Recognition of Teaching Activities from University Lecture Transcriptions

Daniel Diosdado(✉), Alberto Romero, and Eva Onaindia

VRAIN, Universitat Politècnica de València, Valencia, Spain
{dadiolo,alrofer}@inf.upv.es, onaindia@dsic.upv.es

Abstract. Research on language acquisition for academic purposes is not extensive. In this work, we propose to build a system for recognizing teaching activities from automatic transcriptions of classroom audio and video recordings centered on the professor's discourse. To this end, we identified the main teaching activities that cover the nature of the lecturer discourse when giving a course e.g. 'theoretical explanation', real-world practical example', interaction lecturer-student', 'course-related asides', etc. We labeled a dataset of lecture transcriptions from a repository with an approximate length of 50 h and we build a classifier by fine-tuning the XLM-RoBERTa model with a classification head on top of it. The results will show that our proposal is a promising step ahead towards recognition of discourse activities in academic contexts.

Keywords: Spoken academic lecture · Text classification · Transformer models

1 Introduction

This paper centers around recognition of human activity where low-level data come in the form of transcriptions of audio recordings and the objective is to identify the nature of the discourse. More specifically, we aim at recognizing teaching activities from automated lecture transcriptions of university classes.

Research on language acquisition for academic purposes is not extensive. The Language ENvironment Analysis (LENA) system is one of the few existing tools that records the language environment of small children by combining a wearable audio recorder with automated vocal analysis software [8,9]. Data collected from teachers wearing the LENA system while teaching regular mathematics lessons has been used to identify three common discourse activities: teacher lecturing, whole class discussion and student group work [13]. LENA provides timely feedback for teachers to improve their skills in classroom discourse management but it uses a proprietary voice identification and transcription system, and it is particularly limited to small children.

The project Decibel Analysis for Research in Teaching (DART) analyzes the volume and variance of STEM (Science Technology Engineering Mathematics) course audio recordings to predict how much time is spent on single voice (e.g.,

E. Alba et al. (Eds.): CAEPIA 2021, LNAI 12882, pp. 226–236, 2021.
https://doi.org/10.1007/978-3-030-85713-4_22

lecture), multiple voice (e.g., pair discussion), and no voice (e.g., clicker question thinking) activities [12]. DART aims at studying patterns of active learning by comparing lecture and non-lecture activity (multiple and no voice) in courses for STEM majors versus non-STEM majors.

Our work focuses on recognizing teaching activities in university classes, which typically are more of the type of lecture-based learning, and builds on automated transcriptions of the class recordings. LENA and DART provide a discourse activity classification based on speech processing from audio recordings. Our proposal however exploits language modeling, what allows us achieving a richer teaching activity classification based on the analysis of the nature of the discourse and not merely on the audio-recording data.

We used video and audio recordings of lectures registered at our university prior to 2020, where the voice of students is hardly audible, and so our proposal for the recognition of teaching activities focuses on analyzing the nature of the teacher discourse. That said, we present a system based on text classification to recognize the type of speech of a university lecturer out of automated transcriptions of class recordings. We identify a set of categories that characterize the spoken academic discourse and we design a classifier using a transformer-based language model, specifically a version of the BERT family transformer models [5]. We underscore our system aims for recognizing discourse activities across a variety of different university subjects, thus the focus is not on topic modeling but on analyzing the discourse of a speech communication where the communicative intent and modality of the lecturer matter. Our ultimate goal will eventually be to cross-check the classification results with the academic evaluation surveys of the lecturers and study correlations between teaching activities and student satisfaction.

The paper is organized as follows. Section 2 briefly summarizes the main characteristics of the spoken academic lecture. The following section presents the procedure for the data segmentation and labeling. Section 4 explains the construction of the classification model and Sect. 5 shows the experimental results. Finally, we present a discussion and conclusions in Sect. 6.

