MultiLing 2015: Multilingual Summarization of Single and Multi-Documents, On-line Fora, and Call-center Conversations

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

Jeff Kubina
U.S. Dep. of Defense
Ft. Meade, MD
jmkubin@tycho.ncsc.mil

John M. Conroy
IDA/Center for Comp. Sciences
Bowie, MD
conroy@super.org

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

Benoit Favre
University of Marseille
Marseille, France
benoit.favre
@lif.univ-mrs.fr

Mijail Kabadjov,
Udo Kruschwitz,
Massimo Poesio
University of Essex
Colchester, UK
{malexa,udo,poesio}
@essex.ac.uk

Abstract

In this paper we present an overview of MultiLing 2015, a special session at SIGdial 2015. MultiLing is a community-driven initiative that pushes the state-of-the-art in Automatic Summarization by providing data sets and fostering further research and development of summarization systems. There were in total 23 participants this year submitting their system outputs to one or more of the four tasks of MultiLing: MSS, MMS, OnForumS and CCCS. We provide a brief overview of each task and its participation and evaluation.

1 Introduction

Initially text-summarization research was fostered by the evaluation exercises, or tasks, at the Document Understanding and Text Analysis Conferences that started in 2001. But within the past five years a community of researchers have formed that push forward the development of text-summarization methods by creating evaluation tasks, dubbed MultiLing, that involve many languages (not just English) and/or many topical domains (not just news). The MultiLing 2011 and 2013 tasks evolved into a community-driven initiative that pushes the state-of-the-art in Automatic Summarization by providing data sets and fostering further research and development of summarization systems. The aim of MultiLing (Giannakopoulos et al., 2015) at SIGdial 2015 is the same: provide tasks for single and multi-document multilingual summarization and introduce pilot tasks to promote research in summarizing human dialog in online fora and customer call centers. This report provides an outline of the four tasks MultiLing supported at SIGdial; specifically the objective of each task, the data sets used by each task, and the level of participation and success by the research community within the task.

The remainder of the paper is organised as follows: section §2 briefly presents the Multilingual Single-document Summarization task, section §3 the Multilingual Multi-document summarization task, section §4 the Online Forum Summarization task, section §5 the Call-center Conversation summarization task, and finally we draw conclusions on the overall endeavour in section §6.

2 Multilingual Single Document Summarization Task

2.1 Task Description

The multilingual single-document summarization (MSS) task (Kubina and Conroy, 2015a) was created to foster the research and development of single document summarization methods that perform well on documents covering many languages and topics. Historically such tasks have predominantly focused on English news documents, see for example Nenkova (2005). The specific objective for this task was to generate a single document summary for each of the provided Wikipedia featured articles within at least one of the 38 languages provided. Wikipedia featured articles are selected by the consensus of their editors to be examples of some of the best written articles of a Wikipedia that fulfil all the required criteria with respect to accuracy, neutrality, completeness, and

Proceedings of the SIGDIAL 2015 Conference, pages 270–274,
Prague, Czech Republic, 2-4 September 2015.

style. Such articles make an excellent source of test data for single document summarization methods since they each have a well written summary (one of the style criterion), cover many languages, and have a diverse range of topics.

2.2 Participation, Evaluation, and Results

Participation in the 2015 MSS task was excellent, 23 summarization systems were submitted by seven teams. Four of the teams submitted summaries for all 38 languages and the remaining three submitted summaries covering four languages. English was the only language for which all participating systems submitted summaries.

For the evaluation a simple baseline summary was created from each article using the initial text of the article's body truncated to the size of the articles human summary. Its purpose, since it is so easy to compute, is to provide a summary score that participating systems should be able to exceed. An oracle summary was computed for each article using a covering algorithm (Davis et al., 2012) that selected sentences from the body text that covers the words in the summary using a minimal number of sentences until their aggregate size exceeds the summary. The oracle summary scores provide an approximate upper bound on the achievable summary scores and were, as expected, much higher than any submitted systems score.

The baseline, oracle, and submitted summaries were scored against the human summaries using ROUGE-2, -3, -4 (Lin, 2004) and MeMoG (Giannakopoulos et al., 2008). Details of the preprocessing applied to the text and the performance of each submitted system are in (Kubina and Conroy, 2015b), but overall 14 of the 23 systems did better than the baseline summary for at least half of the languages they partook in.

