ACL 2013 MultiLing Pilot Overview

Jeff Kubina
U.S. Department of Defense
9800 Savage Rd., Ft. Meade, MD 20755
jmkubin@tycho.ncsc.mil

John M. Conroy, Judith D. Schlesinger
IDA/Center for Computing Sciences
17100 Science Dr., Bowie, MD
conroy@super.org, drj1945@gmail.com

Abstract

The 2013 Association for Computational Linguistics MultiLing Pilot posed a task to measure the performance of multilingual, single-document, summarization systems using a dataset derived from many Wikipedias. The objective of the pilot was to assess automatic summarization of multilingual text documents outside the news domain and the potential of using Wikipedia articles for such research. This report describes the pilot 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 subject of research and development. The ACM Digital Library has about 806 reports on the subject published since 1993, with over half of them appearing in the last five years. While the impetus for much of this research is the annual Text Analysis Conference (TAC) workshop on document summarization, there is a growing demand in the consumer market for news summarization applications being met by tablet and smart-phone applications such as Clipped[1], Summoner[2], TLDR[3], and Yahoo News. Yahoo and Google even acquired two companies developing such applications, Summly (Stelter, 2013) and Wavii (Tsotsis, 2013) respectively, earlier this year. While summarization technology for news sources is coming to fruition, the performance of such technology on non-English documents outside the news domain has not been throughly assessed and may need further research. Since the datasets used by the TAC summarization workshops have predominately been English news articles, with some exceptions (Giannakopoulos et al., 2011), the objective of the 2013 ACL MultiLing Pilot was to assess the performance of automatic multilingual single-document summarization systems on non-English text outside the news domain and to determine the potential of using Wikipedia articles for such research.

This report starts with a description of the task and dataset, the methods used to evaluate the submitted summaries, the performance of each participating system, and concludes with an assessment of the pilot and potential future work.

2 Task and Dataset Description

The objective of each participant system of the pilot was simple: compute a summary for each document in at least two of the datasets languages. No restrictions were placed on the languages that could be chosen nor was any target summary size specified.

The dataset was derived from a corpus created in 2010 to measure the performance of the CLASSY (Conroy et al., 2009) summarization algorithm on non-English documents outside the news domain. At the time such a corpus did not exist so one was created from the Wikipedias. To date there are Wikipedias in 285 languages comprising over 75 million pages. Some of the Wikipedias maintain a list of *Feature Articles,* which are articles reviewed and voted upon by editors as the best that fulfill Wikipedia's requirements in accuracy, neutrality, completeness, and style. One such requirement is that the article have a lead section that should

> ...be able to stand alone as a concise overview. It should ...summarize the most important points ...[and] material in the lead should roughly reflect its im-

[1] http://goo.gl/dFKD9
[2] http://goo.gl/0QFaZ
[3] http://goo.gl/qEgCs

Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pages 29–38,
Sofia, Bulgaria, August 9 2013.

portance to the topic . . . [4]

So the lead section of a featured article is an excellent summary of it, hence, the featured articles were used to create the corpus. In 2010 there were 41 Wikipedias with more than nine featured articles. The Perl module Text::Corpus::Summaries::Wikipedia[5] was developed to automatically create the corpus from the featured articles of those Wikipedias. The corpus is publicly available (Kubina, 2010) and the Perl module can be used to create an updated corpus.

The dataset for the pilot was created from a subset of the 2010 corpus. This was done to ensure that each language had 30 articles and that the size of each article's body text was sufficiently large. First, for each article the summary and body were compressed to approximate their information content size. For example, given a Chinese and English article with the same character length the Chinese article will usually contain more information than an English article and their compressed sizes will approximation their true information content. Next, if the compressed body size of an article was less than five times its compressed summary size, then the article was discarded. The factor of five was simply chosen to ensure the body of each article was sufficiently large relative to the summary size. For 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. This filtering reduced the corpus from 12,819 articles in 41 languages to the dataset containing 1,200 articles in 40 languages. For each language in the dataset Table 1 contains the mean size of the articles, their bodies, and their summaries, in characters.

