3, 4, and SU4 and MeMoG) we first test if the median score for all the submitted systems and the baseline were the same, i.e., we run a non-parametric analysis of variance test after removing the scores of the oracle system. The last row of Table 3 displays the fraction of times the null hypotheses that the median ROUGE-2, ROUGE-4, and MeMoG scores were equal was rejected, using a rejection threshold of 0.05. Note, in particular for ROUGE-4, there were only 10 out of the 38 languages where the equal median hypothesis was rejected. The remaining rows of the table give the fraction of the time when the null hypothesis was rejected that a given system significantly outperformed the baseline. These tests are performed using a paired Wilcoxon test, which is know to have more statistical power to discriminate between systems. We show ROUGE-2 since it it is widely used and include ROUGE-3 and ROUGE-4 since it provides more statistical power to discriminate between high performing systems Rankel et al. (2011). Based on an analysis of the 2013 multilingual summarization both ROUGE-3 and MeMoG also have good statistical power to predict significant differences in human metrics in the multilingual summarization setting.[1]

Figure 1 gives a scatter plot of the ROUGE-2 scores for the languages where the ANOVA's null hypothesis is rejected. The blue × gives the scores for the oracle system, which is significantly greater than the best system. Figure 2 gives a similar plot without the oracle system to better see the spread between the systems as measured by ROUGE-4. Finally, Figure 3 gives the scatter plot of the system scores in the MeMoG metric.

6 Conclusion

Running the MMS task presents many challenges. Creating the dataset for the task is an arduous process since each Wikipedia lists featured articles differently and preprocessing and scoring all the submissions in a timely manner is always a logistical challenge. But it is well worth the effort in advancing the research and development of better algorithms for automatic document summarization. This year's task had seven teams submit 23 systems—14 of them performed better than the baseline summary in half of the languages they summarized. Further, a human evaluation is

[1] ROUGE-4 scores were not available.

