TAC 2011 MultiLing Pilot Overview

G. Giannakopoulos[*], NCSR Demokritos, Greece
M. El-Haj, University of Essex, UK
B. Favre, Aix-Marseille University, France
M. Litvak, Sami Shamoon Academic College of Engineering, Israel
J. Steinberger, JRC, Italy
V. Varma, IIIT Hyderabad, India

Abstract

The Text Analysis Conference MultiLing Pilot of 2011 posed a multi-lingual summarization task to the summarization community, aiming to quantify and measure the performance of multi-lingual, multi-document summarization systems. The task was to create a 240–250 word summary from 10 news texts, describing a given topic. The texts of each topic were provided in seven languages (Arabic, Czech, English, French, Greek, Hebrew, Hindi) and each participant generated summaries for at least 2 languages. The evaluation of the summaries was performed using automatic (AutoSummENG, Rouge) and manual processes (Overall Responsiveness score). The participating systems were 8, some of which providing summaries across all languages. This paper provides a brief description for the collection of the data, the evaluation methodology, the problems and challenges faced, and an overview of participation and corresponding results.

1 Introduction

Multi-document summarization has received the focus of attention for several years in the summarization community. Especially within the Text Analysis Conference (TAC) series of workshops multi-document summarization has a long history, including such tracks as update summarization, guided summarization and cross-lingual summarization. The missing piece in this mosaic of summarization-related research methods was multi-lingual, multi-document summarization.

The MultiLing Pilot introduced in TAC 2011 was a combined community effort to present and promote multi-document summarization apporaches that are (fully or partly) language-neutral. To support this effort an organizing committee across more than six countries was assigned to create a multi-lingual corpus on news texts, covering seven different languages: Arabic, Czech, English, French, Greek, Hebrew, Hindi.

This document describes: the task and the data (Section 2) of the pilot; the evaluation methodology of the participating systems (Section 3); the problems and challenges faced on the organizational aspect of the pilot (Section 4); the participation and system results, as an overview (Section 5). The document is concluded (Section 6) with a summary and future steps related to the MultiLing effort.

2 Task and Data

The task was aiming at the real problem of summarizing news topics, parts of which may be described or happen in different moments in time. We consider that news topics can be seen as *event sequences*: 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.

[*] Full author names: George Giannakopoulos, Mahmoud El-Haj, Benoît Favre, Marina Litvak, Josef Steinberger, Vasudeva Varma

The task for the MultiLing Pilot was defined as follows.

> The 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 will be called to provide summaries for a range of different languages, based on corresponding corpora. Participating systems will be required to apply their methods on a minimum of two languages. Evaluation will favour systems that apply their methods in more languages.
>
> The MultiLing 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 a given range of 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 250 words at most.

The target summary size was finally set between 240 and 250 words. The aim of this target size was to allow systems to include into a summary several aspects of a topic (e.g., the event and the reasons behind it). It is interesting in itself to see how much overlap there would be in what summarizers consider important information.

To support the defined task, we needed a dataset of freely available news texts (to allow reuse), covering news topics that would contain event sequences. We determined that each event sequence in the corpus should contain at least three distinct atomic events. The dataset created was based on the WikiNews site[1], which covers a variety of news topics, while allowing the reuse of the texts based on the Creative Commons Licence. An example topic with two sample texts derived from the original WikiNews documents is provided in Figure 1.

The creation of the corpus was broken down to the following individual steps:

English texts selection We first gathered an English corpus of 10 topics, each containing 10 texts. We made sure that each topic contained at least one event sequence. From the original HTML text we only kept unformatted content text, without any images, tables or links.

Translation The English texts were translated using a sentence-by-sentence approach to each of the other languages: Arabic, Czech, French, Greek, Hebrew, Hindi.

This whole set of documents was considered to be the *Source Document Set*. Given the creation process, the Source Document Set contains a total of 700 texts: 7 languages, 10 topics per language, 10 texts per topic.

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, as described below, and the automatic evaluation used the ROUGE and AutoSummENG-MeMoG methods to provide a grading of performance.

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 (ROUGE1, ROUGE2, ROUGE-SU4) [Lin, 2004] and the MeMoG variation [Giannakopoulos and Karkaletsis, 2010] of AutoSummENG [Giannakopoulos

[1] See http://www.wikinews.org.

```
2005/01/08 Tsunami aid donations in 2005 deductible for 2004 in the U.S.
Saturday, January 8, 2005
U.S. citizens donating in 2005 to help tsunami victims may write off their donations
on their 2004 tax returns, thanks to a bill quickly passed in the U.S. House of
Representatives and the U.S. Senate on a voice vote, and signed into law by president
George W. Bush.
Without the new law, contributors would have waited until 2006 and their 2005 tax
returns to be able to write off their charitable donations. The law is intended to
promote donating towards the tsunami relief effort.
CBS News reports Indiana University’s Center on Philanthropy is estimating approximately
322 million U.S. dollars in goods and cash have been donated by private U.S. citizens
and corporations, in addition to the 350 million that was promised by the government.
An AP/ISOS poll has found three in ten U.S. citizens have donated to Tsunami Aid
organizations.
```

(a) Text 1

```
Aid pledges rise; Japan promises 500,000,000 USD
Saturday, January 1, 2005
In an abrupt about-face, the world’s wealthiest nations have begun pouring funding
into the Earthquake/Tsunami damaged region. Promised funds have doubled in the past
24 hours, to nearly 2 Billion U.S. dollars (USD).
Japan tops the U.S.
After the U.S. increased it’s funding donation to 350 million USD, Japanese Prime
Minister Junichiro Koizumi announced a half-billion dollar donation on Saturday,
Jan. 1. China has promised 60.5 million USD, after Japan and the U.S., the United
Kingdom and Sweden for largest single-nation donation. Norway increased it’s funding
donation to 180 million USD
U.N. warns of delays
Despite the encouraging promises, the UN Office for the Coordination of Humanitarian
Affairs in Indonesia chief, Michael Elmquist, warned that logistics of securing
the funds, purchasing supplies and shipping them to stricken regions will take time,
possibly weeks. In the meantime, the confirmed death toll will continue to climb,
as may deaths due to dehydration, disease, and starvation.
```

(b) Text 2

Figure 1: Topic Sample (Indian Ocean Tsunami)

et al., 2008] were used to automatically evaluate the summarization systems.

