An Experimental Approach For Information Extraction in Multi-Party Dialogue Discourse

Pegah Alizadeh, Peggy Cellier, Thierry Charnois, Bruno Crémilleux, Albrecht Zimmermann

An Experimental Approach For Information Extraction in Multi-Party Dialogue Discourse

Pegah Alizadeh [1], Peggy Cellier [2], Thierry Charnois [3],
Bruno Crémilleux [1], and Albrecht Zimmermann [1]

(1) GREYC, UMR 6072, UNICAEN/CNRS/ENSICAEN, 14000 CAEN, FRANCE
(2) IRISA, université de Rennes1, 35000 Rennes, FRANCE
(3) LIPN-UMR CNRS 7030, PRES Sorbonne Paris-cité, FRANCE
{pegah.alizadeh,bruno.cremilleux,albrecht.zimmermann}@unicaen.fr
peggy.cellier@irisa.fr thierry.charnois@lipn.univ-paris13.fr

Abstract. In this paper, we address the task of information extraction for transcript of meetings. Meeting documents are not usually well structured and are lacking formatting and punctuations. In addition, the information are distributed over multiple sentences. We experimentally investigate the usefulness of numerical statistics and topic modeling methods on a real data set containing multi-part dialogue texts. Such information extraction can be used for different tasks, of which we consider two: contrasting *thematically related but distinct* meetings from each other, and contrasting meetings involving *the same participants* from those involving other. In addition to demonstrating the difference between counting and topic modeling results, we also evaluate our experiments with respect to the gold standards provided for the data set.

Keywords: Information Extraction, Dialogue Texts, Topic Modeling, Term Weighting

1 Introduction

Many organizations (and people) spend large amount of their professional time on (or in) meetings. When the meetings are finished, they should be analyzed w.r.t. different aspects such as preparing meeting summaries, lists of envisioned projects, decisions, problems, action points etc. Preparing an automated approach for analyzing meetings would therefore be expected to be a boon for both meeting participants and non-meeting actors, e.g. managers, auditors etc. For the sake of capturing as rich a data set as possible, meetings can be recorded and stored in audio and/or video form. Thanks to new technologies such as Speech-to-Text[1], all registered dialogue during the meetings can be transformed into textual transcripts. In general, speech recognition or natural language processing on dialogue based conversations is difficult because they consist of multi-part

[1] Tool example: http://www.vocapia.com

interactions with extreme variability. On the other hand, the transcript of meetings – *output* of speech recognition methods – usually includes unstructured word streams, weak punctuation, formatting or capitalizations.

In this paper we experimentally evaluate how to extract information and topics from a meeting based corpus, namely AMI [2, 1] (See Section 3). We consider this kind of information as the first stepping stone towards summarizing a meeting. It is not obvious which approaches are well-suited to this question but the research literature on extracting representative terms is vast. We chose to evaluate representatives of two different paradigms: a counting statistic, *tf-idf* often used in information retrieval and event detection, and a topic modeling approach, *NMF*.

Term extraction from meetings is obviously not a means to itself but a stepping stone towards more complex tasks. The questions we ask are therefore:

1. Can we identify the terms specific for an individual meeting in comparison to *thematically related but distinct* meetings?
2. Can we identify which meetings should be grouped together, e.g. by identifying common names of participants?
3. To what degree do extracted sets of words agree with the provided gold-standard summarizations?

We contribute to answer these questions by showing that tf-idf approaches perform well on some tasks (e.g. meeting characterisations and speakers identification) within a meeting acted out by a single group speaking about the same scenario. Whereas topic modeling is better for summarization and theme extraction within a set of thematically related meetings acted out by different groups of participants. In the other words, the two approaches complement each other.

In the following section, we give an overview of the related work, and in Section 3 we describe the corpus used in the paper. We describe and define the approaches we have evaluated in Section 4. Section 5 describes the experimental evaluation and discusses the results for the three settings described above in detail. In the last section, we conclude and outline perspectives.

2 Background and Related Work

In this paper we are interested in spoken multi-party human meetings and our goal is to extract information such as as summary, important events, decisions taken during the meeting, identified problems, proposed solutions and decided action points, etc. We are also interested in knowing which topics where discussed in a particular meeting as well as having making the discussion of particular topics accessible.