2 Spoken Academic Lecture

In linguistics, the term **genre** refers to types of spoken and written discourse recognized by a community; e.g. lectures, conversations, advertisements, novels, shopping lists, interviews and many more. Since our work is devoted to the speech used in university classes, our focus is on the **spoken academic discourse** genre, particularly on classroom genres, which are regarded as paramount for both students and faculty. Among the classroom genres, the seminar, tutorial, presentation and oral exams typically involve a high level of interaction between the presenter and the audience the activity is addressed to [6]. The **academic lecture**, however, is mostly considered an expository genre where interaction and communication between teachers and students are less frequent.

The academic lecture is becoming more and more relevant due to the increasing internationalization of higher education both from the point of view of students and lecturers [7]. Lectures have a highly informational focus, similarly to academic prose, and, at the same time, have interactive features as they are delivered under on-line production circumstances that resemble face-to-face conversations in the spoken mode [1,4]. Hence lectures are categorized by features that capture the informational purpose of the speech and by features displaying the spoken discourse. Some researchers examined the macro-structure of university lectures and the micro-features that contribute to this structure [15]:

- **Interaction**: important feature that indicates to which extent lecturers maintain contact with their audience so as to reduce the distance between themselves and their listeners as well as to ensure that what has been taught is in fact understood.
- **Theory or Content**: this is used to reflect the lecturer's purpose, which is to transmit theoretical information.
- **Examples of practical application**: in this phase speakers illustrate theoretical concepts through concrete examples familiar to students.

Besides the three aforementioned identifying features of a lecture, more recent research also point at the ability of lecturers to **express their attitudes**, to **relate personal experience** to the content of the lectures, to talk about **evaluation of materials**, and to use formal and informal languages, spoken and written (text in slideshows or other forms of text) [11].

3 Data Segmentation and Labeling

The data used in this work is a collection of automated transcriptions of class recordings of university subjects given in Spanish. We used an online transcription and translation platform for automated and assisted multilingual media subtitling that provides support for the transcription of video, audio and content of courses[1].

We selected a total of 27 audio recordings of lectures that covered scientific as well as technical matters, e.g. Statistics, Electronic Devices, Mathematics, Microprocessors, etc. All together, the selected recordings feature 3000 transcription minutes, half corresponding to male lecturers and the other half to female lecturers. Additionally, 6 out of the 27 selected videos were manually reviewed so the transcriptions of these lectures are much more reliable and accurate to the original discourse of the speaker. Table 1 shows an excerpt of the output file returned by the transcription platform. A viewer will see the text of section 23 on screen as a caption, then the text of section 24 and so on.

[1] MLLP transcriptions. https://ttp.mllp.upv.es/index.php?page=faq.

Table 1. Excerpt of a transcription.

Spanish	English
23	23
00:09:07,890 --> 00:09:11,020	00:09:07,890 --> 00:09:11,020
¿Vale? Como bien sabéis ya de	Okay? As you know from
teoria de circuitos no existen	circuit theory there aren't
24	24
00:09:11,020 --> 00:09:15,220	00:09:11,020 --> 00:09:15,220
resistencias de cualquier valor. ¿Vale?	resistors of any value. Okay?
Sino que los tenemos tabulados.	But we have them tabulated.
25	25
00:09:15,220 --> 00:09:18,560	00:09:15,220 --> 00:09:18,560
En el laboratorio tenemos resistencias	In the laboratory we have resistors
de la serie E doce y	of the E twelve series and

3.1 Academic Labels

We manually segmented the transcription files by identifying context switching in the text and deciding whether said context change was associated to a change in the teacher's discourse. Data segmentation was done along with data labeling; we previously decided on the academic labels to classify the discourse activities so that a context switching is detectable as a change of label. We reviewed each other's work to ensure consistency in the labeling process.

On the basis of the investigations on the academic lecture genre, we performed an exhaustive analysis of the audio & transcription files and put forward the hierarchy of academic labels shown in Fig. 1. The white nodes denote parts

