1. Introduction

Medical training programs aim at enabling students to diagnose diseases correctly. Nevertheless, today up to 20 percent of patients are misdiagnosed (Berner and Graber, 2008). In order to prevent misdiagnoses and improve the quality of patient treatment, practical courses for doctor-patient interviews[1] are put into the medical study curriculum. Usually, the university pays actors to take on the role of a standardized patient with a defined *persona*, i.e., life history and disease. These actors have to memorize a persona description that contains personal details and answers to typical questions that a doctor may ask the patient in an medical interview. Even professional actors are challenged by transferring the medical student's question to the sample questions from the persona description and remembering all relevant answers. Also, actors are quite expensive and not always available for the live interview sessions.

Modern solutions for simulating doctor-patient interviews include e-learning tools like computer-based simulations for practicing interviews with virtual patients. One sample is the internet platform *USC Standard Patient Hospital*[2], where students can select a virtual patient and interview him/her by typing questions in a chat window. In such simulations, the dialogue is usually structured as a sequence of questions (asked by the trainee) followed by a response of the virtual patient, typically a predefined answer, e.g. in the form of a short video clip.

Various end-to-end systems for virtual doctor patient interviews have been proposed. Narayanan et al. (2004) has developed a multi lingual system with speech recognition for doctor-patient interviews to bridge the language barrier. They describe the components required for such system but focus on dialogue management rather than classification. (Kenny et al., 2007) build a simliar architecture but select appropriate responses by translating the question to an answer representation based on the known mappings between sample questions and answers.

However, due to the lack of training and evaluation data, many e-learning systems for this task have either relied on interfaces that only allow the selection of pre-formulated, specific questions (Manyuk, 2016) or the question is processed by inflexible, rule-based systems (Hirumi et al., 2016). Other approaches like Talbot et al. (2012) use statistical models for language classification and focus on sub-conversations within the dialogue. All of them make only limited use of dialogue context and don't conduct detailed evaluation.

The current state of e-learning tools for patient interviews still has room for improvement. Especially the classification of free-form questions entered by the medical students can be improved by better utilizing sentence and dialog context. Through a more precise and natural understanding of human language in the interviews the interaction and education quality will improve as well.

We propose novel models based on Information Retrieval and Transfer Learning techniques to enable a more natural interaction between medical student and virtual patient. Specifically, the models can understand a freely phrased question by the medical student and respond with the related video clip containing the respective answer by the virtual patient. Moreover, the proposed models take into account the question itself and the context in the current dialogue. Another important property of those models is their ability to learn from a small data set and make precise predictions.

The major contributions of this work can be summarized as follows:

- We release a novel data set with medical questions and answers in German, labeled with 62 categories that are relevant to medical interviews.
- We evaluate different models based on Information Retrieval and Transfer learning algorithms and show that they can perform well even on small amounts of training data.
- We investigate contextual features (from preceding questions) for answer retrieval, and show that they can improve the performance.

In section 2 we present existing approaches and techniques from the field of Information Retrieval and Deep Learning. Specifically we focus on approaches that work well on data sets that contain classes with just a few samples. The data set and training/test procedure is described in section 6 Next, we describe models for the baseline performance scores in section 5 Two kinds of algorithms will be used for model development and comparison, hence we will take a closer look on their comparability and establish shared metrics. The results and their discussion (Section 7) show the empirical effectiveness of the approach.

2. Related Work

The training of doctor-patient interviews is an essential part of the medical course of studies. To reduce costs and improve the training, the idea of virtual patients and interview systems has been around for several decades (Manyuk, 2016). Rule-based systems such as *NERVE* (Hirumi et al., 2016) show important components of e-learning experience in a standardized environment. Initial research on using natural language processing, and more specifically, intent recognition has been described in G.Tavarnesi et al. (2018). The system is based on two components to identify the intent and return the corresponding answer. The first component does a ranking overall intent samples based on similarity with the input, the second checks for a semantic match between the top ranked sample and the input. This match was determined by approximating the the N most important words based on frequency in the corpus and and their overlap between input and original utterance. However, no data set or metrics have been provided to indicate the performance of those components.

