Multi-document multilingual summarization corpus preparation, Part 1: Arabic, English, Greek, Chinese, Romanian

Lei Li
BUPT, China
leili@bupt.edu.cn

Corina Forascu
RACAI, Romania
UAIC, Romania
corinfor@info.uaic.ro

Mahmoud El-Haj
Lancaster Univ., UK
m.el-haj@lancaster.ac.uk

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 Arabic, Chinese, English, Greek, and Romanian languages. A second part, covering the remaining languages, is available as a distinct paper in the MultiLing 2013 proceedings.

1 Introduction

Summarization has recently received the focus of media attention (Cahan, 2013; Shih, 2013), due to a set of corporate buy-outs related to summarization technology companies. This trend of applying summarization is the result of a long research effort related to summarization. Previously, especially within the Text Analysis Conference (TAC) series of workshops (Dang, 2005; Dang, 2006; Dang and Owczarzak, 2008), multi-document summarization has covered aspects of summarization such as update summarization, guided summarization and cross-lingual summarization. In TAC 2011 the MultiLing Pilot (Giannakopoulos et al., 2011) was introduced: 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.

The Pilot gave birth to an active community of researchers, who provided the effort and know-how to realize a continuation of the original effort: MultiLing 2013. The MultiLing 2013 Workshop, taking place within ACL 2013, built upon the existing corpus of MultiLing 2011 to provide additional languages and challenges for summarization systems. This year 3 new languages were added: Chinese, Romanian and Spanish. Furthermore, more texts were added to most existing corpus languages (with the exception of French and Hindi).

In the following paragraphs we first overview the MultiLing tasks, for which the corpus was built (Section 2). We then describe the rationale and strategy applied for the corpus collection and creation (Section 3). We continue with special comments for the English, Greek, Chinese and Romanian languages (Section 4). Finally, we summarize the findings at the end of this paper (Section 5). We note that a second paper (Elhadad et al., 2013) describes the language-specific notes related to the rest of the MultiLing 2013 language contributions (Czech, Hebrew, Spanish).

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 lan-

Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pages 1–12,
Sofia, Bulgaria, August 9 2013. ©2013 Association for Computational Linguistics

guages 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.

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 of the generation of the corpus are as follows:

- Selection of a source corpus in a single language (see Section 3.1).
- Translation of the source corpus to different languages (see Section 3.2).
- Human summarization of corpus topics per language (see Section 3.3).
- Evaluation of human summaries, as well as of submitted system runs (see Section 3.4).

We should note here that the translation is meant to provide a parallel corpus of texts across different languages. The main ideas behind this first approach are that:

- the corpus will allow performing 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.
- we may be able to study topic-related or domain-related summarization difficulty across languages.
- the parallel corpus highlights language-specific problems (such as ambiguity in word meaning, named entity representation across languages).
- the parallel corpus 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 (please see per language notes in the corresponding sections) may lead to a comparable (vs. parallel) corpus in future MultiLing endeavours.

Given the tasks at hand, the Contributors first performed the selection of the texts that would be used for the MultiLing tracks, as described below.

3.1 Selecting the corpus

To support the summarization task, we needed a dataset of freely available news texts (to allow reuse), covering news topics that would contain event sequences. Based on the — apparently good — decisions of the MultiLing 2011 Pilot, we determined that each event sequence in the corpus should contain at least three distinct atomic events, to imply an underlying story.

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. It can be seen clearly that the event in the example has significantly different aspects, since an earthquake caused a radiation leak, via a series of interactions in the real world. Systems would normally be expected to express both aspects of the event with adequate information.

