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 Fallings, 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., 2020) 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. It is also directly installable from PyPI.

2. Related Work
This section studies existing annotation tools and evaluation strategies for meeting summarization.

2.1. Annotation Tools
Table 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
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, 2001a; Zechner, 2001b; Zechner, 2001c; Goo and Chen, 2018; Nihei et al., 2018; Lai et al., 2013) which usually assesses the lexical similarity between produced text samples and the original text.
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; Mccurney 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 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 (Cano and Bojar, 2020). Relevance represents how closely 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 (Krysciński 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., 2019).

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 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.,

[https://www.qt.io/qt-for-python](https://www.qt.io/qt-for-python)
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 \( n \rightarrow 1 \) 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, we propose the evaluation on two levels: (1) document level and (2) each dialogue act (DA) level. At the document level, the evaluation is based only on 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 16.0 and 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. 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.

4. Use Case and 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
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.

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

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Summarization</th>
<th>Alignment</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exper.</td>
<td>E1 E2 E3 C1 C2 C3</td>
<td>E1 E2 E3 C1 C2 C3</td>
<td>E1 E2 E3 C1 C2 C3</td>
</tr>
<tr>
<td>Experienced</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>Total time</td>
<td>52 83 70</td>
<td>66 90 55</td>
<td>52 83 70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Summary points</td>
<td>15 11 19</td>
</tr>
<tr>
<td>#Alignments</td>
<td>378 378 203</td>
</tr>
</tbody>
</table>

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

## 6. Bibliographical References


