Speaker Role Recognition using question detection and characterization

Thierry Bazillon[1], *Benjamin Maza*[2], *Michael Rouvier*[2], *Frederic Bechet*[1], *Alexis Nasr*[1]

[1]Aix Marseille Université, LIF-CNRS, Marseille, France
[2]Université d'Avignon, LIA-CERI, Avignon, France
[1]firstname.lastname@lif.univ-mrs.fr [2]firstname.lastname@univ-avignon.fr

Abstract

Speech Data Mining is an area of research dedicated to characterizing audio streams that contain speech from one or more speakers, using descriptors related to the form and the content of the speech signal. Besides the word transcription, information about the type of audio stream and the role and identity of speakers is also crucial to allow complex queries such as: "seek debates on X", "find all the interviews of Y", etc. In this framework we present a study performed on broadcast conversations that focuses on the way speakers express their questions in conversations. The initial intuition is that the type of questions asked can help identify the role (anchor, guest, expert, etc.) of a speaker in a conversation. By tagging these questions with a set of labels and using this information in addition to the commonly used descriptors to classify users' role in broadcast conversations, we improve the role classification accuracy and validate our initial intuition.

Index Terms: Spoken Language Understanding, Speech Data Mining, Speaker Role classification, Question detection

1. Introduction

Speech data mining aims at characterizing a speech signal of one or several speakers, based on some descriptors of the shape and the content of the signal. The most important of such descriptors is, of course, the automatic word transcription of the utterance. When dealing with multi-speaker dialog, as can be found in radio or television broadcast news, several other descriptors based on other aspects of the dialog can be taken into account, such as the role or the identity of the speaker, the type of the programme: interview, debate, news ... Extracting and representing such information can help answer complex queries as: "look for debates on theme X", "find all interview of Mr Y". It can also help the automatic speech recognition process in selecting the proper language and acoustic models for the specific type of programme being processed.

In this general framework, the subtask of identifying the speakers *roles* in a conversation, has been the subject of many recent studies. Such identification is generally based on lexical choices performed by the different speakers, as well as on acoustic characteristics of the speech signal. We propose in this study to add a new type of clue for identifying the speakers roles, which is the way speakers formulate their questions in the conversation. The underlying intuition is that the type of a question partially reveals the role of the speaker who asked it. We will propose a hierarchical classification of questions types and add such information as new descriptors for identifying speakers roles.

2. Related work

The labeling of the speech signal with roles is done after speaker diarization and automatic transcription. The whole process is decomposed as follows: the speech stream is first segmented into *speech turns*, each of which corresponds to a speaker. Such turns are further segmented everytime a pause duration between two words exceeds a given threshold. The segments thus obtained are then processed by the automatic speech transcription system. Eventually, speech turns are labelled with roles selected from a list of the possible roles.

One of the first studies on such labeling was realized and published in year 2000 by [1]. Many other followed, among which [2, 8, 10] and [5]. Such studies differ with respect to several dimensions, among which the number of roles taken into account (from 3 to 6), the type of the programme (debate, interview, report ...) as well as the segmentation level used for evaluation: segments or speech turns. Several parameters are taken into account for labeling: acoustic and prosodic parameters [2]; lexical parameters [1, 8] or combinations of both [5]. Several levels of supervision can be considered when producing the parameters that will be used for labeling, ranging from complete supervision: manual segmentation and transcription, as in [10]; to no supervision at all, as in [5] where both segmentation and transcription were performed automatically. Our work contrasts with previous studies by introducing new parameters for characterizing speaker roles. As mentioned above, these new parameters concern the type of the questions asked by the speakers.

The task of automatic question detection in oral utterances has focused on conversational speech corpora [11] as well as on meeting recordings [3]. In both cases, they were performed on manual segmentation and transcription of the utterances. The task boils down to a binary classification of speech segments into interrogative and affirmative classes. Question labeling can be considered as a sub-task of a more general labeling of conversations in *dialog acts* which aims at segmenting discourse in discursive units such as: affirmation, question, confirmation, negation ... Several lists of dialog acts have been proposed such as *DAMSL* [4]. Such lists are mainly based on lexical, prosodic and syntactic clues. We propose, in this paper, to refine the binary classification by adding the 8 following types of questions: *adverb*, *complex*, *determiner*, *est-ce-que*, *inversion*, *pronoun*, *si* and *intonation*, that are defined in the next section.

3. A corpus annotated with speaker roles and question types

The current study has been carried out on a section of the EPAC corpus [6] which contains transcription and annotation for approximately one hundred hours of conversational speech. Inside

the EPAC corpus, *Le Téléphone sonne* radio programme represents about 20 hours of data, divided into 32 shows. This corpus has been manually segmented in speaker turns and transcribed.