The ROUGE and MeMog scoring methods provide an automatic measure of summaries, which are good predictors of human judgements. A human evaluation of the summaries, that is currently underway, will measure the responsiveness and readability of each teams best performing system.

3 Multilingual Multi-Document Summarization Task

3.1 Task Description

This multilingual multi-document summarization (MMS) (Giannakopoulos, 2015) task aims to evaluate the application of partially or fully language-independent summarization algorithms. Each system participating in the task was called upon to provide summaries for a range of different languages, based on corresponding language-specific corpora. Systems were to summarize texts in at least two of the ten different languages: Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian, Spanish.

The task aims 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 previous MultiLing efforts (Giannakopoulos et al., 2011; Li et al., 2013) 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 occurrence 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 multi-document summarization task required participants to generate a fluent and representative summary from the set of documents describing an event sequence. The language of each document set belonged to one of the aforementioned set of languages and all the documents in a set were of the same language. The output summary was expected to be in the same language and between 240 and 250 words, with the exception of Chinese, where the output summary size was expected to be 333 characters (i.e., 1000 bytes in UTF-8 encoding).

The task corpus is based on a set of WikiNews English news articles comprising 15 topics, each containing ten documents. Each English document was translated into the other nine languages to create sentence-parallel translations. (Li et al., 2013; Elhadad et al., 2013).

3.2 Participation, Evaluation, and Results

Ten teams submitted 18 systems to the MMS task. Three randomly chosen topics (namely topics M001, M002, M003) out of the 15 topics, were provided as training sets to the participants for the task and were excluded when ranking of the systems.

The ranking was based on automatic evaluations methods using human model summaries provided by fluent speakers of each corresponding language (native speakers in the general case).

ROUGE variations (ROUGE-1, ROUGE-2) (Lin, 2004) and the AutoSummENG-MeMoG (Giannakopoulos et al., 2008) and NPowER (Giannakopoulos and Karkaletsis, 2013) methods were applied to automatically evaluate the summarization systems. There was a clear indication that ROUGE measures were extremely sensitive to different preprocessing types and that different implementations (taking into account multilinguality or not during tokenization) may offer significantly different results (even different order of magnitude in the score). Thus, the evaluation was based on the language-independent MeMoG method.

On average 12 system runs were executed per language, with the least popular language being Chinese, and the most popular being English. On average across all languages, except for Chinese, 13 of the 18 systems surpassed the baseline, according to the automatic evaluation. The systems employed a variety of approaches to tackle the multi-document summarization challenge as described in the following paragraphs.

The approaches contained various types of preprocessing, from POS tagging and extraction of POS patterns, to the representation of documents to language-independent latent spaces before the summarization or reduced vector spaces (e.g. through PCA (Jolliffe, 2002)). It is also interesting to note that more than 10 different tools were used in various preprocessing steps, such as stemming, tokenization, sentence splitting, due to the language dependence limitations of many such tools. Overall, in comparison to the previous MultiLing MMS challenge, this time it appears that reuse of existing tools for such preprocessing was increased (as detailed in individual system reports).

Subtopics were identified in some cases through various methods, such as the use of bag-of-word vector space representation of sentences and cosine-similarity-based clustering, or probabilistic clustering methods (e.g. hLDA (Blei et al., 2004)).

For the sentence scoring, cosine similarity was also used as a means for sentence selection, where the topic(s) of a document group was projected in a vector space (either bag-of-words or latent topic space). Some of the MMS participants' systems used supervised optimization methods (e.g. polytope model optimization, genetic algorithms) on rich feature spaces to either maximize coverage of the output summaries, or train models for sentence scoring. The feature spaces went beyond words to linguistic features, position features, etc. Other systems used graph methods, relying on the "importance" of sentences as indicated by methods such as PageRank (Page et al., 1999).

Finally, redundancy was tackled through cosine similarity between sentences, or in the sentence selection process itself as penalty to optimization cost functions.

Overall, once again the multi-document, multilingual task showed that multilinguality implies a need for many linguistic resources, but is significantly helped by the application of machine learning methods. It appears that these latter approaches transfer the burden to the annotation of good training corpora.