As indicated in the task, the acceptable limits for the word count of a summary were between 240 and 250 words[2] (inclusive). However, some submissions included texts outside the word limit. To avoid rejecting these summaries completely, while penalizing their out-of-limit word count we devised and used the Length-Aware Grading measure (LAG). Given a summary S of length $|S|$ (in words) assigned a grade g, a lower word limit count l_{min} and an upper word limit count l_{max}, then LAG is defined as follows:

$$LAG(g,S) = g * \left(1 - \frac{\max(\max(l_{min}-|S|,|S|-l_{max}),0)}{l_{min}}\right)$$

In our specific evaluation, $l_{min} = 240$, $l_{max} = 250$. LAG simply provides a linearly diminishing weight to grades diverging from the limits. We note that, for extreme text sizes ($|S| > l_{min}+l_{max}$), LAG may even have a negative value. Of course, such a case never appeared in the MultiLing pilot. The LAG function was applied to the Overall Responsiveness score in the analysis of performance, therefore LAG in the following sections implies LAG of the Overall Responsiveness.

4 Problems and Challenges

During the creation of the corpus, many of the difficulties faced across all languages were related to human subjectivity and point-of-view. Others, were related to the specifics of a single language.

For example, in the Hebrew language setting, human experts have encountered several problems during the corpus preparation (in translation, summarization, and evaluation parts, respectively). The first concerns translation of names, acronyms, idioms, and foreign ranks/positions:

- A person's name should most probably be translated phonetically, but it is not as clear when it comes to names of places and objects, especially those containing proper words-New-York should probably be named ניו-יורק, but should "Dome of the Rock" be named דום-אוף-דה-רוק ? Should there be a criterion besides what 'sounds right' or what is commonly used?

- The same goes for acronyms, with UN and GMT as opposite examples. Additionally, they can be translated phonetically or using equivalent letters, which also goes for ranks and positions: should "Corporal" be translated as רב"ט or קורפורל ? An equivalent doesn't always exist, as is in the case of "Specialist", making it harder to choose a guiding rule and maintain consistency.

- Idioms should be translated to idioms with similar meanings, which requires some creativity (watch some TV series with subtitles to see how badly this can come out). This can turn out to be very difficult with certain idiom-language combinations (though irrelevant if the final outcome is a summarization).

The second problem concerns the handling of time orientations within a series of texts. An event dating between different texts will be referenced in a different tense in each one. Should the summary be written from the point of view of the latest text, a fixed-in-the-future point, or at the most referenced point in time (so a single, later text, describing few events, won't determine the description used for a large number of the previous events). This becomes more difficult with vaguely described time intervals, where an event might or might not occur later than latest text.

Third, it was not obvious how the summaries should be evaluated: whether summary should cover all news articles in a given set, whether it should cover all the relevant events described in the set or just include the most significant ones, how to decide about the significance of the described events in the set, etc. These uncertainties caused a very wide range of grades for most of submitted systems and human experts.

This problem is dual to another problem in the whole creation and evaluation process: how can one define that would avoid producing bias. It was important to allow people to summarize as they saw fit. However, it may be the case that no one summarizes without previous bias. As Karen Sparck-Jones [Jones, 2007] mentioned

[2] The count of words was provided by the *wc -w* linux command.

> The need to cross the old intrinsic/extrinsic boundary and address summary purpose more directly is clear.

Thus, the first question posed is “do we need to define a more specific purpose to be able to evaluate better”? Furthermore, is there a generic information need, or do people always define a specific purpose for their summarization needs? What about emergent topics, i.e., topics that emerge when one reads a set of related documents? May it not be the case that important points emerge *while* reading the documents?

In the evaluation of summaries there existed languages where human (model) summaries were graded as really bad (e.g., see Arabic language overview in Section 5). It is really challenging to determine whether different user needs (each user estimates importance differently) have caused bad grades for some human summaries. Another possibility, also discussed with Greek and English evaluators, is that people tend to be more strict with human peers than with automatic peers. Even though evaluators were not told about which systems were human and which not, it was almost trivial for them the nature of the summarizer in many cases.

In the case of Greek and English, this imbalance was avoided by a direct guideline for evaluation: “Never take into account whether a system is a human or a machine. It may be the case that we asked a human to follow an automated process to create the summary. Thus, one should grade only based on the coverage and fluency criteria, as described in the original guidelines”.

In the Greek subcorpus, there was also the case where summaries showed that the summarizer had not really understood the sequence of events or had partially misinterpreted statements. This kind of summary error may be very tiresome for an evaluator to pinpoint. Such errors also make grading more difficult (since the impact may be from trivial to severe to the summary coherence).

There was also another question about whether out-of-limit summaries should be truncated or not. We decided to penalize out-of-limit summaries (using the LAG function) and not reject them, because in the real world it may make sense to (slightly) bend the rules to provide (significantly) better information.

Overall, the multi-lingual summarization problems do not vary a lot from the single language summarization. The major new challenges are: the problem of (ambiguity in) translation; the summary word limits — which may need to vary across languages; and, very importantly, the organizational burden of supplying a corpus across a multitude of languages. The rest of the problems are common with existing summarization research: guidelines, evaluation and human subjectivity can cause problems in the evaluation of systems, however within an acceptable margin given enough evaluators.

In the following paragraphs we provide the overview of participation in the MultiLing pilot of TAC 2011, also discussing briefly the performance on the peer systems.

5 Participation and Overview of Results

This section provides a per-language overview of participation and of the evaluation results. In each language — unless otherwise indicated — there exists a “topline” system, provided by the co-organizer for that specific language. The advantage that these systems were provided with was the knowledge of the corpus before the submission of the 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. In the *Notes* column we indicate which systems are co-organizers (indicated as *Coorg*) for a specific language — thus, having an advantage over others on that specific language — and which are not (indicated as *Peer*).

Moreover, for the MultiLing pilot we created two systems, one acting as a global baseline (System ID9) and the other as a global topline (System ID10). These two systems are described briefly in the following paragraphs.

5.1 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* — ID9 — represents the documents of a topic in vector space using a bag-

Participant	System ID	Arabic	Czech	English	French	Greek	Hebrew	Hindi	Notes
CIST	ID1	✓	✓	✓	✓	✓	✓	✓	Peer
CLASSY	ID2	✓	✓	✓	✓	✓	✓	✓	Peer
JRC	ID3	✓	✓	✓	✓	✓	✓	✓	Coorg (Czech)
LIF	ID4	✓	✓	✓	✓	✓	✓	✓	Coorg (French)
SIEL_IIITH	ID5			✓	✓			✓	Coorg (Hindi)
TALN_UPF	ID6	✓		✓	✓			✓	Peer
UBSummarizer	ID7	✓	✓	✓	✓	✓	✓	✓	Peer
UoEssex	ID8	✓		✓					Coorg (Arabic)
Baseline	ID9	Centroid baseline for all languages							Coorg (All)
Topline	ID10	Using model summaries for all languages							Coorg (All)

Table 1: Participation per language

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* — ID10 — uses the (human) model summaries as a given (thus cheating). These documents are represented using n-gram graphs and merged into a representative graph (see the MeMoG method [Giannakopoulos and Karkaletsis, 2010] for more). Then, an algorithm produces random summaries by combining sentences from the original texts. The summaries are evaluated by their MeMoG score with respect to the model summaries. In other words, the more similar the n-gram graph of the random summary is to the merged model graph of the model summaries, the better it is considered.