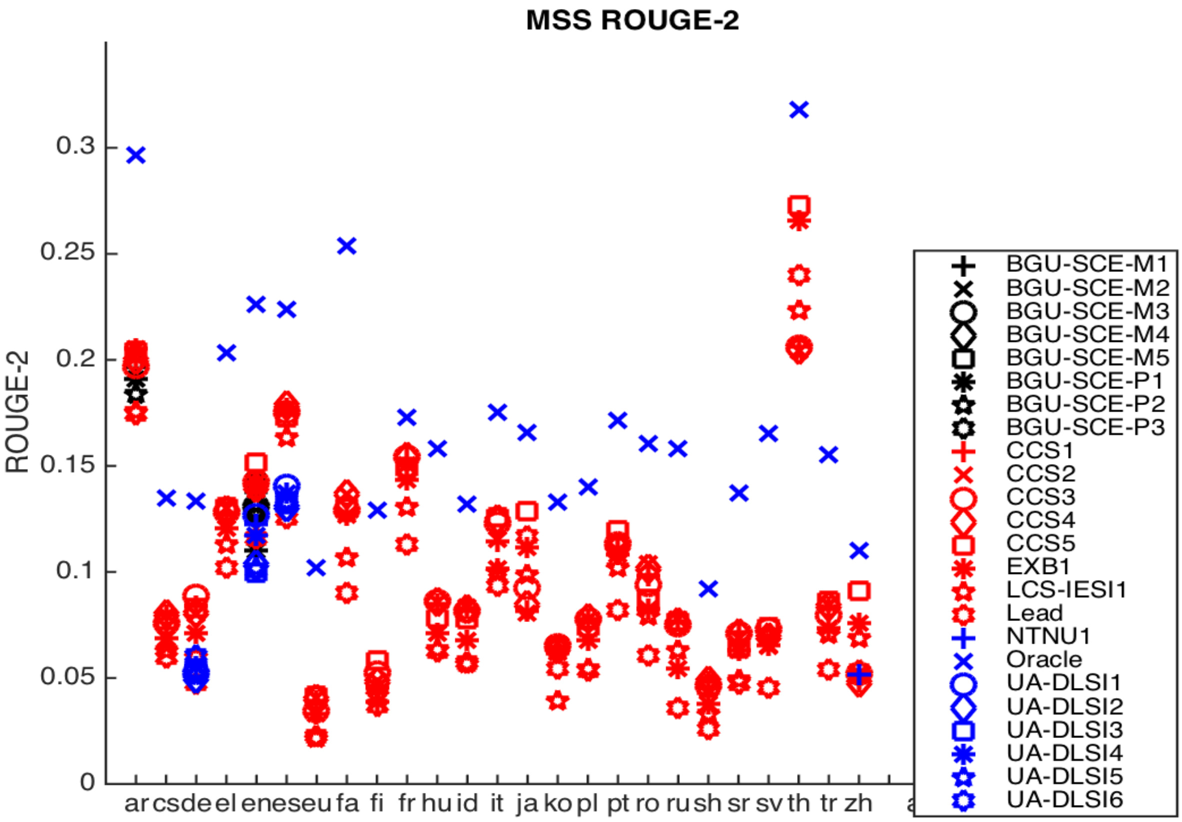

Figure 1: ROUGE-2 scores for the MSS participant systems. The high scores for Arabic and Thai are likely due to the tokenization and lemmatization performed by the Basis Rosette package.

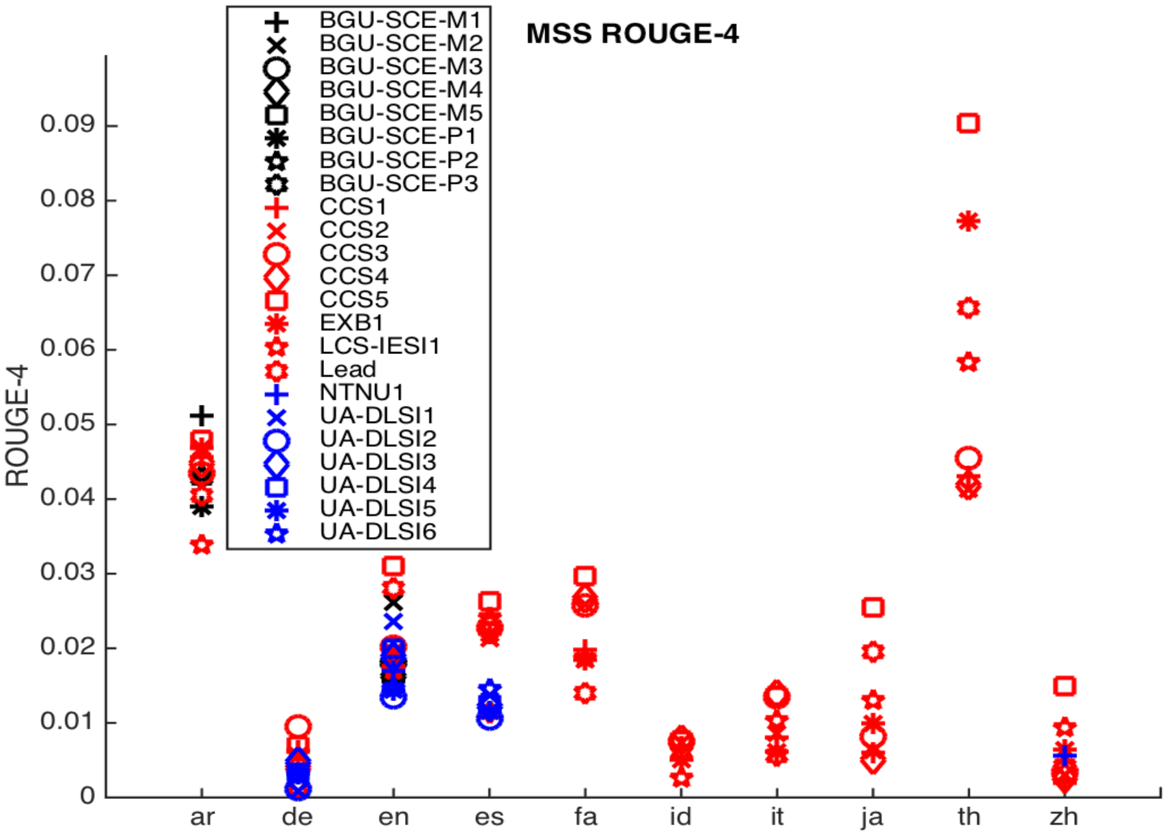

Figure 2: ROUGE-4 scores for the MSS participant systems.

