UNIVERSITÉ DE LORRAINE

L'INSTITUT DES SCIENCES DU DIGITAL MANAGEMENT COGNITION

SUPERVISED PROJECT

What are you saying?
Dialogue act annotation

Authors:
Albert MILLERT,
Anar YEGINBERGENOVA

Supervisors:
Chuyuan LI,
Maria BORITCHEV,
Maxime AMBLARD

A bibliography part of the Supervised Project

Academic year 2020-2021

December 17, 2020

Contents

Introduction

Depression, as a complex mental condition, requires at least an equally complex system for its early detection. The lack of ready to use end-to-end solutions and several related research has inspired us to explore the topic and to decide to develop an early detection system supplementary to specialists' diagnoses.

We've narrowed down the broad subject to primarily address interviews' transcriptions. In addition to the main problem of properly identifying depression cases among patients, we focus on creating an abstraction of the dialogue through parsing. Despite the problem constituting merely the input to the classification problem, it's a crucial step in the project as a whole. It also serves as the domain's knowledge base which helps the interpretability.

It's important to utilize tools accordingly to the set of problems they've specifically been designed to address. We believe that an automatic dialogue parser, given a very detailed and descriptive set of heuristics, will yield at least as good results as its alternative stochastic methods. However, probabilistic methods shine when applied to any prediction-related problems. Additionally, such decomposition of the use of the mathematical tools and the use of respective methods should drastically improve the model's (and its results') interpretability.

The retrieved model should be universal enough to be then put to the test with Tweets of the self-proclaimed depressed individuals and (or) interviews' transcriptions in different languages and regarding slightly different patients' mental conditions.

In the first 1 section, we would like to briefly discuss depression as an illness, provide some linguistic approaches used in psychotherapy, and the need for an accurate detection tool. Sections 2, 3, 4 provide the necessary theory to understand the most important concepts in regards to the system, which proposition is presented in the 6 section. 5 section regards the dataset and gives some initial insights into the data. We finish off by discussing the hypotheses and summarizing the research project in the final section 8.

1 Depression Today

Depression is one type of mental health disorder. It affects human psychology, thoughts, behavior, and overall well-being after, usually long-lasting, negative feelings, sleep troubles, loss of interest in everything, etc. In the worst case, it may lead to suicide. In the era of technology and social media, where human-human interaction decreases and everyone connected in digital world more and more, people get affected by depression, especially, teenagers and young adults. An ever-growing number of people affected by this disease makes it one of the most popular and, at the same time, the most dangerous psychological problem of human beings at this moment. Therefore, it is vital to detect it at an early stage and get appropriate help.

1.1 Detection

Depression can affect anyone. In fact, a person could not be aware of having it, but depression might bring several changes in the everyday life of an individual, such as change of personality, mood fluctuations, the difference in the manner of communication, and in any social interaction. Nowadays, some approaches are used to try to detect depression on the early stage. Such methods may include questionnaires, interviews or even analysis of social media activities. After detecting these symptoms the right thing to do is consulting an expert. The therapy includes some tests and interviews to identify the disease and helps to find a treatment solution.

1.2 Language patterns

The connection between language and depression has been under research in order to discover the exclusive features of an individual. The linguistic features like lexical diversity, average sentence length, grammatical patterns and classes of words might be crucial in analyzing it.

Key factors of depressed language are social skill deficit, responses are short and vague, and self-focus indicating detachment from the community. In fact, the linguistic pattern of depression consists of the use of words that carry negative emotions such as, *lonely*, *sad*, or *miserable*. They also excessively use first-person singular pronouns - *I*, *me*, *myself* - which indicates that depressed people are more focused on themselves. (Bucci and Freedman, 1981)

1.3 Need for early diagnoses

Depression in recent years has become a major issue (especially) among teenagers and young adults. Depression is the primary world's disease. The amount of people concerned with this illness is still growing, and the trend doesn't seem to come to an end any time soon. Unfortunately, all mental health issues are still stigmatized in society, and often it's hard for people to seek real help.

It's been estimated by *WHO* (World Health Organization - a specialized agency of the United Nations responsible for international public health) that 4.4% of the whole population suffers from depression (WHO et al., 2017). In spite of depression being a way more common illness among females compared to males (5.1% vs. 3.6%), the suicide rate (regardless of average living conditions in the country) is drastically higher among men than women.

The topic is very delicate and must be addressed with exceptional care. We think it's a great idea to try to create a tool for early detection of depression symptoms based on human speech, more specifically interviews' transcription in case of our project. We strongly believe there are certain language nuances that could provide indicators leading to a proper patients' classification.

2 Dialog Act Annotation

The project is directly related to interview dialogues so it's crucial to establish common notions and definitions of key terms to avoid unnecessary misunderstanding. The general purpose of dialogue act annotation is marking up stretches of dialogue with information about the dialogue acts, which is often limited to marking up the communicative functions (see 2.2).

2.1 Dialogue act basic units

Dialogue act can be considered an update operation on one's information state; the process is commonly denoted as an information-state (context-change) update approach. It consists of at least two participants:

1. speaker (sender) - whose communicative behavior is interpreted and who intends to occur in the information state of a dialogue participant;
2. participant (addressee, recipient) - to whom the speaker is communicating and whose information state is being influenced.

In the case of the conducted research, we're dealing with the interviews with patients (see section 5 for more details); therefore, distinguishing between the two types is sufficient; however, generally, there could be more participants, whose roles in dialogue may vary.

Stretches of communicative behavior produced by one speaker, bounded by periods of her inactivity are commonly denoted as *turns*. They're traditionally used to segment spoken dialogues; however, they're often too lengthy and coarse to have a communicative function assigned. *Functional segments*, in turn, are minimal functionally relevant stretches of communicative behavior. They are a better fit for the purpose.

Dialogue acts are usually dependent on several previous dialogue acts, hence they're responsive character. In some cases, the system may require marking up the relations to antecedents on which the meaning depends. Many dialogue act annotations schemas ignore altogether indirect speech acts, occurring when the speaker communicates something that has a different meaning than it appears (Bunt et al., 2010).

The metamodel in figure 1 is a diagram representation of key concepts involved in the dialogue act annotations based on the Bunt et al. (2010). Each dialogue act is related to one functional segment, which in turn can relate to many dialogue acts - possible multifunctionality. The numbers denote cardinality between the objects: *1* asserts that only a single object can be present in certain relation, *1..** denotes at least one object, * - any number of objects.

