RC25673 (IRE1803-019) March 21, 2018
Computer Science

IBM Research Report

Preparing a Dataset for Extracting Decision Elements from a Meeting Transcript Corpus

Tuan Tran
L3S Research Center
Hannover, Germany

Francesca Bonin, Léa A. Deleris, Debasis Ganguly, Killian Levacher
IBM Research
IBM Dublin Technology Campus, Bldg. 3
Damastown Industrial Estate
Mulhuddart
Dublin 15, Ireland

Research Division
Almaden – Austin – Beijing – Cambridge – Dublin - Haifa – India – Melbourne - T.J. Watson – Tokyo - Zurich

Preparing a Dataset for Extracting Decision Elements from a Meeting Transcript Corpus

Tuan Tran, Francesca Bonin, Léa A. Deleris, Debasis Ganguly, Killian Levacher
L3S Research Center, IBM Research Ireland
Hannover (Germany), Dublin (Ireland)
ttran@L3S.de, fbonin@ie.ibm.com, debasis.ganguly@ie.ibm.com, killian.levacher@ibm.com, lea.deleris@ie.ibm.com

Abstract

This work describes the construction of a new dataset for the purpose of decision element extraction from a meeting corpus. Specifically, the corpus consists of annotated text spans of alternatives and criteria of decisions undertaken during spoken conversations. The annotations were conducted with the help of crowd sourcing and finally curated by a domain expert. Our experiments show that the curated dataset can lead to more consistent predictions in comparison to the crowdsourced one. The aim of the released dataset is to encourage further studies in automated information extraction for decision analysis, e.g. evaluating the effectiveness of supervised models for this task.

1. Introduction

Decision Analysis is the scientific discipline that formally studies decision processes: procedures, methods, and tools for identifying, representing, and assessing important aspects of a decision and decision process, ultimately to recommend actions to the decision maker (Matheson and Howard, 1977). In multi-party dialogues, while some decisions (often very strategic ones) are framed through controlled and structured approaches (e.g. facilitated by decision analysis consultants), many others are informally discussed during conversations, making the data unstructured and difficult to process. As a result, most decision discussions do not benefit from the structure and insights brought by decision analysis.

With the proliferation of multimodal resources in our professional and personal lives (e.g., teleconferencing, recording of meetings, skype calls, slack channels), it would be helpful to develop automated tools to automatically extract decision related concepts such as alternatives and criteria from these unstructured natural conversations.

Automatic identification of such decision oriented entities is likely to improve the effectiveness of extractive summarization focused on decision elements. It would provide insights into how decisions are made in practice. Specifically, such extracted entities would enable to : *1)* inform people who were not in the meeting, *2)* remind a decision maker of the arguments that were raised so s/he can make her/his decisions after the discussion, *3)* record the decision related process information in a structured way, including how the decision was made and what were the consensus and dissent expressed. In the long term, this process could enable decision analysts to develop new techniques for decision facilitation and to apply structure to decision-making sessions *without attending in person*. In order to achieve this, the first step is to build appropriate resources.

This paper reports our work to annotate a subset of the Augmented Multi-party Interaction (AMI) Meeting Corpus (Carletta et al., 2005). The annotation considers two types of decision elements: **Alternatives** i.e., options which were discussed during meetings, and *Criteria* i.e., arguments for and against each alternative. Therefore, in the utterance

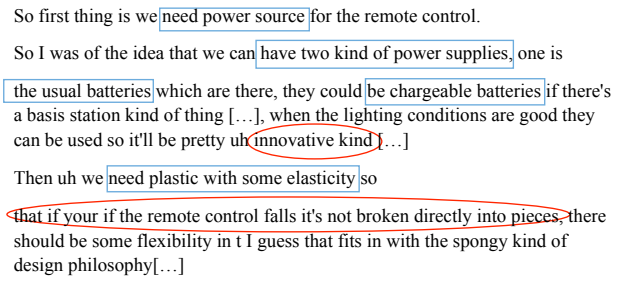

Figure 1: Example *Alternatives* in blue rectangles and *Criteria* in red ovals.

