Multi-document multilingual summarization and evaluation tracks in ACL 2013 MultiLing Workshop

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

Abstract

The MultiLing 2013 Workshop of ACL 2013 posed a multi-lingual, multi-document summarization task to the summarization community, aiming to quantify and measure the performance of multi-lingual, multi-document summarization systems across languages. The task was to create a 240–250 word summary from 10 news articles, describing a given topic. The texts of each topic were provided in 10 languages (Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian, Spanish) and each participant generated summaries for at least 2 languages. The evaluation of the summaries was performed using automatic and manual processes. The participating systems submitted over 15 runs, some providing summaries across all languages. An automatic evaluation task was also added to this year's set of tasks. The evaluation task meant to determine whether automatic measures of evaluation can function well in the multi-lingual domain. This paper provides a brief description related to the data of both tasks, the evaluation methodology, as well as an overview of participation and corresponding results.

1 Introduction

The MultiLing Pilot introduced in TAC 2011 was a combined community effort to present and promote multi-document summarization aporaches that are (fully or partly) language-neutral. This year, in the MultiLing 2013 Workshop of ACL 2013, the effort grew to include a total of 10 languages in a multi-lingual, multi-document summarization corpus: Arabic, Czech, English, French, Greek, Hebrew, Hindi from the old corpus, plus Chinese, Romanian and Spanish as new additions. Furthermore, the document set in existing languages was extended by 5 new topics. We also added a new track aiming to work on evaluation measures related to multi-document summarization, similarly to the AESOP task of the recent Text Analysis Conferences.

This document describes:

- the tasks and the data of the multi-document multilingual summarization track;
- the evaluation methodology of the participating systems (Section 2.3);
- the evaluation track of MultiLing (Section 3).
- The document is concluded (Section 4) with a summary and future steps related to this specific task.

The first track aims at the real problem of summarizing news topics, parts of which may be described or happen in different moments in time. The implications of including multiple aspects of the same event, as well as time relations at a varying level (from consequtive days to years), are still difficult to tackle in a summarization context. Furthermore, the requirement for multilingual applicability of the methods, further accentuates the difficulty of the task.

The second track, summarization evaluation, is related the corresponding, prominent research problem of how to automatically evaluate a summary. While commonly used methods build upon a few human summaries to be able to judge automatic summaries (e.g., (Lin, 2004; Hovy et al., 2005)), there also exist works on fully automatic evaluation of summaries, without human "model" summaries (Louis and Nenkova, 2012; Saggion et al., 2010). The Text Analysis Conference has a separate track, named AESOP (e.g. see (Dang

Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pages 20–28,
Sofia, Bulgaria, August 9 2013. ©2013 Association for Computational Linguistics

and Owczarzak, 2009)) aiming to test and evaluate different automatic evaluation methods of summarization systems. We perform a similar task, but in a multilingual setting.

2 Multi-document multi-lingual summarization track

In the next paragraphs we describe the task, the corpus, the evaluation methodology and the results related to the summarization track of MultiLing 2013.

2.1 The summarization task

This MultiLing task aims to evaluate the application of (partially or fully) language-independent summarization algorithms on a variety of languages. Each system participating in the task was called to provide summaries for a range of different languages, based on corresponding corpora. In the MultiLing Pilot of 2011 the languages used were 7, while this year systems were called to summarize texts in 10 different languages: Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian, Spanish. Participating systems were required to apply their methods to a minimum of two languages.

The task was aiming at the real problem of summarizing news topics, parts of which may be described or may happen in different moments in time. We consider, similarly to MultiLing 2011 (Giannakopoulos et al., 2011) that news topics can be seen as *event sequences*:

Definition 1 *An event sequence is a set of atomic (self-sufficient) event descriptions, sequenced in time, that share main actors, location of occurence or some other important factor. Event sequences may refer to topics such as a natural disaster, a crime investigation, a set of negotiations focused on a single political issue, a sports event.*

The summarization task requires to generate a single, fluent, representative summary from a set of documents describing an event sequence. The language of the document set will be within the given range of 10 languages and all documents in a set share the same language. The output summary should be of the same language as its source documents. The output summary should be between 240 and 250 words.

