Multi-document multilingual summarization corpus preparation, Part 2: Czech, Hebrew and Spanish

Michael Elhadad
Ben-Gurion Univ. in the Negev, Israel
elhadad@cs.bgu.ac.il

Sabino Miranda-Jiménez
Instituto Politécnico Nacional, Mexico
sabino_m@hotmail.com

Josef Steinberger
Univ. of West Bohemia, Czech Republic
jstein@kiv.zcu.cz

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

Abstract

This document overviews the strategy, effort and aftermath of the MultiLing 2013 multilingual summarization data collection. We describe how the Data Contributors of MultiLing collected and generated a multilingual multi-document summarization corpus on 10 different languages: Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian and Spanish. We discuss the rationale behind the main decisions of the collection, the methodology used to generate the multilingual corpus, as well as challenges and problems faced per language. This paper overviews the work on Czech, Hebrew and Spanish languages.

1 Introduction

In this document we present the language-specific problems and challenges faced by Contributors during the corpus creation process. To facilitate the reader we repeat some information found in the first part of the overview (Li et al., 2013): the MultiLing tasks and the main steps of the corpus creation process.

2 The MultiLing tasks

There are two main tasks (and a single-document multilingual summarization pilot described in a separate paper) in MultiLing 2013:

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.

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.

Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pages 13–19,
Sofia, Bulgaria, August 9 2013.

The first task was aiming 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 task, summarization evaluation has come to be a prominent research problem, based on the difficulty of the summary evaluation process. 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 (Dang and Owczarzak, 2009) aiming to test and evaluate different automatic evaluation methods of summarization systems.

Given the tasks, a corpus needed to be generated, that would be able to:

- provide input texts in different languages to summarization systems.
- provide model summaries in different languages as gold standard summaries, to also allow for automatic evaluation using model-dependent methods.
- provide human grades to automatic and human summaries in different languages, to support the testing of summary evaluation systems.

In the following section we show how these requirements were met in MultiLing 2013.

3 Corpus collection and generation

The overall process of creating the corpus of MultiLing 2013 was, similarly to MultiLing 2011, based on a community effort. The main processes consisting the generation of the corpus are as follows:

- Selection of a source corpus in a single language.
- Translation of the source corpus to different languages.
- Human summarization of corpus topics per language.
- Evaluation of human summaries, as well as of submitted system runs.

4 Language specific notes

In the following paragraphs we provide language-specific overviews related to the corpus contribution effort. The aim of these overviews is to provide a reusable pool of knowledge for future similar efforts.

In this document we elaborate on Czech, Hebrew, and Spanish languages. A second document (Elhadad et al., 2013) elaborates on the rest of the languages.

4.1 Czech language

The first part of the Czech subcorpus (10 topics) was created for the multilingual pilot task at TAC 2011. Five new topics were added for Multiling 2013. In total, 14 annotators participated in the Czech corpus creation.

The most time consuming part of the annotation work was the translation of the articles. The annotators were not professional translators and many topics required domain knowledge for correct translation. To be able to translate a person name, the translator needs to know its correct spelling in Czech, which is usually different from English. The gender also plays an important role in the translation, because a suffix 'ová' must be added to female surnames.

Translation of organisation names or person's functions within an organisation needs some domain knowledge as well. Complicated morphology and word order in Czech (more free but sometimes very different from English) makes the translation even more difficult.

For the creation of model summaries the annotator needed to analyse the topic well in order to decide what is important and what is redundant. Sometimes, it was very difficult, mainly in the case of topics which covered a long period (even 5 years) and which contained articles sharing very little information.

The main question of the evaluation part was how to evaluate a summary which contains a readable, continuous text — mainly the case of the

Group	SysID	Avg Perf
a	B	4.75
a	A	4.63
ab	C	4.61
b	D	4.21
b	E	4.10

Table 1: Czech: Tukey's HSD test groups for human summarizers

baseline system with ID6) — however not important information from the article cluster point of view.