Speakers role labels have been added to the annotations. These labels characterize each speaker according to his status and his position in the radio show: anchor, guest, consultant, special correspondent, etc. Identifying speakers roles is a crucial step to fully understand a radio or television programme. Depending on the show organization, the number and roles of each speaker may be different.

In *Le Téléphone sonne*, we identified 4 speaker roles:

1. the anchor: he is the one who leads the radio programme. He has to make sure it goes smoothly, while trying to make all his guests talk.
2. the experts: they are guests who may be inside the radio studio or talking by phone. They answer listeners' questions, but they also debate about the main topic of the programme.
3. the callers: they are selected before the radio programme and then called back in order to ask their questions. However, they don't really debate with the experts and the anchor.
4. the journalists: their role, in this show, is to select and read questions written by email by the listeners.

In addition to these role annotations, we have also annotated all the interrogative sentences found inside this corpus. Each question has been manually annotated with specific tags, depending on which interrogative marker it contains:

- interrogative pronouns (*qui*, *que*)
- interrogative adverbs (*quand*, *comment*, *pourquoi*)
- interrogative determiners (*quel*, *quelle*)
- complex structures (*qu'est-ce que*, *qu'est-ce qui*, *à qui*, *depuis quand...*)
- *est-ce que* marker
- *si* marker (*je voudrais savoir si vos intervenants sont d'accord?*)
- subject inversion
- intonation only (*tu viens ?*)

This categorization is comparable to the ones defined for the English language in [9], although some specificities of the French language have been taken into account. The *intonation only* questions refer to Stolcke's *declarative* ones, while *wh- questions* are in French the ones starting with an interrogative pronoun, adverb, determiner or complex structure. The *est-ce que* marker is very French-specific and has been treated apart, even though it is close to subject inversion as both introduce what is traditionnally named a "yes-no-question" or "close question" (*question totale* in French).

Table 1 presents the first results of speakers' roles annotation, mixed with questions' annotation. It shows that the anchor asks more than half of the questions contained in *Le Téléphone sonne*. This predominance is due to his status of *mediator* that we mentionned before, and which will be detailed more precisely in the next paragraph.

In order to get a deeper analysis, the table 2 shows, for each type of question found in our corpus, their distribution depending on speakers' roles. As we previously said the anchor asks most of the questions, which are mostly intonation-only. They

Role	Number of questions	Frequency (%)
Anchor	791	50.97
Expert	323	20.81
Caller	304	19.59
Journalist	134	8.63
TOTAL	1552	100

Table 1: Number of questions distribution depending on speakers' roles

Type (%)	Anchor	Expert	Caller	Jour.
into. only (587)	93.02	5.96	1.02	0
est-ce que (215)	28.84	34.42	33.49	3.25
adverb (224)	16.52	39.29	29.46	14.73
subject inv. (164)	32.32	10.97	20.12	36.59
complex struct. (134)	19.40	44.03	27.61	8.96
pronoun (89)	34.83	34.83	19.10	11.24
determiner (93)	36.56	18.28	32.26	12.90
si (46)	4.35	2.17	93.48	0

Table 2: Type of questions distribution depending on speakers' roles

are mainly used to distribute speech to his guests, and as a result they represent about 70% of all the questions related to the anchor role. In contrast, experts, which rank at the second place concerning the total amount of asked questions (table 1), use many different interrogative structures with close frequencies. Most of their questions are based on interrogative adverbs, interrogative pronouns, *est-ce que* marker and complex structures (*qu'est-ce que*, *qu'est-ce qui*, *duquel*, *lequel*, etc.). Intonation-based questions are poorly represented. The same observations can be made with callers, as they ask very few questions without any grammatical marker. On the opposite side, they usually use *est-ce que* and indirect structures (whatever the interrogative marker is).

4. Automatic question classification

Among the 8 different question types considered in this study and presented in the previous section, 7 of them are related to the syntactic surface form of the sentences expressing the questions. They will be refered as *syntactic questions*. The last one, the *intonation only* type, corresponds to a specific prosodic pattern and will be refered as *prosodic questions*. We have, on the 32 radio shows of our corpus, a total of 13,224 sentences among which 973 are considered as *syntactic questions* and 562 as *prosodic questions*. Two different processes have been developped for detecting and classifiying questions: one based on lexical and syntactic features; the other one based on prosodic features.

4.1. Classifiying syntactic questions

The goal of this classification process is first to label as interrogative/affirmative each sentence contained in the speakers turns. We use in this study the sentence segmentation labels given by the annotators during the manual transcription process. The use of an automatic sentence segmentation process is currently studied. For each sentence classified as a *syntactic question*, its type is guessed. The sentences are described by the set of all the sequences of one and two consecutive words (1-grams and 2-grams) occuring in them. In addition to words, the 1-grams and

2-grams on the corresponding Part-Of-Speech (POS) labels obtained with the tagger MACAON[1] are also added.

