Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1771–1779
Marseille, 20-25 June 2022

ALIGNMEET: A Comprehensive Tool for Meeting Annotation, Alignment, and Evaluation

Peter Polák, Muskaan Singh, Anna Nedoluzhko, Ondřej Bojar
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
{polak,singh,nedoluzhko,bojar}@ufal.mff.cuni.cz

Abstract

Summarization is a challenging problem, and even more challenging is to manually create, correct, and evaluate the summaries. The severity of the problem grows when the inputs are multi-party dialogues in a meeting setup. To facilitate the research in this area, we present ALIGNMEET, a comprehensive tool for meeting annotation, alignment, and evaluation. The tool aims to provide an efficient and clear interface for fast annotation while mitigating the risk of introducing errors. Moreover, we add an evaluation mode that enables a comprehensive quality evaluation of meeting minutes. To the best of our knowledge, there is no such tool available. We release the tool as open source. It is also directly installable from PyPI.

Keywords: meeting summarization, meeting minutes, minuting, annotation, evaluation

1. Introduction

Meeting summarization into meeting minutes is primarily focused on topical coverage rather than on fluency or coherence. It is a challenging and tedious task, even when meeting summaries are created manually. The resulting summaries vary in the goals, style, and they are inevitably very subjective due to the human in the loop. Also, the awareness of the context of the meeting is essential to create adequate and informative summaries.

1.1. Motivation

First, there is a *scarcity of large-scale meeting datasets*: There are a few meeting corpora, such as AMI (McCowan et al., 2005) and ICSI (Janin et al., 2003), which are rather small, on the order of a few dozens of hours each as represented in Table 1. Due to this fact, meeting summarization models are usually trained on news (Grusky et al., 2018), stories (Hermann et al., 2015), Wikipedia (Frefel, 2020; Antognini and Faltings, 2020), and other textual corpora, relating poorly to meetings.

Second, when one tries to create such a collection or when a new meeting is to be processed, a *reliable transcript* is needed, which is often impossible for the current automatic speech recognition systems (ASR). It usually requires a large amount of processing to make the transcript suitable for summarization.

Third, meeting transcripts are usually *long text documents* consisting of multi-party dialogues (see Table 1) with multiple topics. Moreover, meeting summaries are also longer compared to text summaries. The manifold structure and length of meeting transcripts and summaries make it difficult to traverse and follow the information for human annotators. Even training is difficult for a neural attention summarization model (Zhu et al., 2020b) with such input complexities.

Finally, *evaluation of meeting summarization* requires immediate access to the meeting transcript and sometimes even to the original sound recording to assess the quality of a particular summary point. The length of meeting transcripts and the amount of information quantity contained in a meeting itself easily cause a significant cognitive overload.

1.2. Contribution

We present an efficient, clean, and intuitive comprehensive alignment and evaluation tool which brings the following contributions:

- An annotation platform for data creation and modification with multiple speaker support.
- Alignment between parts of a transcript with corresponding parts of summary.
- A novel evaluation strategy of meeting summaries which we integrate into the tool.

We release the tool as open source.[1] It is also directly installable from PyPI.[2]

2. Related Work

This section studies existing *annotation tools* and *evaluation strategies* for meeting summarization.

2.1. Annotation Tools

Table 2 compares ALIGNMEET with other recent annotation tools for dialogue, conversation and meeting data. Most of the tools were designed for data curation. However, only some of them allow modifying the underlying datasets (see column *D*). Segmenting the dialogues or turns is possible in some tools (see column *A*) while speaker annotation is possible in almost all tools (column *B*). ALIGNMEET provides additional features of alignment and evaluation of meeting summaries.

DialogueView (Heeman et al., 2002) is a tool for annotation of dialogues with utterance boundaries, speech

[1] `https://github.com/ELITR/alignmeet`
[2] `pip install alignmeet`

Category	Dataset	# Meetings	Avg Words (trans)	Avg Words (summ)	Avg Turns (trans)	Avg # of speakers
Meeting	AutoMin (English) (Ghosal et al., 2021)	113	9,537	578	242	5.7
	AutoMin (Czech) (Ghosal et al., 2021)	53	11,784	292	579	3.6
	ICSI (Janin et al., 2003)	61	9,795	638	456	6.2
	AMI (McCowan et al., 2005)	137	6,970	179	335	4.0
Dialogue	MEDIASum (Zhu et al., 2021)	463,596	1,554	14	30	6.5
	SAMSUM (Gliwa et al., 2019)	16,369	84	20	10	2.2
	CRD3 (Rameshkumar and Bailey, 2020)	159	31,803	2,062	2,507	9.6
	DiDi (Liu et al., 2019)	328,880	-	-	-	2.0
	MultiWoz (Budzianowski et al., 2018)	10,438	180	92	14	2.0

Table 1: Dialogue and meeting summarization datasets statistics. The number of words for dialogue, summary, turns, and speakers are averaged across the entire dataset. The meeting dataset statistics have been calculated and dialogue dataset statistics have been derived from Zhu et al. (2021).