Information retrieval methods provide effective relevancy estimation, and the relevancy is typically approximated by the information overlap between query and document and

[1] In medical terms called Anamnesis or history taking. For consistency and simplicity, we use the phrase ”doctor-patient interviews“ in this paper.

[2] https://www.standardpatient.org/

the importance of overlapping terms (Schütze et al., 2008). Relevance estimation methods like TF/IDF (Salton and Buckley, 1988) or BM25 (Robertson et al., 2009) can also be applied to text classification, by comparing the relevance of test samples (with unknown labels) to training samples (where the labels are known, see Zhang et al. (2011) and Marcelo (2012)).

Recently, neural network based models for joint intent detection and slot filling were proposed (see (Zhang and Wang, 2016) and (Haihong et al., 2019)). Both approaches do not consider the conversation context and history as features for classification. These neural approaches require huge data sets to have enough training material.

For problems and domains, where availability of annotated training data is a problem, small data sets can still be leveraged by utilizing pre-trained language models such as BERT (Bidirectional Encoded Representations by Transformers) (Devlin et al., 2019). The models are well-suited for low resource scenarios as they are based on the principle of transfer learning. An algorithm is trained on a task A and then the same model is used to do predictions on a similar task B. Additionally, they can better encode word representations in context of a sentence. Chen et al. (2019) build a joint model for intent recognition and slot filling, showing its effectiveness for this task.

Leveraging conversation history and contextual features for classification decisions has been explored in various settings. Liu et al. (2017) presents several ways to incorporate conversational context for the classification of dialogue acts. Jin and Szolovits (2018) compute a sentence encoding from word embeddings with RNNs followed by an attention and pooling mechanism. This network is extended to encode abstract context features and this results in improvements compared to the non-contextualized variant.

3. Data set

For enabling the development and evaluation of systems for simulating doctor-patient interviews, we provide a dialogue data set in German that is annotated with one or more intent classes per utterance.

This data set was gathered in a study that investigated the assessment of diagnostic abilities in live and video simulations of medical interviews (Fink et al., 2019). Learners were in the role of a physician who conducted the medical interview and actors played patients, answering according to the case description of a specific persona. In the live simulations, the doctors and patients could freely interact with each other and all utterances were recorded on video. In the video simulations, learners selected questions from a menu that contained the same questions and categories as the coding scheme of the live simulations (see table 2).

The study employed a repeated measures design in which students where randomly assigned to either first take part in three cases in the live simulations and then three other cases in the video simulations or vice versa. Hence, the data set of this study is balanced between one group of participants who completed computer simulations before taking part in the live simulations and another group of participants who did not take part in computer simulations before taking part in the live simulations.

After data collection, the utterances of the participating medical students were transcribed by student research assistants in their sequential order. The transcripts were annotated by the assistants according to a coding scheme based on the classifications of communication strategies in doctor-patient interviews as proposed by Roter et al. (2002) and Jefferson et al. (2013). The resulting coding scheme is displayed with its major codes and definitions in table 1.

Code	Category	Definition
DQ	Questioning the patient	Direct questions to the patient.
PINF	Informing the patient	Providing general information, specific instructions, biomedical or psychosocial information to the patient.
DESC	Organizing and structuring the interview	Transitions in the medical interview, information about the structure etc.
SUM	Summarizing	Summarizing symptoms and information for the patient.
REL	Relationship building	Establishing rapport and relationship building with the patient.
MET	Meta information	Information on the medical interview, the physician and the standardized patient.

Table 1: Coding scheme of communication strategies

Important annotation guidelines for this scheme include that each utterance of a physician are coded separately as one unit and that utterances of actors displaying the patient are not transcribed and coded. Within the category *DQ* (direct questions to the patient), the coding scheme is more fine-grained and based on a content and structure analysis of history-taking forms by Bornemann (2016). According to this coding scheme for direct questions (*DQ*), questions can stem from seven different categories. The classification is illustrated with one example per category in table 2.