"*Should I make the party in the garden or inside?*" *Alternatives* are "*in the garden*" and "*inside*", *Criteria* are the personal preferences, i.e. "*space*" and "*warmer*" in the following sentence "*We would have more space in the garden but it would be warmer inside*". Figure 1 shows an excerpt of an AMI meeting, with highlighted *Alternatives* and *Criteria*.

We collaborated with a Decision Analysis expert to identify such elements in natural conversations and developed an associated annotation scheme. The annotation process was performed both in an expert-based and crowdsourced fashion, and the resulting annotated meetings are made available together with this paper.

The rest of the paper is structured as follows: Section 2. introduces related works, section 3. describes the AMI corpus and section 4. the annotation scheme. In section 5., we describe the annotation process which we evaluate in section 6..

2. Related works

Decisions are one of the most important outcomes of business meetings. The work of (Banerjee et al., 2005) shows that updates about the decisions of a meeting are beneficial for persons who had missed the meeting to prepare for the next one. In (Whittaker et al., 2006) , authors reported that users tend to take two kinds of notes in meetings, one of which pertains to the decisions taken.

Interest on meeting developments is shown also by the large amount of corpus collections on the topic, e.g., ICSI (Janin et al., 2004) , AMI (Carletta et al., 2005), CHIL (Mostefa et al., 2007) or VACE (Chen et al., 2006). See (Strauß and Minker, 2010) for an extensive description of such corpora. While some annotations in these corpora consider decisions from meetings, these annotated labels (text spans) are either too specific (dialogue acts) or too general (meeting summaries), and likely not useful to study how decisions are framed.

Research in argumentation meeting annotation was the Twente Argument Schema (TAS) coding scheme (Rienks and Verbree, 2005), a model that formalises observations related to argumentation patterns in meetings. The idea is to capture the most important conversational moves in dialogues where participants discuss the pros and cons of certain solutions to a problem, providing arguments in favour of or against the various solutions. However, the TAS scheme does not suit our task since it does not distinguish concepts that are essential in decision analysis for a formal investigation of the decision-making process (i.e. alternatives and criteria).

Some studies have been conducted on automatic detection of decisions. Hsueh and Moore (2007) attempted to identify patterns of the decision gists, relying on the relevant annotated dialogue acts (DAs) in meeting transcripts. They used 50 meetings and found on average four decisions per meeting with a Cohen's kappa ranging from 0.5 to 0.8. Fernández et al. (2008) extended the annotations with new decision-related DAs, and formulated the problem as a classification problem for each class of DAs. They designed an annotation scheme that takes into account the different roles that different DA play in the decision-making process (decision DAs, called DDAs). For instance, DDAs to initiate a discussion by raising a topic, DDAs to propose a resolution, DDAs to express agreement. Note however, in all this work, the objective was to detect the span of the conversation where the decision is taken. By contrast, our intention, instead, is to annotated elements that belongs to the *content* of decision-making process, whether or not a final decision is taken.

Our work is closest to the effort from (Cadilhac et al., 2012). While focusing more specifically on the representation of preferences, they have proposed a corpus and associated approach to extract what we refer to as alternatives and which in their framework is described as outcomes. However, they do not pursue the extraction of criteria.

3. Corpus

The AMI (Augmented Multi-party Interaction) Meeting Corpus is an English multi-modal data set consisting of 100 hours of meeting recordings (Carletta et al., 2005). The dataset is constituted of real meetings, as well as scenario-driven meetings, designed to elicit several realistic human behaviours. We selected 43 meetings under the criteria that those meetings needed to have some form of decision-making in them, in line with (Fernández et al., 2008). We report the statistics over these meetings in Table 1. We use the manually annotated transcripts for this study. Each meeting has four participants, and the same subjects meet over four different sessions to discuss a design project. Figure 1 provides an example of text from the AMI corpus.