We use the MeMoG 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, average 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.

In the following sections we provide more details per language, related both to the organizational part and the performance of the participating systems. We also provide information on statistically significant performance differences (based on Tukey HSD tests). We provide HSD tests for the original grades in the subsections. To further elaborate on the comparsion between systems, we have also included in the Appendix (Section A) LAG-based HSD tables for systems only.

5.2 Arabic Language

The preparation of the Arabic corpus for the TAC–2011 MultiLing Summarisation Pilot was organised by the university of Essex. A number of 12 people participated in translating the English corpus into Arabic and in summarising the set of related Arabic articles. The participants were paid using Amazon vouchers. The amount of the vouchers varies depending on the task performed. The total amount of amazon vouchers paid to the participants was £250 as 3 of the participants volunteered to do the tasks.

For the Arabic language, there were 7 participants (peers) in addition to the two baseline systems, for a total of 9 runs. According to the results the baseline performed better than the topline; so do at least four of the peers (ID1, ID3, ID7, ID8).

The average time for reading the English news articles by the Arabic native speakers participants was 4.76 minutes. The average time it took them to translation those articles into Arabic is 25.36 minutes, and to validate each of the translated Arabic articles

Group	SysID	Avg Perf
a	B	4.04
ab	ID1	3.77
ab	ID9	3.73
ab	ID8	3.70
abc	ID3	3.43
abc	ID7	3.30
bc	ID10	3.20
bc	ID2	3.10
bc	ID6	3.10
cd	ID4	2.77
de	C	2.23
e	A	1.92

Table 2: Arabic: Tukey's HSD test groups

Group	SysID	Avg Perf
a	B	4.89
a	C	4.80
a	A	4.73
a	D	4.50
b	ID3	3.40
b	ID9	3.30
bc	ID1	3.00
cd	ID2	2.70
cd	ID10	2.68
d	ID7	2.20
e	ID4	1.48

Table 3: Czech: Tukey's HSD test groups

the participants took 6.07 on average.

For the summarisation task the average time for reading the set of related articles (10 articles per each set) was 17.2 minutes. The average time for summarisation process of each set was 24.03 minutes.

Figures 2a, 2b illustrate the overall responsiveness and LAG of all the systems including the human peers. We believe that the reason behind the difference between the human and computer peers grades is due to the evaluators expectations. As in the first task the participants were aware that the summaries they are to evaluate are human summaries generated by native Arabic speakers, the same for the second task where the participants were aware they are evaluating system summaries, therefore the expectations varies between high for human and low for system summaries.

5.3 Czech Language

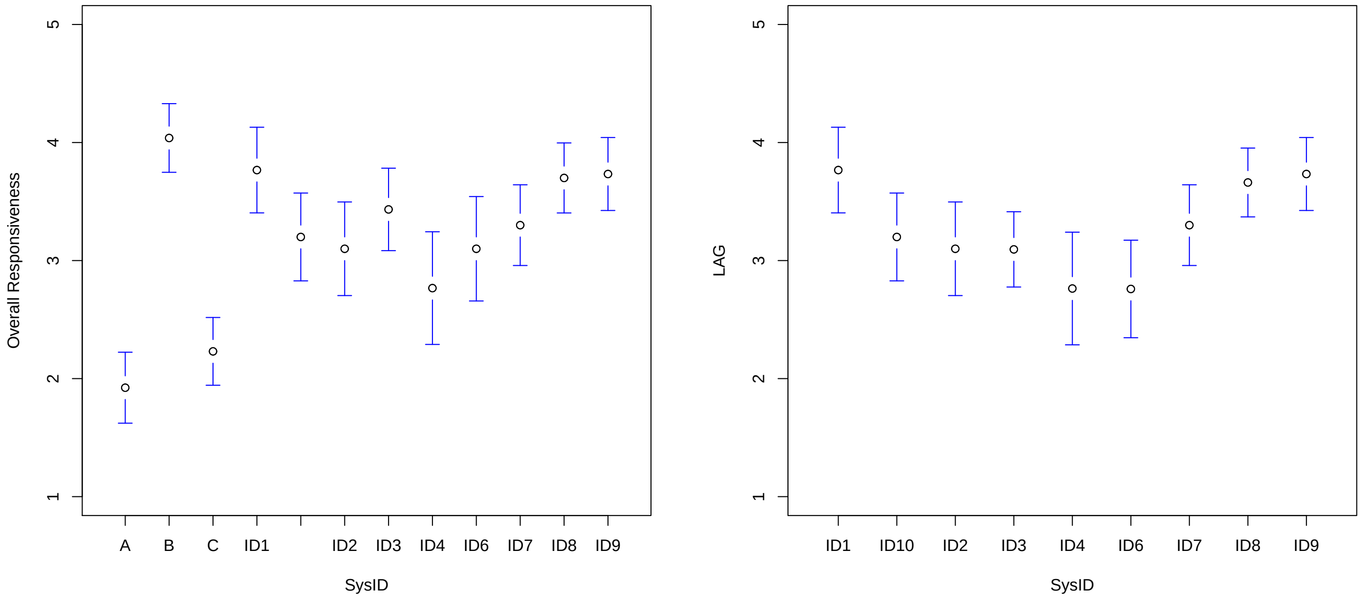
There were 5 participating systems for the Czech language, together with the two baselines systems it made a total of 7 runs. The system with ID3 was submitted by the group working on the evaluation of the Czech part of the task. Thus, it serves as a baseline only. An overview of the Overall Responsiveness and the corresponsing Length Aware Grade (LAG) of these participants can be seen in Figures 3a and 3b.

We see that all human summarizers performed significantly better than all systems. The centroid baseline (ID9) performed well, mainly because it contains continuous text and thus it's readability part of the responsiveness score is high. On the other hand, systems had problems with readability, mainly because of incorrect sentence splitting and shuffled sentence order. Baseline ID3 received highest grades among the system runs, however, it was penalized in the length-aware grading bacause its summaries were several times shorter then 240 words. The penalty moved this system after the centroid baseline. Three systems (ID2, ID4, ID7) and surprisingly the top baseline ID10 as well performed significantly worse than the two baselines (ID3 and ID9). System ID1 was ranked between them, but it was within statistical uncertainty equivalent to the top system runs (see table 3).