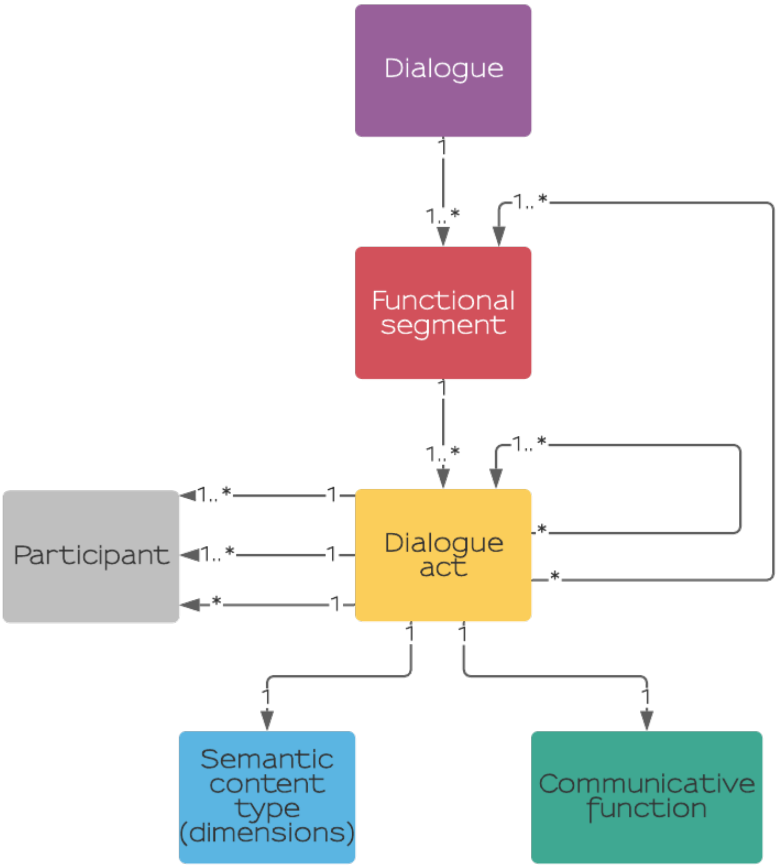

FIGURE 1: Dialog act annotation metamodel diagram

2.2 Communicative function

Communicative function is a specification of the way an addressee uses the semantic content to update her information state when she understands the part of the dialogue. Intuitively, it corresponds to the type of the performed action. The term *dialogue act annotation* is often referred to as the process of assigning the communicative function labels to considered stretches of dialogue.

Existing dialogue act annotations schemas define communicative functions either in terms of (1) intended effects on dialogue participant; (e.g. questions, confirmations, promises are usually defined as such); or - (2) properties of the signals in use (e.g. repetitions, openings, closings).

The success of the communication is dependent on the participant's ability to interpret the communicative functions introduced by the speaker in the way intended by the speaker.

Semantic content specifies objects, relations, actions, events, etc., that the dialogue acts regard.

2.3 Multifunctionality and dimensions

Dialogue's utterances (considered stretches) are often multifunctional - they have multiple communicative functions. Participants share information introduced through the dialogue act, but they also share information about the processing of each other's messages, e.g. about the allocation of turns, contact, and attention, use of time.

A term *dimension* refers to various types of semantic content or the types of communicative activity concerned with these types of information. Each dimension (1) has a clear empirical basis, corresponding to observed forms of behavior in dialogue; (2) is theoretically justified, corresponding to a well-established class of communicative activities, such as taking turns or giving feedback; (3) is recognizable with acceptable precision by human analyst and by dialogue understanding and dialogue annotation systems; (4) can be addressed by dialogue acts independently from addressing other dimensions (independent or orthogonal). Core dimension is present in many existing dialogue act annotation schemes; there are nine core dimensions proposed by Petukhova and Bunt (2009). There are several requirements/propositions for choosing the core dimensions in the domain of interest:

1. empirical validity - for every communicative function there exist linguistic or nonverbal means that can be used to indicate that one's behavior has a certain function;
2. theoretical validity - every communicative function has a definition which semantically distinguishes it from others;
3. a set of communicative functions applicable in a certain dimension provides good coverage of the phenomena in that dimension;
4. the communicative functions should be recognizable comparably well both by human annotators and the system;
5. the core communicative functions occur in many existing annotation schemas;
6. if one communicative function applies to a given segment, then the other one mustn't; or one is the other one's specialization.

Many systems for the dialogue act annotation take into account the multifunctionality of the dialogues' utterances. An example of such a system is *DIT++* described in a bit more detail in the following subsection where we will briefly discuss its key principles and utilized techniques. More methods can be found in Fort (2012) in which author compared and analyzed them.

2.4 DIT++

There have already been several annotation schemata proposed in recent years, each of them created for slightly different purposes and specific application domains, e.g. *DIT++* (Bunt, 2009) which works with data categories. *DIT++* is an acronym for *Taxonomy of Dialogue Acts, Annotation Scheme, and DiAML Markup Language*. By the principle, it's based on the semantic analyses of inter-human and human-machine dialogues. *DIT++* has been developed with respect to the *ISO* standards; and in some applications it's considered a standard approach itself. Given that, our solution (see section 6) has been partially inspired by some of the proposed methods.

DIT++ takes advantage of segments' categorization which, in turn, is the main principle behind description logic. In our case, the first order predicate logic appears to suffice (as described in 3). To successfully represent categories one must have access to posets (lattice structure). It can be obtained by the outside ontologies or by utilizing *WordNet* (Fellbaum, 2012).

3 Parsing an Abstract Dialogue Structure

A successful dialogue structure parser must address several smaller subtasks. Primarily, the system segments the dialogue into smaller chunks, which then serve as primary units for further processing or information extraction (see 3.3). Meaning representations are formal structures that capture the semantic contents of linguistic expressions. It's achieved by assigning the meaning representation to the linguistic inputs within the domain world of the represented dialogue. The theory behind this task and more details regarding it are addressed in 3.1 section. Most of the above tasks are addressable by means of the pure first-order predicate logic (for more detail refer to 3.2 section).