3 Evaluation Methods and Results

Four teams submitted the results of six summarization systems. The teams are denoted by AIS, LAN, MD, and MUS; the MD team submitted three systems. Throughout this report the systems are denoted by AIS, LAN, MD1, MD2, MD3, and MUS. Table 2 contains the list of languages submitted for each system and the mean size, in characters, of the summaries submitted.

[4] http://en.wikipedia.org/wiki/Wikipedia:LEAD
[5] http://goo.gl/ySgOS

For the evaluation a baseline summary was extracted from the each article in the dataset that is the prefix substring of the article's body text with the same length as the text in the lead section of the article. For the remainder of this report the lead section of an article is called the *human summary*. An oracle summary was also computed for each article by heuristically extracting sentences from its body text to maximize its ROUGE-2 score against the human summary until its size exceeded the human summary, upon which it was truncated.

Submitted summaries were automatically evaluated against the human summary of each article using ROUGE-1, ROUGE-2 (Lin, 2004) and MeMoG (Giannakopoulos et al., 2008). For ROUGE, the languages Chinese, Japanese, Korean, and Thai were tokenized into individual characters. For MeMoG the character n-gram size used for each language is listed in Table 3, which is the n-gram size that maximized the standard deviation divided by the mean of the n-gram frequency distribution of the language in the dataset. So the selected n-gram size maximizes the variability of the distribution values relative to their mean. A shorter n-gram size would inflate the MeMoG scores because of their inherent frequent co-occurrence and conversely a longer size would penalize MeMoG scores due to their infrequent co-occurrence.

Each scoring method was performed twice, first by truncating, if necessary, each system summary to the size of the human summary, which is called HSS-scoring. The second set of scores were computed by truncating *all* the summaries of an article, including the human summary, to the size of the shortest summary amongst the system and human summaries for the article, which is called SSS-scoring. For HSS-scoring the system summaries shorter that the human summary are penalized since ROUGE is *recall oriented.* Alternately, SSS-scoring gives preference to shorter system summaries that have their best content (extracted sentences) first.

The performance for HSS-scoring of the systems on the seven languages that at least two teams submitted summaries for are given in Figures 1, 2, and 3. Table 4 gives an overview of how often significant differences in each of the three automatic metrics was observed. In particular, the last row gives the fraction of times that an non-parametric analysis of variance (ANOVA) indicated that the