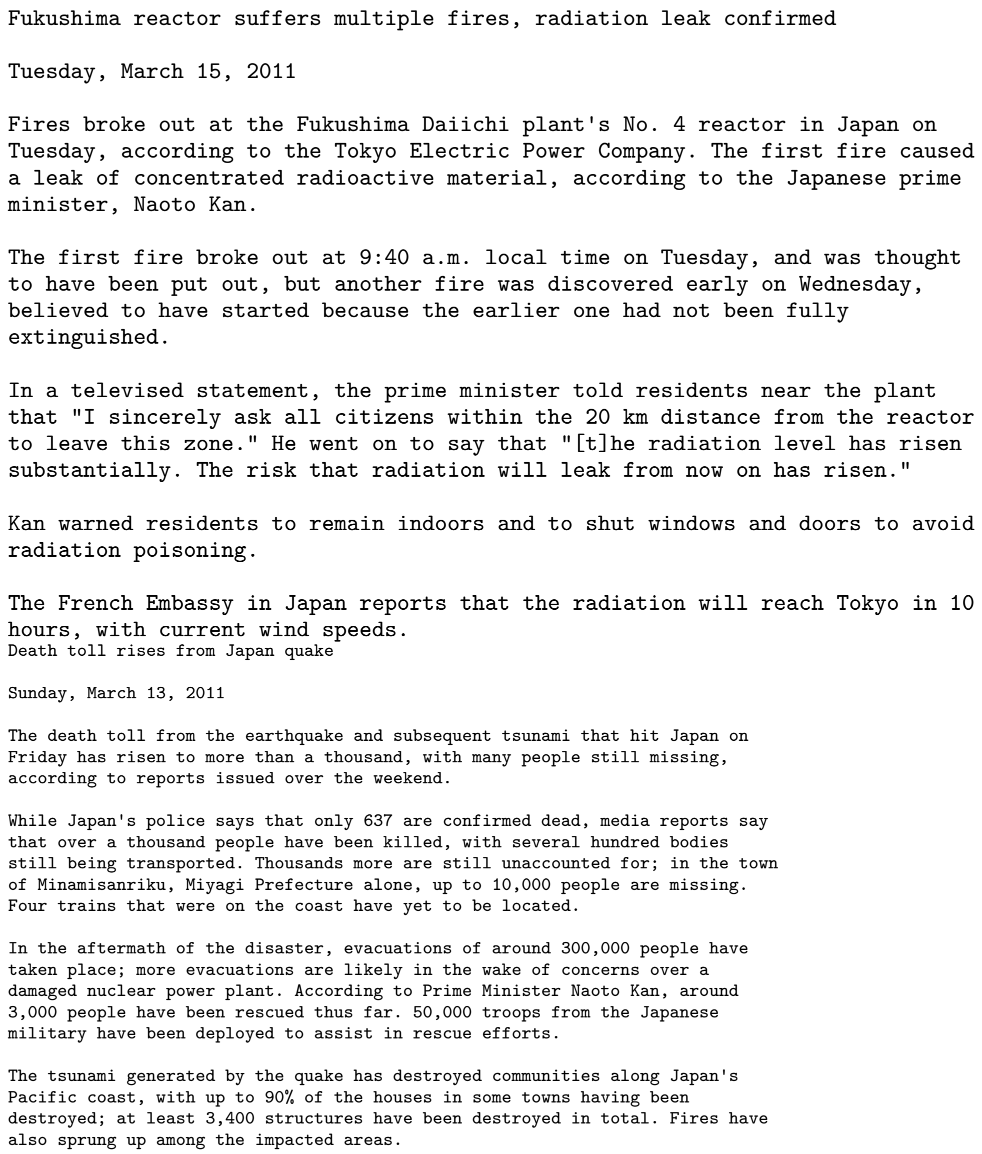
During the selection of the source texts, we first gathered an English corpus of 15 topics (10 of which were already available from MultiLing 2011), each containing 10 texts. We made sure that each topic contained at least one event sequence. From the original HTML text we only kept unformatted content text, without any images, tables or links.

While choosing topics we made sure that there existed topics:

- with varying time granularity. Some topics happen within days (e.g., sports events), while others within years (e.g., Iranian nuclear policy and international negotiations).
- covering various domains. There existed topics related to international politics, sports, natural disasters, political campaigns and elections.
- with a varying number of apparent actors. Some topics focus on specific individuals (e.g., campaign of Barack Obama) while others refer to numerous participants (e.g., para-Olympics and participating athletes).
- with numeric aspects, that would change over time. Such examples are natural disasters (with the number of estimated victims, or the estimated magnitude of earthquakes) and sports events (number of medals per country).
- with an important time dimension. For example during the Egyptian riots, the order of events is non-trivial to determine from text. Determining the order of events is also very challenging while following multi-day sports events. Ignoring the time dimension in such topics is expected to worsen the performance of summarization systems.