A meeting analyzer system called CALO has been proposed in [19, 18] that first transcribes meeting speech to text automatically, and then analyses the meeting in several phases: dialogue act segmentation and tagging, topic segmentation and identification, problem and decision detection, and summarization. With respect to this paper's scope, they identify topic and segment meetings

using a generative topic model derived from Latent Dirichlet Allocation (LDA) method to learn the topic model automatically [13]. Using an unsupervised learning approach, they produce the meeting segmentations simultaneously to learn *what* (meeting topic) and *when* (meeting segmentation) people talk about during meetings. An alternative approach for segmenting meetings consists of tracking changes in lexical distributions: [5] and [6] translate the lexical distributions into a discriminative classifier and apply it on meeting transcripts.

Action items and decision extraction from meetings is another concern of this paper. A structural approach for detecting items and decisions has been proposed in [18, 19]. They classify meeting utterances w.r.t their roles in the process: task definition, agreement, and acceptance of responsibility. Next, they detect actions [12] and decisions [4] from the role patterns. Another approach in the literature extracts important words (related to problems) using classifiers or sequence models using a lexical approach [3].

Tur *et al.* [19, 18] extract different shortened version of meetings according to different evaluation measures. They introduce a method that computes *oracles* of summaries and selects the one with maximum performance according to "ROUGE" [16].

An alternative view consists of considering concepts such as decisions, problem identifications, and dialogue acts as *events* occurring during a meeting. Event detection is a well-established [9] and active [7] research field, which in recent years has focused on social media platforms, particularly Twitter [8, 17, 14, 20, 10]. Compared to meeting dialogue, Twitter is an interesting case because it shares characteristics with it, such as weak punctuation or non-standard use of words. Topic modeling has been used for event detection [14, 10]. Yet other approaches focus on statistical properties of the terms that occur in different documents, identifying "bursty" terms [8, 20], i.e. terms that during a given period occur much more often than before or after, and/or clustering them [17] to identify coherent topics.

3 AMI in Focus

The AMI meeting corpus contains 100 meeting hours captured using many synchronised recording devices[2]. All meeting participants, both native or non-native speakers, speak in English. In total, there are 171 meetings divided into two groups: **scenario-based meeting** and **non-scenario based meeting**. A scenario-based meeting simulates the discussions of team of four participants that are developing a remote control from start to prototype. For each simulation, the design process is divided into four parts which are held during a single day [2]. The "non-scenario based" meetings include real recorded meetings about designing various systems, as well as a small subset of scenarios that are different from the remote control one.

The AMI documents have been transcribed orthographically correct (while maintaining contractions) with annotated subsets of meetings including name

[2] Available here: `http://groups.inf.ed.ac.uk/ami/download/`

entities, dialogues acts, abstractive and extractive summaries [2]. Abstractive summaries are texts of about 200 words for each meeting, consisting of free text giving a general abstract of the meeting plus specific explanations of the decisions, problems or issues encountered during the meeting. Extractive ones identify the parts of meeting related to the abstractive summary. Note that each sentence in an abstractive summary can be referred to several dialogue acts in the corresponding extractive summary, and each dialogue act can refer to several sentences in the abstractive summary.

4 Extracting Information from Meeting Transcripts

The aim of the paper is to extract information that can be used to derive a meeting summary of a meeting including several participants. We are interested in an approach allowing us to extract the "important" words from a document, which can then be used in further steps, such as summarizing the meeting.

We calculate a weight for each word in a meeting document, where the weight expresses the relative word importance for the meeting with respect to a set of comparison meeting documents. Our hypothesis is that in any meeting words with higher weights are more likely to provide relevant information about and characteristics of the meeting. In this section, we present the two approaches we have evaluated: Term Frequency-Inverse Document Frequency (*tf-idf*), a statistical measure of a word's relative importance, as a reference method, and Non-negative Matrix Factorization, a topic modeling approach that groups words into sets and assigns weights (probabilities) to each topic.

4.1 Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (*tf-idf*) is a measure that gives a higher weight to a term (a word) that appears frequently in a particular document but infrequently in the entire corpus. The term frequency ($\mathrm{tf}(w, d)$) is simply the number of times that a word w appears in a document d. Inverse document frequency (idf) is the logarithmically-scaled proportion of documents in the corpus in which the word w appears. More formally:

$$\mathrm{idf}(w, D) = \log \frac{N}{1 + |d \in D \; s.t. \; w \in d|} \tag{1}$$

where N is the total number of documents in corpus and the denominator is the (laplace-corrected) number of documents d in which a word w appears in corpus D. Thus, the *tf-idf* is:

$$\text{tf-idf}(w, d, D) = \mathrm{tf}(w, d)\mathrm{idf}(w, d)$$

As an example for meeting documents, a stop-word such as "and" may appear often in a document, but not have a high *tf-idf* value because it is repeated in almost every meeting. However, a word such as "lunch" has a high weight,

because it appears frequently in a particular document but less often in the entire corpus. *tf-idf* is, for instance, the measure used to quantify terms' importance in [17], and a slight modification is used in [8].