To generate the data set for training and testing the models, we obtain the utterances of the doctor from the raw transcripts, and filter them by their type so that only those are retained which were assigned DQ and REL (those directed to the patient, all other classes correspond to non-verbal or pre/post interview information).

Utterances with multiple classes assigned are be split into several single label instances in the training/test procedure. Specifically, each utterance u with labels $l_1, l_2, ..., l_n$ is copied into multiple instances $S_u = (u, l_1), (u, l_2), ..., (u, l_n)$. During preprocessing, we add the contextual features containing text and class of the preceding utterance in the interview.

Each entry in the dialogue sequence contains following fields:

- utterance - The text of the original utterance
- class - One of the classes assigned to the utterance
- position - The position within the dialogue sequence (e.g. 3 = third utterance from the doctor)
- previous_utterance - The text of the preceding utterance.
- previous_utterance_class - The classes assigned to the preceding utterance.

A class label consists of two parts: The symptom category and a question id. Symptom category determines the general area of symptoms, the question id the specific intent within that group. For example, *PH10* is in the category medical history (*PH*, see table 2) and the intent number 10 covers specific questions for heart diseases in child age. Seven different categories exist in the data (see table 2), but we focus on those directly related to symptoms. The other utterances belonging to category *IQ* and *OQ* are collected and covered by an artificial class *OTHER* in order to reduce complexity.

Symptom category	Code	Sample
Main symptoms	**MS**	Do you experience the complaints for the first time? [MS01]
Prior history	**PH**	Do you know of any pre-existing conditions? [PH01]
Allergies and medication	**AM**	Do you frequently have infections against which you take antibiotics? [AM01]
Social and family history	**SF**	Have your parents or other relatives of your family passed away at a rather young age? [SF01]
System review	**SR**	Has your weight changed within the last weeks? [SR01]
Inquiry	IQ	Inquiry to a question posed previously.
Other questions	OQ	Questions that were not included in the predefined set of history-taking questions.

Table 2: Symptom categories and sample questions (categories for training the models are marked in bold, samples are translated manually from German)

A small section of a medical interview from the data set is presented in table 3. A dialogue usually starts with a greeting such as ”Guten Tag” (eng.: Good day!), then the patient is asked some basic questions from the category of main symptoms (*MS*). The further the dialogue progresses the more detailed the questions become. After questions at positions 2,3,4 concerned with current pain and shortness of breath of the patient, position 5 is directed to the history of infects the patient may have had.

Pos.	Utterance	Class
1	Good morning Ms. Klein, I'm in charge of you in the ER today. What brings you to me?	OTHER
2	Do you have pain ?	MS01
3	When inhaling and exhaling ?	MS01
4	Did that come all of a sudden? Or has it been a long time since you had any shortness of breath?	MS06
5	Have you had an infection lately?	PH13
...	...	...

Table 3: Sample utterances of a medical interview

Table 4 gives an overview of the distribution within the data set and highlights some of its quantitative properties. The number of classes results from 62 symptom classes from the encoding scheme and one additional class *OTHER* which is assigned to all utterances which belong to the utterance type *REL*. A sample is one record from a stored conversation. Overall the data set contains 2627 samples distributed over 63 classes. Only 101 samples have more than one class assigned (approximately 4% of the data).

Number of classes	63
Number of samples	2627
Samples with multiple classes	101
Tokens in vocabulary	1789
Average utterance length	11.14

Table 4: Data set statistics

On spot checks we discovered the multi class samples often contain multiple questions from the same symptom category which ask for different aspects, e.g. the doctors asks for preceding diseases and allergies at the same time. Here is an example of a multi-class utterance:
Und das kam dann ganz plötzlich das Sie Luftnot bekommen haben und einen Schwindel? (eng.: ”And then all of a sudden, you got shortness of breath and felt dizzy?”)