We use a classification method based on the Adaboost[2] algorithm that consists of a linear combination of weak classifiers. Each weak classifier corresponds to a 1-level decision tree on the occurence of a 1-gram or 2-gram on the words and the Part-Of-Speech labels. During the learning process, each iteration chooses the most discriminant weak classifier for performing the classification task. The weights of the wrongly labelled examples are updated after each iteration following the Adaboost algorithm. The number of iterations is determined on a development corpus.

Due to the limited size of our annotated corpus we have used a *Leave-One-Out* experimental setup at the radio show level in order to evaluate this syntactic questions classifier:

- the whole annotated corpus C contains 32 radio shows: $C = \{e_1, e_2, \ldots, e_{32}\}$;
- at each iteration i we use the show e_i as the test corpus T_i; the show e_{i+1} as the development corpus D_i and the remaining 30 shows $A_i = C - \{e_i, e_{i+1}\}$ as the training corpus A_i;
- a classifier B_i is trained on A_i; the number of boosting iterations is chosen on D_i and the sentences of T_i are labelled automatically by B_i and stored in T'_i ;
- after the 32 iterations, the corpus $C' = \bigcup_{i=1}^{32} T'_i$ contains the whole corpus C automatically labelled with question types.

The results on C' are presented in table 3 with the standard Precision (*P*), Recall (*R*) and F-measure (*F-mes*) metrics.

type	# sent	*P*	*R*	*F-mes.*
interrogative	*995*	94.2	85.1	89.4
affirmative	*12229*	98.8	99.6	99.2
adverb	*223*	96.1	87.9	91.8
complex	*139*	79.0	67.6	72.9
determiner	*99*	87.6	78.8	83.0
est-ce-que	*209*	96.7	97.6	97.1
sub. inversion	*159*	82.5	53.5	64.9
pronoun	*94*	80.9	58.5	67.9
si	*45*	83.3	66.7	74.1

Table 3: Automatic classification of sentences into question types

As we can see the *affirmative/interrogative* classification works well (89.4 F-mes), however there is some diversity in the classification performance at the question-type level. The *subject inversion* and *pronoun* questions are the most difficult to retrieve, with a recall lower than 60. This was expected, as detecting such questions would require a deeper parsing process than a simple POS tagging. We are currently adressing this issue. However, even with just words and POS features, the classification precision for all types of questions ranges between 79 and 96, therefore we consider that this classification process is robust enough to provide good features for our speaker role classifier.

[1] http://macaon.lif.univ-mrs.fr/

[2] The Icsiboost [7] implementation of Adaboost is used in the experiments

4.2. Classifiying prosodic questions

All the acoustic features used in this study are obtained on the F0 curve computed directly from the signal, with time-windows of 10ms. Other parameters can be extracted from this curve. We propose in this paper a set of 15 of them divided in 3 classes: statistic (6 parameters), trajectory (5 parameters) and shape (4 parameters). This is a short description for each of them:

Statistic is made of six parameters related to the fundamental frequency: minimum, maximum, range, mean, median and standard deviation, in time windows of 300 or 700ms.

Trajectory groups together five parameters that indicate whether the pitch is raising or falling. Traditionaly trajectory of the pitch was modeled by slope (computed simply using the begining and end points). Recently, additional parameters have been added: *raising sum*, *raising count*, *falling sum*, *falling count* and , *is raising?* (if *raising sum* > *falling sum*).

Shape consists in four parameters modelling the shape of the pitch through a polynomial interpolation of Lagrange. Different orders of polynomial were tested. Empirical experiments showed that best results are obtained with a polynomial order of 2. We extracted 3 parameters (a, b, c of the polynomial $ax^2 + bx + c$). The last parameter is the error interpolation of approximation function.

Once the features are extracted, for a given speech segment, the decision *question* / *non question* is taken by a classifier using all the acoustic features presented. We use as a classifier a Multi-Layer Perceptron (MLP). MLP is trained by the standard back-propagation algorithm. It is a 3 layer network with respectively 15, 17 and 2 networks. The 15 inputs networks represents the parameters calculated on F0. Each of the 2 outputs networks corresponds to the class : *question* and *non-question*.

The performances of the MLP on the EPAC corpus using a 5-fold cross-validation are reported in Table 4. The results show the impact of the combination of the various parameters in the classification results. As we can see, the *Combo* version that combined all the parameters achieves better results than each of them taken separately.

	Precision	Recall	F-Measure
Shape+Statistics	62	37	46
Shape+Trajectory	58	32	41
Statistics+Trajectory	58	33	42
Combo	58	41	48

Table 4: Combination of F0 features for the binary classification of speech segments as question/non question

The prosodic and the syntactic classifiers are combined in the following way: a segment is first presented to the syntactic classifier. If it is classified as interrogative then it is considered as a syntactic question; otherwise, if the prosodic classifier labels it as a question, it is considered as a prosodic question.