4.2 Topic Modeling: Non-negative Matrix Factorization

There exist various topic modeling methods allowing us to identify the topics present in text documents. Among the considerable number of proposals for probabilistic methods such as Latent Dirichlet Allocation (LDA) or non-probabilistic methods such as Non-negative Matrix Factorization (*NMF*), we focus on the *NMF* in this paper [11]. Indeed, preliminary experiments (not given in this paper) showed that algebraic approaches perform better than probabilistic methods on our meeting documents.

In an *NMF* context, we assume that the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ represents m unique words in a corpus of n meetings (documents). The goal is to decompose the $\mathbf{A}$ matrix into two non-negative matrices such that $\mathbf{A} \sim \mathbf{WH}$. The columns in $\mathbf{W} \in \mathbb{R}^{m \times k}$ are the topics and the rows are the words represented in the topics. Each item in matrix identifies a non-negative weight for a word in relation to a topic. The matrix $\mathbf{H} \in \mathbb{R}^{k \times n}$ indicates to what degree a meeting is related to the k of topics.

5 Preliminary Experiments with AMI Corpus

To evaluate the effectiveness of the two methods in treating meeting data, we perform several different experiments. In a first experiment, we assess whether *tf-idf* can be used to identify the characteristics of particular meetings when compared to other meetings in the same context. In a second experiment, we evaluate whether *tf-idf* and *NMF* can be used to help in *grouping* meetings. Finally, we assess the agreement of extracted sets of words with the two types of existing summaries (extractive and abstractive) for the meetings in the AMI corpus.

5.1 AMI Corpus Preparation

We evaluate the effectiveness of our proposed techniques on the AMI data set [1, 2] containing multi-party dialogues[3]. The AMI data set delivers its meeting documents in different divisions such as divided meeting in terms of words, time, speakers and etc. We generated the plain text files manually using these segmentations and divisions which are available online[4].

To study the scenario-based meetings, we first classify meeting files w.r.t their subject. According to [2], each 3 to 5 meetings documents are acted out by a single group speaking about the same scenario, divided into several parts.

[3] `http://groups.inf.ed.ac.uk/ami/download/`
[4] `https://github.com/pegahani/AMI-prep`

For this reason, we regroup the 138 scenario-based meetings into 34 "effective" meeting blocks. On the other hand, there exist 33 non-scenario based meetings as well that will be studied in our future works.

5.2 Characterizing individual meetings

As we have described before, the underlying motivation for this work is to help summarize meetings, in particular identifying decisions taken, action items defined, problems identified, and responsibilities assigned. The scenario-based AMI meetings are intended to simulate the trajectory of a particular project (designing a remote control) from start to finish, condensed into a single day. All 34 groups acting out this scenario follow the same script but will, of course, vary in the concrete implementation of the script. However, if there is something inherent in each stage of the project, i.e. each individual meeting of the same block, we would expect those characteristics to show up in the terms extracted by *tf-idf*.

To evaluate our hypothesis, we use the following experimental setup: we treat a given block of meetings, i.e. meetings acted out by the same group, as the corpus and identify for each individual meeting in this block the twenty highest-scoring words according to *tf-idf*. For each of these words, we then count how often it occurred in the set of words derived from the same stage, e.g. we gather the sets of words derived from the first meeting of each block (M_1) and count duplicates. Table 1 shows the results for each stage, after thresholding word frequency at 4, i.e. 10%, rounded up[5]. We remind the reader that we did not use stemming for those texts. The number at the beginning of each cell therefore reports the modulo stemming, e.g. "cats" counts towards "cat", the number in parentheses the actual count of the word. There are 34 different groups that acted out the scenarios, which means that, for instance, the term "animal" was used in 73.5% of all first meetings of this scenario.