This sample was assigned to the classes *MS06* and *MS09*. In this case, *MS* stands for main symptom category and *06/09* for shortness of breath and dizziness respectively. Sometimes the questions are combined in one sentence using ”and”. Sometimes, it is difficult to identify the part of the utterance which is referring to a specific intent if the utterance is too generic. For example, ”Mit dem Herzen irgendwas mal gewesen?” (eng.: Have you had any heart issues previously?) gets classes *PH08* and *PH10* assigned because the question about heart history is so broad both heart issues in the childhood (*PH10*) and any heart issues at all (*PH08*) are covered. The questions are phrased in various ways and often share some common nouns or verbs.
Next, we present different ways to express a question about drinking alcohol from the data set. The last sample in the following enumeration presents a multi class expression.

- *Und Alkohol? Wie viel trinken Sie da? (eng.: And alcohol? How much do you drink?)*

- *Und mit dem Alkohol? Trinken Sie ab und zu? Oder hatten Sie da was getrunken? (eng.: Do you drink from time to time? Or did you drink something?)*
- *Wie ist es mit Alkohol oder Drogen? (eng.: What about alcohol or drugs?)*

This also demonstrates the variability of length of the utterances. The distribution of samples between the classes varies heavily. The result is a high class imbalance and hence a harder classification problem. To visualize the distribution, we plotted the resulting counts as equal-width bins with width *w = 10* in figure 1.

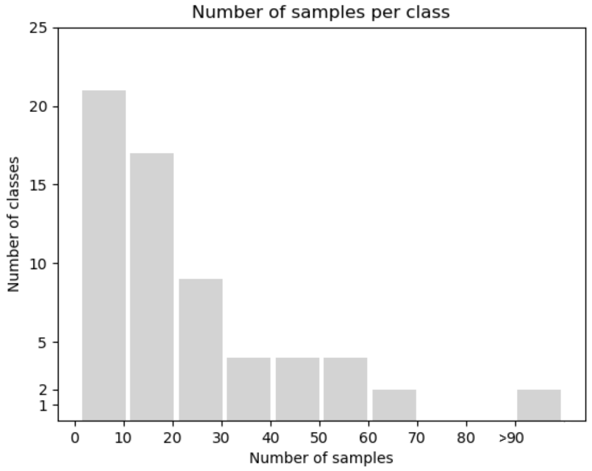

Figure 1: Grouping of classes by how many samples they cover in the data set. Most of the classes have only 10 or fewer samples assigned.

Classes with a small amount of samples (up to 20) make over half of the data set. Samples with a size of $\leq$ 10 are most frequent. Two classes contain more than 90 samples: *OTHER* and *AM02*. *Other* is a catch all class for utterances with out-of-scope labels and therefore has many samples. In fact this class contains 1321 samples overall, which shows the huge amount of utterances produced by the doctor to build a relationship with the patient.

The data set is prepared once and then used in all experiments with the same preprocessing and data split. The data is partitioned in three sets in a stratified manner: training, development and test data. There are classes that do not have enough samples for multiple splits, meaning less than three. The first set contains the data from all classes with 80% of the overall available samples. It is used to learn the weights of the networks and optimize the parameters of the information retrieval system. Second, the development data set contains about 10% of the data. With this data the model performance between each epoch is measured. The best models for testing are chosen exclusively based on the highest accuracy score on this set. The remaining 10% of the data are hidden in test set and only used for the final evaluation of an optimized model.

Due to the short and diversely distributed samples and the imbalanced semantic classes for symptoms, we see high potential in this data set for further search. The dataset was published on `Github` (Munich, 2020).

4. Evaluation Metrics

The models presented in the following sections can be categorized into two types: classification models that are optimized to predict a specific target class, and ranking models that aim at providing a ranking of all classes. In order to establish comparability, we evaluate classification approaches also in a ranking setting (by considering the ranking they induce if one orders all classes by the scores the classifier assigns to them), and we evaluate ranking approaches also in a classification setting (by taking their top-ranked class as the prediction).