2.2 Summarization Corpus

The summarization corpus is based on a gathered English corpus of 15 topics (10 of which were already available from MultiLing 2011), each containing 10 texts. Each topic contains at least one event sequence. The English corpus was then translated to all other languages (see also (Li et al., 2013; Elhadad et al., 2013)), trying to generate sentence-parallel translations.

The input documents generated are UTF8-encoded, plain text files. The whole set of translated documents together with the original English document set will be referred to as the *Source Document Set*. Given the creation process, the Source Document Set contains a total of 1350 texts (650 more than the corpus of the MultiLing 2011 Pilot): 7 languages (Arabic, Czech, English, Greek,) with 15 topics per language and 10 texts per topic for a total of 1050 texts; 3 languages (Chinese, French, Hindi) with 10 topics per language and 10 texts per topic for a total of 300 texts.

The non-Chinese texts had an average *word* length of approximately 350 words (and a standard deviation of 224 words). Since words in Chinese cannot be counted easily, the Chinese text length was based on the *byte* length of the corresponding files. Thus, Chinese texts had an average *byte* length of 1984 bytes (and a standard deviation of 1366 bytes). The ratio of average words in non-Chinese texts to average bytes in Chinese texts shows that on average one may (simplisticly) expect that 6 bytes of Chinese text are adequate to express one word from a European language.

We note that the measurement of Chinese text length in words proved a very difficult endeavour. In the future we plan to use specialized Chinese tokenizers, which have an adequately high performance that will allow measuring text and summary lengths in words more accurately.

2.3 Evaluation Methodology

The evaluation of results was perfromed both automatically and manually. The manual evaluation was based on the Overall Responsiveness (Dang and Owczarzak, 2008) of a text. For the manual evaluation the human evaluators were provided the following guidelines:

> Each summary is to be assigned an integer grade from 1 to 5, related to the overall responsiveness of the summary. We consider a text to be worth a 5, if

> it appears to cover all the important aspects of the corresponding document set using fluent, readable language. A text should be assigned a 1, if it is either unreadable, nonsensical, or contains only trivial information from the document set. We consider the content and the quality of the language to be equally important in the grading.

The automatic evaluation was based on human, model summaries provided by fluent speakers of each corresponding language (native speakers in the general case). ROUGE variations (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4) (Lin, 2004) and the AutoSummENG-MeMoG (Giannakopoulos et al., 2008; Giannakopoulos and Karkaletsis, 2011) and NPowER (Giannakopoulos and Karkaletsis, 2013) methods were used to automatically evaluate the summarization systems. Within this paper we provide results based on ROUGE-2 and MeMoG methods.

2.4 Participation and Overview of Results

This section provides a per-language overview of participation and of the evaluation results.

For an overview of participation information see Table 1. In the table, one can find the mapping between participant teams and IDs, as well as per language information. An asterisk in a cell indicates systems of co-organizers for the specific language. These systems had early access to the corpus for their language and, thus, had an advantage over others on that specific language.

Moreover, for the MultiLing pilot we created two systems, one acting as a global baseline (System ID6) and the other as a global topline (System ID61). These two systems are described briefly in the following paragraphs.

2.5 Baseline/Topline Systems

The two systems devised as pointers of a standard, simplistic approach and of an approach taking into account human summaries were implemented as follows.

The *global baseline system* — ID6 — represents the documents of a topic in vector space using a bag-of-words approach. Then it determines the centroid C of the document set in that space. Given the centroid, the system gets the text T that is most similar to the centroid (based on the cosine similarity) and uses it in the summary. If the text exceeds the summary word limit, then only a part of it is used to provide the summary. Otherwise, the whole text is added as summary text. If the summary is below the lower word limit, the process is repeated iteratively adding the next most similar document to the centroid.