An overview of the Overall Responsiveness and the corresponding average grades of the human summarizers can be seen in Table 1. We note that on average the human summaries are considered excellent (graded above 4 out of 5), but that there exist statistically significant differences between summarizers, essentially forming two distinct groups.

4.2 Hebrew language

This section describes the process of preparing the dataset for MultiLing 2013 in Hebrew: translation of source texts from English, and the summarization for the translated texts, by the Ben Gurion University Natural Language Processing team.

4.2.1 Translation Process

Four people participated in the translation and the summarization of the dataset of the 50 news articles: three graduate students, one a native English speaker with fluent Hebrew and the other two with Hebrew as a mother tongue and very good English skills. The process was supervised by a professional translator with a doctoral degree with experience in translation and scientific editing.

The average times to read an article was 2.5 minutes (std. dev 1.2min), the average translation time was 30 minutes (std. dev 15min), and the average proofing time was 18.5min (std. dev 10.5min).

4.2.2 Translation Methodology

We tested two translation methodologies by different translators. In some of the cases, translation was aided with Google Translate[1], while in other cases, translation was performed from scratch.

In the cases where texts were first translated using Google Translate, the translator reviewed the text and edited changes according to her judgment. Relying on the time that was reported for the proofreading of each translation, we could tell that texts that were translated using this method, required longer periods of proofreading (and sometimes more time was required to proofread than to translate). This is most likely because once the automatic translation was available, the human translator was biased by the automatic outcome, remaining anchored' to the given text with reduced criticism and creativity.

Translating the text manually, aided with online or offline dictionaries, Wikipedia and news site on the subject that was translated, showed better quality as analysis of time shows, where the ratio between the time needed to proofread was less than half.

In addition, we found, that in most cases the time that the translation took for the first texts of a given subject (for each article cluster), tends to be significantly longer than the subsequent articles in the same cluster. This reflects the 'learning phase' experienced by the translators who approached each cluster, getting to know the vocabulary of each subject.

4.2.3 Topic Clusters

The text collection includes five clusters of ten articles each. Some of the topics were very familiar to the Hebrew-speaking readers, and some subjects were less familiar or relevant. The Iranian Nuclear issue is very common in the local news and terminology is well known. Moreover, it was possible to track the articles from the news as they were published in Hebrew news websites at that time; this was important for the usage of actual and correct news-wise terminology. The hardest batch to translate was on the Paralympics championship, which had no publicity in Hebrew, and the terminology of winter sports is culturally foreign to native Hebrew speakers.

4.2.4 Special Issues in Hebrew

A couple of issues have surfaced during the translation and should be noted. Many words in Hebrew have a foreign transliterated usage and an original Hebrew word as well. For instance, the Latin word Atomic is very common in Hebrew and, therefore, it will be equally acceptable to use it in the Hebrew form, אטומי / 'atomi' but also the Hebrew word גרעיני ('gar' ini' / nuclear). Traditional Hebrew News Agencies have for many

[1] See http://translate.google.com/.

Summarizer	Reading time	Summarization
A	43 min	49 min
B	22 min	84 min
C	35 min	62 min

Table 2: Summarization process times (averaged)

Group	SysID	Avg Perf
a	A	4.80
ab	B	4.40
b	C	4.13

Table 3: Hebrew: Tukey's HSD test groups for human summarizers

years adopted an editorial line which strongly encourages using original Hebrew words whenever possible. In recent years, however, this approach is relaxed, and both registers are equally accepted. We have tried to use a 'common notion' in all texts using the way terms are written in Wikipedia as the voice of majority. In most cases, this meant using many transliterations.

Another issue in Hebrew concerns the orthography variations of plene vs. deficient spelling. Since Hebrew can be written with or without vocalization, words may be written with variations. For instance, the vocalized version of the word 'air' is אֲוִיר ('avir') while the non-vocalized version is אוויר ('avvir'). The rules of spelling related to these variations are complicated and are not common knowledge. Even educated people write words with high variability, and in many cases, usage is skewed by the rules embedded in the Microsoft Word editor. We did not make any specific effort to enforce standard spelling in the dataset.