4 OnForumS Task

4.1 Task description

The Online Forum Summarization (OnForumS) pilot task (Kabadjov and Steinberger, 2015) investigated how the mass of comments found on news providers web sites (e.g., The Guardian) can be summarized. We posited that a crucial initial step towards that goal is to determine what comments link to either specific news snippets or comments of other users. Furthermore, a set of labels for a given link is articulated to capture phenomena such as agreement and sentiment with respect to the comment target. Solving this labelled-linking problem can enable recognition of salience (e.g., snippets/comments with most links) and relations between comments (e.g., agreement). For instance, comment sentences linked to the same article sentence can be seen as forming a "cluster" of sentences on a specific point/topic. Moreover, having labels capturing argument structure and sentiment enables computing statistics within such topic clusters on how many readers are in favour or against the point raised by the article sentence and what is the general 'feeling' about it.

The task included data in two languages, English and Italian, provided by the FP7 SENSEI project.[1]

4.2 Participation, Evaluation and Results

Four research groups participated in the OnForumS, each submitting two runs. In addition, two baseline system runs were included making a total of ten different system runs.

[1] http://www.sensei-conversation.eu/

Submissions were evaluated via crowdsourcing on Crowd Flowerwhich is a commonly used method for evaluating HLT systems (Snow et al., 2008; Callison-Burch, 2009). The crowdsourcing HIT was designed as a validation task (as opposed to annotation), where each system proposed link and labels are presented to a contributor for their validation.

The approach used for the OnForumS evaluation is IR-inspired and based on the concept of *pooling* used in TREC (Soboroff, 2010), where the assumption is that possible links that were not proposed by any system are deemed irrelevant. Then from those links proposed by systems, four categories are formed as follows:

(a) links proposed in 4 or more system runs
(b) links proposed in 3 system runs
(c) links proposed in 2 system runs
(d) links proposed only once

Due to the volume of links proposed by systems, a stratified sample was extracted for evaluation based on the following strategy: all of the **a** and **b** links[2] and a third of each **c** and **d** links selected at random.

Once the crowdsourcing exercise was completed, correct and incorrect links were counted.[3] From those links validated as correct, the correct and incorrect argument and sentiment labels were counted. Using these counts precision scores were computed. System runs were then ranked based on these precision scores. For the linking task no system surpassed the baseline algorithm based on overlap and scores were substantially higher for English than for Italian.

A recall-based evaluation was also carried out on a smaller gold standard set created from the validated data by taking all 'yes' validations of links as gold links and then all labels for argument and sentiment with 'yes' validations as the gold labels for those links.

5 CCCS Task

5.1 Task description

The call-center conversation summarization pilot task consists in automatically generating abstractive summaries of spoken conversations between a customer and an agent solving a problem over the phone. This task is different from news summarization in that dialogues need to be analysed in a deeper manner in order to recover the problem being addressed and how it is solved, and convert spontaneous utterances to reported speech. Generating such summaries, called conversation synopses, in this framework, is challenging for extractive approaches, and therefore should make participants focus on abstractive summarization. The task leverages a corpus of French and Italian conversations as well as English translations of those dialogues. The data is provided by the FP7 SENSEI project. For more details on the CCCS task see (Favre et al., 2015).

5.2 Participation, evaluation and results

Four systems have been submitted to this first edition of the CCCS task, by two research groups. In addition, three extractive baselines were evaluated for comparison purposes. The official metric was ROUGE-2. Evaluation on each of the languages shows that the submitted systems had difficulties beating the extractive baselines, and that human annotators are consistent in their synopsis production (for more details see (Favre et al., 2015)). We will focus on extending the evaluation in order to overcome the limitations of ROUGE, and assess the abstractiveness of the generated synopses.

6 Conclusion

MultiLing has been running for a few years now and has proved a successful evaluation campaign for automatic summarization. MultiLing 2015 is the third chapter of the campaign and participation was excellent with 23 participants submitting two or more system runs across the four tasks that the campaign comprises.

The next steps for the classical tasks MSS and MMS is to continue expanding the corpora in size and across languages, whereas for the pilot tasks is to further precise the boundaries of the new tasks and bridge the gaps in the evaluation methodologies by overcoming the limitations of ROUGE in order to assess abstractiveness and minimizing the effect of 'cheating' workers in crowdsourcing (e.g., by incorporating a probabilistic model of annotation, such as the one put forward by (Passonneau and Carpenter, 2013) to filter better noisy crowdsourcing data).