Given the English texts, we now needed to provide corresponding texts in all the languages used in MultiLing. To this end, we organized a translation process, which is elaborated below.

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



Fukushima reactor suffers multiple fires, radiation leak confirmed

Tuesday, March 15, 2011

Fires broke out at the Fukushima Daiichi plant's No. 4 reactor in Japan on Tuesday, according to the Tokyo Electric Power Company. The first fire caused a leak of concentrated radioactive material, according to the Japanese prime minister, Naoto Kan.

The first fire broke out at 9:40 a.m. local time on Tuesday, and was thought to have been put out, but another fire was discovered early on Wednesday, believed to have started because the earlier one had not been fully extinguished.

In a televised statement, the prime minister told residents near the plant that "I sincerely ask all citizens within the 20 km distance from the reactor to leave this zone." He went on to say that "[t]he radiation level has risen substantially. The risk that radiation will leak from now on has risen."

Kan warned residents to remain indoors and to shut windows and doors to avoid radiation poisoning.

The French Embassy in Japan reports that the radiation will reach Tokyo in 10 hours, with current wind speeds.

Death toll rises from Japan quake

Sunday, March 13, 2011

The death toll from the earthquake and subsequent tsunami that hit Japan on Friday has risen to more than a thousand, with many people still missing, according to reports issued over the weekend.

While Japan's police says that only 637 are confirmed dead, media reports say that over a thousand people have been killed, with several hundred bodies still being transported. Thousands more are still unaccounted for; in the town of Minamisanriku, Miyagi Prefecture alone, up to 10,000 people are missing. Four trains that were on the coast have yet to be located.

In the aftermath of the disaster, evacuations of around 300,000 people have taken place; more evacuations are likely in the wake of concerns over a damaged nuclear power plant. According to Prime Minister Naoto Kan, around 3,000 people have been rescued thus far. 50,000 troops from the Japanese military have been deployed to assist in rescue efforts.

The tsunami generated by the quake has destroyed communities along Japan's Pacific coast, with up to 90% of the houses in some towns having been destroyed; at least 3,400 structures have been destroyed in total. Fires have also sprung up among the impacted areas.

Figure 1: Topic Sample (Japan Earthquake and Nuclear Threat)

3.2 Translating the corpus

The English texts selected in the selection step were translated using a sentence-by-sentence approach to each of the other languages: Arabic, Chinese, Czech, French, Greek, Hebrew, Hindi, Romanian, Spanish. This year there was no support for the Hindi and French languages, which still contain 10 topics. Also the Chinese language covers 10 topics. All the remaining languages cover 15 topics.

During the translation process, the guidelines were minimal:

> Given the source language text A, the translator is requested to translate each *sentence* in A, into the target language. Each target sentence should keep the meaning from the source language.

Some additional, optional guidelines (provided in the Appendix) were provided by the Romanian language Contributors, proposing ways to react to date formatting, name translations, etc.

During the translation process, the translators were also asked to keep track of the time spent on different stages of the process: first full reading of the source document, translation and verification.

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 (vs. 700 from MultiLing 2011): 7 languages with 15 topics per language, 10 texts per topic for a total of 1050 texts; 3 languages with 10 topics per language, 10 texts per topic for a total of 300 texts.

This Source Document Set was provided to participating systems as input for their summarization systems. It was also provided to human summarizers, so that they would provide human, model summaries on each topic and each language. The human summarization process is described in the following section.

3.3 Summarizing topics

In the summarization step of the corpus creation different summarizers were asked to generate one summary per topic in each language. The following guidelines were provided to help the summarizers:

> The summarizer will read the whole set of texts at least once. Then, the summarizer should compose a summary, with *a minimum size of 240 and a maximum size of 250 words*. The summary should be in the same language as the texts in the set. The aim is to create a summary that covers all the major points of the document set (what is major is left to summarizer discretion). The summary should be written using fluent, easily readable language. No formatting or other markup should be included in the text. The output summary should be a self-sufficient, clearly written text, providing no other information than what is included in the source documents.