5. Speaker role labelling

The main goal of this study is to verify our intuition that the type of questions expressed by a speaker during a conversation partially reveals its role within a broadcast show. In order to validate this assumption, we train a classifier for labelling speaker

role	*turns*			*speakers*		
F-measure	*baseline*	*question(gold)*	*question(auto)*	*baseline*	*question(gold)*	*question(auto)*
caller	66.4	65.9	66.5	88.2	90.2	90.2
expert	73.7	74.8	73.8	83.9	83.6	83.3
anchor	81.2	82.5	81.2	66.7	77.7	71.7
journalist	37.2	56.7	52.4	45.7	68.3	57.9

Table 5: Results in speaker role labelling at the turn and speaker levels. Comparison with or without queston features (automatic or gold)

turns into one of the 4 speaker roles of our corpus (*anchor*, *expert*, *caller*, *journalist*).

The classifier we use for this task is the same boosting-based classifier presented in section 4 for classifiying questions. We use also the *Leave-One-Out* experimental setup at the radio show level presented in section 4. The classifier is applied at the speaker turn level. Three systems are compared:

- *baseline*: in this system the features describing each speaker turn are the word and POS 2-gram features of the transcriptions and the turn duration (in seconds);
- *question(gold)*: we add to this system, on top of the baseline features, the set of question type labels contained in the turn, according to the manual (gold) annotation of the transcriptions;
- *question(auto)*: we use in this system the automatically predicted question type labels, as described in section 4.

The results are given in table 5 according to two modalities: evaluation at the speaker turn level; evaluation at the speaker level. The first simply takes the classifier decision on the speaker roles for each turn of the test corpus. The standard F-measure metric is then computed for each speaker role. In the second modality, all turns of a given speaker are considered, the classifier is applied to each of them, the speaker is then labeled with the majority role label over all the decisions taken for all the turns. As we can see, adding question features in the classification process helps significantly the speaker role classification process at the speaker level. The improvement is not as significant at the turn level, except for the journalist role. This can be explained by the fact that most turns (except the journalist ones that always contain at least one question as explained in section 3) do not contain questions and our system is inadequate for labeling them with a speaker role. Using automatic labels instead of *gold* labels for questions degrades, as expected, the performance. However the results remain significanlty better than the baseline system for 3 over the 4 roles of our corpus.

6. Conclusion

We have shown in this paper that detecting and classifiying interrogative sentences in conversational speech can help characterizing the roles of the speakers in a broadcast show. Some interrogative patterns can even be a signature of a given role. By adding features relative to question types in our speaker role classifier we have significantly improved the classification accuracy on our broadcast conversation corpus, even with question type automatically labelled. However in this study we did not face the issue of sentence segmentation: we assumed that we had already a segmentation of each turn into "sentences" given by human annotators. We are currently working on methods that can segment and detect questions jointly in a stream of words.

7. Acknowledgment

This work is supported by the French agency ANR, Project DECODA, contract no 2009-CORD-005-01, and the French business clusters Cap Digital and SCS. For more information about the DECODA project, please visit the project home-page, http://decoda.univ-avignon.fr/

8. References

[1] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. The rules behind roles: Identifying speaker role in radio broadcasts. In *Proc. of AAAI*, 2000.

[2] B. Bigot, J. Pinquier, I. Ferrané, and R. André-Obrecht. Looking for relevant features for speaker role recognition. In *Proc. of Interspeech*, 2010.

[3] Kofi Boakye, Benoit Favre, and Dilek Hakkani-Tür. Any Questions? Automatic Question Detection in Meetings. In *ASRU, Merano (Italy)*, 2009.

[4] M. Core and J. Allen. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35. Citeseer, 1997.

[5] Géraldine Damnati and Delphine Charlet. Robust speaker turn role labeling of tv broadcast news shows. In *ICASSP'2011*, 2011.

[6] Yannick Estève, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas. The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news. In *LREC*, Malta, 2010.

[7] Benoit Favre, Dilek Hakkani-Tür, and Sebastien Cuendet. Icsiboost. http://code.google.come/p/icsiboost.

[8] B. Hutchinson, B. Zhang, and M. Ostendorf. Unsupervised broadcast conversation speaker role labeling. In *Proc. of ICASSP*, 2010.

[9] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C.V. Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[10] S. Yaman, D. Hakkani-Tur, and G. Tur. Social role discovery from spoken language using dynamic bayesian networks. In *Proc. of Interspeech*, 2010.

[11] J. Yuan and D. Jurafsky. Detection of questions in Chinese conversational speech. In *2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 47–52, 2005.