Table 1: Dataset Languages and Sizes

ISO	LANGUAGE	ARTICLE	BODY	SUMMARY
af	Afrikaans	24752 (10214)	23448 (10230)	1303 (196)
ar	Arabic	27845 (9490)	26354 (9530)	1491 (220)
bg	Bulgarian	23965 (9248)	22981 (9250)	984 (134)
ca	Catalan	30611 (15248)	29322 (15274)	1289 (140)
cs	Czech	26300 (10453)	24777 (10414)	1522 (190)
de	German	32023 (12522)	31160 (12530)	862 (53)
el	Greek	26072 (11113)	24937 (11096)	1134 (224)
en	English	26572 (9010)	24860 (9013)	1712 (114)
eo	Esperanto	22295 (10031)	21304 (10022)	990 (106)
es	Spanish	40467 (19563)	38726 (19533)	1740 (113)
eu	Basque	17886 (9845)	17231 (9821)	655 (91)
fa	Persian	15132 (7630)	14099 (7217)	1032 (517)
fi	Finnish	27379 (11783)	26353 (11805)	1025 (105)
fr	French	41578 (21952)	40186 (21959)	1392 (73)
he	Hebrew	18492 (8283)	17697 (8283)	794 (82)
hr	Croatian	21132 (11094)	20276 (11113)	855 (96)
hu	Hungarian	26256 (12161)	25175 (12139)	1081 (90)
id	Indonesian	18550 (9131)	17649 (9124)	901 (148)
it	Italian	39189 (19235)	38042 (19220)	1146 (80)
ja	Japanese	14352 (11890)	14131 (11895)	221 (38)
ka	Georgian	15282 (9570)	14558 (9551)	723 (124)
ko	Korean	17140 (7899)	16416 (7889)	724 (175)
ml	Malayalam	27329 (10645)	26158 (10639)	1170 (331)
ms	Malay	19346 (16577)	18436 (16348)	909 (411)
nl	Dutch	29575 (16346)	28580 (16363)	994 (89)
nn	Norwegian-Nynorsk	16107 (8056)	15384 (7917)	722 (297)
no	Norwegian-Bokmal	30225 (17652)	29218 (17594)	1006 (125)
pl	Polish	23028 (12853)	22067 (12861)	960 (66)
pt	Portuguese	30967 (17998)	29310 (18004)	1657 (110)
ro	Romanian	21921 (12812)	20782 (12773)	1139 (108)
ru	Russian	34069 (13792)	33134 (13771)	934 (70)
sh	Serbo-Croatian	21776 (21469)	21060 (21341)	716 (308)
sk	Slovak	21694 (10067)	20983 (10071)	711 (169)
sl	Slovenian	17900 (7222)	17077 (7194)	823 (135)
sr	Serbian	30239 (9812)	28927 (9764)	1312 (176)
sv	Swedish	23476 (10169)	22314 (10156)	1162 (99)
th	Thai	27041 (8312)	25425 (8291)	1616 (226)
tr	Turkish	32956 (16423)	31346 (16338)	1610 (257)
vi	Vietnamese	35376 (16099)	33857 (16050)	1518 (161)
zh	Chinese	10110 (4341)	9608 (4357)	501 (42)

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 entire article, their bodies, and their summaries. For example, for English the mean size of the human summaries is 1,712 characters.

Table 2: Mean Summary Size For Submitted Languages of Systems

ISO	LANGUAGE	AIS	LAN	MD1	MD2	MD3	MUS	SUM
af	Afrikaans			966	953	967		1303
ar	Arabic		1461	876	858	874	2232	1491
bg	Bulgarian	1302		969	946	967		984
ca	Catalan			911	921	925		1289
cs	Czech			1061	1020	1062		1522
de	German	1492		1072	1037	1087		862
el	Greek	1367		989	979	991		1134
en	English	1262	1551	944	957	958	1197	1712
eo	Esperanto			947	933	956		990
es	Spanish			922	916	927		1740
eu	Basque			1154	1151	1167		655
fa	Persian			793	792	800		1032
fi	Finnish			1328	1284	1323		1025
fr	French			936	930	952		1392
he	Hebrew			871	867	876	1098	794
hr	Croatian			979	954	976		855
hu	Hungarian			1092	1064	1089		1081
id	Indonesian			1091	1085	1091		901
it	Italian			981	952	975		1146
ja	Japanese			546	564	563		221
ka	Georgian			1180	1195	1218		723
ko	Korean			663	638	656		724
ml	Malayalam			670	648	676		1170
ms	Malay			1089	1089	1098		909
nl	Dutch			994	974	1000		994
nn	Norwegian-Nynorsk			928	908	929		722
no	Norwegian-Bokmal			967	937	977		1006
pl	Polish			1086	1056	1083		960
pt	Portuguese			942	936	939		1657
ro	Romanian	1311		938	940	948		1139
ru	Russian			1095	1046	1078		934
sh	Serbo-Croatian			969	955	983		716
sk	Slovak			1026	997	1031		711
sl	Slovenian			967	949	981		823
sr	Serbian			990	954	979		1312
sv	Swedish			997	990	1006		1162
th	Thai			553	566	563		1616
tr	Turkish			1166	1132	1152		1610
vi	Vietnamese			696	684	691		1518
zh	Chinese			523	559	552		501

Table 2: The mean summary size, in characters, for each language submitted by each system including the mean of the human summaries in the last column named SUM.