4. Annotation scheme

We analysed this subset of 43 meetings with the domain expert. Annotators were asked to freely highlight text spans that represent a decision element. We noticed that, at a syntactic level, *Alternatives* and *Criteria* typically consist of noun phrase (NP), adjective phrases (ADJP) and, sometimes though only within relative clauses, in verb phrases (VP). In some cases *Criteria* can be expressed in relation to *Alternatives* as in example 1, while in other cases they can be absolute as in example 2.

Ex. 1 *just* ***plastic*** *<Alternative> because that's always* ***the lightest*** *<Criteria>*

Ex. 2 *It needs* ***to be trendy*** *<Criteria>*

The distinction between *Alternatives* and *Criteria* is not always transparent and there may be cases of ambiguity, as in 3.

Ex. 3 *And so for this product it's gonna be* ***television only***

It can be seen from Example 3 that without the contextual information, it is not clear if the phrase "television only" can be considered as an alternative, e.g., a choice between DVR+television or television only, or an expression of the preference of the speaker. In these cases, annotators have been instructed to take into consideration the context, and mark the chunk according to the context. If context did not resolve the ambiguity, they would mark the chunk as *ambiguous*.

Given these features we developed the following annotation scheme: The annotators were given the following definitions: Alternative as being when *the speaker expresses what he/she could do*, and Criteria as being when *the speaker expresses how he/she evaluates*. Then, they were provided with a set of rules:

1. check if the chunk corresponds to the definition of *Alternatives* or *Criteria*
2. check if the chunk is NP, ADJP or VP
3. if VP check if it is a relative clause:
   → if yes: annotate
   → if no: not annotate
4. if in doubt between *Alternatives* or *Criteria* , check the context
   → clarified=yes; stop;
   → clarified=no; mark as ambiguous;

# Meetings:	43
# Tokens:	234,607
# Sentences:	22,903
# Utterances:	1,193
# Median of tokens per utterance:	65

Table 1: AMI dataset general statistics.

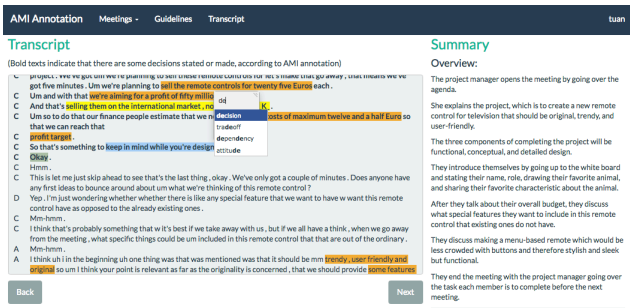

Figure 2: Domain-expert annotation interface

5. Hybrid Annotation Process

While previous efforts in decision annotation label entire utterances or sentences (Fernández et al., 2008; Somasundaran et al., 2007), we consider labeling arbitrary sequences of words. This approach produces more fine-grained annotations. This is necessary since our aim is to annotate the chunks related to the decision process rather than the segments of meetings where decisions are taken. In fact, this choice adds to the difficulty of the labelling task.

We employ a *hybrid* annotation process that exploits both domain expert-knowledge and crowdsourcing. We first develop an in-house annotation system for domain experts to annotate a subset of the data (**Phase 1**). Then, we design crowdsourced annotation tasks for *Alternatives* and *Criteria* for the entire dataset, using domain-expert annotations as quality control (**Phase 2**). Finally, annotations made by the crowd are reviewed by a domain expert (**Phase 3**). The role of the first phase is to generate a sufficient number of gold standard annotations for quality check during the second phase.