4.2.5 Summarization Process

Each cluster of articles was summarized by three persons, and each summary was proof-read by the other summarizers. Most of the summarizers read the texts before summarization, while translating or proofreading them, and, therefore, the time that was required to read all texts was reduced.

The time spent reading and summarizing was extremely different for each of the three summarizers, reflecting widely different summarization strategies, as indicated in the Table 2 (average times over the 5 new clusters of MultiLing 2013):

The trend indicates that investing more time up front reading the clusters pays off later in summarization time.

The instructions did not explicitly recommend abstractive vs. extractive summarization. Two summarizers applied abstractive methods, one tended to use mostly extractive (C). The extractive method did not take markedly less time than the abstractive one. In the evaluation, the extractive summary was found markedly less fluent.

As the best technique to summarize efficiently, all summarizers found that ordering the texts by date of publication was the best way to conduct the summaries in the most fluent manner.

However, it was not completely a linear process, since it was often found that general information, which should be located at the beginning of the summary as background information, appeared in a later text. In such cases, summarizers changed their usual strategy and consciously moved information from a later text to the beginning of the summary. This was felt as a distinct deviation – as the dominant strategy was to keep track of the story told across the chronology of the cluster, and to only add new and important information to the summary that was collected so far.

The most difficult subject to summarize was the set on Paralympic winter sports championship which was a collection of anecdotal descriptions which were not necessarily a developing or a sequential story and had no natural coherence as a cluster.

4.2.6 Human evaluation

The results of human evaluation over the human summarizers are provided in Table 3. It is interesting to note that even between humans there exist two groups with statistically significant differences in their grades. On the other hand, the human grades are high enough to show high quality summaries (over 4 on a 5 point scale).

4.3 Spanish language

Thirty undergraduate students, from National Institute Polytechnic and Autonomous University of the State of Mexico, were involved in creating of Spanish corpus for MultiLing 2013.

The Spanish corpus built upon the Text Analysis Conference (TAC) MultiLing Corpus of 2011. The source documents were news from WikiNews website, in English language. The source corpus for translating consisted of 15 topics and 10 documents per topic. In the following paragraphs, we

show the measured times for each stage and problems that people had to face during the generation of corpus that includes translation of documents, multi-document summarization, and evaluation of human (manual) summaries.

At the translation step, people had to translate sentence by sentence or paraphrase a sentence up to completing the whole document. When a document was translated, it was sent to another person to verify the quality of the translated document. The effort was measured by three different time measurements: reading time, translation time, and verification time.

The reading average at document level was 7.6 minutes (with a standard deviation of 3.4 minutes), the average translation of each document was 19.2 minutes (with a standard deviation of 7.8 minutes), and the average verification was 14.9 minutes (with a standard deviation of 7.7 minutes). The translation stage took 104.5 man-hours.

At summarization step, people had to read the whole set of translated documents (topic) and create a summary per each set of documents. The length of a summary is between 240 and 250 words. Three summaries were created for each topic. Also, reading time of the topic and time of writing the summary were measured.

The average reading of a set of documents was 31.6 minutes (with a standard deviation of 10.2 minutes), and the average time to generate a summary was 27.7 minutes (with a standard deviation of 6.5 minutes). This stage took 44.5 man-hours.

At evaluation step, people had to read the whole set of translated documents and assess its corresponding summary. The summary quality was evaluated. Three evaluations were done for each summary. The human judges assessed the overall responsiveness of the summary based on covering all important aspects of the document set, fluent and readable language. The human summary quality average was 3.8 (on a scale 1 to 5) (with a standard deviation of 0.81). The results are detailed in Table 4. It is interesting to note that all humans have no statistically significant differences in their grades. On the other hand, the human grades are not excellent on average (i.e. exceeding 4 out of 5) which shows that the evaluators considered human summaries non-optimal.