After summarization, human evaluation was performed. The evaluation covered human summaries, but also summarization system submissions. The details are provided in the following paragraphs.

3.4 Evaluating the summaries

The evaluation of summaries was performed 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 (Lin, 2004) and AutoSummENG-MeMoG (Giannakopoulos et al., 2008; Giannakopoulos and Karkaletsis, 2011) and NPowER (Giannakopoulos and Karkaletsis, 2013) 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.

As indicated in the task, the acceptable limits for the word count of a summary were between 240 and 250 words[2] (inclusive). In the case of Chinese there was a problem determining the number of words. Based on the model summaries gathered we (arbitrarily) set the upper limit of length in *bytes* of the UTF8-encoded summary files to 750 bytes.

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 Arabic, English, Greek, Chinese and Romanian languages. A second document (Elhadad et al., 2013) elaborates on the rest of the languages.

4.1 Arabic language

The preparation of the Arabic corpus for the 2013 MultiLing Summarization tasks was organised jointly by Lancaster University and the University of Essex in the United Kingdom. 20 people participated in translating the English corpus into Arabic, validating the translation and summarising the set of related Arabic articles. The participants are studying, or have finished a university degree in an Arabic speaking country. The participants' age ranged between 21 and 32 years old.

The participants translated the English dataset into Arabic. For each translated article another translator validated the translation and fixed any errors. For each of the translated articles, three manual summaries were created by three different participants (human peers). Amid the summarisation process the participants evaluated the quality of the generated summary by assigning a score between one (unreadable summary) and five (fluent and readable summary). No self evaluation was allowed.

The average time for reading the English news articles by the Arabic native speaker participants was 5.58 minutes. The average time it took them to translate these articles into Arabic was 42.18 minutes and to validate each of the translated Arabic articles the participants took 5.25 minutes on average.

For the summarisation task the average time for reading the set of related articles (10 articles per each set) was 34.44 minutes. The average time for summarising each set was 25.41 minutes.

4.1.1 Problems and Challenges

Many difficulties arose during the creation of the gold-standard summaries. Some are language-dependent and relate to the complexity of the Arabic language. This required a special attention to be paid while creating the summaries.

One problem concerns the handling of month names in Arabic. There are two ways of translating month names into Arabic:

- using the Arabic transliteration of the Aramic (Syriac) month names (e.g. “*May*”, “أَيَّار”, “Ayyar”).
- using the Arabic transliteration of the English month names (e.g. “*May*”, “مَايُو”, “Mayo”).

Some of the participants found it difficult to translate sentences where they believe they contain an ambiguous structure. For example: “She said Iranian security Chief Saeed Jalili had requested a meeting in a telephone call” . The translators (who are Native Arabic speakers) found it a bit hard to choose between two translations:

- “Saeed Jalili asked to schedule a telephone meeting”
- “Saeed Jalili phoned to request a meeting” .

Arabic sentence structure is highly complex and therefore great attention must be paid when moving forward or pushing back phrases within a sentence, as such shifts are likely to change the overall meaning. In addition, the use of passive voice, metaphors and idioms in the original English text has captured the translators attention, as the meaning in such cases takes precedence over the literal translation.

During the summarisation process, a summariser found that ordering a set of related articles (discussing the same topic) in chronological order simplifies the summarisation process.

Many participants found it difficult to meet the 250 summary word-limit as they believe 250 is not enough to cover all the essential information derived from a given set of documents.

Another problem concerns 'proper nouns' when translating into Arabic. The Arabic electronic discourse would sometimes show two variants of one

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

English proper noun, as in the case with the name 'Francois Hollande'. Mostly in such cases, the variant used in popular websites such as the Arabic version Wikipedia was adopted.

Finally, there were many questions by the participants on whether to create abstractive or extractive summaries.