Tool	A	B	C	D	E	F	G	H
ALIGNMEET (ours)	✓	✓	✓	✓	✓	✓	✓	Python
ELAN (Brugman et al., 2004)	✓	✓	✓	✓	✓		✓	
EXMARaLDA (Schmidt and Wörner, 2009)	✓	✓	✓	✓			✓	
MATILDA (Cucurnia et al., 2021)	✓	✓	✓	✓				Python
metaCAT (Liu et al., 2020)	✓	✓	✓					Python
LIDA (Collins et al., 2019)	✓	✓	✓					Python
INCEpTion (Klie et al., 2018)		✓						Java
DOCCANO (Nakayama et al., 2018)		✓						Python
BRAT (Stenetorp et al., 2012)		✓						Python
NITE (Kilgour and Carletta, 2006)		✓	✓	✓			✓	Java
SPAACy (Weisser, 2003)	✓	✓	✓				✓	Perl/Tk
DialogueView (Heeman et al., 2002)	✓	✓						Tcl/Tk
ANVIL (Kipp, 2001)	✓	✓	✓				✓	Java
NOMOS (Gruenstein et al., 2005)	✓	✓	✓	✓			✓	Java
TWIST (Plüss, Brian, 2012)	✓							-

Table 2: Annotation Tool Comparison Table. Notation: A – Turn/Dialogue Segmentation, B – Edit Speaker Annotation, C – Data Curation, D – Data Modifications, E – Alignment, F – Evaluation, G – Audio/video playback, H – Programming Language.

repairs, speech act tags, and discourse segments. It fails to capture inter-annotator reliability. TWIST (Plüss, Brian, 2012) is a tool for dialogue annotation consisting of turn segmentation and content feature annotation. The turn segmentation allows users to create new turn segments. Further, each segment can be labeled by selecting from a pre-defined feature list. This limits the user to pre-defined values. BRAT (Stenetorp et al., 2012) and DOCCANO (Nakayama et al., 2018) are simple web-based annotation tools where you can only edit the dialogue and turns. BRAT also provides the user with automated recommendations. INCEpTion (Klie et al., 2018) is a platform for annotation of semantic resources such as entity linking. It provides automated recommendations to the user for annotation. NOMOS (Gruenstein et al., 2005) is an annotation tool designed for corpus development and various other annotation tasks. Its main functionality includes multi-channel audio and video playback, compatibility with different corpora, platform independence and presentation of temporal, non-temporal, and related information. This tool is difficult to use by non-technical users and also lacks extensibility. ANVIL (Kipp, 2001) allows multi-modal annotation of dialogues with the granularity in multiple layers of attribute-value pairs. It also provides the feature of statistical processing but lacks the flexibility to add information to the annotation. NITE (Kilgour and Carletta, 2006) is another multi-modal annotation tool aiding in corpora creation. The tool supports the time-alignment of annotation entities such as transcripts or dialogue structure. SPAACy (Weisser, 2003) is a semi-automated tool for annotating dialogue acts. It aids in corpus creation with tagging such as topic, mode, and polarity. In addition, it produces transcriptions in XML format that require a little post-editing. LIDA (Collins et al., 2019) is one of the most prominent tools for modern task-oriented dialogues with recommendations. However, LIDA does not support more than two speakers in the conversation or additional labeling (e.g., co-reference annotation). MATILDA (Cucurnia et al., 2021) and metaCAT (Liu et al., 2020) address some of the downsides. They add features such as inter-annotator disagreement resolution, customizable recommendations, multiple-language support, and user administration. They still lack support for multiple speakers.

All these annotation tools provide annotation for dialogues, but for various textual phenomena. Our tool ALIGNMEET is specifically designed for meeting data creation or modification, alignment of meeting transcript regions with the corresponding summary items, and their evaluation. We also support dialogue and conversational datasets.