For each utterance i, we denote the rank (obtained by ranking models, or by sorting classification scores) of the correct class as R_i. The Mean Reciprocal Rank (*MRR*) is computed for the entire test set of size n in the following manner:

$$\mathit{MRR} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{R_i} \tag{1}$$

Similarly, the accuracy can be computed by counting how often the top-ranked element is the correct class ($\mathbf{1}_{[\cdot]}$ denotes the indicator function):

$$\mathit{Acc} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{[R_i=1]} \tag{2}$$

5. Methods

5.1. Ranking models

Two different ranking models are proposed, built on the standard retrieval frameworks with only one modification: the incorporation of transition probabilities. Each class c in the ranking is modeled by a document d_c in the index containing all utterances for this specific class c. For the TF/IDF method, the term frequency $tf_{t,c}$ for term t and class c can be calculated from the training corpus by counting the term occurrences per class. The IDF weight for class c can be calculated as follows:

$$idf_t = \log\frac{N}{df_t} \tag{3}$$

where df_t denotes the number of utterances containing the term t and N the number of classes in the collection. Finally, when calculating the relevance score for a class c given the query q, the *term frequency* is multiplied with the *inverse document frequency* for each term in the query:

$$\mathit{TF/IDF}(q,c) = \sum_{t\in q} tf_{t,c} \times idf_t \tag{4}$$

The second ranking method, Okapi BM25, is a probabilistic scoring model for IR systems. The relevancy for class c given query q in BM25 is calculated as follows (Sparck Jones et al., 2000):

$$\mathit{BM25}(q,c) = \sum_{t\in q} idf_t \frac{df_t(k_1+1)}{df_t + k_1 \times (1 - b + b \times \frac{dl}{avl})} \tag{5}$$

where k_1, b are free parameters (b = 0.75, k_1 = 1.2) and avl is the average length and dl the current length of the combined utterances in d_c. We concatenate all sample utterances in one field of the document d_c which is later used for retrieval.

A search index is built by collecting all training samples of utterances per class c and store them in a single document d_c. Additionally, the context of the interview should be taken into account to improve the recognition performance. This is based on the assumption that questions of the medical student rely on each other, e.g. if a doctor asks the patient if he had pain already in the last weeks, then he might follow up with a question about the intensity and region of the pain on the body. In order to model that relationship the probability $p(c_i|c_{i-1})$ is estimated by counting the transitions from each class to each other in the transcripts. To work with transitions that have never been seen before (as they would introduce zero probabilities), a smoothing factor β is added the probability estimation:

$$p(c|c_{i-1}) = \frac{trans(c_{i-1}, c) + \beta}{\sum_c trans(c, c_{i-1}) + \beta N} \tag{6}$$

where c_{i-1} is the class predicted in the previous time step and c is one of the possible classes to predict at the current time step. $trans(a, b)$ is a function that counts the number of times class a transitions to b in the training data. The final ranking score is calculated as follows:

$$score(u, c, c_{i-1}) = p(c|c_{i-1}) \times rank\_score(u, c) \tag{7}$$

with $rank\_score(u, c)$ being the score of the respective ranking function scoring the relevance of the utterance u for class c. If the utterance is the first in the interview, the previous class is set to a special class symbol *START*.

5.2. Classification models

For all classification models the *Cross Entropy Loss* is calculated as the discrepancy between predictions and actual class labels:

$$L(y|x) = -log(\frac{exp(z_y)}{\sum_j exp(z_j)}) \tag{8}$$

where $z = \mathit{FFN}(\mathit{BERT}(x_u), \theta)$ with x_u being the text of the utterance of sample x and *FFN* a feed forward network with parameters θ. Adaptive Moment Estimation (Adam) is used for parameter optimization (Kingma and Ba, 2014). The classification approach requires a model that can learn even with only few samples per class and as utterances are quite short, the amount of information and features to be used for learning is limited. Pretrained language models should work well in this scenario because of their pretrained weights. Hence we modify BERT using a Feed Forward Layer on top of the last output layer. This layer produces an output for each class as a probability using softmax (see figure 2). We add 20% dropout to the output of the language model, forcing it to adapt during the training time and allowing the classification layer to be more robust against diverse inputs. Following the guidance of Peters et al. (2019) we apply fine tuning on BERT representations instead of simply extracting contextual features.