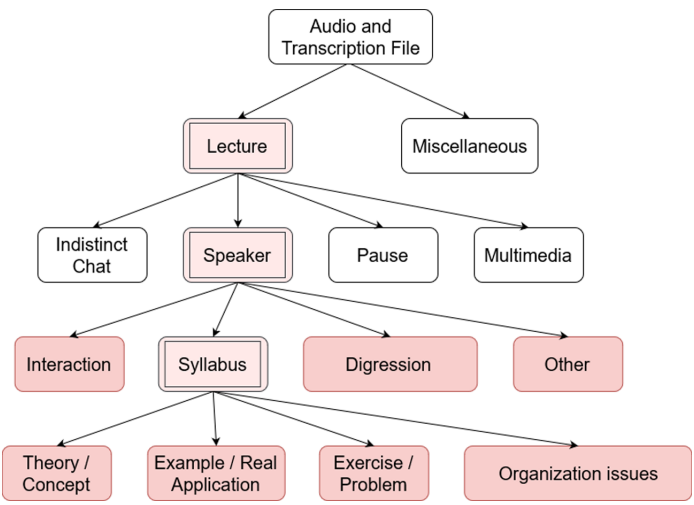

Fig. 1. Hierarchy of academic labels

of the audio file which do not have a readable transcription. The seven dark coloured leaf nodes are the academic activities we used to label the discourse segments. The meaning of each level is as follows:

Level 1: filtering out sounds from the audio file. The audio files of some recordings contain corrupted sections or unwanted sounds due to a sudden cutoff of the recording, background noise, errors in the recording or microphone feedback. We identify these damaged sections of the audio file as *Miscellaneous* and the rest is classified as audio that belongs to the *Lecture*.

Level 2: speaker identification. We distinguish the parts of the file in which the speaker (the lecturer) is talking from those in which they are not. The labels *Indistinct Chat*, *Pause* and *Multimedia* are used to mark sections of the audio file that contain an indistinguishable speaker.

Level 3: lecture-audience relationship. All the features at level 3 and 4 can be extracted from the transcription file since we end up with a file exclusively comprised of the discourse of the speaker after filtering the labels at level 1 and 2. The four labels at level 3 denote different ways for the lecturer to address the students. The key label *Syllabus* comprises the entire academic discourse around the specialized subject. *Interaction* denotes an exchange of communication between the lecturer and students; *Digression* is when a lecturer shifts to a more personal self and offers course-related asides; and *Other* refers to a speech unclassifiable under the other three labels which usually refers to the overall functioning of delivering the class and to non-course-related matters.

Level 4: content-based lecture structure. It includes the phases of a regular expository class around the syllabus of a subject, namely *Theory/Concept*, *Example/Real Application*, *Exercise/Problem* and *Organization issues*. Two remarks are worth mentioning: (a) the label *Exercise/Problem* accounts for a common practice in scientific/technical subjects but can be ignored in humanities and social science subjects; (b) *Organization issues*, which encompasses general course information like schedules, teaching practice or grading policy of interest for the carrying out of the syllabus, could also be classified as a subcategory of *Speaker* if we assume that students generally put much attention when the lecturer talks about organization matters.

We show now in Table 2 two examples of text segmentation and labeling in Spanish, and their English's translation.

4 Text Classification

For building our classifier for academic transcription segments, as we do not have much training data or the necessary hardware to train a NLP model from scratch, we employ transfer learning by using a pre-trained model and fine-tuning it to our task. To this end, we chose XLM-RoBERTa, a multi-lingual model that achieves a performance comparable to monolingual models in a variety of tasks such as named entity recognition, question answering, sentiment analysis, natural language inference, etc. The pre-trained XLM-RoBERTa model was downloaded from HuggingFace's Transformers repository [14]. We used *xlm-roberta-base* instead of the large version due to hardware restrictions.