2.2. Manual Evaluation

Several researchers working on summarization have considered qualitative summary evaluation. The qualitative parameters include *accuracy* (Zechner, 2001b; Zechner, 2001a; Goo and Chen, 2018; Nihei et al., 2018; Lai et al., 2013) which usually assesses the lexical similarity between produced text samples and the reference ones utilizing standard metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004). The accuracy is easily computed in some of the applications when reference texts are available. *Grammaticality* measures the capability of a model to produce grammatically correct texts (Liu and Liu, 2009; Mehdad et al., 2013). It is mostly assessed by counting the different types of errors. *Adequacy* (D'Haro et al., 2019; Ma and Sun, 2017; McBurney and McMillan, 2014; Arumae and Liu, 2019; Libovický et al., 2018) rates the amount of meaning expressed in the generated sample given a reference sample. Human participants and categorical scales dominate the assessment process. *Topicality* expresses how well does the generated sample topic match one of the reference samples (Riedhammer et al., 2008; Arumae and Liu, 2019; Fang et al., 2017). *Naturalness* shows the likelihood of a text being natural or written by a human being rather than automatically generated (Çano and Bojar, 2020). *Relevance* represents how closely are the documents related (Bhatia et al., 2014; Erol et al., 2003; Murray et al., 2010; Zhu et al., 2020a; Zhang and Fung, 2012; Zhu et al., 2020b; Lee et al., 2020). *Consistency* represents the degree of agreement with the original content (Kryściński et al., 2019; Wang et al., 2020; Lee et al., 2020). *Fluency* represents the quality of expression (Oya, 2014; Wang and Cardie, 2013; Oya et al., 2014; Lee et al., 2020). *Coverage* determines how much of the important content is covered from the source document in the summary (Sonjia and Gina-Anne, 2008; Gillick et al., 2009; Li et al., 2019; Mehdad et al., 2013). *Informativeness* represents the importance of the content captured in the summary (Zhang et al., 2021; Liu and Liu, 2009; Oya et al., 2014; Oya, 2014). Besides accuracy, the rest of the above quality criteria are assessed manually by human experts or survey participants (Zhu and Penn, 2006; Shirafuji et al., 2020).

2.3. Automatic Evaluation

The current automatic evaluation of various text summarization tasks (including minuting) is mostly based on ROUGE or similar metrics that utilize n-gram comparisons (from single words to long patterns). While automatic and fast, these metrics are often not able to reflect the quality issues of the text samples (See et al., 2017). Some of the typical problems they miss are grammatical discrepancies, word repetitions, and more. Novikova et al. (2017; Reiter (2018) also report that automatic metrics do not correlate well with human evaluations. To overcome these limitations, it is important to simultaneously run human evaluations (following a systematic protocol) of meeting summaries and augment the automatic metric scores with the manual ones.

3. The ALIGNMEET Annotation Tool

ALIGNMEET is a comprehensive annotation and evaluation tool. It supports all stages of the preparation and/or evaluation of a corpus of multi-party meetings, i.e., creation and editing of meeting transcripts, annotating speakers, creating a summary, alignment of meeting segments to a summary, and meeting summary evaluation.
The tool is written in Python using PySide[3] for GUI which makes the tool available on all major platforms (i.e., Windows, Linux, and macOS).

3.1. Design Choices

We represent a meeting with its transcript and summary in Figure 1. The transcripts are long documents consisting of multi-party dialogues (refer to the left side of the tool window). The meeting summary is a structured document. We decided to break down the meeting summary into separate *summary points*. A summary point roughly represents a line in a summary document (refer to the right part of the tool window). The meeting usually has more versions of transcripts (e.g., generated by ASR and a manual one) and more versions of summaries (e.g., supplied by meeting participants created during the meeting and others provided by an annotator). We add drop-down lists to select a specific version of the transcript and summary. If the user changes the version of one, the program loads the appropriate version of the alignment automatically.
We segment the transcript into dialogue acts (DAs). A DA represents the meaning of an utterance at the level of illocutionary force (Austin, 1975). In the context of our tool, a DA represents a continuous portion of a transcript uttered by one speaker on a single topic. We believe that for better readability, the DA might be further broken down into smaller units.
Optionally, the meeting might have an audio or video recording. The meeting recording is helpful during the meeting annotation (i.e., creating/editing the meeting transcript and summary). The tool offers an embedded player. Then, the annotator does not have to switch between the annotation tool and a media player. Also, if the transcripts come with timestamps, the annotator can easily seek in the player by double-clicking the particular DA.
Many annotation tools we reviewed in Section 2.1 provide automated suggestions. We decided not to include this feature as we believe it would bias the annotators.
ALIGNMEET is designed with two modes: Annotation and Evaluation. We further elaborate them in Sections 3.2 and 3.3.