A special extension to this simple classifier are context features which can be extracted from the dialogue. Specifically, the previous utterance is is encoded through the BERT model into a vector representation and concatenated to the one produced for the current input. A feed forward layer serves as the classification layer and produces the probabilities for each class:

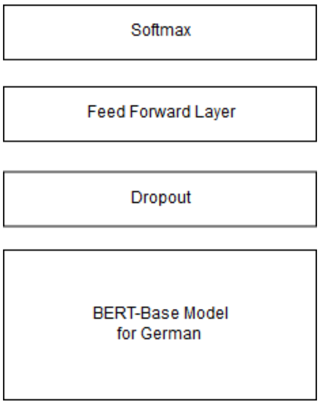

Figure 2: BERT classifier Architecture

$$\begin{gathered} score(c|u, u_{i-1}) = \\ \mathit{FFN}(\mathit{BERT}(u) \oplus \mathit{BERT}(u_{i-1}) \oplus p(c|c_*), \theta) \end{gathered} \tag{9}$$

with $\oplus$ being the concatenation operator combining all vector representations. $p(c|c_*)$ returns a vector containing transition probabilities from all classes to the particular class c. The probability for each class is calculated using a Softmax function:

$$p(c|u, u_{i-1}) = \frac{exp(score(c|u, u_{i-1}))}{\sum_i exp(score(c_i|u, u_{i-1}))} \tag{10}$$

where u is the utterance of the current and previous time step i and c_{i-1} the preceding class.

6. Experimental Setup

For comparison, we conduct initial experiments using only the pretrained model without any context features. Then the available context features are combined in the following three configurations:

1. Only preceding utterance encoding $BERT(u_{i-1})$
2. Only transition probabilities $p_{trans}(c)$
3. Both combined

The class imbalance poses a challenge to each of the models because they may start to overfit on majority classes in order to improve accuracy. To mitigate this problem, counter measures are applied at model development: Avoiding dominance of majority classes is achieved through the definition of a maximum group size S_c (number of samples per group). Downsampling can be used to randomly remove samples form the clusters which exceed S_c (Albon, 2018). The only disadvantage of this method is the chance of removing samples which are in the center of the group and can represent the class best. Through multiple data splits and following performance evaluation the impact can be measured and quantified. For all splits downsampling with $S_c = 100$ is applied because we found it works best through prelimiary experiments.

Next, L2 Regularization is employed and adjusted in multiple training runs to avoid large model weights. This regularizer adds a penalty of the squared summed update value to the loss function:

$$L(x, y) = \sum_{i=1}^{n} L(x_i, y_i) + \lambda||\theta||_2^2 \quad (11)$$

with a summed loss over all elements using the loss function L_i and θ being the update values for each parameter of the network. We also apply early stopping with a maximum number of 10 iterations according to accuracy.

7. Results

First, we compare Information retrieval approaches TF/IDF and BM25, and we study the effect of including transition probabilities into the model. The lambda value for smoothing was determined in preliminary experiments on development data. It is set to $\lambda = 1000$ for TF/IDF and $\lambda = 200$ for BM25. These λ values are very high, perhaps only a few transitions occur often and consistently in the data set. Another aspect is the assumption of a relationship between the questions which may not by very static. We leave the investigation of transitions between classes open for further research.