This section is primarily devoted to introduce the mathematical tools, approaches that help address this issue. Montague (1970) introduced the theory that became *Montague grammar* stating that natural and formal languages can be treated in much the same way. To better understand the purpose of this section, it can be useful to first help the reader understand what exactly does the term *abstract dialogue structure* denote. Dialogue is, in most cases, a spontaneous act of speech among its participants. Expecting it to be structured in any way may, intuitively, seem to be counterintuitive for some readers. It partially holds true if one considers a structure to be a fully organized set of objects; however, such an ordering is possible to retrieve at some level of abstraction. The represented information may differ depending on the application it's designed for.

3.1 Meaning representations

Meaning representations serve as a key building block in the abstract dialogue structure. They are responsible for conveying the semantic information of the analyzed units.

The following are characteristics of the meaning representations and the benefits they provide (Jurafsky and Martin, 2000):

- **verifiability** - one must understand what is the exact object of the discussion, and whether the utterance makes sense in the context; verifiability allows one to compare the information described by a representation to the information in the knowledge base;

- **unambiguous representations** - utterences' components, in general, may have different meaning representations depending on their occurrence contexts; however, the meaning representations themselves must be unambiguous so that system can reason over representations and choose the right answer;
- **canonical form** - distinct inputs with the same meaning should have the same meaning representation; it simplifies reasoning by narrowing down the reasoning space; however, it complicates the semantic parsing process by enforcing the system to know how to map distinct parse structures to the same meaning representation;
- **inference and variables** - the meaning representation of the request should be connected with the fact about the world in the knowledge base; the inference process is needed to draw conclusions based on the meaning representations of inputs and its background knowledge (for more details refer to 3.2.1 section); the system must be able to conclude prepositions represented (even) implicitly in the knowledge base, but which are nonetheless derivable; this type of request requires the use of variables,
- **expressiveness** - the representation scheme must be expressive enough to handle a wide range of subjects; it's almost impossible to represent the whole knowledge within some domain but (nonetheless) first-order-logic provides tools that allow it.

A *model* is a formal construct that represents a state in the world. Expressions can be mapped to the objects, their properties, or relations between them. If the model correctly captures the facts of one's interest, then the mapping between the meaning representations and the model provides a bridge between meaning representations and the world.

Non-logical vocabulary consists of an open-ended set of names for the objects, properties, relations of the represented world, all of which appear in various schemes represented as predicates, nodes, labels on links, etc. Elements of the non-logical vocabulary must correspond to a specific part of the model. *Logical vocabulary* consists of a closed set of symbols, operators, quantifiers, links, etc. that compose expressions in a given meaning representation language.

The *domain of the set* is the set of the represented *objects*. *Objects* denote elements of the domain, *properties* - sets of elements of the domain, *relations* - sets of tuples of elements of the domain.

Interpretation is the mapping from meaning representation to the corresponding denotations; function from a non-logical vocabulary of meaning representation to the denotations in the model.

A model describing a certain domain must be represented in an interpretable way so that one can correctly extract information from it and reason about it. A mathematical tool to address this issue is first order predicate logic.

3.2 First order predicate logic

The whole syntax of the *first-order logic* (*predicate logic*) representations can be described by the context-free grammar specification:

- **term** is a device for representing objects; one distinguishes:
  - **constant** which is a specific object in the world, often denoted as a capitalized letter, or a capitalized word (noun); a constant refers to only one object but a single object can have multiple constants referring to it;
  - **function** is syntactically the same construct as a 1-arity predicate, but in fact, it's still a term because it refers to a single object; functions provide a convenient way of referring to specific object without having to associate a named constant with it;
  - **variable** is a single lowercase letter (unlike in Prolog) which allows one to make assertions and draw inferences about objects without having to make reference to any particular named object; this ability to make statements about anonymous objects is used when making statements about a particular unknown object or - all the objects in some arbitrary world of objects;
- **predicate** is a sequence of symbols referring, or naming the relations that hold among some fixed number of objects in a given domain;
- **logical connectives** allow larger composite representations to be put together; recursiveness of the grammar allows one to create an infinite number of logical formulas through the use of the connectives; meaning that finite devices can be used to create an infinite number of representations: **conjunction**, **disjunction**; **implication**;
- **quantifiers** allow two use-cases for the variables and instruct how to interpret the variable in the context of the sentence: **existential quantifier** denoted as $\exists$ is often signified by the presence of indefinite noun phrase in English; **universal quantifier** denoted as $\forall$ asserts that there must be at least one object, st if a variable was substituted by a specific object, the resulting sentence would hold true.

Capturing the meaning of a sentence requires identifying terms and predicates. **Terms** are the objects in the world, they denote elements in the domain. **Atomic formulas** are captured either as the sets of domain elements for properties, or sets of tuples of elements for relations. There's a slight difference between logical and, or, if and their English correspondents.

$$\textit{I really want to go to Mexico and Colombia.} \tag{1}$$

$$\textit{I want to do it either today or tomorrow.} \tag{2}$$

$$\textit{I want to get a haircut tomorrow or on the weekend.} \tag{3}$$

In the (1) the phrase doesn't necessarily entail that a person wouldn't enjoy a travel to either country; therefore, *and* could have been loosly translated into logical *or*, rather than *and*. The *either ... or ...* grammatical construct commonly implies exclusiveness in speech (2); however, depending on the context, textual *or* may in fact represent the logical *xor* (3). The examples prove that one must be careful while translating linguistic connectives to their logical equivalents.

3.2.1 Inference

As previously mentioned, inference is a mathematical tool needed to draw conclusions of all the information present in the knowledge base. Even the information mentioned implicitly must be derivable.

Modus ponens is a formal tool which allows logical inference. The (4) example could be understood as *given P; if P, then Q; entails Q* which is equivalent to the notation in (5).

$$\frac{P \quad P \rightarrow Q}{Q} \tag{4}$$

$$P \rightarrow Q; P \vdash Q \tag{5}$$

Modus ponens occurs in two main algorithmic forms, namely - *forward* and *backward chaining* which differ significantly.

In forward chaining facts are already present in the knowledge base when needed which significantly reduces the time needed to answer subsequent queries since they should amount to a lookup; however, facts that will never be needed may be inferred and stored.