4.2 Chinese language

Below we provide an overview of the organizational effort and comments on a variety of problems related to the preparation of the Chinese corpus for MultiLing 2013.

4.2.1 Organization

First, the Chinese language team translated two texts from English to Chinese together in order to make an original unified example for each translator, including file format, title format, date format, named entity translation, etc. Second, we assigned different set of news texts as specific task for each translator. For each news topic, we usually split the ten texts to two different translators at least, so as to bring more thoughts from different viewers and prepare enough for later discussion. During the process of each translator, they were asked to note any problems in a 'problem file', including the source English part and the target Chinese part. Third, we summed up a big problem file from each translator. After a series of discussions, we classified the problems into different categories and solved some of the problems successfully. The remaining problems were noted down in a detailed report to the organizer of the MultiLing 2013 Workshop of ACL 2013, as a knowledge pool for future efforts. Fourth, we performed the verification task. During the process, we made sure that for each text, the verifier was different from the translator. Also each verifier was demanded to log any problems. Fifth, we did another discussion for new problems coming from the verification phase. Some problems were solved; others were added to the detailed report. Sixth, we generated the needed result files and made sure that they were in the requested format (e.g., UTF8, no-BOM, plain text files for summaries).

For the process of summarization and human evaluation, first, we assigned three summarizers, each of which needed to read all the ten topics and write a summary for each topic. Second, we assigned three evaluators, making sure that for each summary, the evaluator was different from the summarizer. Third, we made a discussion about the process of summarization and evaluation. All agreed that summarization and evaluation were much easier than translation.

There were mainly two common problems. One was about the summary length. So we set a unified method for length checking. The other problem was more complex, which was that there were many different information in the original ten texts, but the result summary was limited to 250 words, so it was very difficult to choose the most important information. As a result, some information could be lost in final summaries. At the same time, we also found minor problems regarding the translation, improved the translation files and updated the detailed report about the problems we faced.

4.2.2 Problems and proposed solutions

In fact, related problems mainly came up from the task of translation. Most of them were common questions of the translators and language-dependent problems that needed special care. Here we only list the main categories of problems [3]. First, there were problems with the translation of person names. There are several sub-problems here:

- There are some person names which are not so popular, we could not find a result, so we finally keep the unknown English words among Chinese words.
- There is no specific separator between first name, middle name and family name in English, only normal space. But in Chinese, we usually add a separator “·” between them.
- There is also some ambiguity in person name to us, since we may be not quite familiar with some specific knowledge of news related domain.
- There are also some person names which seem to contain non-English characters. These names are more difficult for us, so we just keep most of them as the original format in English news.
- There are some person names with only one capitalized character and a dot in the middle

[3] A more detailed report has been submitted to the organizer of the Workshop.

> part. It’s really difficult for us to find a corresponding Chinese translation for it, so we just keep it as the original English format in the Chinese translations and keep the original English name in the following brackets.

Second, the translation for the English name of some websites, companies, organizations, etc, can cause problems. Since the full name may be too long for news reports, most of them also have occurred in corresponding simple format of abbreviation. Some of them are famous enough that we have a popular Chinese translation for them, while others are not so popular. So we decided that for unknown ones, we just reserve the English name, but for those known ones, we add the Chinese translation and keep some of the English abbreviation.

Third, the translation of time expressions is nontrivial. In English, the order usually used is: Weekday, Month Day, Year. But according to Chinese habit, we mention time usually in the following order: Year Month Day, Weekday.

Fourth, translation of locations names may not exist. There are many location names in these news texts. We tried to find their Chinese translation from many resources, but there are still some difficult ones left.

Fifth, there are some English words in the source texts which seem to be unrelated to other sentences in the news text (these may be text captions of photos in the source WikiNews articles). We just left them as they were.

Sixth, there are some sentences which are difficult to understand clearly because the context and structure are ambiguous. In these cases, we made a Chinese translation which seems best to us.

The above problems conclude the Chinese language contribution language-specific notes.