3.2. Annotation

The annotation task consists of several sub-tasks. We envision the following sub-tasks: (1) transcript annotation, (2) summary annotation, and (3) alignment.

3.2.1. Transcript Annotation

Transcripts may be either generated by an ASR or manually created. The tool supports both scenarios, i.e.,

[3] `https://www.qt.io/qt-for-python`

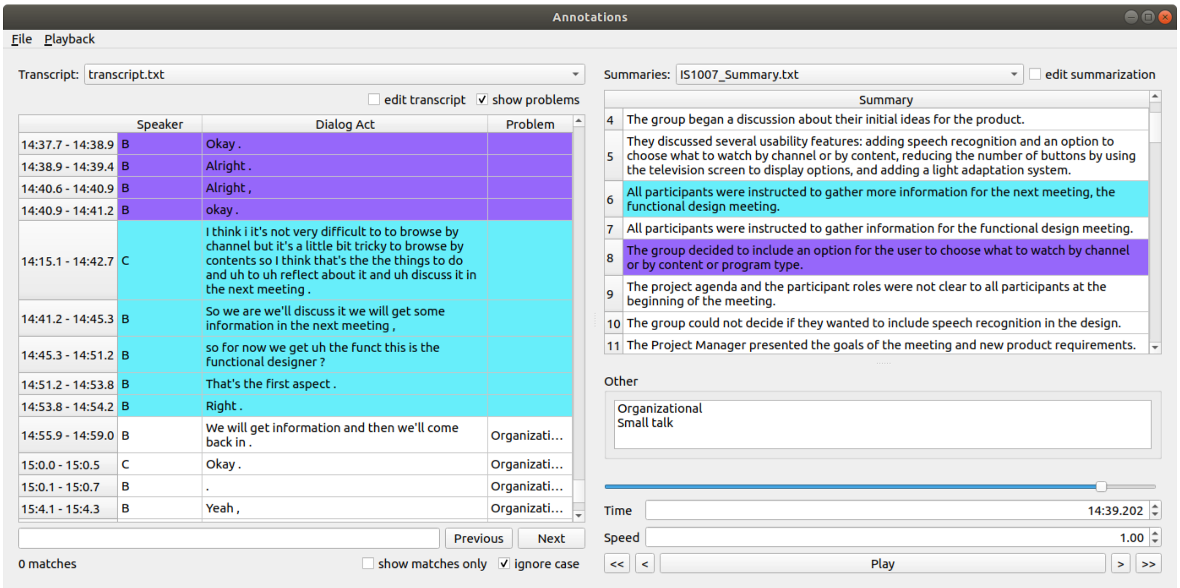

Figure 1: The ALIGNMEET main view in the annotation mode. The left column contains the meeting transcript broken down to dialogue acts. The right column contains a summary, and the player. The alignment between dialogue acts and the summary point is shown using colors.