Table 1 (including 10% of twenty highest scoring words for the whole 34 meeting blocks) demonstrates most of the words are appeared in the first stage or the last one while the two stages in the middle have a small share. For instance in M_1 stage, "animal", "cat" and "favourite" have repeated 29, 19 and 12 times respectively. But in the M_2 stage less words appear, for instance "age", "lunch" and "teletext" have appeared each for 9 times. We see that the first and last meetings of a block have more related words than the second and third meetings. This shows that the script aligns the different groups quite closely in the beginning and in the end, but that they diverge in the middle two meetings. The first meeting seems always to be a brain-storm related to animals, and the fourth meeting about how to evaluate the success of the project. In the case of the third meeting, we can speculate that the discussion turned around what materials to use, and the presence of "lunch" as a highly-ranked word in the second meeting reminds us that this meeting took place at the end of the morning.

[5] For full results, we refer the readers to: `https://github.com/pegahani/Event_detection/blob/master/result/result_4_4.txt`

M_1	M_2	M_3	M_4
25: animal	9: age, lunch, teletext	8: solar, wood	17: criteria
19: cat (15)	8: percent, young (4)	6: titanium	13: seven
12: favourite	6: pay	5: spongy, concepts	12: evaluation
11: tool (5), dog (7)	5: settings, zap, seventy, users	4: dark, sample, doublecurved, materials, sensor, banana, circuit, cases, vegetables, fruit	9: sample
7: training, draw	4: set, messages, group, mode, infrared, recognition, speech		8: special, false
6: rabbit, profit, fish			6: evaluate, process
4: friendly, bird, tail, whiteboard, width, characteristics, elephant, morning			5: prototype, leadership, creativity, under
			4: scale, single, fifteen, team, average, budget, curve

Table 1. The most often repeated high-scoring words (according to *tf-idf*) for individual meetings in scenario-based blocks

5.3 Grouping meetings

Earlier, we have contrasted individual meetings against other meetings within a particular process. In the real world it's not clear how to group meetings, however. In the best case scenario, whoever starts the meeting recording would indicate what project or issue the meeting is related to but we cannot rely on this information being available. A slightly weaker assumption relies on participants stating their names. For example, if *Ada, Billy, Christine* and *Dolores* work together on one project, and *Elise, Frank, Gerd* and *Heather* on a second one, we would assume that certain combinations of names identify *groups* of meetings belonging together.

We therefore change the underlying document for calculating the *tf-idf* score, merging all meetings of a single block into one document, and treating the other merged blocks as comparison documents. The second column from the left of Table 2 shows examples of what happens when we contrast the 3-5 meetings enacted by a particular group against *all other meetings* in the corpus and as we expected, several of the extracted terms are names[6]. When contrasting individual sessions in a block against the other ones in the same block, as in the preceding section, names weren't highly ranked, which is to be expected given that the entire scenario is acted out by the same group of participants.

This indicates that one can contrast meetings against the entire available corpus and group them with other meetings showing the same combination of names, if they have not been grouped together by the person in charge of recording. The results need improvement, however: while we know that each scenario was acted out by four people, the average number of names recovered by *tf-idf* is only 2.44, with a standard deviation of 1.44. Main contributors to this standard deviation are blocks 25, 27, and 31, neither of which gave rise to a single name. It is possible that participants in those blocks did not state their names (often),

[6] For full results, we refer the readers to: `https://github.com/pegahani/Event_detection/blob/master/result/result_4_block_scen.txt`

meeting block	Extracted words using *tf-idf*	Extracted words using Topic Modeling (*NMF*)
meeting block 1	**matthew**, **mael**, **anna**, exce, ip, doctor, decline, assemble, streamed, customizing, asian, voter, undes, **nanne**, highperformance, fik, protec, underlie, provin, zebras	keys, **matthew**, browse, innovation, functionalities, v_c_r_, **mael**, **anna**, sixteen, perfect, demographic, r_c_, receiver, surf, present blinking, cents, store, movie, presented
meeting block 2	incremental, linear, orangutan, barks, squarey, lightening, caramel, computation, parameter, shocks, selled, multidevice, dommage, lawnmower, chec, discus, orangutans, fulf, buck, **olivier**	access, management, incremental, whistle, participant, receive , pear, ami, advantage, ergonomic, define, commands, fulfil, robust, technological, command, cent, task, continue, financial
meeting block 3	**pedro**, midmarket, backlit, crossover, exclusivity, trainable, silvers, **pedros**, uniqueness, paperwork, scheduling, offhand, ditching, consciousness, axes, marketability, ballpark, twelvefifty, synergy, advantageous	cradle, create, niche, sorta, **pedro**, locator, unique, terms, identified, flip, televisions, nah, environmentally, management, mode,shell, marketable, lapel, ta, perspective
meeting block 4	mushroom, **jordan**, coarse, baba, alimentation, mush, gestures, **kemy**, institute, **frahan**, **florent**, laser, longmund, sleeping, ada, saucer, trois, ecological, hmmm, eatable	controller, mushroom, gesture, google, pineapple, powerful, **david**, base, wireless, traditional, lemon, **jordan**, wooden, sophisticated, wire, vocal, participant, ball, bulb, recognise
meeting block 5	turbo, **mando**, panther, indian, spiders, spider, trunk, capsicum, outlier, pepper, kitsch, granularity, playback, spotting, templates, ons, hunt, stylised, epinions, permanantly	station, handy, turbo, **mando**, basis, base, technologies, targeting, wheels, coffee, elephant, sitting, elephants, phrase, milan, instance, morning, crazy, traditional, recognize