4.3 English and Greek languages

The effort related to the organization of the English and Greek languages was essentially equivalent to the MultiLing 2011 pilot (Giannakopoulos et al., 2011). This year 5 new topics were added to the two languages. The effort for English was reduced because no translation was needed. In the following subsections we elaborate on the organization details and the problems faced during the different subprocesses of the corpus creation.

4.3.1 Organization

A total of 7 people (being either MSc students, or researchers, all with fluency in English and Greek) were recruited for the two languages. An initial meeting was held to provide the basic guidelines and discuss questions on the translation process. Subsequently, e-mail communication and periodic conferences were used to assign the next tasks, related to summarization and evaluation.

For the purposes of meaningful assignment we created and used an automatic assignment script, that allows pre-allocating specific texts to workers (for any of the required tasks), while it automatically distributes work according to the availability of workers. The script avoids assigning workers to texts/tasks more than once.

In the evaluation process, we made sure (through pre-assignments) that no human would judge their own summary. It would have increased efficiency, if we had ascertained that human summarization would occur right after the translation of the texts.

The average time for reading the English news articles by the Greek native speaker participants was around 8 minutes. The average time it took them to translate these articles into Greek was around 48 minutes on average (with a couple of extreme cases exceeding 100 minutes, due to technical terminology, which was difficult to translate). The summarization time of the new topics in English was around 24 minutes per topic (plus an average of 8 minutes allocated to reading the source texts). For Greek the summary time was around 50 minutes per topic (we note that the summarizers’ groups for English and Greek were only minimally overlapping). In the Greek case, some deeper search showed that a single summarizer heavily biased the distribution of times to higher values.

To follow the progress of tasks, a generic project management tool was used. However, the tool proved insufficient in the micro-planning of the effort (individual assignments tracking). It would clearly make sense to use an ad-hoc designed system for planning and implementation of the effort.

4.3.2 Problems and proposed solutions

The main problems identified by contributors for Greek and English translation were related to well-known translation problems: named entity translation, date formatting, highly technical or domain specific terminology, ambiguous terms in the source text. Additional effort from translators provided solutions to these problems according to common practice in the translation domain.

The summarization effort indicated a few interesting points. Even though summarizers have their individual method for summarizing, some common practices and notes arise:

- A non-thorough glimpse of the source texts helps determine the overall topic.
- Time ordering is important in several cases, thus time ordering of the source texts is applied before the summarization process itself. The process is non-trivial even for humans.
- An initial summary which may be longer than the target size is created and several reductive transformations are applied. The 250 word limit proved critical and challenging, in that it forced summarizers to carefully choose information, essentially not covering the whole set of information from the source documents.
- Syntactic compression and rewriting is the last line of summarization, when it is obvious that more compression is needed.

As related to the evaluation process, we noted that there exists an inherent tendency for evaluators to determine whether a human or a machine performed the summarization. There were cases where evaluators altered their grading, because they inferred that not all texts were from humans or not all were from machines. We had noted this phenomenon also in MultiLing 2011. There are several cases where the evaluator also tries to determine the strategy of the system and, when one understands the underlying strategy, this may bias the grade. It would be interesting to evaluate this bias in the future.

Some additional notes are related to problems with the organization of the effort:

- A distributed work environment that would help track the progress of individuals and assignment of new tasks without significant communication effort, would have been very helpful.
- The assignment script was really critical in facilitating the organization of the effort and we plan to make it publicly available to allow reuse.

Overall, the collection and generation of the corpus was a very challenging effort, both in terms of organization and individual questions arising. However, next steps can build upon the lessons learnt, if the effort is well documented and the documents are freely and openly shared.

4.4 Romanian language

At MultiLing 2013, Romanian was addressed as a language for the first time. Following the Call for Contributors launched by the MultiLing organizers and based on the experience in the QA @ CLEF[4] evaluation campaign (Peñas et al., 2012), we started the data collection process working with a group of ten MSc students in Computational Linguistics from our Faculty, later adding another MSc student to the working group. Below we provide some notes on the translation and generation of human summaries processes:

- The translation, including verification, of the 150 WikiNews text documents from English into Romanian, was performed in a distributed context, theoretically based on an architecture like the one described in (Alboaie et al., 2003). Each student received one topic (10 documents) to be translated, based on a set of guidelines. We devised guidelines to tackle any language-dependent problems that need special care, and they were improved after each solution received from the students and based also on their questions. The full guidelines are provided in the Appendix of this document.