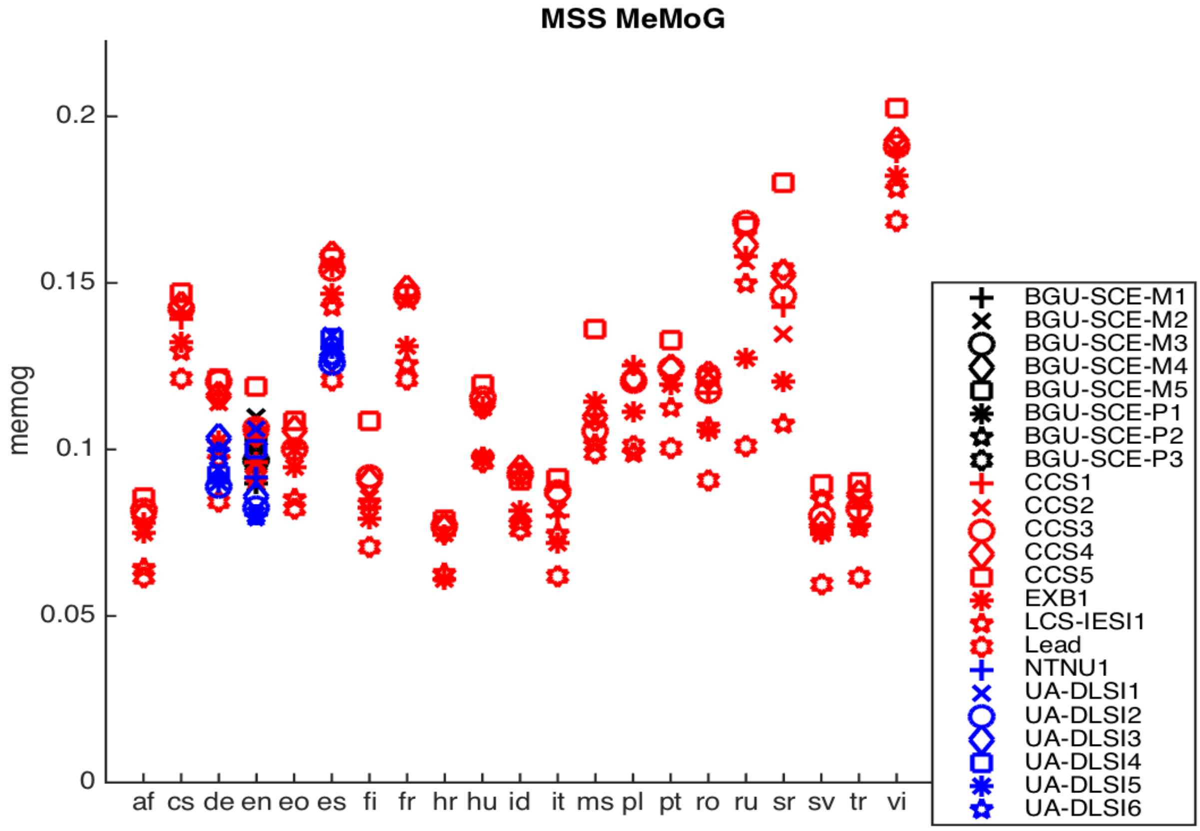

Figure 3: MeMoG scores for the MSS participant systems.