The *global topline system* — ID61 — uses the (human) model summaries as a given (thus cheating). These documents are represented in the vector space similarly to the global baseline. Then, an algorithm produces random summaries by combining sentences from the original texts. The summaries are evaluated by their cosine similarity to the centroid of the model summaries.

We use the centroid score as a fitness measure in a genetic algorithm process. The genetic algorithm fitness function also penalizes summaries of out-of-limit length. Thus, what we do is that we search, using a genetic algorithm process, through the space of possible summaries, to produce one that mostly matches (an average representation of) the model summaries. Of course, using an intermediate, centroid representation, loses part of the information in the original text. Through this method we want to see how well we can create summaries by knowing a priori what (on average) must be included.

Unfortunately, the sentence splitting module of the topline, based on the Apache OpenNLP library[1] statistical sentence splitted failed due to a bug in our code. This resulted in an interesting phenomenon: the system would maximize similarity to the centroid, using fragments of sentences. This is actually an excellent way to examine what types of text can cheat n-gram based methods that they are good, while remaining just-not-good-enough from a human perspective. In the system performance analysis sections we will see that this expectation holds.

In the Tables of the following section we provide MeMoG and Overall Responsiveness (OR) statistics per system and language. We also provide information on statistically significant performance differences (based on Tukey HSD tests).

2.6 Language-specific Tables

The tables below illustrate the system performances per language. Each table contains three columns: 'Group', 'SysID' and 'Avg Perf'. The Group column indicates to which statistically

[1] See `http://opennlp.apache.org/`.

Participant	Run IDs	Arabic	Chinese	Czech	English	French	Greek	Hebrew	Hindi	Romanian	Spanish
Maryland	ID1, ID11, ID21	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CIST	ID2	✓	✓ *	✓	✓	✓	✓	✓	✓	✓	✓
Lancaster	ID3	✓ *			✓						
WBU	ID4	✓	✓	✓*	✓	✓	✓	✓	✓	✓	✓
Shamoon	ID5, ID51	✓			✓			✓*			
Baseline	ID6	Centroid baseline for all languages									
Topline	ID61	Using model summaries for all languages									

Table 1: Participation per language. An asterisk indicates a contributor system, with early access to corpus data.

Group	SysID	Avg Perf
a	ID61	0.2488
ab	ID4	0.2235
abc	ID1	0.2190
abc	ID11	0.2054
abc	ID21	0.1875
abc	ID2	0.1587
abc	ID5	0.1520
bc	ID51	0.1450
bc	ID6	0.1376
c	ID3	0.1230

Table 2: Arabic: Tukey's HSD test MeMoG groups

equivalent groups of performance a system belongs. If two systems belong to the same group, they do not have statistically significant differences in their performance (95% confidence level of Tukey's HSD test). The SysID column indicates the system ID and the 'Avg Perf' column the average performance of the system in the given language. The caption of each table indicates what measure was used to grade performance. In the Overall Responsiveness (OR) tables we also provide the grades assigned to human summarizers. We note that for two of the languages — French, Hindi — there were no human evaluations this year, thus there are no OR tables for these languages. At the time of writing of this paper, there were also no evaluations for human summaries for the Hebrew and the Romanian languages. These data are planned to be included in an extended technical report, which will be made available after the workshop at the MultiLing Community website[2], as an addenum to the proceedings.