Backward chaining runs in reverse to prove specific prepositions (queries). The process begins by checking if a query formula is true by determining if it's already present in the knowledge base. In case it's not, the algorithm begins to search for applicable implication rules present in the knowledge base - rules that are consequent if the rule matches the query formula. The algorithm decides that the query is provable if the antecedent of any one of them can be proved. The process is being performed recursively. This inference type of inference is utilized for instance in Prolog.

Resolution is an alternative inference technique - it's both sound and complete, as opposed to both forward and backward chaining methods which are sound but incomplete, but it's way more computationally expensive.

3.2.2 Event and state representations

We've already introduced how the dialogue's structure can be represented and reasoned about. This part covers a great portion of mathematical tools required to tackle our problem but it's not enough.

Plain representations of events and states consist of a single predicate which arity is fixed - limited to a constant number which is not representative in case of many sentences. An *event variable* is a formal tool that addresses this problem. It simply requires refactoring by introducing an existentially quantified variable as the only argument. With this simple trick, one can provide as many predicates as one desires.

$$I\ have. \tag{6}$$

$$I\ have\ been\ diagnosed. \tag{7}$$

$$I\ have\ been\ diagnosed\ with\ depression. \tag{8}$$

$$I\ have\ been\ diagnosed\ with\ depression\ last\ month. \tag{9}$$

Sentences from (6) through (9) present the problem of logical representation as a fixed-arity predicate. Each consecutive sentence adds some information on top of the previous one. This phenomenon cannot be addressed easily without event variables.

$$\exists e\ ConfirmAnswer(e) \tag{10}$$

$$\exists e\ ConfirmAnswer(e)\ \wedge\ DiagnosedPerson(e, PatientA) \tag{11}$$

$$\begin{aligned}\exists e\ ConfirmAnswer(e)\ &\wedge\ DiagnosedPerson(e, PatientA)\\ &\wedge\ Illness(e, Depression)\end{aligned} \tag{12}$$

$$\begin{aligned}\exists e\ ConfirmAnswer(e)\ &\wedge\ DiagnosedPerson(e, PatientA)\\ &\wedge\ Illness(e, Depression)\\ &\wedge\ RelativeDate(e)\\ &\wedge\ Month(Last)\end{aligned} \tag{13}$$

Logical representations from (10) through (13) correspond to the sentences (6) through (9) respectively. The issue of fixed-arity predicates has been addressed by utilizing existentially qualified variable which is present in the subsequent predicates logically conjunct. This type of representation is known as a *neo-Davidsonian* event representations. These types of constructs follow the rules guidelines:

- events are captured with predicates that take a single event variable as an argument;

- there's no need to specify a fixed number of arguments for a predicate since as many roles and fillers can be glued on as are provided in the input;
- no more roles are postulated than are mentioned in the input;
- logical connections among related inputs sharing the same predicated are satisfied without the need for additional inference.

3.3 Information extraction

Firstly, one must detect all the entities in the dialogue. Named entities are entities' type that can be referred to by a name. Each named entity should fall under some category. Some applications require defining domain-specific entity types on top of the most common ones.

Identifying entities in the dialogue is the initial step in question answering systems. It's also a crucial step to successfully link information as structured knowledge. Interviewing requires, at some point, one agent to ask, and the other interlocutor - to address those questions. The dialogue parser formalizes the obtained knowledge, which can be further processed.

Named entity recognition means finding spans of text with proper names (entities) and then classifying their types. An entity can fall into several categories, and it's a matter of choosing the correct one given its occurrence context.

Specific relations tend to occur among particular types of entities. Such relations are often represented using *Resource Description Framework - RDF* - triples, s.t. $subject - predicate - object$. Extracting relations can be achieved by matching certain patterns.

4 Ultimate Illness Classification

Dialogue act and dialogue structure representation can assist the semantic interpretation of utterances and can help to understand the spoken language. In our task of depression classification, we will infer to this approach in order to understand the specificity of patients' language in the interview. These conversations hold lexical, syntactic, and semantic information that we could analyze, which might be crucial since there are some unique patterns in the language of the person under scrutiny. Order of words, context, turns in conversation, the meaning of words taken into consideration in order to predict and classify the mental health illness. Having the abstract representation of the dialogue will help us facilitate the classification of depression. Moreover, we will be able to explore peculiarities in the language of inputs within different classes more accurately. Similar tasks were tested before and the most popular approach is using machine learning approaches or infers to neural networks. Different machine learning algorithms handle classification problems differently, but still, the effectiveness of them may vary from task to task. One such algorithm is Decision Tree, which is usually represented as a graph the root of which is the starting point of the task, and each following node contains a set of conditions to evaluate with possible outcomes from that conditions. Neural Networks propose the possibility to discover unexpected features or confirm/refuse the initial hypothesis and perform a classification tasks accordingly.

According to the statistics provided by World Health Organization(WHO), any psychological disorders are a wide-spread issue of public health today (WHO et al., 2017). While getting professional treatment seems to be the only way out, it is still not affordable to everyone, and it requires studies to better understand why it happens and how to treat it more efficiently.

Although depression is a vital topic to analyze, it is a subgroup of the general mental health disease group. Mental health issues include plenty of different illness types, some of them might correlate with each other or be absolutely distinct. There are a lot of studies done in the area of detecting different kinds of mental health diseases, such as DSM IV, (American Psychiatric Association, 2000) which is a publication that describes different types and classifications of mental disorders. This section will take a look into several aspects that were considered in the mental health area.

A major part of psychological disorders can be triggered by constant changes in mood. Usually, mood fluctuations can be traced from any textual data, e.g. interviews, everyday interactions in society, and even social media posts. Even though this task is more inclined into the sentiment analysis problem, it is also important to take into account that the geographic and temporal variation play a more vital role in mental health identification. The happiness rate of the location can be considered as the indicator for the overall psychological well-being of the region (Dodds et al., 2011)

Textual data is easily accessible, especially, in the times of social media. In order to collect and predict psychological disorders data collected from self-reported surveys, SMS, social media posts, and interviews are used for this task. It is considered that medical well-being surveys are the most reliable source of data in order to predict mental illness, but this type of resource also considered financially unreasonable. More and more experiments are done by analyzing social media as the popularity grows dramatically and people share their feelings and emotions publicly for everyone.