5.4 English Language

3 summarizers were allocated for the English subcorpus human summarization. According to some preliminary measurements, the reading time per topic set (10 documents) was 20min (standard deviation 7min), whereas the summarization itself took 48min on average (with a standard deviation of 13min).

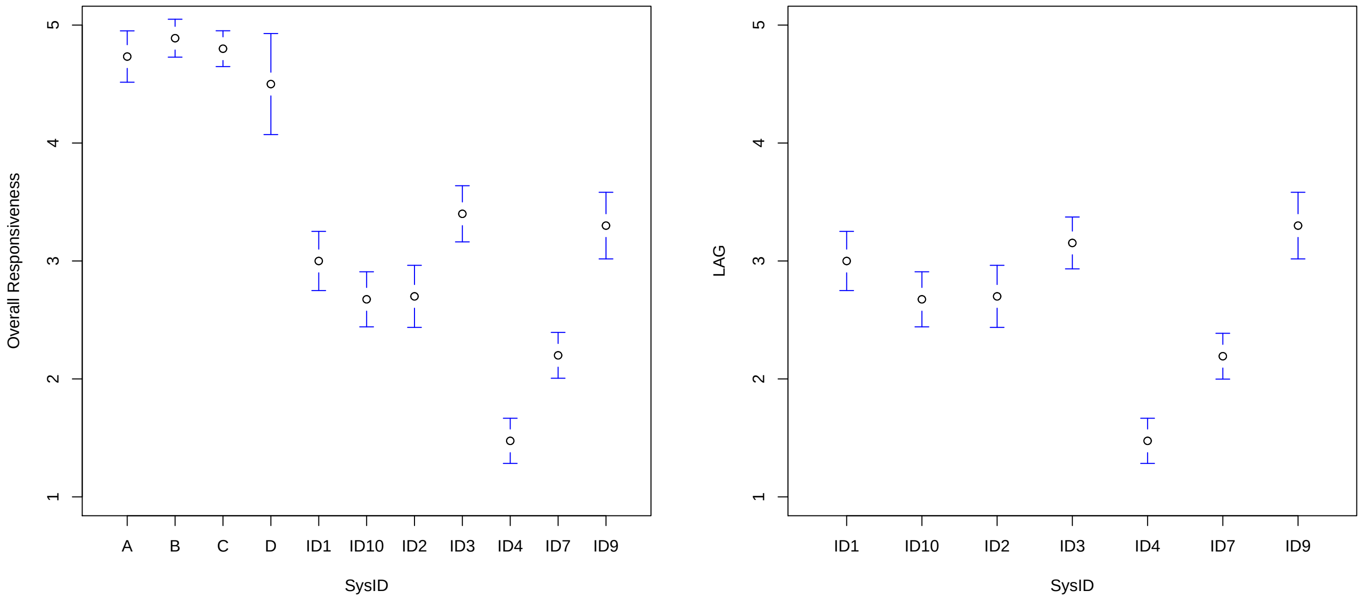
The English part of the MultiLing pilot had the highest participation, because summarization on English texts has been the focus of attention of the summarization community for several years. There were 8 systems participating for the English language, which added to the two baselines systems made a total of 10 runs. There was no system that acted as



(a) Overall Responsiveness — All Peers

(b) LAG for Overall Responsiveness — Systems Only

Figure 2: Arabic: Performance overview



(a) Overall Responsiveness — All Peers

(b) LAG for Overall Responsiveness — Systems Only

Figure 3: Czech: Performance overview

Group	SysID	Avg Perf
a	A	4.37
a	B	4.27
ab	ID3	3.83
abc	C	3.55
abc	ID2	3.53
bcd	ID1	3.20
bcd	ID10	3.20
bcd	ID5	3.03
cde	ID8	2.73
cde	ID6	2.67
de	ID9	2.50
de	ID7	2.30
e	ID4	2.03

Table 4: English: Tukey's HSD test groups

a language-specific baseline for the English subcorpus.

In Figures 4a, 4b we provide an overview of the system performances on the English subcorpus. We can see that systems ID2 and ID3 did remarkably well, performing similarly to human summarizers (see also Tukey's HSD test in Table 4). The baseline system (ID9) performed rather bad, as expected. The topline system (ID10) was good-enough but was — on average — outperformed by two peers (ID2, ID3), even if not significantly. System ID1 also performed well, its performance lying extremely close to the topline.

5.5 French Language

5 peers were involved in the creation of the gold-standard summaries for French. The average document-level reading was 4 minutes. It took on avearge 28.8 minutes to translate each document and then 9.95 minutes to verify them. The average cluster-level reading time for the summarization task was 18.82 minutes. The average summarization time was 35.12 minutes. Human-written summary evaluation time was not measured properly but can be estimated at about 3 to 5 minutes per summary including a portion of the overhead for reading the set of documents. A rough estimate is that a little less than 100 hours of human time were required to create the corpus and evaluate the submissions.

Group	SysID	Avg Perf
a	F	4.67
ab	C	4.33
ab	A	4.22
ab	D	4.17
abc	E	3.89
bc	B	3.47
c	ID3	3.23
d	ID1	2.30
de	ID2	2.20
de	ID6	2.20
de	ID10	2.10
de	ID7	2.07
de	ID9	2.03
de	ID5	1.90
e	ID4	1.33

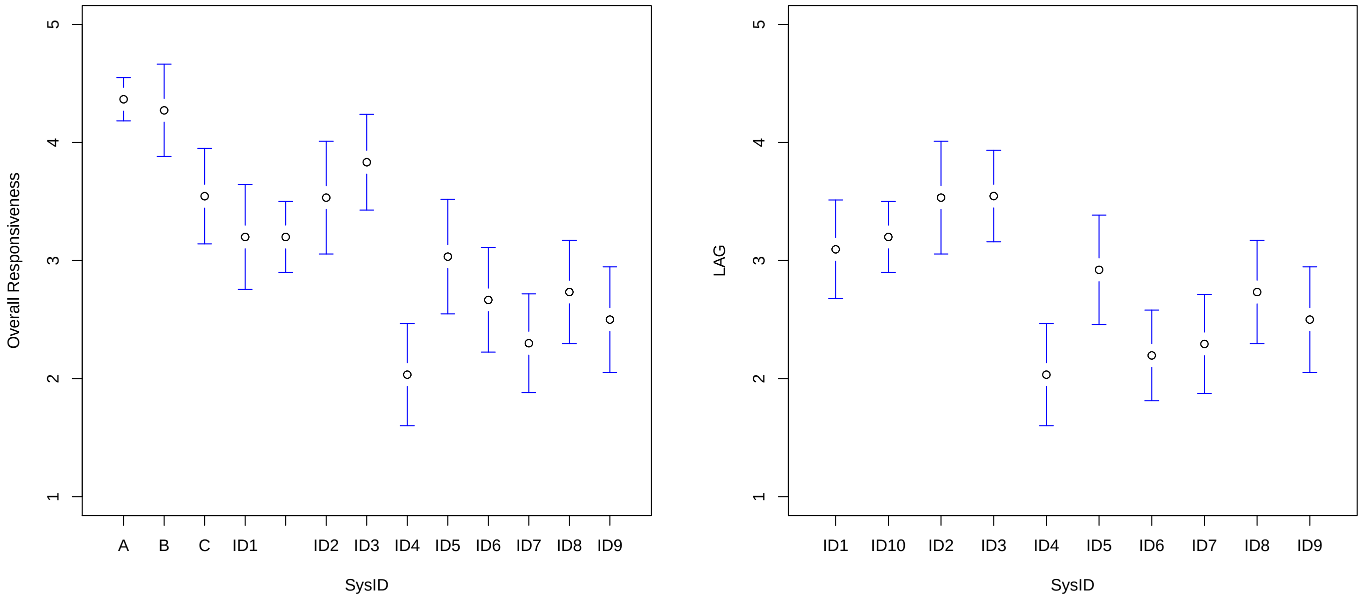
Table 5: French: Tukey's HSD test groups