Table 3: N-gram Size Per Language for MeMoG

ISO	LANGUAGE	SIZE	ISO	LANGUAGE	SIZE
af	Afrikaans	5	ka	Georgian	3
ar	Arabic	3	ko	Korean	1
bg	Bulgarian	4	ml	Malayalam	3
ca	Catalan	4	ms	Malay	4
cs	Czech	4	nl	Dutch	4
de	German	4	nn	Norwegian-Nynorsk	4
el	Greek	4	no	Norwegian-Bokmal	4
en	English	5	pl	Polish	4
eo	Esperanto	4	pt	Portuguese	4
es	Spanish	4	ro	Romanian	4
eu	Basque	4	ru	Russian	4
fa	Persian	4	sh	Serbo-Croatian	3
fi	Finnish	4	sk	Slovak	4
fr	French	4	sl	Slovenian	4
he	Hebrew	3	sr	Serbian	4
hr	Croatian	4	sv	Swedish	5
hu	Hungarian	4	th	Thai	3
id	Indonesian	5	tr	Turkish	5
it	Italian	5	vi	Vietnamese	5
ja	Japanese	1	zh	Chinese	1

Table 3: The table lists the n-gram size used for each language when evaluating summaries using MeMoG, which is the n-gram size that maximized the standard deviation divided by the mean of the n-gram frequency distribution of the language in the dataset.

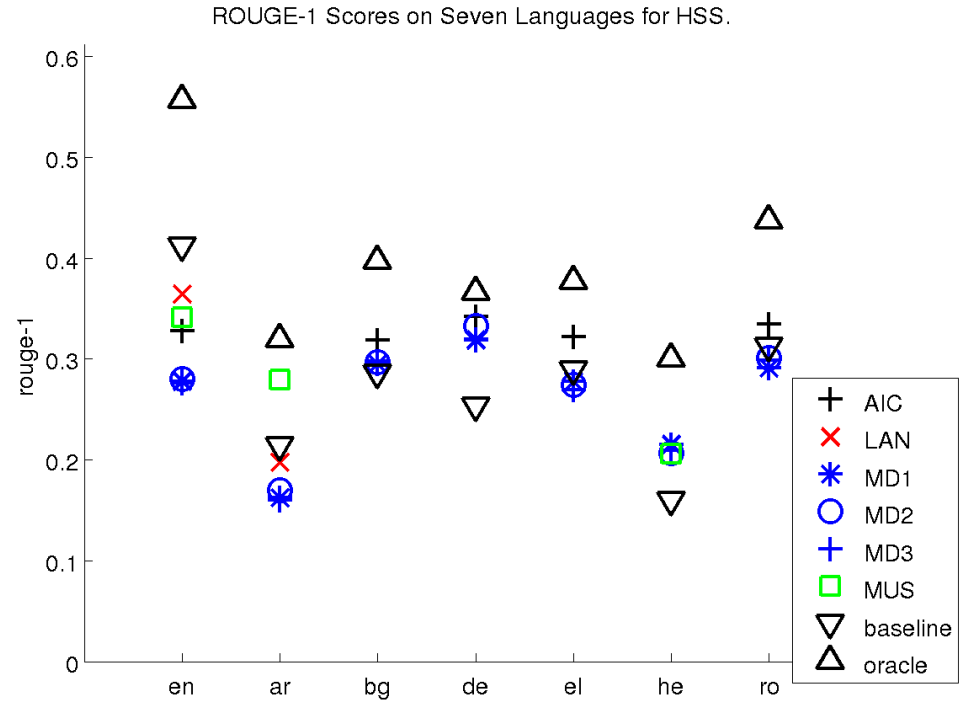

Figure 1: ROUGE-1 scores for HSS.