Data pre-processing. We started with the segmentation of conversations. Segmenting long documents into smaller chunks is crucial in annotation tasks, especially in a crowdsourcing setup, because crowdsourced workers are normally engaged with a sentence or small paragraph (Sabou et al., 2014). On the other hand, the decision annotation is sensitive to anaphora in the document, and to context in general, so the segmentation can result in "breaking the context", or cross-chunk anaphora, thus reducing the quality of annotations. To address this issue, we relied on the following heuristics: we set an arbitrary maximum length for each segment (1500 tokens), but keep utterances of correlated dialogue acts (i.e. a response) even if it goes beyond the maximum length limit of a segment.

Phase 1. We randomly selected 75 segments of transcripts for domain-expert annotation. Three domain experts annotated the segments for *Alternatives* and *Criteria* resulting in 3081 annotated chunks in total. The inter-annotator agreement using Fleiss' Kappa was 0.41 (Fleiss, 1971). Those annotations were used for quality control in Phase 2. Fig. 2 shows the interface developed for the domain expert annotation.

Phase 2. We used the crowdsourced annotation platform CrowdFlower[1] (CF). CF is an online platform that allows to hire non-expert annotators for specific tasks. The annotators are called contributors and can be chosen according to their native language or country of residence. In order to avoid malicious annotations, i.e., spammers that can degrade the quality of annotation (Raykar and Yu, 2011), contributors need to pass a test and are continuously evaluated during the task via *test questions*, questions with known answers hidden among actual questions. While setting the task, one needs to upload sufficiently many test questions to check the ongoing quality of the contributor annotations. As mentioned, this was the purpose of Phase 1.

We designed two separate tasks, one for *Alternatives* and one for *Criteria* annotation. This is justified by studies which report that separation reduces the crowdworkers' cognitive load and enhances both engagement and response quality (Bontcheva et al., 2014). The contributors were presented with conversations in natural language format, and were asked to freely highlight any phrases that were considered to be *Alternatives* and *Criteria* following the guidelines of Section 4.

Annotators in CrowdFlower are ranked according to a score that maps their experience as CrowdFlower annotators (level 1 being expert and level 3 beginners). We selected annotators with minimum level 2 experience, from English speaking countries. We had at least 3 annotators annotating each segment.

As a quality control setting, we forced any contributor to stay at least 10 seconds on each segment. In addition, we required a constant accuracy of 70% on the test questions throughout the task. Any annotator failing to maintain this accuracy was excluded from the task. We selected the test questions from the annotations obtained in Phase 1 by only taking the Phase 1 annotations with perfect agreement among the three annotators. With this filter, we overcome the low inter-annotator agreement (IAA) issue of Phase 1. At the end of this phase, we obtained 558 responses for *Alternatives* and 749 responses for *Criteria*, with a IAA of 0.55 Fleiss' Kappa for *Alternatives* annotation task and 0.31 for *Criteria* annotation task.[2]

Phase 3. Finally, to improve the quality of the annotations obtained, we had a domain expert revising all the annotations from Phase 2. The expert carried out three tasks: (i) confirming the correct annotations, (ii) rejecting incorrect annotations, and (iii) redefining phrase boundaries where needed. Altogether, in Phase 3, 335 chunks (1651 tokens) are confirmed as *Alternatives* and 249 chunks (824 tokens) are confirmed as *Criteria*.

These steps allow for a double level of quality control in the sense that only good quality annotations are retained in Phase 2, and they are further confirmed in Phase 3. This process allows us to exploit the crowd and reduce the time and the cost of the domain expert consultation.

[1] www.crowdflower.com

[2] While those are fair agreement, it is debatable to assess crowdsourced annotations using standard IAA scores. Perreault and Leigh (1989) among others consider the κ values (Fleiss and Cohen's) to be conservative in the case of crowd annotation as the contributors might have different level of expertise on the task (some might be linguists, some might be decision analysts). However, it is not possible to account for the level of the expertise, applying some weighted IAA score, because this information is not available.