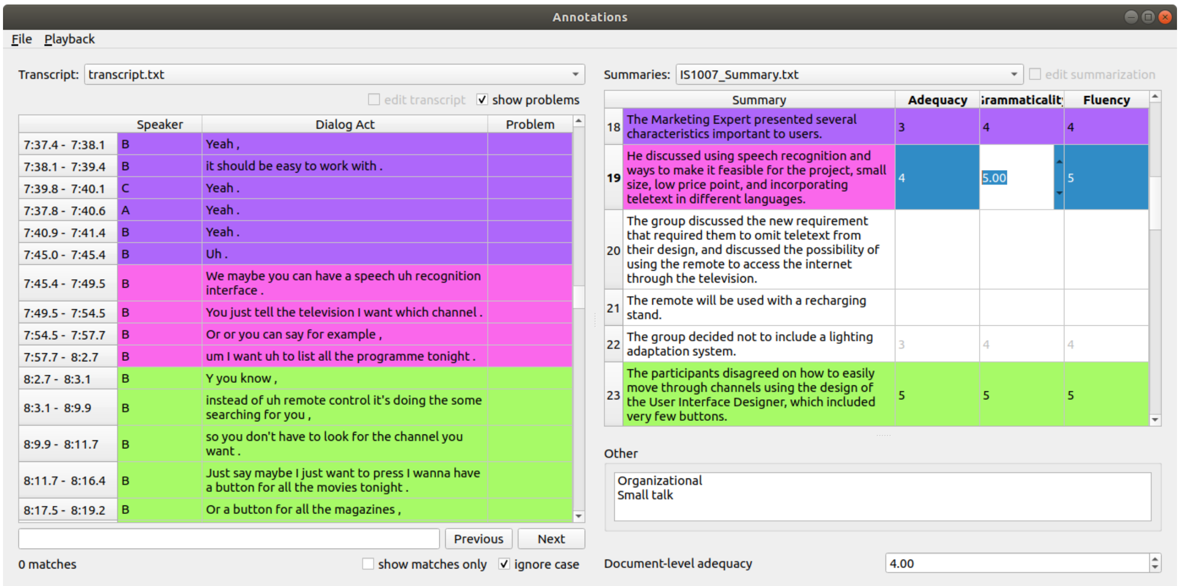

Figure 2: The ALIGNMEET main view in evaluation mode. The left column contains the meeting transcript broken down into dialogue acts. The right column contains a summary, problem flags, and document-level adequacy. Evaluation, i.e., the assignment of scores to a particular summary point, is enabled only for the summary points where the corresponding DAs are visible in the transcript view.

transcribing the recording, post-editing, splitting the transcript into dialogue acts, and speaker annotation. We introduce a set of keyboard shortcuts that make simple tasks like creating/deleting or even splitting DAs very quick. Additionally, we offer a search toolbar supporting regular expressions.

3.2.2. Summarization Annotation

Summarization annotation involves the creation or possible modification of an existing meeting summary. The tool provides a convenient platform to add more points to an existing summary by simply clicking the "add" or "delete" buttons.

Except for summary points, we intentionally do not enforce any precise summary structure and provide users with the flexibility to design their summary. Though, we support indentation as a form of horizontal structuring (with a user-defined indentation symbol).

3.2.3. Alignment

The alignment captures which dialogue acts are associated with a particular summary point. We call a set of DAs belonging to a summary point a *hunk*. DAs which do not correspond to a summary point may be assigned meta-information (i.e., marked as small talk or organizational).

ALIGNMEET supports only n-to-1 alignments because we believe that aligning multiple summary points to a DA would further increase the difficulty of the alignment task. It would also cause a "summary point fragmentation", as the annotator might address the same information in separate summary points. When a DA includes more information that fits in a single summary point, we suggest splitting the DA accordingly.

The matching background color of a hunk and a summary point represents a single alignment (see Figure 1). To make the interface more clean and readable for the annotator, we color only summary points whose hunks are currently visible in the transcript view.

Aligning DA(s) to a particular summary point or meta-information item is very intuitive:

1. *Select DA(s)* in the transcript view. The selection can be contiguous and also discontiguous. Standard GUI gestures are supported (i.e., dragging over items, `[Ctrl]/[Shift]` + clicking/dragging).
2. *Select a summary point* by double-clicking an item in the summary view or choose a *problem label* in the meta-information list.

Resetting the alignment is also possible by selecting DA(s) or a summary point and selecting an action in the context menu or keyboard shortcut.

In this way, ALIGNMEET facilitates the annotation and mitigates potential errors. The annotator has a clear overview of which parts of a meeting are already annotated and makes any revisions straightforward.

3.3. Evaluation Mode

We reviewed several quality criteria for a summary evaluation in Sections 2.2 and 2.3 based on which we formulate a novel manual evaluation scheme. We integrated the evaluation into the tool (see Figure 2).

For the evaluation, we utilize *adequacy*, *grammaticallity* and *fluency*. We think that evaluating these criteria at the document level is challenging and error-prone. Therefore, we propose the evaluation on two levels: (1) manually assigning the hunk level (based on alignment) and (2) automatically aggregating it on the document level. At the hunk level, the evaluation is based only on the aligned part of the transcript and a corresponding summary point.