Table 2. Example of words extracted using *tf-idf*t (left) and *NMF* (right) for the first five scenario-base meeting blocks

a situation that might arise in real-world situations, especially if a certain group of people has already been working together for a while.

In this context, we would also like to point out an interesting observation in our experiments: for the block grouping $M_{52}, M_{53}, M_{54}, M_{55}$, one of the highest-ranking words is "shoulda". This is a non-formal contraction of "should have" and the fact that it is tagged as having one of the 20th highest scores for that block (actually the 8th-highest) indicates that certain participants' *manner of speech* might be enough to group meetings. Other contractions that we have extracted include "you'll" and "ain't". We intend to explore this phenomenon in the future.

As a comparison approach to using *tf-idf* on merged meetings, we also evaluated *NMF*. We use *NMF* to assign a topic to each scenario-based meeting block. As has been mentioned earlier in this section, there are 34 scenario based meeting blocks in total. For this reason we use *NMF* topic modeling with exactly 34 topics to classify the scenario-based meeting blocks, i.e. each block of meetings is assigned to exactly one topic.

To implement the *NMF* method, we used the gensim package [15]. After classification, each meeting should belong to a topic class. As for *tf-idf*, each topic is represented by the 20 most important candidate words according to the induced model. Referring to Table 2, the extracted words for five meeting blocks 1 to 5 is given in the last column[7]. While we can recover names using *tf-idf*, this is not reliably the case when using *NMF* – the average number of recovered names is 0.88, with a standard deviation of 0.91. This means that the information derived the *tf-idf* score and the topic model are complementary. Interestingly enough, when looking at meeting block 4 in Table 2, we find "david"

[7] For more results you can visit: `https://github.com/pegahani/Event_detection/blob/master/result/Topic_modeling_nmf_block_34_topics.txt`

and "jordan" in the *NMF* results, who refer to the *same* person in the acting group.

5.4 Quality of extracted sets of words

In order to evaluate the extracted information of the meetings, we compare them with the abstractive and extractive summaries as follows.

Given a set of words S extracted from a meeting M, we define the accuracy S w.r.t the abstractive summary (extractive summary) of M, $s_{abs}(M)$ ($s_{ext}(M)$), as:

$$acc_{abs}(S, M) = \frac{|S \cap s_{abs}(M)|}{|S|} \; (acc_{ext}(S, M) = \frac{|S \cap s_{ext}(M)|}{|S|}), \text{ i.e.,}$$

number of words of S that appear in the abstractive (extractive) summary divided by the number of words in S. We match words modulo stemming, for instance, if S contains "painting" and s_{abs}/s_{ext} a similar word such as "paint", the word is considered to have appeared.

We assess the quality of the extracted sets of words in two ways: Figures 1 and 2 show how sets of words extracted as described in Section 5.2 perform w.r.t. to abstractive and extractive summaries.

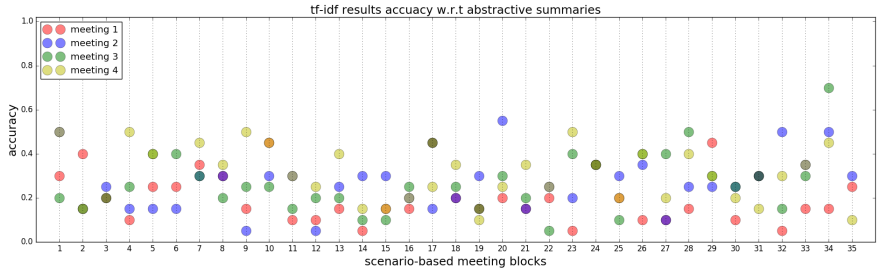

Fig. 1. Abstractive accuracy for sets of words extracted from individual meetings, compared to the other meetings in the same block, using *tf-idf*.

