Call Centre Conversation Summarization: A Pilot Task at Multiling 2015

Benoit Favre[1], Evgeny Stepanov[2], Jérémy Trione[1], Frédéric Béchet[1], Giuseppe Riccardi[2]
[1] Aix-Marseille University, CNRS, LIF UMR 7279, Marseille, France
[2] University of Trento, Via Sommarive 5, Trento, Italy
benoit.favre@lif.univ-mrs.fr

Abstract

This paper describes the results of the Call Centre Conversation Summarization task at Multiling'15. The CCCS task consists in generating abstractive synopses from call centre conversations between a caller and an agent. Synopses are summaries of the problem of the caller, and how it is solved by the agent. Generating them is a very challenging task given that deep analysis of the dialogs and text generation are necessary. Three languages were addressed: French, Italian and English translations of conversations from those two languages. The official evaluation metric was ROUGE-2. Two participants submitted a total of four systems which had trouble beating the extractive baselines. The datasets released for the task will allow more research on abstractive dialog summarization.

1 Introduction

Speech summarization has been of great interest to the community because speech is the principal modality of human communications, and it is not as easy to skim, search or browse speech transcripts as it is for textual messages. Speech recorded from call centres offers a great opportunity to study goal-oriented and focused conversations between an agent and a caller. The Call Centre Conversation Summarization (CCCS) task consists in automatically generating summaries of spoken conversations in the form of textual synopses that shall inform on the content of a conversation and might be used for browsing a large database of recordings. Compared to news summarization where extractive approaches have been very successful, the CCCS task's objective is to foster work on abstractive summarization in order to depict what happened in a conversation instead of what people actually said.

The track leverages conversations from the Decoda and Luna corpora of French and Italian call centre recordings, both with transcripts available in their original language as well as English translation (both manual and automatic). Recordings duration range from a few minutes to 15 minutes, involving two or sometimes more speakers. In the public transportation and help desk domains, the dialogs offer a rich range of situations (with emotions such as anger or frustration) while staying in a coherent and focused domain.

Given transcripts, participants to the task shall generate abstractive summaries informing a reader about the main events of the conversations, such as the objective of the caller, whether and how it was solved by the agent, and the attitude of both parties. Evaluation has been performed by comparing submissions to reference synopses written by quality assurance experts from call centres. Both conversations and reference summaries are kindly provided by the SENSEI project.

This paper reports on the results of the CCCS task in term ROUGE-2 evaluation metric. Two participants have submitted four systems to the task. In addition, we provide three baselines which frame the performance that would be obtained by extractive systems. The results are analysed according to language, human annotator coherence and the impact of automatic translation.

The remaining of the paper is organized as follows: Section 2 describes the synopsis generation task. Section 3 describes the CCCS corpus. Section 4 describes the results from the systems of the participants. Section 5 discusses future research avenues.

2 Task

The CCCS task consists in creating systems that can analyse call centre conversations and generate written summaries reflecting why the customer is calling, how the agent answers that query, what are the steps to solve the problem and what is the resolution status of the problem.

Unlike news summarization which focuses on locating facts in text written by journalists and selecting the most relevant facts, conversation synopses require an extra level of analysis in order to achieve abstraction. Turn taking from the speakers has to be converted to generic expression of their needs, beliefs and actions. Even though extractive systems might give a glimpse of the dialogs, only abstraction can yield synopses that tell the story of what happens in the conversations.

Contrary to previous research on meeting summarization (Gillick et al., 2009; Erol et al., 2003; Lai and Renals, 2014; Wang and Cardie, 2012) (among others), we expect that the fact that conversations are focused and goal oriented will enable to foster research on more abstractive methods, such as (Murray, 2015; Mehdad et al., 2013) and deeper analysis of the conversations.

Participants to the CCCS task could submit system output in any of the supported languages, and could submit a maximum of three runs per language. For each conversation, they had to submit one synopsis of length 7% of the number of words of the transcript of that conversation.

3 Corpus description

The CCCS task draws from two call centre conversation corpora, the Decoda corpus in French and the Luna corpus in Italian. Subsets from both corpora have been translated to English.

Decoda corpus The French DECODA corpus consists in conversations between customers and one or more agent recorded in 2009 in a call centre of the public transport authority in Paris (Bechet et al., 2012). The topics of the conversations range from itinerary and schedule requests, to lost and found, to complaints (the calls were recorded during strikes). The dialogues, recorded in ecological conditions, are very spontaneous and focused on the objective of the caller. They are very challenging for Automatic Speech Recognition due to harsh acoustic conditions such as calling from mobile phones directly from the metro. For the CCCS task, manual transcripts were provided to the participants.

While the original language of the conversations is French, the SENSEI project provided manual translations in English by professional translators which were trained to keep the spontaneous aspects of the originals (a very challenging task according to them). 97 conversations were manually translated, on which an automatic translation system based on Moses was trained in order to produce automatic translations for the remaining of the corpus.