Group	SysID	Avg Perf
a	C	3.867
a	B	3.778
a	A	3.667

Table 4: Spanish: Tukey's HSD test groups for human summarizers

4.3.1 Problems during Generation of Spanish Corpus

During the translation step, translators had to face problems related to proper names, acronyms, abbreviations, and specific themes. For instance, the proper name "United States" can be depicted with different Spanish words such as "EE. UU." [2], "Estados Unidos", and "EUA" — all of them are valid words. Even though translators know all the correct translations, they decided to use the frequent terms in a context of news (the first two terms are frequently used).

In relation to acronyms, well-known acronyms were translated into equivalent well-known (or frequent) Spanish translations such as UN (United Nations) became into ONU (Organización de las Naciones Unidas), or they were kept in the source language, because they are frequently used in Spanish, for example, UNICEF, BBC, AP (the news agency, Associated Press), etc.

On the contrary, for not well-known acronyms of agencies, monitoring centers, etc., translators looked for the common translation of the proper name on Spanish news websites in order to create the acronym based on the name. Other translators chose to translate the proper name, but they kept the acronym from the source document beside the translated name. In cases where acronyms appeared alone, they kept the acronym from source language. It is a serious problem because a set of translated documents has a mix of acronyms.

Abbreviations were mainly faced with ranks such as lieutenant (Lt.), Colonel (Col.), etc. Translators used an equivalent rank in Spanish. For instance, lieutenant (Lt.) is translated into "teniente (Tte.)"; however, translators preferred to use the complete word rather than the abbreviation.

In case of specific topics, translators used Spanish websites related to the topic in order to know the particular vocabulary and to decide what (tech-

[2] The double E and double U indicate that the letter represents a plural: e.g. EE. may stand for Asuntos Exteriores (Foreign Affairs).

nical) words should be translated and how they should be expressed.

As regards at text summarization step, summarizers dealt with how to organize the summary because there were ten documents per topic, and all documents involved dates. Two strategies were employed to solve the problem: generating the summary according to representative dates, or starting the summary based on a particular date.

In the first case, summarizers took the chain of events and wrote the summary considering the dates of events. They gathered important events and put together under one date, typically, the latest date according to a part of the chain of events. They grouped all events in several dates; thus, the summary is a sequence of dates that gather events. However, the dates are chosen arbitrary according to the summarizers.

In the second case, summarizers started the summary based on a specific date, and continued writing the sequence of important events. The sequence of events represents the temporality starting from a specific point of time (usually, the first date in the set of documents). Finally, in most cases, evaluators think that human summaries meet the requirements of covering all important aspects of the document set, fluent and readable language.

5 Conclusions and lessons learnt

The findings from the languages presented in this paper appear to second the claims found in the rest of the languages (Li et al., 2013):

- Translation is a non-trivial process, often requiring expert know-how to be performed.
- The distribution of time in summarization can significantly vary among human summarizers: it essentially sketches different strategies of summarization. It would be interesting to follow different strategies and record their effectiveness in the multilingual setting, similarly to previous works on human-style summarization (Endres-Niggemeyer, 2000; Endres-Niggemeyer and Wansorra, 2004). Our find may be related to the (implied) effort of taking notes while reading, which can be a difficult cognitive process (Piolat et al., 2005).
- The time aspect is important when generating a summary. The exact use of time (a simple timeline? a grouping of events based on time?) is apparently arbitrary.

We remind the reader that extended technical reports recapitulating discussions and findings from the MultiLing Workshop will be available after the workshop at the MultiLing Community website[3], as an addenum to the proceedings.

