AUTOMATIC SUMMARIZATION OF CALL-CENTER CONVERSATIONS

E. Stepanov[1], *B. Favre*[2], *F. Alam*[1], *S. Chowdhury*[1], *K. Singla*[1], *J. Trione*[2], *F. Béchet*[2], *G. Riccardi*[1]

[1]Signals and Interactive Systems Lab, University of Trento, Trento, Italy
[2]Aix-Marseille Université, CNRS, LIF UMR 7279, 13000, Marseille, France

ABSTRACT

This paper presents the SENSEI approach to automatic summarization which represents spoken conversation in terms of factual descriptors and abstractive synopses that are useful for quality assurance supervision in call centers. We demonstrate a browser-based graphical system that automatically produces these summary descriptors and synopses.

Index Terms— Summarization, Speech Processing

1. INTRODUCTION

Speech is the primary medium of human communication. With the expansion of the call center industry, spoken conversation data is being generated in overwhelming amounts. Large corporations often outsource their customer support, and hosting call centers either monitor the calls in real time or record them for later review. Human reviewers are able to evaluate only a small random portion of the data (much less that 1%). However, they are required to produce reports addressing various aspects of the service they are providing. These manual evaluation and analysis services are very expensive and do not scale to the quantity of data generated by call centers. The SENSEI project addresses this problem with automatic summarization of spoken conversations.

A summary is a reduced form of an original document that carries only important information. The importance of information varies with respect to the purpose of a summary. In call centers, where the goals are to evaluate the expertise of operators, as well as to understand the content of the call in terms of topics, callers' concerns and emotions, an automatic summary should contain a range of indicators that are useful for monitoring call quality addressing all these aspects. In this paper, we present the SENSEI prototype for automatic spoken conversation summarization where a summary is organized in terms of several dimensions – objective conversation descriptions – such as factual metrics, emotional labels, discourse, and written synopses. While the main component is the synopsis, an abstractive summary of the content of the conversation, the other levels provide a wider perspective on the conversation.

The research leading to these results has received funding from the European Union - Seventh Framework Program (FP7/2007-2013) under grant agreement nº 610916 SENSEI

2. APPROACHES AND METHODS

Abstractive Conversation Summarization: Call-center conversation synopses are short summaries of the events taking place during a conversation between a caller (or user) and one or more agents. Such a synopsis should contain a description of the user need or problem, and how the agent solves that problem. It might also describe the attitude of the caller and the agent. Our approach to abstractive summarization of spoken conversations uses domain knowledge to fill hand-written templates from entities detected in the transcript of the conversation using topic-dependent rules. For example, for the public transportation domain, we first cluster conversations by topic, and then write a template for each topic. Each template is a regular language with optional and repeatable parts. Slots are expressed as cross-template variables which need to be filled from the conversation.

We performed evaluation on a subset of templates on the CCCS Shared Task for the Decoda corpus [1] using the ROUGE-2 evaluation metric [2]. The abstractive summarization systems are compared to extractive and abstractive baselines. The extractive baselines are the longest turn of the conversation, the longest turn in the first quarter of the conversation and Maximal Marginal Relevance (MMR). The first abstractive baseline consists in replacing the slot values with a bogus token, which will not be matched by Rouge during evaluation, in order to simulate the worst slot filling system. The second baseline is based on the assumption that named entities play an important role in synopses: it consists in concatenating conversation named entities until the length constraint, without repetition. This baseline achieves a very bad readability, as expected. The topline consists in replacing the slot values with those manually annotated in the reference synopses. Results are summarized in Table 2.

Template Generation: In addition to hand-written templates, which fit well-structured conversations, we address unexpected events through template generation. Following [3], additional templates are learned by extracting frequent patterns from hand-written synopses, generalizing slot variables and filling the templates with entities extracted from

Topic	Template
Itinerary	Query for itinerary (using $TRANSPORT)? from $FROM to $TO (without using $NOT_TRANSPORT)?. (Take the $LINE towards $TOWARDS from $START_STOP to $END_STOP.)*. Query for location $LOCATION.
Navigo pass	Query for (justification|refund|fares|receipt) for $CARD_TYPE. Customer has to go to offices at $ADDRESS.
Lost&found	$ITEM lost in $TRANSPORT (at $LOCATION)? (around $TIME)?. (Found, to be retrieved from $RETRIEVE_LOCATION |Not found).