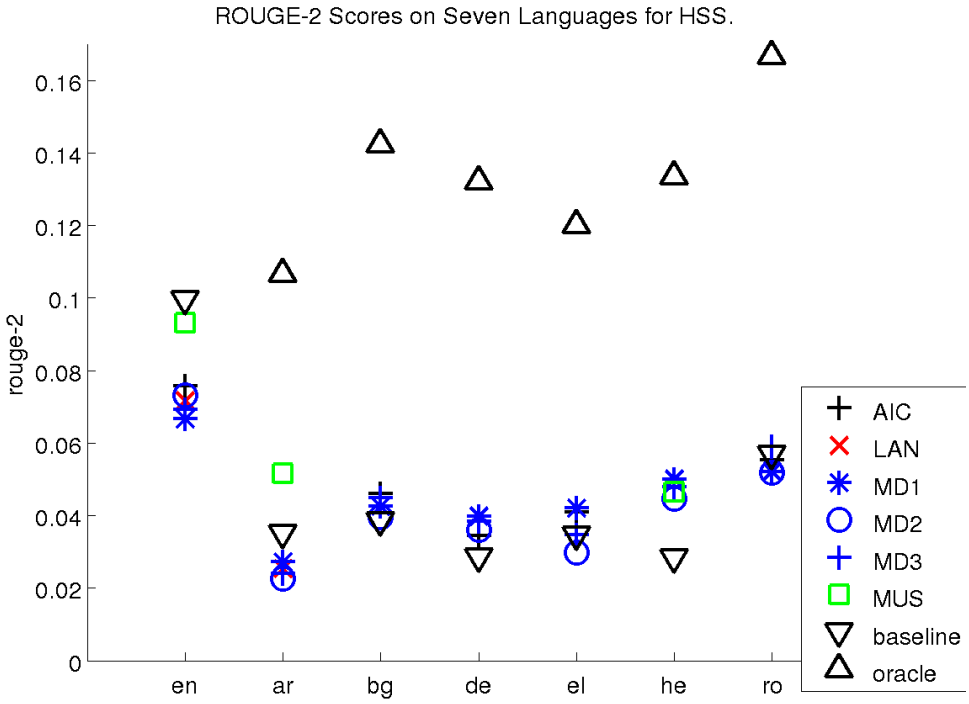

Figure 2: ROUGE-2 scores for HSS.