(Schwartz et al., 2014) created users' continuous depression scores across Facebook users based on their activities during a year. This study indicated seasonal fluctuations of depression that people are more depressed during winter. Based on the Twitter posts of people who had depression on self-declared diagnosis(those who explicitly share posts like "I was diagnosed depression today"), showed that estimating the users' age allows more accurately predict if the user has any mental issues. However, it also shows that these data largely overlaps with language that predicts the personality, it indicates that users with particular personality voluntarily post their diagnosis publicly (Guntuku et al., 2017). (Choudhury and Gamon, 2013) developed a classifier for depression using texts from Twitter and proposed the Social Media Depression Index (SMDI), which is based on the group of tweets generated by users or by people from the same geographical location and they achieved outstanding results that correlated with the results from Centre for Disease Control (CDC) based solely on Twitter posts. Alongside with this, they built a classifier to estimate the risk of depression before it happens. According to the results, depressed people are less active in responding and using third-person pronouns and more inclined to be active during the night and use more first-person pronouns.

Clinical data is another source of information to analyze mental problems better. One such source is Electronic Health Records (EHS). It helps to classify different kinds of mental diseases with the main focus on schizophrenia and bipolar disorder using different tools to annotate data and predict the illness. Studies on these datasets imply that the symptoms expressed by patients are more reliable to indicate the disorder rather than relying on the symptoms of the description of the illness (Jackson et al., 2017).

Regarding professional treatment, it is important for experts to identify any mental decline in a patient's behavior and respond quickly to provide appropriate help. Their success rate with patients measured by their ability to adapt to the circumstances and provide individual help by being more creative in conversations rather than using templated responses, especially, when treatments are undertaken in textual format (e.g. SMS) (Althoff, Clark, and Leskovec, 2016)

Overall, with the development of different approaches to get insights about mental illness and with the growing amount of data more and more studies are taking place to retrieve some useful features to help with early diagnosis or more accurately recognize the type of disease in general.

5 Dataset

Distress Analysis Interview Corpus dataset (Gratch et al., 2014) consists of English transcripts of the interviews conversations of a virtual assistant with patients. The dataset is especially encouraging since it's similar in its nature to the corpus dataset used in the *Sémagramme* team's *SLAM - Schizophrénie et Langage : Analyse et Modélisation* (Amblard, Musiol, and Rebuschi, 2015) research project. That makes us believe that the conducted research project may be universal.

The *SLAM* corpus and *Sémagramme's* research around it have greatly inspired the idea for this project. In this project we're trying to address a very similiar set of problems but the methods have been chosen independently from the ones utilized in the *SLAM*. Additionally, the two datasets are small, and they could have additionally been biased by the expert/avatar carrying out the interview. Even without paying much attention to this fact, certain responses may influence the way interviewers continue the conversations.

The two most noticeable differences regard interviews' languages (French in *SLAM* and English in *DAIC*), as well as the detected mental illness (schizophrenia in *SLAM* and depression in *DAIC*).

5.1 Details

The attributes present in the *_TRANSCRIPT.csv* files are: *start_time*, *stop_time*, *speaker*, *value*. The *speaker* feature shall be used as a feature to detect turns[1] in the conversation; it takes either of the two values: *Ellie*, or *Participant*, indicating interviewing virtual assistant and a patient respectively. The *value* feature corresponds to the actual utterence spoken by the interlocutor. The whole dataset constitutes 192 interviews' folders. The paper mentioned 170 interviews - there's some inconsistency with the number of the actual interviews found in the database. Out of the original 170 interviews, 49 patients have been identified as depressed. It confirms the fact that the dataset is small and imbalanced since both categories are so irregular in size. However, this is the usual amount of data at the disposal of doctors in many research.

5.2 Statistics

The statistics have been calculated on a subset (roughly $\frac{1}{3}$) of the whole dataset. The main focus of the calculations was to explore the turns in the conversations and compare the speakership between patients and virtual avatar. The statistics would be way more insightful if there was a specialist's diagnoses preannotation provided or any additional information about patients.

TABLE 1: Juxtaposition of the simple statistics between patients' and overall turns in the interviews

Turns	Min	Quartile I	Quartile II	Quartile III	Mean	Std	Max
Patients	42	97.5	121.5	157.75	139.5323	74.1953	386
Overall	83	178	212.5	249	226.7581	82.2847	473

TABLE 2: Patient's speakership share in total length of the interview

Turns	Min	Quartile I	Quartile II	Quartile III	Mean	Std	Max
Patients	38.93%	53.37%	57.55%	64.44%	59.12%	9.36%	81.61%

Table 1 provides simple statistics regarding patients' speakerships in the interviews. This information gives a general insight into the size of the dataset in terms of speakership turns.

The shortest (turn-wise) interview of a total of 83 turns consists of 42 patient's turns contributing to over 50% of patient's speakership share in the interview. However, the minimal patient's speakership share of 38.93% has been observed for the interview consisting of 175 turns in total; meaning that the shortest interview doesn't exactly relate to the lowest patient's speakership share in the interview. The interview with maximum length of the total amount of turns 473 is at the same time the

[1] Throughout the section, terms *turn* and *speakership* are being loosely used interchangeably but as a word of disclaimer, generally, by definition, a *turn* is bounded by speaker's potential inactivity.

interview with the highest share of patient's speakership of 81.6%. This indicates that in the majority of cases, for interviews with patients' speakership over 75%, the overall amount of turns doesn't drop below 400 turns level.

In general, no interview with patient's share under 50% exceeded the length of 200 turns in total. This indicates (more or less) that shorter interviews have a higher chance of having been done with a bit less talkative patient (turn-wise). Without very deep analyses, one can naïvely draw a pair of hypotheses that often (1) shorter interviews (turn-wise) correspond to the lower share of patient's speakership in the whole interview, whilst in the opposite situation of (2) longer interviews - the patient's speakership share tends to be higher.

On average, the share of patient's speakership in the total interview is close to 60% (around 140 turns), whilst the average interview consists of roughly 230 turns. Given the values of the 3^{rd} quartile, one can observe that the lengths of the interviews in the $\frac{3}{4}$ of all documents and the amount of patients' speakership are really close in values to the mean. This hints that the values closer to maximum (with very talkative patients) are a bit more rare cases than the others.