What can definitely be derived from all the effort and discussion related to the gathering of summarization corpora is that it is a research challenge in itself. If the future we plan to broaden the scope of the MultiLing effort, integrating all the findings in tools that will support the whole process and allow quantifying the apparent problems in the different stages of corpus creation. We have also been considering to generate comparable corpora (e.g., see (Saggion and Szasz, 2012)) for future MultiLing efforts. We examine this course of action to avoid the significant overhead by the translation process required for parallel corpus generation. We should note here that so far we have been using parallel corpora to:

- allow for secondary studies, related to the human summarization effort in different languages. Having a parallel corpus is such cases can prove critical, in that it provides a common working base.
- be able to study topic-related or domain-related summarization difficulty across languages.
- highlight language-specific problems (such as ambiguity in word meaning, named entity representation across languages).
- fixes the setting in which methods can show their cross-language applicability. Examining significantly varying results in different languages over a parallel corpus offers some background on how to improve existing methods and may highlight the need for language-specific resources.

On the other hand, the significant organizational and implementaion effort required for the translation may turn the balance towards comparable corpora for future MultiLing endeavours.

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

Acknowledgments

MultiLing is a community effort and this community is what keeps it alive and interesting. We would like to thank contributors for their organizational effort, which made MultiLing possible in so many languages and all volunteers, helpers and researchers that helped realize individual steps of the process. A more detailed reference of the contributor teams can be found in the Appendix.

The MultiLing 2013 organization has been partially supported by the NOMAD FP7 EU Project (cf. http://www.nomad-project.eu).

References

[Dang and Owczarzak2009] Hoa Trang Dang and K. Owczarzak. 2009. Overview of the tac 2009 summarization track, Nov.

[Elhadad et al.2013] 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.

[Endres-Niggemeyer and Wansorra2004] Brigitte Endres-Niggemeyer and Elisabeth Wansorra. 2004. Making cognitive summarization agents work in a real-world domain. In *Proceedings of NLUCS Workshop*, pages 86–96. Citeseer.

[Endres-Niggemeyer2000] Brigitte Endres-Niggemeyer. 2000. Human-style WWW summarization. Technical report.

[Giannakopoulos et al.2011] 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.

[Hovy et al.2005] E. Hovy, C. Y. Lin, L. Zhou, and J. Fukumoto. 2005. Basic elements.

[Li et al.2013] 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.

[Lin2004] 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.

[Louis and Nenkova2012] Annie Louis and Ani Nenkova. 2012. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300, Aug.

[Piolat et al.2005] Annie Piolat, Thierry Olive, and Ronald T Kellogg. 2005. Cognitive effort during note taking. *Applied Cognitive Psychology*, 19(3):291–312.

[Saggion and Szasz2012] Horacio Saggion and Sandra Szasz. 2012. The concisus corpus of event summaries. In *LREC*, pages 2031–2037.

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

Appendix: Contributor teams

Czech language team

Team members Brychcín Tomáš, Campr Michal, Fiala Dalibor, Habernal Ivan, Habernalová Anna, Ježek Karel, Konkol Michal, Konopík Miloslav, Krčmář Lubomír, Nejezchlebová Pavla, Pelechová Blanka, Ptáček Tomáš, Steinberger Josef, Zíma Martin.

Team affiliation University of West Bohemia, Czech Republic

Contact e-mail jstein@kiv.zcu.cz

Hebrew language team

Team members Tal Baumel, Raphael Cohen, Michael Elhadad, Sagit Fried, Avi Hayoun, Yael Netzer

Team affiliation Computer Science Dept. Ben-Gurion University in the Negev, Israel

Contact e-mail elhadad@cs.bgu.ac.il

Spanish language team

Team members Sabino Miranda-Jiménez, Grigori Sidorov, Alexander Gelbukh (Natural Language and Text Processing Laboratory, Center for Computing Research, National Institute Polytechnic, Mexico City, Mexico)

Obdulia Pichardo-Lagunas (Interdisciplinary Professional Unit on Engineering and Advanced Technologies (UPIITA), National Institute Polytechnic, Mexico City, Mexico)

Contact e-mail sabino_m@hotmail.com