Automatic Community Creation for Abstractive Spoken Conversation Summarization

Karan Singla[1,2], **Evgeny A. Stepanov**[2], **Ali Orkan Bayer**[2],
Giuseppe Carenini[3], **Giuseppe Riccardi**[2]
[1]SAIL, University of Southern California, Los Angeles, CA, USA
[2]Signals and Interactive Systems Lab, DISI, University of Trento, Trento, Italy
[3]Department of Computer Science, University of British Columbia, Vancouver, Canada
singlak@usc.edu,carenini@cs.ubc.ca
{evgeny.stepanov,aliorkan.bayer,giuseppe.riccardi}@unitn.it

Abstract

Summarization of spoken conversations is a challenging task, since it requires deep understanding of dialogs. Abstractive summarization techniques rely on linking the summary sentences to sets of original conversation sentences, i.e. communities. Unfortunately, such linking information is rarely available or requires trained annotators. We propose and experiment automatic community creation using cosine similarity on different levels of representation: raw text, WordNet SynSet IDs, and word embeddings. We show that the abstractive summarization systems with automatic communities significantly outperform previously published results on both English and Italian corpora.

1 Introduction

Spoken conversation summarization is an important task, since speech is the primary medium of human-human communication. Vast amounts of spoken conversation data are produced daily in call-centers. Due to this overwhelming number of conversations, call-centers can only evaluate a small percentage of the incoming calls (Stepanov et al., 2015). Automatic methods of conversation summarization have a potential to increase the capacity of the call-centers to analyze and assess their work.

Earlier works on conversation summarization have mainly focused on extractive techniques. However, as pointed out in (Murray et al., 2010) and (Oya et al., 2014), abstractive summaries are preferred to extractive ones by human judges. The possible reason for this is that extractive techniques are not well suited for the conversation summarization, since there are style differences between spoken conversations and human-authored summaries. Abstractive conversation summarization systems, on the other hand, are mainly based on the extraction of lexical information (Mehdad et al., 2013; Oya et al., 2014). The authors cluster conversation sentences/utterances into communities to identify most relevant ones and aggregate them using word-graph models.

The graph paths are ranked to yield abstract sentences – a template. And these templates are selected for population with entities extracted from a conversation. Thus the abstractive summarization systems are limited to these templates generated by supervised data sources. The template selection strategy in these systems leverages on the manual links between summary and conversation sentences. Unfortunately, such manual links are rarely available.

In this paper we evaluate a set of heuristics for automatic linking of summary and conversations sentences, i.e. 'community' creation. The heuristics rely on the similarity between the two, and we experiment with the cosine similarity computation on different levels of representation – raw text, text after replacing the verbs with their WordNet SynSet IDs, and the similarity computed using distributed word embeddings. The heuristics are evaluated within the template-based abstractive summarization system of Oya et al. (2014). We extend this system to Italian using required NLP tools. However, the approach transparently extends to other languages with available WordNet, minimal supervised summarization corpus and running text. Heuristics are evaluated and compared on AMI meeting corpus and Italian LUNA Human-Human conversation corpus.

The overall description of the system with the more detailed description of the heuristics is provided in Section 2. In Section 3 we describe the corpora, evaluation methodology and the commu-

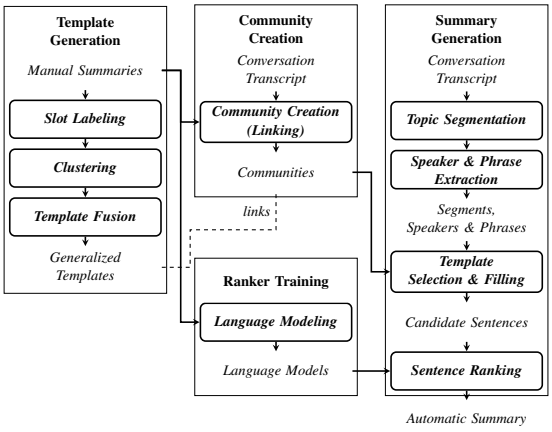

Figure 1: Abstractive summarization pipeline.

nity creation experiments. Section 4 provides concluding remarks and future directions.