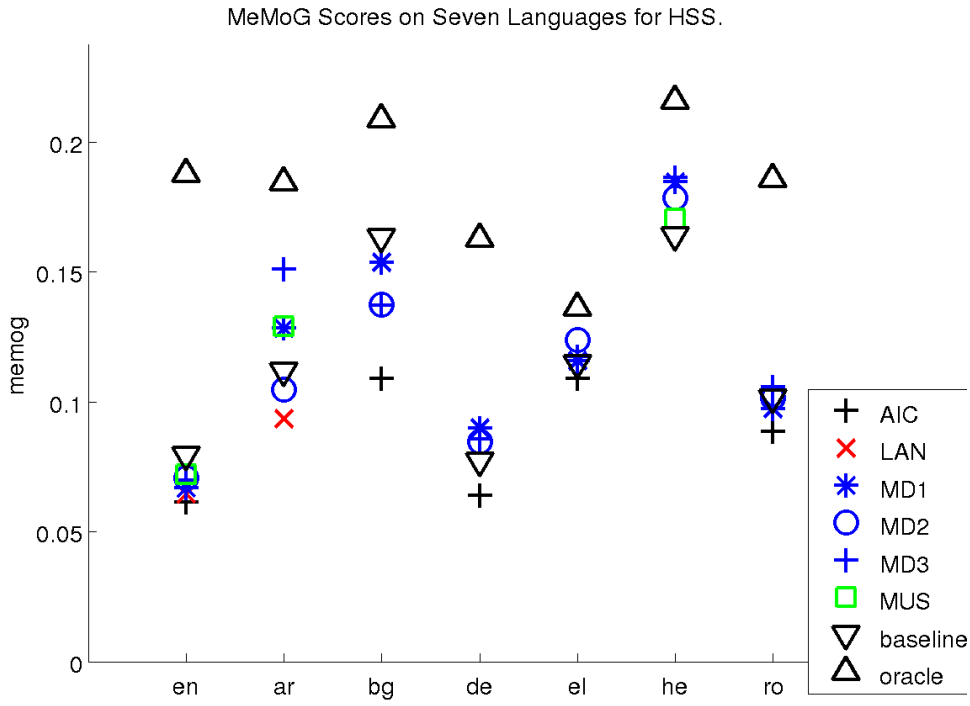

Figure 3: MeMoG scores for HSS.

Table 4: Fraction of time a system beat the baseline for HSS.

System	ROUGE-1	ROUGE-2	MeMoG
AIC	2/5	0/5	0/5
LAN	0/2	0/2	0/2
MD1	15/40	4/40	2/39
MD2	16/40	4/40	0/39
MD3	15/40	4/40	0/39
MUS	2/3	1/3	0/3
ANOVA	28/40	13/40	5/39

Table 4: The table gives the fraction of languages each system significantly outperform the baseline. The last line gives the number of times an ANOVA rejected the null hypothesis, indicating significance.

medians of the system scores were not the same, using a rejection threshold of 0.05. Also, the fraction of time that each system significantly outperformed the lead baseline is also recorded. A paired Wilcoxon test was invoked whenever the ANOVA indicated a significant difference was present, with a threshold of 0.05.

Lastly, each systems performance for SSS-scoring is provided in Figures 4, 5, and 6. Surprisingly, the results change little. Lastly Table 5 contains the number of times that each system beat the baseline summary with a 95% confidence measured as a result of the non-parametric ANOVA and the Wilcoxon paired sign rank test. The results show that the number of significant differences go down for ROUGE scores and up for MeMoG.

4 Summary

Overall, the authors believe the pilot was successful in that it exposed researchers to the potential for using Wikipedia articles for summarization research and demonstrated that generating summaries for the genre of Wikipedia articles is a more challenging task than newswire documents. Notably, no system outperformed the baseline for English! In hindsight this is not too surprising since news articles have a prose style[6] significantly different from Wikipedia articles[7]. Wikipedia articles are written as expositions having a topical flow that can vary significantly between sections but news articles are written in a style[8] that addresses the most important information first—the

[6] http://en.wikipedia.org/wiki/News_style
[7] http://en.wikipedia.org/wiki/MOS:
[8] http://en.wikipedia.org/wiki/Inverted_pyramid