In overall, abstractive accuracy results are always less than 0.6 but in many cases they vary between 0.0 to 0.4. It is noticeable that for almost half of the blocks, the abstractive accuracy for the first meeting is lowest. As we saw in Table 1, the brain-storming in the first meeting involved concrete animal names that are not necessarily presented in the abstractive meeting.

Most extractive accuracies are above 0.6 and noticeable cases have the accuracy more than 0.8. As expected extractive accuracy is higher than the abstractive one. This is because our accuracy measure computes the percentage of appearing words in each summary. Since the extractive summary includes

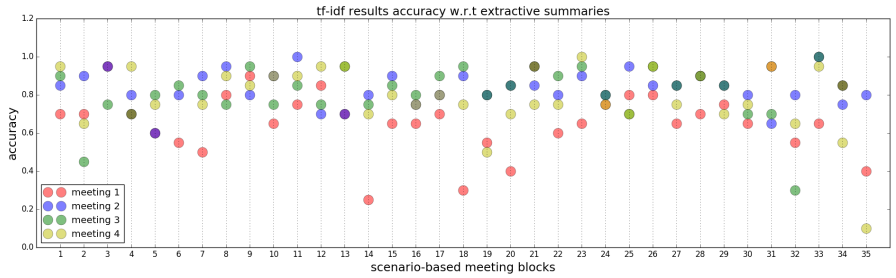

Fig. 2. Extractive accuracy for sets of words extracted from individual meetings, compared to the other meetings in the same block, using *tf-idf*.

more words as a subset of the meeting documents, the accuracy value for extractive summaries is far higher than the abstractive one. In addition, abstractive summaries by definition *abstract* from the actual contents of the meeting.

The second assessment uses the sets of words extracted according to the description in Section 5.3, and accuracies are calculated w.r.t. the *merged* summaries of all meetings in a block.

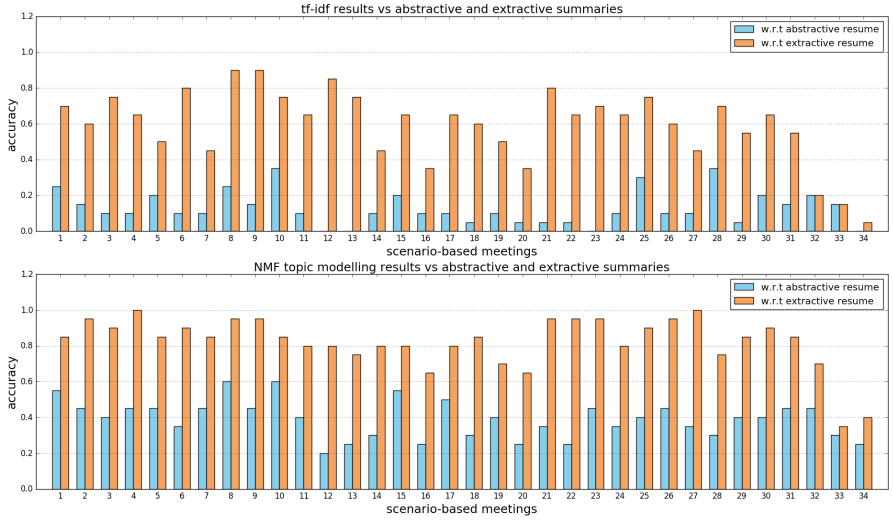

Fig. 3. The upper graph (lower graph) indicates the abstractive and extractive accuracies for sets of words extracted from scenario-based meeting blocks using *tf-idf* (*NMF*).

The upper graph in Figure 3 reports the accuracies of sets of words extracted from each meeting block using *tf-idf*. Similar to the preceding results, extractive

accuracy is higher than the abstractive one. According to the graph, the extractive accuracy is more than 0.6 while the abstractive accuracy is less than 0.4 for all the scenario-based meeting blocks.