It's important to point out that the values in table 1 are independent of each other, i.e. minimum values in *Patients* row and *Overall* row don't necessarily refer to the same interviews (however, in this case, they do). The table shouldn't be interpreted pairwise (single columns *min*, *max*) are not obtained for the same interview - these are two independent values, even though they may turn out to refer to the same interview. To obtain the information about the ratio of patients' share to the overall length of the interview (turn-wise), one should refer to the table 2 in which values correspond to the patient's speakership within the interview.

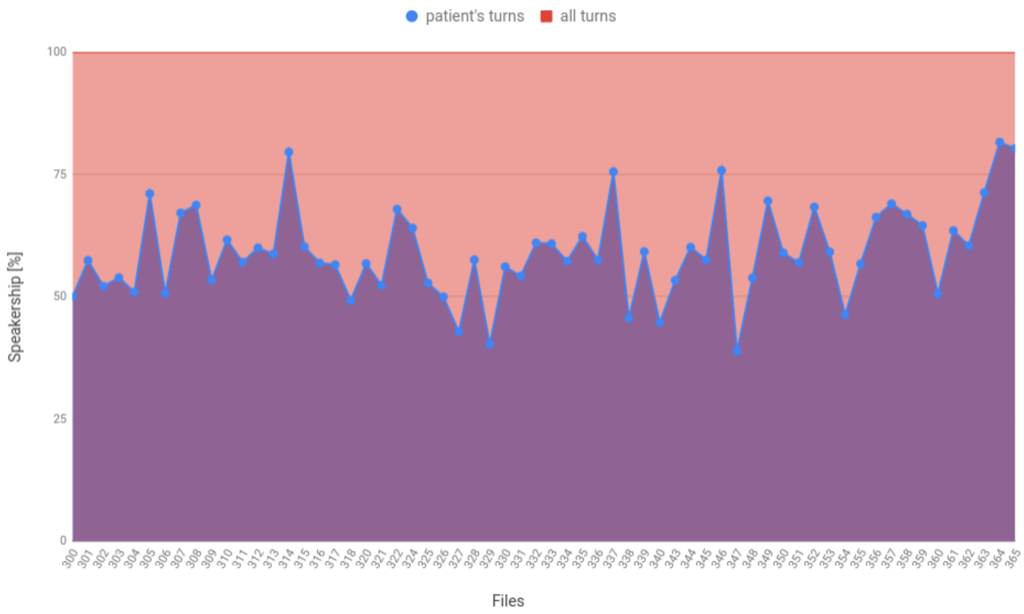

FIGURE 2: Juxtaposition of patient's speakership to the overall length of the conversation

The average token's length observed in the patients' turns is 3.6 long. Words of lengths 4, 2, 3, 5 have the biggest share among other word lengths. 4-character words make up 23.78%, 2-character - 23.26%, 3-character - 19.84%, 5-character - 4.66%. This group of the most common words' lengths altogether makes up roughly 72% of all the tokens. The average amount of tokens within a single patients' turn is 9.56, with minimal value - 1, and maximum - 125. It appears that the shorter turns are a lot more probable to occur in patients' utterances. Single token utterances make up to 19.99%, 2-token - 9.19%, 3-token - 7.21%, 4-token - 6.05%. It's important to note that many of the turns consisting of single tokens appear to be responses to commonly known *yesquestions*. We've observed 1729 of such single token utterances in the *DAIC* dataset sample. Table 3 constitutes the most common tokens found in this category of single token patients' turns.

The calculated and analyzed statistics regard solely to the $patient - avatar$ interviews. The dataset, additionally, consists of the interviews with the control group. Including comparison between the statistics obtained among real patients and the control group would be very interesting and insightful. This task will be addressed and presented in the future work.

TABLE 3: Most common tokens and their share among the category of single-token patients' turns

Token	Tokens share in the category %
um	25.56
yeah	8.16
no	8.1
uh	7.35
yes	6.83
<laughter>	4.45
mhm	3.53
so	2.78
mm	2.55
okay	1.91

6 System Proposition for the Problem

The main concern is to explore the tools and methods for automatic abstract dialog structure parsing as part of the depression detection tool. The detected objects, alongside the relations that they occur at - serve as enriched dataset features for the final classification process. We wish to prove that formal tools and stochastic methods can be used together and complement each other.

Formal symbolic methods have been used in the past to solve AI problems because of the lack of computational power (Ben-Nun and Hoefler, 2019). The trend, however, has changed in recent years, and deep learning-based methods have taken over. Either of the methods is good, but for a slightly different subset of problems. Our other task is to correctly assign the methods to the tasks to allow interpretability in the final solution.

6.1 Problem's description

On the very high-level abstraction, a specialist uses the system to help her early diagnosis regarding a patient's condition. She provides a transcription of the interview to the system. The transcript must be processed accordingly to (1) extend the current knowledge base, (2) successfully reason about the interview. Once the dialogue's structure has been obtained and the transcript has been correctly reasoned about, the output is passed to the classifier, which makes a prediction and provides an answer to the specialist. The proposed system's visualization (see figure 3) is described in detail in the 6 section.

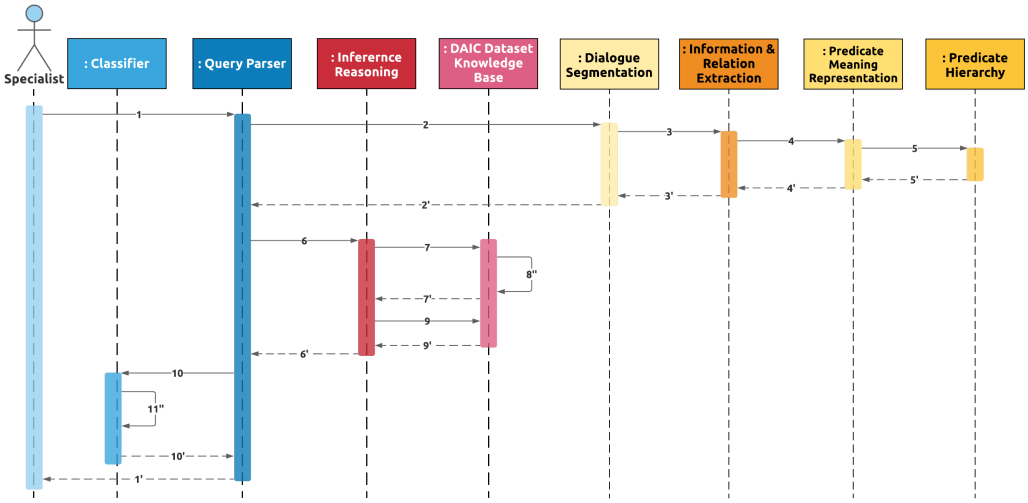

FIGURE 3: UML-like sequence diagram visualization of the system proposition

The UML-like sequence diagram (Fig. 3) visualizes the communication process between the main actors/modules in the system. Forward messages are denoted by a single number (e.g. *1* denotes the first message); responses to the messages are denoted as a single number prime (e.g. *1′* denotes a response to the first message); the recursive self-loop is denoted as double primed single number(e.g. *8″* denotes a process within the same actor/module). A simple use-case scenario reads as follows:

1. A specialist provides the system with interview transcription with a desire to obtain patients depression pre-diagnosis. The request propagates to the *Query Parser* module.
2. *Query Parser* module propagates message further, initiates the dialogue's processing by invoking the segmentation in the Dialogue Segmentation module.
3. *Dialogue Segmentation* module requests information/relation extraction from the *Extractor* module.
4. *Information & Relation Extractor* module requests meaning representation by invoking *Predicate Meaning Representation* builder, which parses dialogue's representation at this step in terms of predicates.

5. Predicates must have a special representation which allows effective reasoning; hence, *Predicate Hierarchization* builder builds a structure among predicates after being requested.
6. The responses propagate back up until they reach *Query Parser* about the successful operation and requests directly to update the knowledge in the knowledge base.
7. *Query Parser* invokes a sequence of requests-responses messages with the *Inference Reasoning* module and *DAIC Dataset Knowledge Base* leading to knowledge base update and inferring response to the query from the knowledge base.
8. The information along with the obtained structure is provided as the input to the classification model which makes a prediction and provides obtained response to back to the *Specialist*.
9. *Specialist* can then use obtained information along with her expert knowledge and make an ultimate decision about the patient's condition, or make a decision to continue examination at the following interviewing sessions.

6.2 Simplified architecture

Figure 4 represents a very high-level visualization of the system's proposition. The modules and interconnection among them are coherent with the figure 3 - system's sequence diagram.

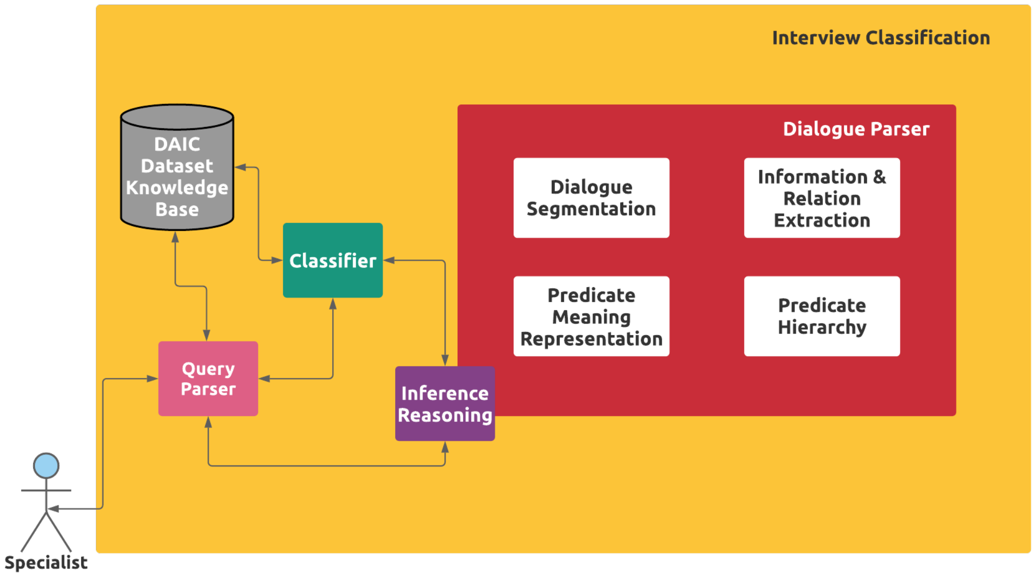

FIGURE 4: UML-like simplified system's architecture visualization

The four submodules - *Dialogue Segmentation*, *Information & Relation Extractor*, *Predicate Meaning Representation*, *Predicate Hierarchy* builders are enclosed under *Dialogue Parser* module; whereas *Inference Reasoning* module is represented as an intersection between *Dialogue Parser* and *Interview Classification* itself since it directly interacts with modules from either of the two. *Specialist* actor interacts directly with the *Query*

Parser which is interconnected with all the submodules from within *Interview Classification* and intermediate *Inference Reasoning* module.

7 Related Work

Studies on dialogue parsing include building abstract representations by using either formal languages or, especially in recent times, utilizing machine learning approaches. Formal methods allow building a comprehensive semantic representation of texts (both monologue and dialogue). (Montague, 1970) stated that human language can be interpreted in language of logic and that they can and should be based on the same principles. However, Montague could not provide a universal logical representation of discourse.

Later Discourse Representation Theory(DRT) was introduced by (Hans, Genabith, and Reyle, 1981). This theory provided formal representation of discourse with the consideration of the dynamics of language. Instead of examining inputs only sentence by sentence, it considers the sequence of sentences. It examines how the representation of new discourse affects the already processed data. DRT constructs a logical representation from which the original text could be derived. That paradigm is counted as classical formal semantics, which considers two assumptions: (1) the hearer builds the mental representation of sentence, (2) every next sentence is an addition to that representation. As further were concluded, the above assumptions cannot be true at the same time.

Following the motivation of DRT, (Asher and Lascarides, 2003) introduced Segmented Discourse Representation Theory (SDRT), which adds discourse coherence theories along with DRT. SDRT proposes discourse relations, such as Narration, Contrast, Explanation, Elaboration, Correction that are used in order to connect sentences and produce a fully coherent structure.

DIT++ provides comprehensive annotation of text with information about dialogue acts in dialogue segments, it consists of a set of 10 orthogonal dimensions to which dialogue act may belong, different relations between dialogue acts (Bunt, 2009). DIT++ was built as a extension of Dynamic Interpretation Theory (DIT) which was developed for various dialogue studies (Bunt, 1995).