Table 2. Examples of text segmentation and labeling

Spanish	English
¿Qué es lo que mide C M R R? Es la cantidad, lo que mide es la cantidad de ruido que un amplificador operacional es capaz de eliminar. ¿Vale? Al final normalmente las señales que queremos medir muchas de las señales que queremos de mi, medir son diferenciales.¿Vale? [Theory]	What does C M R R measure? It is the quantity, what it measures is the amount of noise that an operational amplifier is capable of removing. Okay? In the end, normally the signals that we want to measure, many of the signals that we want to me, to measure are differential. Okay? [Theory]
un ejemplo, la señal de electrocardiograma, encefalograma todas las señales biomédicas son diferenciales [Example]	an example, the electrocardiogram signal, encephalogram all biomedical signals are differential [Example]
¿De acuerdo? ¿Lo entendéis? ¿Sí o no? Sí. [Interaction]	Agreed? Do you understand it? Yes or no? Yes. [Interaction]
Vale, siguientes transparencias que os las dejaré, ahora lo escucháis un poquitín. A ver el video, el video, el video, vale, el vídeo. Si es que tengo una, unas, tengo unas preparadas, pero vale. [Organization]	Okay, next slides that I'll leave you, now you listen to it a little bit. Let's see the video, the video, the video, okay, the video. If I have one, some, I have some ready, but okay. [Organization]
¿Sí tengo una función de dos variables qué habéis deducido en cuanto a las derivadas? ¿A ver, quién me dice algo? ¿Quién me queda hacer un pequeño resumen, a ver, me hace alguien un pequeño resumen, a ver del vídeo? [Interaction]	If I have a function of two variables, what have you deduced regarding the derivatives? Let's see, who tells me something? Who is left for me to make a small summary, let's see, does someone make me a small summary, let's see of the video? [Interaction]

We set the maximum sequence length to 512 tokens, which is the maximum length supported by XLM-RoBERTa. The longer the sequence, the easier to classify a segment due to the larger amount of context information comprised in it. If the segment contains less than 512 tokens we apply padding, and for segments longer than 512 tokens we split them using a sliding window with a stride of 0.8 * max_seq_length (410 tokens).

We added a classification head on top of the pre-trained XLM-RoBERTa model. This classification head takes as input the segment representation contained in the embedding of the *classifier token* (a special token added at the beginning of the segment). The classification head consists of a dense layer of hidden size (768 units) with *tanh* activation function followed by a dense layer of seven units (one unit for each label/class associated to one academic activity) with *softmax* activation. The output of the classification head is a list containing the probability that the input segment belongs to each of the seven classes. We used the *Adam* algorithm with weight decay fix as optimizer and *categorical cross-entropy* as our loss function for fine-tuning.

The hyperparameters were tuned with the *Weights and Biases* framework [2] according to the model performance using Bayesian Optimization. The final values of the hyperparameters are: learning rate = 0.00005, 40 epochs, batch size of 8 segments (due to memory constraints), gradient accumulation steps of 32 (for a simulated batch size of 256 segments) and weight decay of 0.0007. We trained our model with a Nvidia Geforce RTX 3090.

Table 3. Distribution of our data by label. The number of tokens was obtained by using XLM-Roberta's tokenizer.

Label	Num segments	Total tokens	Avg. tokens	Max tokens
Theory/Concept	454	115885	255.25	3019
Exercise/Problem	537	77866	145.01	1481
Example/Real Appl.	347	63239	182.24	1845
Organization	260	37083	142.63	1989
Interaction	567	66326	116.98	3878
Digression	118	11857	100.48	647
Other	112	5416	48.36	297
Total	2395	377672	157.69	3878

Table 3 shows the composition of our dataset by class. For each class, we report the number of text segments, the total number of tokens, the average token length of the segments and the maximum length in tokens of a segment. As we can see in Table 3, our dataset is rather imbalanced, as some classes like Theory/Concept or Exercise/Problem are much more frequent than other classes like Digression or Other. This is reasonable and in line with the nature of the academic lecture, wherein the most part of the teacher's speech is devoted to the contents of the syllabus of the subject. As a result, the number of segments and the total number of tokens of the most populated classes is obviously higher.

We can draw some conclusions about the composition and structure of each class. The three academic labels *Theory/Concept*, *Exercise/Problem* and *Example/Real Application* make up 56% of the total number of segments, and 68% of the total number of tokens in the dataset. These classes also share a substantial part of their vocabulary with one another. The *Other* class is fairly regular as it is mostly composed of relatively short segments. This is indicated by the lowest value of the maximum number of tokens (297) in a segment and also by the lowest average number of tokens (48.36). In contrast, *Interaction* is highly irregular because it contains segments of variable length. *Interaction* is the class with the largest number of segments (567) and the longest segment (3878) while this class has about half the number of tokens of *Theory/Concept*, and its average number of tokens is closer to the less populated classes. We also observed that *Interaction* appears more frequently among segments of *Exercise/Problem*, either by the student asking for clarification or the teacher querying the students.