At the hunk-level, annotators evaluate *adequacy*, *grammaticality* and *fluency* using a 5-star scale (Likert, 1932) with 1 being the worst and 5 the best. At the document level, we automatically aggregate the hunk-level judgments with a simple average. Aside from averaging hunk-level adequacy across the document, we also independently ask annotators to report the overall accuracy of the minutes. We call this score '*Doc-level adequacy*' in the following. Finally, we compute *coverage*, i.e., the number of aligned DAs divided by the total number of DAs.

4. Use Case and Pilot Study

In this section, we present a use case and conduct a small-scale pilot study.

4.1. Use Cases

We organized the First Shared Task on Automatic Minuting (Ghosal et al., 2021) on creating minutes from multi-party meetings. As a part of the shared task, we made available a minuting corpus, which is now being released publicly (Nedoluzhko et al., 2022). ALIGNMEET was created during the annotation process. We have started with a modified NITE (Kilgour and Carletta, 2006) tool, but the annotators faced many issues, including the need to make changes to the transcript and minutes. Hence, we decided to create a new tool to meet the annotators' requirements. We used agile development, i.e., we constantly improved ALIGNMEET following the annotators' comments.

Before annotation, each meeting consisted of a recording, ASR-generated transcript, and meeting minutes assembled by the meeting participants (often incomplete). First, we asked the annotators to revise the ASR transcript. Later, we asked the annotators to provide minutes and alignment. We have observed different styles of minuting among the annotators. Therefore, many of the meetings have two or more versions of minutes provided by different annotators.

4.2. Pilot Study

To assess ALIGNMEET, we conduct a simple experiment similar to Collins et al. (2019) for both modes of tool: (1) annotation and (2) evaluation. We evaluate all the results across two different meeting corpora, AMI (McCowan et al., 2005) for English and AutoMin for Czech. We considered one meeting per language from each corpus (the selected English meeting has 205 DAs and the selected Czech meeting has 153 DAs; both are approximately 16 minutes long). The task was to create an abstractive summary, align the transcript with the corresponding parts of the reference summary, and finally evaluate the reference summary relying on the constructed alignment. Each of the three annotators had a different experience level and report their timings in Table 3. The summarization stage took on average 40.7 minutes and 33.0 minutes for English and Czech, respectively. The alignment took on average 16.0 and 19.7 minutes and evaluation on average of 11.7 and 17.7 minutes for English and Czech data, respectively. In other words, this particular meeting needed about 2–3 times its original time to summarize, its duration to align, and finally somewhat less than its duration to evaluate. Based on this minimal study, a factor of 4 or

	English			Czech		
Annotator	E1	E2	E3	C1	C2	C3
Experienced	✗	✓	✓	✗	✓	✓
Summarization	37	45	40	23	45	31
Alignment	5	23	20	18	30	11
Evaluation	10	15	10	25	15	13
Total time	52	83	70	66	90	55

Table 3: Pilot study: annotator experience and time in minutes each annotator spent on each task.

more has to be expected when processing meetings by annotators who have not taken part in them.

The evaluation results are in Table 4. Adequacy is deemed average (3.98±0.62 on average), with the document-level manual judgment being similar (3.83±0.37), while grammaticality and fluency are somewhat higher (4.32±0.39 and 4.63±0.31, resp.). Additionally, we report the inter-annotator agreement (IAA). Our definition of IAA is rather strict, we count the number of DAs that were aligned to the same summary point by all annotators divided by the total number of DAs.

If we consider the recorded pace of our annotators, the AMI meeting corpus consisting of 137 meetings and 45,895 DAs in total (see Table 1), it would take 9,105 minutes to summarize, 3,582 minutes to align, and 2,613 minutes to evaluate using our tool, or 255 hours in total. We infer from Table 3 that the time spent on the task does not necessarily depend on the annotator's experience but rather on the personal preferences and thoroughness of the annotator. Despite the limited size of the experiment, we believe that the results suggest the tool is intuitive and facilitates fast annotation.

5. Conclusion

We presented ALIGNMEET, an open-source and intuitive comprehensive tool for meeting annotation. Its main goal is to facilitate alignment between parts of a transcript with the corresponding part of the summary. We also integrate the proposed evaluation strategy of meeting summaries in the tool.

In the future, we will add the support for automatic transcript generation with timestamps, user-defined problems in the list of explicit problem labels, and a quick onboarding tutorial integrated into the user interface. Finally, we hope ALIGNMEET will generally improve as annotators will provide their feedback.