The lower graph in the figure shows abstractive and extractive accuracies for sets of words extracted by *NMF*. By comparing this graph with the *tf-idf* results, we see that *NMF* gives better results for both summary types. This is the flip side of *NMF*'s inability to pick out names – instead of picking up particularities of the group – *NMF* learns topics describing the meeting block, matching summaries better.

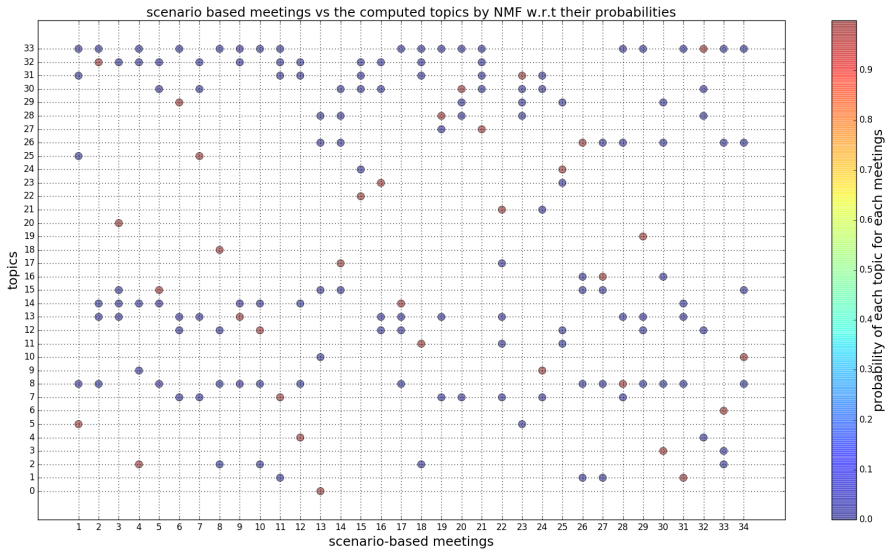

Fig. 4. The x axis indexes the 34 scenario-based meeting blocks while the Y axis indicates the list of 34 topics. The figure shows the 5 highest-ranked topics for each meeting block. The heat map on the right hand side maps probabilities to the colors used in the plot.

Since topic modeling methods compute the probability with which each text is related to each topic, Figure 4 shows for each meeting block how probable it is that it concerns a particular topics. The color bar in the right side of the figure shows the probabilities i.e. a point with a color closer to red is more highly probable. For instance, meeting 1 concerns topic 5 with a probability of more than 0.9 while, it is about topics $8, 25, 31$ and 3 with a very low probability (less than 0.1). It shows that there are no ambiguities – each meeting block has one topic assigned to it with very high probability (in excess of 0.9), and all others are at less than 0.1 probability.

6 Conclusions and Perspectives

In this article, we tested representatives of two paradigms for term extraction from documents – numerical statistics and topic modeling – to extract important information from texts in the context of recorded and transcribed multiparty dialogues, more specifically the AMI data set. We evaluated three settings: characterizing individual meetings in comparison to other thematically related meetings, identifying names of participants tying together related meetings, and matching extracted sets of words to provided gold-standard summaries.

In terms of the first setting, *tf-idf* performs rather well, consistently characterizing first and last meetings of a scenario, and showing good extractive accuracies (and to a lesser degree abstractive ones), i.e. comparisons of the extracted sets to gold-standard summaries. When it comes to identifying names, *tf-idf* shows better performance than *NMF* but they are arguably not good enough to reliably group meetings, explicit tagging before recording is therefore probably required.

Topic modeling via *NMF*, on the other hand shows better performance w.r.t the provided gold standard (abstractive and extractive summaries) when we attempt to characterize a full scenario, i.e. a block of thematically related meetings acted out by one group.

A conclusion to be drawn from our results is therefore that the two approaches complement each other, providing results that are useful for different aspects of the larger task of characterizing and summarizing meetings. To continue in this direction, automatically grouping meetings seems to be the most important issue to improve on because working on related meetings is foundational for other tasks, i.e. characterizing different types of meetings and/or identifying trends that help in understanding the progression of a particular project. A second question is how to arrive at summaries from terms sets – the process of extracting sentences related to extracted terms, and translating those into abstractive summaries in turn, is not an obvious one but needs to be tackled.

Acknowledgement

This work is supported by the FUI 22 (REUs project) and the ANR (French Research National Agency) funded project NARECA ANR-13-CORD-0015.