Systems ID1, ID2, ID3, ID4, ID5, ID6, ID7, the baseline and the topline were evaluated by human judges. Results are illustrated in Figures 5a, 5b, where Overall Responsiveness and LAG are depicted. Intrestingly the topline (ID10) performed worse than some of the systems, in particular ID3. This is due to the fact the ID10 was generated by pasting sentences together without spaces, which created typographic artifacts, and therefore degrading the perceived quality of the summary.

In Table 5 we provide the groups deterined based on a Tukey's HSD test.

5.6 Greek Language

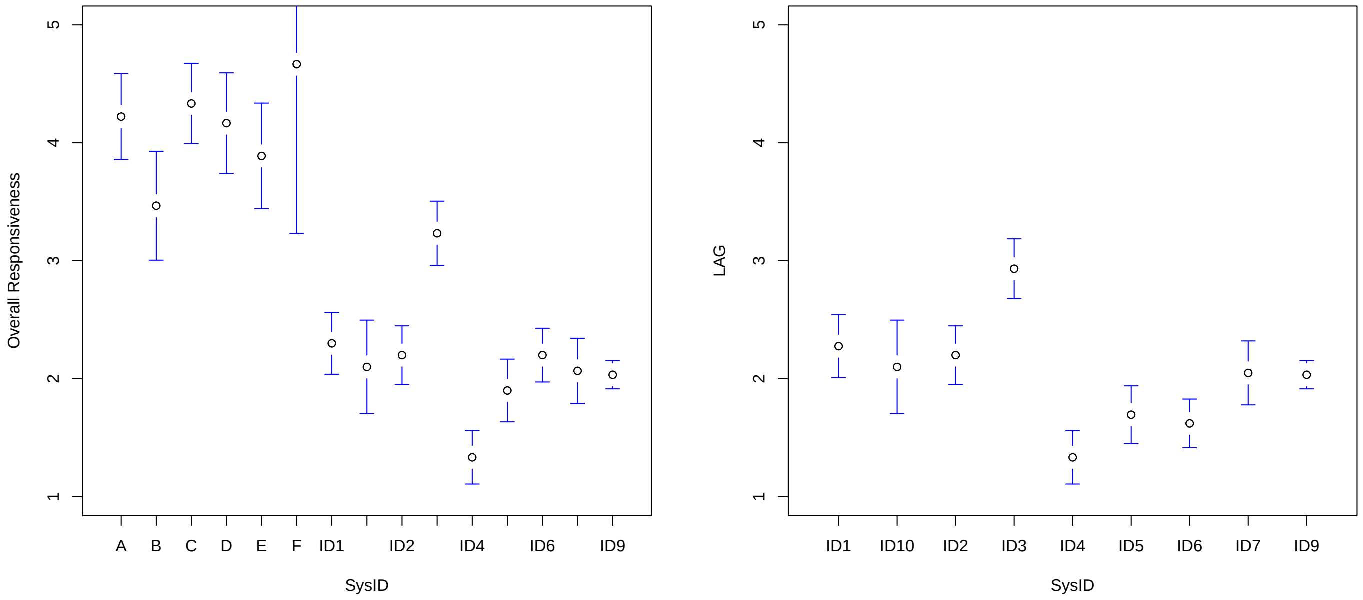
3 translators were employed for the translation from the English texts. Per text, the reading time averaged at 4.7min (with a significant standard deviation of 3.7min). The translation time was on average 33.6min (with a standard deviation of 18min). Thus, per topic (10 documents), approximately 7 hours were required for translation, including verification. 3 summarizers were allocated for the Greek Language human summarization. According to some preliminary measurements, the reading per topic set (10 documents) took an average of 22min (with a standard deviation of 8min), whereas the summarization itself took 55mins on average (with a standard deviation of 24min) . Overall, an approximate total of 110h were required for the creation of the



(a) Overall Responsiveness — All Peers

(b) LAG for Overall Responsiveness — Systems Only

Figure 4: English: Performance overview



(a) Overall Responsiveness — All Peers

(b) LAG for Overall Responsiveness — Systems Only

Figure 5: French: Performance overview

Group	SysID	Avg Perf
a	A	4.43
ab	B	3.93
abc	ID3	3.63
bc	C	3.41
bc	ID2	3.33
bc	ID10	3.30
bc	ID9	3.13
c	ID1	3.00
d	ID7	2.10
d	ID4	1.97

Table 6: Greek: Tukey's HSD test groups

Group	SysID	Avg Perf
a	A	4.56
ab	ID3	3.87
bc	B	3.56
bc	C	3.56
bc	ID1	3.29
bc	ID2	3.29
bc	ID9	3.16
bc	ID7	3.06
c	ID10	3.03
d	ID4	2.19

Table 7: Hebrew: Tukey's HSD test groups

corpus.

For the Greek language, there were 5 participants (peers), plus the two baselines systems, for a total of 7 runs. An overview of the Overall Responsiveness and the corresponsing Length Aware Grade (LAG) of these participants can be seen in Figures 6a, 6b. We note that the baseline and topline systems, both perform really well, but so do at least three of the peers (ID1, ID2, ID3). However, we notice that System ID3 has used words outside the word limits, thus being penalized in LAG grading.

An one-way analysis of variance (ANOVA) with pairwise comparison of means (Tukey's HSD test) shows that, as related to Overall Responsiveness, some systems were within statistical uncertainty equivalent to human summarizers (even though one human significantly outperforms all systems). The same stands for our topline summarizer (ID10). An overview of this analysis can be seen in Table 6.