Table 1. Example of templates manually created for the Decoda corpus (translated from French). We use a regular-expression formalism for denoting optional an repeatable parts.

System	Rouge-2
Longest turn extract	0.04030
Longest turn @ 25%	0.04594
MMR extract	0.04490
Hand-written templates + Bogus slots	0.02228
Named entities concatenation	0.09337
Hand-written templates + auto slots	0.10084
Abstractive topline	0.18067

Table 2. Rouge-2 results of the Decoda synopsis generation systems on a subset of the CCCS test set.

the conversation transcript. To achieve that, training synopses are aligned to conversation sentences sharing the same semantic frames; generalizable words are replaced by their WordNet synset; and sentences are clustered to form the final templates. The module depends on various NLP tasks: NER, PoS-tagging, chunking, and dependency parsing.

Other Description Levels: The other levels of conversation description are factual metrics such as conversation length, waiting times, amount of overlapping speech, and the labels provided from emotion recognition system [4]. Overlapping speech is a frequently occurring event in human-human conversations and it indicates the level of co-operation between the speakers. Emotion recognition – identification of basic and complex emotions such as anger, frustration, empathy and satisfaction – on the other hand, has a straightforward application to the evaluation of the call itself, as well as the operator expertise in handing situations. The user interface for these descriptors is shown in Figure 1.

3. CONCLUSIONS AND FUTURE DIRECTIONS

This paper describes the SENSEI approach to spoken conversation summarization that besides summarizing the content also provides several conversation descriptors. We plan on extending the system with other dimensions of spoken conversations, such as sentiment target and polarity, and argumentative structure. Both are useful for mining conversations. We will also focus on global views that can represent large sets of conversations. We plan to run extrinsic evaluation of our approach with call-center professionals.

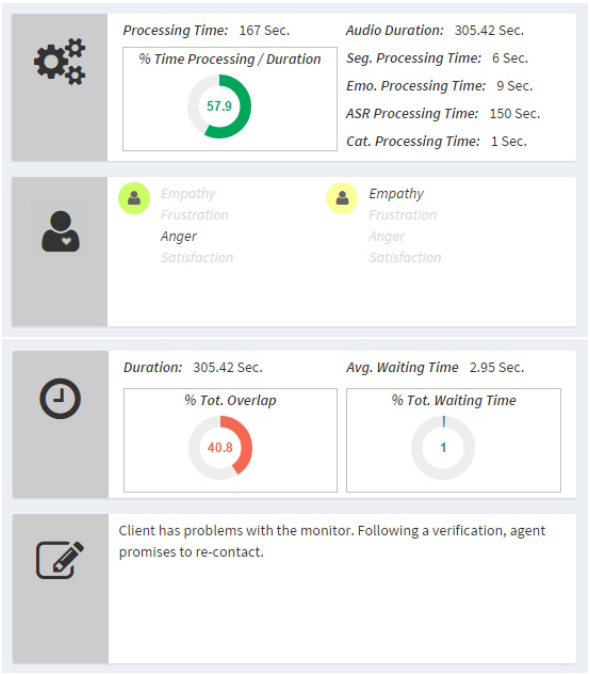

Fig. 1. Screenshots of user interface for various summary dimensions (metadata, emotion, duration, synopsis).

4. REFERENCES

[1] Benoit Favre, Evgeny Stepanov, Jeremy Trione, Frederic Bechet, and Giuseppe Riccardi, "Call centre conversation summarization: A pilot task at multiling 2015," in *Sigdial*, 2015.

[2] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL Workshop*, 2004, pp. 74–81.

[3] Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng, "A template-based abstractive meeting summarization: Leveraging summary and source text relationships," in *INLG*, 2014.

[4] Morena Danieli, Giuseppe Riccardi, and Firoj Alam, "Emotion unfolding and affective scenes: A case study in spoken conversations.," in *(ERM4CT 2015) ICMI*, 2015.