Acknowledgements

This work has received support from the project "Grant Schemes at CU" (reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935), the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460 (ELITR), and 19-26934X (NEUREM3) of the Czech Science Foundation, and partially supported by SVV project number 260 575.

	English			Czech		
Annotator	E1	E2	E3	C1	C2	C3
Experienced	✗	✓	✓	✗	✓	✓
#Summary points	15	11	19	23	14	21
#Alignments	378	378	203	282	282	282
IAA		0.21*			0.63	
Avg. adequacy	3.71	3.71	3.17	3.67	4.93	4.67
Avg. grammaticality	3.86	4.21	4.08	5.00	4.13	4.67
Avg. fluency	4.71	4.07	4.92	5.00	4.53	4.53
Doc.-level adequacy	3.00	4.00	4.00	4.00	4.00	4.00
Coverage	1.00	0.94	0.54	0.64	0.54	0.30

Table 4: Pilot study: annotator experience, number of produced summary points and alignments, and evaluation score.

∗ If we remove the second annotator, we obtain agreement 0.59.

6. Bibliographical References

Antognini, D. and Faltings, B. (2020). GameWikiSum: a novel large multi-document summarization dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6645–6650, Marseille, France, May. European Language Resources Association.

Arumae, K. and Liu, F. (2019). Guiding extractive summarization with question-answering rewards. *arXiv preprint arXiv:1904.02321*.

Austin, J. L. (1975). *How to do things with words*, volume 88. Oxford university press.

Bhatia, S., Biyani, P., and Mitra, P. (2014). Summarizing online forum discussions–can dialog acts of individual messages help? In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2127–2131.

Brugman, H., Russel, A., and Nijmegen, X. (2004). Annotating multi-media/multi-modal resources with elan. In *LREC*, pages 2065–2068.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November. Association for Computational Linguistics.

Çano, E. and Bojar, O. (2020). Human or machine: Automating human likeliness evaluation of nlg texts. *arXiv preprint arXiv:2006.03189*.

Collins, E., Rozanov, N., and Zhang, B. (2019). Lida: Lightweight interactive dialogue annotator. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 121–126.

Cucurnia, D., Rozanov, N., Sucameli, I., Ciuffoletti, A., and Simi, M. (2021). Matilda-multi-annotator multi-language interactivelight-weight dialogue annotator. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 32–39.

D'Haro, L. F., Banchs, R. E., Hori, C., and Li, H. (2019). Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. *Computer Speech & Language*, 55:200–215.

Erol, B., shyang Lee, D., and Hull, J. (2003). Multimodal summarization of meeting recordings. In *In Proceedings of the IEEE International Conference on Multimedia & Expo*, Baltimore, MD, July.

Fang, C., Mu, D., Deng, Z., and Wu, Z. (2017). Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems with Applications*, 72:189–195.

Frefel, D. (2020). Summarization corpora of wikipedia articles. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6651–6655.

Ghosal, T., Singh, M., Nedoluzhko, A., and Bojar, O. (2021). Overview of the first shared task on automatic minuting (automin) at interspeech 2021. In *In print*.

Gillick, D., Riedhammer, K., Favre, B., and Hakkani-Tur, D. (2009). A global optimization framework for meeting summarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4769–4772. IEEE.

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (2019). SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November. Association for Computational Linguistics.

Goo, C. and Chen, Y. (2018). Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742, Athens, Greece, Dec. IEEE Xplore.

Gruenstein, A., Niekrasz, J., and Purver, M. (2005). Meeting structure annotation: Data and tools. In *6th SIGdial Workshop on Discourse and Dialogue*.

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June. Association for Computational Linguistics.

Heeman, P. A., Yang, F., and Strayer, S. E. (2002). Dialogueview-an annotation tool for dialogue. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 50–59.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Kilgour, J. and Carletta, J. (2006). The nite xml toolkit: Demonstration from five corpora. In *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*.

Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. (2019). Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551.

Lai, C., Carletta, J., and Renals, S. (2013). Detecting summarization hot spots in meetings using group level involvement and turn-taking features. In *INTERSPEECH 2013 14th Annual Conference of the International Speech Communication Association*, pages 2723–2727, Lyon, France. ICSA.

Lee, D., Shin, M., Whang, T., Cho, S., Ko, B., Lee, D., Kim, E., and Jo, J. (2020). Reference and document aware semantic evaluation methods for korean language summarization. *arXiv preprint arXiv:2005.03510*.