5.7 Hebrew Language

For the Hebrew language, there were 5 participants (peers), plus the two baselines systems, for a total of 7 runs. An overview of the Overall Responsiveness and the corresponding Length Aware Grade (LAG) of these participants can be seen in Figures 7a, 7b. Generally considering the grades data, we note that the overall grading is quite unstable: there are systems having very wide distribution of grades (for example, ID3) having grades in the range from 1 to 5.

We note that the baseline and topline systems both perform similarly to two of the human peers (within statistical error). Equivalently high grades are assigned to several other automatic peer systems. System ID3 has used words outside the word limits, thus being penalized in LAG grading. Also, the baseline ID9 and topline ID10 were penalized in LAG for a one set (M001 and M002, respectively).

One-way analysis of variance (ANOVA) with pairwise comparison of means (Tukey's HSD test) shows that, as related to Overall Responsiveness, System ID3 performs as well (within statistical error) as the best human. An overview of this analysis can be seen in Table 7.

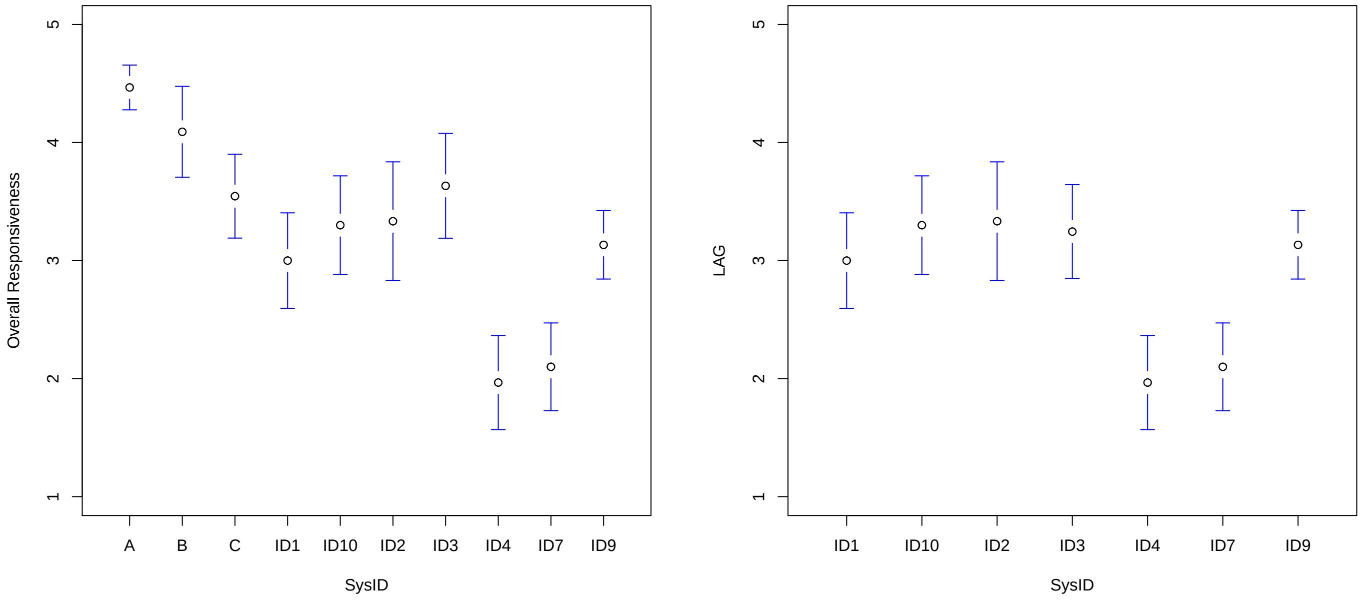
5.8 Hindi Language

For the Hindi language, there were 7 participants (peers), along with the two baselines systems, making a total of 9 runs. Overall Responsiveness of the systems is calculated by three human evaluators. These evaluations and the corresponding LAG is presented in Figures 8a, 8b respectively.

We noted that systems ID1, ID2, ID3 and ID5 performed very well compared to the baseline systems. We have also observed that ID4 submitted an empty file for M009. The groups formed by a Tukey's HSD test on the system performances are indicated in Table 8. The human summarizers form a group on their own, having significantly better performance than any of the other systems (whether baseline, topline or peer).

5.9 System Ranking Across All Languages

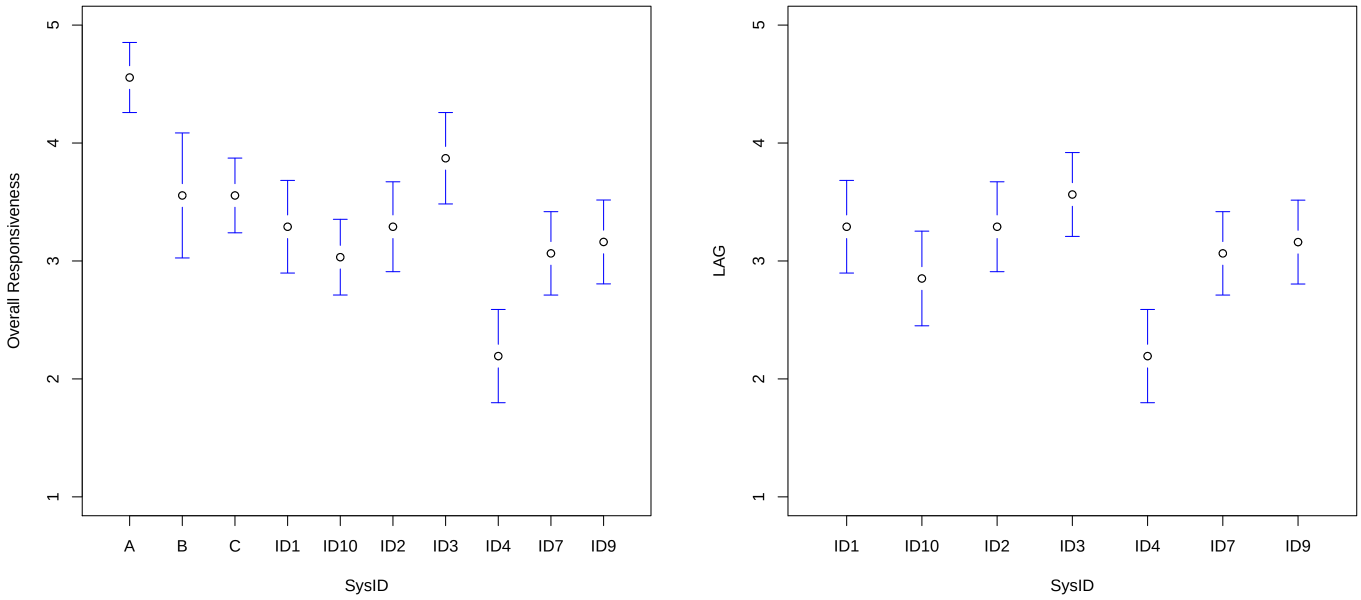
In this section we provide an overall system ranking across all languages, providing a bonus to methods that were tested on more languages. To this