  We started with the following workflow: student A receives 10 English documents to be translated and summarized and sends the results to the organizer; another student, B, receives the English documents and the Romanian translations (made by student A) and s/he verifies the translations and prepares another summary. Finally, another student, C, receives from the organizer the 10 Romanian documents and s/he prepares the third summary of a given topic.

  Since the task proved to be very time-consuming for the students, all the last five topics (the ones introduced this year) were given to one student and then the translations were verified by the organizer.

[4] See http://celct.fbk.eu/ResPubliQA/index.php for more information.

- The generation of human summaries was performed immediately after the translation. For each topic, the aim was to create a summary that covers all the major points of the topic (what is major was left to summarizer' s discretion), being a self-sufficient, clearly written text, providing no other information than what is included in the source documents. The students were given no specific recommendations regarding the type of summary they should produce, e.g. an abstract versus an extract (Mani and Maybury, 1999), but they were specifically instructed to understand the main aspects of summarization.

5 Conclusions and lessons learnt

The corpus generated throughout the MultiLing corpus preparation provides a benchmark dataset for multilingual summarization. It tries to captured interesting, representative events, covering a variety of well-known news events around the world. The recent corporate interest in summarization, in conjunction with the ever-present increase of information flow from the Web and information redundancy, show that having a scientifically plausible set of evaluation tools for systems can help bring useful summarization systems to a wide audience. MultiLing functions as a focus point for multilingual summarization research and this document described the methods used to create a commonly accepted multilingual, multi-document summarization corpus.

Concerning thoughts on the future work of MultiLing, there are some points that have been raised by Contributors that we reproduce in the following sentences:

- In the translation phase, it would be useful to have translators for different languages discuss directly about some difficult cases, such as some ambiguous words, phrases and sentences, especially when they are expressed in some language-specific way.
- It would be very interesting to exploit the potential of comparable corpora, and not only of the parallel ones, especially if we consider the multilingual setting of MultiLing 2013. This means that the data should be collected starting from a given topic and each language contributor should find 10 documents on that given topic in his/her language.
- Creating a collaborative platform for building and improving summarization corpora could significantly facilitate the corpus building process for future efforts.

We remind the reader that a second paper (Elhadad et al., 2013) addresses the problems and challenges faced in the remaining languages actively contributed to in MultiLing 2013 (Czech, Hebrew and Spanish), thus completing the lessons learnt from the MultiLing 2013 contribution effort. Extended technical reports recapitulating discussions and findings from the MultiLing Workshop will be available after the workshop at the MultiLing Community website[5], as an addenum to the proceedings.

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 Appendix A.

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

References

[Alboaie et al.2003] Lenuta Alboaie, Sabin C Buraga, and Sınica Alboaie. 2003. tuBiG–a layered infrastructure to provide support for grid functionalities. *Omega*, 2:3.

[Cahan2013] Adam Cahan. 2013. Yahoo! To Acquire Summly http://yodel.yahoo.com/blogs/general/yahoo-acquire-summly-13171.html, March 25th.

[Dang and Owczarzak2008] 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.

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

[Dang2005] H. T. Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the*

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

Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005).

[Dang2006] H. T. Dang. 2006. Overview of DUC 2006. In *Proceedings of HLT-NAACL 2006*.

[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.

[Giannakopoulos and Karkaletsis2011] George Giannakopoulos and Vangelis Karkaletsis. 2011. Autosummeng and memog in evaluating guided summaries. In *TAC 2011 Workshop*, Maryland MD, USA, November.

[Giannakopoulos and Karkaletsis2013] George Giannakopoulos and Vangelis Karkaletsis. 2013. Summary evaluation: Together we stand npower-ed. In *Computational Linguistics and Intelligent Text Processing*, pages 436–450. Springer.

[Giannakopoulos et al.2008] 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.

[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.

[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.