There are several notable findings in the tables:

- In several languages (e.g., Arabic, Spanish) there were systems (notable system ID4) that

Group	SysID	Avg Perf
a	B	4.07
ab	C	3.93
ab	A	3.80
ab	ID6	3.71
ab	ID2	3.58
ab	ID3	3.58
ab	ID4	3.49
ab	ID1	3.47
abc	ID11	3.33
bcd	ID21	3.11
cde	ID51	2.78
de	ID5	2.71
e	ID61	2.49

Table 3: Arabic: Tukey's HSD test OR groups

Group	SysID	Avg Perf
a	ID4	0.1019
ab	ID61	0.0927
bc	ID2	0.0589
bc	ID1	0.0540
bc	ID11	0.0537
c	ID21	0.0256
c	ID6	0.0200

Table 4: Chinese: Tukey's HSD test MeMoG groups

Group	SysID	Avg Perf
a	B	4.47
a	C	4.30
a	A	4.03
b	ID2	3.40
c	ID4	2.43
c	ID61	2.33
c	ID21	2.13
c	ID11	2.13
c	ID1	2.07
d	ID6	1.07

Table 5: Chinese: Tukey's HSD test OR groups

[2] See http://multiling.iit.demokritos.gr/pages/view/1256/proceedings-addenum)

Group	SysID	Avg Perf
a	ID61	0.2500
a	ID4	0.2312
ab	ID11	0.2139
ab	ID21	0.2120
ab	ID1	0.2026
b	ID2	0.1565
b	ID6	0.1489

Table 6: Czech: Tukey's HSD test MeMoG groups

Group	SysID	Avg Perf
a	B	4.75
ab	A	4.633
ab	C	4.613
ab	D	4.215
b	E	4.1
c	ID4	3.129
d	ID1	2.642
d	ID11	2.604
de	ID21	2.453
e	ID61	2.178
e	ID2	2.067
f	ID6	1.651

Table 7: Czech: Tukey's HSD test OR groups

Group	SysID	Avg Perf
a	ID4	0.2220
a	ID11	0.2129
a	ID61	0.2103
ab	ID1	0.2085
ab	ID21	0.1903
ab	ID6	0.1798
ab	ID2	0.1751
ab	ID5	0.1728
b	ID3	0.1590
b	ID51	0.1588

Table 8: English: Tukey's HSD test MeMoG groups

Group	SysID	Avg Perf
a	A	4.5
a	C	4.467
a	B	4.25
ab	D	4.167
ab	ID4	3.547
b	ID11	3.013
b	ID6	2.776
bc	ID21	2.639
bc	ID51	2.571
bc	ID61	2.388
bc	ID5	2.245
bc	ID1	2.244
bc	ID3	2.208
c	ID2	1.893

Table 9: English: Tukey's HSD test OR groups

Group	SysID	Avg Perf
a	ID4	0.2661
ab	ID61	0.2585
ab	ID1	0.2390
ab	ID11	0.2353
ab	ID21	0.2180
ab	ID6	0.1956
b	ID2	0.1844

Table 10: French: Tukey's HSD test MeMoG groups

Group	SysID	Avg Perf
a	ID61	0.2179
ab	ID11	0.1825
ab	ID1	0.1783
ab	ID21	0.1783
ab	ID4	0.1727
b	ID2	0.1521
b	ID6	0.1393

Table 11: Greek: Tukey's HSD test MeMoG groups

Group	SysID	Avg Perf
a	A	3.889
a	ID4	3.833
a	B	3.792
a	C	3.792
a	D	3.583
ab	ID11	2.878
ab	ID6	2.795
ab	ID1	2.762
ab	ID21	2.744
ab	ID61	2.717
b	ID2	2.389

Table 12: Greek: Tukey's HSD test OR groups

Group	SysID	Avg Perf
a	ID61	0.219
ab	ID11	0.1888
ab	ID4	0.1832
ab	ID21	0.1668
ab	ID51	0.1659
ab	ID1	0.1633
ab	ID5	0.1631
b	ID6	0.1411
b	ID2	0.1320

Table 13: Hebrew: Tukey's HSD test MeMoG groups

Group	SysID	Avg Perf
a	ID11	0.1490
a	ID4	0.1472
a	ID2	0.1421
a	ID21	0.1402
a	ID61	0.1401
a	ID1	0.1365
a	ID6	0.1208