(a) Overall Responsiveness — All Peers

(b) LAG for Overall Responsiveness — Systems Only

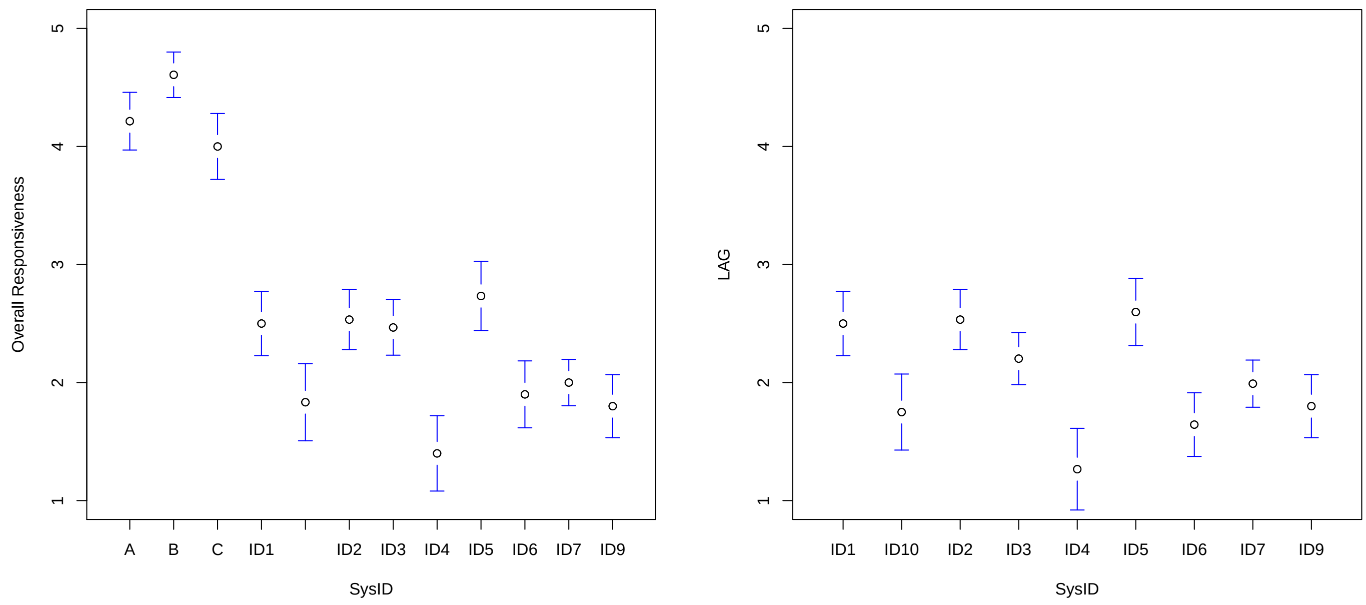
Figure 6: Greek: Performance overview



(a) Overall Responsiveness — All Peers

(b) LAG for Overall Responsiveness — Systems Only

Figure 7: Hebrew: Performance overview



(a) Overall Responsiveness — All Peers

(b) LAG for Overall Responsiveness — Systems Only

Figure 8: Hindi: Performance overview

Group	SysID	Avg Perf
a	B	4.61
a	A	4.21
a	C	4.00
b	ID5	2.73
bc	ID2	2.53
bcd	ID1	2.50
bcd	ID3	2.47
cde	ID7	2.00
de	ID6	1.90
e	ID10	1.83
e	ID9	1.80
e	ID4	1.40

Table 8: Hindi: Tukey's HSD test groups

end, we consider a system to have a grade of one (1) for languages it did not participate in. The *combined multi-lingual performance (CMP)* of a system is then calculated as the average LAG performance over all languages for a given system. Thus, if $g_s(l)$ is the LAG grade of system s in a given language l from the full set of languages L, we consider that the CMP of a system s is calculated as follows.

$$\mathrm{CMP}_s = \frac{\sum\limits_{l \in L} g_s(l)}{|L|} \qquad (1)$$

Table 9 illustrates the CMP grades of the participants. In addition to the CMP performance, we provide the standard error of the mean performance across languages as a measure of possible instability of a system across languages. For the calculation of the instability, we take into account only the languages for which a system provided summaries. Thus, if a system s participated in the set L_s of languages, $L_s \subset L$, and the standard deviation of its LAG grades in these languages is σ_s, then its insta-

System	CMP	Instability (StdErr)
ID1	2.99	0.19
ID2	2.95	0.18
ID3	3.13	0.18
ID4	1.86	0.21
ID5	1.60	0.48
ID6	1.60	0.34
ID7	2.41	0.19
ID8	1.63	0.78
ID9 (Baseline)	2.81	0.27
ID10 (Topline)	2.71	0.22

Table 9: Combined Multi-lingual Performance and Instability per System

bility is measured as follows.

$$\text{Instability}_s = \frac{\sigma_s}{\sqrt{|L_s|}} \tag{2}$$

Higher instability values indicate more uncertainty on whether the system is expected to perform near its mean performance in a new language.

It is important to note that only System ID3 performed on average above a grade of 3 (indicating average performance). Furthermore, the "overall baseline" proved to be better than the "overall topline" on average across all languages. However, this latter observation may be related to some minor bugs that existed in the topline code, causing zero-length summaries in some cases (assigned a grade of zero, when using LAG).

5.10 Automatic Evaluation

The question reposed 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 scores (ROUGE1, ROUGE2 and ROUGE-SU4), as well as the AutoSummENG method (MeMoG variation) to grade summaries. We used ROUGE2 because it has been robust and highly used for several years in the DUC and TAC communities. From AutoSummENG we preferred the MeMoG variation to the original version (which may have been more robust), because the original version correlated highly to ROUGE2 and it might not offer additional information. Given these two metrics, we tested whether there was a correlation between the automatically assigned grades and the Overall Responsiveness (OR) scores.

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 10.

There are some important lessons to be learnt from the (lack of significant) correlation results:

- There is almost no statistically significant correlation (except for English) between the Overall Responsiveness scores and the automatic scores. This may be due to the rather small set of the observations per language. We need more (and stronger) evidence to determine whether using one of these methods is good enough to replace human evaluation.

- The ROUGE2 and MeMoG are indeed overall not very strongly correlated, which may imply that they cover different aspects of performance.