6. Evaluation

In this section, we carry out standard experimental setup to evaluate the effectiveness of the created dataset. The objective of the experiments is to investigate whether the curated dataset is more *consistent* in terms of experimental observations than the non-curated one.

We treat the problem of decision extraction from the spoken transcript as a supervised sequence prediction task. To simplify the supervised prediction task, we treat each decision element as a separate task. For each classification type, i.e. alternatives and criteria, each word is associated with a binary label which is indicative of whether the word is a decision element or not.

In our experimental setup, with given training sequences of words and their associated labels, we build supervised models to predict the labels for unknown sequences. The supervised models that we use for our experiment are standard ones - namely the maximum entropy (MaxEnt) (Berger et al., 1996) based classifier and the conditional random fields (CRF) (Lafferty et al., 2001). For both MaxEnt and CRF implementations, we use the Java API from the Stanford NLP toolkit.

To simplify our setup, we use the lexical features, i.e. character n-gram ($n = 2$ to 5) and word features. In addition, the CRF also uses Part-of-speech tags. The CRF classifier takes into account the context of the previous words while making predictions. For CRF, we set the window size to 2. Since MaxEnt is not a sequential model, the previous context of two word windows is provided as an additional input feature to it.

We run the sequence prediction experiments on the two sets of corpora: (i) the non-curated one obtained after Phase 2, denoted by C for Crowd in Table 2, and (ii) the curated one, obtained after Phase 3 i.e. after revision by the domain expert, denoted as H for Hybrid in Table 2.

We conduct supervised sequence prediction experiments on identical batches of each dataset (i.e. crowd and hybrid) with identical settings. Specifically we generate 1000 iterations of each prediction problem. For each iteration, we randomly select 80% of the data to train a supervised model and conduct testing on the remaining 20%.

Note that a higher effectiveness in the decision element prediction task does not necessarily imply that the labels are more consistent and meaningful. In contrast, it is expected that consistent labeling across the whole dataset would reduce train-test bias and produce more consistent observations of the evaluation metrics. Hence, to measure the consistency of the experiments, we compare the variability in the precision, recall and F-score values computed over the batches. Specifically, we look at variance but also at the lowest and highest values as measured through quantiles (2.5% on one side and 97.5% quantile on the other side of the distribution). We expect that the curated dataset will result in lower variability in the measured evaluation metrics for the decision element prediction task than its non-curated counterpart.

In this abstract, we only report the corpora comparison results for the *Criteria* prediction task using the Maximum entropy classifier. In this case, as shown on Table 2 we observe that there is indeed a significant different in variance between the crowd annotated dataset and the hybrid dataset for precision. While we also see reduced variance with the curated dataset for Recall and F-score, that difference is less significant than for Precision. Such patterns are not observed when looking at CRF classifier or looking at *Criteria*. Specifically, in all other cases, the Levene's test reveals no significant difference between the crowd and hybrid samples.

	Precision		Recall		F-score	
	C	H	C	H	C	H
2.5%	0.364	0.366	0.286	0.286	0.322	0.324
97.5%	0.465	0.460	0.389	0.386	0.418	0.416
Mean	0.413	0.411	0.337	0.334	0.371	0.368
Var.	$6.7e^{-4}$	$5.9e^{-4}$	$7.0e^{-4}$	$6.4e^{-4}$	$5.8e^{-4}$	$5.3e^{-4}$
p-value	0.031		0.147		0.117	

Table 2: Comparison between Crowd and Hybrid Datasets for the Classification of *Criteria* using the MaxEnt classifier. The p-value reported is associated with Levene's test of equal variance.

Entity	Classifier	Precision	Recall	F-score
Criteria	MaxEnt	0.411	0.334	0.368
Criteria	CRF	**0.784**	**0.705**	**0.742**
Alternatives	MaxEnt	0.487	**0.402**	0.440
Alternatives	CRF	**0.572**	0.394	**0.465**

Table 3: Average Performance Values for the *Alternatives* and *Criteria* classification tasks based on the hybrid corpus.