Table 14: Hindi: Tukey's HSD test MeMoG groups

Group	SysID	Avg Perf
a	ID61	0.2308
a	ID4	0.2100
a	ID1	0.2096
a	ID21	0.1989
a	ID11	0.1959
a	ID6	0.1676
a	ID2	0.1629

Table 15: Romanian: Tukey's HSD test MeMoG groups

Group	SysID	Avg Perf
a	ID4	4.336
ab	ID6	4.033
bc	ID11	3.433
c	ID1	3.329
c	ID21	3.207
c	ID61	3.051
c	ID2	2.822

Table 16: Romanian: Tukey's HSD test OR groups

Group	SysID	Avg Perf
a	ID4	0.2516
a	ID61	0.2491
ab	ID11	0.2399
ab	ID1	0.2261
ab	ID21	0.2083
ab	ID2	0.2075
b	ID6	0.187

Table 17: Spanish: Tukey's HSD test MeMoG groups

Group	SysID	Avg Perf
a	C	3.867
a	ID4	3.844
a	B	3.778
ab	A	3.667
abc	ID6	3.444
bc	ID2	3.067
c	ID11	3.022
c	ID1	2.978
c	ID21	2.956
c	ID61	2.844

Table 18: Spanish: Tukey's HSD test OR groups

reached human level performance.

- The centroid baseline performed very well in several cases (e.g., Spanish, Arabic), while rather badly in others (e.g., Czech).

- The cheating topline system did indeed manage to reveal a blind-spot of automatic evaluation, achieving high MeMoG grades, while performing badly in terms of OR grade.

We note that detailed results related to the performances of the participants will be made available via the MultiLing website[3].

3 Automatic Evaluation track

In the next paragraphs we describe the task, the corpus and the evaluation methodology related to the automatic summary evaluation track of MultiLing 2013.

3.1 The Evaluation Task

This task aims to examine how well automated systems can evaluate summaries from different languages. This task takes as input the summaries generated from automatic systems and humans in the Summarization Task. The output should be a grading of the summaries. Ideally, we would want the automatic evaluation to maximally correlate to human judgement.

3.2 Evaluation Corpus

Based on the Source Document Set, a number of human summarizers and several automatic systems submitted summaries for the different topics in different languages. The human summaries were considered model summaries and were provided, together with the source texts and the automatic summaries, as input to summary evaluation systems. There were a total of 405 model summaries and 929 automatic summaries (one system did not submit summaries for all the topics). Each topic in each language was mapped to 3 model summaries.

The question posed in the multi-lingual context is whether an automatic measure is enough to provide a ranking of systems. In order to answer this question we used the ROUGE-2 score, as well as the ”n-gram graph”-based methods (AutoSummENG, MeMoG, NPowER) to grade summaries. We used ROUGE-2 because it has been robust and highly used for several years in the DUC and TAC communities. There was only one additional participating measure for the evaluation track — namely the Coverage measure — in addition to the above methods.

In order to measure correlation we used Kendall's Tau, to see whether grading with the automatic or the manual grades would cause different rankings (and how different). The results of the correlation per language are indicated in Table 19. Unfortunately, the Hebrew evaluation data were not fully available at the time of writing and, thus, they could not be used. Please check the technical report tha twill be available after the completion of the Workshop for more information[4].

4 Summary and Future Directions

Overall, the MultiLing 2013 multi-document summarization and summary evaluation tasks aimed to provide a scientifically acceptable benchmark setting for summarization systems. Building upon previous community effort we managed to achieve two main aims of the MultiLing Pilot of 2011 (Giannakopoulos et al., 2011): we managed to increase the number of languages included to 10 and increase the number of topics per language.

We should also note that the addition of Chinese topics offered a fresh set of requirements, related to the differences of writing in this specific language from writing in the rest of the languages in the corpus: not even tokenization is easy to transfer to Chinese from other,e.g. European languages.

The main lessons learned from the multi-document and evaluation tracks were the following:

- multi-document summarization is an active domain of research.