5 Experimental Results

We evaluated our model on the whole dataset using 10-fold cross-validation with stratify (each partition holds approximately one tenth of each class). We report the aggregated confusion matrix of the results in Fig. 2 and the values of precision, recall and F-score for each class in Table 4.

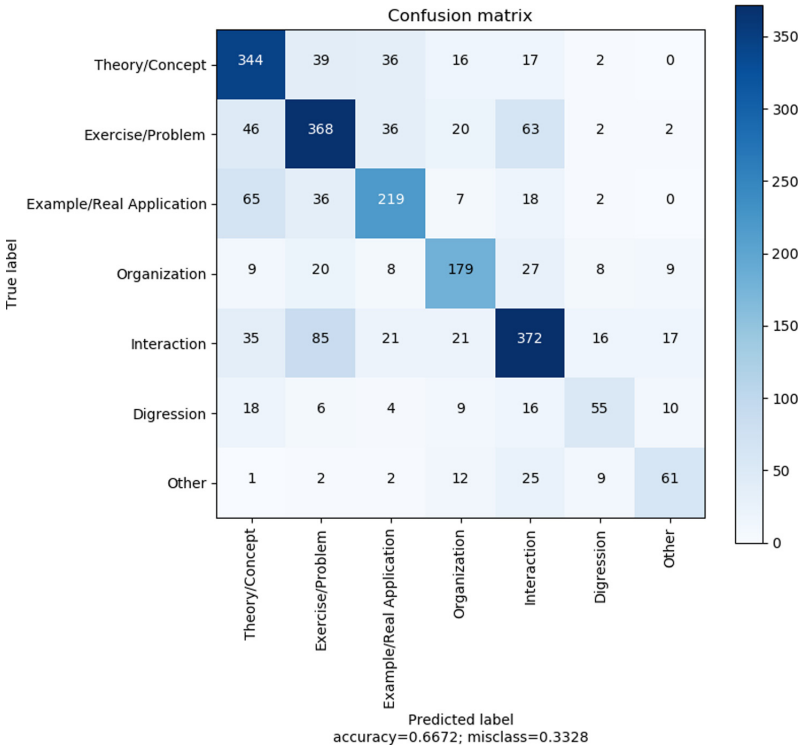

Fig. 2. Aggregated confusion matrix of 10-fold cross-validation

In the confusion matrix of Fig. 2, rows show the true label of the segments, i.e., the label we manually assigned to the segments, and columns represent the predicted class by our model. The values on the diagonal are the number of True Positives (TP); for each class, the values in the columns show the False Positives (FP) and the values in the rows show the number of False Negatives (FN).

Table 4 shows that the metrics of the class *Digression* fall behind the rest of the classes, followed by the class *Other*. *Digression* has the lowest recall among all classes, which is also confirmed by the high number of FN in Fig. 2 relative to the number of samples of this class. This reveals the difficulty of the model to correctly classify segments of *Digression*. We believe the reason for the poor performance of the model with classes *Digression* and *Other* is due to the low

number of samples of these two classes in our dataset as well as the difficulty we experienced to correctly label the *Digression* samples.

It is also noticeable that the precision value of *Interaction* is higher than its recall value, and higher than the precision of the other classes. The fact that *Interaction* is the class with the largest proportion of samples correctly classified responds to the ability of the model to recognize the distinguishing characteristics of the discourse style in this class. In *Interaction* the teacher frequently uses the second person to address the students, for querying them or answering their questions. Interestingly, the model is able to identify the interaction teacher-student even not yet having transcriptions of the students' speech.

Table 4. Precision, Recall and F-Score by class.