Another way of representing textual utterances was presented by (Mann and Thompson, 1988), they introduced a method of representing relations in the discourse in the form of a tree by segmenting inputs, but this approach is only applicable for written text. (Shi and Huang, 2018) introduced a deep sequential model for parsing dialogues from (Asher et al., 2016) which consisted of multi-party dialogues generated by players of the game. The results of their work allowed them to predict dependency relations and construct a discourse structure jointly and alternately using deep learning methods.

(Alexandersson and Reithinger, 1997) introduced their method of dialogue structure by creating a tree-like structure with different levels: dialogue act level, turn level, phase level, and dialogue level, each level holds other corresponding classes. Regarding dialogue act annotation and classification, (Amanova, Petukhova, and

Klakow, 2016) built an automatic annotator of dialogue acts in in speakers intentions. In (Raheja and Tetreault, 2019), dialogue acts classifier using question-answer corpus was developed by using self-attention recurrent neural networks and discussed the impact of utterance-level representation learning for semantic text representation.

8 Conclusion

Depression as a primary world's disease must be detected with great precision. It's important to be able to detect it as soon as possible to treat it properly; hence, the need for an early diagnosis tool as a supporting decision for the expert's opinion. The system should be able to classify patients as either depressed or not, while simultaneously updating the knowledge base which would be used to support the decision. It's crucial to correctly process the interviews' transcripts to retrieve relevant information from it. The information doesn't necessarily have to be obvious to the reader/listener but, nonetheless, must be interpretable which is easier achievable by logical reasoning rather than blindly "believing" the neural network's output. The neural network can be utilized as an ultimate step in the classification process while given highly interpretable data representing hierarchically categorized interview's abstract structure.

In this report, we've provided the most relevant theory and methods relevant in terms of tackling the problem. Firstly, we've provided some facts and intuition around depression as a disease. Then, we've introduced key concepts of the dialogue act annotation process since it constitutes the main portion of the system. The abstract structure of the dialogue in terms of hierarchized predicates (FOPL) provides semantic information of the information by stating multi-level relations. The dialogue's structure can serve as features of the data provided as the input for the neural network classification model. In further part, we've analyzed some simple statistics from the *DAIC* dataset, mainly regarding turns and patient's speakership in the interviews. Finally, we've introduced a high-level overview of the system's proposition and described briefly the problem we're dealing with.

To successfully implement the proposed system, it's important to be comfortable with the methods presented in the report. Thanks to the high modularization of the system, the work can successfully be split between the team members. The use of the exact tools and frameworks is still to be decided; however, there are some pretendents for each task.

Once the system is working, the future work could focus on improving the tool to make it more universal, i.e. be compatible with different types of datasets - not only interviews, not exclusively regarding the depression, etc. The available *DAIC* dataset is slightly biased and limited in terms of its size; therefore, the system could be trained on a bigger dataset. Building a domain's knowledge base is a task that is being tackled along the way and could be improved separately; the increase of input dataset and use of more relevant data should drastically improve its quality.

References

Alexandersson, Jan and Norbert Reithinger (1997). *Learning dialogue structures from a corpus.*

Althoff, Tim, Kevin Clark, and Jure Leskovec (2016). *Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health.*

Amanova, Dilafruz, Volha Petukhova, and Dietrich Klakow (2016). *Creating Annotated Dialogue Resources: Cross-Domain Dialogue Act Classification.*

Amblard, Maxime, Michel Musiol, and Manuel Rebuschi (2015). "SLAM Schizophrénie et Langage: Analyse et Modélisation". In: *Journée de restitution CNRS PEPS HuMaIn.*

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders(4th edition).*

Asher, Nicholas and Alex Lascarides (2003). *Logics of conversation.*

Asher, Nicholas et al. (2016). *Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus.*

Ben-Nun, Tal and Torsten Hoefler (2019). "Demystifying parallel and distributed deep learning: An in-depth concurrency analysis". In: *ACM Computing Surveys (CSUR)* 52.4, pp. 1–43.

Bucci, W. and N. Freedman (1981). *The language of depression.*

Bunt, H. et al. (2010). "Towards an ISO Standard for Dialogue Act Annotation". In: *LREC.*

Bunt, Harry (1995). *Dynamic Interpretation and Dialogue Theory.*

— (2009). "The DIT++ taxonomy for functional dialogue markup". In: *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pp. 13–24.

Choudhury, Munmun de and Michael Gamon (2013). *Predicting Depression via Social Media.*

Dodds, Peter Sheridan et al. (2011). *Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter.*

Fellbaum, Christiane (2012). "WordNet". In: *The encyclopedia of applied linguistics.*

Fort, K. (2012). "Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus. (Annotated resources, a key issue in content analysis : towards a methodology for manual corpus annotation)". In:

Gratch, Jonathan et al. (2014). "The distress analysis interview corpus of human and computer interviews." In: *LREC*, pp. 3123–3128.

Guntuku, Sharath Chandra et al. (2017). *Detecting depression and mental illness on social media: an integrative review.*

Hans, Kamp, Josef van Genabith, and Uwe Reyle (1981). *Discourse Representation Theory.*

Jackson, Richard G et al. (2017). *Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project.*

Jurafsky, Dan and James H. Martin (2000). "Speech and language processing - an introduction to natural language processing, computational linguistics, and speech recognition". In: *Prentice Hall series in artificial intelligence.*

Mann, William C. and Sandra A. Thompson (1988). *Rhetorical Structure Theory: Toward a functional theory of text organization.*

Montague, R. (1970). *Universal grammar*.

Petukhova, V. and H. Bunt (2009). “Dimensions of communication”. In: *Neuroscience Letters*.

Raheja, Vipul and Joel Tetreault (2019). *Dialogue Act Classification with Context-Aware Self-Attention*.

Schwartz, H. Andrew et al. (2014). *Towards Assessing Changes in Degree of Depression through Facebook*.

Shi, Zhouxing and Minlie Huang (2018). *A Deep Sequential Model for Discourse Parsing on Multi-Party Dialogues*.

WHO et al. (2017). *Depression and other common mental disorders: global health estimates*. Tech. rep. World Health Organization.