The original corpus consists of 1513 conversations (about 70h of speech). 1000 conversations have been distributed without synopses for unsupervised system training. 50 conversations were distributed with multiple synopses from up to five annotators. The test set consists of 47 manually translated conversations and corresponding synopses, and 53 automatically translated conversations and corresponding synopses. The data for training and testing is also provided in French.

Statistic	**FR**	**EN**
Conversations	100	100
Turns	7,905	7,909
Words	42,130	41,639
Average length	421.3	416.4
Lexicon size	2,995	2,940
Number of synopses	212	227
Average synopsis length	23.0	26.5

Table 1: Decoda test set statistics.

The human written synopses are very diverse and show a high degree of abstraction from the words of the conversation with third person writing, telegraphic style and analysis of the conversations. Examples:

- *A man is calling cause he got a fine. He is waiting for a new card so he used his wife's card. He must now write a letter asking for clemency.*
- *A user wants to go to the Ambroise Paré clinic but the employee misunderstands and gives her the wrong itinerary. Luckily the employee realises her mistake and gives the passenger the right information in the end.*
- *School bag lost on line 4, not found.*

Luna corpus The Italian human-human Luna corpus (Dinarelli et al., 2009) consists of 572 dialogs (≈ 26.5K turns & 30 hours of speech) in the hardware/software help desk domain, where a client and an agent are engaged in a problem solving task over the phone. The dialogs are organised in transcriptions and annotations created within the FP6 LUNA project. For the CCCS shared task, manual transcriptions were used.

Within the FP7 SENSEI project, 100 dialogs were translated from Italian to English using professional translation services according to the methodology described in (Stepanov et al., 2014). For more accurate translations, manual transcriptions were converted to an 'annotated' text format, which contained mark-up for overlapping turns, fillers, pauses, noise, partial words, etc.; and translators received detailed guidelines on how to handle each phenomenon in translation. Additionally, the translators were required to translate the speech phenomena such as disfluencies as closely as possible to the source language maintaining 'naturalness' in the target language.

Five native Italian speakers have annotated 200 Luna dialogs with synopses so that each dialog was processed by every annotator.[1] Synopses of the 100 translated dialogs were also manually translated to English.

The translated and annotated dialogs were equally split into training and test sets for the CCCS task. The training dialogs were used to automatically translate additional Luna dialogs and synopses for both training and testing. Similar to the DECODA corpus, for the unsupervised training of the systems a supplementary set of 261 dialogs was automatically translated and provided to the participants without synopses. Dialogs and their associated synopses were provided both in English and Italian. The statistics for Luna manual English test set are provided in Table 2.

Statistic	**IT**	**EN**
Conversations	100	100
Turns	4,723	4,721
Words	34,913	32,502
Average length	349.1	325.0
Lexicon size	3,393	2,451
Number of synopses	500	500
Average synopsis length	17.4	15.4

Table 2: Luna test set statistics.

4 Results

Metric Evaluation is performed with the ROUGE-2 metric (Lin, 2004). ROUGE-2 is the recall in term of word bigrams between a set of reference synopses and a system submission. The ROUGE 1.5.5 toolkit was adapted to deal with a conversation-dependent length limit of 7%, had lemmatization disabled and stop-words kept, to be as language independent as possible [2]. Jackknifing and resampling is used in order to compute confidence estimate intervals.

Participation Seven research groups had originally expressed their intention to participate to the CCCS task. Four groups downloaded the test data, and two groups actually submitted system output at the deadline. Those two groups generated four runs: `NTNU:1`, `NTNU:2`, `NTNU:3`, `LIA-RAG:1`. The technical details of these submissions are described in their own papers.

In addition to those four runs, we provide three baselines which serve to calibrate participant performance. The first baseline is Maximal Marginal Relevance (`Baseline-MMR`) (Carbonell and Goldstein, 1998) with $\lambda = 0.7$. The second baseline is the first words of the longest turn in the conversation, up to the length limit (`Baseline-L`). The third baseline is the words of the longest turn in the first 25% of the conversation, which usually corresponds to the description of the caller's problem (`Baseline-LB`). Those baselines are described in more details in (Trione, 2014).

In order to estimate the overlap between human synopses, we remove each of the human synopses in turn from the reference and compute their performance as if they were systems. Across languages, 11 annotators (denoted `human-1` to `human-5` for IT/EN, and `human-A` to `human-G` for FR/EN) produced from 5 to 100 synopses. Note that some annotators only worked on English conversations.

Performance Performance of the systems is reported in Table 3. It shows that in the source languages, the extractive baselines were difficult to beat while one of the systems significantly outperformed the baselines on English (the EN test set

[1] Few (2) synopses were found to address dialog dimensions other than the task and were removed.