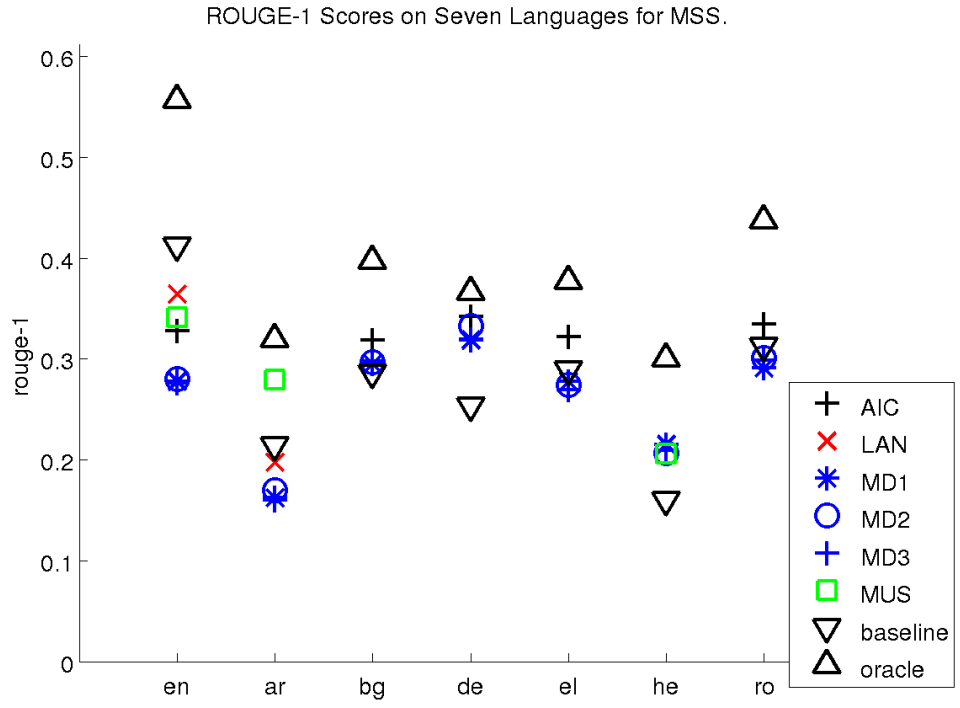

Figure 4: ROUGE-1 scores for SSS.

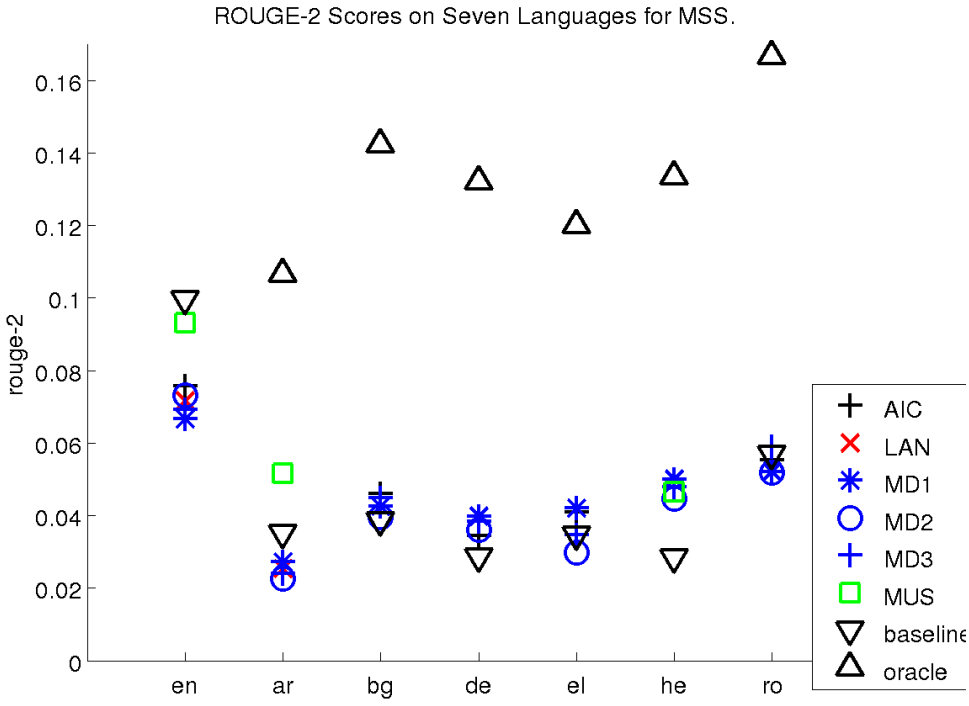

Figure 5: ROUGE-2 scores for SSS.