Bibliography

[1] Augmented multi-party interaction. http://www.amiproject.org (2010). [Online]

[2] Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. Language Resources and Evaluation **41**, 181–190 (2007)

[3] Fernández, R., Frampton, M., Dowding, J., Adukuzhiyil, A., Ehlen, P., Peters, S.: Identifying relevant phrases to summarize decisions in spoken meetings. In: Proceedings of Interspeech08. Brisbane (2008)

[4] Fernández, R., Frampton, M., Ehlen, P., Purver, M., Peters, S.: Modelling and detecting decisions in multi-party dialogue. In: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, SIGdial '08, pp. 156–163. Association for Computational Linguistics, Stroudsburg, PA, USA (2008)

[5] Galley, M., McKeown, K., Fosler-Lussier, E., Jing, H.: Discourse segmentation of multi-party conversation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03. Association for Computational Linguistics, Stroudsburg, PA, USA (2003)

[6] Georgescul, M., Clark, A., Armstrong, S.: Exploiting structural meeting-specific features for topic segmentation. In: TALN/RECITAL, pp. 15–24. Toulouse (France) (2007)

[7] Gurin, Y., Szymanski, T., Keane, M.T.: Discovering news events that move markets. In: Intelligent Systems Conference 2017 (IntelliSys2017), London, United Kingdom, 7-8 September 2017 (2017)

[8] He, Q., Chang, K., Lim, E.P.: Analyzing feature trajectories for event detection. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, pp. 207–214. ACM, New York, NY, USA (2007)

[9] Kleinberg, J.M.: Bursty and hierarchical structure in streams. Data Min. Knowl. Discov. **7**(4), 373–397 (2003). DOI 10.1023/A:1024940629314. URL http://dx.doi.org/10.1023/A:1024940629314

[10] Lau, J.H., Collier, N., Baldwin, T.: On-line trend analysis with topic models:#twitter trends detection topic model online. Proceedings of COLING 2012 pp. 1519–1534 (2012)

[11] Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. Nature **401**, 788–791 (1999)

[12] Purver, M., Dowding, J., Niekrasz, J., Ehlen, P., Noorbaloochi, S., Peters, S.: Detecting and summarizing action items in multi-party dialogue. In: In Proc. of the 9th SIGdial Workshop on Discourse and Dialogue (2007)

[13] Purver, M., Griffiths, T.L., Körding, K.P., Tenenbaum, J.B.: Unsupervised topic modelling for multi-party spoken discourse. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-

44, pp. 17–24. Association for Computational Linguistics, Stroudsburg, PA, USA (2006)
[14] Ramage, D., Dumais, S.T., Liebling, D.J.: Characterizing microblogs with topic models. ICWSM **10**(1), 16 (2010)
[15] Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta, Malta (2010). `http://is.muni.cz/publication/884893/en`
[16] Riedhammer, K., Favre, B., Hakkani-Tür, D.: Packing the Meeting Summarization Knapsack. In: Interspeech, Brisbane (Australia). Unknown, Unknown or Invalid Region (2008). URL `https://hal-amu.archives-ouvertes.fr/hal-01194290`
[17] Sayyadi, H., Hurst, M., Maykov, A.: Event detection and tracking in social streams. In: In Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009). AAAI (2009)
[18] Tur, G., Stolcke, A., Voss, L., Dowding, J., Favre, B., Fernandez, R., Frampton, M., Frandsen, M., Frederickson, C., Graciarena, M., Hakkani-Tur, D., Kintzing, D., Leveque, K., Mason, S., Niekrasz, J., Peters, S., Purver, M., Riedhammer, K., Shriberg, E., Tien, J., Vergyri, D., Yang, F.: The CALO meeting speech recognition and understanding system. In: 2008 IEEE Spoken Language Technology Workshop, pp. 69–72 (2008)
[19] Tur, G., Stolcke, A., Voss, L., Peters, S., Hakkani-Tur, D., Dowding, J., Favre, B., Fernandez, R., Frampton, M., Frandsen, M., Frederickson, C., Graciarena, M., Kintzing, D., Leveque, K., Mason, S., Niekrasz, J., Purver, M., Riedhammer, K., Shriberg, E., Tien, J., Vergyri, D., Yang, F.: The calo meeting assistant system. IEEE Transactions on Audio, Speech, and Language Processing **18**, 1601–1611 (2010)
[20] Weng, J., Lee, B.S.: Event detection in twitter. ICWSM **11**, 401–408 (2011)