Label	Precision	Recall	F-Score
Theory/Concept	0.664	0.758	0.708
Exercise/Problem	0.662	0.685	0.673
Example/Real Application	0.672	0.631	0.651
Organization	0.678	0.689	0.683
Interaction	0.692	0.656	0.673
Digression	0.585	0.466	0.519
Other	0.616	0.545	0.578

In Fig. 2, the majority of high values in the columns other than the diagonals are concentrated in three classes: *Theory/Concept*, *Example/Real Application* and *Exercise/Problem*. We can thus say our model has a certain bias towards this group of classes which otherwise seems reasonable since they are the most representative classes of the academic discourse of a lecturer and share a substantial part of their vocabulary. We also observe a significant amount of misclassifications between *Exercise/Problem* and *Interaction*. This happens because it is typically the case that students get more engaged during problem-solving than theory explanations.

The class *Theory/Concept* shows the highest recall but a lower precision value (low values in its row in Fig. 2 compared to the values in its column). This means the model is fairly successful in correctly classifying many of the segments labeled as *Theory/Concept*, but it also tends to classify as *Theory/Concept* segments that do not actually belong to this class. This reveals a slight bias towards this class which is probably due to *Theory/Concept* being the largest class in the dataset. The second highest F-score of the class *Organization* is likely explained by the particular vocabulary of this class, which easily distinguishes it from other classes, such as terms that denote dates, weekdays, grading system, explanations about laboratory activities, etc.

6 Conclusions and Future Work

Despite the low number of labeled samples for this type of NLP task, our model achieves a significant performance that confirms the adequacy of our labeling scheme for recognizing teaching activities from automated transcriptions. We observed that the three academic classes *Theory/Concept*, *Exercise/Problem* and *Example/Real Application* concentrate a large part of the errors because these classes make for more than half of the dataset and share a large part of their vocabulary. Misclassfications between *Interaction* and *Exercise/Problem* happen because many interactions student-lecturer take place during exercise solving in class. Additionally, the model achieves good results for the *Organization* class thanks to its distinctive and recognizable vocabulary (dates, grading, etc.).

A straightforward way to increase the performance of our model is by augmenting the size of the dataset [3,10]. We plan to develop an automated segmentation process and use our classification model to help us augment the dataset. Additionally, we will consider using a Language Model to spellcheck the text and thus improve the quality of the transcriptions. Lastly, we intend to test XLM-RoBERTa-large and check if the superior performance of the large version over the base model [3] translates into a improvement in our classification model.

References

1. Biber, D.: Dimensions of Register Variation: A Cross-linguistic Comparison. Cambridge University Press, New York (1995)
2. Biewald, L.: Experiment tracking with weights and biases (2020). https://www.wandb.com/. software available from wandb.com
3. Conneau, A., et al.: Unsupervised Cross-lingual Representation Learning at Scale (2020)
4. Csomay, E.: Academic lectures: an interface of an oral/literate continuum. NovELTy **7**(3), 30–48 (2000)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805
6. Fortanet-Gómez, I.: Honoris Causa speeches: an approach to structure. Discourse Stud. **7**(1), 31–51 (2005)
7. Fortanet-Gómez, I., Bellés-Fortuño, B.: Spoken academic discourse: an approach to research on lectures. Revista española de lingüística aplicada **1**(8), 161–178 (2005)
8. Ganek, H., Eriks-Brophy, A.: Language ENvironment analysis (LENA) system investigation of day long recordings in children: a literature review. J. Commun. Disord. **72**, 77–85 (2018)
9. LENA Research Foundation: The LENA research foundation (2014). http://www.lenafoundation.org/
10. Liu, Y., et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019)
11. Malavska, V.: Genre of an academic lecture. Int. J. Lang. Lit. Cult. Educ. **3**(2), 56–84 (2016)

12. Owens, M.T., et al.: Classroom sound can be used to classify teaching practices in college science courses. Proc. Nat. Acad. Sci. **114**(12), 3085–3090 (2017). https://doi.org/10.1073/pnas.1618693114
13. Wang, Z., Pan, X., Miller, K.F., Cortina, K.S.: Automatic classification of activities in classroom discourse. Comput. Educ. **78**, 115–123 (2014)
14. Wolf, T., et al: Huggingface's transformers: state-of-the-art natural language processing. CoRR abs/1910.03771 (2019)
15. Young, L.: University Lectures - Macro-structure and Micro-features, pp. 159–176. Cambridge University Press, Cambridge Applied Linguistics (1995). https://doi.org/10.1017/CBO9781139524612.013