- current systems seem to perform well-enough to provide more than basic, acceptable services to humans in a variety of languages. However, there still exist challenging languages.

- there are languages where systems achieved human-grade performance.

- automatic evaluation of summaries in different languages in far from an easy task. Much more effort must be put in this direction, to facilitate summarization research.

[3] See `http://multiling.iit.demokritos.gr`

[4] See `http://multiling.iit.demokritos.gr/pages/view/1256/proceedings-addenum)`

Language	R2 to OR	MeMoG to OR	Coverage to OR
Arabic	-0.11	0.00	-0.07
Chinese	**-0.38**	**0.46**	**0.41**
Czech	**0.38**	**0.30**	**0.26**
English	**0.22**	**0.24**	**0.26**
Greek	0.07	0.07	0.03
Romanian	**0.15**	**0.16**	**0.12**
Spanish	0.01	0.05	0.04
All languages	**0.12**	**0.18**	**0.14**

Table 19: Correlation (Kendall's Tau) Between Gradings. Note: statistically significant results, with p-value < 0.05, in **bold**.

The main steps we plan to take, based also on the future steps inherited from the MultiLing Pilot of 2011 are:

- to find the funds required for the evaluation process, in order to support the quality of the endeavour.
- to use the top performing evaluation system as the main evaluation measure in future MultiLing workshops.
- to create a piece of support software that will help implement and track all corpus generation processes.
- to study the possibility of breaking down the summarization process and asking systems to make individual components available as (web) services to other systems. This practice aims to allow combinations of different components into new methods.
- to check the possibility of using the corpus for cross-language summarization. We can either have the task of generating a summary in a different language than the source documents, or/and use multi-language source documents on a single topic to provide a summary in one target language.
- to start a track aiming to measure the effectiveness of multi-lingual summarization as a commercial service to all the world. This track would need a common interface, hiding the underlying mechanics from the user. The user, in turn, will be requested to judge a summary based on its extrinsic value. Much conversation needs to be conducted in order for this task to provide a meaningful comparison between systems. The aim of the track would be to illustrate the current applicability of multilingual multi-document summarization systems in a real-world task, aiming at non-expert users.

Overall, the MultiLing effort enjoys the contribution of a flourishing research community on multi-lingual summarization research. We need to continue building on this contribution, inviting and challenging more researchers to participate in the community. So far we have seen the MultiLing effort grow from a pilot to a workshop, encompassing more and more languages and research groups under a common aim: providing a commonly accepted benchmark setting for current and future multi-lingual summarization systems.

References

H. T. Dang and K. Owczarzak. 2008. Overview of the TAC 2008 update summarization task. In *TAC 2008 Workshop - Notebook papers and results*, pages 10–23, Maryland MD, USA, November.

Hoa Trang Dang and K. Owczarzak. 2009. Overview of the tac 2009 summarization track, Nov.

Michael Elhadad, Sabino Miranda-Jiménez, Josef Steinberger, and George Giannakopoulos. 2013. Multi-document multilingual summarization corpus preparation, part 2: Czech, hebrew and spanish. In *MultiLing 2013 Workshop in ACL 2013*, Sofia, Bulgaria, August.

George Giannakopoulos and Vangelis Karkaletsis. 2011. Autosummeng and memog in evaluating guided summaries. In *TAC 2011 Workshop*, Maryland MD, USA, November.

George Giannakopoulos and Vangelis Karkaletsis. 2013. Summary evaluation: Together we stand npower-ed. In *Computational Linguistics and Intelligent Text Processing*, pages 436–450. Springer.

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

G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, and V. Varma. 2011. TAC 2011 MultiLing pilot overview. In *TAC 2011 Workshop*, Maryland MD, USA, November.

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

Lei Li, Corina Forascu, Mahmoud El-Haj, and George Giannakopoulos. 2013. Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian. In *MultiLing 2013 Workshop in ACL 2013*, Sofia, Bulgaria, August.

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.

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

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