Finally, as a mean to establish a baseline for the decision analysis entity detection task, we report in Table 3 the performance of each model for each task based on the hybrid corpus (As we mentioned, performance is sensibly similar). Our experiments reveal that both models perform similarly for the *Alternatives* prediction task with F-score in $[0.440 - 0.468]$. However, for the prediction of *Criteria* the CRF model is strictly superior in all dimensions. Note, however, that the variances of the CRF models on the Criteria tasks are 3 to 4 times higher than those of the MaxEnt models.

7. Conclusions

In this paper, we described the construction of a dataset of extracted decision elements from transcripts of spoken conversations. We release the annotation of a subset of meetings of the AMI corpus together with this paper at `http://researcher.watson.ibm.com/researcher/view_group.php?id=8178`. We developed a specific annotation scheme together with domain experts. As part of future work, we plan to extend our annotation to other decision analysis elements such as expressions of constraints and trade-offs.

8. Bibliographical References

Banerjee, S., Rosé, C. P., and Rudnicky, A. I. (2005). The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *INTERACT*.

Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, March.

Bontcheva, K., Derczynski, L., and Roberts, I. (2014). Crowdsourcing named entity recognition and entity linking corpora.

Cadilhac, A., Asher, N., Benamara, F., Popescu, V., and Seck, M. (2012). Preference extraction from negotiation dialogues. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 211–216. IOS Press.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2005). The ami meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 28–39. Springer.

Chen, L., Rose, R. T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T. X., Tu, J., Huang, Z., Harper, M., Quek, F., Xiong, Y., McNeill, D., Tuttle, R., and Huang, T. (2006). Vace multimodal meeting corpus. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, pages 40–51, Berlin, Heidelberg. Springer-Verlag.

Fernández, R., Frampton, M., Ehlen, P., Purver, M., and Peters, S. (2008). Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 156–163. ACL.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Hsueh, P.-Y. and Moore, J. D. (2007). Automatic decision detection in meeting speech. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 168–179. Springer.

Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macas-guarasa, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C., and Wrede, B. (2004). The icsi meeting project: Resources and research. In *in Proc. of ICASSP 2004 Meeting Recognition Workshop*. Prentice Hall.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML '01*, pages 282–289.

Matheson, J. E. and Howard, R. A. (1977). An introduction to decision analysis. In Miller Howards, Matheson, editor, *Readings in Decision Analysis*, chapter 1, pages 9–43. Stanford Research Institute, California.

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S., Tyagi, A., Casas, J., Turmo, J., Cristoforetti, L., Tobia, F., Pnevmatikakis, A., Mylonakis, V., Talantzis, F., Burger, S., Stiefelhagen, R., Bernardin, K., and Rochet, C. (2007). The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language resources and evaluation*, 41(3):389–407, 01/2008.

Perreault, W. D. and Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. In *Journal of Marketing Research (JMR)*, volume 26, pages 135–148.

Raykar, V. C. and Yu, S. (2011). Ranking annotators for crowdsourced labeling tasks. In *Advances in Neural Information Processing Systems*, pages 1809–1817.

Rienks, R. and Verbree, D. (2005). Twente argument schema annotation manual v 0.99b. University of Twente.

Sabou, M., Bontcheva, K., Derczynski, L., and Scharl, A. (2014). Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866.

Somasundaran, S., Ruppenhofer, J., and Wiebe, J. (2007). Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6.

Strauß, P. and Minker, W. (2010). *Proactive Spoken Dialogue Interaction in Multi-Party Environments*. SpringerLink : Bücher. Springer US.

Whittaker, S., Laban, R., and Tucker, S., (2006). *Analysing Meeting Records: An Ethnographic Study and Technological Implications*, pages 101–113. Springer Berlin Heidelberg, Berlin, Heidelberg.