Li, M., Zhang, L., Ji, H., and Radke, R. J. (2019). Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196.

Libovický, J., Palaskar, S., Gella, S., and Metze, F. (2018). Multimodal abstractive summarization of open-domain videos. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL). NIPS*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22:55.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, F. and Liu, Y. (2009). From extractive to abstractive meeting summaries: Can it be done by sentence compression? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 261–264.

Liu, C., Wang, P., Xu, J., Li, Z., and Ye, J. (2019). Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965.

Liu, X., Xue, W., Su, Q., Nie, W., and Peng, W. (2020). metacat: A metadata-based task-oriented chatbot annotation tool. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–25.

Ma, S. and Sun, X. (2017). A semantic relevance based neural network for text summarization and text simplification. *arXiv preprint arXiv:1710.02318*.

McBurney, P. W. and McMillan, C. (2014). Automatic documentation generation via source code summarization of method context. In *Proceedings of the 22nd International Conference on Program Comprehension*, pages 279–290.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al. (2005). The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100. Citeseer.

Mehdad, Y., Carenini, G., Tompa, F., and Ng, R. (2013). Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 136–146.

Murray, G., Carenini, G., and Ng, R. (2010). Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference*.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Nedoluzhko, A., Singh, M., Hledíková, M., Ghosal, T., and Bojar, O. (2022). ELITR Minuting Corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC-2022)*, Marseille, France, June. European Language Resources Association (ELRA). In print.

Nihei, F., Nakano, Y. I., and Takase, Y. (2018). Fusing verbal and nonverbal information for extractive meeting summarization. In *Proceedings of the Group Interaction Frontiers in Technology*, GIFT'18, New York, NY, USA. Association for Computing Machinery.

Novikova, J., Dušek, O., Cercas Curry, A., and Rieser, V. (2017). Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark, September. Association for Computational Linguistics.

Oya, T., Mehdad, Y., Carenini, G., and Ng, R. (2014). A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53.

Oya, T. (2014). Automatic abstractive summarization of meeting conversations. Master's thesis, University of British Columbia, Vancouver, Canada.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Plüss, Brian. (2012). Annotation study materials. `http://mcs.open.ac.uk/nlg/non-cooperation/resources/study-materials.pdf`.

Rameshkumar, R. and Bailey, P. (2020). Storytelling with dialogue: A critical role dungeons and dragons dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134.

Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, September.

Riedhammer, K., Favre, B., and Hakkani-Tur, D. (2008). A keyphrase based approach to interactive meeting summarization. In *2008 IEEE Spoken Language Technology Workshop*, pages 153–156. IEEE.

Schmidt, T. and Wörner, K. (2009). Exmaralda–creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19(4):565–582.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July. Association for Computational Linguistics.

Shirafuji, D., Kameya, H., Rzepka, R., and Araki, K. (2020). Summarizing utterances from japanese assembly minutes using political sentence-bert-based method for qa lab-poliinfo-2 task of ntcir-15. *arXiv preprint arXiv:2010.12077*.

Sonjia, W. and Gina-Anne, L. (2008). Topic summarization for multiparty meetings.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Wang, L. and Cardie, C. (2013). Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405.

Wang, A., Cho, K., and Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Weisser, M. (2003). Spaacy–a semi-automated tool for annotating dialogue acts. *International journal of corpus linguistics*, 8(1):63–74.

Zechner, K. (2001a). Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *IN PROC. ACM SIGIR*, pages 199–207, New Orleans, USA. ACM.

Zechner, K. (2001b). *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA.

Zhang, J. J. and Fung, P. (2012). Automatic parliamentary meeting minute generation using rhetorical structure modeling. *IEEE transactions on audio, speech, and language processing*, 20(9):2492–2504.

Zhang, X., Zhang, R., Zaheer, M., and Ahmed, A. (2021). Unsupervised abstractive dialogue summarization for tete-a-tetes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14489–14497, May.

Zhu, X. and Penn, G. (2006). Summarization of spontaneous conversations. In *Ninth International Conference on Spoken Language Processing*.

Zhu, C., Xu, R., Zeng, M., and Huang, X. (2020a). End-to-end abstractive summarization for meetings. *CoRR*, abs/2004.02016.

Zhu, C., Xu, R., Zeng, M., and Huang, X. (2020b). A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online, November. Association for Computational Linguistics.

Zhu, C., Liu, Y., Mei, J., and Zeng, M. (2021). Mediasum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934.