Model	MRR	Accuracy
TF/IDF	66.39	54.24
BM25	73.25	**62.15**
TF/IDF+TProbs	66.37	53.67
BM25+TProbs	**73.88**	**62.15**

Table 5: Performance of the IR Models on test data

Next we report the final performance scores of IR models on the test data (see table 5). The scores demonstrate that BM25 is superior to TF/IDF in this scenario. All measures are better for BM25 compared to TF/IDF and also adding the transition probabilities works slightly better for BM25. However, the transition probabilities only slightly improve the performance of the BM25 model in terms of MRR. The accuracy is not effected and does even slightly decrease for the TF/IDF model. It seems for both models that they start to overfit more on the data set when using the transition probabilities. The scores overall show that the approach is effective and in almost three quarter of the cases the prediction is correct.

Model	MRR	Accuracy
BERT	**67.08**	64.08
+ Previous Utterance	65.87	**71.35**
+ Transition Probabilities	66.02	70.31
+ Both	65.29	64.58

Table 6: Performance of BERT-based models

We report the same score for the classification models using the pretrained, German BERT model. The basic model without any context information achieves the highest MRR score. The gap to the other configurations is quite small suggesting that all configurations do not differ much by their ability to rank correct classes to the top. Regarding the accuracy, the models with context information perform better. The model using the encoded previous utterance performs best overall, the second best uses the previously used transition probabilities. If both feature vectors are used for prediction, the accuracy goes down, suggesting the model starts to overfit on the training set. All metrics are determined as class-weighted average, meaning the number of available instances per class for prediction is the weighting factor for the average, so that classes that only appear rarely in the test set should have the same impact on the result as classes with many samples.

8. Discussion

The ranking and classification models performed significantly higher than a random baseline ($\frac{1}{63} = 1,58\%$). This shows that all models can learn successfully from the data set (even though it contains only few samples for many classes) and approximate the distribution of the utterances. The best performing ranking and classification model are compared in terms of accuracy and MRR scores on test data. The accuracy is higher for the classification model, achieving the best score at 71,35% accuracy. The ranking models can achieve higher MRR scores with 73,88% as the best result. This indicates that they provide relevance rankings that are more useful across the entire spectrum of ranking positions. From a model design point of view the results suggest that if the task is concerned only with the best class, a classification model is suited better and if the task requires multiple answers, a ranking model might work better.

9. Conclusion

We investigated contextual intent recognition in doctor-patient interviews in the medical domain. A data set was created which allows feature generation from the context of the interviews. Also, semantic labels of the respective utterance content were added. The experimental results show a good performance of the Information Retrieval techniques for this kind of data. Classification models performed superior in terms of accuracy and established the state-of-the-art baseline.

Virtual patient systems like those described in Kenny et al. (2007) and e-learning tools can benefit from our proposed model by incorporating it into their dialog and classification components. The experimental results indicate that those models should recognize the correct intent most of the time and hence could trigger the correct responses and behavior of the virtual patient. Especially advanced medical students may possibly gain higher and longer-lasting diagnostic competences from such a free formulation of interview questions. Further research can use the provided data set and baseline models to investigate the task of contextual classification within medical interviews.

Acknowledgement

This work was supported by the Institute for Medical Education at the University Hospital, LMU Munich. We thank Martin Fischer, Frank Fischer, Matthias Siebeck and Victoria Reitmeier for the conception of the underlying study and their commitment. They granted permission for usage of the valuable medical interview corpus that was gathered as part of a project funded by the German Research Association (Deutsche Forschungsgemeinschaft, DFG) (Project number FOR2385). Moreover, we thank Ana Maria Semm, Renke Biallas, Johannes Kissel and Jessica Feichtmayr for their support in organizing, conducting and recording the study.

10. Bibliographical References

Albon, C. (2018). *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*. O'Reilly Media, Inc., 1st edition.

Berner, E. and Graber, M. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American journal of medicine*, 121:S2–23, June.

Bornemann, B. (2016). *Dokumentationsbögen der Inneren Medizin und der Chirurgie für Anamnese und körperliche Untersuchung für die studentische Lehre in Deutschland*. Ph.D. thesis, February.