2 Methodology

In this section we describe the conversation summarization pipeline that is partitioned into community creation, template generation, ranker training, and summary generation components. The whole pipeline is depicted in Figure 1.

2.1 Template Generation

Template Generation follows the approach of (Oya et al., 2014) and, starting from human-authored summaries, produces abstract templates applying slot labeling, summary clustering and template fusion steps. The information required for the template generation are part-of-speech (POS) tags, noun and verb phrase chunks, and root verbs from dependency parsing.

For English, we use Illinois Chunker (Punyakanok and Roth, 2001) to identify noun phrases and extract part-of-speech tags; and the the tool of (De Marneffe et al., 2006) for generating dependency parses. For Italian, on the other hand, we use TextPro 2.0 (Pianta et al., 2008) to perform all the Natural Language Processing tasks.

In the slot labeling step, noun phrases from human-authored summaries are replaced by WordNet (Fellbaum, 1998) SynSet IDs of the head nouns (right most for English). For a word, SynSet ID of the most frequent sense is selected with respect to the POS-tag. To get hypernyms for Italian we use MultiWordNet (Pianta et al., 2002).

The clustering of the abstract templates generated in the previous step is performed using the WordNet hierarchy of the root verb of a sentence. The similarity between verbs is computed with respect to the shortest path that connects the senses in the hypernym taxonomy of WordNet. The template graphs, created using this similarity, are then clustered using the Normalized Cuts method (Shi and Malik, 2000).

The clustered templates are further generalized using a word graph algorithm extended to templates in (Oya et al., 2014). The paths in the word graph are ranked using language models trained on the abstract templates and the top 10 are selected as a template for the cluster.

2.2 Community Creation

In the AMI Corpus, sentences in human-authored summaries are manually linked to a set of the sentences/utterances in the meeting transcripts, referred to as communities. It is hypothesized that a community sentence covers a single topic and conveys vital information about the conversation segment. For the automatic community creation we explore four heuristics.

- *H1* (baseline): take the whole conversation as a community for each sentence;
- *H2*: The 4 closest turns with respect to cosine similarity between a summary and a conversation sentence.
- *H3*: The 4 closest turns with respect to cosine similarity after replacing the verbs with WordNet SynSet ID.
- *H4*: The 4 closest turns with respect to cosine similarity of averaged word embedding vectors obtained using word2vec for a turn.(Mikolov et al., 2013).

The number of sentences selected for a community is set to 4, since it is the average size of the manual community in the AMI corpus.

We use word2vec tool (Mikolov et al., 2013) for learning distributed word embeddings. For English, we obtained pre-trained word embeddings trained on a part of Google News data set (about 3 billion words)[1]. The model contains 300-dimensional vectors for 3 million words and phrases. For Italian, we use the word2vec to train word embeddings on the Europarl Italian corpus (Koehn, 2005)[2]. We empirically choose 300, 5, and 5 for the embedding size, window length, and word count threshold, respectively.

[1] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

[2] http://www.statmt.org/europarl/

2.3 Summary Generation

The first step in summary generation is the segmentation of conversations into topics using a lexical cohesion-based domain-independent discourse segmenter – LCSeg (Galley et al., 2003). The purpose of this step is to cover all the conversation topics. Next, all possible slot 'fillers' are extracted from the topic segments and are ranked with respect to their frequency in the conversation.

An abstract template for a segment is selected with respect to the average cosine similarity of the segment and the community linked to that template. The selected template slots are filled with the 'fillers' extracted earlier.

2.4 Sentence Ranking

Since the system produces many sentences that might repeat the same information, the final set of automatic sentences is selected from these filled templates with respect to the ranking using the token and part-of-speech tag 3-gram language models. In this paper, different from (Oya et al., 2014), the sentence ranking is based solely on the n-gram language models trained on the tokens and part-of-speech tags from the human-authored summaries.

3 Experiments and Results

We evaluate the automatic community creation heuristics on the AMI meeting corpus (Carletta et al., 2006) and Italian and English LUNA Human-Human corpora (Dinarelli et al., 2009).