[Mani and Maybury1999] Inderjeet Mani and Mark T Maybury. 1999. *Advances in automatic text summarization*. the MIT Press.

[Peñas et al.2012] Anselmo Peñas, Eduard H. Hovy, Pamela Forner, Álvaro Rodrigo, Richard F. E. Sutcliffe, Caroline Sporleder, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2012. Overview of qa4mre at clef 2012: Question answering for machine reading evaluation. In *CLEF (Online Working Notes/Labs/Workshop)*.

[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.

[Shih2013] Gerry Shih. 2013. Sound Familiar? After Yahoo Buys Summly, Google Buys News Summarization App Wavii http://www.huffingtonpost.com/2013/04/24/google-wavii_n_3143116.html, April 23rd.

[Tufis et al.2004] Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.

Appendix A: Contributor teams

Arabic language team

Team members Mahmoud El-Haj (Lancaster University, UK); Ans Alghamdi, Maha Althobaiti (Essex University, UK); Ahmad Alharthi (King Saud University, Saudi Arabia)

Contact e-mail m.el-haj@lancaster.ac.uk

Chinese language team

Team members Lei Li, Wei Heng, Jia Yu, Yu Liu, Qian Li

Team affiliation Center for Intelligence Science and Technology (CIST), School of Computer Science,Beijing University of Posts and Telecommunications,

Postal Address P.O.Box 310, Beijing University of Posts and Telecommunications, Xitucheng Road 10, Haidian District, Beijing, China

Contact e-mail leili@bupt.edu.cn

English and Greek languages team

Team members Zoe Angelou, Argyro Mavridakis, Valentini Mellas, Efrosini Zacharopoulou, George Kiomourtzis, George Petasis, George Giannakopoulos

Team affiliation NCSR “Demokritos”

Postal Address Institute of Informatics and Telecommunications, Patriarchou Grigoriou and Neapoleos Str., Aghia Paraskevi Attikis, Athens, Greece

Contact e-mail ggianna@iit.demokritos.gr

Romanian language team

Team members Corina Forascu, Raluca Moiseanu; Ana Maria Timofciuc, Alexandra Cristea, Alexandrina Sbiera, Bogdan Puiu, and Tudor Popoiu; other contributors to the task were Monica Ancuța, Romică Iarca, Claudiu Popa, and Cosmin Vlăduțu

Team affiliation UAIC, Romania

Contact e-mail corinfor@info.uaic.ro

Appendix B: Romanian guidelines

1. Translation equivalents belonging to the same part of speech should be used. The Romanian words should be as "closest" as possible to their English equivalents: If the English word has as equivalent a cognate in Romanian, this one should be used. The Romanian wordnet[6] (Tufis et al., 2004) should be used for problematic situations. If the English word doesn' t have a Romanian cognate, then the translator should not try to paraphrase it. Example: The English "sporadic" will be translated into 'sporadic', even though the translator would be tempted to use instead 'izolat' or 'rar' . It is not recommended to give translations such as 'mai puțin' or 'mai rar' .

2. English words should not be omitted and words which are not in the original English text should not be added because of stylistic reasons. Example: "The Telegraph" will be not translated when it refers to the newspaper and, moreover, the translators will not introduce an explanation, like 'cotidianul The Telegraph' [English: The Telegraph newspaper].

3. The Romanian diacritics have to be used, in UTF-8 encoding.

4. The translators must preserve as much as possible the tenses of the English verbs. Any disagreement from the English tense is allowed for linguistic reasons only (Romanian specific constructions), and not for stylistic ones.

5. The translators will preserve the format of dates, times, numbers. For example, for the issuing date of an article being "March 25, 2010" , the Romanian translation will be ' 25 martie 2010' and NOT 'Martie, 25, 2010' OR '25 Martie, 2010' .

6. The format of the numbers should follow the Romanian convention with respect to the decimal separator, which is comma (,), and not the period (.), like in English-speaking countries.

7. The unclear or unsure situations encountered by the translators will be separately recorded in a file, indicating the provenance of the document, the ID used for the problematic sentence and the commentaries/suggestions.

[6] See http://www.racai.ro/wnbrowser/.