- Over all languages both metrics perform almost identically (in terms of correlation to overall responsiveness): not well enough. However, there exist languages where performance differences are very obvious (e.g., Arabic or Greek).

- There exists a case where the correlation between MeMoG and Overall Responsiveness is negative (even though this correlation may be attributed to chance), which may prove to be an interesting negative result. In fact this is the case of the Arabic language, where 2 out of 3 human peers got very low grades (on average below 2.5), thus making these summaries *anti-models*. Therefore, high similarity to bad models (high MeMoG value) implies *low performance*. One question that may come from this observation is: "Can we perhaps improve summarization evaluation using also anti-models, in addition to the model summaries?".

6 Summary and Future Directions

Overall, the MultiLing pilot was successful, in that:

Language	ROUGE2 to OR	MeMoG to OR	ROUGE2 to MeMoG
Arabic	0.25	-0.36	0.11
Czech	0.33	-0.04	0.24
English	**0.56**	**0.47**	**0.47**
French	0.42	0.37	**0.50**
Greek	0.14	0.33	0.24
Hebrew	0.52	0.05	-0.24
Hindi	0.18	0.33	0.13
All languages	0.12	0.12	**0.42**

Table 10: Correlation (Kendall's Tau) Between Gradings. Note: statistically significant results, with p-value < 0.1, in **bold**.

- it bootstrapped multilingual summarization research as a community effort, by bringing together researchers from a variety of institutions and countries, aiming to tackle the same problem.

- it provided a method and an estimated cost for the creation of a multilingual summarization dataset.

- it provided such a benchmark dataset in 7 languages, using openly and freely available texts. The dataset is itself provided freely, upon request, and the specifics of making its dissemination and use even easier is ongoing.

- it provided two baseline systems, that can be easily reused in future efforts.

- it indicated that there exist systems that perform good-enough summarization in several languages.

The main lessons learnt from the pilot were the following:

- There is need of automation for all the processes of the creation of a dataset, to avoid communication problems and easily keep track of the process.

- The translation of texts across languages may be non-trivial and may induce some bias in itself. Given also the cost of translating one set of documents, it may be needed to lift the requirement of using the same texts across languages.

- Automatic multi-lingual summarization is feasible.

- There is significant space for improvement in the output summaries of most languages.

- There exist systems that function well on several languages.

- There exist languages, where all available automatic systems consistently perform much worse than in other languages. This poses the following question: "Is this a problem related to the idiosyncracy of the specific language, or was there a technical problem that caused such performance fall?".

For the future, we plan to continue the MultiLing effort. The main steps we plan to take are:

- to create subcorpora for more languages, especially taking into account languages with strong variation from the ones already in the corpus (e.g., Chinese).

- to gather more freely available texts per language, possibly by searching for texts that cover the same topic in different languages (instead of performing translations from a single source language).

- to find the funds required for the corpus creation process, in order to support the quality of the endeavour.

- to create a piece of support software that will help implement and track the (sub)corpus creation process.

- to examine the feasibility and usability of crowdsourcing.

- 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.

Overall, we hope and believe that the MultiLing effort has set the foundations for a flourishing community on multi-lingual summarization research. What remains to be done is build on these foundations, inviting and challenging more researchers to participate in this community. The fruit of this collaboration will be boosting the MultiLing effort to a global effort, providing a commonly accepted benchmark setting for current and future multi-lingual summarization systems.

Acknowledgements

We would like to thank Hoa Dang, Karolina Owczarzak, Vangelis Karkaletsis and Mark Last for their support and help.

We owe most of the things achieved to volunteers and employees acting as translators, summarizers or evaluators (or even all of the above).

We would deeply like to thank the participants for their patience and, especially, their participation which made the MultiLing pilot reality.

References

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

[Giannakopoulos and Karkaletsis, 2010] Giannakopoulos, G. and Karkaletsis, V. (2010). Summarization system evaluation variations based on n-gram graphs. In *TAC 2010 Workshop*, Maryland MD, USA.

[Giannakopoulos et al., 2008] Giannakopoulos, G., Karkaletsis, V., Vouros, G., and Stamatopoulos, P. (2008). Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.*, 5(3):1–39.

[Jones, 2007] Jones, K. S. (2007). Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449 – 1481. Text Summarization.

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

Group	SysID	Avg Perf
a	ID1	3.77
a	ID9	3.73
a	ID8	3.67
ab	ID7	3.30
ab	ID10	3.20
ab	ID2	3.10
ab	ID3	3.10
b	ID4	2.76
b	ID6	2.76

Table 11: Arabic: LAG-based Tukey's HSD test groups

A LAG HSD Tables

In this section of the Appendix we provide a set of tables illustrating statistically significantly different performances per language. We provide tables for all the languages: Arabic (Table 11), Czech (Table 12), French (Table 14), English (Table 13), Greek (Table 15), Hebrew (Table 16) and Hindi (Table 17).

Group	SysID	Avg Perf
a	ID9	3.30
ab	ID3	3.15
ab	ID1	3.00
b	ID2	2.70
bc	ID10	2.68
c	ID7	2.19
d	ID4	1.48

Table 12: Czech: LAG-based Tukey's HSD test groups

Group	SysID	Avg Perf
a	ID3	3.55
a	ID2	3.53
ab	ID10	3.20
abc	ID1	3.10
abcd	ID5	2.92
abcd	ID8	2.73
bcd	ID9	2.50
bcd	ID7	2.29
cd	ID6	2.20
d	ID4	2.03

Table 13: English: LAG-based Tukey's HSD test groups

Group	SysID	Avg Perf
a	ID3	2.93
b	ID1	2.28
bc	ID2	2.20
bcd	ID10	2.10
bcd	ID7	2.05
bcd	ID9	2.03
cde	ID5	1.69
de	ID6	1.62
e	ID4	1.33

Table 14: French: LAG-based Tukey's HSD test groups

Group	SysID	Avg Perf
a	ID2	3.33
a	ID10	3.30
a	ID3	3.25
a	ID9	3.13
a	ID1	3.00
b	ID7	2.10
b	ID4	1.97

Table 15: Greek: LAG-based Tukey's HSD test groups

Group	SysID	Avg Perf
a	ID3	3.56
a	ID1	3.29
a	ID2	3.29
a	ID9	3.16
a	ID7	3.06
ab	ID10	2.85
b	ID4	2.19

Table 16: Hebrew: LAG-based Tukey's HSD test groups

Group	SysID	Avg Perf
a	ID5	2.598333
ab	ID2	2.533333
ab	ID1	2.5
abc	ID3	2.203
bc	ID7	1.990667
cd	ID9	1.8
cd	ID10	1.750333
cd	ID6	1.644333
d	ID4	1.266

Table 17: Hindi: LAG-based Tukey's HSD test groups