3.1 Data Sets

The two corpora used for the evaluation of the heuristics are AMI and LUNA. The AMI meeting corpus (Carletta et al., 2006) is a collection of 139 meeting records where groups of people are engaged in a 'roleplay' as a team and each speaker assumes a certain role in a team (e.g. project manager (PM)). Following (Oya et al., 2014), we removed 20 dialogs used by the authors for development, and use the remaining dialogs for the three-fold cross-validation.

The LUNA Human-Human corpus (Dinarelli et al., 2009) consists of 572 call-center dialogs where a client and an agent are engaged in a problem solving task over the phone. The 200 Italian LUNA dialogs have been annotated with summaries by 5 native speakers (5 summaries per dialog). For the Call Centre Conversation Summarization (CCCS) shared task (Favre et al., 2015) a set of 100 dialogs was manually translated to English. The conversations are equally split into training and testing sets as 100/100 for Italian, and 50/50 for English.

3.2 Evaluation

ROUGE-2 metric (Lin, 2004) is used for the evaluation. The metric considers bigram-level precision, recall and F-measure between a set of reference and hypothesis summaries. For AMI corpus, following (Oya et al., 2014), we report ROUGE-2 F-measures on 3-fold cross-validation. For LUNA Corpus, on the other hand, we have used the modified version of ROUGE 1.5.5 toolkit from the CCCS Shared Task (Favre et al., 2015), which was adapted to deal with a conversation-dependent length limit of 7%. Unlike the AMI Corpus, the official reported results for the CCCS Shared Task were recall; thus, for LUNA Corpus the reported values are ROUGE-2 recall.

For statistical significance testing, we use a paired bootstrap resampling method proposed in (Koehn, 2004). We create new virtual test sets of 15 conversations with random re-sampling 100 times. For each set, we compute the ROUGE-2 score and compare the system performances using paired t-test with $p = 0.05$.

3.3 Results

In this section we report on the results of the abstractive summarization system using the community creation heuristics described in Section 2.

Following the Call-Center Conversation Summarization Shared Task at MultiLing 2015 (Favre et al., 2015), for LUNA Corpus (Dinarelli et al., 2009) we compare performances to three extractive baselines: (1) the longest turn in the conversation up to the length limit (7% of a conversation) (*Baseline-L*), (2) the longest turn in the first 25% of the conversation up to the length limit (*Baseline-LB*) (Trione, 2014), and (3) Maximal Marginal Relevance (*MMR*) (Carbonell and Goldstein, 1998) with $\lambda = 0.7$. For AMI corpus, on the other hand, we compare performances to the abstractive systems reported in (Oya et al., 2014).

The performances of the heuristics on AMI corpus are given in Table 1. In the table we also report the performances of the previously published summarization systems that make use of the manual communities – (Oya et al., 2014) and (Mehdad et al., 2013); and our run of the system of (Oya et al., 2014). With manual communities we have

Model	ROUGE-2
Mehdad et al. (2013)	0.040
Oya et al. (2014) (15 seg.)	0.068
Manual Communities	0.072
(H2) Top 4 turns: token	0.076
(H3) Top 4 turns: SynSetID	0.077
(H4) Top 4 turns: Av. WE	**0.079**

Table 1: Average ROUGE-2 F-measures on 3-fold cross-validation for the abstractive summarization systems on AMI corpus.

Model	EN	IT
Extractive Systems		
Baseline-L	0.015	0.015
Baseline-LB	0.023	0.027
MMR	0.024	0.020
Abstractive Systems		
(H1) Whole Conversation	0.019	0.018
(H2) Top 4 turns: token	0.039	0.021
(H3) Top 4 turns: SynSetID	0.041	0.025
(H4) Top 4 turns: Av. WE	**0.051**	**0.029**

Table 2: ROUGE-2 recall with 7% summary length limit for the extractive baselines (Favre et al., 2015) and abstractive summarization systems with the community creation heuristics on LUNA corpus.

obtained average F-measure of 0.072. From the table, we can observe that all the systems with automatic community creation heuristics and the simplified sentence ranking described in Section 2 outperform the systems with manual communities. Among the heuristics, average word embedding-based cosine similarity metric performs the best with average F-measure of 0.079. All the systems with automatic community creation heuristics (*H2*, *H3*, *H4*) perform significantly better than the system with manual communities.