The next MultiLing is planned for 2017.

[2]The popular links (**a** and **b**) were not that many, hence, we chose to include all.

[3]Based on CrowdFlower's aggregated judgements.

Acknowledgements

The research leading to these results has received funding from the European Union - 7th Framework Programme (FP7/2007-2013) under grant agreement 610916 SENSEI. The research leading to these results has received funding from the European Regional Development Fund of the European Union and from national funds in the context of the research project 'SentIMAGi - Brand monitoring and reputation management via multimodal sentiment analysis' (ISR_2935) under the Regional Operational Programme Attica (Priority Axis 3 Improving competitiveness, innovation and digital convergence) of the 'Bilateral R&D Cooperation between Greece and Israel 2013-2015' of the Action of national scope 'Bilateral , Multilateral and Regional R&D Cooperation'.

References

D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. 2004. Hierarchical topic models and the nested chinese restaurant process. *NIPS*, 16:17.

C. Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazons mechanical turk. In *Proceedings of EMNLP*, volume 1, pages 286–295.

S. T. Davis, J. M. Conroy, and J. D. Schlesinger. 2012. Occams - an optimal combinatorial covering algorithm for multi-document summarization. In *ICDM Workshops*, pages 454–463. IEEE Computer Society.

M. Elhadad, S. Miranda-Jiménez, J. Steinberger, and G. Giannakopoulos. 2013. Multi-document multilingual summarization corpus preparation, Part 2: Czech, Hebrew and Spanish. In *MultiLing 2013 Workshop in ACL 2013*, Sofia.

B. Favre, E. Stepanov, J. Trione, F. Béchet, and G. Riccardi. 2015. Call Centre Conversation Summarization: A Pilot Task at Multiling 2015. In *SIGDIAL*.

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

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

G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, and V. Varma. 2011. TAC2011 MultiLing Pilot Overview. In *TAC 2011 Workshop*.

G. Giannakopoulos, J. Kubina, J. Conroy, J. Steinberger, B. Favre, M. Kabadjov, U. Kruschwitz, and M. Poesio. 2015. MultiLing 2015. `http://multiling.iit.demokritos.gr/pages/view/1516/multiling-2015`. [Online; accessed 19-July-2015].

G. Giannakopoulos. 2015. MMS MultiLing 2015 Task. `http://multiling.iit.demokritos.gr/pages/view/1540/task-mms-multi-document-summarization-data-and-information`. [Online; accessed 19-July-2015].

I. Jolliffe. 2002. *Principal component analysis*. Wiley Online Library.

M. A. Kabadjov and J. Steinberger. 2015. OnForumS MultiLing 2015 Task. `http://multiling.iit.demokritos.gr/pages/view/1531/task-onforums-data-and-information`. [Online; accessed 19-July-2015].

J. Kubina and J. Conroy. 2015a. MSS MultiLing 2015 Task. `http://multiling.iit.demokritos.gr/pages/view/1532/task-mss-single-document-summarization-data-and-information`. [Online; accessed 19-July-2015].

J. Kubina and J. Conroy. 2015b. SIGDIAL 2015 Multilingual Single-Document Summarization Task Overview. In *MultiLing 2015 Addendum*.

L. Li, C. Forascu, M. El-Haj, and G. Giannakopoulos. 2013. Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian. In *MultiLing 2013 Workshop in ACL 2013*, Sofia.

C.-Y. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

A. Nenkova. 2005. Automatic text summarization of newswire: lessons learned from the document understanding conference. In *Proceedings of AAAI*, pages 1436–1441. AAAI Press.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford InfoLab.

R. J. Passonneau and B. Carpenter. 2013. The Benefits of a Model of Annotation. In *Proceedings of the 7th LAW at ACL*, pages 187–195, Sofia.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. 2008. Cheap and fast – but is it good?: Evaluating nonexpert annotations for natural language tasks. In *Proceedings of EMNLP*, pages 254–263.

I. Soboroff. 2010. Test Collection Diagnosis and Treatment. In *Proceedings of the Third International Workshop on Evaluating Information Access (EVIA)*, pages 34–41, Tokyo.