Table 2: Language, Software Package, and Lemmatization

Iso	Language	Package	Lemma	Iso	Language	Package	Lemma
af	Afrikaans	Basis	✓	ja	Japanese	Basis	✓
ar	Arabic	Basis	✓	ka	Georgian	NLTK	
bg	Bulgarian	NLTK		ko	Korean	Basis	✓
ca	Catalan	NLTK		ms	Malay	NLTK	
cs	Czech	Basis	✓	nl	Dutch	Basis	✓
de	German	Basis	✓	no	Norwegian-Bokmal	Basis	✓
el	Greek	Basis	✓	pl	Polish	Basis	✓
en	English	Basis	✓	pt	Portuguese	Basis	✓
eo	Esperanto	NLTK		ro	Romanian	Basis	✓
es	Spanish	Basis	✓	ru	Russian	Basis	✓
eu	Basque	NLTK		sh	Serbo-Croatian	NLTK	
fa	Persian	NLTK		sk	Slovak	NLTK	
fi	Finnish	NLTK	✓	sl	Slovenian	NLTK	
fr	French	Basis	✓	sr	Serbian	NLTK	
he	Hebrew	NLTK		sv	Swedish	Basis	✓
hr	Croatian	NLTK		th	Thai	Basis	✓
hu	Hungarian	Basis	✓	tr	Turkish	Basis	✓
id	Indonesian	NLTK		vi	Vietnamese	NLTK	
it	Italian	Basis	✓	zh	Chinese	Basis	✓

Table 2: The table lists the software package used to process each language and whether or not lemmatization was performed on the extracted tokens.

planned for each team's highest scoring system to provide a quality and readability score of the systems and, hopefully, with enough data to better understand which automatic scoring methods correlate best with human judgments of good summaries.

Acknowledgments

The authors would like to thank Basis Technologies for the use their Rosette natural language toolkit and (Tange, 2011) for the development of the GNU Parallel program, both sped up much of the processing.

References

Basis Technology. 2015. Rosette base linguistics (rbl-je) version 7.12.0. http://www.basistech.com.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Sashka T. Davis, John M. Conroy, and Judith D. Schlesinger. 2012. Occams - an optimal combinatorial covering algorithm for multi-document summarization. In Jilles Vreeken, Charles Ling, Mohammed Javeed Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoffrey I. Webb, and Xindong Wu, editors, *ICDM Workshops*, pages 454–463. IEEE Computer Society.

George Giannakopoulos, Vangelis Karkaletsis, George Vouros, and Panagiotis Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.*, 5(3):5:1–5:39, October.

Jeff Kubina, John M Conroy, and Judith D Schlesinger. 2013. Acl 2013 multiling pilot overview. *MultiLing 2013*, page 29.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Peter Rankel, John Conroy, Eric Slud, and Dianne O'Leary. 2011. Ranking human and machine summarization systems. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 467–473, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

O. Tange. 2011. Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47, Feb.

System	ROUGE-2	ROUGE-3	ROUGE-4	MeMoG
BGU-SCE-M1	2/3	2/3	1/3	1/3
BGU-SCE-M2	1/2	1/2	0/2	1/2
BGU-SCE-M3	1/1	0/1	0/1	1/1
BGU-SCE-M4	1/1	1/1	0/1	1/1
BGU-SCE-M5	1/1	1/1	0/1	1/1
BGU-SCE-P1	0/3	0/3	0/3	0/3
BGU-SCE-P2	2/3	0/3	0/3	1/3
BGU-SCE-P3	2/3	1/3	0/3	0/3
CCS1	20/38	8/38	3/38	19/38
CCS2	21/38	7/38	4/38	19/38
CCS3	21/38	8/38	3/38	19/38
CCS4	20/38	8/38	2/38	20/38
CCS5	23/38	10/38	7/38	20/38
EXB1	15/38	5/38	1/38	11/38
LCS-IESI1	6/38	3/38	2/38	6/38
NTNU1	1/2	0/2	0/2	0/2
UA-DLSI1	2/3	1/3	0/3	2/3
UA-DLSI2	0/3	0/3	0/3	0/3
UA-DLSI3	0/3	0/3	0/3	2/3
UA-DLSI4	1/3	0/3	0/3	2/3
UA-DLSI5	0/3	0/3	0/3	1/3
UA-DLSI6	1/3	0/3	0/3	0/3
ANOVA	25/38	12/38	10/38	21/38

Table 3: The entires in the table show the fraction of times each participant system significantly outscored the lead baseline in ROUGE-2, ROUGE-3, ROUGE-4 and MeMoG.