For Italian, the extractive baseline that selects the longest utterance from the first quarter of a conversation, is the strong baseline with ROUGE-2 recall of 0.027. It is not surprising, since the longest turn from the beginning of the conversation is usually a problem description, which appears in human-authored summaries. In the CCCS Shared Task, none of the submitted systems was able to outperform it. The system with a word embedding-based automatic community creation heuristic, however, achieves recall of 0.029, significantly outperforming it.

Using word embeddings allow us to exploit monolingual data, which helps to avoid the problem of data sparsity encountered using WordNet, which allows for better communities on out-of-domain data set and better coverage. This fact can account for the wider gap in performance between using H2 – H4 heuristics.

For the 100 English LUNA dialogs, we observe the same pattern as for Italian dialogs and AMI corpus: the best performance is observed for the similarity using word embeddings (0.051). However, for English LUNA, the best extractive baseline is weaker, as *H2* and *H3* heuristics are able to outperform it.

The additional observation is that the performance for English is generally higher. Moreover, word embeddings provide larger boost on English LUNA. Whether this is due to the properties of Italian or the differences in the amount and domain of data used for training word embeddings is a question we plan to address in the future. We also observe that English WordNet gives a better lexical coverage than the Multilingual WordNet used for Italian. Thus, it becomes important to explore methods which does not rely on WordNet, as now the Italian system may be suffering from the data sparsity problem due to it.

Overall, the heuristics with word embedding vectors perform the best on both corpora and across-languages. Consequently, we conclude that automatic community creation with word embedding for similarity computation is a good technique for the abstractive summarization of spoken conversations.

4 Conclusion

In this paper we have presented automatic community creation heuristics for abstractive spoken conversation summarization. The heuristics are based on the cosine similarity between conversation and summary sentences. The similarity is computed as different levels: raw text, text after verbs are replaces with WordNet SynSet IDs and average word embedding similarity. The heuristics are evaluated on AMI meeting corpus and LUNA human-human conversation corpus. The community creation heuristic based on cosine similarity using word embedding vectors outperforms all the other heuristics on both corpora, as well as it outperforms the previously published results.

We have observed that the systems generally perform better on English; and the performance differences among heuristics is less for Italian. The Italian word embedding were trained on Europarl, that is much smaller in size than the data that was used to train English embeddings. In the future we plan to address these issues and train embeddings on a larger more diverse corpus.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pages 335–336.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2006. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, Springer, pages 28–39.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*. pages 449–454.

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 *Proc. of EACL Workshop on the Semantic Representation of Spoken Language*. Athens, Greece, pages 34–41.

Benoit Favre, Evgeny A. Stepanov, Jérémy Trione, Frédéric Béchet, and Giuseppe Riccardi. 2015. Call centre conversation summarization: A pilot task at MultiLing 2015. In *The 16th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. ACL, Prague, Czech Republic, pages 232–236.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, pages 562–569.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*. Citeseer, pages 388–395.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches out: Proc. of the ACL-04 Workshop*. volume 8.

Yashar Mehdad, Giuseppe Carenini, Frank W. Tompa, and Raymond T. Ng. 2013. Abstractive meeting summarization with entailment and fusion. In *Proc. of European Natural Language Generation Workshop (ENLG)*. pages 136–146.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics, pages 105–113.

Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proc. of the 8th International Natural Language Generation Conference (INLG 2014)*. pages 45–53.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the first international conference on global WordNet*. volume 152, pages 55–63.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The textpro tool suite. In *Proc. of LREC*. ELRA.

Vasin Punyakanok and Dan Roth. 2001. The use of classifiers in sequential inference. *arXiv preprint cs/0111003* .

Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22(8):888–905.

Evgeny A. Stepanov, Benoit Favre, Firoj Alam, Shammur Absar Chowdhury, Karan Singla, Jérémy Trione, Frédéric Béchet, and Giuseppe Riccardi. 2015. Automatic summarization of call-center conversations. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) - Demo Papers*. IEEE, Scottsdale, Arizona, USA.

Jérémy Trione. 2014. Méthodes par extraction pour le résumé automatique de conversations parlées provenant de centres d'appels. In *16éme Rencontre des ètudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*. pages 104–111.