Chen, Q., Zhuo, Z., and Wang, W. (2019). Bert for joint intent classification and slot filling.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Fink, M. C., Reitmeier, V., Fischer, F., Siebeck, M., and Fischer, M. R. (2019). Assessment of diagnostic competences with standardized patients and interactive video simulations: Results from a study on history-taking. In *Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA)*, Frankfurt, September. Gesellschaft für Medizinische Ausbildung.

G.Tavarnesi, A.Laus, R.Mazza, Ambrosini, L., Catenazzi, N., Vanini, S., and Tuggener, D. (2018). Learning with virtual patients in medical education. In *EC-TEL Practitioner Proceedings 2018: 13th European Conference On Technology Enhanced Learning*, EC-TEL 2018.

Haihong, E., Niu, P., Chen, Z., and Song, M. (2019). A novel bi-directional interrelated model for joint intent detection and slot filling. In *ACL*.

Hirumi, A., Kleinsmith, A., Johnsen, K., Kubovec, S., Eakins, M., Bogert, K., Rivera-Gutierrez, D. J., Reyes, R. J., Lok, B., and Cendan, J. (2016). Advancing virtual patient simulations through design research and interplay: part i: design and development. *Educational Technology Research and Development*, 64(4):763–785, August.

Jefferson, L., Bloor, K., Birks, Y., Hewitt, C., and Bland, M. (2013). Effect of physicians' gender on communication and consultation length: a systematic review and meta-analysis. *Journal of Health Services Research & Policy*, 18(4):242–248. PMID: 23897990.

Jin, D. and Szolovits, P. (2018). Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *EMNLP*.

Kenny, P., Parsons, T. D., Gratch, J., Leuski, A., and Rizzo, A. A. (2007). Virtual patients for clinical therapist skills training. In Catherine Pelachaud, et al., editors, *Intelligent Virtual Agents*, pages 197–210, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Liu, Y., Han, K., Tan, Z., and Lei, Y. (2017). Using Context Information for Dialog Act Classification in DNN Framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, Copenhagen, Denmark, September. Association for Computational Linguistics.

Manyuk, L. (2016). Virtual patients as the tools of professional communicative training in the higher medical education of USA. *EUREKA: Social and Humanities*, 0(5):60–68, sep.

Marcelo, M. (2012). A new term-weighting scheme for naïve bayes text categorization. *International Journal of Web Information Systems*, 8(1):55–72, January.

Narayanan, S., Ananthakrishnan, S., Belvin, R., Ettelaie, E., Gandhe, S., Ganjavi, S., Georgiou, P. G., Hein, C. M., Kadambe, S., Knight, K., Marcu, D., Neely, H. E., Srinivasamurthy, N., Traum, D., and Wang, D. (2004). The transonics spoken dialogue translator: An aid for english-persian doctor-patient interviews. *AAAI Conference on Artificial Intelligence*.

Peters, M. E., Ruder, S., and Smith, N. A. (2019). To tune or not to tune? adapting pretrained representations to diverse tasks. In *RepL4NLP@ACL*.

Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Roter, D. L., Hall, J. A., and Aoki, Y. (2002). Physician gender effects in medical communication: A meta-analytic review. *JAMA*, 288(6):756–764, 08.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). Introduction to information retrieval. In *Proceedings of the international communication of association for computing machinery conference*, page 260.

Sparck Jones, K., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management*, 36(6):779–808, November.

Talbot, T., Sagae, K., John, B. S., and Rizzo, A. (2012). Designing Useful Virtual Standardized Patient Encounters. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, Orlando, FL, December.

Zhang, X. and Wang, H. (2016). A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2993–2999. AAAI Press.

Zhang, W., Yoshida, T., and Tang, X. (2011). A comparative study of tf*idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758 – 2765.

11. Language Resource References

LMU Munich. (2020). *Transcribed Doctor Utterances from Doctor-Patient Interviews*. Med LMU, 1.0.