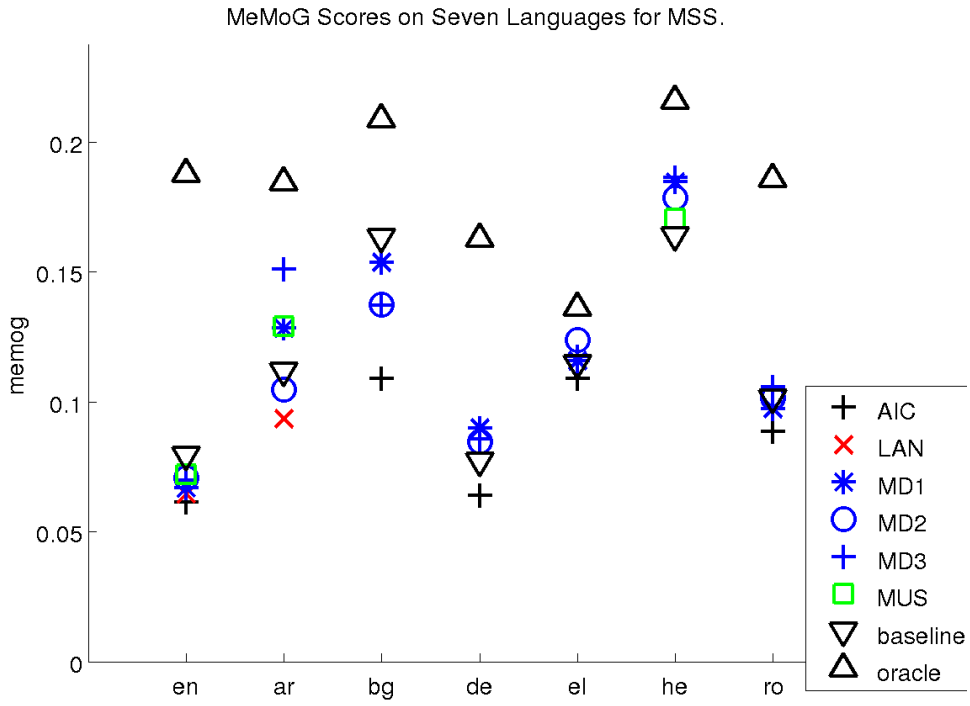

Figure 6: MeMoG scores for SSS.

Table 5: Fraction of the time a system beat the lead baseline for SSS.

System	ROUGE-1	ROUGE-2	MeMoG
AIC	0/5	0/5	0/5
LAN	0/2	0/2	0/2
MD1	8/40	2/40	6/39
MD2	10/40	2/40	2/39
MD3	7/40	2/40	4/39
MUS	0/3	0/3	0/3
ANOVA	11/40	5/40	7/39

Table 5: The table gives the fraction of languages that each system significantly outperform the baseline on. The last line contains the number of times an ANOVA rejected the null hypothesis, indicating significance.

who, what, when, where and why—with the subsequent text providing more details. Hence news articles have a more even topical flow. The authors hope these results stimulate research and development of summarization algorithms outside the news domain.

As for the metrics, ROUGE-1 observed the most significant differences among the systems and MeMoG observed the least as measured by a non-parametric ANOVA. However, a human evaluation of the summaries generated would be needed to detemine which of the automatic metrics is best at predicting significant differences among systems for such data.

References

John M. Conroy, Judith D. Schlesinger, and Dianne P. O'leary. 2009. Classy 2009: summarization and metrics. In *Proceedings of the text analysis conference (TAC)*.

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.

George Giannakopoulos, Mahmoud El-Haj, Benoît Favre, Marina Litvak, Josef Steinberger, and Va-

sudeva Varma. 2011. Tac 2011 multiling pilot overview.

Jeff Kubina. 2010. Wikipedia featured article corpus. http://goo.gl/AmMGN. [Online; accessed 30-May-2013].

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.

Brian Stelter. 2013. He has millions and a new job at yahoo. soon, he'll be 18. *New York Times*.

Alexia Tsotsis. 2013. Google buys wavii for north of $30 million. *TechCrunch*.