[2] The options for running ROUGE 1.5.5 are `-a -l 10000 -n 4 -x -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0`

corresponds to the union of manual and automatic translations).

System	EN	FR	IT
NTNU:1	0.023	0.035	0.013
NTNU:2	0.031	0.027	0.015
NTNU:3	0.024	0.034	0.012
LIA-RAG:1	-	0.037	-
Basline-MMR	0.029	0.045	0.020
Basline-L	0.023	0.040	0.015
Basline-LB	0.025	0.046	0.027

Table 3: ROUGE-2 performance of the submitted systems and baselines for each of the languages. Confidence intervals are not given but are very tight (±0.005).

An analysis of the consistency of human synopsis writers is outlined in Table 4. Consistency is computed by considering in turn each of the human synopses as system output, and computing ROUGE-2 performance. Humans have much better scores than the systems, showing that they are consistent in producing the gold standard. However, human annotators suffer from a much higher performance variance than systems (for which confidence intervals are 4-5 times smaller). This partly comes from the low number of manual synopses which is greater impacted by resampling than if there were hundreds of references for each conversation. It also comes from local inconsistencies between humans on a given conversation, resulting in diverging choices in term of which information is important.

Annotator	FR	IT
human-1	-	0.121 ±0.023
human-2	-	0.213 ±0.023
human-3	-	0.175 ±0.022
human-4	-	0.073 ±0.014
human-5	-	0.125 ±0.018
human-A	0.194 ±0.029	-
human-B	0.207 ±0.036	-
human-D	0.077 ±0.048	-
human-F	0.057 ±0.039	-
human-G	0.113 ±0.054	-

Table 4: ROUGE-2 performance of the human annotators along with confidence intervals. Note that human-C and human-E only produced synopses in English.

Table 5 shows the impact of automatic translation on system performance for the English set. This experiment is hard to interpret as the set of conversations for automatic and manual translations is different. However, it seems that processing MT results leads to better ROUGE scores, probably due to the consistency with which the MT system translates words for both conversations and synopses (reference synopses are automatic translations of source language synopses for those conversations).

Annotator	EN-man	EN-auto
NTNU:1	0.018	0.023
NTNU:2	0.019	0.031
NTNU:3	0.015	0.024
Baseline-MMR	0.024	0.033
Baseline-L	0.015	0.030
Baseline-LB	0.023	0.027

Table 5: ROUGE-2 performance on English according to whether the conversations have been manually translated or automatically translated

5 Conclusion

The objective of the CCCS pilot task at Multiling'15 was to allow work on abstractive summarization of goal-oriented spoken conversations. This task involved generating synopses from French and Italian call centre recording transcripts, and English translations of those transcripts. Four systems were submitted by two participants, and obtained reasonable results but had trouble exceeding the performance of the extractive baselines.

Clearly, ROUGE evaluation is limited for abstractive summarization in that the wording of generated text might be very different from system to system, and from reference to reference, while conveying the same meaning. In addition, ROUGE does not assess fluency and readability of the summaries.

Future work will focus on proposing better evaluation metrics for the task, probably involving the community for manually evaluating the fluency and adequacy of the submitted system output. In addition, work will be conducted in evaluating and insuring the consistency of the human experts who create the gold standard for the task.

Acknowledgments

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement n.610916 - SENSEI.

References

Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Beze, Renato De Mori, and Eric Arbillot. 2012. Decoda: a call-centre human-human spoken conversation corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM.

Marco Dinarelli, Silvia Quarteroni, Sara Tonelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In *Proceedings of EACL Workshop on the Semantic Representation of Spoken Language*, Athens, Greece.

Berna Erol, D-S Lee, and Jonathan Hull. 2003. Multimodal summarization of meeting recordings. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 3, pages III–25. IEEE.

Daniel Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2009. A global optimization framework for meeting summarization. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4769–4772. IEEE.

Catherine Lai and Steve Renals. 2014. Incorporating lexical and prosodic information at different levels for meeting summarization. In *Fifteenth Annual Conference of the International Speech Communication Association*.

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

Yashar Mehdad, Giuseppe Carenini, Frank W Tompa, and Raymond T NG. 2013. Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 136–146.

Gabriel Murray. 2015. Abstractive meeting summarization as a markov decision process. In *Advances in Artificial Intelligence*, pages 212–219. Springer.

Evgeny A. Stepanov, Giuseppe Riccardi, and Ali Orkan Bayer. 2014. The development of the multilingual luna corpus for spoken language system porting. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2675–2678, Reykjavik, Iceland, May.

Jeremy Trione. 2014. Mthodes par extraction pour le rsum automatique de conversations parles provenant de centres dappel. In *RECITAL*.

Lu Wang and Claire Cardie. 2012. Focused meeting summarization via unsupervised relation extraction. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 304–313. Association for Computational Linguistics.